Dinâmica Adaptativa, Genealogias e Testes Estatísticos de Neutralidade em Evolução Molecular

Leonardo Paulo Maia

Tese apresentada ao Instituto de Física de São Carlos, da Universidade de São Paulo, para obtenção do título de Doutor em Ciências: Física Básica

Orientador: Prof. Dr. José Fernando Fontanari

São Carlos - 2004

*

*

Subversão! Caos! Sabonete!

"Fight Club", filme de David Fincher

"Dããã, dãããããã! Eu sou um sabão!
Eu sou um sabão! Eu sabo tudo!"

Cebolinha, após uma experiência
que deveria torná-lo inteligentíssimo,
para que pudesse derrotar a Mônica
(personagens de Maurício de Souza)

Agradecimentos

Sou bastante grato ao Fontanari por sua orientação. Ele sempre esteve preocupado em indicar problemas interessantes e viáveis para que eu desenvolvesse meu doutorado e, ao mesmo tempo, deixou-me livre para desenvolver minhas próprias idéias. Parece obrigação, mas nem sempre funciona assim. Obrigado pela confiança!

Faço um agradecimento especial aos colegas de sala, Alexandre Colato, pela amizade de longa data, e Giovano. Sem a ajuda de vocês, o dia-a-dia teria sido bem mais complicado.

Também merecem destaque a Dani (Botelho), Paulo PRAC, Vivi e Cláudia, pela longa convivência e trabalhos em conjunto. Mas todos que integraram o grupo do Fontanari durante meu doutorado tem participação nos resultados obtidos nesses três anos: Edvaldo, Wim, Danielle, Fábio, Fernandinho, Crepaldi, Daniel, Milton, Rosas. Esse agradecimento se estende também à Cris (que ajuda!), ao Vivaldo e ao Antônio Seridônio e, principalmente, ao Paulo PAC e ao Antônio Francisco, pelos papos e "quebras-de-galho" generalizadas.

Agradeço à FAPESP pelo apoio financeiro.

Pai, Mãe, Rafa, Xande, obrigado por toda a proteção, respeito e carinho. Amo vocês. Não há distância que nos separe.

Tá, ... obrigado, querida!

Resumo

Esta tese aborda diversos temas em evolução molecular, usando extensivamente o formalismo de funções geratrizes para obter resultados analíticos sempre que possível. Em primeiro lugar, apresenta-se a solução exata para o comportamento dinâmico de uma população infinita de seqüências infinitamente longas (não há mutações reversas) evoluindo sob a ação de mutações deletérias em um relevo adaptativo multiplicativo ou truncado.

Além disso, foi estudado o comportamento de uma população submetida a sucessivas diluições de intensidades arbitrárias, como ocorre em alguns protocolos de evolução experimental. Foram obtidas expressões matemáticas que, em princípio, podem ser úteis na caracterização de populações reais de microorganismos.

Demonstrou-se também que um processo estocástico de ramificação multidimensional generalizado é uma excelente ferramenta para analisar numericamente os efeitos da degeneração mutacional (especificamente, de um fenômeno denominado catraca de Muller) em populações sob variadas condições de crescimento exponencial.

Finalmente, simulações foram extensivamente utilizadas para analisar a história evolutiva de populações finitas e averiguar a possibilidade de certas grandezas, como certas medidas da topologia de árvores genealógicas, serem empregadas na elaboração de testes estatísticos capazes de detectar as marcas deixadas pela seleção natural.

Abstract

This thesis discusses some topics of molecular evolution, extensively using generating function methods to find analytical results whenever possible. In first place, it gives the exact solution for the dynamics of an infinite population of infinitely long sequences (no back mutations) evolving under the action of deleterious mutations on either multiplicative or truncated fitness landscapes.

In addition, the behavior of a population subject to sucessive dilutions of arbitrary intensity, just like some experimental evolution protocols, is found. The mathematical expressions, in principle, may prove useful in characterizing real populations of microorganisms.

It was also demonstrated that a generalized multidimensional branching process is a nice tool in numerically studying mutational degeneration effects (specifically a phenomenon called Muller's ratchet) in populations under a wide variety of exponential growth settings.

Finally, the evolutionary history of finite populations was studied by simulations to probe the viability of certain statistics, like some topological measures in genealogical trees, being incorporated in statistical tests to detect the fingerprints of natural selection.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Relações entre a Física e a Biologia

Nas últimas décadas do século XX, o mundo observou notáveis desenvolvimentos nas ciências médicas e biológicas. Após muitos anos de supremacia indiscutível da Física como "rainha das ciências", atraindo a maior parte dos recursos destinados à pesquisa mundial em projetos como os dos grandes aceleradores de partículas, a Biologia, merecidamente, também alcançou a condição de *big science* ("ciência grande", de grande porte), fato bem ilustrado por projetos da envergadura do Genoma, cujo objetivo inicial era seqüenciar o código genético humano, e do Proteoma, que enfoca a identificação de proteínas.

No entanto, sem desmerecer o trabalho dos profissionais das áreas biológicas e da saúde, é preciso destacar que esse progresso foi largamente baseado em desenvolvimentos tecnológicos de outras disciplinas. Toda a caracterização de proteínas, por exemplo, está baseada em métodos químicos e físicos, como a cristalografia. As modernas técnicas de diagnóstico que tanto impulsionaram a Medicina surgiram de estudos de fenômenos físicos, como a ressonância magnética. Mas as contribuições de outras disciplinas à Biologia não se restringem, em absoluto, somente a ferramentas tecnológicas.

Embora a realização do Projeto Genoma fosse impossível com os limitados recursos computacionais da década de 1970, os modernos processadores seriam extremamente subutilizados se o seu desenvolvimento não tivesse sido acompanhado pelo aprimoramento de

técnicas estatísticas e computacionais para o seqüenciamento e alinhamento do material genético. Na verdade, há uma abundância cada vez maior de dados e continuamente surgem novas possibilidades de estudos experimentais envolvendo sistemas complexos. É um fato amplamente reconhecido que projetos como o Genoma constituem apenas o primeiro passo rumo ao controle dos processos moleculares responsáveis por diversas doenças, é preciso também conhecer os mecanismos de interação entre as partes da maquinaria celular.

Portanto, acima de tudo, é preciso desenvolver técnicas para organizar e analisar toda essa informação e elaborar modelos quantitativos para tentar entender cada problema e, se possível, encontrar princípios gerais que governem a complexidade de alguns tipos de sistemas. Essa integração com técnicas quantitativas (em um sentido bem geral) é fundamental nas previsões dos futurólogos de ciência que afirmam que o século XXI será "da Biologia". Um recente fascículo da revista *Science*, do último 27 de fevereiro, enfatiza em seu editorial [92] e traz vários artigos sobre a "nova onda", demonstrando que esta é uma época de verdadeira efervescência, em que os currículos dos cursos estão sendo completamente reformulados e há uma explosão de pesquisa, extensamente baseada no trabalho conjunto dos profissionais de formação biológica com estatísticos, cientistas de computação, matemáticos e ... físicos! Os físicos teóricos naturalmente adquirem habilidades variadas na tentativa de compreender sistemas bem complicados. Acredita-se que uma forte cooperação entre a Biologia e a Física, em particular, pode levar a uma verdadeira revolução científica nos anos vindouros [1].

Muitos físicos estatísticos, em particular, sentem-se bastante atraídos pelo campo da evolução molecular, haja vista que técnicas como equações de difusão, relevos de energia, autômatos celulares e simulações computacionais de agentes em interação são utilizadas em ambas as áreas, entre outras semelhanças [79]. Entendendo-se evolução simplesmente como descendência com modificação, é claro que a evolução molecular apresenta uma considerável superposição com a genética de populações, que estuda os processos e mecanismos inerentes às mudanças responsáveis pela diversidade genética entre representantes de uma mesma espécie ou entre espécies diferentes, constituindo-se no fundamento de toda

a biologia evolucionária. O adjetivo molecular indica que os objetos de estudo não são macro-espécies como *Homo sapiens* (homem) ou *Drosophila melanogaster* (uma mosca) e sim variantes dentro de uma certa população de genes.

Os resultados apresentados nesta tese constituem-se em mais uma incursão de um físico estatístico na evolução molecular. Dessa forma, antes de introduzir os conceitos relevantes e apresentar o conteúdo deste trabalho, é conveniente oferecer alguma perspectiva histórica acerca da evolução biológica. Os leitores que eventualmente desconheçam alguns termos técnicos biológicos utilizados no texto podem encontrar auxílio no glossário que se encontra no fim da tese.

1.2 Um panorama da teoria da evolução e da genética de populações

1.2.1 Desenvolvimento histórico

A primeira teoria sistemática da evolução foi elaborada pelo naturalista francês Jean Baptiste Lamarck (1744-1829). Embora sua idéia de que as características que um organismo adquirisse pelo uso pudessem ser transmitidas aos seus descendentes tenha se mostrado completamente equivocada, Lamarck deixou uma importante contribuição, pois introduziu o conceito de adaptação dos indivíduos ao meio em que vivem, fundamental para os desenvolvimentos posteriores.

Quando Charles Darwin (1809-1882) concebeu sua teoria da seleção natural no século XIX, ele jamais poderia imaginar as possíveis aplicações de sua descoberta. Afinal, embora o autor de "A Origem das Espécies" (1859) tivesse plena consciência das revolucionárias implicações científicas, filosóficas e religiosas de seu trabalho, a própria idéia de átomo era considerada uma heresia científica pela maior parte dos pesquisadores daquela época. Mesmo assim, sua idéia de que modificações que representem maior chance de sobrevivência e/ou fertilidade tendam a tornar-se cada vez mais presentes em uma população, devido à vantagem competitiva de seus portadores e à sua transmissão à prole, ainda é o

fundamento de toda pesquisa em evolução. Esse aprimoramento da população via seleção natural é denominado adaptação evolucionária, e sempre ocorre de uma forma condicionada, dependente das condições ambientais. Para fazer justiça, é importante lembrar que, independente e simultaneamente a Darwin, Alfred Russel Wallace (1823-1913) chegou a conclusões semelhantes. Eles chegaram a apresentar conjuntamente seus resultados à Sociedade Lineana de Londres, em uma histórica reunião em 1858. Entretanto, como Darwin dispunha de uma quantidade muito maior e mais detalhada de observações que Wallace, aquele foi considerado o principal autor da idéia da seleção dos mais aptos.

O mecanismo responsável pela herança genética foi descoberto por um contemporâneo de Darwin, o monge Gregor Mendel (1822-1884), embora seu artigo de 1865 tenha permanecido desconhecido até 1900. Estudando variações fenotípicas em experimentos com ervilhas realizados entre 1856 e 1863, ele concluiu que a informação genética era transmitida segundo um mecanismo de herança *discreta* das características paternas. As ervilhas de Mendel eram verdes ou amarelas, lisas ou rugosas, não havia meio termo. Em linguagem moderna, pode-se dizer que Mendel mostrou que os dois alelos de cada gene se separam na formação dos gametas, em um processo denominado segregação.

Quando redescoberto, o trabalho de Mendel não teve aceitação imediata por boa parte dos biológos evolucionistas, apesar da importância hoje atribuída a alguns desenvolvimentos a ele relacionados, como o teorema de Hardy-Weinberg (1908). Enquanto a herança discreta mendeliana enquadrava-se perfeitamente no saltacionismo, corrente que defendia que as mudanças evolucionárias ocorreriam via "saltos" de magnitude apreciável, ela era tida como incompatível com as idéias do gradualismo, defendidas pelo próprio Darwin, pelas quais as características fenotípicas deveriam ser medidas em uma escala contínua e apresentariam variações suaves e praticamente imperceptíveis. Essa visão parecia mais adequada quando se analisava características como altura e peso, e tinha afinidade com a noção intuitiva de um filho ser a "mistura de seus pais". Um dos maiores defensores do gradualismo (apesar de uma certa ambigüidade em alguns momentos) foi Francis Galton (1822-1911), um primo de Darwin, que foi um pioneiro na aplicação de técnicas estatísticas à Biologia e conseqüentemente considerado o fundador da Biometria.

Na verdade, as idéias de Galton e Mendel não eram incompatíveis. De fato, a relação entre genótipo e fenótipo é obscurecida pela complexa interação entre genes e ambiente na determinação da fisiologia, desenvolvimento e comportamento. A complexidade é ainda maior na biologia evolucionária porque o ponto principal é a habilidade *relativa* dos organismos em sobreviver e reproduzir-se em seus ambientes. Embora alguns trabalhos tenham abordado essa questão na primeira década do século XX e passado despercebidos, considera-se que o começo da reconciliação teórica entre o gradualismo e o saltacionismo foi determinada por um artigo de 1918 do criador da Estatística moderna, Ronald Aylmer Fisher (1890-1962), que mostrou que um modelo com múltiplos fatores mendelianos interagentes era capaz de gerar variações contínuas em um fenótipo. Matematicamente, o que se observa é apenas uma manifestação do teorema do limite central [1] , razão pela qual as distribuições observadas empiricamente por Galton não somente eram contínuas, mas tinham a famosa forma de sino de uma curva gaussiana.

Na verdade, o casamento entre o darwinismo clássico e o mecanismo de herança genética mendeliana foi o resultado de três décadas de pesquisas empíricas e teóricas, dando origem à teoria sintética da evolução, também conhecida como neo-darwinismo. O período que se estende do fim da década de 10 até o fim da década de 40 é conhecido como a "era dourada" [30, 94] da genética de populações e foi fundamental para que essa disciplina adquirisse um conjunto sistemático e sofisticado de princípios matemáticos, exclusivamente devido aos esforços de Fisher, Sewall Wright (1889-1988) e John Burdon Sanderson Haldane (1892-1964). Experimentalmente, destacam-se os trabalhos decisivos de geneticistas como Theodosius Dobzhansky (1900-1975) e Ernst Mayr (1904-), demonstrando a base mendeliana da variação genética, tanto contínua quanto descontínua, e a impossibilidade da transmissão das características adquiridas.

Entretanto, nos anos 50, ainda havia muito pouca evidência empírica acerca da intensidade da variação genética. Havia consenso quanto à proeminência da seleção natural

[1] Principal teorema da teoria de probabilidades, segundo o qual a distribuição de uma soma de variáveis aleatórias quaisquer, mas com médias e variâncias finitas, aproxima-se de uma distribuição normal à medida que o número de variáveis cresce.

entre as forças evolucionárias, mas os geneticistas dividiam-se entre aqueles que defendiam seu papel purificador, removendo mutações deletérias e reduzindo a variabilidade, e os defensores de uma seleção de balanço. Aqueles, os partidários da escola clássica, acreditavam que os membros de uma população seriam homozigotos na grande maioria dos seus *loci* (um alelo comum para cada *locus*). A escola do balanço, por outro lado, defendia que os indivíduos são heterozigotos em uma parcela considerável dos *loci* gênicos. Desenvolvimentos experimentais, como o seqüenciamento e a eletroforese de proteínas, resolveram essa questão ao evidenciarem grande variabilidade não apenas entre espécies mas também dentro de populações. Contudo, surgiu um problema muito mais sério, até hoje não resolvido.

Simplificando bastante, pode-se dizer que a variabilidade observada era bem maior do que a esperada, em um nível aparentemente inconsistente com a existência de seleção natural. Nesse momento, destaca-se o pesquisador japonês Motoo Kimura (1924-1994) lembrado, acima de tudo, por ser um dos criadores [98] (independente e simultaneamente a J. L. King e T. H. Jukes [104]) e, sem dúvida, o maior divulgador da teoria da evolução neutra. Essa teoria afirma que a maior parte das mutações é seletivamente neutra ou aproximadamente neutra [101], de modo que a deriva genética, e não a seleção natural, seria a principal responsável pelo polimorfismos inter e intraespecíficos. Essa hipótese é controversa, pois embora realmente haja maior variabilidade genética do que seria esperado na presença de seleção natural em certos sistemas e saiba-se atualmente que apenas cerca de 4% do genoma dos mamíferos codifica proteínas, aparentemente dando ampla margem para a ocorrência de mutações neutras, os defensores da seleção também têm dados a seu favor [67, 144]. Mas Kimura desempenhou um papel incontestável [116] na unificação da evolução molecular com a genética de populações, criando modelos que certamente ainda serão usados por muitos anos.

Um importante contemporâneo de Kimura cuja obra permaneceu desconhecida por muito tempo foi o francês Gustave Malécot (1911-1998). Depois deles, matemáticos como Karlin, Ewens, Watterson, Moran e Nagylaki fizeram contribuições relevantes em modelos de muitos genes, dinâmica espaço-temporal de populações e modelos genéticos estocásticos

diversos. Alguns deles também anteciparam [108, 171] argumentos que viriam a fazer parte do conjunto de técnicas conhecido como teoria do coalescente, elaborado e sistematizado [107, 105, 106] por J. F. C. Kingman um pouco antes e independentemente de R. R. Hudson [87] e F. Tajima [169]. Essa metodologia permitiu grandes progressos no estudo do polimorfismo genético em populações nas duas últimas décadas e baseia-se em duas propriedades do regime neutro de notável simplicidade mas importantes conseqüências.

Em primeiro lugar, é possível separar o processo de mutação do processo genealógico quando as mutações são neutras. Afinal, por definição, tais variações não afetam o sucesso reprodutivo de seus portadores. Dessa forma, a evolução de uma população poderia ser simulada sem incluir a incidência de defeitos e, ao término do processo, as mutações poderiam ser distribuídas ao longo dos ramos da árvore genealógica, embora esse procedimento não seja particularmente útil por envolver a simulação de toda a população, como a formulação original. Neste aspecto, destaca-se a segunda propriedade, pela qual é possível descrever a genealogia de um grupo de indivíduos de frente para trás (*backwards*) ao longo do tempo, sem considerar o restante da população. Geração por geração, a árvore genealógica desse grupo é rastreada para detectar coalescências entre linhagens até que o ACMR seja encontrado, quando então retoma-se o sentido rotineiro da evolução temporal para implementar o processo de mutação.

Essa abordagem leva diretamente a algoritmos extremamente eficientes, que descartam a maior e irrelevante parte da história evolutiva. Mas, acima de tudo, trata-se de uma teoria elegante e de grande valor heurístico, que afeta profundamente a forma como se analisa a dinâmica evolucionária no regime neutro. O coalescente possibilita o cálculo das probabilidades de configurações amostrais sob diversos modelos de genética de populações (mesmo modelos neutros mais complexos freqüentemente convergem para os resultados de Kingman) e análises de verossimilhança de dados de polimorfismo [73, 113, 167, 172]. Como não poderia deixar de ser, esse sucesso tem estimulado a busca de uma metodologia semelhante para descrever genealogias sob seleção [89, 93, 112, 139], embora os resultados ainda não sejam tão expressivos. Embora por muito tempo tenha sido difícil encontrar discussões adequadas na literatura [43, 88], alguns artigos de revisão bem recentes [59,

130, 141, 158], enfocando diferentes aspectos da teoria, oferecem uma boa perspectiva da relevância do coalescente.

1.2.2 Aplicações modernas e especulações sobre o futuro

Ao longo do tempo, a teoria da evolução começou a separar-se da genética de populações e hoje ela não somente se constituiu em uma linha de pensamento que unifica todas as áreas da Biologia, como também permeia diversas áreas do pensamento moderno. Discussões gerais acerca do significado da evolução e suas implicações podem ser encontradas nos livros de Richard Dawkins [32, 33, 34] e Daniel C. Dennett [38].

A perspectiva evolucionária deu origem a fortes controvérsias entre cientistas sociais, psicólogos e etologistas quanto à natureza dos aspectos comportamentais em animais e humanos, considerando-se questões como a seleção sexual de parceiros [29, 129] e o surgimento e permanência do altruísmo [154]. Termos como sociobiologia [188, 189] e psicologia evolucionista [191] estão cada vez mais difundidos. Alguns desses problemas começaram a ser passíveis de análises quantitativas quando o biólogo John Maynard Smith, modelando o comportamento de animais em disputas, generalizou [125] a teoria econômica de jogos de John Von Neumann and Oskar Morgenstern [175]. Ao introduzir o conceito de estratégia evolucionariamente estável, o cientista inglês deu origem à teoria evolucionária de jogos, que revolucionou o estudo do comportamento, inclusive na área econômica, que já viu mais de um prêmio Nobel contemplar esse campo de pesquisa.

Por outro lado, é praticamente consenso que as esperanças do homem poder compreender melhor o funcionamento de sua mente residem na união [150] entre as ciências cognitivas [64] e o pensamento darwiniano. O desenvolvimento da inteligência artificial enfoca cada vez mais as técnicas de otimização e computação bio-inspiradas, como os algoritmos genéticos [68] e demais algoritmos evolucionários.

Uma aplicação um pouco mais relacionada a este trabalho é a recuperação da história de populações com base na análise da estrutura molecular de seus membros. A nova área da filogeografia [110] baseia-se no fato de que eventos demográficos e biogeográficos muito antigos, como migrações ou fenômenos naturais que extinguem muitas espécies, podem ser

inferidos a partir da composição genética de populações. Dessa forma, algumas hipóteses históricas podem ser testadas de uma forma objetiva, potencialmente esclarecendo questões acerca do surgimento de novas espécies em um ecossistema ou sobre a co-evolução de microorganismos com o homem, um tema de interesse tanto histórico [40] quanto de saúde pública. Em tempos de bioterrorismo e na iminência de uma pandemia global de gripe, como a que vitimou mais de 20 milhões de pessoas em todo o mundo em 1918 [65], o estudo da dinâmica evolucionária de vírus e bactérias, de forma geral, certamente pode dar significativas contribuições à epidemiologia.

Além disso, nas próximas décadas grandes avanços serão obtidos na integração de dados acerca de taxas e padrões de evolução genômica e fenotípica. Afinal de contas, as diferenças entre espécies devem-se a milhares de mutações acumuladas ou a algumas poucas (dezenas?) mutações reguladoras? Por outro lado, será possível que algum dia o homem possa criar formas de vida artificiais? É incrível que experimentos com organismos digitais [2, 115, 184, 187] já permitam especulações acerca da ecologia, genética e evolução dessas eventuais formas de vida e possam levar a uma melhor compreensão da vida orgânica. Questões fundamentais como essas certamente serão alvo de muitos esforços no futuro próximo e inspirarão novas idéias acerca da vida e da evolução.

1.3 Conteúdo e organização da tese

Este volume discute detalhadamente alguns problemas em evolução molecular analisados pelo autor durante os 3 anos do seu doutoramento. O capítulo 2 apresenta vários conceitos e modelos que são utilizados em todo o resto da tese, constituindo-se em uma breve introdução à evolução molecular pelo olhar de um físico.

No capítulo 3, mostra-se que o comportamento de uma população de tamanho infinito, constituída por indivíduos sujeitos à ação de mutações deletérias e seleção, pode ser determinado exatamente quando cada indivíduo é representado por uma seqüência infinita de genes (de modo que não haja mutações reversas) e se a seleção atuar de forma multiplicativa ou truncada. Embora a solução de equilíbrio, no caso multiplicativo, já fosse

conhecida há mais de vinte anos, sua contrapartida dinâmica ainda era desconhecida. No caso truncado, o autor não encontrou na literatura expressões analíticas para a concentração de equilíbrio das seqüências mutantes e nem para o comportamento dinâmico de toda a população. Esses resultados também são discutidos nas referências [121] e [120], respectivamente.

Os capítulos seguintes são dedicados a problemas em que os efeitos de flutuações estocásticas são importantes. Isso é demonstrado claramente no capítulo 4, dedicado à generalização de dois modelos, concebidos por A. Colato e J. F. Fontanari, em que a variabilidade no tamanho da população é essencial na caracterização da degeneração mutacional. No primeiro, uma população é submetida a transferências seriais arbitrárias, enquanto, no segundo, um processo de ramificação generalizado foi utilizado para estudar numericamente os efeitos da catraca de Muller em populações sob crescimento exponencial.

No capítulo 5, simulações computacionais foram extensivamente utilizadas para estudar várias características de populações finitas usadas em testes estatísticos capazes de detectar as marcas deixadas pela seleção natural, denominados testes de neutralidade. Esse estudo exigiu a construção computacional de genealogias para analisar alguns testes baseados no polimorfismo genético em seqüências de DNA e outros elaborados a partir de características topológicas das genealogias [122].

As conclusões finais são apresentadas no capítulo 6, sendo seguidas por apêndices que abordam tópicos diversos, aprofundando algumas discussões ou simplesmente introduzindo conceitos presumivelmente desconhecidos de alguns leitores. Mas é preciso ressaltar que o último apêndice é um pequeno glossário de termos biológicos, que pode ser muito útil.

Capítulo 2

Conceitos básicos

2.1 Introdução

Vários conceitos e modelos são utilizados em mais de uma ocasião nesta tese. O objetivo deste capítulo é apresentar esses tópicos de uma forma integrada, assumindo pouco conhecimento prévio de evolução e genética de populações. Dessa forma, este capítulo constitui-se em uma introdução à evolução molecular, enfatizando modelos simples de seleção e mutação. Recombinação [55] e fenômenos demográficos complexos, como migrações e subdivisão de populações, não são abordados nesta tese, embora sejam fatores reconhecidamente importantes na preservação da variabilidade genética.

Obviamente, só é possível adquirir conhecimento aprofundado na área mediante a leitura dos livros textos adequados. Os fundamentos matemáticos da genética de populações clássica são discutidos detalhadamente nos clássicos [53] e [31]. Para uma visão geral com ênfase na interpretação biológica, recomendam-se os livros [77] e [117], que também abordam os desenvolvimentos mais recentes, embora de forma apenas semi-quantitativa.

Por outro lado, alguns artigos foram escritos justamente com o intuito de introduzir físicos ao tema da evolução biológica. As relações entre o conceito de relevo adaptativo e a física estatística dos sistemas desordenados foram discutidas em dois trabalhos [146, 147]. E enquanto a literatura sobre modelos de coevolução e macroevolução, além do papel de relevos adaptativos em evolução molecular, foi revista em [45], a referência [7] oferece uma

perspectiva dos fundamentos da genética de populações, destacando também trabalhos recentes em modelos de mutação e seleção.

2.2 Forças evolucionárias

O estado de uma população é o resultado da interação de várias forças evolucionárias. De fato, a diversidade genética observada em um sistema deve-se aos efeitos conjuntos de processos elementares como mutações, seleção e deriva genética.

Antes do advento do neodarwinismo, acreditava-se que a teoria da seleção natural enfrentava o sério desafio de explicar a grande diversidade biológica observada na natureza. Afinal, a seleção irrefreavelmente tende a aumentar a participação dos indivíduos mais aptos na composição de uma população. Esse problema começou a dissipar-se quando se descobriu que o mecanismo de herança discreta não somente permite que a variabilidade genética seja mantida na ausência de seleção, como faz com que a escala de tempo envolvida na eliminação de alelos pior adaptados possa ser da ordem de centenas de gerações [53].

Portanto, na presença de mecanismos capazes de introduzir variação em uma população em intervalos de tempo relativamente curtos, a seleção não é capaz de eliminar a diversidade populacional. A inevitável ocorrência de mutações é um desses mecanismos. Diversos modelos apresentam um estado de equilíbrio estável, denominado **balanço mutação-seleção**, para a composição de uma população sob a ação dessas duas forças antagônicas.

Na verdade, há muitos outros fatores envolvidos nessa questão. Por exemplo, enquanto freqüências alélicas não se alteram em populações infinitas (ou suficientemente grandes, depende do problema) em equilíbrio, elas podem sofrer fortes flutuações devido a erros de amostragem em populações finitas. Esse fenômeno chama-se **deriva genética** e pode levar à perda de algumas classes de indivíduos, especialmente aquelas menos numerosas, reduzindo a variabilidade populacional.

Por outro lado, não é razoável esperar que um modelo elaborado para descrever uma

população de tamanho constante seja válido irrestritamente. Um modelo com essa característica pode ser útil para analisar um sistema em escalas de tempo curtas, comparadas ao tempo necessário para o número de indivíduos variar apreciavelmente, ou para estudar algumas propriedades específicas do sistema modelado. Portanto, dependendo do objetivo do estudo, pode ser preciso considerar explicitamente modelos de populações com tamanhos variáveis, o que também afeta a diversidade genética.

2.3 Modelos de mutação

2.3.1 Modelo de seqüências

Genericamente, denomina-se mutação qualquer alteração em um trecho de material genético e que, portanto, pode ser transmitida a um descendente. Tais alterações podem ocorrer, por exemplo, pela remoção ou inserção de novos segmentos na seqüência original. Esses tipos de mutações são importantes, haja vista que podem alterar completamente o ordenamento dos aminoácidos em uma proteína. Mas as mutações mais estudadas, tanto por relevância quanto por conveniência na modelagem, são as pontuais, em que alguns nucleotídeos isolados são substituídos quando uma nova seqüência é copiada incorretamente a partir da original, mas o número total de nucleotídeos permanece constante.

Nessas condições, é conveniente introduzir um modelo de seqüências, em que genótipos são identificados com arranjos lineares de L sítios. O estado σ_i de cada sítio i é representado por um dos símbolos pertencente a um conjunto V_i, denominado alfabeto. Na maior parte dos casos, o mesmo alfabeto, V, aplica-se a todos os sítios e o genótipo $\boldsymbol{\sigma}$ pertence ao espaço L-dimensional V^L.

No contexto molecular, cada sítio representa um nucleotídeo em uma seqüência de DNA, em que o alfabeto seria constituído pelas bases nitrogenadas adenina (A), guanina (G), citosina (C) e timina (T), ou RNA, em que a timina é substituída pela uracila (U). Muitas vezes, distinguem-se apenas as purinas (A e G) e as pirimidinas (C e T), e um alfabeto binário $V = \{0, 1\}$ revela-se suficiente. Neste caso, V^L é denominado espaço de

seqüências.

Mas há uma outra interpretação possível [7] para o termo sítio na discussão acima. No chamado contexto clássico, o termo genótipo refere-se a um conjunto de L genes. Assim, cada sítio é identificado com um gene, ou *locus* gênico, e o alfabeto, em princípio, seria o conjunto de todos os possíveis alelos. Contudo, essa escolha tornaria o modelo intratável, sem falar na inadequação do uso do mesmo alfabeto para todos os sítios. Dessa forma, muitas vezes é conveniente adotar mais uma vez um alfabeto binário, em que símbolos indicam apenas se um certo sítio está ocupado por um mutante ou pelo alelo que ali se encontrava originalmente, denominado alelo selvagem (*wildtype*).

Em ambos os casos, costuma-se impor que os sítios sofram mutações de forma independente. Além disso, em um modelo sem superposição de gerações, também é comum admitir que, a cada passo temporal, haja uma mesma probabilidade u de que uma mutação ocorra em cada sítio. Se o alfabeto V contiver A símbolos e um sítio, ao sofrer uma mutação, assumir aleatoriamente um novo elemento de V, então a probabilidade de um genótipo passar do estado $\boldsymbol{\sigma}$ ao estado $\boldsymbol{\sigma}'$ em uma geração é

$$P(\boldsymbol{\sigma} \to \boldsymbol{\sigma}') = \left(\frac{u}{A-1}\right)^{d(\boldsymbol{\sigma},\boldsymbol{\sigma}')} (1-u)^{L-d(\boldsymbol{\sigma},\boldsymbol{\sigma}')}, \tag{2.1}$$

em que $d(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ é o número de sítios que distinguem os dois genótipos, denominado **distância de Hamming**. Claramente, a distância de Hamming é simétrica em relação aos genótipos e conseqüentemente o mesmo ocorre com a probabilidade de transição acima.

2.3.2 Infinitos alelos e infinitos sítios

Apesar da simplicidade do modelo de seqüências, em muitos casos faz-se necessária uma formulação alternativa. Um bom exemplo é a situação realística em que há um número enorme de possíveis alelos mutantes em comparação com o único alelo (ou com os membros de uma pequena classe de alelos mais aptos) originalmente presente em cada sítio, de modo que é improvável que um sítio ocupado por um mutante venha a perder essa condição.

Em 1964, Kimura e James F. Crow conceberam um modelo para descrever um *locus*

gênico em que cada nova mutação dá origem a um alelo nunca antes presente na população, sem considerar qualquer estrutura molecular explícita. Na linguagem desta seção, isso corresponderia a adotar o contexto clássico e fazer $L = 1$ e $A \to \infty$ na Eq. (2.1), o que mostra apenas que um alelo permanece inalterado com probabilidade $1 - u$ mas que não é possível fazer previsões sobre o estado final de qualquer transição. Com base nesse **modelo de infinitos alelos** [102], eles previram os efeitos conjuntos das mutações e da deriva genética sobre a heterozigosidade (proporção de genótipos heterozigotos) de uma população finita de diplóides. Dessa forma, o modelo de Kimura e Crow naturalmente sugeria um **teste de neutralidade**, pois outras forças evolucionárias, além da deriva e da pressão mutacional, deveriam ser responsabilizadas caso um gene apresentasse heterozigosidade diferente da prevista.

Entretanto, o modelo de infinitos alelos não era capaz de incorporar a informação contida no grau de similaridade entre alelos distintos. Em 1969, Kimura desenvolveu um novo modelo [99] para suprir essa deficiência. O pesquisador japonês notou que, se a probabilidade de mutação por sítio u for pequena em relação ao comprimento L de uma seqüência, poucos sítios sofreriam mutações a cada geração e seria necessário muito tempo até que um sítio já alterado sofresse uma nova mutação. Além disso, esse comportamento é tanto mais intenso quanto maior for L (a probabilidade de uma mutação ocorrer em um certo sítio é $1/L$), sugerindo os limites $L \to \infty$ e $u \to 0$. O termo **modelo de infinitos sítios** surgiu apenas em 1971 [100].

É importante notar que, quando se considera uma coleção dessas longas seqüências, o comportamento acima descrito aplica-se a toda a população, de modo que é muito improvável (zero, quando $L \to \infty$) haver mais do que uma mutação em um conjunto de sítios homólogos (um em cada seqüência, todos na mesma posição). Dessa forma, tendo em vista a homologia, a maioria dos sítios é monomórfica e todos os sítios polimórficos são segregantes para apenas dois símbolos, o original e um mutante. Esse padrão é justamente o que se observa nos dados de variação alélica em seqüências de DNA. Embora essa seja a aplicação mais direta e importante do modelo de infinitos sítios, ele não é incompatível, em absoluto, com a visão clássica do modelo de seqüências.

É claro que há uma certa semelhança entre esses modelos. Afinal, se no caso de infinitos sítios cada nova mutação ocorre em um sítio até então monomórfico, não pode haver mutações reversas e sempre surgem seqüências inéditas, como no caso de infinitos alelos. Portanto, o modelo de infinitos sítios reproduz os resultados do seu antecessor mas, como deve ter ficado claro, possibilita estudos bem mais gerais. Embora qualquer transição entre configurações bem definidas tenha probabilidade nula, como no caso dos alelos, no modelo de infinitos sítios é possível pensar em transições entre *classes* de estados. Como já foi dito anteriormente, em algumas situações basta utilizar um alfabeto binário e distinguir entre sítios mutantes e não mutantes. Além disso, em muitos casos, um genótipo é classificado pelo número de mutações que carrega consigo. Esse procedimento também é útil quando as seqüências são finitas, pois reduz o número de possíveis configurações de 2^L a $L+1$, quando o alfabeto é binário.

No apêndice A, mostra-se que o modelo de seqüências pode ser usado para calcular a probabilidade de um certo genótipo com k defeitos se transformar em *algum* outro com k' defeitos. Além disso, prova-se também que, se $uL \to U$ quando $L \to \infty$ e $u \to 0$, não há mutações reversas e que o acréscimo n na carga mutacional durante a transição é dado por uma variável aleatória com distribuição de Poisson e valor médio U, com probabilidade

$$M_n = e^{-U}\frac{U^n}{n!}. \tag{2.2}$$

2.4 Modelos de seleção: relevos adaptativos

Existem inúmeros tipos de seleção [77, 117], distinguindo a atuação do mesmo princípio, "sobrevivência do mais apto", em diferentes situações. Uma discussão detalhada desses tópicos foge ao escopo desta tese. O único conceito realmente relevante para os desenvolvimentos posteriores neste trabalho é o de **valor adaptativo** (correspondente a *fitness*, segundo [62]) de um indivíduo a um certo ambiente, que é uma medida da capacidade conjunta de sobrevivência (viabilidade) e reprodução (fecundidade) do organismo em dadas condições, um "sucesso relativo esperado".

Em modelos de tempo contínuo, o valor adaptativo normalmente é associado à dife-

rença entre as taxas de nascimento e morte do indivíduo. Quando o tempo é discreto e não há superposição de gerações, muitas vezes o valor adaptativo é dado pelo tamanho médio da prole que um organismo deixa para a geração seguinte. Isso não quer dizer que a viabilidade seja desconsiderada, pois dois indivíduos com a mesma fecundidade podem ter chances de sobrevivência distintas e, conseqüentemente, aquele com maior viabilidade tem maior número médio de descendentes. Esse conceito de valor adaptativo foi adotado por S. Wright e, atualmente, é denominado valor adaptativo absoluto. Na verdade, excetuando-se casos como aqueles em que uma população inteira pode se extinguir, a seleção depende apenas da adaptação *relativa* de um indivíduo. Portanto, em geral, a dinâmica evolucionária não é afetada se os valores adaptativos absolutos dos membros de uma população forem multiplicados por um mesmo fator de escala, quase sempre escolhido de modo que o indivíduo mais apto tenha valor adaptativo igual a um.

A determinação do valor adaptativo é um problema empírico. Face às imensas dificuldades experimentais, especialmente na associação de estimativas de valores adaptativos a genótipos específicos, muitos trabalhos teóricos têm analisado diversas funções que associam um valor adaptativo a cada configuração do espaço de genótipos, esperando descobrir fenômenos já observados empiricamente e propriedades universais desses mapeamentos. Uma função com essa finalidade é denominada um **relevo adaptativo**. Esse conceito foi introduzido por S. Wright [192] e pode ser entendido como um tipo de "energia potencial" inerente à dinâmica adaptativa de otimização evolucionária. É importante esclarecer que, embora Wright tenha usado o termo *adaptive landscape* originalmente, esta expressão tem outro sentido atualmente [66]. Assim, relevo adaptativo, ou paisagem adaptativa [62], equivale ao termo em inglês *fitness landscape*.

Provavelmente, a presença de uma intricada estrutura de picos e vales, para representar a notória complexidade da relação genótipo-fenótipo, deve ser uma condição necessária a qualquer relevo pretensamente realístico. Entretanto, para viabilizar os estudos analíticos, freqüentemente são escolhidos relevos com estruturas mais simples, mas ainda capazes de revelar propriedades interessantes. Nesta tese, são empregados apenas relevos estáticos (para modelos com dependência temporal, ver [186]) em que o valor adaptativo depende

apenas do número total de mutações (ou carga mutacional) de uma seqüência. Especificamente, são utilizados o relevo multiplicativo, incluindo sua alteração para representar efeitos de **epistase**, a interação entre genes na determinação do valor adaptativo, e o relevo truncado (ou de seleção truncada), que é uma generalização do famoso relevo de pico agudo. Considera-se, nesses relevos, que as mutações são sempre deletérias. Portanto, há somente um máximo, em ambos os casos.

2.4.1 Relevos multiplicativo e truncado

Para não sobrecarregar a notação, o valor adaptativo de um indivíduo com j mutações será sempre denotado por w_j, independentemente do relevo. O contexto sempre vai deixar claro qual relevo adaptativo estará sendo considerado.

Se $s \in [0, 1]$ é um coeficiente seletivo, define-se um relevo em que o valor adaptativo de uma seqüência com j mutações é dado por

$$w_j = (1 - s)^{j^\alpha}. \tag{2.3}$$

Uma eventual interação entre os sítios da seqüência é determinada pelo parâmetro de epistase, α. Quando $\alpha = 1$, obtém-se o relevo multiplicativo, em que não há interação entre os genes, pois cada nova mutação leva à mesma redução relativa do valor adaptativo do indivíduo, independentemente do número de mutações acumulado até então. Se $0 < \alpha < 1$, cada novo erro reduz o valor adaptativo de forma mais amena do que as mutações anteriores. Esse fenômeno é denominado epistase atenuante ou antagonística. Por outro lado, as mutações tornam-se cada vez mais prejudiciais quando $\alpha > 1$, efeito conhecido como epistase sinergética.

Esse relevo adaptativo tem sido muito utilizado [56, 190] em estudos da **catraca de Muller** (*Muller's ratchet*), um processo de degradação mutacional que pode ocorrer em populações finitas sujeitas à atuação conjunta de mutação e seleção. Esse fenômeno, que será discutido em detalhes no capítulo 4, é a perda inexorável dos indivíduos melhor adaptados devido a efeitos de deriva genética, levando a um decréscimo do valor adaptativo médio. Especula-se que esse comportamento possa ser responsável por uma forte pressão

seletiva em favor do surgimento de mecanismos responsáveis pela manutenção da aptidão de uma população, como a recombinação na reprodução sexuada [22].

O outro relevo utilizado neste trabalho, o de seleção truncada, incorpora epistase de uma forma extrema, alternando efeitos positivos e negativos. A aptidão de um indivíduo é inalterada por mutações enquanto a carga mutacional não atinge um certo valor, K. A mutação seguinte reduz o valor adaptativo pelo fator $1 - s$ e, a partir de então, as mutações tornam-se inócuas novamente. Em símbolos matemáticos,

$$w_j = \begin{cases} 1, & \text{se } j \leq K \\ 1 - s, & \text{se } j > K \end{cases} . \tag{2.4}$$

Nota-se que, se $s = 1$, uma quantidade de mutações maior que K é fatal para um indivíduo. Quando $K = 0$ e $s < 1$, obtém-se um relevo em que apenas um genótipo tem vantagem seletiva sobre os demais, sendo todos os mutantes igualmente pior adaptados. Há uma profusão de terminologias para esse caso especial na literatura internacional, entre as quais *sharply-peaked landscape*, *single sharp peak landscape*, *singly-peaked landscape*, *isolated peak landscape* ou ainda *master sequence landscape*. Optou-se por chamá-lo relevo de pico agudo, nome que deixa menor margem a interpretações equivocadas acerca da natureza do relevo. O genótipo mais apto é denominado seqüência mestra, o que justifica o último termo em inglês.

Acredita-se que o relevo de seleção truncada possa desempenhar um papel importante no estudo da dinâmica evolucionária de seqüências repetitivas em eucariontes [23]. Já a aplicabilidade do caso particular de pico agudo é bastante questionável. Sua possível relevância biológica só pode ser defendida quando se sugere seu uso no estudo de regiões restritas do genoma (por exemplo, os poucos sítios no centro ativo de uma enzima, cuja função pode ser facilmente comprometida por qualquer eventual mutação). Mesmo assim, o relevo de pico agudo tem sido amplamente estudado. Ele foi utilizado em estudos de evolução em alguns tipos de vírus [143], embora tenha sido proposto originalmente por Manfred Eigen [47], em sua teoria de quase-espécies para evolução pré-biótica (para uma excelente introdução, ver [159]). A literatura acerca da teoria de quase-espécies até 1989, em geral, e do relevo de pico agudo, em particular, foi revista em [48, 49]. Trabalhos mais

recentes são discutidos em [7] e [45].

2.4.2 Limiar de erro

Desde seu surgimento [47], o relevo de pico agudo tem sido associado [168] à determinação do chamado **limiar de erro** (*error threshold*). O limiar de erro é um tipo de ponto crítico em uma transição de fase. Trata-se de um limite superior para a taxa de mutação que uma população de seqüências pode suportar, sob certa intensidade de seleção. Alternativamente, quando os genomas são finitos, é possível dizer que há um tamanho máximo dos indivíduos, sob dadas intensidades de mutação e seleção, acima do qual não é possível evitar um acúmulo irrefreável de erros, e a população é levada à extinção. Esse resultado é conhecido como crise da informação, pois as macromoléculas incipientes degradar-se-iam muito antes que pudessem codificar informação suficiente para possibilitar o desenvolvimento de sistemas mais complexos. Esse problema levou M. Eigen e Peter Schuster a sugerirem que a informação genética poderia ser armazenada em um *conjunto* de moléculas que auxiliariam umas às outras, constituindo um ciclo catalítico denominado hiperciclo [50], embora posteriormente tenham sido apontados alguns problemas em relação à estabilidade dessa estrutura.

Entretanto, há alguns pontos controversos a destacar a respeito de limiares de erro. Muitas estimativas baseiam-se na hipótese, quase certamente inadequada, de que todo o genoma é descrito pelo relevo de pico agudo. Além disso, dependendo das aproximações adotadas (não há solução analítica exata para o modelo de quase-espécies nesse relevo), é até possível afirmar que, se o efeito da seleção for forte o suficiente, qualquer taxa de mutação é suportável [3]. De modo geral, essas inconsistências foram discutidas em [176] e por Thomas Wiehe em [181], que mostra que limiares de erro são ambíguos, no sentido de dependerem não somente do relevo considerado, mas também do critério adotado para seu cálculo. Ainda por cima, muitos cálculos de limiares de erro são baseados no modelo de infinitos sítios. Embora isso não impeça a obtenção de uma taxa máxima de mutação admissível, não faz sentido falar em crise de informação em uma população de indivíduos com genoma infinito. Wiehe chega a afirmar que a relevância de limiares de erro como

fatores limitantes da evolução molecular tem sido bastante superestimada.

Transições de fase legítimas só podem ser observadas no limite termodinâmico, o que corresponde a uma população de tamanho infinito. Não obstante esse fato e as ambigüidades inerentes ao problema mesmo no contexto usual, efeitos de limiares de erro têm sido descritos em populações finitas. O trabalho pioneiro nessa linha foi desenvolvido por Martin Nowak e P. Schuster [142]. Posteriormente, destacam-se [182] e alguns trabalhos de José F. Fontanari e colaboradores [4, 16, 17], que empregaram técnicas de escala de tamanho finito (*finite-size scaling*) para caracterizar o limiar de erro.

Finalmente, é importante destacar que há alguma evidência experimental em favor da existência de limiares de erro em populações reais. Observa-se uma relação inversa entre a intensidade da taxa de mutação e o tamanho do genoma em diversas espécies [44], sugerindo que a evolução possa "ajustar" a taxa de mutação de modo a evitar um limiar de erro [126]. Evidências mais diretas vêm de experimentos de evolução em microorganismos [51], especialmente vírus de RNA. Quando se induz um aumento na incidência de mutações em tais organismos, o valor adaptativo médio diminui e a população pode se extinguir [41, 42], supostamente devido à ultrapassagem do limiar de erro.

2.5 Dinâmica evolucionária: modelos de Wright-Fisher e de Moran

Ainda falta descrever a dinâmica populacional, ou seja, como varia ao longo do tempo a composição de uma população de indivíduos com diferentes habilidades competitivas e sujeitos a mutações. Antes de tudo, é preciso definir se a dinâmica deve ser descrita em tempo contínuo ou discreto, escolhas geralmente associadas [1] a haver ou não superposição de gerações, respectivamente.

Independentemente, S. Wright (1931) e R. A. Fisher (1930) conceberam um modelo de tempo discreto que se tornou o paradigma nos estudos sobre deriva genética. O **modelo**

[1] Com tempo contínuo, necessariamente há superposição de gerações, mas o mesmo não ocorre em modelos de tempo discreto.

de Wright-Fisher consiste simplesmente em uma população de N indivíduos que é atualizada segundo um esquema de amostragem probabilística com reposição. Dessa forma, são realizados N sorteios para construir a geração no instante t. Em cada um deles, se não houver mutação, o novo membro da população (filho) é idêntico a um indivíduo da geração $t-1$ (pai), escolhido com probabilidade proporcional ao seu valor adaptativo.

Mais precisamente, a probabilidade de que um certo indivíduo seja o pai de um específico membro de uma nova geração é a razão entre seu valor adaptativo e a soma dos valores adaptativos de toda a população. Se todos os indivíduos com mesmo valor adaptativo w_j forem agregados, a probabilidade de algum dos N_j candidatos do tipo j ser selecionado é

$$P_j = \frac{w_j N_j}{wN}, \tag{2.5}$$

em que

$$w = \frac{1}{N}\sum_k N_k w_k \tag{2.6}$$

é o valor adaptativo médio da população. Além disso, se m_{ij} for a probabilidade de que um genótipo do tipo j sofra uma mutação para o tipo i, então a probabilidade de um membro de uma nova geração ser do tipo i é

$$p_i = \sum_j m_{ij} P_j. \tag{2.7}$$

Dessa forma, a probabilidade de que a partição $\mathbf{N} = (N_{i_1}, \ldots, N_{i_l})$ dê origem à partição $\mathbf{N}' = (N_{i'_1}, \ldots, N_{i'_{l'}})$ na próxima geração é dada pela distribuição multinomial

$$P(\mathbf{N} \to \mathbf{N}') = N! \prod_{k=1}^{l'} \frac{p_{i'_k}^{N_{i'_k}}}{N_{i'_k}!}. \tag{2.8}$$

A notação carregada é necessária para representar todas as possíveis atualizações da população. São essas possíveis flutuações na composição da população que fazem do modelo de Wright-Fisher um boa ferramenta matemática para estudar a deriva genética. Na verdade, em todas as aplicações nesta tese, faz-se uso do modelo de infinitos sítios, de modo que a variabilidade por erros de amostragem é menor do que sugere a formulação geral exposta acima. O índice do tipo de um indivíduo é dado por sua carga mutacional

e a matriz de mutação é tal que $m_{ij} = M_{i-j}$, usando a Eq. (2.2). Em palavras, um filho adquire todas as mutações de seu pai e, além disso, um certo número de mutações adicionais, provenientes de falhas no mecanismo de replicação.

Um importante conceito relacionado ao modelo de Wright-Fisher é o de **tamanho efetivo da população**, N_e, mais um conceito concebido por S. Wright. A deriva genética naturalmente ocorre em qualquer população finita, mas é intensificada por outros fatores, como mutação, seleção e superposição de gerações. O tamanho efetivo da população, geralmente bem menor que N, é o número de indivíduos de uma população que sofre a mesma magnitude de deriva genética que a população real, mas em condições idealizadas, sem ação de quaisquer outros fatores. Uma população que tenha passado por uma recente expansão, por exemplo, tem N_e significativamente menor que seu tamanho real. A dinâmica populacional depende de N_e e, em muitos trabalhos, quando se escreve N, na verdade tem-se N_e em mente. Maiores informações podem ser obtidas em [77, p. 289] e [53, p. 104].

Finalmente, é preciso descrever uma dinâmica evolucionária em que seja possível haver superposição de gerações [156]. Um exemplo típico é o **modelo de Moran** [133]. Quando há reprodução, o que pode ocorrer tanto em tempo discreto quanto em tempo contínuo, ocorrem 2 sorteios. Primeiro, um membro da população é sorteado para gerar um novo indivíduo. Em seguida, há um novo sorteio, do qual o recém-ingresso não participa, para determinar um indivíduo que será eliminado da população. Segundo [53], esse mecanismo de reprodução (caracterizado pela superposição de gerações) e haploidia são as características definidoras de um modelo de Moran. Outras propriedades, como tipos possíveis de indivíduos, mutação e seleção (responsável por eventuais não uniformidades nos sorteios), podem ser definidas arbitrariamente.

Capítulo 3

Dinâmica de populações infinitas em relevos adaptativos simples

3.1 Introdução

Como em diversas outras áreas que têm sistemas complexos como objetos de estudo, um modelo em genética de populações precisa ser cuidadosamente elaborado de modo a revelar algumas características do sistema real que representa, embora essa representação seja, em geral, extremamente simplificada. Em muitos experimentos onde são cultivados indivíduos simples como vírus ou bactérias, por exemplo, esses organismos são tão numerosos que flutuações na composição da população são imperceptíveis experimentalmente. Nessas condições, o tamanho da população pode, efetivamente, ser considerado infinito. Assim, na ausência de deriva genética, o problema torna-se completamente determinístico, e é possível aplicar ferramentas da teoria de sistemas dinâmicos [82, 83] com bastante proveito.

Um bom exemplo da utilidade dessa metodologia é o célebre teorema de Hardy-Weinberg, descoberto independentemente pelo matemático inglês Godfrey Harold Hardy (1877-1947) e pelo fisiologista alemão Wilhelm Weinberg (1862-1937) em 1908 [76, 180]. Esse teorema garante que, em uma população infinitamente grande de indivíduos diplóides em que os cruzamentos ocorrem ao acaso e sobre a qual não há atuação de fatores evolutivos, como mutação, seleção natural e migrações, as freqüências gênicas e geno-

típicas não se alteram ao longo das gerações. Essa propriedade é válida para qualquer número de alelos associados ao *locus* gênico em questão [82, 83] e é usada para caracterizar populações em equilíbrio.

O princípio de Hardy-Weinberg baseia-se em condições biologicamente muito restritivas. Nos problemas modernos, mesmo com a adoção de hipóteses adicionais que permitam simplificações matemáticas (por exemplo, mutação e/ou seleção fracas, justificando expansões perturbativas), muitas vezes só é possível obter o comportamento estacionário de uma população. Embora esse resultado seja freqüentemente útil, há diversas situações experimentais em que a população só pode ser observada ao longo de períodos curtos quando comparados com o tempo médio de vida dos indivíduos, de modo que o estado de equilíbrio nunca é atingido. Além disso, acredita-se que o relevo adaptativo de uma população seja uma estrutura dinâmica, extremamente dependente de alterações no ambiente [186]. Portanto, a própria idéia de um sistema biológico evoluindo, sob as mesmas condições ambientais, por tempo suficiente para atingir um estado estacionário, pode ser inapropriada para a análise de alguns sistemas. Sendo assim, é fundamental conhecer o comportamento dinâmico de uma população. São escassos os trabalhos nessa linha, mas destacam-se os recentes [35] e [90, 91], que estudam a evolução da taxa de mutação.

Nesse contexto, os resultados apresentados neste capítulo são especialmente interessantes. Foram obtidas as soluções analíticas exatas (sem aproximações) para a dinâmica de populações infinitas de organismos haplóides, de genoma infinito, reproduzindo-se assexuadamente nos relevos multiplicativo e truncado. Inicialmente, é descrito o formalismo matemático geral, empregado em ambos os casos, seguido dos resultados particulares de cada relevo.

3.2 Formulação matemática da dinâmica determinística

Se $C_i(t)$ denotar a fração das infinitas seqüências que, no tempo t, carregam i mutações, cada uma com valor adaptativo w_i, a probabilidade de que alguma das seqüências com j

mutações seja escolhida para reproduzir-se é

$$P_j(t) = \frac{C_j(t)w_j}{w(t)}, \tag{3.1}$$

em que

$$w(t) = \sum_{k=0}^{\infty} C_k(t)w_k \tag{3.2}$$

é o valor adaptativo médio no tempo t. As duas expressões acima são idênticas às Eqs. (2.5) e (2.6) do capítulo anterior quando as classes são indexadas pela carga mutacional, pois claramente também seria possível definir concentrações $C_j = N_j/N$ naquele contexto. Nesta discussão, é fundamental representar explicitamente a dependência temporal (tempo discreto).

A evolução da população é descrita pela equação de convolução

$$C_i(t) = \sum_{j=0}^{i} P_j(t-1)M_{i-j} = \sum_{j=0}^{i} \frac{C_j(t-1)w_j}{w(t-1)}\, e^{-U} \frac{U^{i-j}}{(i-j)!}, \tag{3.3}$$

obtida originalmente por Kimura e Maruyama [103] e baseada na Eq. (2.2). O lado direito desta expressão é $p_i(t)$, dado pela Eq. (2.7). Mesmo assim, apesar da evidente interpretação estocástica, a evolução ocorre de forma completamente determinística. A igualdade $C_i(t) = p_i(t)$ só ocorre quando $N_i(t) \to \infty$, $N \to \infty$ e $N_i(t)/N \to C_i(t)$, de modo que a concentração $C_i(t)$ não é mais uma variável aleatória.

Nesta tese, a equação (3.3) foi resolvida mediante a utilização de uma função geratriz, $G(z,t)$, para as concentrações $C_i(t)$,

$$G(z,t) = \sum_{i=0}^{\infty} C_i(t)z^i. \tag{3.4}$$

O procedimento consiste em multiplicar os dois lados da equação (3.3) por z^i e somá-los para valores de i de 0 a ∞, gerando uma equação de recorrência para a função geratriz. Em geral, não é possível obter uma solução simples para essa recorrência, iterando-a até $t = 0$. Felizmente, os relevos multiplicativo e truncado são exceções (talvez as únicas) a essa regra. A determinação da distribuição de concentrações pode então ser facilmente

obtida porque, no instante t, a concentração de indivíduos com i mutações é o coeficiente de z^i na expansão de $G(z,t)$ em série (de Taylor) de potências de z,

$$C_i(t) = \frac{1}{i!}\left\{\frac{\partial^i}{\partial z^i}G(z,t)\right\}_{z=0}. \tag{3.5}$$

3.3 Solução em um relevo multiplicativo

3.3.1 Solução assintótica

Em 1978, embora John Haigh estivesse interessado principalmente em estudar as propriedades da catraca de Muller, fenômeno que ocorre em populações finitas, ele obteve [74] expressões assintóticas para a distribuição de concentrações,

$$C_i(\infty) = e^{-U/s}\frac{(U/s)^{i-m}}{(i-m)!}, \tag{3.6}$$

e para o valor adaptativo médio,

$$w(\infty) = (1-s)^m e^{-U}, \tag{3.7}$$

que também são válidas para populações infinitas. Admite-se, nas expressões acima, que a seqüência mais apta tenha m mutações. Obviamente, esses resultados são recuperados como um caso particular da solução dinâmica apresentada logo abaixo.

3.3.2 Solução dinâmica

A equação de recorrência obtida para a função geratriz a partir da Eq. (3.3) no caso do relevo multiplicativo, dado pela Eq. (2.3) quando $\alpha = 0$, é

$$G(z,t) = \frac{e^{zU}\,G[z(1-s),t-1]}{e^{U}\,G[1-s,t-1]}, \tag{3.8}$$

após uma conveniente inversão da ordem nos somatórios durante as manipulações algébricas. Essa equação pode ser resolvida recursivamente, e o resultado final é

$$G(z,t) = \frac{e^{zU\theta_t}\,G[z(1-s)^t,0]}{e^{U\theta_t}\,G[(1-s)^t,0]}, \tag{3.9}$$

em que θ_t é simplesmente a soma dos t termos de uma progressão geométrica finita de razão $1-s$ e primeiro termo 1,

$$\theta_t = \frac{1}{s}\left[1-(1-s)^t\right]. \tag{3.10}$$

Usando a equação (3.5) e a definição da função geratriz (3.4),

$$C_i(t) = \frac{e^{-U\theta_t}}{\sum_{j=0}^{\infty} C_j(0)(1-s)^{jt}} \sum_{j=0}^{i} \frac{(U\theta_t)^{i-j}}{(i-j)!} C_j(0)(1-s)^{jt}. \tag{3.11}$$

O valor adaptativo médio, definido de forma geral em (3.2), assume uma forma bem simples no relevo multiplicativo,

$$w(t) = G(1-s,t) = e^{-sU\theta_t} \frac{\sum_{j=0}^{\infty} C_j(0)(1-s)^{j(t+1)}}{\sum_{j=0}^{\infty} C_j(0)(1-s)^{jt}}. \tag{3.12}$$

Nota-se que, quando $t \to \infty$, os somatórios nas Eqs. (3.11) e (3.12) são dominados pelas menores concentrações não nulas inicialmente, recuperando os resultados clássicos de Haigh. Além disso, se todos os indivíduos tiverem o mesmo número m de mutações inicialmente, ou seja, $C_i(0) = \delta_{i,m}$, a dinâmica da população é governada pela equação (3.6) com $1/s$ substituído por θ_t.

3.4 Solução em um relevo de seleção truncada

3.4.1 Solução dinâmica

Neste caso, os cálculos são bem mais extensos e as principais manipulações algébricas estão discutidas no apêndice B. Em um relevo truncado (2.4), a função geratriz obedece à equação

$$G(z,t) = \frac{e^{zU}\left[s\sum_{j=0}^{K} z^j C_j(t-1) + (1-s)G(z,t-1)\right]}{e^{U}\left[s\sum_{j=0}^{K} C_j(t-1) + (1-s)\right]}, \tag{3.13}$$

obtida a partir de (3.3). Recursivamente,

$$G(z,t) = \left\{ s\, e^{zU} \sum_{j=0}^{K} \sum_{i=0}^{t-1} \sum_{l=0}^{K-j} z^j C_j(0) [e^{zU}(1-s)]^{t-1-i} \frac{(iUz)^l}{l!} \right.$$
$$\left. + [e^{zU}(1-s)]^t G(z,0) \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{K} \sum_{i=0}^{t-1} \sum_{l=0}^{K-j} C_j(0) [e^{U}(1-s)]^{t-1-i} \frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\}. \tag{3.14}$$

A distribuição de concentrações é obtida mediante a equação (3.5),

$$C_m(t) = \left\{ s \sum_{j=0}^{\min(K,m)} \sum_{i=0}^{t-1} \sum_{l=0}^{\min(K,m)-j} C_j(0) \frac{U^{m-j}}{(m-j)!} (1-s)^{t-1-i} \right.$$
$$\left. \times \begin{pmatrix} m-j \\ l \end{pmatrix} i^l (t-i)^{m-j-l} + (1-s)^t \sum_{j=0}^{m} C_j(0) \frac{(Ut)^{m-j}}{(m-j)!} \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{K} \sum_{i=0}^{t-1} \sum_{l=0}^{K-j} C_j(0) [e^{U}(1-s)]^{t-1-i} \frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\}. \tag{3.15}$$

Em particular, a concentração dos indivíduos melhor adaptados admite uma expressão bem mais simples,

$$C_{m \leq K}(t) = \left\{ \sum_{j=0}^{m} C_j(0) \frac{(Ut)^{m-j}}{(m-j)!} \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{K} \sum_{i=0}^{t-1} \sum_{l=0}^{K-j} C_j(0) [e^{U}(1-s)]^{t-1-i} \frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\}. \tag{3.16}$$

O valor adaptativo médio, por sua vez, é

$$w(t) = 1 - s + s \sum_{m=0}^{K} C_m(t) = 1 - s + \left\{ s \sum_{j=0}^{K} C_j(0) \sum_{l=0}^{K-j} \frac{(Ut)^l}{l!} \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{K} \sum_{i=0}^{t-1} \sum_{l=0}^{K-j} C_j(0) [e^{U}(1-s)]^{t-1-i} \frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\} \tag{3.17}$$

e nota-se que, em qualquer instante t, as duas equações acima dependem somente das concentrações iniciais das seqüências com vantagem seletiva. No caso geral, essas expressões não podem ser simplificadas, mas se $K = 0$ ou $t \to \infty$, algum progresso analítico ainda é possível.

3.4.2 Dinâmica no relevo de pico agudo

No relevo de pico agudo, a concentração da seqüência mestra é

$$C_0^*(t) = \frac{[1 - e^U(1-s)]C_0(0)}{s\,e^U C_0(0) + \{1 - e^U + s\,e^U[1 - C_0(0)]\}\{e^U(1-s)\}^t} \tag{3.18}$$

e o valor adaptativo médio é dado por

$$w^*(t) = 1 - s + sC_0^*(t), \tag{3.19}$$

ou, explicitamente, por

$$w^*(t) = \frac{sC_0(0) + (1-s)\left\{1 - e^U + s\,e^U[1 - C_0(0)]\right\}\left\{e^U(1-s)\right\}^t}{s\,e^U C_0(0) + \{1 - e^U + s\,e^U[1 - C_0(0)]\}\{e^U(1-s)\}^t}. \tag{3.20}$$

As freqüências dinâmicas dos mutantes seguem a equação (3.15),

$$\begin{aligned} C_{m>0}(t) &= \left\{1 - e^U(1-s)\right\} \\ &\times \left\{\frac{s}{1-s}C_0(0)\frac{U^m}{m!}\sum_{j=1}^{t} j^m(1-s)^j + (1-s)^t\sum_{j=0}^{m} C_j(0)\frac{(Ut)^{m-j}}{(m-j)!}\right\} \\ &\div \left\{s\,e^U C_0(0) + \left\{1 - e^U + s\,e^U[1 - C_0(0)]\right\}\left\{e^U(1-s)\right\}^t\right\}. \end{aligned} \tag{3.21}$$

Em princípio, essa equação poderia ser simplificada ainda mais, pois seu primeiro somatório pode ser formalmente expresso em termos de objetos denominados números de Euler generalizados [124]. Mas, para t qualquer, esse procedimento não se mostra realmente útil, pois a própria definição desses números não é simples. Entretanto, uma simplificação bem interessante ocorre no estado estacionário.

Solução assintótica no relevo de pico agudo

Quando $t \to \infty$, a equação (C.2), do apêndice C, permite que se escreva a concentração dos mutantes como

$$C_{m>0}(\infty) = \left\{\frac{1 - e^U(1-s)}{s\ e^U}\right\}\left\{\frac{(U/s)^m}{m!}\right\}\mathcal{P}_m(1-s), \tag{3.22}$$

quando a taxa de mutação está abaixo de um limiar de erro, determinado pela relação $e^U(1-s) < 1$. Nessa equação,

$$\mathcal{P}_m(x) \equiv \sum_{k=0}^{m-1} E_{m,k}x^k \tag{3.23}$$

e os números de Euler $E_{m,k}$ são definidos [1] no apêndice C.

Sabe-se (ver, por exemplo, [181]) que, em uma população infinita, a concentração assintótica de uma seqüência mestra infinita em um relevo de pico agudo é

$$C_0^*(\infty) = \frac{1 - e^U(1-s)}{s\, e^U} \tag{3.24}$$

e que o valor adaptativo médio estacionário é

$$w^*(\infty) = e^{-U}. \tag{3.25}$$

Esses dois resultados pressupõem que $C_0(0) \neq 0$ e que a taxa de mutação é suficientemente pequena, $U < -\ln(1-s)$; se pelo menos uma dessas condições não for satisfeita, $C_0^*(\infty) = 0$ e $w^*(\infty) = 1 - s$. Portanto, conforme exposto na subseção 2.4.2, pode existir um limiar de erro mesmo quando se considera uma população de indivíduos de genoma infinito. É fácil ver que, além do limiar de erro, a solução de equilíbrio também pode ser obtida mediante uma análise do comportamento das equações (3.18) e (3.19) quando $t \to \infty$.

3.4.3 Estado estacionário

A solução geral para $t \to \infty$ quando $K \neq 0$ é uma conseqüência direta do fato que, usando a equação (3.16) para $0 \leq m < n < K$, vale a relação

$$\lim_{t\to\infty} \frac{C_m(t)}{C_n(t)} = 0. \tag{3.26}$$

Portanto, assintoticamente, todas as seqüências com menos de K mutações são extintas, quaisquer sejam o coeficiente seletivo e a taxa de mutação. Conseqüentemente, o estado estacionário da população no relevo de seleção truncada é o mesmo obtido para o relevo de pico agudo, considerando os portadores de K mutações como seqüências mestras. É claro que cálculos explícitos confirmam esses resultados. Finalmente, se $e^U(1-s) < 1$ e

[1] Crédito a quem merece: tentando encontrar a concentração assintótica dos mutantes, descobri os números de Euler, mas eu não os conhecia. Não os encontrei em nenhuma referência, procurava cegamente. Até o dia em que pensei: www.google.com! Digitei a seqüência: 1,11,26,57,66 (ver apêndice). O Google achou a referência [124].

$\mathcal{P}_0(x) \equiv 1$, $C_{m<K}(\infty) = 0$ e

$$C_{m\geq K}(\infty) = \left\{\frac{1-e^U(1-s)}{s\ e^U}\right\}\left\{\frac{(U/s)^{m-K}}{(m-K)!}\right\}\mathcal{P}_{m-K}(1-s). \tag{3.27}$$

Esse resultado admite uma interpretação simples. O relevo truncado é plano abaixo de K, e sabe-se que uma população infinita não pode se manter localizada indefinidamente em um relevo plano. Apenas a barreira entrópica em K pode garantir a sobrevivência de indivíduos com um número finito de mutações (K, no mínimo), desde que o limiar de erro não tenha sido ultrapassado. Em outras palavras, os mutantes com menos de K imperfeições não possuem qualquer vantagem seletiva em relação às seqüências com K mutações que lhes permita contrabalançar os efeitos degenerativos da taxa de mutação.

3.5 Comparações entre os dois relevos adaptativos

Com algum esforço numérico, a solução por recorrência da Eq. (3.3) já possibilitava o estudo do comportamento dinâmico de uma população infinita em *qualquer* relevo adaptativo. Mesmo assim, é interessante aproveitar os resultados deste capítulo para ilustrar as propriedades dos relevos multiplicativo e truncado, lembrando que apenas este apresenta limiar de erro. Na figura 3.1, estão representadas as dependências temporais do genótipo selvagem e do valor adaptativo médio para quatro combinações dos parâmetros s e U, em três relevos diferentes: multiplicativo, de pico agudo e truncado com $K = 10$.

Como previsto na Eq. (3.26), a concentração de seqüências mestras eventualmente se anula em relevos com $K > 0$, como pode ser visto nas curvas em verde dos gráficos de $C_0(t)$. Além disso, verifica-se que as soluções assintóticas nos relevos multiplicativo e de pico agudo dadas pelas Eqs. (3.6) e (3.24) também são praticamente nulas nos gráficos a, b e d. Somente no gráfico c, quando $s = 0.1$ e $U = 0.1$, nota-se uma fração significativa de genótipos selvagens no estado de equilíbrio. Mas, convém ressaltar, a dinâmica de relaxação no relevo de pico agudo é muito lenta. São necessárias muitas gerações para que $C_0(t)$ atinja o valor previsto de cerca de 0.048, bem menor do que a concentração 0.25 obtida em $t = 30$. Esse efeito também está presente no comportamento do valor

adaptativo médio. Nos gráficos e e f da figura 3.1, observa-se que, em $t = 30$, $w(t)$ no relevo multiplicativo ainda está distante do valores previstos pela Eq. (3.7), 0.90 e 0.61, respectivamente. Essa lentidão é ainda mais marcante nos relevos truncados. Enquanto nos gráficos e, f e h, o limiar de erro é ultrapassado e rapidamente o valor adaptativo médio atinge a previsão $1 - s$ (ver comentário abaixo da Eq. (3.25)), o gráfico g leva a crer que a curva verde já está estabilizada, quando na verdade ela acompanha a solução de pico agudo até o valor 0.9 dado pela Eq. (3.25), após algumas poucas centenas de gerações (resultados não expostos).

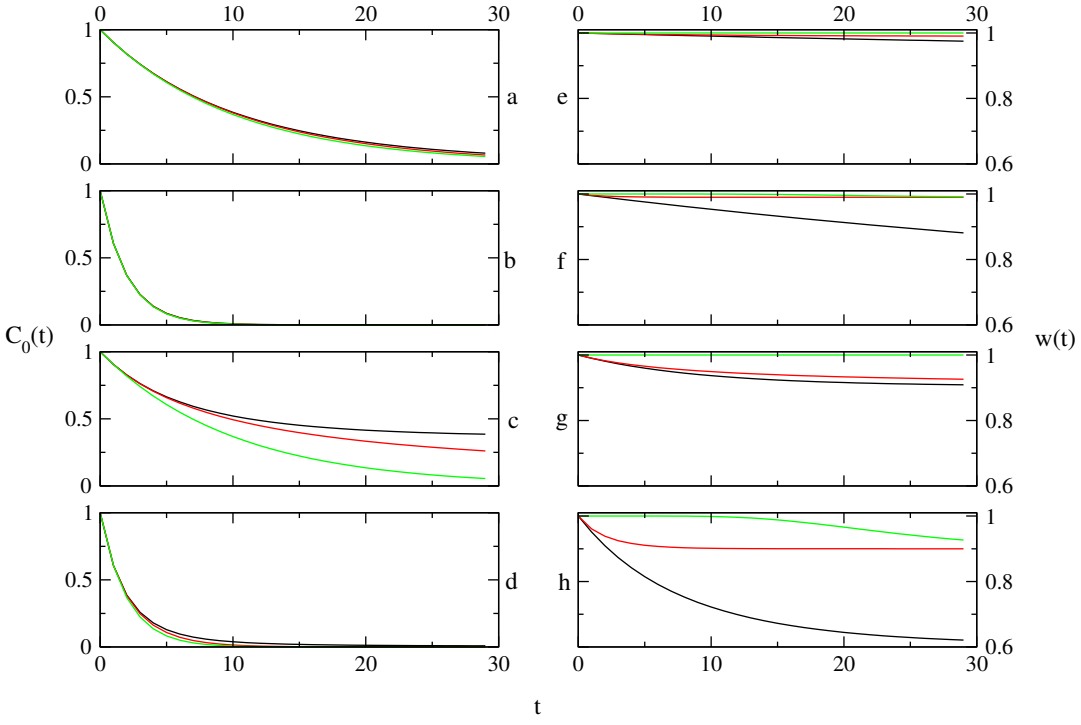

Figura 3.1: Comportamento dinâmico da concentração do genótipo selvagem e do valor adaptativo médio. Na coluna da esquerda, representa-se $C_0(t)$ e, na da direita, $w(t)$. Cada par de gráficos na mesma linha corresponde aos mesmos parâmetros, $s = 0.01$ e $U = 0.1$, $s = 0.01$ e $U = 0.5$, $s = 0.1$ e $U = 0.1$, $s = 0.1$ e $U = 0.5$, de cima para baixo. Em todos os gráficos, foi adotada a condição inicial $C_0(0) = 1$ e as cores preta, vermelha e verde denotam os relevos multiplicativo, de pico agudo e truncado com $K = 20$, respectivamente.

As Eqs. (3.7) e (3.25) mostram que o valor adaptativo médio no estado estacionário é o mesmo em ambos os relevos estudados, se o limiar de erro não tiver sido ultrapassado. Isso não é uma coincidência. Kimura e Maruyama mostraram [103] que o valor adaptativo médio de uma população com reprodução assexuada em equilíbrio é dado pelo produto entre $w(\infty) = e^{-U}$ e o valor adaptativo do indivíduo mais apto, independentemente dos detalhes do relevo adaptativo adotado. Esse resultado também é conhecido como princípio de Haldane-Muller, haja vista que J. B. S. Haldane [75] e H. J. Muller [137] também notaram a irrelevância do impacto seletivo das mutações deletérias e obtiveram a aproximação $1 - U$ para o valor adaptativo médio. Esse fenômeno ocorre porque, no balanço mutação-seleção, quanto maior é o coeficiente seletivo, menor é a freqüência de equilíbrio de indivíduos com muitas mutações.

Mas é importante ressaltar que esse princípio nem sempre é válido. Kimura e Maruyama admitiram que a seqüência de replicação mais rápida na população não teria sítios neutros, ou seja, todas as mutações deveriam afetar seu valor adaptativo. A existência de mutações neutras origina [13, 153, 173, 174, 183] uma pressão seletiva que dirige a população a regiões do espaço de genótipos onde a densidade de seqüências neutras é alta, conferindo robustez contra mutações às seqüências. Esse efeito pode ser intenso a ponto ser mais vantajoso [2] para a população assumir taxas mais baixas de replicação, desde que acompanhadas de maior robustez mutacional [185, 187].

3.6 Conclusões

Em resumo, foram encontradas soluções exatas para o comportamento dinâmico de populações infinitas nos relevos adaptativos mais empregados em modelos para a evolução de organismos simples. Esses resultados podem ser úteis em qualquer situação experimental em que se acredite que a dinâmica evolucionária possa ser descrita pelos relevos multiplicativo e truncado, pelo menos aproximadamente.

[2] Em inglês, esse efeito é descrito pelo trocadilho *survival of the flattest* (sobrevivência do "mais plano", ao invés de *survival of the fittest*, sobrevivência do mais apto).

É importante ressaltar que a solução determinística no relevo multiplicativo é reconhecidamente [111] útil em estudos quantitativos da velocidade da catraca de Muller, embora esse fenômeno ocorra somente em populações finitas. Na verdade, a distribuição estacionária da população infinita é crucial na caracterização [26] da catraca de Muller em alguns experimentos de transferências seriais de gargalo com populações de crescimento rápido, discutidos no capítulo 4. A solução dinâmica certamente poderia ser empregada nesse contexto ou mesmo viabilizar tratamentos teóricos em outros arranjos experimentais.

Embora o próprio conceito de limiar de erro tenha sido alvo de muitas críticas, alguns estudos analisam a possibilidade de que, ao longo da evolução de algumas populações virais, tenha havido uma pressão seletiva sobre a taxa de mutação do sistema de modo a torná-la bem próxima ao limiar de erro [37, 91, 162]. Evidências experimentais são discutidas em [42]. A razão desse fenômeno seria a necessidade de constante adaptação para escapar da reação imunológica nos organismos atacados pela população viral. Os relevos de seleção truncada e de pico agudo ainda são os modelos mais utilizados para descrever tais sistemas ou quaisquer outros em que se acredita haver um limiar de erro.

Capítulo 4

Acúmulo de mutações em populações de tamanho variável

4.1 Introdução

O surgimento da reprodução sexuada entre organismos primitivos ainda é um tema bastante controverso. Como uma população incipiente de mutantes que precisam gastar energia procurando parceiros e, principalmente, cujo processo reprodutivo é metabolicamente mais intenso, poderia sobreviver à competição contra seus rivais asexuados e autônomos? Em 1964, H. J. Muller (1890-1967), ganhador do prêmio Nobel de Medicina e Fisiologia de 1946, concebeu uma possível explicação.

Ele conjecturou [138] que o valor adaptativo médio em linhagens asexuadas, desprovidas de mecanismos de reparo genético, deveria decrescer com o passar do tempo se a taxa com que ocorrem mutações deletérias fosse suficientemente alta em relação às taxas de ocorrência de mutações reversas ou benéficas. Nessas condições, flutuações estocásticas inerentes à finitude da população seriam responsáveis por inevitáveis e sucessivas perdas dos organismos melhor adaptados, em um processo que foi metaforicamente associado [56] ao movimento de uma catraca e denominado a partir de então como catraca de Muller. A figura 4.1 ilustra esse fenômeno, cuja assinatura é o comportamento não decrescente do número mínimo de mutações na população em função do tempo.

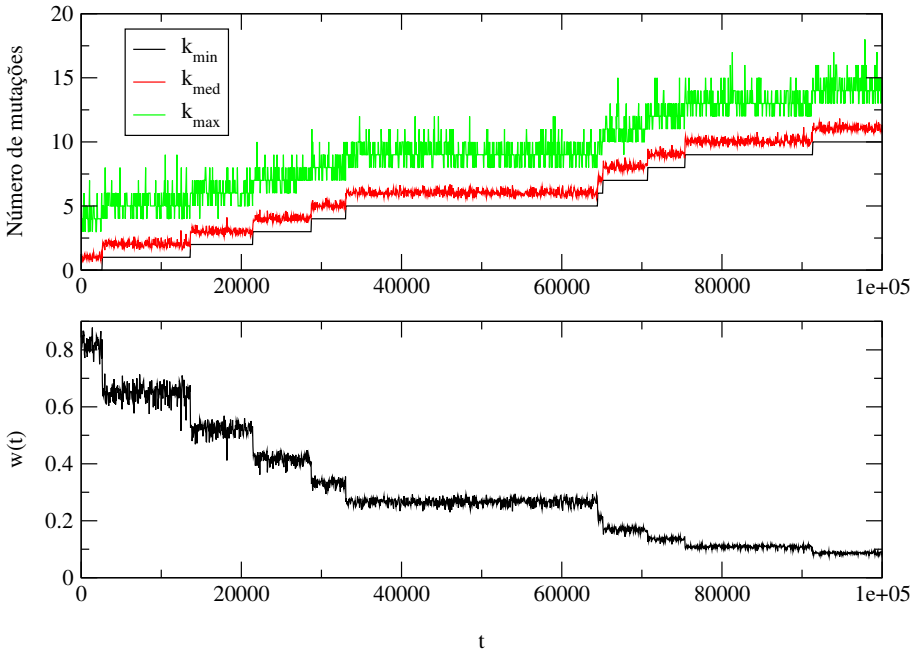

Figura 4.1: Comportamento típico da catraca de Muller. O número mínimo de mutações na população em função do tempo é denotado por k_{min}. Para comparação, também estão expostos os números médio (k_{med}) e máximo (k_{max}) de mutações, além do valor adaptativo médio. Essas médias são calculadas tendo como amostras todos os indivíduos de *uma* população de tamanho constante que obedece uma dinâmica de Wright-Fisher em um relevo multiplicativo. Os saltos irreversíveis de k_{min} justificam o termo catraca.

Assim, parece plausível imaginar a existência de uma pressão seletiva em favor do surgimento do sexo [22], pois a recombinação pode reverter o dano genético decorrente do acúmulo de mutações deletérias e, dependendo da magnitude desse efeito, ele poderia compensar a vantagem metabólica da reprodução asexuada. Embora ainda não haja evidência direta que a catraca de Muller tenha realmente sido um fator importante no surgimento do sexo, ela tem sido utilizada para justificar o decaimento do valor adaptativo em experimentos envolvendo linhagens assexuadas de protozoários [8], bactérias [5], bacteriófagos de RNA [21] e vírus de RNA em mamíferos [46]. Além disso, especula-se também que a catraca pode ser relevante em questões como infertilidade masculina [28] e atenuação no HIV (*human immunodeficiency virus* - vírus da imunodeficiência humana), que pode desencadear a AIDS (*acquired immunodeficiency syndrome* - síndrome da imunodeficiência

adquirida) [193].

Após Felsenstein ter adotado [56] o relevo adaptativo multiplicativo para descrever a catraca de Muller, vários artigos, a começar por [74], utilizaram-no na determinação da velocidade da catraca (taxa de perda das classes mais aptas) em função de s, U e N, principalmente via simulações e aproximações de difusão [81, 166]. Entretanto, nunca foi encontrada alguma expressão analítica de validade geral, em todo o espaço de parâmetros. Recentemente, Brian Charlesworth e colaboradores também se dedicaram a analisar as propriedades dinâmicas da catraca de Muller [24, 70, 71, 72].

O acúmulo de mutações também constitui o fundamento do modelo de cadeias de bits (*bit-strings*), introduzido [148] por Tadeu Penna em 1995. Esse modelo foi elaborado originalmente para estudar processos de envelhecimento biológico em populações com estrutura etária. Entretanto, várias modificações tornaram-no uma ferramenta útil para analisar outros efeitos, como a depressão endogâmica [163] e a evolução do sexo [12]. Esses desenvolvimentos já foram revisados em diversas ocasiões [10, 134, 135, 136, 149, 165].

De fato, há vários processos degenerativos associados ao acúmulo de mutações além da catraca de Muller. Um outro exemplo, claro, é o limiar de erro, discutido nos capítulos anteriores. Mas é importante destacar que, embora em ambos os casos haja perda do genótipo melhor adaptado, as origens dessas degenerações mutacionais são distintas. Embora limiares de erro também tenham sido descritos em modelos teóricos de populações de tamanho finito, isso não é fundamental, pois eles foram observados originalmente em populações infinitas. É inevitável concluir que a competição entre mutação e seleção é o principal fator responsável pelos limiares de erro. Por outro lado, a deriva genética é o elemento chave na manifestação da catraca de Muller, evidenciando seu caráter essencialmente estocástico. Assim, parece razoável que modelos com flutuações mais intensas do que aquelas presentes em populações de tamanho finito, mas fixo, possam revelar efeitos inesperados.

Nessa linha de raciocínio, a maioria dos trabalhos recentes têm apresentado modelos de populações de tamanho variável. Além da função exploratória mencionada logo acima, essa abordagem é promissora por permitir uma representação mais fiel da rea-

lidade. Afinal, expansões e contrações populacionais são ubíquas tanto no laboratório quanto na natureza. Além de extinções e migrações, outros exemplos importantes são a dinâmica populacional de endosimbiontes [155] e patógenos [9, 85, 86] em seus hospedeiros e a propagação de infecções mediante a emissão de gotículas respiratórias no ar, seguida pela rápida multiplicação dos micróbios nelas transportados quando encontram um novo hospedeiro não infectado.

A relação entre o número médio de descendentes de uma geração e a capacidade de suporte do ambiente também pode potencializar os efeitos da variabilidade no tamanho da população. Quando a fecundidade é alta, a amostragem de Wright-Fisher, baseada em valores adaptativos relativos em uma população de tamanho fixo, geralmente revela-se uma ferramenta adequada. No entanto, é possível que a degeneração do valor adaptativo médio no sistema seja acentuado, de modo que os valores adaptativos absolutos tornam-se importantes. Se isso ocorrer, a população extingue-se rapidamente, pois à medida que seu tamanho diminui, a catraca de Muller atua de forma progressivamente mais intensa. Nesse caso, fala-se em derretimento mutacional (*mutational meltdown*). Esse fenômeno foi considerado pioneiramente por Lynch e Gabriel [63, 118, 119] e pode afetar até mesmo pequenas populações que apresentam reprodução sexuada [11]. Por outro lado, há casos em que a catraca pode não operar, mesmo em colônias sujeitas à extinção [128].

Este capítulo é dedicado ao estudo de dois modelos em que a variabilidade no tamanho da população é essencial na caracterização da degeneração mutacional. Em primeiro lugar, determina-se a distribuição de probabilidade do número mínimo de mutações em uma população sujeita a transferências seriais via "gargalos estocásticos". No segundo problema, um processo de ramificação generalizado é utilizado para estudar os efeitos da catraca de Muller em populações sob crescimento exponencial.

4.2 Transferências seriais com gargalos variáveis

O novo campo da evolução experimental está baseado em técnicas [1] que permitem a determinação do valor adaptativo médio de uma população de microorganismos mediante uma comparação de sua taxa de replicação com um grupo de controle, preservado sob condições anti-mutagênicas (resfriamento intenso, entre outros fatores).

Há um protocolo experimental em que, repetidamente, permite-se o crescimento (exponencial, nos primeiros estágios) de culturas de microorganismos por algum tempo, ao fim do qual pequenas amostras da população são colhidas para dar origem a novas culturas. Esse tipo de experimento é denominado, por razões óbvias, transferência serial com gargalo (*bottleneck*). Os gargalos simulam a perda dos indivíduos mais aptos na deriva genética e, após várias transferências, alguns efeitos da dinâmica evolucionária podem ser estudados quando os mutantes são comparados com representantes da primeira cultura.

Experimentos desse natureza, envolvendo vírus de RNA [41], deram origem a fortes evidências [21, 25, 46] em favor da ocorrência da catraca de Muller em sistemas reais. Em particular, efeitos da largura do gargalo têm sido observados, conforme revisto em [42]. De forma geral, nota-se que um grande número de transferências estreitas leva ao acúmulo de mutações deletérias na população viral, enquanto poucas e largas passagens possibilitam o crescimento do valor adaptativo médio nas culturas. Provavelmente, este último comportamento é devido à ocorrência de mutações benéficas e jamais pode ser observado nos modelos de mutações deletérias e irreversíveis empregados nesta tese.

4.2.1 O caso $N = 1$

Inspirados por essa bem estabelecida metodologia, Colato e Fontanari elaboraram um modelo para estudar a catraca de Muller em populações assexuadas [26]. Como usual, eles não consideraram mutações reversas. Além disso, eles admitiram que o período de incubação subseqüente a cada passagem seria longo o suficiente para que todas as novas

[1] Um artigo de revisão acerca das diversas metodologias experimentais para o estudo da evolução em microorganismos foi publicado recentemente [51].

culturas pudessem ser descritas por uma população infinita em equilíbrio e que a classe mais apta não seria perdida no processo de expansão. Com base nessas hipóteses, eles apropriadamente argumentaram que o número de passagens até a perda da classe mais apta, com m mutações, deveria obedecer uma distribuição geométrica [2] com probabilidade de sucesso $p = [1 - C_m(\infty)]^N$, onde $C_m(\infty)$ é a concentração assintótica da classe m em qualquer relevo adaptativo, se N indivíduos são transferidos em cada passagem.

Tendo em vista o experimento de Chao [21], em que os gargalos consistiam na passagem de apenas um vírus, Colato e Fontanari também calcularam a probabilidade $P_\tau(m)$ de que, na τ-ésima transferência em uma população evoluindo no regime multiplicativo, um indivíduo com m mutações seja amostrado para fundar a próxima geração. Isso só é possível se o fundador da própria geração τ tiver n mutações, com $n \leq m$. Nessas condições, sempre será possível sortear um m-mutante a partir da distribuição de equilíbrio adequada. Quando todas as possibilidades são levadas em conta, naturalmente surge a equação

$$P_\tau(m) = \sum_{n=0}^{m} P_{\tau-1}(n) e^{-U/s} \frac{(U/s)^{m-n}}{(m-n)!}, \tag{4.1}$$

em que o termo poissoniano vem da Eq. (3.6). Segundo o procedimento descrito no capítulo 3, a função geratriz

$$G_\tau(z) = \sum_{n=0}^{\infty} z^n P_\tau(n) \tag{4.2}$$

viabiliza uma resolução recursiva da Eq. (4.1), que resulta em $G_\tau(z) = \exp[(z-1)\tau U/s]$ e, conseqüentemente, gera a solução

$$P_\tau(m) = e^{-\tau U/s} \frac{(\tau U/s)^m}{m!}. \tag{4.3}$$

O fundador de cada nova cultura sempre sobrevive e determina [3] a distribuição de equilíbrio das concentrações. Se ele portar m defeitos, o valor adaptativo assintótico

[2] Se há interesse apenas na ocorrência ou não de um evento em diversas tentativas, é conveniente denominar uma das possibilidades como um sucesso, com probabilidade p, ou um fracasso, com probabilidade $q = 1 - p$. A distribuição geométrica $P(n) = q^{n-1}p$ corresponde à probabilidade de que o primeiro sucesso só ocorra na n-ésima tentativa.

[3] Consultar o capítulo 3 para uma discussão mais detalhada.

médio $\widehat{w}$ da população é dado pela Eq. (3.7), escrita na forma

$$\widehat{w} = e^{-U}(1-s)^m. \tag{4.4}$$

Porém, como o processo de transferência faz com que m seja uma variável aleatória, é preciso combinar a equação acima com a Eq. (4.3) para obter a esperança [4] do q-ésimo momento do valor adaptativo médio da população no estado de equilíbrio alcançado após a τ-ésima passagem,

$$\langle \widehat{w}^q \rangle = \sum_{m=0}^{\infty} \left[e^{-U}(1-s)^m\right]^q P_\tau(m) = e^{-qU} e^{-\tau U \theta_q}, \tag{4.5}$$

onde θ_q é dado pela Eq. (3.10).

Colato e Fontanari utilizaram um raciocínio indutivo para obter a Eq. (4.3). Mas, de forma geral, $P_\tau(m)$ pode ser vista como a probabilidade de que o número mínimo de mutações no sistema após a τ-ésima passagem seja m. Esse ponto de vista, juntamente com o formalismo de funções geratrizes, possibilita uma abordagem sistemática do problema, e o objetivo desta seção é justamente generalizar esse estudo para diferentes tamanhos de amostra. Afinal, a passagem de apenas um indivíduo é um caso extremo, tornando questionável qualquer tentativa de extensão de fenômenos observados no modelo a populações reais. Além disso, os autores de [26] já tinham obtido os dois primeiros momentos, $\langle \widehat{w} \rangle = e^{-U} e^{-\tau U}$ e $\langle \widehat{w}^2 \rangle = e^{-2U} e^{-\tau U(2-s)}$, e destacado o fato de $\langle \widehat{w} \rangle$ ser independente de s, o que exige medidas da razão $\langle \widehat{w}^2 \rangle \,/\, \langle \widehat{w} \rangle^2 = e^{\tau U s}$ para determinar o coeficiente seletivo. É interessante investigar como diferentes esquemas poderiam afetar a determinação empírica de s e U.

4.2.2 Formalismo geral

Cada termo no somatório da Eq. (4.1) é o produto de $P_{\tau-1}(n)$ pela probabilidade $P(Y_1 = m \mid n, N)$ do número mínimo de mutações em uma amostra de tamanho N colhida em uma população cujo indivíduo mais apto porta n defeitos, no caso $N = 1$. É

[4] Essa esperança refere-se à média tomada no *ensemble* de populações, em contraposição à média referente aos membros de *uma* população no estado assintótico, $\widehat{w}$.

possível calcular essa probabilidade condicional para qualquer N, mas os desenvolvimentos posteriores exigirão a adoção um relevo multiplicativo. Seja X_i a carga mutacional do i-ésimo membro da amostra. É claro que todos os X_i seguem a mesma distribuição e é conveniente definir

$$\alpha_{m,n} \equiv P(X_i \geq m \mid n) = \sum_{j=m}^{\infty} e^{-U/s} \frac{(U/s)^{j-n}}{(j-n)!}, \tag{4.6}$$

que é a probabilidade de um indivíduo com pelo menos m mutações ser sorteado em uma população fundada por um n-mutante. A distribuição de probabilidade do mínimo de uma amostra aleatória, $Y_1 = \min_i X_i$, $i = 1, \ldots, N$, é bem conhecida [5] , sendo dada por

$$P(Y_1 = m \mid n, N) = \alpha_{m,n}^N - \alpha_{m+1,n}^N. \tag{4.7}$$

Contudo, essa generalização ainda não é suficiente. Também é preciso considerar todos os possíveis valores do tamanho da amostra quando ele é uma variável aleatória N', com distribuição $P(N')$. Portanto, ao invés da Eq. (4.1), a expressão adequada é

$$P_\tau(m) = \sum_{n=0}^{m} \sum_{N'} P_{\tau-1}(n) P(Y_1 = m \mid n, N') P(N'). \tag{4.8}$$

Felizmente, pelo menos no relevo multiplicativo, a equação acima é solúvel para algumas distribuições do tamanho do gargalo que parecem condizentes com o arranjo experimental. Em todos esses casos, a função geratriz (4.2) reduz-se à forma $G_\tau(z) = [g(z)]^\tau$. Se $g(z)$ admite a expansão em série de Taylor $g(z) = \sum_{k=0}^{\infty} g_k z^k$, onde

$$g_k = \left\langle \left\langle \alpha_{k,0}^N - \alpha_{k+1,0}^N \right\rangle \right\rangle \tag{4.9}$$

e $\langle\langle \cdot \rangle\rangle$ denota uma média pela distribuição $P(N')$, então, por igualdade dos coeficientes de potências de mesmo expoente,

$$P_\tau(m) = \sum_{i_1=0}^{\infty} \cdots \sum_{i_\tau=0}^{\infty} \left\{ \delta \left(m \,, \sum_{j=1}^{\tau} i_j \right) \prod_{j=1}^{\tau} g_{i_j} \right\}, \tag{4.10}$$

[5] As *estatísticas de ordem* Y_i, dadas pelo ordenamento crescente da amostra, *não* são equidistribuídas [19, 131]. Pela independência entre os X_i, $P(Y_1 \leq k) = 1 - P(Y_1 > k) = 1 - P(X_1 > k, \ldots, X_N > k) = 1 - [P(X_i > k)]^N$, de modo que $P(Y_1 = m) = P(Y_1 \leq m) - P(Y_1 \leq m-1) = [P(X_i \geq m)]^N - [P(X_i > m)]^N$.

onde $\delta(.,.)$ é a delta de Kronecker [6] . A Eq. (4.10), aparentemente complicada, possibilitaria a construção numérica de $P_\tau(m)$ de uma forma bem mais eficiente [7] do que a Eq. (4.8).

Entretanto, como no modelo original, o indivíduo mais apto entre os membros da amostra que dá origem a uma população infinita sempre sobrevive e determina o estado assintótico. Portanto, o que realmente torna a Eq. (4.10) relevante é o fato dela e a Eq. (4.4) poderem ser combinadas para determinar o q-ésimo momento do valor adaptativo médio da população no estado estacionário após a τ-ésima passagem,

$$\langle \widehat{w}^q \rangle = \sum_{m=0}^{\infty} \left[e^{-U}(1-s)^m \right]^q P_\tau(m) = e^{-qU}(\beta_q)^\tau, \tag{4.11}$$

onde

$$\beta_q \equiv \sum_{i=0}^{\infty} (1-s)^{qi} g_i. \tag{4.12}$$

Há evidência experimental [36] de que o logaritmo de $\langle \widehat{w} \rangle$ decai linearmente com o número de passagens τ. Esse comportamento é previsto pela Eq. (4.11). Quando $q = 1$,

$$\ln \langle \widehat{w} \rangle = -U + \tau \ln \beta_1 \tag{4.13}$$

e a taxa de decaimento, β_1, é positiva mas menor que um, implicando um coeficiente angular negativo. Além disso, nota-se que a taxa de mutação U pode ser obtida mediante o coeficiente linear da reta. Portanto, uma vez conhecido U e fixada uma distribuição de probabilidade para o tamanho dos gargalos, β_1 depende apenas de s e, pelo menos em princípio, o coeficiente seletivo também poderia ser determinado mediante a mensuração apenas de $\langle \widehat{w} \rangle$, em contraste com o caso mais simples.

Os autores de [26] já tinham feito considerações semelhantes, mas o caráter notável dos resultados aqui expostos é que, admitido um relevo multiplicativo, eles demonstram ser possível inferir os parâmetros biologicamente relevantes, s e U, em diversos protocolos de transferências seriais e não apenas no experimento de Chao.

---

[6] $\delta(a,b) = \begin{cases} 1, & \text{se } a = b; \\ 0, & \text{se } a \neq b. \end{cases}$

[7] Na verdade, devido à delta de Kronecker, não há contribuição aos somatórios quando qualquer um dos índices é maior que m.

4.2.3 Caracterização da taxa de decaimento linear

Dessa forma, é preciso estudar como a seleção e as mutações influenciam β_q. Parece que apenas o modelo original admite uma solução analítica geral, dada pela Eq. (4.5). Mas o comportamento de β_q quando s, U e N tendem a seus valores extremos pode ser determinado mediante uma análise cuidadosa das Eqs. (4.6), (4.9) e (4.12). É conveniente fixar o raciocínio no caso dos gargalos com tamanho fixo, embora as conclusões tenham validade geral.

O caso $N = 1$ corresponde a $\beta_q = e^{-U\theta_q}$. Por outro lado, quando $N \to \infty$, $\alpha_{i,0}^N \to \delta_{i,0}$ e $\beta_q \to 1$. Realmente, os indivíduos mais aptos tendem a ser sempre transferidos e $\langle \widehat{w} \rangle$ não decai. Contudo, essa tendência pode não se manifestar se os outros parâmetros impuserem um comportamento diferente a β_q. Por exemplo, enquanto $U \to 0$ tem exatamente o mesmo efeito de $N \to \infty$, $U \to \infty$ faz com que $\alpha_{i,0} \to 1$ e $\beta_q \to 0$. Quando $s \to 0$, também há uma redução no valor de β_q, mas não necessariamente tão intensa. É preciso notar que $\sum_{i=0}^{\infty} g_i = 1$, mas que praticamente toda a contribuição a esse somatório é dada pelos termos i pertencentes a algum intervalo centrado na média U/s e com amplitude proporcional ao desvio padrão $\sqrt{U/s}$. Nessa região, o termo $(1-s)^{qi}$ na Eq. (4.12) é constante,

$$\lim_{s\to 0} (1-s)^{q\left(\frac{U}{s} \pm \eta\sqrt{\frac{U}{s}}\right)} = \lim_{s\to 0} (1-s)^{q\frac{U}{s}} = e^{-qU}, \tag{4.14}$$

onde η é uma constante qualquer. Portanto, quando $s \to 0$, $\beta_q \to e^{-qU}$. Finalmente, se $s \to 1$, é fácil ver que $\beta_q \to 1 - (1 - e^{-U})^N$.

Esta última expressão é condizente com os dois comportamentos extremos de U e com $N \to \infty$. Se U for substituído por $U\theta_q$, os limites $N \to 1$ e $s \to 1$ também passam a ser satisfeitos. Dessa forma, parece razoável conjecturar que a taxa de decaimento β_q possa ser descrita pela equação

$$\beta_q = 1 - \left[1 - e^{-U\theta_q}\right]^{N^{h(s,U,N,q)}}, \tag{4.15}$$

desde que a hipotética função $h(s, U, N, q)$ seja tal que $h \to 0$ quando $s \to 0$ e $h \to 1$ quando $s \to 1$, o que é suficiente para satisfazer todas as restrições apresentadas. Quando N é uma variável aleatória, é preciso calcular a média na distribuição utilizada.

Embora a função $h = h(s) = s^{\gamma}$, com $\gamma \approx 0.5$, aproxime satisfatoriamente o comportamento exato de β_1 para valores pequenos de s, não foi possível obter uma expressão válida em todo o espaço de parâmetros e, mesmo que a forma funcional adotada para β_q esteja correta, é provável que h não admita uma representação em termos de um número finito de funções elementares. De forma geral, essa análise não levou a maiores progressos e o estudo numérico da expressão exata dada pela Eq. (4.12) revelou-se mais produtivo.

No apêndice D, discutem-se em detalhes as propriedades das distribuições de probabilidade empregadas para descrever a largura dos gargalos, inclusive as modificações necessárias para desconsiderar a possibilidade de gargalos de tamanho nulo. Na distribuição binomial (B), em que a variância sempre é menor que a média, a razão entre essas grandezas foi fixada em $0,8$. Por outro lado, o comportamento inverso ocorre na mistura Poisson-gama (PG) e aquela razão foi fixada em 10. A terceira distribuição empregada, a de Poisson (P), é conhecida pela igualdade entre média e variância e apresenta um comportamento intermediário entre as outras duas. Não houve diferenças qualitativas entre seus comportamentos e, para gargalos largos, não houve nem mesmo diferenças quantitativas. Portanto, nas figuras abaixo, uma legenda como $N = PG = 200$ significaria que, além da curva correspondente a um gargalo fixo de tamanho $N = 200$ e daquela que representa uma mistura Poisson-gama de média 200 serem coincidentes, também o são os gráficos das distribuições binomial e Poisson de mesmo valor médio.

A figura 4.2 ilustra o comportamento de β_1 em função de s. Quando $N = 1$, observa-se os valores constantes previstos em [26]. Entretanto, em geral, β_1 é uma função crescente da intensidade de seleção enquanto a catraca de Muller estiver ativa. Como mencionado acima, a incerteza na largura dos gargalos só é relevante quando N é pequeno. Nessa situação, quanto maior é a variância, mais lenta é a degradação mutacional, até mesmo porque as flutuações privilegiam, assimetricamente, amostras maiores que a média. Convém destacar, entretanto, que esse efeito já se mostra menos acentuado quando $N = 2$ do que para gargalos com tamanho médio 1. Para $U = 0.1$ e $U = 1$, $N = 10^3$ já é suficiente para β_1 ser indiferente a eventuais oscilações no processo de transferência serial.

Sem dúvida, o aspecto mais importante da figura 4.2 é o forte crescimento de β_1

quando o coeficiente seletivo é positivo mas muito próximo a 0. Isso mostra que $\langle \widehat{w} \rangle$ é extremamente sensível a variações em s justamente na condição biologicamente relevante, o regime de seleção fraca. Esse efeito torna-se mais acentuado à medida que os gargalos tornam-se mais largos, como mostram as curvas $U = 1$, $N = 10^3$ e $U = 1$, $N = 10^5$.

Taxas de mutação mais altas também aumentam essa sensibilidade. Elas afetam fortemente β_1, como pode ser observado na figura 4.3, onde é utilizada uma escala logarítmica para melhor evidenciar as diferenças entre as propriedades do modelo generalizado e o comportamento exponencial do caso $N = 1$. Como se vê, seleção forte é necessária para haver desvios significativos: gargalos enormes como $N = 10^5$ só começam a ser notados quando s é da ordem de 10^{-4}.

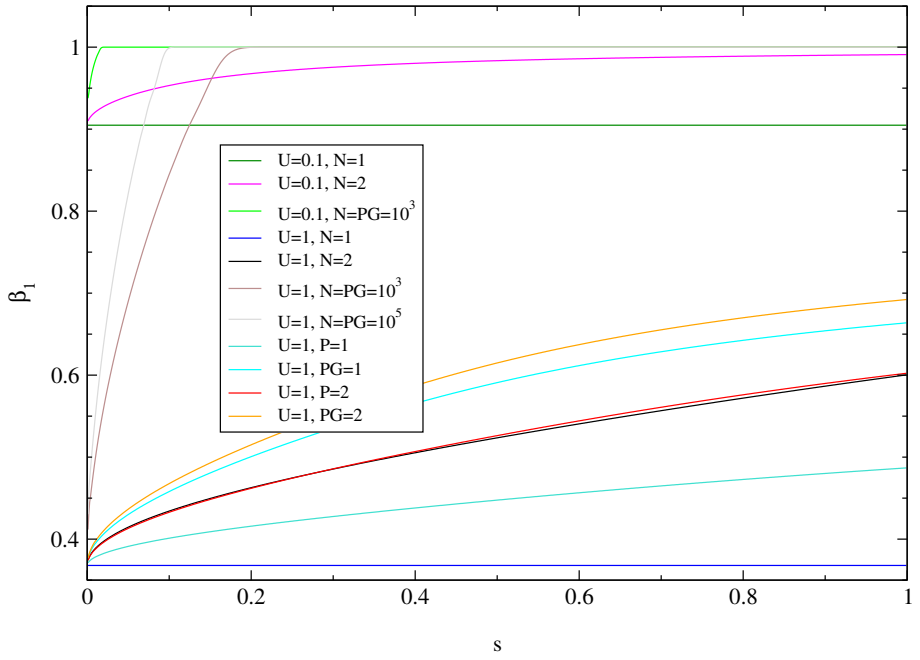

Figura 4.2: Influência da seleção em β_1.

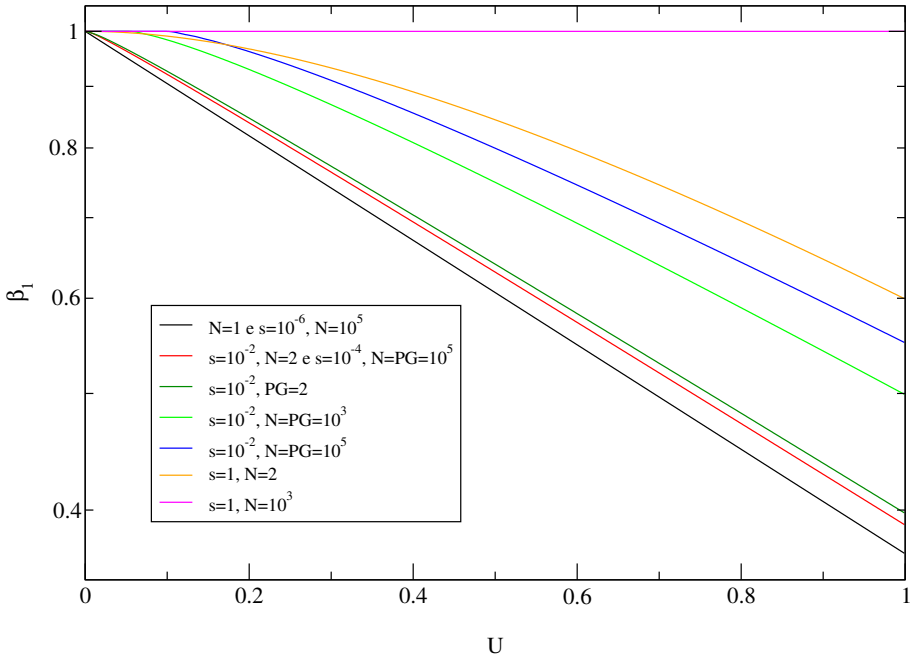

Figura 4.3: Efeitos da mutação em β_1 para s fixo, com escala logarítmica.

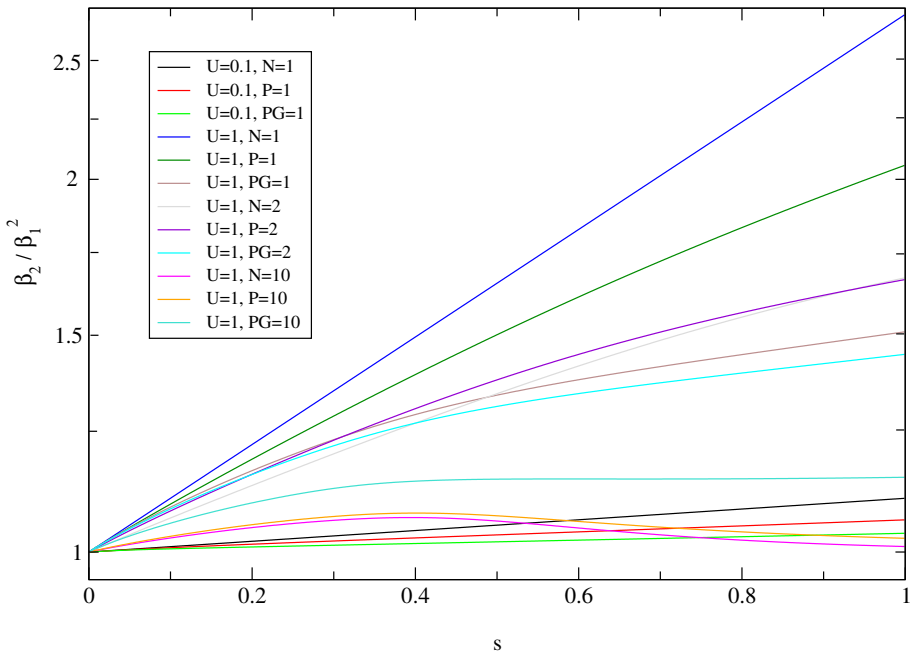

Figura 4.4: Gráfico mono-log da razão $\beta_2 \,/\, \beta_1^2$ em função da seleção.

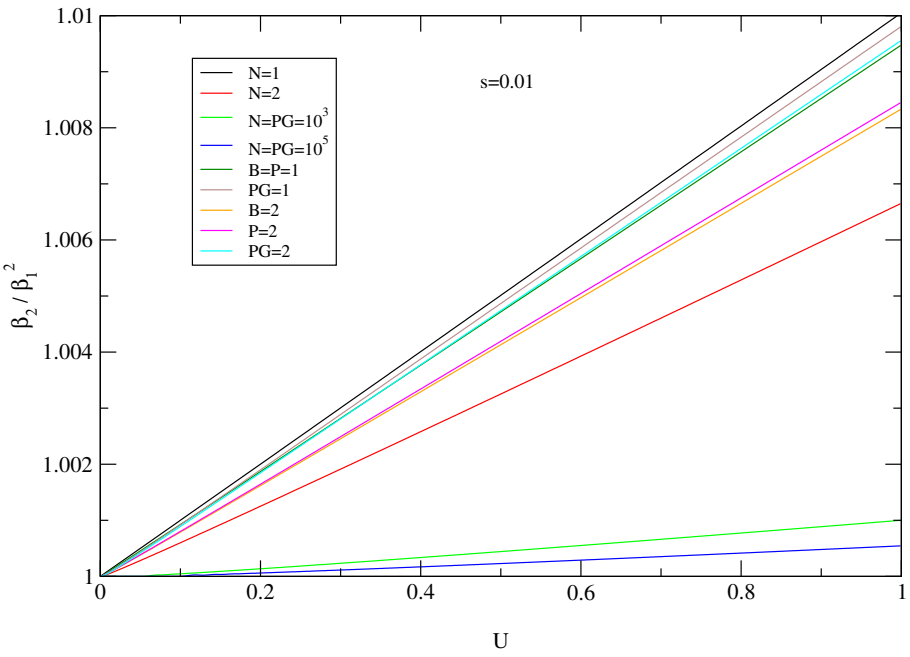

Figura 4.5: Influência da mutação em $\beta_2 \,/\, \beta_1^2$. A dependência é tão suave que a escala logarítmica não foi necessária.

O comportamento da razão $\rho \equiv \beta_2 \,/\, \beta_1^2$, que está diretamente relacionada àquele da variância de $\widehat{w}$, está ilustrado nas figuras 4.4 e 4.5. Considerando o resultado das transferências unitárias, $\rho_{N=1} = e^{sU}$, como uma primeira aproximação para o caso geral, não é surpreendente que ρ seja crescente, tanto em s quanto U, para várias combinações desses parâmetros. Mas esse comportamento não é universal, como pode ser visto nas curvas $U = 1$, $N = 10$ e $U = 1$, $P = 10$ da figura 4.4, embora as discrepâncias ocorram para valores altos de U e excessivamente altos de s. Para valores baixos dessas grandezas, o comportamento é exponencial e ρ é bem próxima de 1, como mostra a figura 4.5, que adota uma intensidade de seleção já considerada alta. Observa-se que passagens largas sempre reduzem a variabilidade no valor adaptativo e que, na situação oposta, as flutuações no tamanho dos gargalos tem um comportamento complexo, em nada lembrando a regularidade no comportamento de β_1.

4.2.4 Conclusões

O modelo analítico discutido nesta seção descreve adequadamente a dinâmica evolucionária de uma população sujeita a transferências seriais com gargalos arbitrários. Com efeito, o fato do decaimento do logaritmo do valor adaptativo com o número de passagens ser linear não depende da distribuição de probabilidade do tamanho dos gargalos, conforme a Eq. (4.13). Esse comportamento, na verdade, decorre da adoção do relevo multiplicativo. Após cada transferência, quando o equilíbrio é atingido, a distribuição assintótica é dada pela Poisson truncada da Eq. (3.6), que tem sempre o mesmo formato, apenas deslocado. Dessa forma, a cada passagem, a catraca dá o mesmo número médio de "cliques", que correspondem à mesma redução relativa do valor adaptativo. Portanto, os decaimentos lineares observados experimentalmente podem estar associados a ausência de epistase.

A generalização do esquema pioneiro de Colato e Fontanari pode ter conseqüências importantes. Como a amostragem de uma fração aleatória de cada cultura é bem descrita pelo novo formalismo, o procedimento experimental torna-se mais simples, pois não é necessário o isolamento de um microorganismo em cada passagem. Além disso, o modelo original requer medidas das flutuações de $\widehat{w}$ para estimar s, mas esses efeitos podem ser pequenos demais (ver figura 4.5) para gerar estimativas confiáveis. Por outro lado, o modelo generalizado só requer medidas de $\langle\widehat{w}\rangle$ para determinar U e s e tem grande sensibilidade ao coeficiente seletivo quando os gargalos são largos, como foi discutido na apresentação dos resultados. Altas taxas de mutação reduzem as incertezas na estimação em ambos os casos e podem ser induzidas artificialmente. Dessa forma, o formalismo introduzido parece ser uma contribuição interessante ao estudo da evolução molecular.

4.3 Processos de ramificação e catraca de Muller em populações sob expansão exponencial

A discussão da seção anterior admite que os mais aptos indivíduos coletados em cada transferência sempre sobrevivem ao processo de expansão, sendo responsáveis pela caracterização do estado eventualmente atingido pela população infinita em equilíbrio. Se os gargalos são estreitos, essa hipótese é certamente inadequada, pois as flutuações estatísticas são mais importantes justamente quando populações são pequenas. Nesta seção, argumenta-se que processos estocásticos de ramificação são uma ferramenta extremamente versátil para descrever os processos de rápido crescimento populacional empregados nessas situações. Para ilustrar esse ponto de vista, esse formalismo é utilizado para estudar analiticamente dois modelos elaborados por outros autores para suprir essa deficiência, mas que só foram explorados mediante simulações computacionais.

Grosseiramente, pode-se descrever um processo de ramificação [96] como um processo estocástico em que um objeto dá origem a um número aleatório de outros objetos, independentemente do resto do sistema. No caso mais simples, todos os objetos são da mesma natureza e, conseqüentemente, não há qualquer estrutura na população. Mas há uma generalização, denominada processo de ramificação de muitos tipos (PRMT), que foi concebida especialmente para lidar com situações em que objetos distintos geram sua "prole" com diferentes distribuições de probabilidade.

Nos modelos considerados nesta seção, a população organiza-se naturalmente em um número infinito, mas enumerável, de classes. Dessa forma, foi concebido um processo de ramificação de infinitos tipos que, embora careça de rigor matemático, apresentou resultados numéricos consistentes. Mesmo assim, resultados importantes podem ser obtidos a partir da teoria básica de processos de ramificacão. As populações consideradas evoluem em tempo discreto, com distribuições estáticas de tamanho da prole.

4.3.1 Processo de ramificação simples (PRS)

A ubíqua função geratriz (FG) revela-se uma ferramenta essencial também no estudo de processos de ramificação. No PRS, é conveniente definir uma FG para o tamanho da prole de um indivíduo,

$$\varphi(z) \equiv \sum_{m=0}^{\infty} P[M = m] z^m, \tag{4.16}$$

e uma outra associada ao tamanho total $N(t)$ da população no instante t,

$$\phi(z,t) \equiv \sum_{n=0}^{\infty} P[N(t) = n] z^n. \tag{4.17}$$

Essas duas funções relacionam-se entre si através da importante expressão

$$\phi(z,t) = \phi[\varphi(z), t-1], \tag{4.18}$$

que é uma conseqüência [96] da independência na reprodução. Quando o processo inicia-se com um fundador, $\phi(z,1) = \varphi(z)$ e a recorrência acima é equivalente à equação

$$\phi(z,t) = \varphi[\phi(z,t-1)]. \tag{4.19}$$

Da Eq. (4.17), segue que a probabilidade de que a população esteja extinta no instante t é dada por

$$\pi(t) \equiv P[N(t) = 0] = \phi(0,t) \tag{4.20}$$

e, a partir da Eq. (4.19), obtém-se a expressão

$$\pi(t) = \varphi[\pi(t-1)]. \tag{4.21}$$

A probabilidade de extinção eventual, $\pi(\infty)$, é tal que $\pi(\infty) = \varphi[\pi(\infty)]$. Pela definição de uma FG, $\pi(\infty) = 1$ sempre é uma raiz dessa equação.

Uma propriedade bem conhecida na teoria de processos de ramificação [96] é que o mapa descrito pela Eq. (4.21) admite um segundo ponto fixo, que corresponde à probabilidade de extinção eventual se, e somente se, $\langle M \rangle > 1$, onde $\langle M \rangle$ é o tamanho médio da prole de um indivíduo. A necessidade dessa condição é intuitivamente óbvia, haja vista que

uma linhagem que tende a ter cada vez menos componentes está fadada à extinção. Uma justificativa formal dessa propriedade é dada pelo limite assintótico da relação entre os valores esperados de M e $N(t)$,

$$\langle N(t)\rangle = \left\{\frac{\partial\phi(z,t)}{\partial z}\right\}_{z=1} = \langle M\rangle^t\,, \tag{4.22}$$

em que a primeira igualdade é uma conseqüência da Eq. (4.17) e a segunda, de qualquer uma das Eqs. (4.18) e (4.19). No ponto de bifurcação $\langle M\rangle = 1$, flutuações conduzem o sistema à extinção.

4.3.2 Processo de ramificação de infinitos tipos (PRIT)

Nesta seção, é conveniente discutir algumas notações detalhadamente. Qualquer vetor $\mathbf{u}$ deve ser visto como um conjunto de um número infinito de componentes cujo primeiro índice é 0, para denotar os indivíduos livres de mutações. Explicitamente, $\mathbf{u} = (u_0, u_1, \ldots)$. Também é conveniente definir $\mathbf{u} = \left(\mathbf{u}^{(j)}, \overline{\mathbf{u}^{(j)}}\right)$, onde $\mathbf{u}^{(j)} = (u_0, \ldots, u_{j-1})$ tem j componentes e $\overline{\mathbf{u}^{(j)}} = (u_j, u_{j+1}, \ldots)$, para qualquer $j \geq 1$. As mesmas regras aplicam-se a vetores constantes. Logo, $\mathbf{0}^{(j)}$ representa um vetor com j componentes nulos e, apesar de $\overline{\mathbf{1}^{(j)}}$ ter infinitos elementos e, portanto, parecer idêntico a $\mathbf{1}$, esses objetos são diferentes, haja vista que o primeiro índice a que se refere $\overline{\mathbf{1}^{(j)}}$ é j, e não 0.

No PRIT, $N_{ij}(t)$ é o número de indivíduos na classe j no instante t, dado um fundador na classe i, e o sistema é completamente caracterizado pela configuração $\mathbf{N_i}(t)$ e sua hierarquia de FGs,

$$\phi_i(\mathbf{z},t) \equiv \sum_{\mathbf{n}} P[\mathbf{N_i}(t) = \mathbf{n}] \prod_{j=0}^{\infty} z_j^{n_j}. \tag{4.23}$$

A FG associada à distribução de probabilidade conjunta da prole de um indivíduo do tipo i é

$$\varphi_i(\mathbf{z}) \equiv \sum_{\mathbf{m}} P[\mathbf{M_i} = \mathbf{m}] \prod_{k=0}^{\infty} z_k^{m_k}, \tag{4.24}$$

se M_{ij} descendentes enquadram-se na classe j. As generalizações adequadas [96] das Eqs. (4.18) e (4.19) para o PRIT são

$$\boldsymbol{\phi}(\mathbf{z},t) = \boldsymbol{\phi}[\boldsymbol{\varphi}(\mathbf{z}), t-1], \tag{4.25}$$

em geral, e

$$\boldsymbol{\phi}(\mathbf{z},t)=\boldsymbol{\varphi}[\boldsymbol{\phi}(\mathbf{z},t-1)], \tag{4.26}$$

quando há somente um fundador. Portanto, a FG que caracteriza a população em um certo instante depende das FGs na geração anterior obteníveis a partir de todos os possíveis tipos de fundadores.

Um resultado rigoroso da teoria do PRMT assegura que toda a população extingue-se assintoticamente se o maior autovalor da matriz $\langle\mathbf{M}\rangle=\{\langle M_{ij}\rangle\}$ com elementos estritamente positivos não exceder um [96]. Embora a matriz $\langle\mathbf{M}\rangle$ neste estudo seja infinita e tenha elementos nulos, como será mostrado adiante, essa propriedade ainda será obedecida.

Mas questões acerca de extinções adquirem um sentido mais amplo no PRIT, pois agora é possível estudar as condições de sobrevivência de algumas classes específicas na comunidade. Sejam

$$\pi_i(j,t)\equiv P\left[\mathbf{N_i}^{(j)}(t)=\mathbf{0}^{(j)}\right] \tag{4.27}$$

a probabilidade de que o menor índice de uma classe populada no instante t seja pelo menos j, dado um fundador na classe i, e

$$\sigma_i(j,t)\equiv P\left[\overline{\mathbf{N_i}^{(j)}(t)}=\overline{\mathbf{0}^{(j)}}\right] \tag{4.28}$$

a probabilidade de que o maior índice de uma classe populada no instante t seja no máximo $j-1$, dado um fundador na classe i. Pelas Eqs. (4.23) e (4.26), $\boldsymbol{\pi}(j,t)$ e $\boldsymbol{\sigma}(j,t)$ obedecem a equações de recorrência idênticas,

$$\boldsymbol{\pi}(j,t)=\boldsymbol{\phi}\left[\left(\mathbf{0}^{(j)},\overline{\mathbf{1}^{(j)}}\right),t\right]=\boldsymbol{\varphi}[\boldsymbol{\pi}(j,t-1)] \tag{4.29}$$

e

$$\boldsymbol{\sigma}(j,t)=\boldsymbol{\phi}\left[\left(\mathbf{1}^{(j)},\overline{\mathbf{0}^{(j)}}\right),t\right]=\boldsymbol{\varphi}[\boldsymbol{\sigma}(j,t-1)], \tag{4.30}$$

respectivamente. As condições iniciais quando há um fundador do tipo i são

$$\pi_i(j,0)=\begin{cases}1, & \text{se } j\leq i\\ 0, & \text{se } j>i\end{cases} \tag{4.31}$$

e $\sigma_i(j,0) = 1 - \pi_i(j,0)$.

Claramente, a probabilidade de qualquer conjunto de classes estar extinto obedece à mesma forma recursiva. Entretanto, os casos definidos acima são particularmente importantes, especialmente $\boldsymbol{\pi}(j,t)$, que determina o comportamento da catraca de Muller mediante a distribução de probabilidade do menor índice j de uma classe presente na geração t,

$$\Pi_{ij}(t) = \pi_i(j,t) - \pi_i(j+1,t), \tag{4.32}$$

se o fundador for do tipo i, e o valor médio do número mínimo de mutações na população,

$$\langle k_{\min}(t)\rangle = \sum_{j=1}^{\infty} j\Pi_{ij}(t) \tag{4.33}$$

conseqüentemente. De forma análoga,

$$\Sigma_{ij}(t) = \sigma_i(j+1,t) - \sigma_i(j,t) \tag{4.34}$$

é a probabilidade do maior índice j de uma classe ocupada na geração t por um descendente de um fundador do tipo i e

$$\langle k_{\max}(t)\rangle = \sum_{j=1}^{\infty} j\Sigma_{ij}(t). \tag{4.35}$$

Em uma analogia direta com o caso unidimensional, o valor médio de $N_{ij}(t)$, obtido de sua FG, Eq. (4.23), é

$$\langle N_{ij}(t)\rangle = \left\{ \frac{\partial \phi_i(\mathbf{z},t)}{\partial z_j} \right\}_{\mathbf{z}=\mathbf{1}}. \tag{4.36}$$

A Eq. (4.25) leva à expressão

$$\langle N_{ij}(t)\rangle = \sum_{k=0}^{\infty} \langle M_{kj}\rangle \langle N_{ik}(t-1)\rangle, \tag{4.37}$$

que, dada $\langle \mathbf{N}(\boldsymbol{t})\rangle = \{\langle N_{ij}(t)\rangle\}$, claramente leva à equação matricial

$$\langle \mathbf{N}(\boldsymbol{t})\rangle = \langle \mathbf{M}\rangle^t, \tag{4.38}$$

considerando que $\mathbf{N(0)} = \mathbf{I}$, em que $\mathbf{I}$ é a matriz identidade.

4.3.3 As linhagens em crescimento de Fontanari *et al*

O modelo

Recentemente, Fontanari, Colato e Howard (FCH) descreveram [57] um possível cenário para o funcionamento da catraca de Muller em linhagens asexuadas em crescimento. Eles elaboraram um modelo estocástico de tempo discreto que evolui a partir de um fundador sem mutações e onde cada membro da população contribui para a próxima geração com um número aleatório de filhos, dado por uma distribuição de Poisson de média R. Como no capítulo 3, o número de mutações em uma seqüência é a soma dos defeitos transmitidos via herança genética com os novos erros de replicação, cuja quantidade é distribuída segundo uma Poisson de média U. Portanto, a carga mutacional dos indivíduos naturalmente organiza a comunidade em um número infinito de classes.

Mas nem todos os filhos sobrevivem para reproduzir-se. No modelo de FCH, considera-se uma probabilidade de sobrevivência que varia inversamente com a carga mutacional, segundo a mesma função que descreve um relevo multiplicativo. Neste caso, entretanto, trata-se de valor adaptativo absoluto, pois o estado do resto da população não influencia o destino de qualquer seqüência. De fato, a essência desse modelo é o fato de cada membro da comunidade reproduzir-se ou morrer de forma completamente independente dos demais. Essa ausência de competição é uma condição necessária para que uma população exiba crescimento exponencial.

Em [57], FCH obtiveram alguns resultados analíticos no caso em que $s = 0$, ou seja, quando todos os filhos certamente sobrevivem, independentemente de seus defeitos. Especificamente, no caso neutro, FCH haviam obtido uma equação recursiva para a grandeza que determina o comportamento da catraca de Muller, a distribuição de probabilidade do número de mutações do indivíduo mais apto da população na geração t. Além disso, no mesmo regime, eles determinaram uma condição necessária para a sobrevivência assintótica da comunidade ($R > 1$) e uma condição necessária e suficiente para o travamento da catraca nas linhagens sobreviventes ($R > e^U$). No caso mais geral, eles afirmaram que a catraca de Muller permanece inativa para qualquer $s > 0$, mas tiveram que recorrer a

simulações para caracterizar o comportamento da população.

Aplicação do PRS

Vários desses resultados podem ser obtidos a partir da teoria básica de processos de ramificação. Inicialmente, considere-se o caso neutro. Se a carga mutacional de cada seqüência for desconsiderada, todos os membros da população são idênticos e a evolução do sistema é descrita exatamente por um PRS. Dessa forma, $R > 1$ é uma condição necessária à sobrevivência de toda a população quando $s = 0$. Além disso, a recorrência para a probabilidade de sobrevivência $\mathcal{S}(t) \equiv 1 - \pi(t)$ de uma linhagem no caso neutro dada em [57],

$$\mathcal{S}(t+1) = 1 - \exp[-R\mathcal{S}(t)], \tag{4.39}$$

é apenas um caso particular da Eq. (4.21) quando $\varphi(z)$ é a FG de uma Poisson de média R.

Um dos objetivos principais de FCH era a determinação das condições necessárias ao travamento da catraca de Muller. Após algum esforço, eles concluíram que $R > e^U$ é um critério necessário e suficiente no caso neutro. Mas esse resultado pode ser obtido simplesmente restringindo a atenção à subpopulação de seqüências mestras. Pelo teorema de decomposição invocado na seção anterior, cada indivíduo sem mutações dá origem a um número médio de $Rf(0 \mid U)g_0 = Re^{-U}$ descendentes idênticos a si próprio, segundo um PRS poissoniano. Como não há mutações reversas, nenhuma outra classe gera seqüências mestras. Portanto, se $Re^{-U} \leq 1$, essa subpopulação certamente se extingue. Quando $s = 0$, o mesmo argumento se aplica a qualquer classe de mutantes que, pela extinção de seus competidores, venha a dispor dos indivíduos com menor carga mutacional na população. A conclusão inevitável é que, se $R > 1$ mas $R \leq e^U$, nenhuma classe é estável e, conseqüentemente, a catraca de Muller nunca travará nas populações que eventualmente sobreviverem. Por outro lado, se $R > e^U$, há uma probabilidade positiva $\mathcal{S}(\infty)$ de que a classe com menor carga mutacional sobreviva ao processo de expansão, dada pela solução assintótica da Eq. (4.39) quando $R \to Re^{-U}$. Cada giro da catraca equivale a reiniciar

o processo, sempre com uma probabilidade positiva de sobrevivência. Portanto, se a população não se extinguir, alguma classe eventualmente sobreviverá e travará a catraca.

Um argumento bem simples esclarece porque a catraca de Muller nunca está ativa quando há decaimento. Por definição, se a catraca não trava, o número mínimo de mutações no sistema cresce irrestritamente. Além disso, a função w_j decresce monotonicamente a zero com a carga mutacional. Logo, para qualquer $s > 0$, em qualquer realização particular em que a catraca estivesse ativa, há um instante (aleatório, mas dependente de s) em que todos os indivíduos têm tantas mutações que a probabilidade de sobrevivência de qualquer um deles é arbitrariamente próxima a zero. Portanto, nenhuma população pode suportar um acúmulo incessante de mutações e a catraca certamente precisa estar travada em qualquer linhagem que sobreviva indefinidamente. Esse argumento obviamente se aplica a qualquer lei de decaimento que se anula assintoticamente. Uma outra conclusão inevitável é que $R > e^U$ revela-se como uma condição necessária para sobrevivência da comunidade na presença de decaimento, pois as seqüências mestras não são afetadas pela seleção. Se elas não tiverem chances de sobreviver, o mesmo ocorrerá com os mutantes. Dessa forma, extrapolando um resultado do PRMT para o PRIT, não parece ser coincidência que o maior autovalor de $\langle \mathbf{M} \rangle$ seja exatamente $\langle M_{00} \rangle = Re^{-U}$.

Aplicação do PRIT

Neste momento, é conveniente adotar uma notação mais explícita para a probabilidade de uma variável aleatória poissoniana de média λ assumir o valor k,

$$f(k \mid \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \tag{4.40}$$

e para sua FG,

$$\psi(z \mid \lambda) \equiv \sum_{k=0}^{\infty} f(k \mid \lambda) z^k = \exp[\lambda(z-1)]. \tag{4.41}$$

Processos de Poisson desfrutam de uma importante propriedade de decomposição [20] que afirma que, se cada objeto gerado por um processo de Poisson de média R for associado a uma certa classe i com probabilidade p_i e independentemente dos demais objetos, então

o tamanho da população da classe i é distribuído segundo uma distribuição de Poisson de parâmetro Rp_i e é independente das demais classes. Essa decomposição corresponde exatamente ao papel da mutação e da seleção por morte no modelo de FCH. A mutação faz uma "pré-classificação" da prole de um indivíduo nas diversas classes acessíveis naquele momento, mas a seleção estocasticamente escolhe alguns desses descendentes para serem direcionados a uma classe extra, que abriga os mortos, que é completamente irrelevante para a descrição do problema. Portanto, p_i na discussão acima deve levar em conta tanto a mutação quanto a probabilidade de decaimento, $w_i = (1-s)^i$.

Dessa forma, no modelo de FCH, M_{ij} é o número de indivíduos com j mutações que descendem diretamente de um i-mutante e

$$P[M_{ij} = m] = f[m \mid Rf(j-i \mid U)w_j] \tag{4.42}$$

quando $j \geq i$ e 0 nos demais casos. A matriz $\langle \mathbf{M} \rangle = \{Rf(j-i \mid U)w_j\}$ é triangular devido à ausência de mutações reversas. Além disso, a independência entre as classes faz com que, se $m_j = 0$ para todo $j < i$,

$$P[\mathbf{M_i} = \mathbf{m}] = \prod_{j=i}^{\infty} f[m_j \mid Rf(j-i \mid U)w_j]. \tag{4.43}$$

A forma explícita da FG do tamanho da prole,

$$\varphi_i(\mathbf{z}) = \prod_{j=i}^{\infty} \exp\left\{R\frac{e^{-U}U^{j-i}}{(j-i)!}w_j[z_j - 1]\right\}, \tag{4.44}$$

é obtida a partir das Eqs. (4.24), (4.41) e (4.43). Pela Eq. (4.29),

$$\pi_i(j,t) = \prod_{k=0}^{j-i-1} \exp\left\{R\frac{e^{-U}U^{k}}{k!}w_{i+k}[\pi_{i+k}(j,t-1) - 1]\right\}. \tag{4.45}$$

O produto torna-se finito porque $\pi_i(j,t) = 1$ quando $j \leq i$. Quando $s = 0$, $\pi_i(j,t)$ depende apenas de $j - i$ e, se $\mathcal{Q}_n(t) \equiv \pi_i(i+n,t)$,

$$\mathcal{Q}_n(t) = \prod_{k=0}^{n-1} \exp\left\{R\frac{e^{-U}U^{k}}{k!}[\mathcal{Q}_{n-k}(t-1) - 1]\right\}. \tag{4.46}$$

A condição inicial para esta equação é $\mathcal{Q}_n(0) = 0$. A evolução de $\boldsymbol{\sigma}(j,t)$ é dada pela Eq. (4.45) com o produto ilimitado.

Neste modelo, o fundador sempre foi escolhido como uma seqüência mestra. As figuras 4.6 e 4.7 ilustram a dependência temporal dos valores médios das cargas mutacionais mínima e máxima na população, respectivamente, considerando apenas as realizações em que a população não esteja extinta. Portanto, os gráficos apresentados são obtidos mediante a divisão das grandezas das Eqs. (4.33) e (4.35) pela probabilidade $1-\pi_0(\infty,t)$ (ou $1-\sigma_0(0,t)$) de alguma classe estar populada no instante t, embora a notação não indique isso explicitamente.

Na figura 4.6, observa-se claramente o travamento da catraca de Muller para qualquer $s>0$, como antecipado por FCH. Além disso, como $R>e^U$, o mesmo ocorre no regime neutro. Entre as curvas apresentadas, nota-se que $\langle k_{\min}\rangle$ anula-se assintoticamente quando $s \geq 0.3$. Há uma explicação simples para esse comportamento, que pode ser usada no cálculo de cotas superiores para $\langle k_{\min}\rangle$. Os indivíduos com apenas uma mutação contribuem para essa grandeza somente se houver realizações em que eles possuam a carga mutacional mínima. Ora, se isso ocorrer, essa classe de indivíduos não recebe contribuições de qualquer outra e só tem chances de sobreviver indefinidamente se gerar, em média, mais de um representante da própria classe a cada geração. Para ilustrar esse argumento, destaca-se que $Re^{-U}(1-s)\approx 1.07$ para $s=0.2$, enquanto $Re^{-U}(1-s)\approx 0.94$ para $s=0.3$. Em geral, $\langle k_{\min}\rangle$ é limitado superiormente pelo menor i tal que $Re^{-U}(1-s)^i<1$.

Como o acúmulo excessivo de mutações resulta em morte quando há decaimento, era previsível que $\langle k_{\max}\rangle$ fosse limitado, como realmente se vê na figura 4.7. Após algum tempo, todos os gráficos para $s>0$ estabilizam-se (a curva para $s=0.03$ não ultrapassa o valor 61, por exemplo). Por outro lado, a carga mutacional máxima cresce indefinidamente quando $s=0$. Como seria esperado intuitivamente, a figura 4.7 mostra que, quanto mais forte é a seleção, menor é a carga mutacional máxima. Mas isso chama a atenção para uma importante peculiaridade da figura 4.6 que ainda não foi discutida.

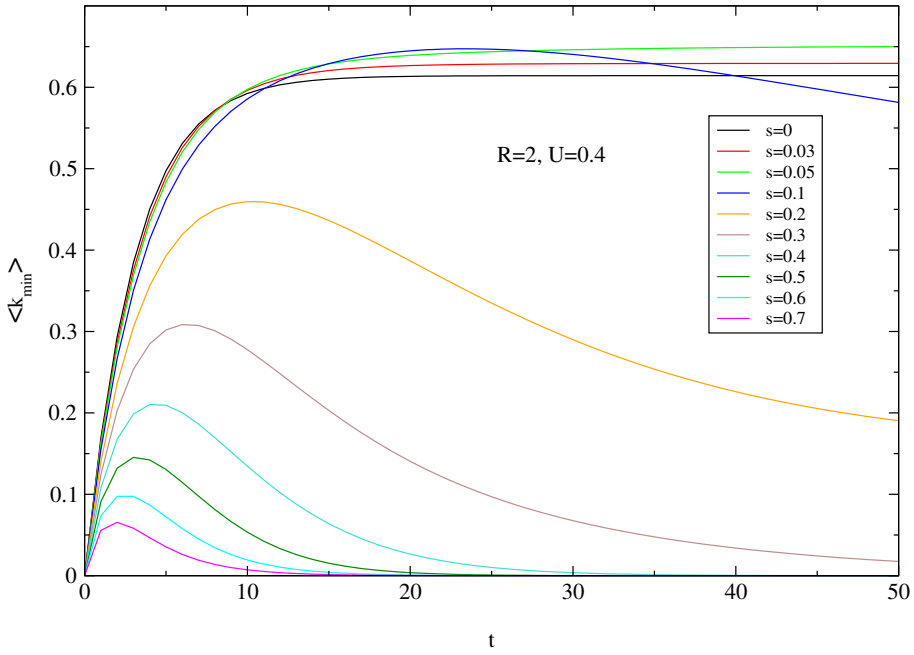

Figura 4.6: Valor médio do número de mutações na classe mais apta em função do tempo, dado que a população não está extinta.

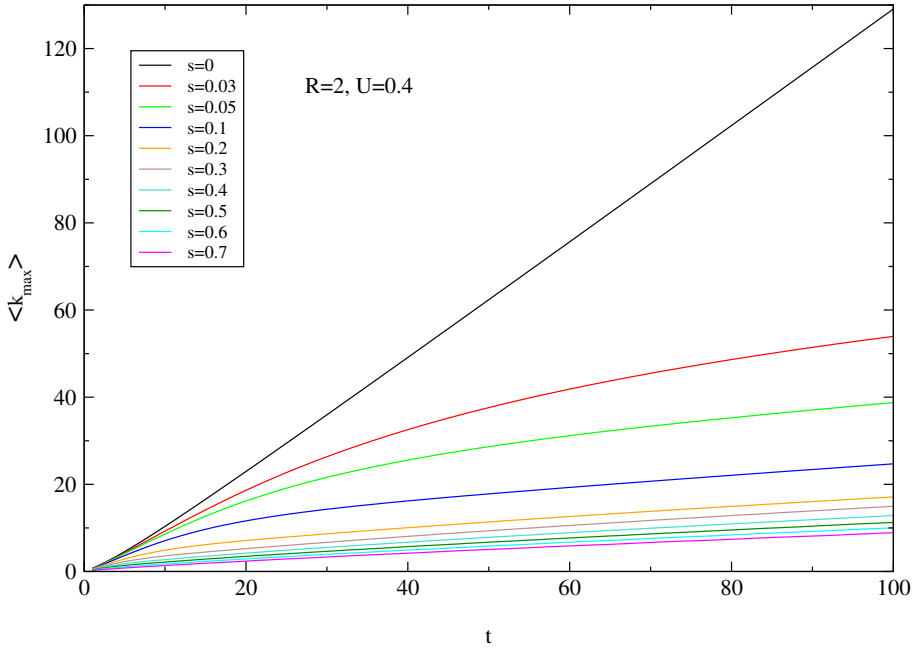

Figura 4.7: Comportamento dinâmico do número médio de mutações do indivíduo mais defeituoso da população, admitida a sobrevivência da população.

Como FCH já haviam observado em suas simulações, é possível que $\langle k_{\min} \rangle$ cresça com s durante algumas gerações, para valores baixos do coeficiente seletivo. Entretanto, esse fenômeno, agora ratificado por uma análise exata, ainda precisa de uma justificativa satisfatória, haja vista que a análise de FCH implicitamente admite competição dentro da população. Antes de tudo, é importante destacar que essa aceleração da catraca em resposta ao crescimento de s também é observada no valor médio incondicional de $\langle k_{\min} \rangle$ e que, implicitamente, o raciocínio intuitivo visualiza a ação dos diferentes níveis de seleção em um mesmo contexto. Entretanto, em cada geração, o decaimento atua em diferentes distribuições do número mínimo de mutações, construídas ao longo do tempo já sob o efeito de s. O ponto fundamental é que, quando s é pequeno, o crescimento da população e a pressão mutacional paulatinamente deslocam o espectro no sentido do acúmulo de mutações, sem que haja muitas mortes. Além disso, há pouca diferença entre as penalizações das diferentes classes quando o decaimento é suave. Esses dois efeitos, combinados, viabilizam a aceleração da catraca, cuja magnitude e efetiva manifestação dependem da combinação de todos os parâmetros do modelo (incluindo o tempo).

Na falta de resultados analíticos, é fundamental estudar numericamente como o decaimento afeta a probabilidade de sobrevivência da população, haja vista que não há sentido em analisar populações inviáveis. Como mostra a figura 4.8, quanto maior s, maior é a chance de extinção, como não poderia deixar de ser. Mas é interessante notar que há uma saturação desse efeito. O valor máximo da probabilidade de sobrevivência assintótica é dado pela solução de equilíbrio da Eq. (4.39), enquanto o valor mínimo também segue da mesma equação, mas quando $R \to Re^{-U}$, porque claramente a curva para $s = 1$ também atinge esse valor e, nesse caso, a população é constituída apenas pelos indivíduos livres de mutações.

No modelo de gargalos seriais de [26] e na primeira metade deste capítulo, admitiu-se que os fundadores de cada nova cultura nunca eram perdidos. Assim, é uma questão natural investigar se o modelo de FCH pode oferecer algum suporte a essa hipótese. Como mostra a figura 4.9, isso não ocorre. Ela mostra que, para valores realísticos de s, próximos a zero, a probabilidade de permanência da classe do fundador pode ser bem menor que 1.

Na verdade, esses resultados estão até mesmo superestimados, de certa forma, pois estão condicionados à sobrevivência da população. Como era esperado, a permanência é mais provável quando a comunidade sobrevive a decaimentos mais fortes.

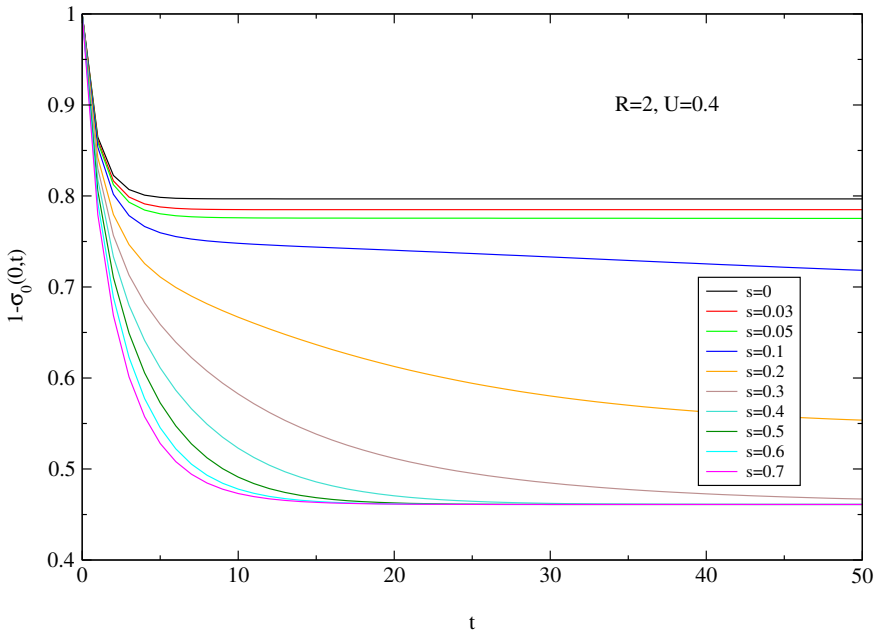

Figura 4.8: Probabilidade de que a população não esteja extinta no instante t.

Todos os resultados discutidos até agora foram baseados no decaimento multiplicativo empregado por FCH, mas a equação recursiva obtida pelo PRIT possibilita o estudo da catraca de Muller para qualquer função w_j. As próximas 4 figuras comparam os efeitos de diferentes relevos adaptativos, no papel de leis de decaimento. Foram considerados relevos truncados em $K = 0$ (pico agudo) e $K = 10$, além de perturbações epistáticas ao relevo multiplicativo, dadas pela Eq. (2.3) e dimensionadas pelo expoente α. A figura 4.10 mostra o travamento da catraca em todos os casos. O decaimento de pico agudo, que penaliza igualmente todas as classes mutantes, leva $\langle k_{\min} \rangle$ a valores maiores do que o decaimento truncado, mais ameno. Mas o comportamento aguardado à primeira vista seria justamente o oposto, como mostra a variação da epistase, de sinergética para atenuante. Isso reforça o argumento apresentado para justificar a aceleração da catraca na figura 4.6.

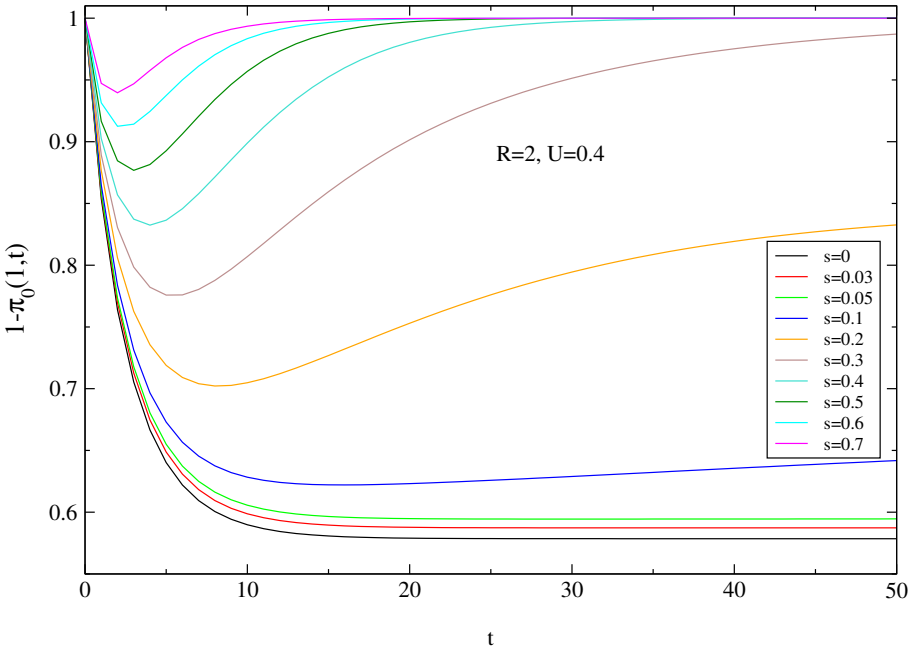

Figura 4.9: Dependência temporal da probabilidade de permanência da classe mais apta, condicionada à sobrevivência da população.

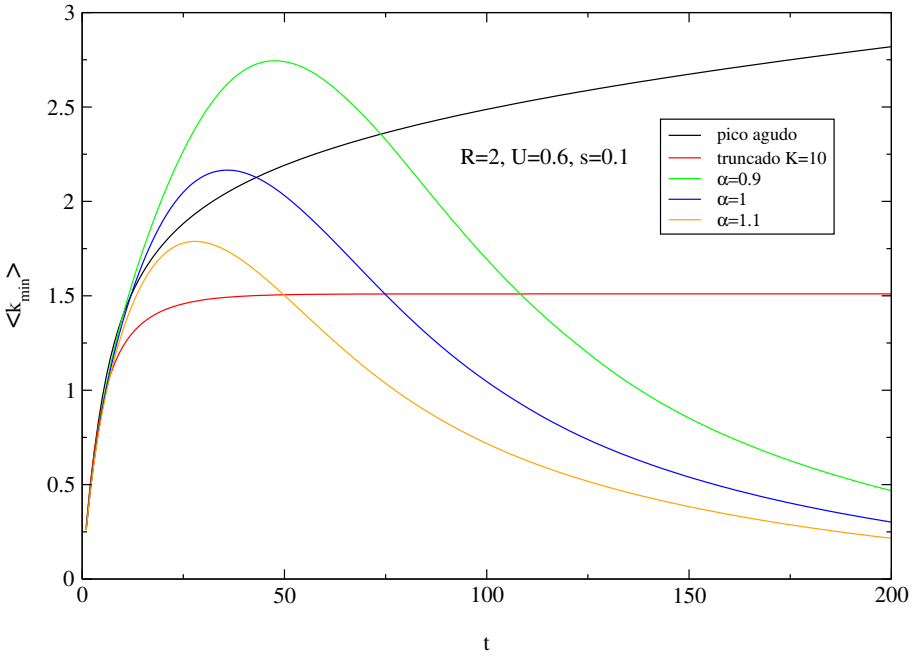

Figura 4.10: Efeitos de diferentes leis de decaimento na dependência temporal do número médio de mutações na classe mais apta, dado que a população não está extinta.

Pela figura 4.11, há combinações de parâmetros em que o acúmulo de mutações parece ser ilimitado nos relevos truncado e de pico agudo, exatamente como ocorre no regime neutro (ver figura 4.7), que também permite, em princípio, a sobrevivência de indivíduos com uma quantidade arbitrária de mutações. Assim como nas figuras 4.12 e 4.13, que ilustram a dependência temporal das probabilidades de sobrevivência da população e de permanência da classe mais apta, respectivamente, a variação gradual na intensidade dos decaimentos (do mais ameno para o mais intenso: truncado, pico agudo, epistase atenuante, multiplicativo, epistase sinergética) induz a esperada resposta monotônica em $\langle k_{\max} \rangle$ na figura 4.11. Portanto, parece que apenas o número mínimo de mutações pode apresentar comportamentos complexos.

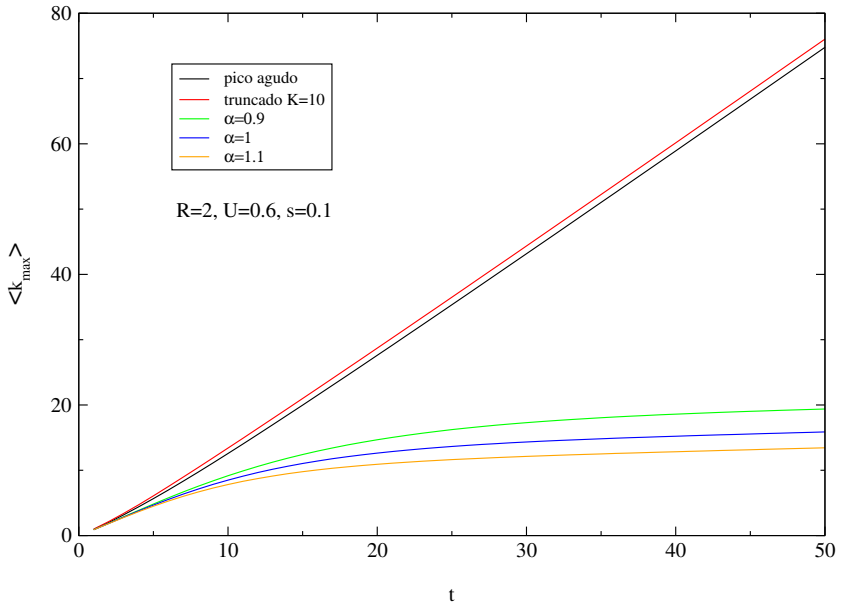

Figura 4.11: Média do número máximo de mutações nas populações que se mantêm ao longo do tempo.

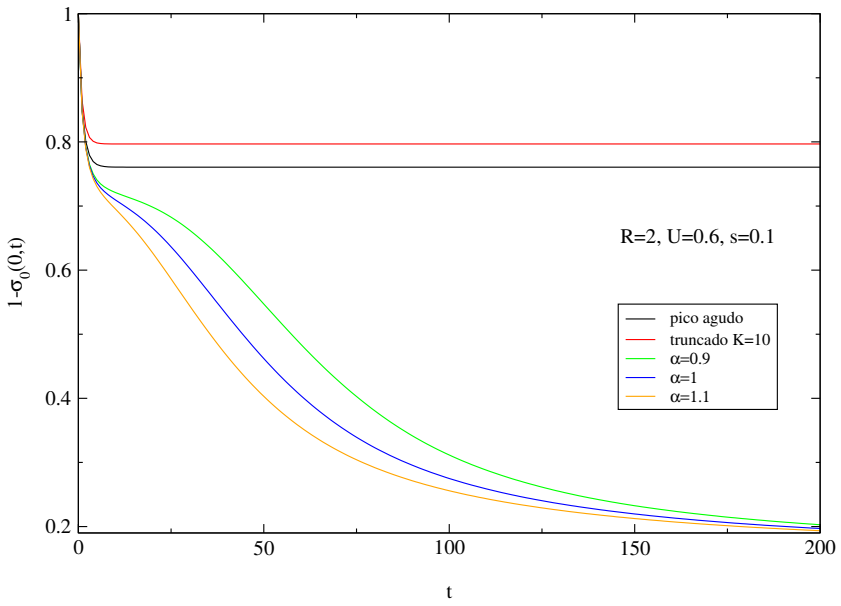

Figura 4.12: Efeitos do decaimento na probabilidade de que a população não esteja extinta no instante t.

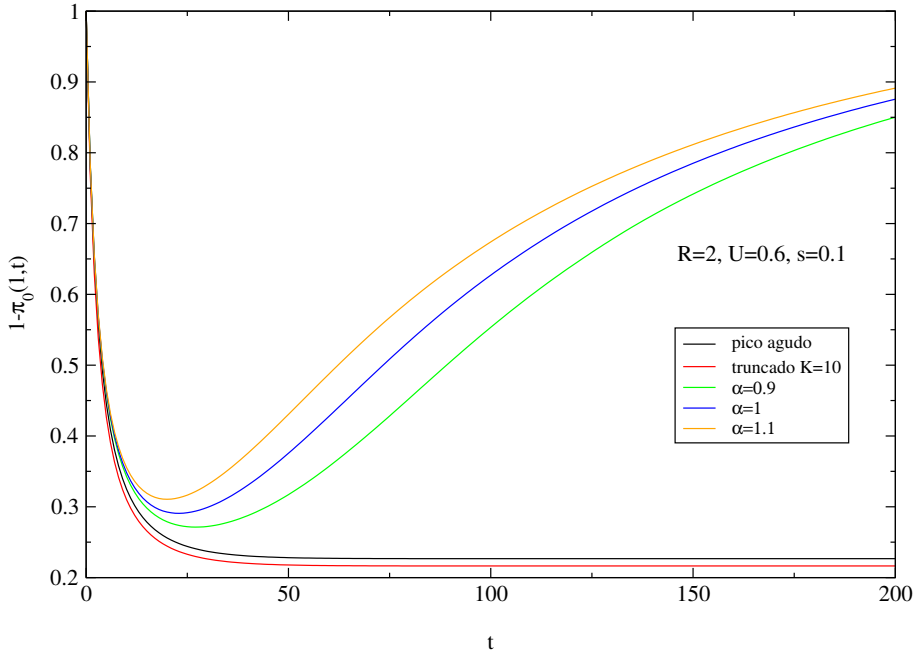

Figura 4.13: Comportamento da probabilidade condicional de permanência da classe mais apta em função do tempo, para vários tipos de decaimento.

A fertilidade média R e a taxa de mutação U foram mantidas constantes ao longo desta exposição, mas seus efeitos merecem alguns comentários. O crescimento na pressão mutacional aumenta as cargas mutacionais máxima e mínima, além de diminuir as probabilidades de sobrevivência e permanência discutidas acima, como era esperado. Por outro lado, maiores valores de R aumentam essas probabilidades, mas induzem crescimento em $\langle k_{\max} \rangle$ e decrescimento em $\langle k_{\min} \rangle$. Isso ocorre porque o tamanho médio da população aumenta em todas as classes, até mesmo naquelas com baixas probabilidades de ocupação, como as mais extremas. As flutuações também aumentam com R e reforçam ainda mais esse efeito. É como um processo difusivo ou uma gota de líquido se espalhando em todas as direções possíveis.

Finalmente, a forma multiplicativa da probabilidade de sobrevivência possibilita uma expressão explícita para $\langle N_{ij}(t) \rangle$. A partir da Eq. (4.37), obtém-se

$$\langle N_{ij}(t) \rangle = \sum_{k=i}^{j} R f(j-k \mid U) g_j \langle N_{ik}(t-1) \rangle \,, \tag{4.47}$$

que pode ser resolvida segundo o procedimento exposto na seção 3.3, ou seja, resolvendo uma recorrência para uma FG. O resultado é

$$\langle N_{ij}(t) \rangle = [R e^{-U} (1-s)^i]^t \frac{[\eta(t)]^{j-i}}{(j-i)!}, \tag{4.48}$$

onde

$$\eta(t) = U(1-s)\theta_t \tag{4.49}$$

é a razão entre o número esperado de mutações na população, $\sum_{j=i}^{\infty} j \langle N_{ij}(t) \rangle$, e o tamanho médio da população, $\sum_{j=i}^{\infty} \langle N_{ij}(t) \rangle$. Além disso, θ_t é dado pela Eq. (3.10).

Aparentemente, FCH interpretaram $\eta(t)$ como a carga mutacional média no instante t, ou seja, como o valor esperado da razão entre o número de mutações na população e seu tamanho, uma grandeza que, sem sombra de dúvida, seria uma cota superior para a carga mutacional do indivíduo melhor adaptado. Como $\eta(t)$ é sempre finita para $s > 0$, conseqüentemente a catraca de Muller estaria sempre inativa. A conclusão está correta, mas a razão entre duas médias pode ser completamente diferente da média da razão, qualquer que seja a relação de dependência entre as variáveis aleatórias.

4.3.4 O modelo de Lázaro *et al*

Dinâmica evolucionária

Em [114], Lázaro, Escarmís, Domingo e Manrubia (LEDM) introduziram um modelo numérico para descrever o comportamento de uma população viral sujeita a transferências seriais com gargalos unitários. Entretanto, enquanto Colato e Fontanari tinham admitido que o fundador de cada nova cultura sempre sobrevivia, LEDM adotaram uma dinâmica de crescimento estocástico semelhante ao modelo de FCH, mas que incorpora a possibilidade da ocorrência de mutações benéficas. A análise que se segue envolve apenas a etapa de expansão desse modelo.

No esquema de LEDM, um indivíduo pertence a uma classe indexada por um número natural k que é uma medida do valor adaptativo de seus membros, pois o tamanho da prole de cada um deles é dado por uma distribuição de Poisson de média k. Um filho pode sofrer uma mutação deletéria com probabilidade p e pertencer à classe $k-1$, ter valor adaptativo $k+1$ por receber alterações benéficas (probabilidade q) ou ter o mesmo valor adaptativo de seu pai, com probabilidade $1-p-q$, se não for alterado ou sofrer apenas mutações neutras.

Mediante simulações, LEDM estudaram o valor adaptativo médio da população em função do número de transferências. Inicialmente, eles observaram um decaimento exponencial idêntico àquele discutido na seção anterior para a carga mutacional mínima. Após muitas passagens, contudo, suas simulações indicaram a presença de um estado estacionário de baixo valor adaptativo e fortes flutuações, previamente observado em experimentos. Posteriormente, os mesmos autores fizeram uma análise de campo médio para uma versão simplificada do modelo, em que o tamanho da prole é determinístico [123].

Aplicação do PRIT

No modelo de LEDM, M_{ij} é o tamanho da prole do tipo j que descende diretamente de um indivíduo de valor adaptativo i. Invocando novamente a decomposição dos processos

de Poisson,

$$P[M_{ij} = m] = f[m \mid i\,T_{ij}], \tag{4.50}$$

onde a probabilidade de um certo filho de um i-indivíduo ser do tipo j é

$$T_{ij} = p\,\delta_{j,i-1} + (1-p-q)\,\delta_{j,i} + q\delta_{j,i+1}. \tag{4.51}$$

Se $m_j = 0$ quando $\mid j - i \mid > 1$,

$$P[\mathbf{M_i} = \mathbf{m}] = \prod_{j=i-1}^{i+1} f[m_j \mid i\,T_{ij}]. \tag{4.52}$$

A FG do tamanho da prole é

$$\varphi_i(\mathbf{z}) = \prod_{j=i-1}^{i+1} \exp\left[i\,T_{ij}(z_j - 1)\right] \tag{4.53}$$

e tanto $\boldsymbol{\pi}(j,t)$ quanto $\boldsymbol{\sigma}(j,t)$ podem substituir $\boldsymbol{\phi}(\mathbf{z},t)$ na equação recursiva

$$\phi_i(\mathbf{z},t) = \prod_{k=i-1}^{i+1} \exp\left\{i\,T_{ik}[\phi_k(\mathbf{z},t-1) - 1]\right\}. \tag{4.54}$$

Embora o modelo de LEDM tenha tido sucesso em reproduzir curvas experimentais do valor adaptativo médio de populações virais, ele não parece ser apropriado ao estudo da catraca de Muller. O problema é que, em qualquer população que sobreviva indefinidamente, todas as classes têm uma quantidade de membros que diverge quando $t \to \infty$. Por menores que sejam p e q (mas positivos), sempre há alguma geração em que a classe de um fundador qualificado (para valores adaptativos a partir de 2, já há uma alta probabilidade de sucesso reprodutivo) fica povoada o suficiente para que as mutações tornem-se corriqueiras. Dessa forma, é evidente que o valor médio $\langle k_{\min} \rangle$ do menor índice de uma classe com representantes na população eventualmente se anula, como mostra a figura 4.14, enquanto a figura 4.15 mostra que os indivíduos mais aptos aprimoram-se incessantemente.

Mesmo assim, vale a pena comentar alguns efeitos interessantes. A figura 4.14 será analisada inicialmente. Em geral, q praticamente não influencia $\langle k_{\min} \rangle$. Mesmo valores altos como $q = 0.1$ (não mostrados) deslocam apenas levemente as curvas. A única

exceção ocorre na rara situação em que um fundador de valor adaptativo 1 inicia uma população que consegue se manter viva, onde uma mudança de $q = 0.0001$ para $q = 0.001$ tem conseqüências perceptíveis. Entretanto, nota-se que esse aumento, que desloca a prole para classes mais altas, favorece o decrescimento de $\langle k_{\min} \rangle$. Esse comportamento é idêntico aquele induzido pelo parâmetro R no modelo de FCH: o crescimento da população favorece o alargamento da distribuição de classes povoadas. Os incrementos em p induzem respostas no sentido convencional.

Na figura 4.15, ocorre exatamente o contrário. Em geral, há baixa sensibilidade a p, enquanto q tem o comportamento esperado de potencializar $\langle k_{\max} \rangle$. A única surpresa é o crescimento observado quando $q = 0.001$, o fundador é da classe 1 e p sobe de 0.05 para 0.15. O fato é que, à beira da "esterilização" (a classe zero é inativa), esse incremento aumenta bastante a probabilidade de extinção da população. Nos raríssimos casos em que a extinção não ocorre, há uma "explosão demográfica" nas primeiras gerações, responsável pelo comportamento observado.

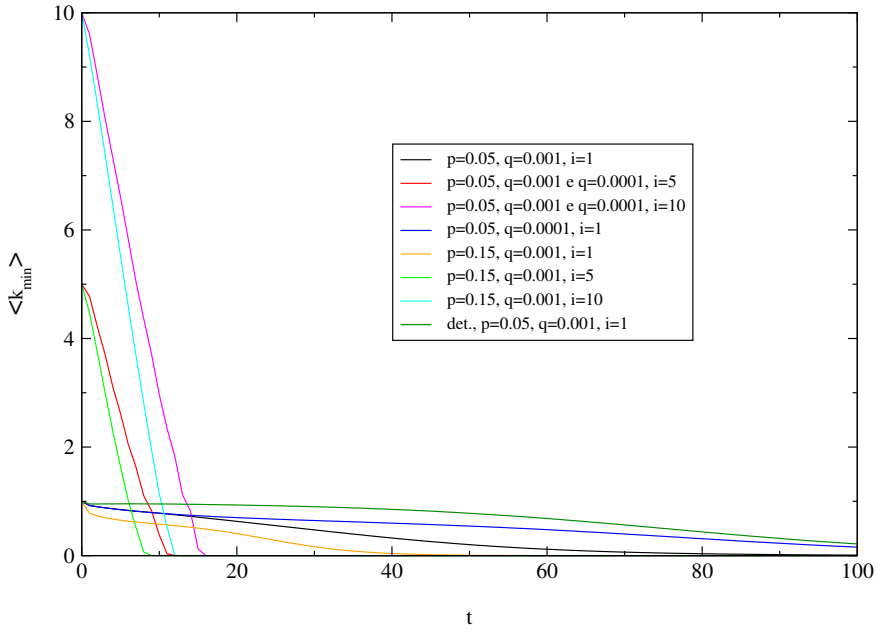

Figura 4.14: Valor médio do menor índice de uma classe com representantes na população em função do tempo, dado que a população não está extinta. O índice i denota a classe do fundador.

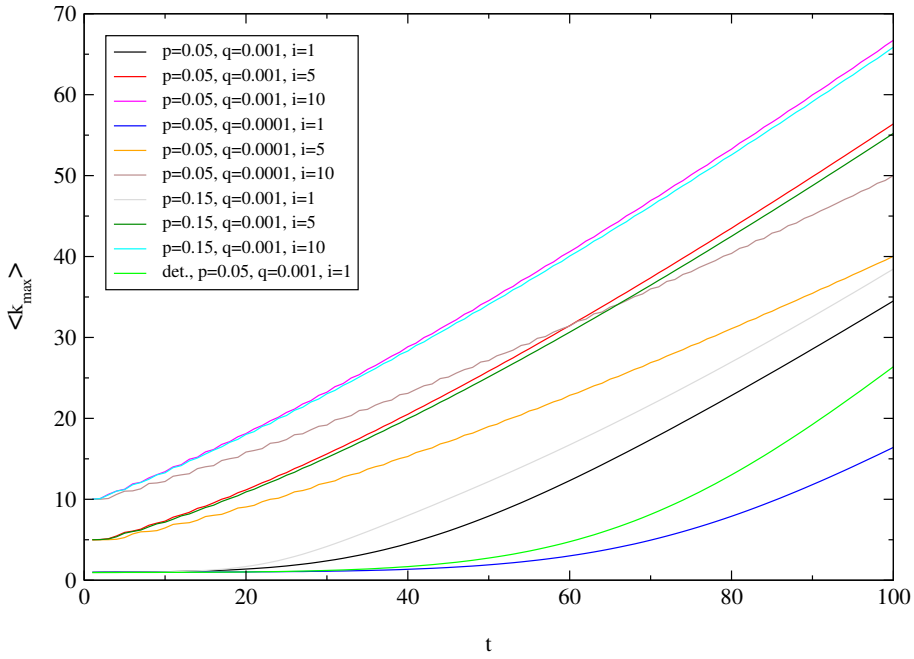

Figura 4.15: Valor médio do maior índice de uma classe com representantes na população em função do tempo, quando a população não está extinta. O índice i denota a classe do fundador.

Ainda é preciso esclarecer o significado das curvas denotadas por *det*. Trata-se da simplificação que LEDM analisaram em um segundo trabalho [123], em que o tamanho da prole de um indivíduo é exatamente igual a seu valor adaptativo e não mais uma variável aleatória. Os efeitos dessa ausência de flutuações só se fazem sentir quando a população é bem pequena e cresce suficientemente devagar. Se o fundador tiver valor adaptativo 5, a diferença já é imperceptível. Nota-se que esse modelo modificado não consegue atingir valores tão extremos em $\langle k_{\min} \rangle$ e $\langle k_{\max} \rangle$ quanto o original, como esperado.

Essa caso determinístico também foi estudado pelo formalismo do PRIT que, afinal de contas, pode ser aplicado a quaisquer distribuições do tamanho de prole, mesmo as triviais. Quando o tamanho da prole é poissoniano, o teorema de decomposição assegura que a distribuição conjunta é igual ao produto das distribuições marginais. Essa propriedade é muito útil, mas não é essencial. O ponto fundamental consiste em saber se é possível calcular a distribuição de probabilidade conjunta $P[\mathbf{M_i} = \mathbf{m}]$ e sua FG. A resposta é

afirmativa no caso determinístico, pois M_{ij} é uma variável binomial,

$$P[M_{ij} = m] = \begin{pmatrix} i \\ m \end{pmatrix} (T_{ij})^m (1 - T_{ij})^{i-m}, \tag{4.55}$$

e, se $m_j = 0$ quando $\mid j - i \mid > 1$, $\mathbf{M_i}$ segue uma distribuição multinomial,

$$P[\mathbf{M_i} = \mathbf{m}] = i! \prod_{j=i-1}^{i+1} \frac{(T_{ij})^{m_j}}{m_j!}, \tag{4.56}$$

cuja FG,

$$\varphi_i(\mathbf{z}) = \left[\sum_{j=i-1}^{i+1} T_{ij}\, z_j \right]^i \tag{4.57}$$

é facilmente obtida pelo teorema multinomial.

Novamente, tanto $\boldsymbol{\pi}(j,t)$ quanto $\boldsymbol{\sigma}(j,t)$ podem substituir $\boldsymbol{\phi}(\mathbf{z},t)$ na equação

$$\phi_i(\mathbf{z},t) = \left[\sum_{k=i-1}^{i+1} T_{ik}\, \phi_k(\mathbf{z}, t-1) \right]^i . \tag{4.58}$$

4.3.5 Conclusões

Os processos de ramificação e o PRIT, em particular, são ferramentas extremamente versáteis. Eles viabilizam o estudo de uma grande variedade de dinâmicas estocásticas de crescimento populacional. Como um exemplo simples, é possível construir um modelo híbrido daqueles de FCH e LEDM, em que as transições entre classes no modelo de FCH são governadas pela matriz de transição de LEDM, ao invés da distribuição de Poisson. Toda a dinâmica da população descrita por esse modelo é descrita pela equação de recorrência

$$\phi_i(\mathbf{z},t) = \prod_{k=i-1}^{i+1} \exp\left\{ R\, T_{ik}\, w_k[\phi_k(\mathbf{z}, t-1) - 1] \right\}. \tag{4.59}$$

Na verdade, foi realizada uma análise desse modelo, mas os resultados observados foram qualitativamente semelhantes àqueles já discutidos e não justificam maior atenção. Mas esse exemplo já mostra como diversas dinâmicas evolucionárias podem ser automaticamente incorporadas em um mesmo formalismo. Em princípio, qualquer matriz $\{T_{ij}\}$ pode ser adotada. Também é interessante a perspectiva de integração dessa técnica a modelos de protocolos experimentais.

Capítulo 5

Genealogias e testes de neutralidade

5.1 Introdução

Como foi discutido no capítulo 1, a controvérsia neutralista-selecionista é um tópico importante na genética de populações desde a década de 60. Isso ocorre porque, muitas vezes, é extremamente difícil detectar sinais da seleção natural no nível molecular. Nesse sentido, a teoria neutra de Kimura desempenha um importante papel, mesmo que eventualmente fique comprovado que ela não seja a melhor resposta ao problema do excesso de polimorfismos na natureza.

O estudo das propriedades estatísticas de modelos neutros em genética de populações é justificado simplesmente pelo fato deles oferecerem bons padrões de comparação para os resultados gerados pela análise de seqüências moleculares reais. Por exemplo, se uma árvore genealógica construída a partir de uma amostra de seqüências de DNA tem características significativamente distintas daquelas previstas pela teoria de evolução neutra, isso sugere atuação de seleção naquela população. Na verdade, tal resultado pode refletir apenas a inadequação de alguma outra hipótese do modelo adotado e a busca por técnicas capazes de discernir entre as possíveis causas das discrepâncias observadas é um tema atual de pesquisa, embora não seja esse o enfoque no presente trabalho.

O termo "significativamente distinto" adquire um sentido bem preciso dentro da metodologia de testes estatísticos de hipóteses. Essa teoria, descrita resumidamente no

apêndice E, permite dizer se eventuais diferenças representam meras flutuações ou se elas realmente sinalizam autênticos desvios da neutralidade. Nesse contexto, a teoria neutra é a hipótese nula, cuja avaliação requer o conhecimento das possíveis alternativas. Portanto, é preciso estudar modelos envolvendo seleção não somente para compreender possíveis mecanismos evolutivos, mas também para identificá-la em situações práticas como uma alternativa à neutralidade.

Dessa forma, este capítulo apresenta simulações computacionais de populações finitas em um relevo multiplicativo empregadas na caracterização de diversas grandezas potencialmente úteis na elaboração de testes estatísticos de neutralidade. Várias estatísticas analisadas tem caráter genealógico, ou seja, seu estudo envolve conhecimento, mesmo que apenas parcial, da história evolutiva da amostra de genes. Salvo menção em contrário, a dinâmica evolucionária foi simulada segundo o modelo de Wright-Fisher.

5.2 Construção computacional das genealogias

É inviável manter registro do estado completo de uma população durante muitas gerações, devido a limitações na capacidade de armazenamento computacional. Mesmo assim, é possível ter acesso a boa parte da história evolutiva de uma população com o conhecimento de duas matrizes, que envolvem tempos de coalescência e distâncias de Hamming entre indivíduos.

Sejam $\Gamma(\mu)$ o pai da seqüência μ e $\Gamma(\mu,\nu)$ o ancestral mais recente de μ e ν. O tempo de coalescência $T_t^{\alpha\beta}$ de dois indivíduos presentes no instante t, α e β, é o número de gerações transcorrido desde a geração em que viveu $\Gamma(\alpha,\beta)$ até o instante t. Essa definição pode ser melhor compreendida com o auxílio da figura 5.1, em que $T_t^{\alpha\beta}=1$ e $T_t^{\alpha\gamma}=T_t^{\beta\gamma}=2$. Nesse exemplo, $\Gamma(\alpha,\beta)=\Gamma(\alpha)=\Gamma(\beta)$, enquanto $\Gamma(\alpha,\gamma)=\Gamma(\beta,\gamma)=\phi$. Nota-se que ϕ desempenha o papel de ancestral mais recente das 3 seqüências no exemplo e é evidente como se define o ancestral *comum* mais recente (ACMR) em uma amostra de tamanho n. A matriz de tempos de coalescência, de dimensão $n\times n$, guarda toda a informação necessária para a determinação da árvore genealógica da população até seu

ACMR. Em particular, o maior valor de qualquer linha ou coluna de T_t indica quantas gerações separam a amostra do seu ACMR. Essa matriz é atualizada mediante a relação

$$T_t^{\alpha\beta} = T_{t-1}^{\Gamma(\alpha)\Gamma(\beta)} + 1, \tag{5.1}$$

desde que $\alpha \neq \beta$, pois, embora $T_0^{\alpha\beta} = 0$ para todos os pares $\alpha\beta$ no começo de uma simulação, a diagonal principal (e apenas ela) mantêm-se nula nas demais gerações. Portanto, só é preciso manter registro da paternidade da geração t até a matriz T_t ser completamente construída, quando então T_{t-1} é apagada. Essa matriz e sua regra de evolução foram introduzidas por Paul Higgs e Bernard Derrida em [80].

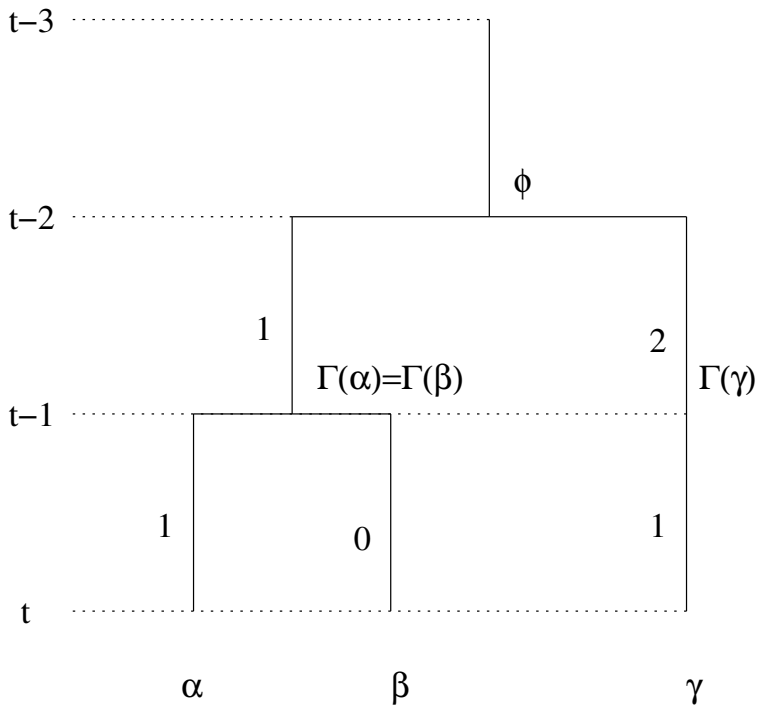

Figura 5.1: Árvore genealógica de uma amostra de 3 indivíduos, para ilustrar a definição das matrizes de tempos de coalescência e de distâncias de Hamming. Os números ao lado dos ramos da árvore representam as mutações que se adicionam às herdadas. A seqüência ϕ é o ACMR da amostra.

A matriz de tempos de coalescência é simétrica. Por outro lado, em certas aplicações, é conveniente adotar uma matriz assimétrica para descrever a variabilidade entre as seqüências. O elemento $D_t^{\alpha\beta}$ é igual a $d(\alpha, \Gamma(\alpha, \beta))$, a distância de Hamming entre α

e seu ancestral mais recente em relação a β, onde α e β são contemporâneos na geração t. Pelo modelo de infinitos sítios, todas as mutações presentes no ACMR são herdadas por seus descendentes e não podem ser revertidas, de forma que cada nova mutação que um indivíduo adquira aumenta em uma unidade sua distância de Hamming em relação a qualquer outra seqüência. Então,

$$D_t^{\alpha\beta} = D_{t-1}^{\Gamma(\alpha)\Gamma(\beta)} + d_\alpha, \tag{5.2}$$

em que d_α é o número aleatório de mutações que α adquire em adição às herdadas diretamente de seu pai. Obviamente, $d(\alpha, \beta) = D_t^{\alpha\beta} + D_t^{\beta\alpha}$. Em relação à figura 5.1, $D_t^{\alpha\beta} = 1$, $D_t^{\beta\alpha} = 0$, $D_t^{\alpha\gamma} = 2$, $D_t^{\gamma\alpha} = 3$, $D_t^{\beta\gamma} = 1$ e $D_t^{\gamma\beta} = 3$. A matriz de distâncias, assim como a de tempos de coalescência, é nula em $t = 0$ e sempre tem elementos nulos em sua diagonal principal.

É importante notar que, conceitualmente, as regras de atualização discutidas acima são válidas em qualquer dinâmica evolucionária, como os modelos de Wright-Fisher ou de Moran, embora cada caso tenha suas próprias peculiaridades no que se refere à implementação computacional. A dinâmica de Wright-Fisher foi empregada para caracterizar o tempo de coalescência e a distância de Hamming entre duas seqüências de uma população finita evoluindo em um relevo multiplicativo em [18].

5.3 Testes baseados em polimorfismo genético de um *locus*

Nesta seção, são discutidos vários testes de neutralidade que envolvem a distribuição alélica ou níveis de variabilidade amostral. Os primeiro deles baseiam-se em um dos trabalhos seminais da genética de populações, a teoria de amostragem de W. J. Ewens [52] para alelos seletivamente neutros. Ewens adotou o modelo de infinitos alelos, que atinge um estado de equilíbrio dinâmico (*steady-state*) onde continuamente novas mutações surgem na população mas alelos antigos são perdidos por deriva genética, de forma que não somente o número de alelos, mas também a distribuição de freqüências alélicas,

mantêm-se aproximadamente constante ao longo do tempo. Explicitamente, a amostra é caracterizada pela invariância (estatística) do conjunto de freqüências $\{p_i\}$, onde i é o índice do alelo com a i-ésima maior freqüência observada, embora a identidade do alelo a ocupar cada posto sempre seja temporária.

O apêndice F apresenta, entre outros resultados, a famosa fórmula de amostragem de Ewens, que expressa a distribuição de probabilidade conjunta de $\{p_i\}$. Conseqüentemente, essa fórmula também caracteriza a homozigosidade amostral $h = \sum_{i=1}^{k} p_i^2$ no regime neutro, que é simplesmente a probabilidade de dois alelos escolhidos aleatoriamente em uma certa amostra serem idênticos, dado que ela contém k alelos no total.

Como os modelos de infinitos alelos e de infinitos sítios apresentam o mesmo comportamento quanto à geração da diversidade genética, a teoria de Ewens pode ser aplicada, com pequenos ajustes, às simulações neutras desta tese. Dois alelos (seqüências) são idênticos se, e somente se, a distância de Hamming entre eles for nula. A probabilidade de mutação por indivíduo é $u = 1 - e^{-U}$, onde U tem seu sentido usual, como a média da distribuição poissoniana de mutações. Se $U \ll 1$, $u \approx U$. Entretanto, quando se pensa em inferência, é preciso lembrar que a taxa de mutação não pode ser observada diretamente, bem como o tamanho efetivo de qualquer população. Em diversos modelos de genética de populações, a dinâmica evolucionária é governada pelo parâmetro

$$\theta = 2Nu, \tag{5.3}$$

onde N é o tamanho efetivo da população e u é a probabilidade de mutação por indivíduo por geração. Para indivíduos diplóides, $N \to 2N$.

Taxas mais altas de mutação e seleção fraca geram maior variabilidade em uma população, como se vê nas figuras 5.2, que ilustra os efeitos da competição entre mutação e seleção no número de alelos, e 5.3, que descreve o comportamento da homozigosidade em função de s. Na figura 5.2, além dos resultados para $s = 0$ confirmarem as previsões do apêndice F, também é possível prever o número médio de alelos sob seleção extrema. Quando $s = 1$, qualquer amostra é composta por seqüências cujos pais são idênticos, sem mutações, caso contrário não teriam se reproduzido. Portanto, eventuais diferenças

devem-se a mutações recentes. Como cada nova mutação introduz um novo alelo, o número de alelos na amostra é simplesmente a quantidade de indivíduos que sofreu pelo menos uma mutação, mais um, correspondente àqueles que nada sofreram. O número de mutantes obedece uma distribuição binomial com probabilidade de mutação $1 - e^{-U}$ e, assim, $E(k) = 1 + n(1 - e^{-U})$ quando $s = 1$.

Em 1977, G. A. Watterson elaborou [178] uma metodologia que tornou-se conhecida como teste de Ewens-Watterson, que consiste simplesmente em comparar a homozigosidade observada e a distribuição neutra condicionada ao número observado de alelos. Se o ajuste for bom, considera-se que não há evidência significativa de atuação de seleção. Embora ele tenha se tornado bastante popular, hoje em dia é notório que o teste de Ewens-Watterson é muito conservador, ou seja, ele apresenta baixo poder na detecção de desvios da neutralidade.

Essa ineficiência é compreensível, haja vista que o modelo de infinitos alelos simplesmente rotula algum par de alelos como distintos quando assim é preciso, mas não avalia o porquê dessa diferença. Portanto, aquele modelo não é capaz de aproveitar a massiva quantidade de informação contida no polimorfismo observado empiricamente, ao contrário de sua versão aprimorada, o modelo de infinitos sítios. Naturalmente, testes baseados neste modelo revelam-se mais poderosos. Além disso, como foi discutido no capítulo 1, a teoria do coalescente tornou-se uma ferramenta imprescindível no estudo do polimorfismo genético e levou ao desenvolvimento de inúmeros testes estatísticos de neutralidade.

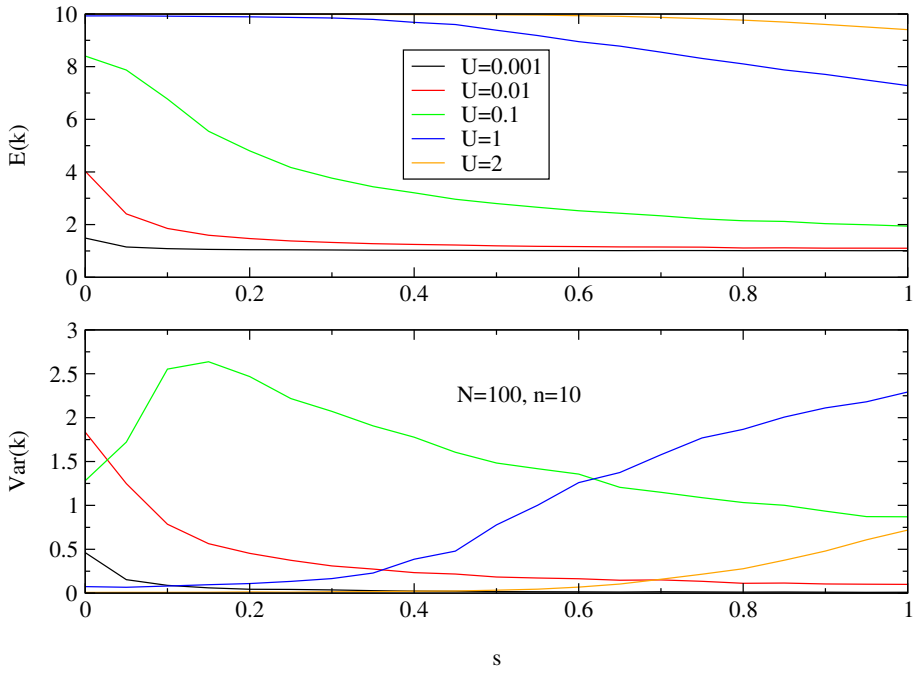

Figura 5.2: Efeitos da seleção na média e variância do número k de alelos em uma amostra de 10 indivíduos de uma população de tamanho 100.

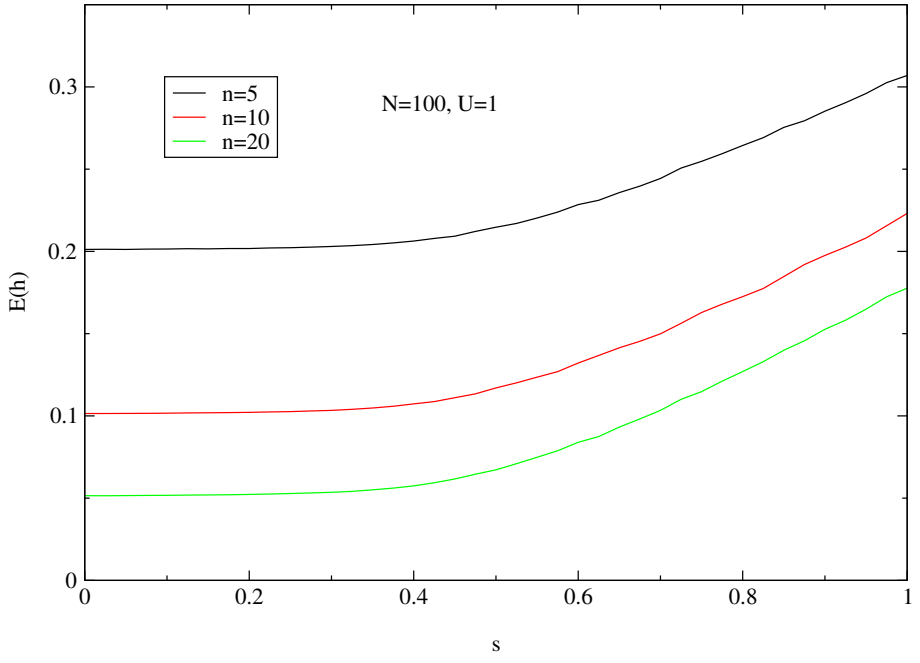

Figura 5.3: Valor médio da homozigosidade em função da seleção, para diferentes tamanhos das amostras.

Em uma amostra de seqüências de DNA, sítios que são ocupados por pelo menos dois nucleotídeos diferentes são denominados segregantes. Nas 4 seqüências moleculares apresentadas logo abaixo, há 7 sítios segregantes. Além disso, denominam-se singletos os sítios em que estão presentes exatamente dois nucleotídeos e, além disso, um deles aparece em uma única seqüência. Observam-se 6 singletos neste exemplo.

$$AC\mathbf{T}GG\mathbf{CT}AAG\mathbf{CG}CAT\mathbf{A}CT\mathbf{A}G$$

$$AC\mathbf{T}GG\mathbf{CG}AAG\mathbf{CC}CAT\mathbf{G}CT\mathbf{A}G$$

$$AC\mathbf{C}GG\mathbf{TG}AAG\mathbf{TC}CAT\mathbf{G}CT\mathbf{T}G$$

$$AC\mathbf{C}GG\mathbf{CG}AAG\mathbf{CC}CAT\mathbf{G}CT\mathbf{A}G$$

Na verdade, pelo modelo de infinitos sítios, seria impossível observar mais de dois tipos de nucleotídeo em um sítio segregante, pois jamais haveria incidência de mais de uma mutação em uma mesma posição [1] . Dessa forma, as mutações do ACMR estão presentes em todos os indivíduos da população e o número de sítios segregantes em uma amostra é igual à quantidade de mutações sofridas pelos membros da amostra e por todos os seus ancestrais, até o ACMR. Na figura 5.1, portanto, há 5 sítios segregantes.

Em meados da década de 70, Watterson estudou [177] a distribuição de probabilidade do número K de sítios segregantes em uma amostra de tamanho n de uma população sujeita a mutações neutras. Entre outras propriedades, ele obteve a média

$$E[K] = \theta a_n, \tag{5.4}$$

onde

$$a_n = \sum_{j=1}^{n-1} \frac{1}{j}. \tag{5.5}$$

De fato, espera-se que K cresça com o tamanho da amostra. Por outro lado, também considerando neutralidade, F. Tajima mostrou [169] que a distância de Hamming média (na amostra) entre duas seqüências, $\widehat{d}$, tem valor esperado

$$E[\widehat{d}] = \theta \tag{5.6}$$

[1] Em seqüências reais, há alguns raros desvios desse comportamento.

e, portanto, independe de n (embora o mesmo não ocorra com sua variância).

Mas a principal diferença entre K e $\widehat{d}$ é o fato dessas grandezas se comportarem de forma distinta quando há mudanças na composição populacional. Portadores de mutações deletérias ocorrem em baixas freqüências em uma população, porque tendem a ser eliminadas pela seleção. Mas $\widehat{d}$ é pouco afetada por esses raros mutantes, eles dão uma contribuição minoritária a essa estatística. Por outro lado, o comportamento de K é menos sensível à forma da distribuição de freqüências dos alelos na população. Não é preciso haver muitos mutantes em uma amostra para afetar fortemente o número de sítios segregantes. Isso não quer dizer que K não seja sensível a desvios da neutralidade, como pode ser observado na figura 5.4. Mas é importante ressaltar que não é possível testar a teoria neutra contando simplesmente os sítios segregantes em uma amostra porque, como a taxa de mutação não é diretamente observável, um valor baixo de K pode ser justificado por uma baixa taxa de mutação e ser considerado compatível com a neutralidade.

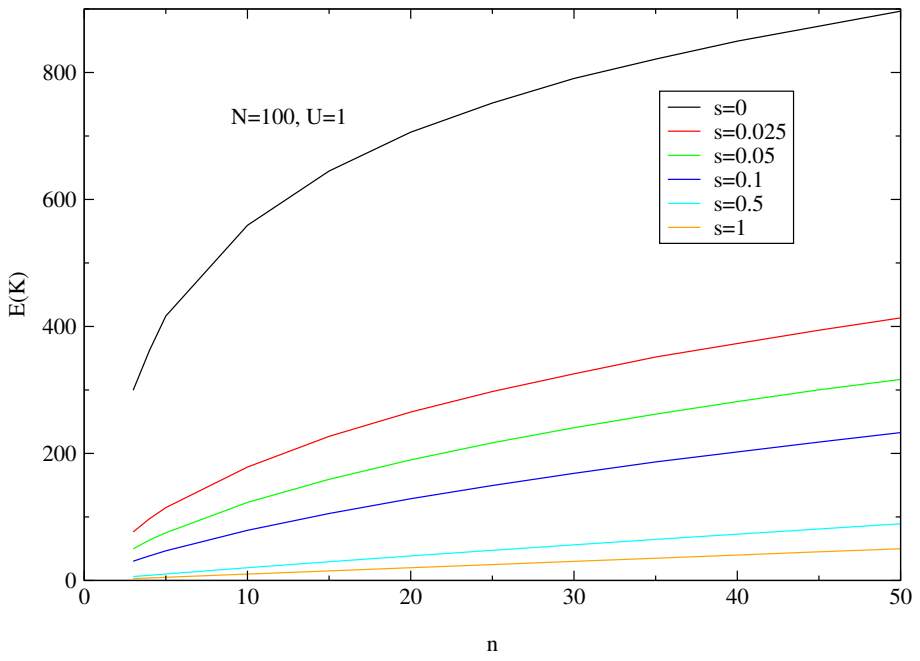

Figura 5.4: Influência do tamanho da amostra no número médio de sítios segregantes, em vários níveis de seleção em um relevo multiplicativo.

Anos mais tarde, Tajima explorou essa diferença de comportamento para criar um dos mais importantes testes de neutralidade baseados no polimorfismo de apenas um *locus* [170]. Ele adotou a estatística de teste

$$T = \frac{\widehat{d} - K/a_n}{\sigma_T}, \tag{5.7}$$

onde σ_T é o desvio padrão de $\widehat{d} - K/a_n$, cujo valor também deve ser estimado empiricamente, obedecendo uma complicada expressão apresentada no apêndice G. No regime neutro, T não é exatamente nulo, devido a um "efeito colateral" da divisão por σ_T [170]. Além disso, qualquer alteração na distribuição de freqüencias alélicas, em particular aquelas provocadas pelas diversas formas de seleção, afetam essa expectativa. É apropriado dizer, portanto, que a estatística T de Tajima mede o deslocamento do espectro de freqüências alélicas na população. Pela adoção do relevo multiplicativo e das mutações exclusivamente deletérias, a seleção é purificadora nas populações simuladas neste capítulo, afetando fortemente o espectro de freqüências e tornando T negativo. Outros mecanismos evolucionários, contudo, podem levar a valores positivos da estatística de Tajima.

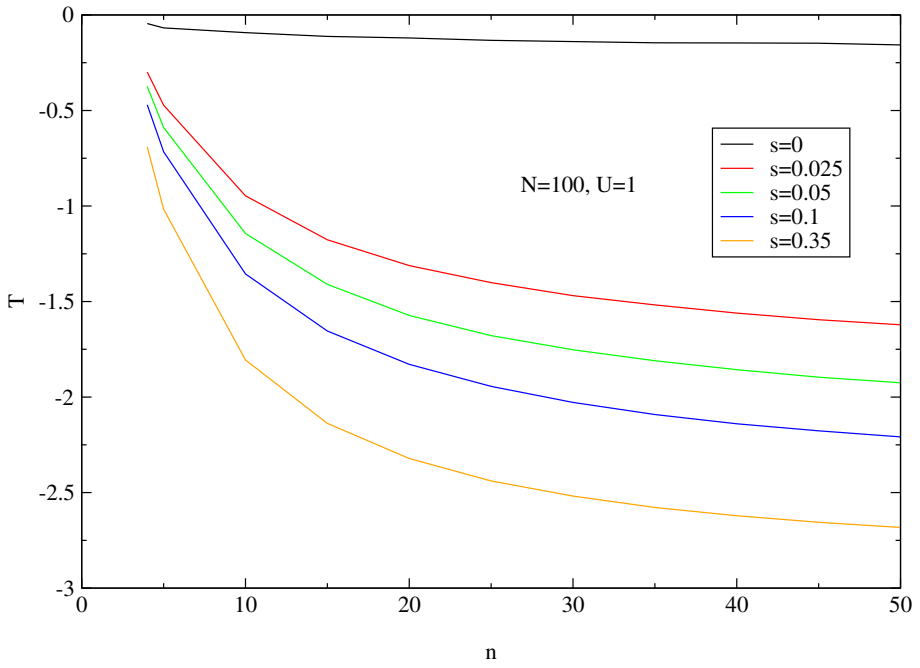

Figura 5.5: A dependência da estatística T de Tajima em relação ao tamanho da amostra.

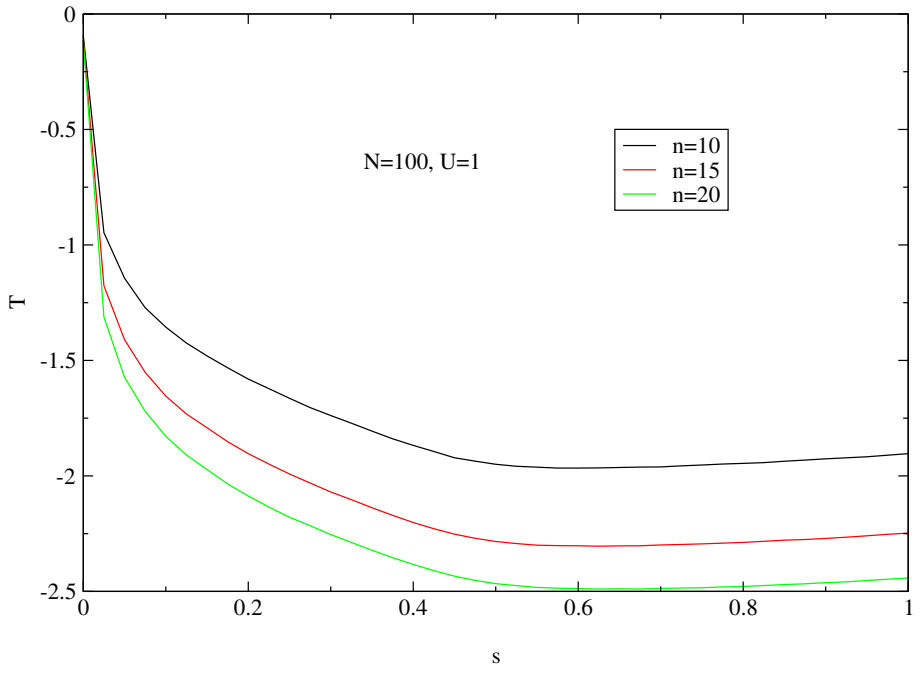

Figura 5.6: Efeito da seleção no comportamento de T.

Os efeitos do tamanho da amostra nessa grandeza estão ilustrados na figura 5.5. Já na figura 5.6, destaca-se o rápido decréscimo em T em resposta a suaves desvios da neutralidade. Afinal, essa sensibilidade é justamente o que se espera de um bom teste de neutralidade. Tajima notou que sua estatística podia ser aproximada por uma distribuição beta de probabilidade e usou esse fato para obter regiões críticas e formalizar seu teste, de forma bem convencional.

Neste capítulo, os testes considerados serão construídos numericamente. Em cada caso, o complemento do intervalo mais curto que inclui 95% da massa de probabilidade da distribuição neutra da estatística em questão é adotado como região crítica. Portanto, todos os testes adotados têm nível de significância de 5% e, via de regra, são bilaterais (tanto valores altos demais quanto baixos demais podem levar à rejeição da hipótese nula). Esse procedimento foi empregado para avaliar a eficiência do teste de Tajima mediante o cálculo da função poder, exposta na figura 5.7. Amostras maiores tornam mais fácil a detecção da seleção, embora se saiba que há uma saturação nesse comportamento.

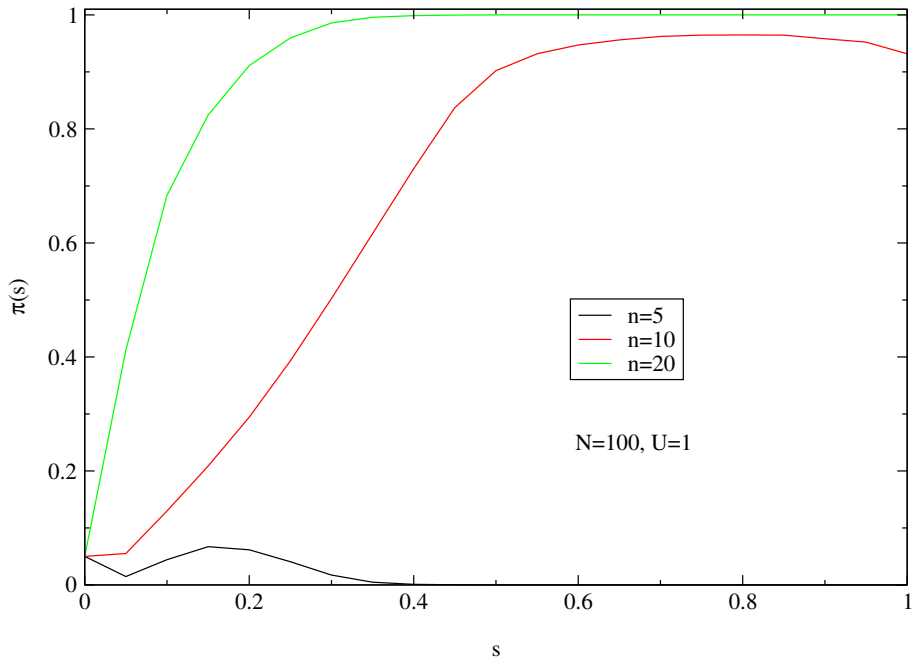

Figura 5.7: Poder do teste de Tajima em rejeitar a teoria neutra quando a população evolui em um relevo multiplicativo com parâmetro s.

Convém ressaltar que o poder de rejeição só começa a ser aceitável (melhor do que decidir no "cara ou coroa") para $s \approx 0.1$, uma condição certamente não realística. Isso não quer dizer que o teste não seja útil [2] . Quando se observa um valor alto ou baixo demais para a estatística T de Tajima e as condições de validade do modelo (ausência de fenômenos demográficos, entre outras) são respeitadas, há grande probabilidade de que a seleção natural tenha afetado a evolução populacional. O problema é que, pelo baixo poder, há uma grande ocorrência de "falsos negativos", casos em que a neutralidade deveria ser rejeitada mas a evidência empírica não é forte o suficiente para sustentar essa decisão.

Isso justifica a busca intensiva por testes de neutralidade mais poderosos que tem ocorrido nos últimos 15 anos. Y-X. Fu e W-H. Li [58] separaram os polimorfismos em recentes e ancestrais, com base em uma árvore genealógica, para tentar obter testes mais eficientes que o de Tajima. Eles argumentaram que os ramos externos, que ligam cada seqüência da amostra ao seu mais próximo ancestral em comum com algum outro membro da amostra,

[2] O artigo de Tajima [170] tem mais de 950 (!) citações até julho de 2004.

como mostra a figura 5.8, devem apresentar majoritariamente mutações recentes, embora essa mesma figura mostre um possível ramo externo que se estende por várias gerações. Por outro lado, as mutações mais antigas devem ser encontradas principalmente nos ramos internos (todos aqueles que não são externos) da genealogia.

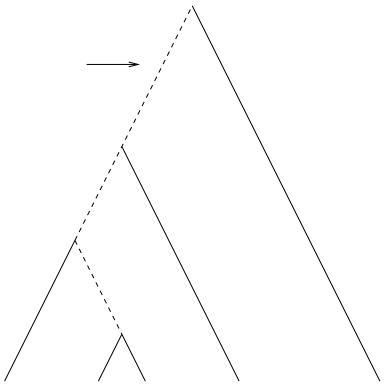

Figura 5.8: Distinção entre ramos externos (linha cheia) e internos (tracejados). A seta indica um ramo interno onde cada mutação gera um singleto, devido à existência de um ramo externo ligado diretamente ao ACMR.

Interpreta-se o comprimento de um ramo que une dois indivíduos como o número de gerações que os separa. Portanto, nas simulações aqui discutidas, o comprimento de qualquer ramo pode ser obtido a partir da matriz de tempos de coalescência T_t e, como a taxa de mutação é constante ao longo do tempo, o número médio de mutações em um ramo é uma função linear de sua extensão. Fu e Li mostraram que os valores médios da soma dos comprimentos de todos os ramos externos (CRE) e da soma dos comprimentos de todos os ramos internos (CRI) em uma genealogia neutra são

$$E(\text{CRE}) = 2N \tag{5.8}$$

e

$$E(\text{CRI}) = 2N(a_n - 1), \tag{5.9}$$

respectivamente. Surpreendentemente, $E(\text{CRE})$ é independente do tamanho da amostra. A figura 5.9, que mostra a influência do tamanho da amostra sobre os ramos externos em

vários níveis de seleção, parece contradizer essa afirmação. Entretanto, há uma justificativa para a leve tendência de crescimento observada. Os resultados de Fu e Li foram obtidos para árvores binárias, mas o modelo de Wright-Fisher, empregado na simulação, só se aproxima dessa condição quando as populações são suficientemente grandes, o que levanta dúvidas especialmente sobre os resultados obtidos para as maiores amostras, de tamanho comparável ao da população completa. Dessa forma, pelo que se observa na figura 5.9, é possível que $E(\mathrm{CRE})$ seja independente do tamanho da amostra também quando $s = 1$. Essa conjectura parece ainda mais razoável quando se considera a figura 5.10. Além disso, essa figura também ressalta uma grande sensibilidade de CRE à seleção, especialmente em amostras pequenas.

O comportamento dos ramos internos é qualitativamente semelhante ao dos ramos externos, como pode ser visto nas figuras 5.11 e 5.12. Em ambos os casos, nota-se que a seleção sempre diminui CRE e CRI em comparação com o regime neutro. Isso é uma conseqüência direta de que os membros da população cujas linhagens sobrevivem à seleção assim o fazem deixando mais descendentes a cada geração do que ocorreria sob neutralidade. Um ramo externo só pode se estender ao longo de várias gerações se os indivíduos desse ramo gerarem somente um descendente, o que é menos provável no regime seletivo do que no neutro. Portanto, CRE deve ser máximo quando $s = 0$. Além disso, as árvores genealógicas sob seleção ficam mais curtas, pois as coalescências tornam-se mais prováveis. Assim, CRI também deve diminuir quando ocorrem pequenos desvios da neutralidade. Esses argumentos justificam até mesmo o eventual crescimento nos comprimentos de ramos quando s se aproxima de 1, afinal a seleção extrema garante homogeneidade no valor adaptativo dos indivíduos que se reproduzem (embora com uma população menor, assegurando comprimentos de ramos menores do que no caso neutro).

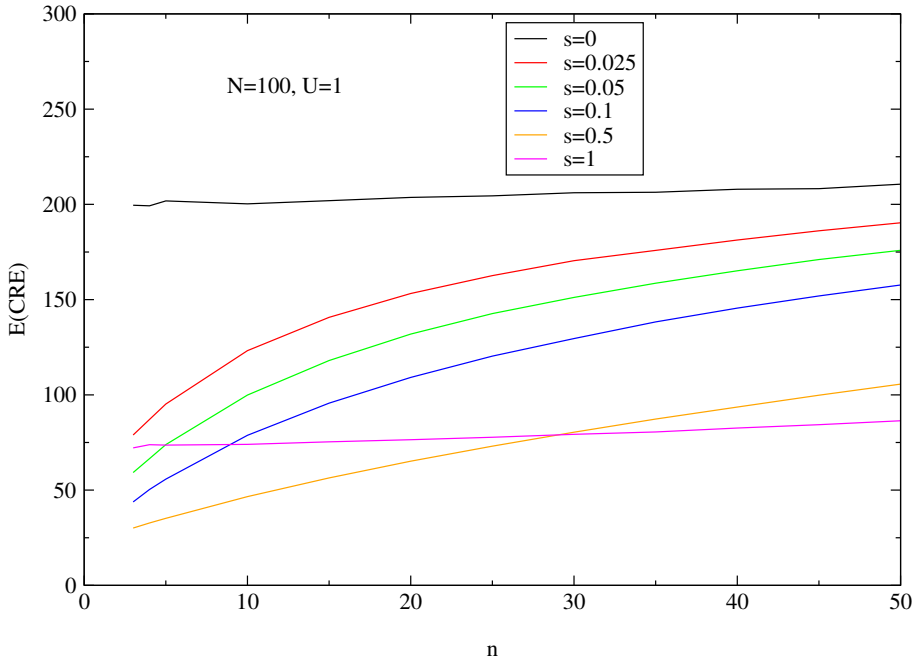

Figura 5.9: Distância de ramos externos em função do tamanho da amostra.

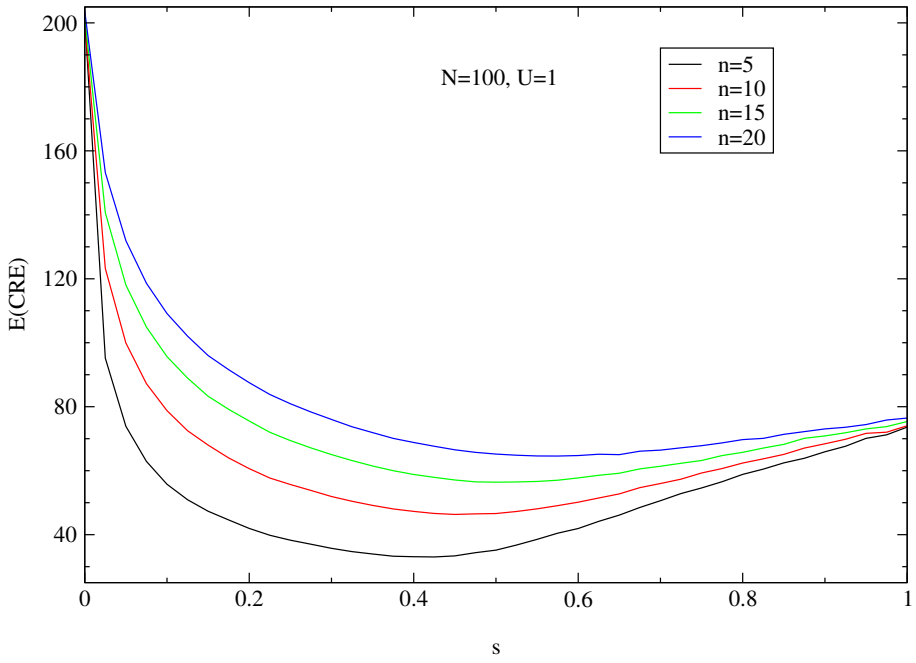

Figura 5.10: Efeitos da seleção sobre a distância de ramos externos.

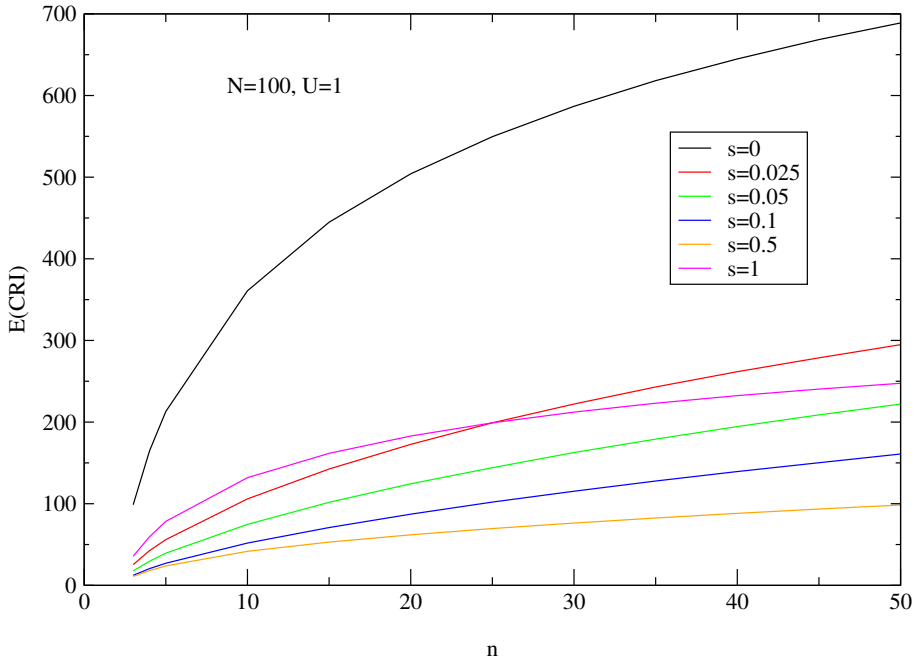

Figura 5.11: Influência do tamanho da amostra na distância de ramos internos.

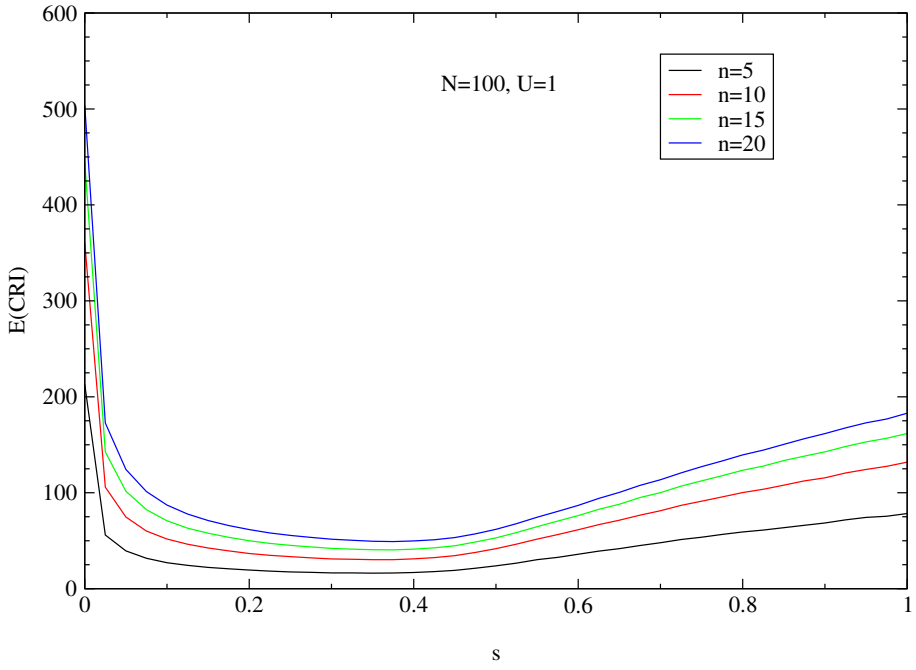

Figura 5.12: Distância de ramos internos em função de s.

Esses resultados são inéditos, afinal Fu e Li apenas estudaram analiticamente o caso neutro. Em [69], os comprimentos de ramos foram estudados nas genealogias obtidas a partir de um modelo de dois alelos (apenas dois possíveis valores adaptativos, 1 e $1-s$) com probabilidades simétricas de transformarem-se um no outro. O autor argumentou que os testes de neutralidade desenvolvidos em [58] seriam pouco poderosos porque s exerce pouquíssima influência sobre CRE e CRI naquele esquema. Mas esse comportamento é radicalmente diferente daquele apresentado nas figuras 5.10 e 5.12, e a relevância das conclusões de [69] dependem da existência de algum *locus* em que o modelo de 2 alelos possa ser considerado mais apropriado do que outros concorrentes, entre os quais a presente combinação do modelo de infinitos sítios com um relevo multiplicativo.

Assim como ocorre com o número de sítios segregantes, os comprimentos de ramos não permitem testar a hipótese neutra porque dependem da desconhecida taxa de mutação. Pior ainda, a genealogia não é evidente a partir do polimorfismo observado e precisaria ser inferida, introduzindo flutuações ainda maiores em qualquer procedimento nessa linha. Entretanto, das Eqs. (5.8) e (5.9), segue facilmente que o número de mutações presentes em ramos externos, K_e, tem média

$$E(K_e) = \theta, \tag{5.10}$$

enquanto o número de mutações em ramos internos, K_i, é tal que

$$E(K_i) = \theta(a_n - 1). \tag{5.11}$$

O número de sítios segregantes é a soma dessas duas grandezas, $K = K_e + K_i$. Além disso, Fu e Li mostraram que o valor médio do número de singletos, K_s, é

$$E(K_s) = \left(\frac{n}{n-1}\right)\theta. \tag{5.12}$$

Com base nesses resultados, eles propuseram 4 testes, baseados nas estatísticas

$$D = \frac{K - a_n K_e}{\sigma_D}, \tag{5.13}$$

$$F = \frac{\widehat{d} - K_e}{\sigma_F}, \tag{5.14}$$

$$D^* = \frac{\left(\frac{n}{n-1}\right) K - a_n K_s}{\sigma_{D^*}} \tag{5.15}$$

e

$$F^* = \frac{\widehat{d} - \left(\frac{n-1}{n}\right) K_s}{\sigma_{F^*}}. \tag{5.16}$$

Todos os desvios padrão são dados por expressões complicadas, reservadas ao apêndice G. As duas primeiras estatísticas, D e F, dependem do número de mutações nos ramos externos. Essa grandeza é tão acessível quanto qualquer outra nas simulações, mas só é observável na prática se, além da amostra, estiver disponível uma seqüência adicional cujo ancestral mais próximo em relação a qualquer membro da amostra seja mais antigo que o ACMR das n seqüências [58]. Essa seqüência denomina-se *outgroup*, em inglês, e seu uso, quando possível, potencializa a inferência. À primeira vista, é possível imaginar que K_e seja dado pelo número de singletos, que pode ser facilmente medido. Afinal, todas as mutações em ramos externos geram singletos e mutações em um ramo interno, por definição, perdem sua individualidade quando são compartilhadas por pelo menos dois indivíduos que descendem daquele ramo. Este raciocínio está correto, mas é preciso notar que estas mutações oriundas de um ramo interno podem ser transmitidas a $n-1$ indivíduos quando há um ramo externo ligado diretamente ao ACMR, como na figura 5.8. Nesse caso, um sítio singleto é segregante por não ter mutado e K_s pode superestimar K_e.

As estatísticas de Fu e Li apresentam exatamente o mesmo comportamento qualitativo da estatística T de Tajima, e seus gráficos serão omitidos. A questão relevante consiste em saber se algum dos testes considerados é significativamente mais poderoso do que os demais.

Naturalmente, os testes baseados nas estatísticas D e F sempre são mais poderosos que D^* e F^*, respectivamente. Mas todos os 4 testes de Fu e Li mostraram-se mais eficientes do que o de Tajima, pelo menos para os conjuntos de parâmetros analisados. Um caso típico, obtido em amostras de 10 seqüências, está representado na figura 5.13. O teste D é levemente mais poderoso que os demais nesse caso, mas em vários outros contextos os 4 testes são praticamente equivalentes. Também é preciso ter em mente que o poder de um teste para rejeitar uma hipótese nula sempre depende da alternativa considerada. Formas

de seleção diferentes daquela empregada neste estudo podem afetar os testes de forma imprevisível e torna-se impossível dizer de antemão qual procedimento é mais eficiente.

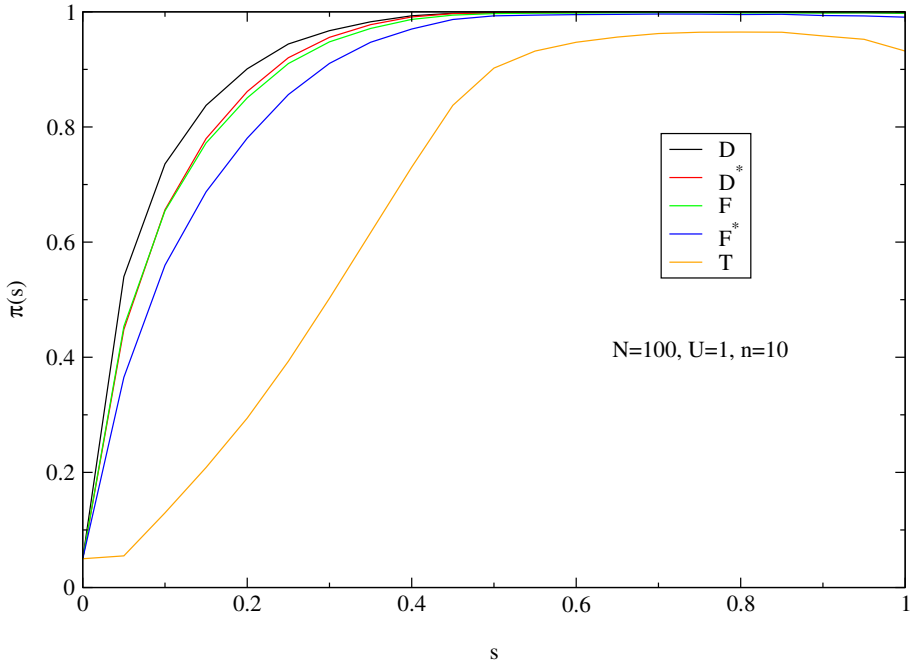

Figura 5.13: Comparação entre as funções poder das estatísticas de Tajima e Fu & Li.

5.4 Efeitos de seleção na topologia de árvores genealógicas

Também é possível construir testes de neutralidade baseados em estatísticas relacionadas à topologia da árvore genealógica de uma amostra de genes. No presente contexto, o termo topologia refere-se apenas ao padrão de ramificação de uma árvore genealógica, ignorando os comprimentos de ramos. Nesse caso, normalmente são empregadas medidas de balanço (ou simetria) das genealogias [132, 160]. Nesta seção, foram adotadas as mesmas estatísticas de balanço empregadas por Kirkpatrick e Slatkin em [109] para estudar fenômenos de especiação e extinção em filogenias e que foram posteriormente utilizadas na caracterização dos relevos energéticos de sistemas desordenados [84, 164].

5.4.1 Modelo evolucionário

Algumas das estatísticas de simetria empregadas só se aplicam a árvores binárias. No modelo de Wright-Fisher, sem superposição de gerações, é quase certo que algum indivíduo tenha 3 ou mais descendentes na geração seguinte (politomia), mesmo no regime neutro. A presença de seleção agrava ainda mais esse quadro. Porém, como só há interesse no padrão de ramificação das genealogias amostrais, os comprimentos de ramos são irrelevantes e basta adotar uma dinâmica populacional em que ocorra um evento reprodutivo a cada geração para contornar esse problema.

Dessa forma, foi adotado um modelo de Moran de tempo discreto que incorpora mutação e seleção, de modo que um indivíduo se reproduza com probabilidade proporcional ao seu valor adaptativo, dado por um relevo multiplicativo. O tamanho da população é constante, pois um dos membros antigos, incluindo o próprio indivíduo que se reproduziu, é escolhido aleatoriamente (o valor adaptativo não afeta este sorteio) para dar lugar ao recém-chegado, que carrega consigo todas as mutações de quem o originou, adicionadas de uma quantidade aleatória dada por uma distribuição de Poisson de média U.

5.4.2 Estatísticas de balanço de árvores

Dada uma amostra de tamanho n da população, seu ACMR, denotado por ϕ, é a raiz da árvore genealógica correspondente. Se i e j forem dois nós quaisquer da árvore, representando dois indivíduos que fazem parte da genealogia da amostra, não necessariamente contemporâneos, $S(i,j)$ é o número de arcos ao longo do caminho (único) que os conecta. Os indivíduos que compõem a amostra, únicos com conectividade unitária nesse grafo especial, são denominados folhas ou nós terminais da árvore que tem ϕ como raiz. Mas qualquer nó interior i (não terminal, pode ser ϕ) pode ser visto como a raiz de uma árvore, T_i, cujos nós são os descendentes de i e ele próprio. Suas folhas constituem um subconjunto da amostra original. Assim, é preciso especificar uma árvore, mediante sua raiz i, para definir a altura de uma folha particular l, $N_{i,l} = S(i,l)$, e a altura de uma árvore, $m_i = \max_{l \in T_i} N_{i,l}$. Por simplicidade, seja $N_l \equiv N_{\phi,l}$. Além disso, de cada nó i

interior da genealogia descendem duas sub-árvores, com r_i e s_i folhas, $r_i \geq s_i$.

Após essas definições, é possível apresentar adequadamente as 5 estatísticas usadas para caracterizar a topologia da árvore genealógica da amostra. Quando conhecidos, seus valores nos casos determinísticos de árvores completamente simétricas ou assimétricas e valores médios no regime neutro são apresentados para comparação. Note-se que só existem genealogias perfeitamente simétricas se o tamanho da amostra for uma potência de 2.

1. A primeira estatística é a altura média de uma folha da árvore,

$$\overline{N} = \frac{1}{n}\sum_{l=1}^{n} N_l. \tag{5.17}$$

Em uma genealogia simétrica, $\overline{N} = \log_2 n$, na assimétrica, $\overline{N} = \frac{(n-1)(n+2)}{2n}$ e Kirkpatrick e Slatkin mostraram [109], pela teoria do coalescente, que o valor esperado de $\overline{N}$ ponderado em realizações possíveis no regime neutro é $\left\langle \overline{N} \right\rangle = 2\sum_{i=2}^{n} \frac{1}{i}$.

2. A segunda é o desvio padrão da altura de uma folha,

$$\sigma_N = \frac{1}{n}\sum_{l=1}^{n}(N_l - \overline{N})^2. \tag{5.18}$$

Na árvore assimétrica, $\sigma_N = \frac{(n-1)(2n-1)}{6} - \frac{(n-1)^2(n^2+4)}{4n^2}$, enquanto na simétrica, $\sigma_N = 0$.

3.

$$C = \frac{2}{n(n-3)+2}\sum_{i=1}^{n-1}(r_i - s_i) \tag{5.19}$$

é uma medida de balanço para árvores com pelo menos 3 folhas. O índice i varre todos os $n-1$ nós internos (ϕ incluso). Desconsiderando a normalização, essa estatística é idêntica à introduzida por Colless em [27], e parecida com uma medida baseada no valor esperado da razão $\frac{r_i - s_i}{r_i + s_i}$ [60, 152].

C varia de 0, para uma árvore perfeitamente simétrica, a 1, no caso oposto. Seu valor médio, na ausência de seleção, é dado pela equação de recorrência [157]

$$\langle C \rangle_n = \frac{2}{[n(n-3)+2][n-1]}\sum_{i=1}^{n-1}\left\{[i(i-3)+2]\langle C \rangle_i + |n-2i|\right\}, \tag{5.20}$$

onde $\langle C \rangle_1 = \langle C \rangle_2 = 0$.

4. A quarta medida de balanço é o valor médio (considerando todas as sub-árvores da genealogia) do inverso da altura das sub-árvores,

$$B_1 = \sum_{i \neq \phi} \frac{1}{m_i}, \tag{5.21}$$

a soma se extendendo pelos $n-2$ nós internos distintos do ACMR da população. Em uma árvore assimétrica,

$$B_1 = \sum_{i=1}^{n-2} \frac{1}{i}, \tag{5.22}$$

enquanto

$$B_1 = \sum_{i=2}^{\log_2 n - 1} \frac{2^i}{\log_2 n - i} \tag{5.23}$$

no caso oposto.

5. A última estatística, B_2, é uma média ponderada das alturas das folhas da árvore, em constraste com a primeira medida. Os pesos foram convenientemente escolhidos de forma que B_2 seja a informação de Shannon-Wiener com probabilidade $P_l = 2^{-N_l}$ de alcançar a folha l a partir da raiz de uma árvore simétrica,

$$B_2 = -\sum_{l=1}^{n} P_l \log_2 P_l = \sum_{l=1}^{n} 2^{-N_l} N_l. \tag{5.24}$$

Assimetria faz com que

$$B_2 = 2\left(1 - \frac{1}{2^{n-1}}\right), \tag{5.25}$$

enquanto

$$B_2 = \log_2 n \tag{5.26}$$

em genealogias simétricas.

Assim, valores altos das duas últimas medidas indicam simetria, ao contrário das estatísticas restantes.

5.4.3 Resultados

Foram realizadas simulações da dinâmica de uma população de $M = 100$ indivíduos haplóides, sujeitos a uma taxa de mutações $U = 1$. O mesmo comportamento qualitativo é observado para outros valores de U e mesmo o comportamento quantitativo depende muito pouco de M quando os tamanhos das amostras são pequenos em comparação com o tamanho da população. A dependência dos valores esperados das 5 estatísticas em relação ao tamanho da amostra, baseados em 5.10^4 realizações independentes, está ilustrada nas figuras 5.14 a 5.18, para vários valores de s. As barras de erro foram omitidas porque são muito pequenas, comparáveis às espessuras das linhas dos gráficos. Os resultados analíticos para $\langle \overline{N} \rangle$ e $\langle C \rangle$ na ausência de seleção coincidiram perfeitamente com as simulações.

Embora a seleção sempre dê origem a genealogias mais assimétricas do que as obtidas no regime neutro, como esperado, nota-se que a assimetria não cresce monotonicamente com s. Para $M = 100$ e $U = 1$, observa-se um comportamento extremo em $s \approx 0.4$ e, daí em diante, os valores esperados das estatísticas aproximam-se novamente de seus comportamentos neutros. Esse efeito pode ser compreendido tendo em vista que, de certa forma, o regime de seleção extrema é semelhante ao neutro [81], embora com uma população menor, pois os indivíduos que podem se reproduzir são seletivamente idênticos. Convém ressaltar que a estatística C parece apresentar um comportamento livre de escala em função do tamanho da amostra, como se vê na figura 5.16. É difícil afirmar se essa lei de potência é característica das genealogias geradas pelo modelo evolucionário adotado, pois é possível que ela decorra de alguma propriedade puramente geométrica das árvores. Essa questão também permeia a interpretação de resultados recentes acerca da emergência de relações de escala alométricas em certas genealogias [15].

Entretanto, como as figuras 5.14 a 5.18 apresentam comportamentos médios, elas não expõem diretamente a eficiência das 5 estatísticas em rejeitar a hipótese de neutralidade em *uma* árvore genealógica gerada na presença de seleção. Para responder a essa questão no contexto de um modelo evolucionário de especiação, Kirkpatrick e Slatkin [109] usaram

cada uma das estatísticas acima para construir testes com nível de significância de 5%, exatamente da forma utilizada ao longo deste capítulo (maximização da extensão da região crítica).

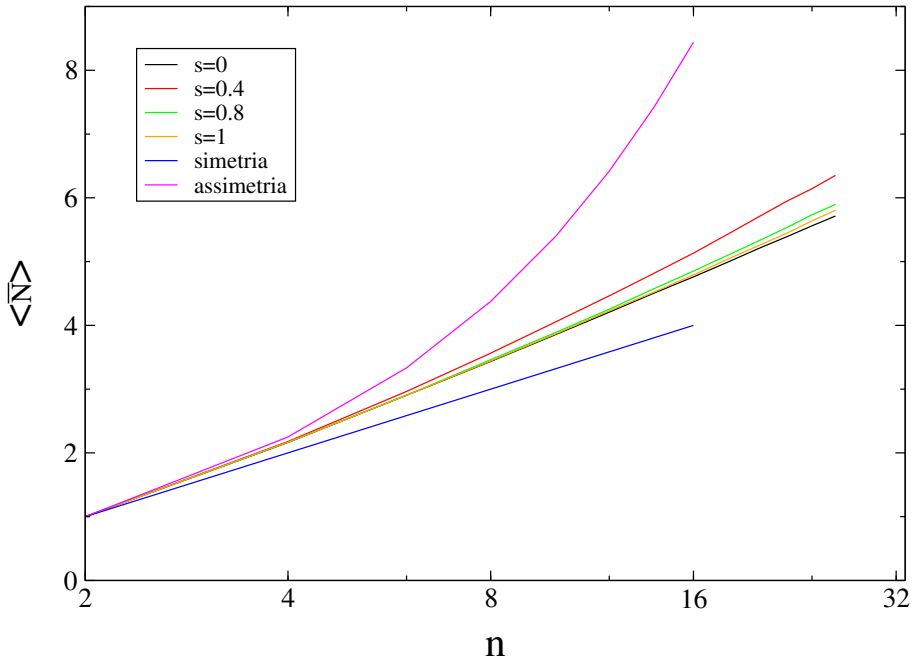

Figura 5.14: Gráfico semi-log do valor esperado da altura média de uma folha de uma árvore genealógica em função do tamanho da amostra.

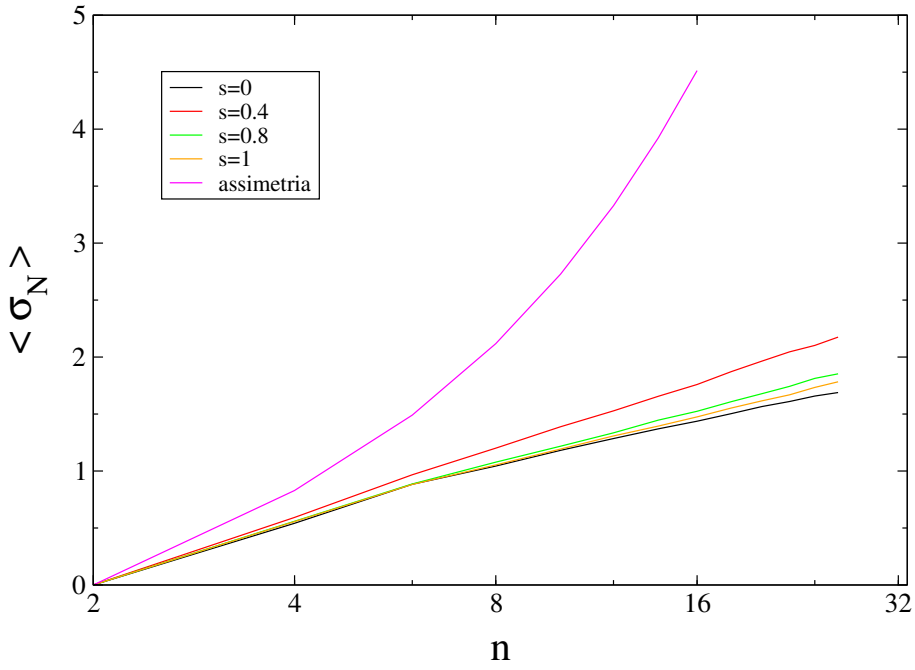

Figura 5.15: Valor médio do desvio padrão da altura de uma folha versus o número de folhas (escala logarítmica). Esta estatística é nula para árvores simétricas.

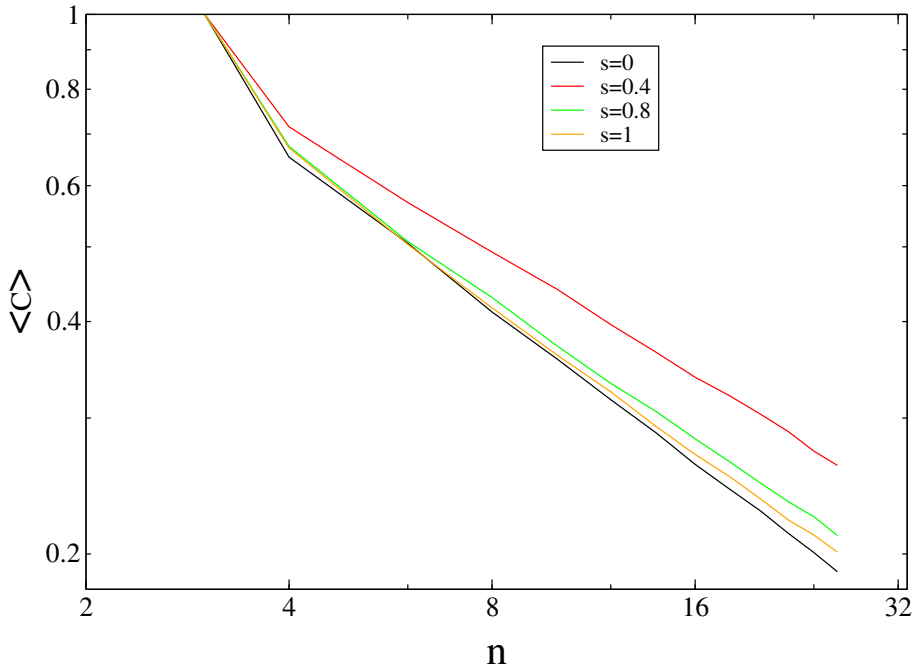

Figura 5.16: Lei de potência para a medida de balanço de Colless em função do tamanho da amostra. Em árvores simétricas, $C = 0$ e $C = 1$ no caso oposto.

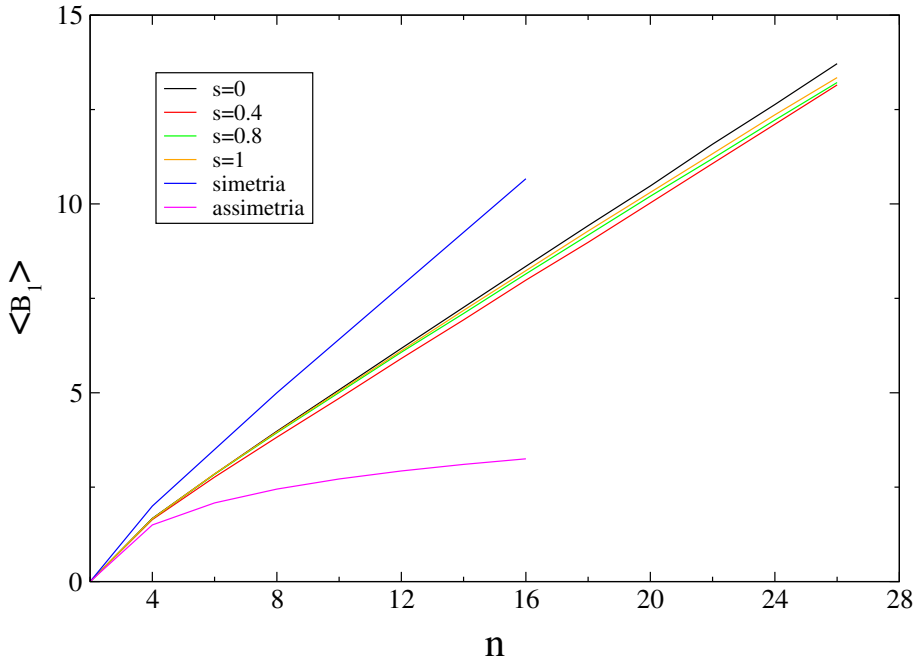

Figura 5.17: Valor esperado do inverso da altura média de uma folha de uma árvore genealógica em função do tamanho da amostra.

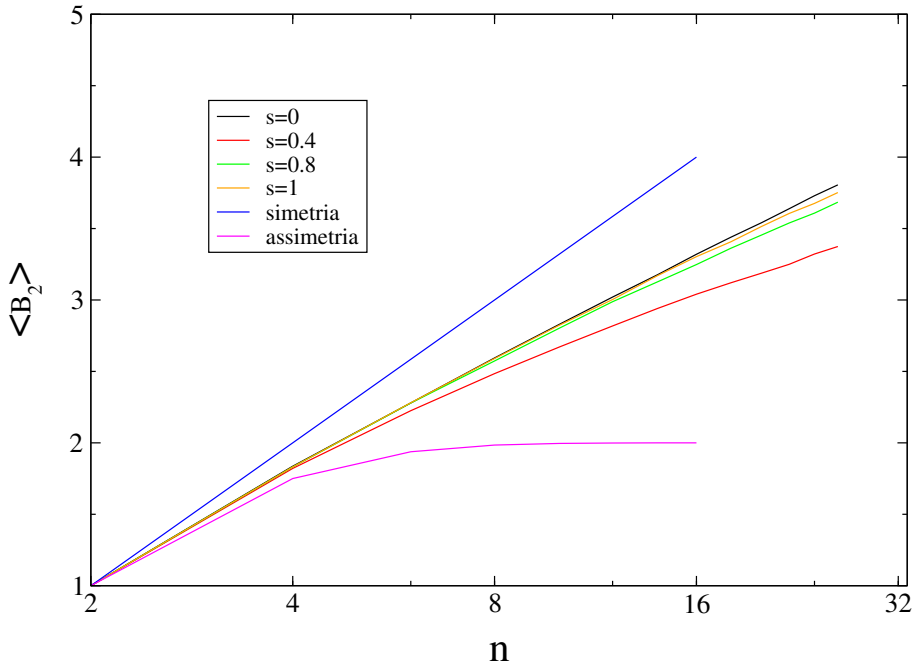

Figura 5.18: Informação de Shannon-Wiener versus tamanho da amostra. O eixo das abscissas está em escala logarítmica.

Esse procedimento foi mais uma vez utilizado para testar a hipótese neutra frente à alternativa de seleção pela dinâmica de Moran. As distribuições de probabilidade das estatísticas dos testes foram obtidas para vários valores de s via simulação de 10^4 árvores para construir as funções poder ilustradas na figura 5.19. Como se vê, o teste baseado na estatística $\overline{N}$ revela-se a melhor escolha para boa parte dos possíveis valores de s. Em seguida, destaca-se B_2, enquanto B_1 quase sempre tem o menor poder. Entretanto, nenhum dos 5 testes apresenta um desempenho aceitável. Até mesmo a maior probabilidade de rejeição da hipótese de neutralidade entre todos os casos considerados, 0.25, é muito baixa. Além disso, também é preciso considerar que esses resultados admitem o pleno conhecimento das genealogias. Em aplicações reais, é preciso inferir a forma da árvore genealógica de qualquer conjunto de seqüências, o que introduz ainda mais incertezas e diminui o poder dos testes considerados. Portanto, sem sombra de dúvida, pode-se dizer que os testes estatísticos elaborados em [109] não são úteis na detecção de desvios de neutralidade em populações evoluindo em um relevo adaptativo multiplicativo.

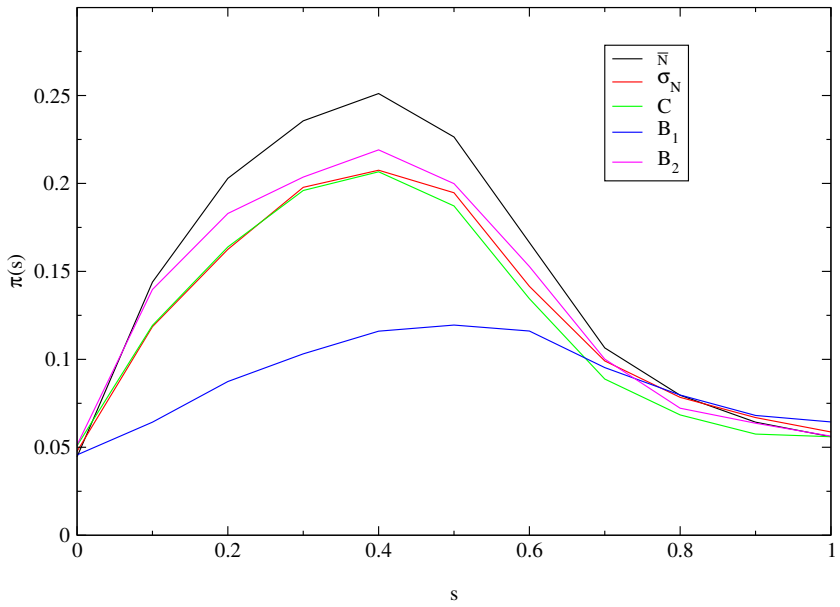

Figura 5.19: Poder dos testes de Kirkpatrick e Slatkin para amostras de tamanho 20.

5.5 Fixações

Esta seção não discute um teste de neutralidade, mas utiliza as genealogias para analisar o processo de substituição de genes. Como todas as mutações do ACMR estão presentes em seus descendentes, diz-se que tais mutações estão fixadas. Só há fixações quando há uma mudança do ACMR da população, embora a recíproca nem sempre seja verdadeira. Esse efeito pode ser observado na figura 5.20. O número de fixações quando há uma troca de ACMR é a quantidade de mutações acumulada entre os ACMR novo e antigo. Se tal número for nulo, não há fixações. Assim, tanto o intervalo de tempo entre mudanças de ACMR quanto a quantidade de fixações são variáveis aleatórias. Isso pode ser visto claramente na figura 5.21.

No caso neutro, a teoria do coalescente com tempo discreto mostra [18, 39, 81] que a distribuição de probabilidade do intervalo temporal entre trocas de ACMR pode ser aproximada, assintoticamente, por uma densidade exponencial, de expoente $\lambda = \frac{1}{N}$, sendo N o tamanho da população. Na presença de seleção, esperava-se observar o mesmo comportamento qualitativo, embora com constantes de decaimento alteradas. A figura 5.22 confirma essa expectativa.

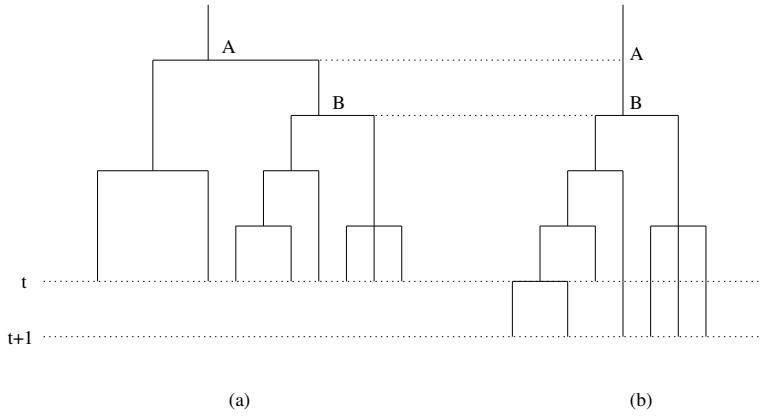

Figura 5.20: Mudança de ACMR. Os 2 indivíduos à extrema esquerda não deixam descendentes na geração $t+1$, havendo uma mudança de ACMR. (a) Árvore genealógica até o instante t. (b) Árvore genealógica estendida até $t+1$.

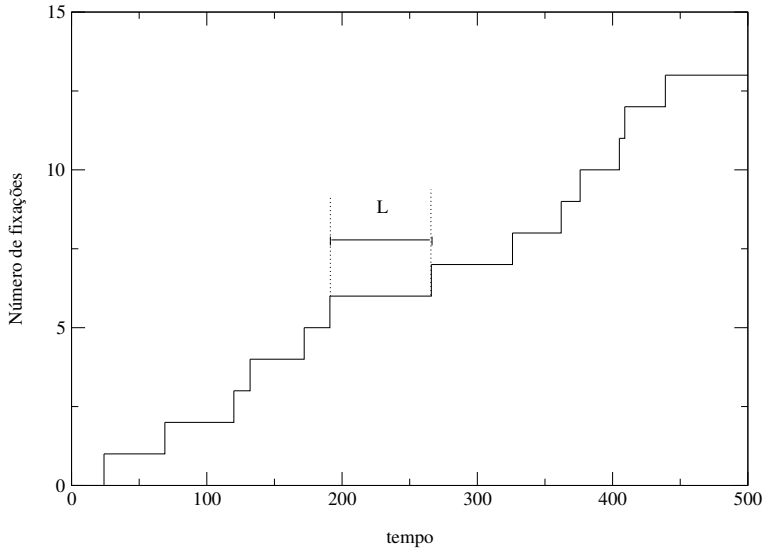

Figura 5.21: Um gráfico (quase) típico do número de fixações em função do tempo. O número de fixações não precisa variar sempre em uma unidade. Também pode haver mudanças de ACMR sem fixações, o que este gráfico, por construção, não pode evidenciar.

O comportamento observado na figura 5.22 era esperado, de certa forma. Em casos que se desviam das condições normais de aplicação do modelo de Wright-Fisher (tamanho total da população não fixado, diferenças sexuais na contribuição para a reprodução, presença de seleção), uma população aparenta um tamanho efetivo N_e menor que seu tamanho verdadeiro, conforme discussão no capítulo 2. Assim, como no caso neutro, o inverso da constante de decaimento exponencial em cada curva da figura 5.22 é uma medida de N_e na presença de seleção. Como qualquer desvio da neutralidade faz com que haja uma maior probabilidade de que vários indivíduos tenham um mesmo pai, é preciso um tempo menor para encontrar o ACMR e, conseqüentemente, a população efetiva deve ser sempre menor que a real. A ausência de monotonicidade na resposta de N_e face à seleção pode ser explicada pela neutralidade entre os indivíduos que podem se reproduzir no regime de seleção extrema ($s = 1$).

A razão entre os tamanhos efetivo e real na população simulada estão ilustrados na figura 5.23, em função de s e para vários valores da taxa de mutação. Um efeito curioso que pode ser observado nessa figura, para altas taxas de mutação, é uma lei de potência em algumas regiões relativamente extensas do domínio do parâmetro s.

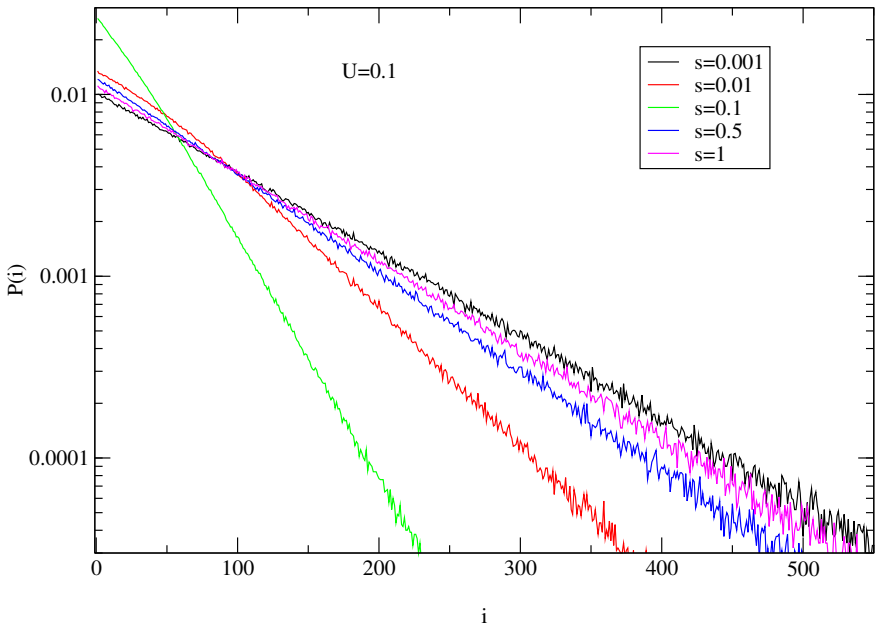

Figura 5.22: Distribuição do intervalo temporal entre mudanças de ACMR em uma população de 100 indivíduos. O comportamento exponencial ocorre apenas assintoticamente e a constante de decaimento exponencial não responde monotonicamente à seleção.

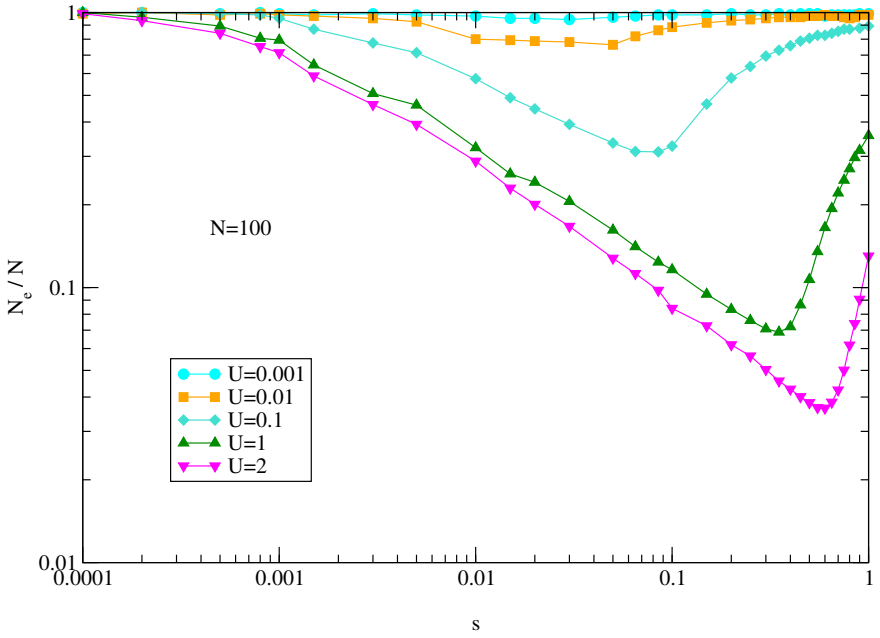

Figura 5.23: Razão entre os tamanhos de população efetivo e real, em função do coeficiente de seleção.

5.6 Conclusões

Este capítulo foi dedicado à análise de diversos testes de neutralidade baseados direta ou indiretamente na história evolutiva de uma população. Afinal de contas, embora os testes de Tajima e Fu & Li utilizem estatísticas sumárias do polimorfismo intra-específico de DNA em um *locus*, toda a teoria [170, 58] que justifica esses procedimentos está fundamentada nas propriedades de árvores genealógicas. Ainda assim, a rejeição da teoria neutra é uma tarefa reconhecidamente difícil. O teste de Tajima é menos poderoso do que os quatro propostos por Fu e Li quando se considera a rejeição de seleção purificadora mediante um relevo multiplicativo, mas mesmo a estatística D só apresenta poder respeitável em condições não realísticas. Outros trabalhos já avaliaram essas estatísticas sob outras formas de seleção [14, 161]. Os outros testes, baseados em propriedades topológicas das genealogias, mostraram-se ainda menos poderosos.

Alguns autores acreditam [140] que a razão desse fracasso é a grande dependência dos testes baseados em distribuições alélicas e/ou níveis de variabilidade em relação à demografia da população. Em aplicações reais, é muito comum haver violações de algumas hipóteses implícitas na elaboração de um modelo neutro. Se a população apresentar alguma estrutura, como separações em comunidades, ou estiver em desequilíbrio, como após extinções em larga escala, é muito difícil distinguir esses efeitos e a seleção natural, apesar de progressos recentes [54, 97].

Mas o que poderia ser visto como um defeito pode ser uma virtude, na realidade. Afinal, fenômenos demográficos também são importantes. Panoramas gerais do potencial de aplicação dos testes de neutralidade são encontrados em [6, 140, 145, 179].

Capítulo 6

Conclusões gerais

O desenvolvimento de técnicas progressivamente mais avançadas para a manipulação de microorganismos e material genético originou uma nova área de pesquisa, a evolução experimental. Entretanto, essa metodologia precisa ser acompanhada pelo avanço na modelagem matemática, necessária para organizar as informações obtidas experimentalmente e até mesmo para sugerir protocolos que se revelem particularmente interessantes pela análise teórica prévia.

Esta tese apresentou algumas contribuições nessa direção. Em alguns experimentos, não se permite que a população analisada atinja um estado de equilíbrio. Nessas situações, é fundamental conhecer a dependência temporal do sistema. Se, além disso, populações razoavelmente grandes estiverem envolvidas, o comportamento dinâmico de populações infinitas obtido no capítulo 3 pode ter alguma utilidade. Em particular, os relevos multiplicativo e de pico agudo estão entre os casos analisados e ainda são os modelos mais populares quando se deseja estudar seleção.

Por outro lado, fortes flutuações são inerentes aos processos considerados no capítulo 4. O modelo de gargalos estocásticos pode ter aplicabilidade na mensuração dos efeitos seletivos e mutacionais em sistemas reais. Na verdade, talvez até seja possível combiná-lo com o modelo de ramificação para populações em crescimento, de forma a obter um método que permita o estudo numérico de dinâmicas evolucionárias mais realísticas, que levem em consideração a probabilidade de perda dos fundadores. Convém ressaltar que

esses resultados estão relacionados a processos de acúmulo de mutações deletérias, como a catraca de Muller, de grande interesse fundamental e aplicado.

Finalmente, os resultados do capítulo 5 corroboram a notória dificuldade na detecção de eventuais marcas da seleção natural no código genético. Tanto estatísticas usualmente estudadas, baseadas no polimorfismo da população, quanto propriedades da topologia de árvores genealógicas, revelam-se incapazes de atingir esse objetivo. Mas esses testes abordam outras questões além da controvérsia neutralista-selecionista, como a identificação de efeitos demográficos no passado de uma população. De forma geral, os testes estatísticos de neutralidade podem vir a ser usados para sistematicamente identificar genes que concederam vantagens seletivas ao longo da evolução do homem moderno e assim consolidarem-se como importantes ferramentas no estudo das diferenças entre espécies e na identificação de regiões de importância médica e funcional no genoma.

Apêndice A

Distribuição de Poisson para incidência de mutações

Neste apêndice, complementar à seção 2.3, mostra-se que a quantidade de mutações adicionais originadas da replicação de uma seqüência infinitamente longa é dada por uma distribuição de Poisson.

Alguns autores, já sabendo que a condição de infinitos sítios impossibilita a ocorrência de mutações reversas, incorporam antecipadamente essa propriedade em seus modelos, mesmo quando as seqüências tem um comprimento finito. Em [78], por exemplo, a probabilidade M_{jk} de uma seqüência com k genes passar a ter j genes na geração seguinte é dada pela distribuição binomial

$$M_{jk} = \begin{pmatrix} L-k \\ j-k \end{pmatrix} u^{j-k}(1-u)^{L-j}, \qquad k \leq j \leq L. \tag{A.1}$$

A equação acima não somente é bem simples como também leva facilmente à Eq. (2.2), pois é um fato bem conhecido, exposto em qualquer texto básico sobre probabilidades, que uma lei binomial caracterizada pelos parâmetros N e p, correspondentes a $L-k$ e u na Eq. (A.1), aproxima-se de uma distribuição de Poisson de média λ quando $N \to \infty$, $p \to 0$ e $Np \to \lambda$.

Mas é ilustrativo considerar como o modelo de seqüências finitas apresentado na seção 2.3 possibilita o cálculo da probabilidade de um genótipo com k mutações transformar-se

em *algum* genótipo com k' mutações, sem restrições a k e k' e ao tamanho do alfabeto. Como não poderia deixar de ser, a distribuição de probabilidade obtida tende à distribuição de Poisson esperada quando são satisfeitas as condições de validade do modelo de infinitos sítios. Nesse limite, as mutações reversas naturalmente revelam-se impossíveis.

A idéia consiste em partir de uma seqüência com k sítios mutantes e $L-k$ não mutantes e obter um genótipo com k' defeitos. Em geral, é possível que m mutantes, $0 \leq m \leq k$, percam essa condição, o que ocorre com probabilidade $u/(A-1)$ para cada um deles. Os $k-m$ restantes podem até mutar, desde que não adquiram o símbolo original. Essa probabilidade é $1-u/(A-1)$. Há $\begin{pmatrix} k \\ m \end{pmatrix}$ formas diferentes disso acontecer. Por outro lado, se só restaram $k-m$ mutantes e é preciso obter k', $k'-(k-m)$ dos $L-k$ sítios não mutantes precisam mudar seus símbolos e os demais, $L-k-[k'-(k-m)] = L-m-k'$, devem permanecer inalterados. As probabilidades envolvidas são u e $1-u$, respectivamente, e existem $\begin{pmatrix} L-k \\ k'-(k-m) \end{pmatrix}$ combinações possíveis. Além disso, é claro que os dois números, $k'-(k-m)$ e $L-m-k'$, precisam ser maiores ou iguais a zero. Dessa forma, além do par de desigualdades $0 \leq m \leq k$, m também precisa satisfazer $k-k' \leq m \leq L-k'$. A imposição simultânea dessas duas condições determina os possíveis valores de m, de modo que a probabilidade desejada é

$$P(k \to k') = \sum_{m=\max(0,k-k')}^{\min(k,L-k')} \begin{pmatrix} k \\ m \end{pmatrix} \begin{pmatrix} L-k \\ k'-(k-m) \end{pmatrix} \times \left(\frac{u}{A-1}\right)^m \left(1-\frac{u}{A-1}\right)^{k-m} u^{m-(k-k')}(1-u)^{L-m-k'}. \tag{A.2}$$

Enquanto o caso em que $A=2$ já é conhecido na literatura [190] há algum tempo, a expressão geral parece ser inédita.

Não é difícil simplificar a equação acima quando $L \to \infty$, $u \to 0$ e $uL \to U$. Basta notar que as potências com base $1-u/(A-1)$ ou $1-u$ e expoente finito tendem a 1 quando $u \to 0$. Há uma contribuição não trival do termo $(1-u)^L = (1-U/L)^L \to e^{-U}$. Uma das potências de u contribui com a potência $L^{k-k'-m}$, que contrabalança os termos dependentes de L na segunda combinação, resultando em um fator unitário. Resta apenas

um termo dependente em L, L^{-m}. O resultado só não é nulo se m puder ser nulo, ou seja, se $k - k' \leq 0 \Rightarrow k' \geq k$. Nessas condições, só o termo $m = 0$ contribui para a soma, elimina qualquer dependência no tamanho do alfabeto e gera os fatores necessários à obtenção da distribuição de Poisson

$$P(k \to k') = e^{-U} \frac{U^{k'-k}}{(k'-k)!}, \qquad k' \geq k. \tag{A.3}$$

Apêndice B

Cálculos detalhados da evolução de uma população infinita em um relevo de seleção truncada

Neste apêndice são apresentados alguns cálculos relacionados à seção 3.4.

Resultados para a obtenção da função geratriz (3.14)

Quando os dois lados da Eq. (3.3) são multiplicados por z^i e somados de $i = 0$ a ∞, surge a expressão $A(k, z, t) \equiv \sum_{j=0}^{\infty} z^j C_j(t) w_j$, após a inversão da ordem dos somatórios no lado direito da equação resultante. Usando a forma do relevo de seleção truncada (2.4) e a definição da função geratriz (3.4),

$$A(k,z,t) = \sum_{j=0}^{k} z^j C_j(t) + (1-s)\left\{\sum_{j=0}^{\infty} z^j C_j(t) - \sum_{j=0}^{k} z^j C_j(t)\right\} = \\ = s\sum_{j=0}^{k} z^j C_j(t) + (1-s)G(z,t). \tag{B.1}$$

Essa expressão aparece na equação de recorrência (3.13) e precisa ser solúvel recursivamente para que seja possível encontrar fórmulas fechadas para o problema. Nota-se que $w(t) = A(k, 1, t)$, justificando as Eqs. (3.17) e (3.19). Definindo $B(k, z, t) \equiv \sum_{j=0}^{k} z^j C_j(t)$

e usando a equação (3.3),

$$B(k,z,t) = \sum_{i=0}^{k} z^i \left\{ \frac{e^{-U}}{w(t-1)} \sum_{j=0}^{i} C_j(t-1) w_j \frac{U^{i-j}}{(i-j)!} \right\} =$$

$$= \frac{e^{-U}}{w(t-1)} \sum_{j=0}^{k} \left\{ z^j C_j(t-1) \sum_{l=0}^{k-j} \frac{(zU)^l}{l!} \right\} \equiv \frac{e^{-U}}{w(t-1)} D(k,z,t-1). \tag{B.2}$$

Essa expressão tem dois somatórios que não podem ser simplificados, em contraste com o único somatório de $A(k,z,t)$. Se a cada iteração, regredindo no tempo, aparecesse um somatório adicional, não haveria solução fechada. Entretanto,

$$D(k,z,t-1) = \sum_{j=0}^{k} \left\{ z^j \left[\frac{e^{-U}}{w(t-2)} \sum_{i=0}^{j} C_i(t-2) w_i \frac{U^{j-i}}{(j-i)!} \right] \left[\sum_{l=0}^{k-j} \frac{(zU)^l}{l!} \right] \right\} =$$

$$= \frac{e^{-U}}{w(t-2)} \sum_{i=0}^{k} \sum_{j=i}^{k} \left\{ z^i C_i(t-2) \frac{(zU)^{j-i}}{(j-i)!} \left[\sum_{l=0}^{k-j} \frac{(zU)^l}{l!} \right] \right\} =$$

$$= \frac{e^{-U}}{w(t-2)} \sum_{i=0}^{k} \left\{ z^i C_i(t-2) \sum_{m=0}^{k-i} \left[\frac{(zU)^m}{m!} \sum_{l=0}^{k-i-m} \frac{(zU)^l}{l!} \right] \right\} =$$

$$= \frac{e^{-U}}{w(t-2)} \sum_{i=0}^{k} \left\{ z^i C_i(t-2) \sum_{j=0}^{k-i} \frac{(2zU)^j}{j!} \right\}. \tag{B.3}$$

O aspecto importante da equação acima é que foi realizada mais uma iteração e o número de somatórios se manteve constante, o que faz com que modelo seja solúvel. A última passagem, não trivial, é uma simples aplicação da identidade

$$f(\alpha,\beta,N) \equiv \sum_{n=0}^{N} \sum_{m=0}^{N-n} \left(\frac{\alpha^n}{n!} \right) \left(\frac{\beta^m}{m!} \right) = \sum_{i=0}^{N} \frac{(\alpha+\beta)^i}{i!}. \tag{B.4}$$

Sua demonstração é simples, basta iterar a equação de recorrência

$$f(\alpha,\beta,N) = f(\alpha,\beta,N-1) + \frac{(\alpha+\beta)^N}{N!}, \tag{B.5}$$

que pode ser facilmente deduzida a partir da definição de $f(\alpha,\beta,N)$.

Derivadas para o cálculo da Eq. (3.15)

É preciso calcular derivadas da função geratriz para obter as concentrações. Nesse contexto, as expressões abaixo são importantes. No que segue abaixo, a m-ésima derivada de uma função $f(z)$ é denotada por $[f(z)]^{(m)}$ e $\theta(x)$ é a função degrau,

$$\theta(x) = \begin{cases} 1, & \text{se } x \geq 0 \\ 0, & \text{se } x < 0 \end{cases} . \tag{B.6}$$

Alguns resultados preliminares úteis são

$$(z^\alpha)^{(i)} = \frac{\alpha!}{(\alpha - i)!} z^{\alpha - i}\theta(\alpha - i) \Rightarrow \left\{(z^\alpha)^{(i)}\right\}_{z=0} = i!\,\delta_{i,\alpha}, \tag{B.7}$$

$$(e^{\beta z})^{(i)} = \beta^i e^{\beta z} \Rightarrow \left\{(e^{\beta z})^{(i)}\right\}_{z=0} = \beta^i \tag{B.8}$$

e a bem conhecida expansão "binomial" para a derivada do produto de duas funções,

$$(f.g)^{(i)} = \sum_{j=0}^{i} \binom{i}{j} f^{(j)} g^{(i-j)}, \tag{B.9}$$

todos facilmente demonstráveis por indução finita.

Assim,

$$\begin{aligned}
\left\{\left[z^{j+l} e^{zU(t-i)}\right]^{(m)}\right\}_{z=0} &= \left\{\sum_{n=0}^{m} \binom{m}{n} (z^{j+l})^{(n)} [e^{zU(t-i)}]^{(m-n)}\right\}_{z=0} = \\
&= \sum_{n=0}^{m} \binom{m}{n} \left\{(z^{j+l})^{(n)}\right\}_{z=0} \left\{[e^{zU(t-i)}]^{(m-n)}\right\}_{z=0} = \\
&= \sum_{n=0}^{m} \binom{m}{n} n!\,\delta_{n,j+l} [U(t-i)]^{m-n} = \\
&= m! \frac{[U(t-i)]^{m-(j+l)}}{[m-(j+l)]!} \theta[m-(j+l)]
\end{aligned} \tag{B.10}$$

e

$$\left\{\left[e^{zUt}G(z,0)\right]^{(m)}\right\}_{z=0} = \left\{\sum_{n=0}^{m}\left(\begin{array}{c} m \\ n \end{array}\right)\left[e^{zUt}\right]^{(m-n)}\left[G(z,0)\right]^{(n)}\right\}_{z=0} =$$

$$= \sum_{n=0}^{m}\left(\begin{array}{c} m \\ n \end{array}\right)\left\{\left[e^{zUt}\right]^{(m-n)}\right\}_{z=0}\left\{\left[G(z,0)\right]^{(n)}\right\}_{z=0} =$$

$$= \sum_{n=0}^{m}\left(\begin{array}{c} m \\ n \end{array}\right)(Ut)^{m-n}n!C_n(0) = m!\sum_{n=0}^{m}C_n(0)\frac{(Ut)^{m-n}}{(m-n)!}, \qquad \text{(B.11)}$$

tendo sido usada a Eq. (3.5) no penúltimo passo.

Apêndice C

Números de Euler

Os números de Euler $E_{n,k}$ são indexados por 2 inteiros, como os coeficientes binomiais. E embora sejam bem menos famosos que seus primos binomiais, os números de Euler constituem um triângulo simétrico, como o de Pascal, e admitem uma interpretação combinatorial bem interessante: $E_{n,k}$ é o número de permutações $\pi_1\pi_2...\pi_n$ de $\{1, 2, ..., n\}$ que têm k posições onde $\pi_j < \pi_{j+1}$ ("subidas"). Assim, k pode assumir qualquer valor de 0 a $n-1$, e $\sum_{k=0}^{n-1} E_{n,k} = n!$. A simetria $E_{n,k} = E_{n,n-1-k}$ decorre da permutação $\pi_1\pi_2...\pi_n$ ter $n-1-k$ "subidas" se e somente se sua reflexão $\pi_n...\pi_2\pi_1$ tiver k "subidas". Por exemplo, se $n = 3$, há $3! = 6$ permutações possíveis, $\{1, 2, 3\}$, $\{2, 3, 1\}$, $\{3, 1, 2\}$, $\{1, 3, 2\}$, $\{2, 1, 3\}$ e $\{3, 2, 1\}$. A primeira apresenta duas "subidas" ($1 \to 2$ e $2 \to 3$), a última, nenhuma, e as demais, uma "subida".

Alternadamente diferenciando e multiplicando por x a série geométrica,

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i, \tag{C.1}$$

onde $| x | < 1$,

$$\frac{x}{(1-x)^{n+1}} \sum_{k=0}^{n-1} E_{n,k} x^k = \sum_{i=1}^{\infty} i^n x^i. \tag{C.2}$$

Esse procedimento revela que os números de Euler satisfazem a recorrência

$$E_{n,k} = (k+1)E_{n-1,k} + (n-k)E_{n-1,k-1}. \tag{C.3}$$

Os primeiros números de Euler estão apresentados na tabela C.1.

Tabela C.1: Triângulo de Euler

n	$E_{n,0}$	$E_{n,1}$	$E_{n,2}$	$E_{n,3}$	$E_{n,4}$	$E_{n,5}$	$E_{n,6}$	$E_{n,7}$	$E_{n,8}$
0	1								
1	1	0							
2	1	1	0						
3	1	4	1	0					
4	1	11	11	1	0				
5	1	26	66	26	1	0			
6	1	57	302	302	57	1	0		
7	1	120	1191	2416	1191	120	1	0	
8	1	247	4293	15619	15619	4293	247	1	0

Apêndice D

Distribuições para o tamanho do gargalo

Neste apêndice são apresentadas as formas explícitas dos coeficientes g_k utilizados nas subseções 4.2.3 e 4.2.2.

Devido à estrutura da Eq. (4.7), cada coeficiente g_k é dado pela diferença de dois termos,

$$g_k = f_k - f_{k+1}, \tag{D.1}$$

onde f_k é a média de $\alpha_{k,0}^{N'}$ na distribuição $P(N')$ considerada. Portanto, é suficiente conhecer a função f_k em cada um dos modelos probabilísticos [19] empregados.

Em termos de f_k, a equação (4.11) poderia ser expressa como

$$\langle \widehat{w}^q \rangle = e^{-qU} \left\{ \frac{1 - [1 - (1-s)^q] \sum_{i=0}^{\infty} (1-s)^{qi} f_i}{(1-s)^q} \right\}^{\tau}. \tag{D.2}$$

Tamanho fixo

Este caso corresponde a $P(N') = \delta(N, N')$, o que faz com que

$$f_k = \alpha_{k,0}^{N}. \tag{D.3}$$

Como foi possível transferir precisamente um vírus a cada gargalo no experimento de Chao [21], este protocolo também deve ser factível.

Binomial

Se uma população de tamanho M estiver dispersa uniformemente em um meio líquido de volume V, do qual é colhida uma amostra de volume ΔV, a probabilidade de um certo indivíduo estar presente na amostra é $u = \frac{\Delta V}{V}$, de forma que a probabilidade da amostra conter exatamente N' indivíduos é $P(N') = \begin{pmatrix} M \\ N' \end{pmatrix} u^{N'}(1-u)^{M-N'}$, $N' = 0, \ldots, M$. O teorema binomial leva ao resultado

$$f_k = [1 + u(\alpha_{k,0} - 1)]^M. \tag{D.4}$$

É importante notar que a variância de N', $Mu(1-u)$, é sempre menor que sua média, Mu, e pode ser arbitrariamente pequena.

Poisson

Considerando o procedimento de amostragem descrito logo acima, o interesse deste estudo reside no caso em que M é um número enorme. Se, além disso, $\Delta V \ll V$, a distribuição poissoniana torna-se uma excelente aproximação à binomial e passa a existir apenas um parâmetro relevante, o tamanho médio da amostra, $N = M\frac{\Delta V}{V}$, que também é igual à variância. Dessa forma, a distribuição de Poisson $P(N') = \frac{e^{-N}N^{N'}}{N'!}$, $N' \in \mathbb{N}$, dá origem ao coeficiente

$$f_k = \exp[N(\alpha_{k,0} - 1)]. \tag{D.5}$$

Mistura Poisson-gama

Por outro lado, o próprio volume ΔV da amostra e, conseqüentemente, o parâmetro do caso poissoniano, podem ser considerados variáveis aleatórias. Nesse caso, recomenda-se a adoção de um modelo hierárquico [19] onde a média da Poisson obedece uma distribuição gama [1] , especificada por dois parâmetros positivos, λ e r. O modelo resultante é chamado mistura Poisson-gama mas, quando $r \in \mathbb{N}$, ele se reduz a uma forma alternativa da distribuição binomial negativa. Em particular, se $r = 1$, obtém-se a distribuição geométrica.

[1] A densidade gama é $P(N) = \lambda^r N^{r-1} e^{-\lambda N}/\Gamma(r)$, com média r/λ e variância r/λ^2.

A densidade de probabilidade Poisson-gama $P(N') = \frac{\Gamma(N'+r)}{N'!\Gamma(r)} p^r q^{N'}$, onde $q = 1 - p$, depende dos parâmetros $p = \lambda/(\lambda+1)$ e r, tendo média rq/p e variância rq/p^2. Um pouco de álgebra mostra que

$$f_k = \left[\frac{p}{1 - q\,\alpha_{k,0}}\right]^r . \tag{D.6}$$

Truncamento das distribuições

Nos experimentos de transferências seriais, culturas que não se desenvolvem não são consideradas no cálculo de $\langle \widehat{w} \rangle$. Essas contribuições, em princípio, poderiam afetar positivamente a confiabilidade na determinação de s e U. No entanto, é possível mostrar que, nas regiões biologicamente plausíveis do espaço de parâmetros, o procedimento empiricamente adotado é plenamente justificado, pois reduz as incertezas na inferência. Esse fato precisa ser incorporado nos resultados acima, pois as distribuições adotadas admitem a possibilidade do tamanho da amostra ser nulo [2] .

Dessa forma é preciso truncar todas as distribuições de forma a excluir $N = 0$. Esse procedimento faz com que

$$f_k \to \frac{f_k - P(N' = 0)}{1 - P(N' = 0)} \tag{D.7}$$

e, conseqüentemente,

$$g_k \to \frac{g_k}{1 - P(N' = 0)}. \tag{D.8}$$

Portanto, dada uma distribuição para o tamanho dos gargalos, todos os β_q tornam-se maiores devido à divisão por um fator comum.

[2]Nos experimentos, a estagnação de algumas culturas deve-se ao fracasso dos fundadores e não ao fato de nenhum indivíduo ser transferido, mas essa possibilidade não foi incorporada no modelo estudado nesta tese.

Apêndice E

Conceitos básicos de testes de hipóteses

Segundo uma bem difundida corrente na epistemologia das ciências naturais [151], uma condição necessária para que uma teoria científica seja considerada científica é a falseabilidade, ou seja, ela precisa fazer previsões passíveis de refutação empírica. Dessa forma, uma metodologia consistente para a tomada de decisões na avaliação de hipóteses é um instrumento fundamental dentro do método científico. Embora ainda seja bem menos conhecida e utilizada do que deveria, a metodologia estatística de testes de hipóteses [19] já desempenha esse papel há bastante tempo.

Via de regra, denomina-se hipótese nula, H_0, a prerrogativa mais relevante entre todas as possibilidades consideradas, que deve ser posta à prova e cuja confiabilidade será conhecida em um sentido quantitativamente bem preciso. Hipóteses, nesse contexto, são afirmações acerca de um modelo probabilístico. No caso mais simples, que atende às necessidades deste trabalho, considera-se uma estatística X cuja distribuição de probabilidade depende de um parâmetro, especificado pela hipótese nula. Nesta tese, H_0 é a hipótese de evolução neutra, $H_0 : s = 0$. É conveniente especificar também a contrapartida que será considerada aceitável caso H_0 seja rejeitada, denominada hipótese alternativa, H_1. Dependendo do problema, a alternativa pode ser completamente especificada ou dada por uma hipótese composta como $H_1 : s \neq 0$.

É preciso definir um critério de decisão para a avaliação das hipóteses, embora eventuais erros sejam inevitáveis. Deseja-se evitar que H_0 seja rejeitada quando é verdadeira, o que

é conhecido como erro do tipo I, e que H_0 não seja rejeitada quando é falsa, o erro do tipo II. É usual fixar a probabilidade do erro do tipo I em um certo valor α, chamado nível de significância do teste, e escolher uma parte do espaço amostral de X, a região crítica (RC), que corresponda a uma fração α da massa de probabilidade de X segundo a distribuição especificada por H_0. Caso o valor observado da estatística em uma amostra pertença à RC, H_0 é rejeitada. Às vezes, alguns princípios são úteis na determinação da RC, mas em geral a escolha é arbitrária. Portanto, uma estatística sozinha não especifica um teste, pois é compatível com RCs arbitrárias.

A determinação da probabilidade β do erro do tipo II demanda uma especificação completa da hipótese alternativa. No presente contexto, $\beta(s) = P(\text{não rejeitar } H_0 \mid s)$. Porém, costuma-se utilizar a função poder $\pi(s) = 1 - \beta(s)$, que corresponde à fração da massa de probabilidade da distribuição de probabilidade de X, parametrizada por s, que está presente na região crítica. Assim, a função poder é realmente uma medida da capacidade de um teste em rejeitar H_0 em situações em que ela realmente deve ser rejeitada.

Apêndice F

Teoria de amostragem de Ewens para alelos seletivamente neutros

Probabilidade de uma amostra ter i alelos

Na aproximação de difusão, admite-se que o tamanho efetivo da população seja infinito, $N_e \to \infty$, mas que a probabilidade de mutação por indivíduo diplóide por geração u seja muito pequena, $U \to 0$, de modo que $4N_e u$ tende a um valor constante, θ, que rege a dinâmica evolucionária. Se os indivíduos fossem haplóides, valeria a equação $\theta = 2N_e u$.

Em [52], Ewens afirma que certos resultados da aproximação de difusão indicam que a função

$$f(x) = \theta x^{-1}(1-x)^{\theta-1} \tag{F.1}$$

é tal que $f(x)\delta x$ é a probabilidade de que um alelo tenha freqüência entre x e $x + \delta x$. Em outras palavras, $f(x)$ é uma densidade de freqüências alélicas. Esse fato permite transformar somatórios em integrais simples, pois estas têm sempre a forma de uma função beta,

$$B(r,s) = \int_0^1 x^{r-1}(1-x)^{s-1}dx = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}, \tag{F.2}$$

onde $\Gamma(n)$ é a função gama, que satisfaz $\Gamma(n+1) = n\Gamma(n)$ em geral e, se n for inteiro, $\Gamma(n+1) = n!$.

A homozigosidade F é simplesmente a probabilidade de dois alelos escolhidos aleatoriamente serem idênticos, $\sum_i p_i^2$, e, na aproximação de difusão, é dada por

$$F = \theta \int_0^1 x^2[x^{-1}(1-x)^{\theta-1}]dx = \theta\frac{\Gamma(2)\Gamma(\theta)}{\Gamma(2+\theta)} = \frac{1}{1+\theta}. \tag{F.3}$$

Além disso, a probabilidade de que o $(j+1)$-ésimo gene (freqüência igual a p_i) a ser retirado da população para compor a amostra seja de um tipo alélico diferente dos j primeiros é $\sum_i (1-p_i)^j p_i$ e vale

$$\theta \int_0^1 (1-x)^j x[x^{-1}(1-x)^{\theta-1}]dx = \theta\frac{\Gamma(1)\Gamma(j+\theta)}{\Gamma(1+j+\theta)} = \frac{\theta}{j+\theta} \equiv g_{j+1}. \tag{F.4}$$

Obviamente, a probabilidade de que tal gene não seja inédito na amostra é $h_{j+1} \equiv 1 - g_{j+1} = \frac{j}{j+\theta}$. Essas duas grandezas permitem que se obtenha π_i, a probabilidade da amostra de $2n$ genes ter i alelos. Para tanto, basta notar que, se $q_{j,i}$ é a probabilidade dos j primeiros genes da amostra pertencerem a exatamente i alelos, então é possível escrever a relação de recorrência

$$q_{j+1,i} = q_{j,i}\frac{j}{j+\theta} + q_{j,i-1}\frac{\theta}{j+\theta}, \tag{F.5}$$

que é uma aplicação da lei da probabilidade total. Os casos extremos são

$$q_{j,1} = h_2 h_3 \ldots h_j = \frac{(j-1)!}{(\theta+1)(\theta+2)\ldots(\theta+j-1)} = \frac{\theta(j-1)!}{\theta\ldots(\theta+j-1)} \tag{F.6}$$

e

$$q_{j,j} = g_2 g_3 \ldots g_j = \frac{\theta^{j-1}}{(\theta+1)(\theta+2)\ldots(\theta+j-1)} = \frac{\theta^j}{\theta\ldots(\theta+j-1)}. \tag{F.7}$$

Os demais casos são obtidos mediante a utilização da função geratriz

$$G_j(z) = \sum_{i=1}^{j} z^i q_{j,i}, \tag{F.8}$$

que, aplicada à equação (F.5), leva à equação

$$G_{j+1}(z) = \frac{j+\theta z}{j+\theta}G_j(z) = \frac{(j+\theta z)\ldots(1+\theta z)}{(j+\theta)\ldots(1+\theta)}G_1(z). \tag{F.9}$$

Como

$$G_1(z) = z q_{1,1} = z, \tag{F.10}$$

então

$$G_{j+1}(z) = \frac{(j+\theta z)\ldots(\theta z)}{(j+\theta)\ldots(\theta)}. \tag{F.11}$$

Para uma amostra de $2n$ elementos, deve-se tomar $j = 2n-1$ e, se for definido o polinômio

$$L_r(x) = x(x+1)\ldots(x+r-1) = l_1 x + l_2 x^2 + \cdots + l_r x^r, \tag{F.12}$$

então

$$G_{2n}(z) = \frac{L_{2n}(\theta z)}{L_{2n}(\theta)} \tag{F.13}$$

e uma simples comparação de potências entre as equações (F.8) e (F.11) permite concluir que

$$\pi_i = q_{2n,i} = \frac{l_i \theta^i}{L_{2n}(\theta)}. \tag{F.14}$$

Os coeficientes l_i são conhecidos como números de Stirling de primeira espécie. A função geratriz dada pela Eq. (F.13) leva facilmente ao número médio de alelos na amostra,

$$E[i] = \left\{ \frac{\partial}{\partial z} G_{2n}(z) \right\}_{z=1} = \theta \sum_{j=0}^{2n-1} \frac{1}{\theta + j}, \tag{F.15}$$

e, após alguma álgebra, à variância

$$Var[i] = \theta \sum_{j=0}^{2n-1} \frac{j}{(\theta + j)^2}. \tag{F.16}$$

Fórmula de amostragem de Ewens

A fórmula de amostragem de Ewens foi demonstrada por S. Karlin e J. McGregor em um adendo [95] ao trabalho original [52]. É notável que, após tantos anos, essa expressão ainda venha encontrando aplicações em diversas áreas, inclusive na determinação do tamanho de aglomerados [127] em Física. De fato, várias medidas de diversidade genética são análogas a grandezas físicas, especialmente na área de sistemas desordenados [79].

A fórmula de amostragem de Ewens expressa a probabilidade de cada possível partição de uma amostra de r genes em k tipos de alelos. Originalmente, as variedades alélicas

foram indexadas de forma decrescente segundo suas freqüências (índice 1 para o alelo mais freqüente, k para o mais raro) e cada partição era expressa pelo conjunto $\{n_1, \ldots, n_k\}$, onde n_i é a quantidade de genes do tipo i na amostra, $i = 1, \ldots, k$. Obviamente, $\sum_{i=1}^{k} n_i = r$. Alternativamente, uma partição pode ser denotada inequivocamente por toda uma família de variáveis como $\alpha(j)$, que representa a quantidade de alelos que contribuem com j genes na amostra. É claro que necessariamente $\alpha(j) = 0$ se $j > r$ (mas não apenas nesse caso) e sempre valem as equações $\sum_{j=1}^{r} \alpha(j) = k$ e $\sum_{j=1}^{r} j\,\alpha(j) = r$. Por exemplo, se $n_1 = 7$, $n_2 = 7$, $n_3 = 6$, $n_4 = 3$, $n_5 = 3$, $n_6 = 2$, $n_7 = 1$, $n_8 = 1$ e $n_9 = 1$, então $k = 9$, $r = 31$, $\alpha(1) = 3$, $\alpha(2) = 1$, $\alpha(3) = 2$, $\alpha(4) = 0$, $\alpha(5) = 0$, $\alpha(6) = 1$, $\alpha(7) = 2$ e $\alpha(i) = 0$ para $i \geq 8$. Com a segunda notação, a célebre fórmula adquire uma forma mais simples,

$$P\left(\{\alpha(1), \ldots, \alpha(r)\} \mid k, \theta\right) = \frac{r!}{1^{\alpha(1)}\alpha(1)! \ldots r^{\alpha(r)}\alpha(r)!} \frac{\theta^k}{L_r(\theta)}. \tag{F.17}$$

Apêndice G

Parâmetros dos testes de Tajima e Fu & Li

Este apêndice apresenta as fórmulas explícitas dos desvios padrão da estatística T de Tajima e das 4 estatísticas de Fu e Li. Sejam

$$a_{n,m} = \sum_{j=1}^{n-1} \frac{1}{j^m}, \tag{G.1}$$

$$b_1 = \frac{n+1}{3(n-1)}, \tag{G.2}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}, \tag{G.3}$$

$$c_n = \begin{cases} 1, & \text{se } n = 2 \\ 2[na_{n,1} - 2(n-1)]/[(n-1)(n-2)], & \text{se } n > 2 \end{cases} \tag{G.4}$$

e

$$d_n = c_n + \frac{n-2}{(n-1)^2} + \frac{2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1,1}-3}{n-2} - \frac{1}{n}\right). \tag{G.5}$$

Então

$$\sigma_T^2 = u_T K + v_T K(K-1), \tag{G.6}$$

onde

$$u_T = \frac{a_{n,1}\, b_1 - 1}{a_{n,1}^2} \tag{G.7}$$

e

$$v_T = \left[b_2 - \frac{n+2}{n\, a_{n,1}} + \frac{a_{n,2}}{a_{n,1}^2} \right] / \left[a_{n,1}^2 + a_{n,2} \right]. \tag{G.8}$$

Além disso, para qualquer estatística X de Fu e Li,

$$\sigma_X^2 = u_X K + v_X K^2, \tag{G.9}$$

e as constantes para cada caso são

$$v_D = 1 + \frac{a_{n,1}^2}{a_{n,1}^2 + a_{n,2}} \left(c_n - \frac{n+1}{n-1} \right) \tag{G.10}$$

e

$$u_D = a_{n,1} - 1 - v_D, \tag{G.11}$$

$$v_{D^*} = \left[\left(\frac{n}{n-1} \right)^2 b_n + a_{n,1}^2\, d_n - 2 \frac{n\, a_{n,1}(a_{n,1}+1)}{(n-1)^2} \right] / \left[a_{n,1}^2 + a_{n,2} \right] \tag{G.12}$$

e

$$u_{D^*} = \frac{n}{n-1} \left(a_{n,1} - \frac{n}{n-1} \right) - v_{D^*}, \tag{G.13}$$

$$v_F = \left[c_n + b_2 - \frac{2}{n-1} \right] / \left[a_{n,1}^2 + a_{n,2} \right] \tag{G.14}$$

e

$$u_F = \left[1 + b_1 - 4 \frac{n+1}{(n-1)^2} \left(a_{n+1,1} - \frac{2n}{n+1} \right) \right] / a_{n,1} - v_F, \tag{G.15}$$

e, finalmente,

$$v_{F^*} = \left[d_n + b_2 - \frac{2}{n-1} \left(4a_{n,2} - 6 + \frac{8}{n} \right) \right] / \left[a_{n,1}^2 + a_{n,2} \right] \tag{G.16}$$

e

$$u_{F^*} = \left[\frac{n}{n-1} + b_1 - \frac{4}{n(n-1)} + 2 \frac{n+1}{(n-1)^2} \left(a_{n+1,1} - \frac{2n}{n+1} \right) \right] / a_{n,1} - v_{F^*}. \tag{G.17}$$

Apêndice H

Glossário

Este glossário foi elaborado com base em [61]. Alguns termos omitidos aqui foram definidos detalhadamente no capítulo 2.

alelo Uma das várias possíveis formas de um mesmo gene, presumivelmente diferindo das demais por mutação da seqüência do DNA.

autossomo Cromossomo não ligado ao sexo.

codominância Manifestação fenotípica conjunta de dois alelos distintos em um heterozigoto.

códon Trinca de bases nitrogenadas que codifica um aminoácido.

diplóide Uma célula ou organismo que possui dois complementos cromossômicos.

dominância Propriedade de um alelo que produz em um heterozigoto o mesmo efeito fenotípico que produziria em um homozigoto.

especiação Processo evolucionário em que há divergência de diferentes linhas genéticas a partir de um ancestral comum.

fenótipo Uma propriedade ou conjunto de propriedades de um organismo que se manifesta pela ação dos genes e do ambiente.

freqüência gênica (ou alélica) Dada uma população de genes, é a proporção observada de um certo alelo.

freqüência genotípica Dada uma população de genótipos, é a proporção observada de um certo genótipo.

gene A unidade funcional de hereditariedade, usualmente uma extensão de DNA que difere em função (codificar uma proteína) das demais.

genótipo O conjunto de genes que um organismo individual possui. Freqüentemente, refere-se à sua composição genética em um loco específico ou conjunto de locos em questão.

haplóide Uma célula ou organismo que possui um único complemento cromossômico e, portanto, uma única cópia gênica a cada loco.

haplótipo Variante da seqüência de DNA, reconhecida por seqüenciamento direto; alelo.

heterozigoto Um organismo individual que possui alelos diferentes para um loco.

homozigoto Um organismo individual que tem o mesmo alelo em cada uma de suas cópias de um loco gênico.

loci Plural de *locus*.

loco (locus) Uma posição em um cromossomo ocupada por um gene específico.

polimorfismo Propriedade de uma população apresentar dois ou mais genótipos para uma certa característica, desde que o mais raro exceda alguma pequena freqüência pré-fixada.

Bibliografia

[1] Can physics deliver another biological revolution? *Nature*, 397:89, 1999.

[2] C. Adami. *Introduction to Artificial Life*. Springer, New York, 1st edition, 1998.

[3] D. Alves and J. F. Fontanari. Population genetics approach to the quasispecies model. *Phys. Rev. E*, 54:4048–4053, 1996.

[4] D. Alves and J. F. Fontanari. Error threshold in finite populations. *Phys. Rev. E*, 57:7008–7013, 1998.

[5] D. I. Andersson and D. Hughes. Muller's ratchet decreases fitness of a DNA-based microbe. *Proc. Natl. Acad. Sci. USA*, 93:906–907, 1996.

[6] P. Andolfatto. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Devel.*, 11:635–641, 2001.

[7] E. Baake and W. Gabriel. Biological evolution through mutation, selection, and drift: an introductory review. In D. Stauffer, editor, *Ann. Rev. Comp. Phys.*, volume 9, pages 203–264. World Scientific, Singapore, 1999.

[8] G. Bell. *Sex and Death in Protozoa: The History of an Obsession*. Cambridge Univ. Press, Cambridge, U.K., 1st edition, 1988.

[9] C. T. Bergstrom, P. McElhany, and L. A. Real. Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proc. Natl. Acad. Sci. USA*, 96:5095–5100, 1999.

[10] A. T. Bernardes. Monte Carlo simulations of biological ageing. In *Ann. Rev. Comp. Phys.*, volume 4, pages 359–395. World Scientific, Singapore, 1996.

[11] A. T. Bernardes. Mutation load and the extinction of large populations. *Physica A*, 230:156–173, 1996.

[12] A. T. Bernardes. Strategies for reproduction and ageing. *Ann. Physik (Leipzig)*, 5:539–550, 1996.

[13] E. Bornberg-Bauer and H. S. Chan. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA*, 96:10689–10694, 1999.

[14] J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140:783–796, 1995.

[15] P. R. A. Campos, V. M. de Oliveira, and L. P. Maia. Emergence of allometric scaling in genealogical trees. *Adv. Complex Syst.*, 7:39–46, 2004.

[16] P. R. A. Campos and J. F. Fontanari. Finite-size scaling of the quasispecies model. *Phys. Rev. E*, 58:2664–2667, 1998.

[17] P. R. A. Campos and J. F. Fontanari. Finite-size scaling of the error threshold transition in finite populations. *J. Phys. A: Math. Gen.*, 32:L1–L7, 1999.

[18] P. R. A. Campos, M. T. Sonoda, and J. F. Fontanari. On the structure of genealogical trees in the presence of selection. *Physica A*, 283:11–16, 2000.

[19] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, Belmont, 1st edition, 1990.

[20] E. Çinlar. *Introduction to Stochastic Processes*. Prentice Hall, Englewood Cliffs, 1st edition, 1975.

[21] L. Chao. Fitness of RNA virus decreased by Muller's ratchet. *Nature*, 348:454–455, 1990.

[22] B. Charlesworth. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res. Camb.*, 55:199–221, 1990.

[23] B. Charlesworth, P. Sniegowski, and W. Stephan. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371:215–220, 1994.

[24] D. Charlesworth and B. Charlesworth. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genet. Res. Camb.*, 70:63–73, 1997.

[25] D. K. Clarke, E. A. Duarte, A. Moya, S. F. Elena, E. Domingo, and J. Holland. Genetic bottlenecks and population passages cause profound fitness diferences in RNA viruses. *J. Virol.*, 67:222–228, 1993.

[26] A. Colato and J. F. Fontanari. Soluble model for the accumulation of mutations in asexual populations. *Phys. Rev. Lett.*, 87:238102, 2001.

[27] D. H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool.*, 31:100–104, 1982.

[28] H. J. Cooke. Y chromosome and male infertility. *Rev. Reprod.*, 4:5–10, 1999.

[29] H. Cronin. *A Formiga e o Pavão*. Papirus, Campinas, 1a. edição, 1995.

[30] J. F. Crow. Population genetics history: a personal view. *Ann. Rev. Genet.*, 21:1–22, 1987.

[31] J. F. Crow and M. Kimura. *An Introduction to Population Genetic Theory*. Harper & Row, New York, 1st edition, 1970.

[32] R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1st edition, 1976.

[33] R. Dawkins. *The Extended Phenotype*. Oxford University Press, Oxford, 1st edition, 1982.

[34] R. Dawkins. *O Relojoeiro Cego*. Companhia das Letras, São Paulo, 1a. edição, 2001.

[35] K. J. Dawson. The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. *Theor. Pop. Biol.*, 55:1–22, 1999.

[36] M. de La Peña, S. F. Elena, and A. Moya. Effect of deleterious mutation-accumulation on the fitness of RNA bacteriophage MS2. *Evolution*, 54:686–691, 2000.

[37] J. A. G. M. de Visser. The fate of microbial mutators. *Microbiol.*, 148:1247–1252, 2002.

[38] D. C. Dennett. *A Perigosa Idéia de Darwin*. Rocco, Rio de Janeiro, 1a. edição, 1998.

[39] B. Derrida and L. Peliti. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, 53:355–382, 1991.

[40] J. Diamond. *Armas, Germes e Aço*. Record, Rio de Janeiro, 3a. edição, 2002.

[41] E. Domingo, C. Escarmis, N. Sevilla, A. Moya, S. F. Elena, J. Quer, I. S. Novella, and J. J. Holland. Basic concepts in RNA virus evolution. *FASEB J.*, 10:859–864, 1996.

[42] E. Domingo and J. J. Holland. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.*, 51:151–178, 1997.

[43] P. Donnelly and S. Tavaré. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.*, 29:401–421, 1995.

[44] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. Rates of spontaneous mutation. *Genetics*, 148:1667–1686, 1998.

[45] B. Drossel. Biological evolution and statistical physics. *Adv. Phys.*, 50:209–295, 2001.

[46] E. Duarte, D. Clarke, A. Moya, E. Domingo, and J. Holland. Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. *Proc. Natl. Acad. Sci. USA*, 89:6015–6019, 1992.

[47] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwiss.*, 58:465–523, 1971.

[48] M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *J. Phys. Chem.*, 92:6881–6891, 1988.

[49] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv. Chem. Phys.*, 75:149–263, 1989.

[50] M. Eigen and P. Schuster. *The Hypercycle: a principle of natural self-organization.*

[51] S. F. Elena and R. E. Lenski. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Rev. Genet.*, 4:457–469, 2003.

[52] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87–112, 1972.

[53] W. J. Ewens. *Mathematical Population Genetics.* Springer-Verlag, New York, 1st edition, 1979.

[54] J. C. Fay and C-I. Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155:1405–1413, 2000.

[55] M. Feldman, S. Otto, and F. Christiansen. Population genetics perspectives on the evolution of recombination. *Annu. Rev. Genet.*, 30:261–295, 1997.

[56] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78:737–756, 1974.

[57] J. F. Fontanari, A. Colato, and R. S. Howard. Mutation accumulation in growing asexual lineages. *Phys. Rev. Lett.*, 91:218101, 2003.

[58] Y. X. Fu and W. H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133:693–709, 1993.

[59] Y. X. Fu and W. H. Li. Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Pop. Biol.*, 56:1–10, 1999.

[60] G. Fusco and Q. C. B. Cronk. A new method for evaluating the shape of large phylogenies. *J. Theor. Biol.*, 175:235–243, 1995.

[61] D. Futuyma. *Evolutionary Biology*. Sinauer, Sunderland, 1st edition, 1986.

[62] D. Futuyma. *Biologia Evolutiva*. Sociedade Brasileira de Genética / CNPq, Ribeirão Preto, 2a. edição, 1997.

[63] W. Gabriel, M. Lynch, and R. Burger. Muller's ratchet and mutational meltdowns. *Evolution*, 47:1744–1757, 1993.

[64] H. Gardner. *A Nova Ciência da Mente*. Edusp, São Paulo, 1a. edição, 1995.

[65] L. Garrett. *A Próxima Peste*. Nova Fronteira, Rio de Janeiro, 1a. edição, 1995.

[66] S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *TREE*, 12:307–312, 1997.

[67] J. H. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, New York, 1st edition, 1991.

[68] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, 1st edition, 1989.

[69] G. B. Golding. The effect of purifying selection on genealogies. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 271–285. Springer, New York, 1997.

[70] I. Gordo and B. Charlesworth. The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics*, 154:1379–1387, 2000.

[71] I. Gordo and B. Charlesworth. On the speed of Muller's ratchet. *Genetics*, 156:2137–2140, 2000.

[72] I. Gordo and B. Charlesworth. The speed of Muller's ratchet with background selection, and the degeneration of y chromossomes. *Genet. Res. Camb.*, 78:149–161, 2001.

[73] R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Stat. Sci.*, 9:307–319, 1994.

[74] J. Haigh. The accumulation of deleterious genes in a population. *Theor. Pop. Biol.*, 14:251–267, 1978.

[75] J. B. S. Haldane. The effect of variation on fitness. *Am. Nat.*, 71:337–349, 1937.

[76] G. H. Hardy. Mendelian proportions in a mixed population. *Science*, 28:49–50, 1908.

[77] D. L. Hartl and A. G. Clark. *Principles of Population Genetics*. Sinauer, Sunderland, 3rd edition, 1997.

[78] P. G. Higgs. Error thresholds and stationary mutant distributions in multi-locus diploid genetics models. *Genet. Res. Camb.*, 63:63–78, 1994.

[79] P. G. Higgs. Frequency distributions in population genetics parallel those in statistical physics. *Phys. Rev. E*, 51:95–101, 1995.

[80] P. G. Higgs and B. Derrida. Genetic distance and species formation in evolving populations. *J. Mol. Evol.*, 35:454–465, 1992.

[81] P. G. Higgs and G. Woodcock. The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection. *J. Math. Biol.*, 33:677–702, 1995.

[82] J. Hofbauer and K. Sigmund. *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, Newcastle, 1st edition, 1991.

[83] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Newcastle, 1st edition, 1998.

[84] W. Hordijk, J. F. Fontanari, and P. F. Stadler. Shapes of tree representations of spin-glass landscapes. *J. Phys. A: Math. Gen.*, 36:3671–3681, 2003.

[85] R. S. Howard and C. M. Lively. Parasitism, mutation accumulation and the maintenance of sex. *Nature*, 367:554–557, 1994.

[86] R. S. Howard and C. M. Lively. The Ratchet and the Red Queen: the maintenance of sex in parasites. *J. Evol. Biol.*, 15:648–656, 2002.

[87] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.*, 23:183–201, 1983.

[88] R. R. Hudson. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, editors, *Oxford Surveys in Evolutionary Biology*, pages 1–44. Oxford University Press, Oxford, UK, 1990.

[89] R. R. Hudson and N. L. Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 120:831–840, 1988.

[90] T. Johnson. The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. R. Soc. Lond. B*, 266:2389–2397, 1999.

[91] T. Johnson. *Theoretical Studies of the Interaction Between Deleterious and Beneficial Mutations*. PhD thesis, University of Edinburgh, Edinburgh, 2000.

[92] F. C. Kafatos and T. Eisner. Unification in the century of Biology. *Science*, 303:1257–1257, 2004.

[93] N. L. Kaplan, T. Darden, and R. R. Hudson. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.

[94] S. Karlin. R. A. Fisher and evolutionary theory. *Stat. Sci.*, 7:13–33, 1992.

[95] S. Karlin and J. McGregor. Addendum to a paper of W. Ewens. *Theor. Pop. Biol.*, 3:113–116, 1972.

[96] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, 1st edition, 1975.

[97] Y. Kim and W. Stephan. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, 155:1115–1427, 2000.

[98] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[99] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903, 1969.

[100] M. Kimura. Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.*, 2:174–208, 1971.

[101] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1st edition, 1983.

[102] M. Kimura and J. F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738, 1964.

[103] M. Kimura and T. Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54:1337–1351, 1966.

[104] J. L. King and T. H. Jukes. Non-Darwinian evolution: Random fixation of selectively neutral mutations. *Science*, 164:788–798, 1969.

[105] J. F. C. Kingman. The coalescent. *Stoch. Proc. Appl.*, 13:235–248, 1982.

[106] J. F. C. Kingman. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland, Amsterdam, 1982.

[107] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob. A*, 19:27–43, 1982.

[108] J. F. C. Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156:1461–1463, 2000.

[109] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47:1171–1181, 1993.

[110] L. L. Knowles. The burgeoning field of statistical phylogeography. *J. Evol. Biol.*, 17:1–10, 2004.

[111] A. S. Kondrashov. Muller's ratchet under epistatic selection. *Genetics*, 136:1469–1473, 1994.

[112] S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theor. Popul. Biol.*, 51:210–237, 1997.

[113] M. K. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140:1421–1430, 1995.

[114] E. Lázaro, C. Escarmís, E. Domingo, and S. C. Manrubia. Modeling viral genome fitness evolution associated with serial bottleneck events: evidence of stationary states of fitness. *J. Virol.*, 76:8675–8681, 2002.

[115] R. E. Lenski, C. Ofria, R. T. Pennock, and C. Adami. The evolutionary origin of complex features. *Nature*, 423:139–144, 2003.

[116] W.-H. Li. Kimura's contributions to molecular evolution. *Theor. Pop. Biol.*, 49:146–153, 1996.

[117] W.-H. Li. *Molecular Evolution*. Sinauer, Sunderland, 1st edition, 1997.

[118] M. Lynch, R. Burger, D. Butcher, and W. Gabriel. The mutational meltdown in asexual populations. *J. Heredity*, 84:339–344, 1993.

[119] M. Lynch and W. Gabriel. Mutation load and the survival of small populations. *Evolution*, 44:1725–1737, 1990.

[120] L. P. Maia. The dynamical way to mutation-selection balance of an infinite population evolving on a truncated fitness landscape. *submetido ao J. Math. Biol.*, 2004.

[121] L. P. Maia, D. F. Botelho, and J. F. Fontanari. Analytical solution of the evolution dynamics on a multiplicative-fitness landscape. *J. Math. Biol.*, 47:453–456, 2003.

[122] L. P. Maia, A. Colato, and J. F. Fontanari. Effect of selection on the topology of genealogical trees. *J. Theor. Biol.*, 226:315–320, 2004.

[123] S. C. Manrubia, E. Lázaro, J. Pérez-Mercader, C. Escarmís, and E. Domingo. Fitness distributions in exponentially growing asexual populations. *Phys. Rev. Lett.*, 90:188102, 2003.

[124] Mathpages. Euler numbers. http://www.mathpages.com/home/kmath464.htm.

[125] J. Maynard Smith. The theory of games and the evolution of animal conflicts. *J. Theor. Biol.*, 47:209–221, 1974.

[126] J. Maynard Smith. *Evolutionary Genetics*. Oxford University Press, Oxford, 1st edition, 1989.

[127] A. Z. Mekjian. Cluster distributions in physics and genetic diversity. *Phys. Rev. A*, 44:8361–8374, 1991.

[128] A. L. Melzer and J. H. Koeslag. Mutations do not accumulate in asexual isolates capable of growth and extinction - Muller's ratchet re-examined. *Evolution*, 45:649–655, 1991.

[129] G. F. Miller. *A Mente Seletiva*. Editora Campus, Rio de Janeiro, 1a. edição, 2001.

[130] M. Mohle. Ancestral processes in population genetics - the coalescent. *J. Theor. Biol.*, 204:629–638, 2000.

[131] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, New York, 3rd edition, 1974.

[132] A. O. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *Quart. Rev. Biol.*, 72:31–54, 1997.

[133] P. A. P. Moran. Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54:60–71, 1958.

[134] S. Moss de Oliveira. A small review of the Penna model for biological ageing. *Physica A*, 257:465–469, 1998.

[135] S. Moss de Oliveira, D. Alves, and J. S. Sá Martins. Evolution and ageing. *Physica A*, 285:77–100, 2000.

[136] S. Moss de Oliveira, P. M. C. de Oliveira, and D. Stauffer, editors. *Evolution, Money, War and Computers*. Teubner, Stuttgart-Leipzig, 1999.

[137] H. J. Muller. Our load of mutations. *Am. J. Hum. Genet.*, 2:111–176, 1950.

[138] H. J. Muller. The relation of recombination to mutational advance. *Mutat. Res.*, 1:2–9, 1964.

[139] C. Neuhauser and S. M. Krone. The genealogy of samples in models with selection. *Genetics*, 145:519–534, 1997.

[140] R. Nielsen. Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86:641–647, 2001.

[141] M. Nordborg. Coalescent theory. In D. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, volume 1, pages 179–212. Wiley, Chichester, UK, 2001.

[142] M. Nowak and P. Schuster. Error thresholds of replication in finite populations: mutation frequencies and the onset of Muller's ratchet. *J. Theor. Biol.*, 137:375–395, 1989.

[143] M. A. Nowak and R. M. May. *Virus Dynamics*. Oxford University Press, New York, 1st edition, 2000.

[144] D. I. Nurminsky. Genes in sweeping competition. *Cell. Mol. Life Sci.*, 58:125–134, 2001.

[145] S. P. Otto. Detecting the form of selection from DNA sequence data. *Trends Genet.*, 16:526–529, 2000.

[146] L. Peliti. Fitness landscapes and evolution. In T. Riste and D. Sherrington, editors, *Physics of materials: Fluctuations, selfassembly and evolution*, page 267. Kluwer, Dordrecht, 1996.

[147] L. Peliti. Introduction to the statistical theory of darwinian evolution, December 1997.

[148] T. J. P. Penna. A bit-string model for biological aging. *J. Stat. Phys.*, 78:1629–1633, 1995.

[149] T. J. P. Penna, A. Racco, and M. Argollo de Menezes. Getting older with computers. *Comp. Phys. Comm.*, 121-122:108–112, 1999.

[150] S. Pinker. *Como a Mente Funciona*. Companhia das Letras, São Paulo, 1a. edição, 1998.

[151] K. R. Popper. *A Lógica da Pesquisa Científica*. Cultrix, São Paulo, 1a. edição, 1974.

[152] A. Purvis, A. Katzourakis, and P.-M. Agapow. Evaluating phylogenetic tree shape: two modifications to Fusco & Cronk's method. *J. Theor. Biol.*, 214:99–103, 2002.

[153] C. Reidys, C. V. Forst, and P. Schuster. Replication and mutation on neutral networks. *Bull. Math. Biol.*, 63:57–94, 2001.

[154] M. Ridley. *As Origens da Virtude*. Record, Rio de Janeiro, 1a. edição, 2000.

[155] C. Rispe and N. A. Moran. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *Am. Nat.*, 156:425–441, 2000.

[156] A. Rogers and A. Prügel-Bennett. Evolving populations with overlapping generations. *Theor. Pop. Biol.*, 57:121–129, 2000.

[157] J. S. Rogers. Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution*, 48:2026–2036, 1994.

[158] N. A. Rosenberg and M. Nordborg. Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nature Rev. Genet.*, 3:380–390, 2002.

[159] G. Rowe. *Theoretical Models in Biology*. Clarendon Press, Oxford, 1st edition, 1994.

[160] K-T. Shao and R. R. Sokal. Tree balance. *Syst. Zool.*, 39:266–276, 1990.

[161] K. L. Simonsen, G. Churchill, and C. F. Aquadro. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141:413–429, 1995.

[162] P. D. Sniegowski, P. J. Gerrish, T. Johnson, and A. Shaver. The evolution of mutation rates: separating causes from consequences. *BioEssays*, 22:1057–1066, 2000.

[163] A. O. Sousa, S. Moss de Oliveira, and A. T. Bernardes. Simulating inbreeding depression through the mutation accumulation theory. *Physica A*, 278:563–570, 2000.

[164] P. F. Stadler, W. Hordijk, and J. F. Fontanari. Phase transition and landscape statistics of the number partitioning problem. *Phys. Rev. E*, 67:056701, 2003.

[165] D. Stauffer, P. M. C. de Oliveira, S. Moss de Oliveira, T. J. P. Penna, and J. S. Sá Martins. Computer simulations for biological ageing and sexual reproduction. e-print cond-mat/0011524.

[166] W. Stephan, L. Chao, and J. G. Smale. The advance of Muller's ratchet in a haploid asexual population: approximate solutions based on diffusion theory. *Genet. Res.*, 61:225–231, 1993.

[167] M. Stephens. Inference under the coalescent. In D. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, volume 1, pages 213–238. Wiley, Chichester, UK, 2001.

[168] J. Swetina and P. Schuster. Self-replication with errors: a model for polynucleotide replication. *Biophys. Chem.*, 16:329–345, 1982.

[169] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

[170] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595, 1989.

[171] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.*, 26:119–164, 1984.

[172] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518, 1997.

[173] E. van Nimwegen and J. P. Crutchfield. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull. Math. Biol.*, 62:799–848, 2000.

[174] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720, 1999.

[175] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1st edition, 1953.

[176] G. P. Wagner and P. Krall. What is the difference between models of error thresholds and Muller's ratchet? *J. Math. Biol.*, 32:33–44, 1993.

[177] G. Watterson. On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.*, 7:256–276, 1975.

[178] G. A. Watterson. Heterosis or neutrality? *Genetics*, 85:789–814, 1977.

[179] M. L. Wayne and K. L. Simonsen. Statistical tests of neutrality in the age of weak selection. *Trends Ecol. Evol.*, 13(6):236–240, 1998.

[180] W. Weinberg. On the detection of heredity in man (em alemão). *Jh. Ver. Vaterl. Naturk. Wurttemb.*, 64:368–382, 1908.

[181] T. Wiehe. Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. *Genet. Res. Camb.*, 69:127–136, 1997.

[182] T. Wiehe, E. Baake, and P. Schuster. Error propagation in reproduction of diploid organisms: a case study on single peaked landscapes. *J. Theor. Biol.*, 177:1–15, 1995.

[183] C. O. Wilke. Adaptive evolution on neutral networks. *Bull. Math. Biol.*, 63:715–730, 2001.

[184] C. O. Wilke and C. Adami. The biology of digital organisms. *Trends Ecol. Evol.*, 17:528–532, 2002.

[185] C. O. Wilke and C. Adami. Evolution of mutational robustness. *Mut. Res.*, 522:3–11, 2003.

[186] C. O. Wilke, C. Ronnewinkel, and T. Martinetz. Dynamic fitness landscapes in molecular evolution. *Phys. Rep.*, 349:395–446, 2001.

[187] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412:331–333, 2001.

[188] E. O. Wilson. *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, 1st edition, 1975.

[189] E. O. Wilson. *Da Natureza Humana*. T. A. Queiroz e Editora da Universidade de São Paulo, São Paulo, 1a. edição, 1981.

[190] G. Woodcock and P. G. Higgs. Population evolution on a multiplicative single-peak fitness landscape. *J. Theor. Biol.*, 179:61–73, 1996.

[191] R. Wright. *O Animal Moral*. Editora Campus, Rio de Janeiro, 1a. edição, 1996.

[192] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366. Ithaca, 1932.

[193] E. Yuste, S. Sánchez-Palomino, C. Casado, E. Domingo, and C. López-Galíndez. Drastic fitness loss in human immunodeficiency virus type 1 upon serial bottleneck events. *J. Virol.*, 73:2745–2751, 1999.