Universidade de São Paulo
Instituto de Física de São Carlos

Análise Estatística da Teoria de Quase-espécies de Evolução Molecular

Domingos Alves

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

Orientador: *Prof. Dr. José Fernando Fontanari*

São Carlos

1999

Alves, D.

Análise Estatística da Teoria de Quase-espécies de Evolução Molecular/Domingos Alves – São Carlos, 1999.

138 p.

Tese (Doutorado)–Instituto de Física de São Carlos, 1999.

Orientador: Prof. Dr. José Fernando Fontanari

1. Quase-espécie. 2. Evolução Molecular. I. Título.

UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos

Av. Dr. Carlos Botelho, 1465
CEP 13560-250 - São Carlos - SP
Brasil

Fone (016) 273-9333
Fax (016) 272-2218

MEMBROS DA COMISSÃO JULGADORA DA TESE DE DOUTORADO DE DOMINGOS ALVES, APRESENTADA AO INSTITUTO DE FÍSICA DE SÃO CARLOS, UNIVERSIDADE DE SÃO PAULO, EM 24/03/1999.

COMISSÃO JULGADORA:

Prof. Dr. JOSÉ FERNANDO FONTANARI *(Orientador)* (IFSC-USP)

Prof. Dr. RICHARD CHARLES GARRATT (IFSC-USP)

Prof. Dr. ROBERTO NICOLAU ONODY (IFSC-USP)

Profa. Dra. CARLA GOLDMAN (IF/USP)

Profa. Dra. RITA M. ZORZENON DOS SANTOS (UFF/RJ)

Dedicado ao meu pai Domingos e às mulheres da minha vida: Néria, Magali e Bárbara.

Agradecimentos

É claro que durante a elaboração de uma tese ficamos em débito com muitas pessoas que, direta ou indiretamente nos auxiliaram, às vezes com um simples bate-papo em um insuspeitável cafezinho. Dentre essas pessoas tenho que destacar algumas que são as fundamentais.

Ao professor Fontanari, mais que um agradecimento, uma dívida de dedicar aos meus possíveis futuros alunos o mesmo profissionalismo e integridade científica que me foram dispensados. Sem sombra de dúvida sua orientação merece um destaque especial no que diz respeito à dedicação. Entretanto, gostaria de fazer menção à sua criatividade e principalmente à sua capacidade de discernir o que é relevante como produção científica.

Agradeço também aos companheiros do grupo, Daniela, Viviane, Ana e Fernando, pela ajuda, convivência amiga e conversa fácil. Particularmente, ao Camilo, ao Osame e ao Paulo pelos debates e questionamentos sempre intrigantes e entusiasmados.

Gostaria de lembrar minha colaboração com o Professor Pedro Marijuan. Suas idéias sobre sistemas biológicos sempre me encantaram e influenciaram, tendo um papel importante em minha formação.

Conhecer e interagir com a professora Rita Zorzenon nestes útimos anos foi extremamente frutífero para me decidir sobre o estudo fascinante de populações virais e suas estratégias para *enganar* o sistema imunológico. A esse respeito também, tenho que agradecer ao professor Paolo Zanotto por um simples e inesquecível seminário sobre evolução viral. Realmente seu entusiasmo científico é contagiante.

Gostaria de agradecer também ao meu irmão Gustavo pela preocupação e pelo apoio. Sinto-me em débito com ele em muitos aspectos, enfim, por todos esses anos como meu irmão mais novo. Ainda com respeito à família - a minha família - Magali e Bárbara, são realmente o começo, o meio e o fim de tudo o que me move e, em particular, as verdadeiras heroínas deste trabalho de tese, enfrentando de tudo, inclusive o meu antológico mau-humor.

Finalmente, agradeço ao Conselho Nacional de Pesquisa e Desenvolvimento Tecnológico (CNPq) que permitiu a realização deste trabalho ao conceder-me uma bolsa de Doutorado.

Lista de Símbolos

$A_i \rightarrow$ taxa de replicação das moléculas do tipo i

$A_{P^iP^j} \rightarrow$ taxa de replicação de indivíduos diplóides

$a \rightarrow$ valor seletivo da seqüência mestra

$a_0 \rightarrow$ valor seletivo das seqüências mutantes em um relevo de um pico

$\alpha \rightarrow$ ordem ou intensidade da interação epistática

$\alpha_{\max} \rightarrow$ valor crítico da intensidade da interação epistática acima da qual não ocorre o limiar de erro

$d(i,j) \rightarrow$ distância de Hamming entre a seqüência j e a seqüência i

$D_i \rightarrow$ taxa de desagregação da molécula do tipo i

$d \rightarrow$ número de zeros em uma dada seqüência

$\overline{d} \rightarrow$ distância de Hamming média normalizada de toda a população

$\mathbf{\Phi_0} \rightarrow$ fluxo de diluição global das macro-moléculas dentro do reator

$\Phi_k \rightarrow$ fluxo de monômeros

$\gamma \rightarrow$ parâmetro que determina o quão suave é o relevo replicativo com seleção truncada

$\gamma_{\max} \rightarrow$ valor crítico de γ abaixo do qual não há o fenômeno do limiar de erro

$h \rightarrow$ coeficiente de dominância

$h_{\max} \rightarrow$ valor crítico do coeficiente de dominância acima do qual não há limiar de erro

$I_i \rightarrow$ molécula do tipo i dentre as κ^L possíveis

$\kappa \rightarrow$ número de monômeros (ou alelos) diferentes possíveis em uma posição de uma dada seqüência (ou locus)

$L \rightarrow$ comprimento de cada seqüência ou genoma

$L_{\max} \rightarrow$ comprimento máximo de uma seqüência permitido pelo limiar de erro

$\lambda_{\max} \rightarrow$ maior auto-valor da matriz de replicação

$M_{PR} \rightarrow$ elemento da matriz de mutação da seqüência com R $1's$ para uma seqüência com P $1's$

$\mu = 1 - q \rightarrow$ taxa de mutação por monômero

$\mu_c = 1 - q_c \rightarrow$ taxa de mutação crítica para manter a seqüência com maior valor seletivo na população (define a localização do limiar de erro no relevo de

um pico)

$N_{ES} \rightarrow$ número de seqüências possíveis no espaço de seqüências

$N \rightarrow$ número de seqüências ou genomas

$N_{\min} \rightarrow$ tamanho da população abaixo do qual o limiar de erro não ocorre

$\mathbf{n} \rightarrow$ número de ocupação

$n_P(t) \rightarrow$ número de seqüências com P $1's$

$P \rightarrow$ número de 1 em uma dada seqüência

$p(t) \rightarrow$ freqüência de monômeros do tipo 1 na população

$p^* \rightarrow$ ponto fixo para a freqüência de monômeros

$\Pi_P(t) \rightarrow$ freqüência das seqüências da classe P na geração t

$\Pi_{P^iP^j}(t) \rightarrow$ freqüência dos indivíduos diplóides

$q \rightarrow$ fidelidade de replicação de cada monômero

$\rho \rightarrow$ parâmetro que mede o tamanho da região chata em um relevo replicativo com seleção truncada

$s \rightarrow$ valor seletivo de um gen deletério

$s^i_\beta \rightarrow$ variável indicando o monômero (ou alelo) de uma molécula do tipo i na posição β

$\sigma_d \rightarrow$ dispersão de d em torno de $\overline{d}$

$\sigma_N \rightarrow$ dispersão que leva em consideração as flutuações em Π_P

$\sigma_N(L) \rightarrow$ espalhamento da seqüência ótima na população

$t \rightarrow$ tempo ou geração

$x_i = [I_i] \rightarrow$ concentração da molécula do tipo i

$\mathbf{W} \rightarrow$ matriz de replicação

$W_{ij} \rightarrow$ elemento da matriz de replicação com erro da seqüência j à seqüência i

$W_P(\mathbf{n}) \rightarrow$ taxa de replicação relativa da seqüência da classe P

$\overline{W} \rightarrow$ *fitness* médio da população

$y_i = x_i / \sum_j x_j \rightarrow$ concentração relativa da seqüência do tipo i

Resumo

Nesta tese propomos e estudamos um modelo alternativo para investigar a evolução de quase-espécies moleculares, no qual supomos que a população seja uma combinação aleatória das moléculas constituintes em cada geração. Essa aleatoriedade deve-se a inclusão de um procedimento adicional de amostragem da população além dos procedimentos usuais de mutação e reprodução diferenciada. O modelo, denominado modelo de amostragens, é baseado em um algoritmo que *mimetiza* procedimentos experimentais que usam técnicas de transferências em série para reproduzir o processo de evolução de microorganismos *in vitro*. Além do modelo reproduzir a solução exata do estado estacionário do modelo de quase-espécies no regime determinístico, ele permite o estudo da evolução da quase-espécie molecular em todo espaço de parâmetros de controle, incluindo o caso em que a população é finita. A generalização dessa formulação alternativa para uma classe geral de relevos de replicação permite-nos realizar um estudo bastante completo do fenômeno do limiar de erro, levando-nos a uma análise crítica sobre a generalidade desse fenômeno.

Abstract

In this thesis we propose and study an alternative model to investigate the evolution of a molecular quasispecies, in which we assume that the population is a random combination of the constituent molecules in each generation. This randomness is due to the inclusion of an additional sampling procedure of the population, besides the usual procedures of mutation and differential reprodution. This model, termed sampling model, is based on an algorithm that *mimics* experimental procedures using serial transfer techniques to study the microbial evolutionary process *in vitro*. Besides yielding the exact steady-state solution of the quasispecies model in the deterministic limit, the sampling model allows the study of the molecular evolution in the full space of the control parameter, including the case where of population is finite. The generalization of this alternative formulation to a general class of fitness landscapes, allows us to investigate thoroughly the error threshold phenomenon, leading us to discuss critically the generality of this phenomenon in molecular evolution.

Conteúdo

...essa é a moral da história.

Nós estamos ainda como as crianças de Newton,

brincando nas praias de um oceano.

Karl Sigmund - Games of Life.

Introdução

Ainda deve parecer estranho a muitos pesquisadores deparar-se com uma monografia de tese de doutorado, apresentada em um instituto de Física, que trata de aspectos teóricos de um tema acreditado ser debatido exclusivamente entre os biólogos: o da teoria da seleção natural. Apesar da aparente novidade, essa pesquisa é motivada e tem o respaldo em muitos programas desenvolvidos em importantes centros de Física. Realmente, muito do que tem sido feito em Física atualmente está direcionado para o entendimento de sistemas mais *complexos* do que os tradicionalmente considerados como objeto de estudos de um físico. Mesmo não havendo um consenso sobre o que define a complexidade de um sistema, todos concordam que o primeiro exemplo de um sistema complexo é um organismo vivo. Então, para ilustrar o tipo de complexidade com o qual um físico tem que lidar ao estudar um sistema vivo, poderíamos dizer que a presente monografia emerge da atividade de uma população em torno 10^{13} células que cooperam de uma maneira organizada e funcionalmente coordenada. Mesmo a atividade de uma única célula simples é o resultado de um conjunto de interações monumentalmente complexas. Na mais simples célula bacteriana, mais de 10^7 macro-moléculas biológicas interagem coerentemente de modo a sustentar o estado ordenado o qual chamamos *vida*. Cada uma dessas moléculas carrega um *programa* que coordena todas as informações disponíveis e que é uma entre mais de $10^{2000000}$ possibilidades. Esses números, entretanto, deixam dúvidas se as idéias e métodos da Física contemporânea são suficientes para explicar sis-

temas com esse tipo de complexidade. Daí a necessidade de restringir nossas considerações a um fenômeno biológico particular, sem o apelo filosófico de um reducionismo da Biologia à Física, mas sustentando a possibilidade de se fazer uma Física do processo evolucionário.

Dessa maneira, a presente monografia se enquadra em um programa de pesquisa cujo objetivo é aplicar métodos da Física Estatística e Sistemas Dinâmicos, concomitantemente ao desenvolvimento de modelos computacionais, para estudar e simular problemas ligados à evolução biológica. Para especificar como pretendemos contribuir com o tema, devemos primeiramente delinear o contexto em que este trabalho é desenvolvido.

Acredita-se que o princípio natural que governa a dinâmica de entidades vivas, isto é, sistemas capazes de se auto-sustentar e auto-reproduzir, ocasionalmente com variações na descendência, seja a evolução por seleção natural proposta por Darwin. De uma maneira muito simplificada, este princípio estabelece que se indivíduos geneticamente distintos competirem por recursos limitados, aqueles mais adequados ao meio produzirão mais descendentes, tornando-se os maiores *acionistas genéticos* da próxima geração. Somado a isso, mutações aleatórias misturam os genes para criar descendentes com novas combinações genéticas e, a cada geração, o crivo da seleção natural elimina as combinações menos eficientes produzindo organismos cada vez mais adaptados ao meio.

Essas idéias podem ser entendidas de uma maneira mais formal através da teoria de quase-espécies proposta por M. Eigen [1], originalmente no contexto de evolução pre-biótica, para descrever a dinâmica de macro-moléculas replicadoras (ácidos nucleicos) sob a influência dos mecanismos de seleção e mutação. Em um sentido mais geral, essa teoria descreve qualquer população de organismos que se auto-reproduzem. Na formulação original de cinética química, o modelo de quase-espécies é descrito por um conjunto de equações diferenciais ordinárias

para as concentrações dos diferentes tipos de moléculas que compõem a população, sendo válido somente no limite em que o número total de moléculas vai a infinito. Talvez o resultado mais contundente dessa teoria é que os efeitos da seleção e mutação levam à coexistência de vários tipos dessas moléculas no equilíbrio: a *quase-espécie*. Outra conclusão importante é que existe um limite superior para a taxa de mutação que pode ser tolerada por essa população de moléculas: o *limiar de erro* de replicação.

Diante desse quadro, organizamos a monografia como segue. No capítulo 1 descrevemos a teoria de quase-espécie em sua formulação original determinística, porém em um contexto um pouco diferente, procurando apresentar suas principais conclusões como sendo válidas para um *sistema darwiniano* geral. Na última seção desse capítulo fazemos uma análise crítica do modelo que servirá de guia para os próximos capítulos. Assim, o capítulo 2 é dedicado a uma formulação alternativa da teoria original que, a despeito de sua extrema simplicidade, produz vários resultados não divisados no modelo original, incluindo os efeitos de população finita. Esse modelo será derivado de um algoritmo para a dinâmica estocástica de moléculas replicadoras, no qual a ênfase é dada à competição entre seleção natural e mutação. Particularmente, o algoritmo é construído *mimetizando* técnicas experimentais de passagem em série de *amostras* de populações de micro-organismos. A robustez desse *modelo de amostragens* é demonstrada na generalização feita no capítulo 3, onde estudamos a evolução de uma *quase-espécie* desde ambientes que selecionam apenas indivíduos extremamente especializados, até ambientes em que é suposto algum tipo de cooperação entre os componentes da população. No capítulo 4 mostramos como podemos utilizar os conceitos de *quase-espécie* e evolução molecular para avançar no entendimento dos modelos de genética de populações para indivíduos haplóides e diplóides com muitos *loci* em seu genoma. O capítulo 5 é escrito com o intuito de apresentar as conclusões

finais deste trabalho. Em particular, discutimos as possibilidades de extensão do modelo desenvolvido a outros mecanismos evolucionários, além dos de seleção e mutação, bem como as perspectivas de aplicabilidade do formalismo desenvolvido ao estudo da evolução de uma população viral.

Capítulo 1

Micro-evolução darwiniana

Neste capítulo vamos descrever as principais características da teoria de quase-espécies proposta por M. Eigen [1] para tentar explicar como a vida se originou na Terra. Neste trabalho, todavia, não vamos nos referir a esse contexto (para uma revisão completa sobre o tema o leitor dispõe de dois ótimos livros [2, 3]). De uma maneira um pouco diferente, vamos construir o modelo a partir de um sistema simples composto de certas moléculas transportadoras de informação e que são formadas, inicialmente ao acaso, por seqüências de monômeros distintos. Estas moléculas quando submetidas a um processo prévio de seleção e evolução molecular organizam-se de uma maneira muito elaborada. Como veremos, isto se deve ao seu potencial de variabilidade, determinado pelo número quase astronômico de seqüências possíveis que se podem formar. Muitos dos mecanismos e inter-relações que iremos postular para a modelagem dessas moléculas, e que conduzem a um processo de auto-organização, estão implicitamente presentes em outros tipos de sistemas de maior complexidade de modo que muitas das conclusões extraídas dessa análise podem ser extrapoladas a esses sistemas e vice-versa.

1.1 Um sistema modelo para a evolução molecular

Genericamente, o problema que vamos discutir é o seguinte: dada uma distribuição inicial de seqüências, que tipos de processos físico-químicos levam a seleção e posterior evolução dessas moléculas até que as mais adequadas para determinados fins sejam predominantes no sistema? Para isso, antes de tudo, devemos determinar quais são as propriedades mínimas que devem ter essas moléculas para que sobre elas possa ocorrer um processo de seleção e evolução darwiniana, ou seja, quais as condições necessárias para que um *sistema* possa ser chamado de *vivo*. Essas propriedades são os critérios adotados em biologia moderna para demarcar a fronteira entre organismos vivos e não vivos. Assim, sem especificar os detalhes dos processos biológicos, postulamos que o sistema de moléculas em questão deve possuir as seguintes propriedades:

1. *As moléculas devem estar submetidas a um metabolismo contínuo*, isto é, existe um processo de síntese e degradação das moléculas. Para manter continuamente este metabolismo é necessário que o sistema se encontre muito longe do equilíbrio, mediante um aporte contínuo de monômeros ricos em energia, já que os resultantes da degradação do polímero não são aproveitáveis para formar novos polímeros.

2. *A síntese das moléculas se dá mediante uma atividade de auto-replicação*, isto é, formam-se cópias das moléculas que existem previamente no sistema. Supõe-se portanto que não haja possibilidade de uma síntese *de novo*, apesar de alguns sistemas virais apresentarem esta propriedade em experimentos *in vitro*.

3. *Durante o processo de cópia pode-se introduzir erros.* Esta propriedade de *mutabilidade* é que dará lugar a possibilidade de evolução dessas moléculas.

4. *Deve existir algum tipo de competição entre as unidades auto-replicativas distintas.* Usualmente (mas nem sempre, como veremos) esta competição é conseqüência da imposição de algum tipo de restrição ao meio. Esta restrição é que conduzirá a um processo de seleção no sistema e a maior ou menor adaptabilidade (*fitness*) de um *fenótipo* particular ao seu entorno.

Todas estas características estão implícitas no *reator de seleção e evolução* ou simplesmente *reator de fluxo* representado na figura (1.1). Desde que pre-

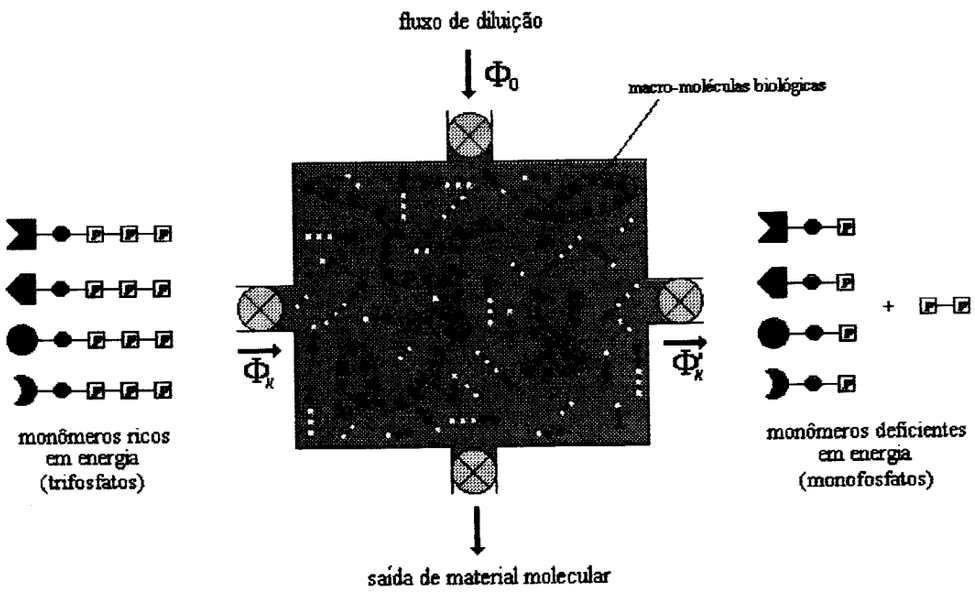

Figura 1.1: Reator de fluxo a ser usado como sistema modelo para o processo de auto-organização molecular. Aqui as macro-moléculas biológicas são continuamente construídas a partir de monômeros ricos em energia. Definidas as condições de reação, pode-se regular o fluxo desses monômeros $\mathbf{\Phi} = \{\Phi_1, \Phi_2, ..., \Phi_\kappa\}$ e o fluxo de diluição global Φ_0, que efetivamente controla a população total de macro-moléculas, fazendo com que o sistema possa evoluir, por exemplo, em fluxo constante ou concentração constante. Esse sistema pode induzir, sob condições apropriadas, competição de seleção entre as várias seqüências e portanto simular o processo básico de evolução molecular.

tendemos definir matematicamente os princípios da seleção e evolução molecular,

iremos primeiramente concentrar nossa atenção na construção desse sistema modelo, o que irá ajudar na visualização do processo.

Basicamente, este aparato contém macro-moléculas biológicas (ácidos nucleicos ou proteínas) que estão constantemente sendo construídas com monômeros ricos em energia. Ainda, podemos supor que o movimento molecular térmico, por exemplo, diminui o tempo de vida de todas moléculas e de todos os estados ordenados resultantes das interações moleculares. Portanto, essas macro-moléculas biológicas decaem depois de algum tempo a monômeros deficientes em energia por simples dissociação. A fim de evitar também que esse sistema volte a um estado de equilíbrio termodinâmico, suporemos que o reator tenha paredes semi-permeáveis, por meio das quais o sistema está perpetuamente trocando energia e matéria com o meio. Através de uma dessas paredes, monômeros ricos em energia são introduzidos, enquanto que através de outra, os produtos do decaimento (monômeros deficientes em energia, etc.) são removidos. Essas condições podem ser mantidas regulando-se o fluxo total de monômeros e o fluxo de diluição Φ_0, que controla a população total de polímeros.

Podemos agora especificar os componentes químicos do nosso reator de fluxo. Uma simples macro-molécula é constituída de L subunidades, as quais, por sua vez, podem existir em κ formas diferentes. Entre as macro-moléculas biológicas, temos $\kappa = 4$ para os ácidos nucleicos (G, A, C, U/T) ou $\kappa = 2$ se os agruparmos nos dois tipos de bases existentes (purinas e pirimidinas), ou ainda, $\kappa = 20$ para as proteínas. Os polímeros podem ser divididos de acordo com seu comprimento, de maneira que uma medida da capacidade de informação para um conjunto de moléculas de comprimento L particular seja o número $N_{es} = \kappa^L$, ou seja, o número de todas as seqüências de L símbolos combinatorialmente possíveis. Logo, estabelecidas as regras dinâmicas a que o sistema está sujeito, este número define a quantidade de pontos do *espaço de seqüências* (ou de genomas) no qual

o sistema pode mover-se. Assim, para evitarmos complicações por trabalharmos com seqüências de comprimentos distintos dentro do reator, no que segue consideraremos somente polímeros com comprimento constante, o que simplifica o tratamento matemático.

1.2 O ansatz da cinética química

Para avançarmos na descrição formal desse sistema, modelamos cada macromolécula replicadora como uma seqüência de L dígitos $I_i = (s_1^i, s_2^i, \ldots, s_L^i)$, com as variáveis s_β^i $(i = 1, ..., \kappa^L; \beta = 1, ..., L)$ tomando κ valores diferentes, cada um representando um tipo diferente de monômero usado para formar a molécula. Desse modo, para uma seqüência particular do tipo i, a série de eventos que podem ocorrer dentro do reator de fluxo pode ser modelada através de passos simples de reações químicas, a saber,

$$(T) + I_i \xrightarrow{W_{ii}} 2I_i \tag{1.1}$$

$$(T) + I_i \xrightarrow{W_{ji}} I_i + I_j \quad j \neq i \tag{1.2}$$

$$I_i \xrightarrow{D_i} (M) \tag{1.3}$$

$$I_i \xrightarrow{\Phi_0} 0. \tag{1.4}$$

Nessas reações, os monofosfatos M (monômeros deficientes em energia) e os trifosfatos T (monômeros ricos em energia) são colocados entre parentesis, desde que suas concentrações não são consideradas como variáveis.

A reação (1.1) representa a auto-replicação fiel da molécula I_i e a reação (1.2) a auto-replicação com erro permitindo a formação da molécula I_j devido a

mutação de I_i. A matriz de replicação W leva em conta a estrutura primária das moléculas, sendo seus elementos dados por

$$W_{ii} = A_i\, q^L \tag{1.5}$$

e

$$W_{ij} = \frac{A_j}{(\kappa - 1)^{d(i,j)}}\, q^{L-d(i,j)}\,(1-q)^{d(i,j)} \qquad i \neq j, \tag{1.6}$$

onde A_i é a taxa de replicação das moléculas do tipo i, $d\,(i,j)$ é a distância de Hamming entre as moléculas i e j, ou seja, dadas duas seqüências i e j, $d(i,j)$ conta o número de dígitos em que essas seqüências diferem. Note que $A_i W_{ij} = A_j W_{ji}$. Aqui, $q \in [0,1]$ é o parâmetro que mede a fidelidade de replicação de cada monômero, que é suposta a mesma para todos monômeros. Assim, auto-replicação e mutação são representadas por duas reações de auto-catálise, homogênea e heterogênea, respectivamente. A reação (1.3) corresponde a degradação da molécula I_i, sendo a constante D_i a taxa de desagregação das moléculas do tipo i. A reação (1.4) representa a saída da molécula I_i por difusão, sendo Φ_0 o coeficiente de difusão global que é suposto o mesmo para todas as moléculas.

As variáveis relevantes desse sistema dinâmico são as concentrações de cada seqüência $x_i = [I_i]$. Assim, dentro do reator, a evolução temporal da concentração x_i das moléculas do tipo $i = 1, 2, \ldots, \kappa^L$ para esse sistema obedece a equação diferencial,

$$\frac{dx_i}{dt} = \sum_j W_{ij} x_j - [D_i + \Phi_0]\, x_i. \tag{1.7}$$

O conjunto de equações diferenciais ordinárias para as concentrações dos diferentes tipos de moléculas que compõe o reator sintetiza o modelo de quase-espécies. Este formalismo, entretanto, é válido somente no limite em que o número total de moléculas N vai a infinito, necessitando de uma reformulação completa para levar em conta os efeitos de população finita. Além disso, a aplicabilidade da

cinética química convencional ao problema de evolução, por si só, é uma questão bastante sutil. Na seção (1.5) retomaremos esse ponto para analisar os limites da teoria convencional de quase-espécies.

Como foi dito anteriormente, para que apareça um processo de seleção é necessário impor algum tipo de restrição ao sistema, que pode aparecer na natureza de muitas formas distintas. Para apreciarmos o efeito da competição em nosso sistema (que pode levar ou não à seleção), vamos imaginar a situação ideal em que a auto-replicação das moléculas se dá sem nenhuma possibilidade de erro $q = 1$. Nesse caso todos os elementos fora da diagonal principal da matriz de replicação W são nulos, e a equação (1.7) reduz-se a forma simples

$$\frac{dx_i}{dt} = [W_{ii} - D_i - \Phi_0]\, x_i. \tag{1.8}$$

É fácil ver que esta equação para a concentração de moléculas da classe i tem uma solução exponencial trivial no caso em que Φ_0 é mantido constante. Desse modo, se existirem k espécies de moléculas distintas dentro do reator, todas aquelas cujas taxas de replicação W_{kk} forem maiores que $D_k + \Phi_0$ terão suas concentrações aumentadas exponencialmente sem limites. Ao contrário, as moléculas cujas taxas de replicação forem menores que $D_k + \Phi_0$ diminuirão suas concentrações até desaparecerem da população. Este comportamento está esquematizado na figura (1.2) (a) para o caso de cinco moléculas binárias ($\kappa = 2$, $s_k = 0, 1$).

Ainda com relação ao caso da replicação sem erro, podemos impor agora um vínculo global ao qual a população de moléculas dentro do reator estará sujeita. Aqui só discutiremos o caso em que a *concentração total de moléculas* dentro do reator, $\sum_i x_i = N$, é mantida *constante*. Isto faz com que o termo Φ_0 varie no tempo, sendo determinado pela condição $\sum_i dx_i/dt = 0$. Portanto, o fluxo global de moléculas dentro do reator terá de ser controlado dependendo da produtividade

e da concentração das moléculas sendo dado por

$$\Phi_0 = \frac{\sum_i (W_{ii} - D_i)x_i}{N}. \tag{1.9}$$

No caso de replicação imperfeita ($q < 1$), considerações similares levam a

$$\Phi_0 = \frac{\sum_i \sum_j W_{ij}x_j - \sum_i D_i x_i}{N}. \tag{1.10}$$

Note que essas equações para Φ_0 tornam a equações (1.7) e (1.8) não lineares.

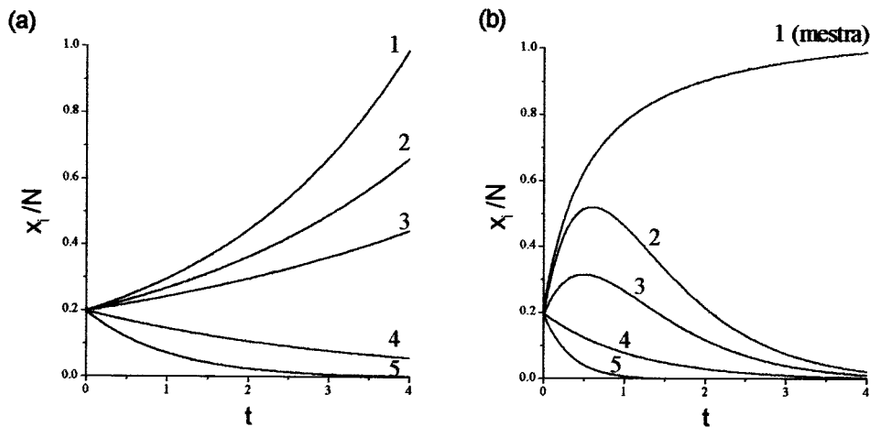

Figura 1.2: Variação da freqüência de 5 tipos de moléculas binárias no reator de fluxo, sem possibilidade de erro no processo de auto-replicação. (a) segregação: dependendo dos valores de W_{ii}, D_i e Φ_0 tem lugar um crescimento exponencial das moléculas. (b) seleção: uma vez submetidas a uma restrição de população constante, somente aquela com maior produtividade, $W_{ii} - D_i$, sobreviverá. No caso (a) mantivemos $\Phi_0 = 1.0$ e constante e em ambos os casos fizemos: $D_i = 0.5$, $W_{11} = 1.9$, $W_{22} = 1.8$, $W_{33} = 1.7$, $W_{44} = 1.2$ e $W_{55} = 0.5$.

A figura (1.2) (b) esquematiza o comportamento temporal da mesma população do exemplo mostrado na figura (1.2) (a) quando impomos a restrição de concentração constante ao sistema. Podemos fazer assim algumas considerações qualitativas, que intuitivamente ajudem a compreender o comportamento do sistema. Se interpretarmos o termo $W_{ii} - D_i$ como a produtividade ou replicação

líquida de cada molécula, o termo Φ_0 será a produtividade média da população, como pode ser visto da equação (1.9). De fato, diferentemente do caso sem restrição em que Φ_0 era mantido constante, ao se impor que a concentração total permaneça constante, o fluxo global deve aumentar cada vez mais. Isto, por sua vez, faz com que a medida que o tempo passa um número maior de tipos de moléculas sejam segregados da população de acordo com (1.8). Ao final desse processo, somente a molécula com produtividade maior que a média será *selecionada* frente a todas as outras da população. O processo de seleção é portanto um efeito do meio através de um vínculo imposto à população.

No estado estacionário desta competição, chamamos de *seqüência mestra* (I_m) àquela que é selecionada quando a população é submetida a algum tipo de restrição. O estado estacionário desse sistema é denominado de *equilíbrio de seleção.*

1.3 O que é uma quase-espécie?

Vamos agora discutir o caso mais geral, descrito pela equação (1.7), em que é levado em consideração a possibilidade de erros no processo de auto-replicação. As soluções no equilíbrio de seleção para essa equação podem ser encontradas em termos dos auto-valores e auto-vetores da matriz de replicação W [5, 6]. Se definirmos o vetor $\mathbf{y} = (y_1, y_2, ...y_L)$, cujas componentes representam as concentrações relativas de cada espécie molecular dentro de toda população, $y_i = x_i / \sum_j x_j$, podemos escrever

$$W\,'\mathbf{y} = \lambda \mathbf{y}, \tag{1.11}$$

onde os elementos da diagonal de W foram modificados de forma a absorver D_i, ou seja, $W'_{ii} = W_{ii} - D_i$. De fato, a transformação do sistema não linear descrito pelas equações (1.7) e (1.10) no sistema linear acima envolve uma álgebra relativamente simples explicada em detalhes nos apêndices do artigo de revisão [4]. Daí, definimos uma quase-espécie precisamente em termos matemáticos como

sendo o auto-vetor dominante $\mathbf{y}_{\max}$ associado ao maior auto-valor $\lambda_{\max}$ da matriz de replicação W'. Este auto-vetor descreve a estrutura exata da população: cada mutante I_i está presente na quase-espécie com uma freqüência y_i ($\sum_i y_i = 1$). O maior auto-valor é exatamente a taxa de replicação média da quase-espécie, $\lambda_{\max} = \sum_i A_i y_i$. Dessa maneira, a freqüência de um dado mutante dentro da quase-espécie não depende apenas de sua taxa de replicação, mas também da probabilidade com que ele é produzido devido a erros na replicação de outras espécies moleculares. Na figura (1.3) ilustramos a estrutura da população das

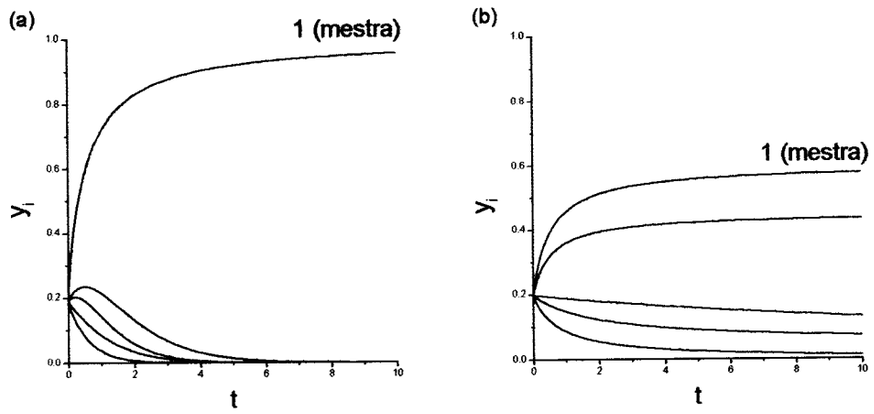

Figura 1.3: Comportamento do sistema das 5 moléculas binárias da figura (1.2) competindo com possibilidade de erro no processo de auto-replicação. (a) Quando a taxa de mutação é baixa ($\mu = 0.01$) a seqüência mestra eventualmente domina a população. (b) Com o aumento da taxa de mutação ($\mu = 0.1$), a mestra e alguns (ou todos) mutantes coexistem como uma quase-espécie no equilíbrio de seleção.

cinco seqüências do exemplo sem mutação (figura (1.2)) para dois valores distintos de taxa de mutação por monômero $\mu = (1 - q)$, mostrando o seu comportamento ao aproximar-se do equilíbrio de seleção.

Assim, um dos resultados cruciais da teoria é que, no equilíbrio, a seleção não leva em geral a uma população homogênea formada por um único tipo de

indivíduo molecular mais apto, mas sim a um *ensemble* de variantes geneticamente distintas mas bastante próximas. Para distinguir do conceito clássico de espécie já existente em biologia, este *ensemble* foi então chamado de *quase-espécie*. Segundo essa teoria, a adaptação é uma propriedade não de uma seqüência particular mas dessa distribuição de mutantes centrada em torno da *seqüência mestra* (aquela com maior *fitness* ou valor reprodutivo do ensemble).

A mutação nesse modelo é o parâmetro que caracteriza a largura da distribuição, ou seja, o quanto a *quase-espécie* está espalhada no espaço de seqüências. Esse novo conceito tem importantes implicações, desde que a evolução é normalmente pensada como a interação entre mutação e seleção, sendo este último fator aquele que favorece mutantes com vantagens reprodutivas que tenham sido gerados puramente ao acaso. Realmente, é considerado um erro pensar em mutações sendo guiadas de outra maneira que não ao acaso. Contudo, uma quase-espécie pode guiar mutações. Isso não significa que haja qualquer correlação entre o fenômeno intrinsicamente probabilístico da mutação e a vantagem seletiva de um mutante, mas que a seleção opera sobre a estrutura de toda a quase-espécie, que por sua vez é adaptada ao seu relevo replicativo ou *fitness landscape* (este termo foi originalmente introduzido por S. Wright [7]). Assim, a evolução pode ser guiada na direção dos *picos* deste relevo, isto acontecendo porque os mutantes mais bem sucedidos (que podem estar perto dos picos do relevo) irão produzir mais descendentes que os outros mutantes na população (que podem estar longe dos picos). Neste sentido, este processo é comumente interpretado [1, 4] como uma *escalada da montanha reprodutiva* das quase-espécies, que ocorre por meio de certos caminhos no espaço de seqüências.

1.4 O paradoxo de Eigen

À medida que a taxa de mutação aumenta, o modelo prevê que a composição da população deixa de estar distribuída como uma *quase-espécie*, passando a um regime no qual a distribuição dos κ^L tipos de moléculas é *uniforme* (todas as seqüências aparecem em proporções iguais). Em outras palavras, como comentamos antes, se a replicação ocorresse livre de erros, nenhum mutante apareceria e a evolução cessaria; contudo a evolução também deve ser impossível se o erro na replicação for muito alto (se houver poucos mutantes na população, isso pode melhorar a adaptação, porém se houver muitos, eles irão levar a deterioração da população). A transição entre esses dois regimes é outro importante resultado da teoria, conhecido por *limiar* ou *catástrofe de erro de replicação.*

É interessante observar como o fenômeno do *limiar de erro* manifesta-se para a equação (1.7) que caracteriza o modelo determinístico de quase-espécies. Para isso, precisamos definir primeiramente o mais simples, e talvez o mais utilizado, relevo de replicação em que aparece esse fenômeno - o *relevo de replicação de um pico* - que pode ser visto como uma aproximação local de um relevo rugoso mas com picos distantes. Matematicamente, atribuímos a taxa de replicação $A_m = a > 1$ à seqüência mestra, e $A_d < a$ às seqüências remanescentes. A figura (1.4) mostra a distribuição de quase-espécies para seqüências binárias de tamanho $L = 30$ no equilíbrio de seleção em função da taxa de mutação $\mu = (1 - q)$. Nesse caso, o número possível de seqüências no reator de fluxo é $N_{es} = 2^{30}$. Portanto, para uma melhor visualização, as seqüências são agrupadas em classes de mutantes: todas as seqüências distantes d mutações da mestra são membros da classe d. Vamos supor que todos os mutantes tenham a mesma taxa de replicação, isto é, $A_d = 1 \ \forall \ d \neq 0$, enquanto que a seqüência mestra ($d = 0$) possui uma taxa maior, $A_0 = a$. Este procedimento de estratificação reduz o número de

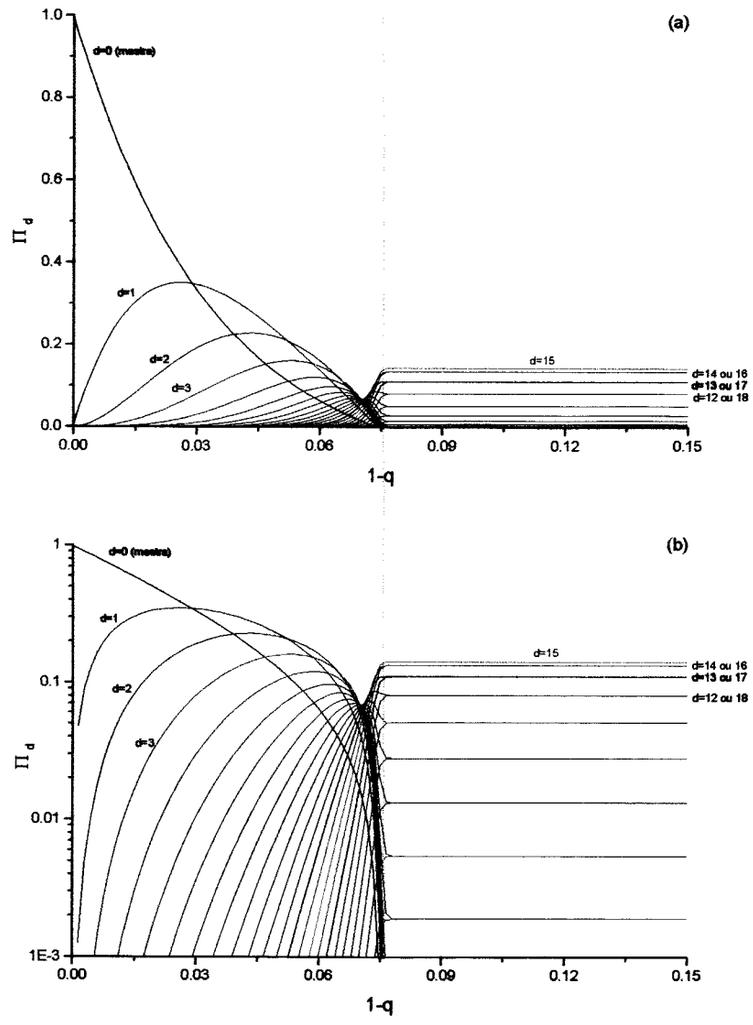

Figura 1.4: Freqüência de classes de seqüências em função da taxa de mutação $\mu = (1 - q)$. Aqui d é o número de mutações entre uma seqüência particular e a mestra. As seqüências binárias têm comprimento $L = 30$ e a taxa de replicação da mestra é $a = 10$. Destacamos os primeiros mutantes da mestra, que já aparecem na população para valores pequenos de μ, bem como alguns mutantes mais distantes que só começam a aparecer para valores de mutação perto do limiar de erro (linha vertical em verde). Após o limiar de erro todas as seqüências (não as classes) são igualmente prováveis. Desde que $\binom{30}{14} = \binom{30}{16}$, por exemplo, as freqüências das duas classes são iguais nessa região. Na parte (b) da figura mostramos o mesmo exemplo em escala logarítmica para ressaltar o caráter abrupto da transição.

equações da representação da cinética química de 2^L para apenas $L+1$ equações diferenciais de primeira ordem acopladas [8], cada qual descrevendo a evolução de uma das classes de moléculas dentro do reator. No caso do relevo de replicação em questão, essas equações são do tipo

$$\frac{d\Pi_d}{dt} = \sum_{R=0}^{L-1} M_{dR}\Pi_R + a\Pi_0 M_{d0} - \Pi_d\left[1 + \Pi_0(a-1)\right], \tag{1.12}$$

onde Π_d denota a concentração de molécula da classe $d = 0, ..., L$, com $\sum_d \Pi_d = 1$. O preço a pagar pela redução no número de equações é o considerável aumento de complexidade da matriz de mutação, entre classes cujos elementos M_{ij} $(i, j = 0, .., L)$ são obtidos no Apêndice A. A figura (1.4) mostra então a freqüência de cada uma das classes de seqüências mutantes e a da mestra (em azul) em função de μ, no equilíbrio de seleção. Cada uma das curvas é obtida fazendo-se $d\Pi_d/dt = 0$ em (1.12). Ainda nessa figura, destacamos a linha de transição separando o regime caracterizado por uma *quase-espécie* e o regime uniforme: *o limiar de erro.*

Esta transição é uma transição de fase termodinâmica genuína do tipo ordem-desordem [14] apenas no caso em que $L \to \infty$ para o relevo replicativo de um pico [15]. Mais recentemente, essa transição foi caracterizada também para L finito [13].

Para derivarmos uma expressão para o fenômeno de *limiar de erro* de uma maneira bastante simples [11, 12, 13], vamos supor que a população consista de duas subpopulações: uma contendo a seqüência mestra (I_m), ou a mais apta dentre todas; e outra formada pela *cauda de erro* da quase-espécies, ou seja, todos os mutantes da distribuição. Estas útimas são substituídas por uma seqüência média (I_{ce}). Vamos também desprezar a pequena probabilidade de que uma seqüência mutante dê origem à uma mestra através de uma mutação reversa, desde que é muito mais provável que elas mutem para outras seqüências na cauda

de erro. Dessa maneira, a probabilidade de que a mestra replique sem erro é q^L. Com essas hipóteses simplificadoras, a equação (1.7) é reescrita para as duas subpopulações como:

$$\frac{dx_m}{dt} = A_m q^L x_m - x_m \left[A_m x_m + A_{ce} x_{ce}\right] \tag{1.13}$$

$$\frac{dx_{ce}}{dt} = A_{ce} x_{ce} + A_m (1 - q^L) x_m - x_{ce} \left[A_m x_m + A_{ce} x_{ce}\right], \tag{1.14}$$

onde fizemos ainda, por simplicidade, $D_i = 0 \ \forall i$ e tomamos $x_m + x_{ce} = 1$, como o vínculo de população constante (uma classe só pode crescer às expensas da outra). As parcelas entre colchetes em ambas equações correspondem a Φ_0.

Estamos interessados na coexistência da mestra e os mutantes no equilíbrio. Igualando ambas derivadas a zero, e impondo que neste equilíbrio tenhamos $x_m \neq 0$ obtemos, no caso de o relevo de replicação ser de um pico, a seguinte condição

$$q^L > \frac{A_{ce}}{A_m} = \frac{1}{a}. \tag{1.15}$$

Isso implica que q tem de ser maior que um valor mínimo $q_{min} = (A_{ce}/A_m)^{1/L}$ para que a seqüência mestra não seja perdida da população. Este valor mínimo é então interpretado como o *limiar de erro* de replicação. Isto leva a uma relação importante entre a precisão da replicação q e o comprimento da seqüência L,

$$L < \frac{\ln a}{(1-q)}, \tag{1.16}$$

que é obtida tomando-se o logaritmo de (1.15) e fazendo-se a aproximação $\ln(q) \approx q - 1$. Esta equação indica que a quantidade de informação que pode ser seletivamente mantida (L) é limitada pela fidelidade de cópia por dígito (q).

A relação descrita pela equação (1.16) é mostrada na figura (1.5) e representada pela área abaixo da hipérbole em azul. Quanto maior for a fidelidade na cópia, maior será a freqüência da seqüência mestra que pode ser seletivamente mantida. Se a cópia mestra cresce perto do *limiar de erro* ela deve deteriorar-se

rapidamente. Abaixo do *limiar de erro* entretanto, a população deve consistir da seqüência mestra circundada por uma *nuvem* de seus mutantes mais próximos: a quase-espécie.

Assim, outra das conseqüências do fenômeno da catástrofe de erro de replicação é o *paradoxo de Eigen*: o tamanho de uma molécula sujeita a esse mecanismo de replicação não pode exceder a um certo comprimento máximo L_{max}, fixado o grau de precisão ou confiabilidade do mecanismo de cópia; por outro lado, o sistema de moléculas não pode aperfeiçoar seu mecanismo de cópia enquanto seu comprimento for limitado, pois para instruir uma enzima, por exemplo, a fazer cópias com alta fidelidade necessita-se de um conteúdo mínimo de informação L_c, e que pode ser maior do que o permitido pelo limiar de erro. Dessa maneira, a

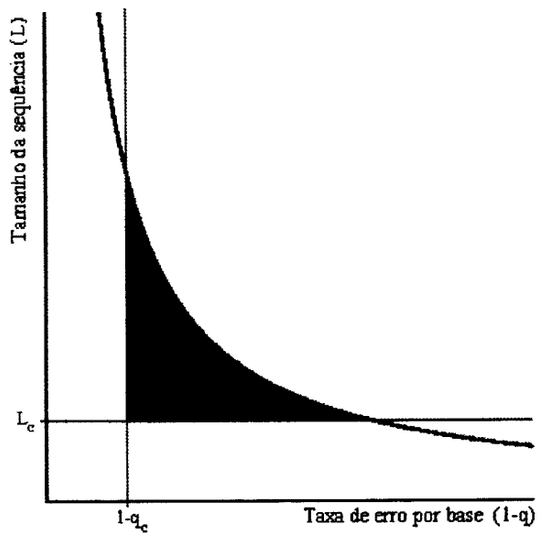

Figura 1.5: Tamanho de seqüência permitido L em função da taxa de mutação $\mu = (1-q)$ representado pela hipérbole construída para $\ln a = 1$, onde a é a vantagem seletiva da seqüência mestra sobre seus mutantes. Um sistema viável, portanto, deve ficar abaixo dessa hipérbole. Entretanto, há outros vínculos: o tamanho da seqüência tem que ser maior que algum limiar L_c para poder codificar informação suficiente e a taxa de erro não pode ser reduzida abaixo de um valor menor que $1 - q_c$ sem um custo excessivo de tempo e energia na replicação.

área destacada em vermelho na figura, representa a região permitida para que o sistema seja viável definida pela hipérbole do limiar de erro e pelos valores mínimos de L e $1 - q$ abaixo dos quais o sistema perde a propriedade de auto-replicação. Esta versão moderna do *paradoxo do ovo e da galinha* levou Eigen e Schuster [16] a proporem um modelo incorporando um mecanismo fechado de colaboração ou catálise entre as moléculas auto-replicantes, o qual chamaram de *hiperciclo*, evitando ou postergando assim a *catástrofe de erro*. A descrição desse modelo foge do contexto deste trabalho.

1.5 Análise crítica

Neste capítulo exploramos um sistema simples de moléculas auto-replicantes e descobrimos que, se o sistema for fechado (no sentido de que uma cópia errônea de uma molécula produz uma das outras espécies presentes) o resultado será uma coleção estável das várias espécies - uma *quase-espécie*. Contudo, o modelo tem muitos inconvenientes.

O modelo além de determinístico é contínuo e, de uma maneira geral, nenhuma dessas propriedades é particularmente interessante para um sistema tal como uma sopa molecular, que consiste de um número relativamente pequeno de moléculas (desprezado o substrato aquoso) em movimento aleatório. Um modelo mais realístico deve ter duas características principais: ser discreto e estocástico.

Para ilustrar o quão errado o modelo baseado em equações diferenciais pode ser, considere um sistema de N moléculas, cada qual de uma espécie diferente. Considere o caso ideal onde $A_i = D_i$ e $q^L = 1$, de maneira que as taxas de nascimentos e de mortes sejam iguais e não haja nenhum erro na reprodução. Nesse caso, todos os elementos da matriz de replicação W são nulos, e o modelo prediz que nada irá mudar no sistema (a previsão é correta para $N \to \infty$). Na realidade, como iremos ver, devemos esperar que um sistema com essas

características eventualmente seja dominado por uma das N espécies inicialmente presentes na população (*deriva genética*).

Ainda, reações químicas ordinárias envolvem *algumas* poucas espécies moleculares, cada uma das quais presentes em *número praticamente infinito* de cópias (da ordem do número de Avogrado 6.02×10^{23}). No caso da evolução acontece o contrário, o número de polinucleotídeos possíveis (e portanto reações) N_{es} é muito maior do que o número de moléculas presentes em qualquer experimento realístico que se queira realizar (por exemplo $2^{100} \approx 10^{30}$) e, portanto, a adoção do formalismo determinístico tem que ser considerada cuidadosamente.

Para produzir o *limiar de erro* como feito a partir do modelo de Eigen, é necessário que o relevo de replicação usado tenha uma forma extrema, ou seja, a relação (1.15) que caracteriza o *limiar de erro* depende fortemente do modelo adotado para o relevo de replicação, sendo que certos relevos não produzem o *limiar de erro* [17]. Assim, pode ser que esse conceito desempenhe um papel menos fundamental do que tem-lhe sido atribuído.

Outras forças evolucionárias como recombinação, por exemplo, também podem prevenir o limiar de erro em sistemas mais realísticos como os vírus. Além disso, não há uma aceitação universal da definição de *limiar de erro* para população e/ou comprimento de seqüências finitos. Várias análises têm caracterizado esse fenômeno como uma propriedade da população inteira enquanto que outras como uma propriedade de uma molécula individual (a mestra).

Nos próximos capítulos discutiremos essas e outras questões, entretanto, sem minimizar a importância dos resultados da teoria clássica do modelo de quase-espécies.

Por último, devemos frisar que existe uma correspondência entre as equações que definem o modelo de quase-espécies e as propriedades de equilíbrio de sistemas de redes de spins. Realmente, como já comentamos, um mapa entre as

equações (1.7) e modelos de redes de spin foi investigado recentemente [14] e uma solução analítica completa foi derivada para um relevo de replicação simples, a partir da transformação das equações cinéticas em um problema de localização de polímeros [15]. Dessa maneira, podemos associar os conceitos ligados ao problema de evolução biológica (como os de relevo replicativo e mutação) aos da física de sistemas de spin, (como relevo de energia e entropia) e ainda nos utilizarmos de várias técnicas oriundas da mecânica estatística. Esse tipo de abordagem esta bem documentada na literatura de uma maneira bastante didática [33] e, portanto, restringimos nossa análise aos aspectos biológicos do problema de evolução darwiniana.

Capítulo 2

O modelo de amostragens

A primeira experiência de evolução darwiniana foi realizada por Sol Spiegelman e colaboradores para a replicação de RNA viral [10], utilizando o método tradicional de passagens em série no qual a pressão de seleção era controlada durante o experimento. O passo radical desse experimento foi a possibilidade da replicação de RNA *in vitro*. Para isso foram necessários um RNA auto-replicador, nucleotídeos ativados e uma enzima do tipo replicase.

Inspirados nesse procedimento experimental, propomos e estudamos neste capítulo um modelo alternativo (e mais geral) ao de quase-espécies, cujas características principais e resultados descreveremos nas próximas seções. A motivação inicial, entretanto, é desenvolver um modelo capaz de descrever a dinâmica evolucionária de uma população finita de indivíduos ou moléculas. Particularmente, desenvolvido o modelo, comparamos os dois formalismos enfatizando as diferenças entre o reator de fluxo e o método de passagens em série. Essas diferenças, como veremos, são cruciais para os modos de modelar.

Finalmente, é importante destacar que para fazer essa comparação, restringiremos o estudo deste capítulo ao relevo de replicação de um pico. Isso, por sua vez, limita as possibilidades analíticas do modelo desenvolvido aqui, desde que é possível estender os resultados a relevos de replicação mais gerais de uma maneira bastante simples, como faremos no próximo capítulo.

2.1 Evolução *in vitro*

A teoria apresentada no capítulo anterior era baseada na idéia de que seleção e evolução no sentido darwiniano devem descrever qualquer processo de auto-organização molecular mantendo para isso as propriedades de um sistema vivo. Qualquer tipo de verificação experimental dessa teoria depende, portanto, de que o sistema eleito para tal fim possua os requisitos apontados, para que seja possível observar um processo de seleção e evolução.

A maioria dos sistemas auto-replicativos reais estão formados por moléculas de DNA que apresentam pouco polimorfismo estrutural devido aos diversos mecanismos de reparo que esses sistemas possuem para minimizar erros de cópia na reprodução. Para ser um bom modelo experimental de algo que tem de ser ao mesmo tempo objeto e sujeito do processo evolutivo, é necessário que o sistema seja encontrado na natureza com uma vasta variabilidade de estruturas, requisito esse encontrado no RNA de uma grande variedade de populações virais.

Os primeiros resultados importantes para evolução molecular foram descritos por Spiegelman [10] utilizando um tipo de vírus de RNA que infecta a bactéria *Escherichia coli* (um bacteriófago), o $Q\beta$. O tipo dominante (wild type) desse fago tem um RNA com mais de 4000 nucleotídeos e foi utilizado como *semente* nos experimentos. Depois de várias rodadas de replicação, foi observado que as moléculas de RNA sobreviventes eram formadas por somente algumas centenas de nucleotídeos. Para explicarmos esses resultados é importante primeiramente descrever a técnica experimental utilizada, que aparece esquematizada na figura (2.1). Este método de passagens consiste de uma série de tubos de ensaio, todos eles contendo monômeros ricos em energia em igual concentração e a enzima replicase que catalisa a replicação. Ao primeiro tubo coloca-se o RNA viral, e depois de um período determinado no qual deu-se a replicação do RNA, uma

alíquota muito pequena do conteúdo do primeiro tubo de ensaio é transferida ao segundo, e assim sucessivamente. Durante o período de incubação aparecem

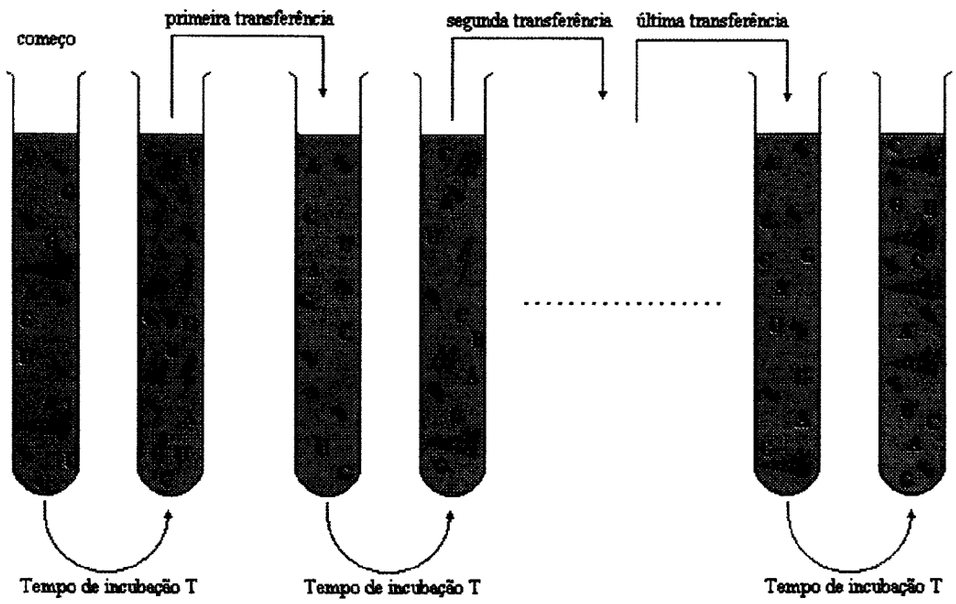

Figura 2.1: Desenho esquemático do método de passagens em série. O tubo inicial contém os quatro tipos de moléculas necessárias para a síntese do RNA, além da enzima replicase do fago $Q\beta$ e uma molécula de RNA utilizada como semente. Esta macromolécula é copiada muitas vezes com alguns erros. Depois de uma incubação por um tempo T, uma pequena amostra da solução (uma gota) é transferida a um novo tubo. Os processos de incubação e transferência são repetidos muitas vezes e as mudanças evolucionárias na população de RNA podem então ser observadas.

variantes de RNA devido aos erros de cópia na replicação, podendo-se observar então um processo de evolução. Além disso, o processo de passagens de um tubo a outro introduz no sistema uma forma sutil de *competição*, já que em cada operação de passagem os RNA de maior concentração, correspondendo àqueles com maior *fitness,* terão maior probabilidade de serem transferidos. O único critério nesse processo para definir o *fitness* das variantes produzidas é a velocidade de replicação e, desde que as moléculas menores replicam-se mais rápido, estas serão selecionadas mais vezes. A razão pela qual estas moléculas não di-

minuem em tamanho indefinidamente deve-se ao fato de que elas necessitam de um tamanho mínimo para manter a atividade de replicação como substratos das replicases na solução.

A prova mais evidente de que são produzidas cópias cada vez *mais adaptadas* à enzima é a observação de que a velocidade de síntese das novas cópias cresce rapidamente nas transferências sucessivas, indicando que cada vez mais aparecem variantes que são melhores substratos da enzima. De fato, originalmente nos experimentos de Spiegelman, foi necessário diminuir o tempo T entre transferências, conforme se avançava na série. Ao cabo de aproximadamente 70 transferências o sistema pareceu estabilizar-se, e não se observou variação aparente no fitness das moléculas produzidas. O RNA resultante deste processo de seleção e evolução tinha perdido toda a sua capacidade infectiva, enquanto aumentou muito sua capacidade replicativa, única pressão de seleção existente nestes experimentos. É interessante notar nesse ponto que simultaneamente ao desenvolvimento experimental com o fago $Q\beta$, porém inicialmente sem nenhuma referência a esses experimentos, implementou-se a teoria e o formalismo dos processos de seleção e evolução darwiniana para sistemas pré-bióticos, desenvolvidos por Eigen e Schuster, utilizando como sistema modelo o reator de fluxo.

Diferentemente daqueles autores, construímos um modelo de micro-evolução com base no procedimento experimental de amostragens utilizado por Spiegelman. Esse modelo, como argumentaremos, mostra-se mais simples e mais geral que o modelo original de quase-espécies. Para realizar isso, discutiremos na próxima seção os passos de um algoritmo que contém as principais características desse procedimento, mas que pode ser estendido a sistemas mais complexos e a condições mais interessantes na prática como a replicação *in vivo*.

2.2 Construção de um algoritmo para a estratégia experimental

O experimento descrito anteriormente ilustra bastante bem os principais aspectos da teoria darwiniana de seleção que vinhamos comentando: como uma pressão de seleção, as vezes sutil, pode dirigir inexoravelmente o processo evolutivo, e como a adaptação do sujeito da evolução aos condicionantes da pressão de seleção se realiza quase, poderíamos dizer, deterministicamente. Portanto, vamos utilizar este experimento como sistema modelo, fazendo-se primeiramente algumas simplificações necessárias, no sentido de não nos distrairmos com detalhes de menor importância.

Como no modelo original de quase-espécies, consideraremos populações com *tamanho constante* (N) e *tamanho de cada seqüência* também *constante* (L). Mais ainda, consideraremos que os caracteres herdáveis de cada indivíduo estão codificados em uma seqüência binária de 0's e 1's ($\kappa = 2$), composta de L símbolos. Cada seqüência é caracterizada pelo número de 1's que ela contém, sem levar em conta a posição desses monômeros dentro da *seqüência.* Como no caso do modelo cinético, isto faz com que existam $L+1$ tipos (ou classes) diferentes de *seqüências* possíveis (*espaço de seqüências*) os quais são identificados aqui pela variável inteira $P = 0, 1, 2, ..., L$. Esta consideração é razoável, desde que as características que distingüem os genomas são suas taxas de replicação que, em muitas análises, têm sido escolhidas de modo a depender de P somente [4, 9, 8]. Na figura (2.2) mostramos uma comparação entre o espaço de seqüências gerado por todas as possíveis mutações e o espaço de seqüências reduzido em classes de acordo com a distância de Hamming d e com o número de 1's que cada seqüência contém. No exemplo, o comprimento das moléculas é fixado em $L = 4$, o que nos leva aos 5 tipos de seqüências possíveis, mostrados como exemplo no capítulo anterior.

Finalmente, vamos supor que todos os indivíduos da população na geração t sejam substituídos por seus descendentes na geração posterior $t+1$. Esse pro-

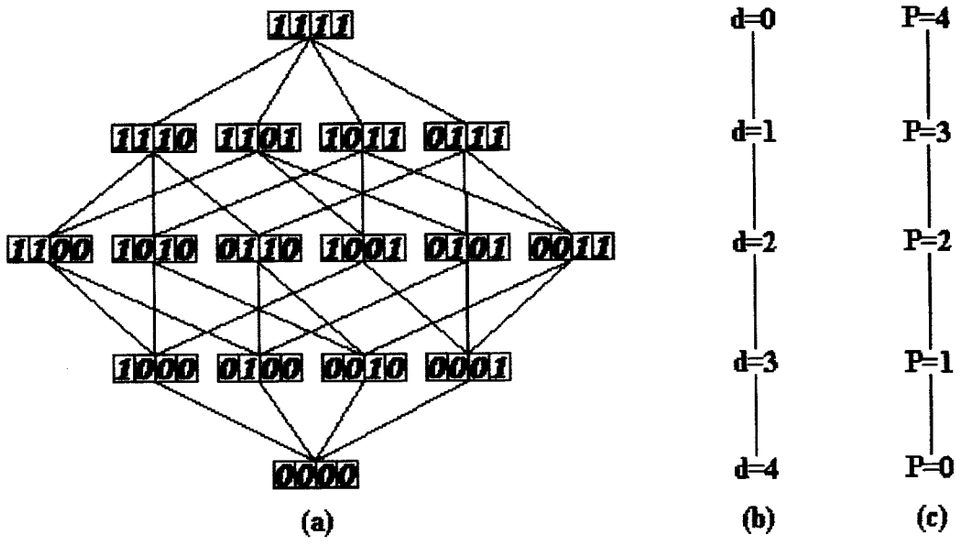

Figura 2.2: Comparação entre os espaços de seqüências. (a) Espaço gerado no modelo original de quase-espécies onde cada seqüência, das 2^4 possíveis, é obtida a partir de mutações puntuais das seqüências mais próximas a mestra {1, 1, 1, 1}. (b) Espaço reduzido onde cada conjunto de degenerescências da posição dos monômeros na seqüência é representado pela distância de Hamming à seqüência mestra (diferença do número de zeros em relação a mestra). (c) Mesmo que (b) para o número de 1's que cada seqüência contém.

cedimento faz com que o tempo seja uma variável discreta que mede o número de gerações (gerações sem superposição). A rigor, aqui uma geração é igual ao tempo de incubação do procedimento experimental e, portanto, engloba várias gerações do modelo original de tempo contínuo. Assim, o estado da população no tempo t pode ser descrito indicando-se o número de indivíduos ou *número de ocupação* para cada um dos $L+1$ pontos do *espaço de seqüências*, denotado por $n_P(t)$. Portanto, dado o estado da população na geração t como sendo o vetor $\mathbf{n}(t) = \{n_P(t)\}$ (de modo que $\sum_P n_P(t) = N$), o processo evolucionário ocorrendo nas próximas gerações pode ser pensado como um *processo estocástico* em três estágios, cujas regras dinâmicas são: *reprodução*, não propriamente pertencente

ao processo evolucionário, mas sua premissa; *seleção natural,* modelada pela taxa de replicação de cada indivíduo A_P; e *mutação*, modelada pela taxa de erro por dígito $\mu = 1 - q$.

Com essas simplificações, e estabelecida a dinâmica a que nosso sistema está sujeito, podemos *mimetizar* a estratégia experimental descrita anteriormente considerando uma *sopa de seqüências* contendo os $L+1$ diferentes tipos de seqüências nas proporções Π_P ($P = 0, 1, ..., L$), da qual retiramos com reposição N indivíduos aleatoriamente. Assim, o estado $\mathbf{n}(t) = (n_0, \ldots, n_L)$ da população está relacionado com as freqüências Π_P através da distribuição de probabilidade mutinomial

$$\mathcal{P}_\Pi(\mathbf{n}) = \frac{N!}{n_0!\, n_1! \ldots n_L!} \Pi_0^{n_0} \Pi_1^{n_1} \ldots \Pi_L^{n_L}. \tag{2.1}$$

Dessa maneira, em cada geração as seqüências dos descendentes são uma amostra da *sopa de seqüências* dos pais, similarmente ao experimento em que transferimos pequenas alíquotas entre os tubos de ensaio. Isto faz com que em cada geração a população seja uma combinação aleatória das seqüências constituintes, ou seja, postulamos um *equilíbrio de ligação* (ou *linkage*) a nível de população. Embora esse procedimento destrua as correlações entre as seqüências, ele não causa qualquer perda significativa da informação genética desde que, nos relevos de replicação estudados nesta tese, a aptidão de uma seqüência para produzir descendentes em um dado meio depende somente do número de 1's em sua representação e que, em média, não é afetado pelo procedimento. Além disso, cabe ressaltar que a reprodução como descrita acima (mais propriamente, um processo de amplificação), diferentemente dos modelos de evolução existentes na literatura, leva em consideração os efeitos de população finita e, como veremos, descreve bastante bem o fenômeno da *deriva genética*, já que, a cada amostragem, a freqüência dos novos indivíduos deve ser um pouco diferente da contida na *sopa de seqüências* original, o que levará à fixação de um dos indivíduos depois de

muitas gerações (ou passagens).

Podemos agora inserir nesse procedimento as mudanças na composição da população **n**. Seguindo a prescrição comumente usada na implementação de algoritmos genéticos [18], consideraremos primeiro o efeito da seleção natural e depois o efeito das mutações. A figura (2.3) mostra esquematicamente o algoritmo de amostragens que define uma passagem no experimento de Spiegelman. Observe-se que a inclusão de mecanismos evolucionários, como seleção e mutação,

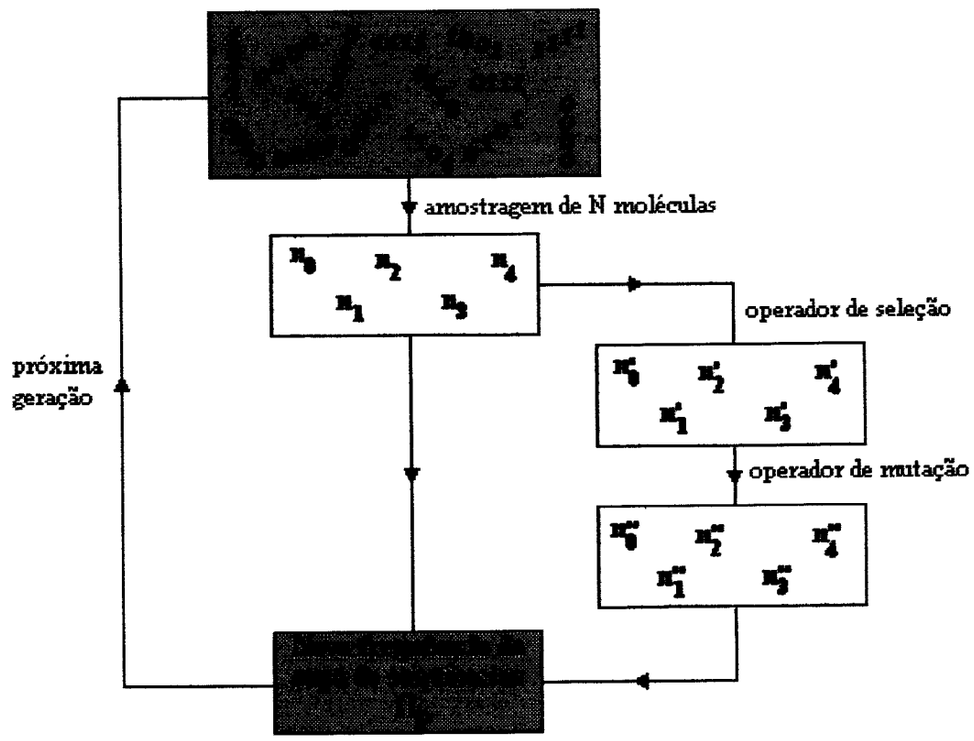

Figura 2.3: A sopa de seqüências inicial (em azul) contém 2^4 tipos diferentes de moléculas nas proporções Π_P (P=0, ..., 4). Depois da amostragem com reposição de N moléculas, cria-se um conjunto $\{n_P\}$ onde n_P é o número de moléculas com P 1's. A proporção de moléculas na sopa, na próxima geração, é diferente da inicial devido às flutuações introduzidas pela amostragem. Podemos inserir nesse procedimento mecanismos evolucionários, como seleção e mutação, que irão produzir mudanças no conjunto $\{n_P\}$.

é feita de maneira independente da amostragem; além disso, o algoritmo não depende do particular *agente* (a replicase, por exemplo) que define a característica

(ou *fenótipo*) particular da seqüência que será selecionada.

Após a amostragem dos N indivíduos, suporemos que o número de descendentes com que uma seqüência do tipo P contribui para a próxima geração seja proporcional a sua taxa de replicação relativa

$$W_P(\mathbf{n}) = \frac{n_P A_P}{\sum_R n_R A_R}. \tag{2.2}$$

Obviamente, os indivíduos com maior taxa de reprodução $W_P(\mathbf{n})$ terão maior probabilidade de aparecer nas próximas gerações, de maneira que a composição da população depois de selecionados os N novos indivíduos passa a ser descrita pelo vetor aleatório $\mathbf{n}'(t) = (n'_0, \ldots, n'_L)$, que por sua vez é distribuído de acordo com a distribuição de probabilidade condicional

$$\mathcal{P}_s(\mathbf{n}' \mid \mathbf{n}) = \frac{N!}{n'_0!\, n'_1! \ldots n'_L!} [W_0(\mathbf{n})]^{n'_0} [W_1(\mathbf{n})]^{n'_1} \ldots [W_L(\mathbf{n})]^{n'_L}. \tag{2.3}$$

Este procedimento de seleção não é trivial para populações finitas, desde que ele introduz correlações na população (essas correlações são responsáveis pela estrutura da árvore genealógica da população). Diferentemente da literatura ligada a algoritmos genéticos [18, 19] ou a genética de populações [20], que tenta manipular essas correlações, nós simplesmente as eliminamos, usando uma amostragem adicional depois do procedimento de seleção.

Podemos agora considerar as mudanças em $\mathbf{n}'$ devido às mutações, através das quais as seqüências herdadas por todos os indivíduos na população sofrem transições aleatórias, de maneira que *cada elemento* de uma *seqüência* é modificado com uma dada probabilidade μ, independentemente dos outros elementos (consideramos somente mutações pontuais). Com esse mecanismo simples, o estado da população depois da mutação é descrito pelo vetor $\mathbf{n}''(t) = (n''_0, \ldots, n''_L)$, cujas componentes são dadas por $n''_P = \sum_{R=0}^{L} n''_{PR}$, onde os inteiros n''_{PR} representam o número de seqüências do tipo R que mutaram para seqüências do tipo P

(claramente, $n'_R = \sum_P n''_{PR}$). É fácil mostrar (ver Apêndice A) que a probabilidade de mutação de uma seqüência do tipo R para uma seqüência do tipo P é dada por

$$M_{PR} = \sum_{Q=Q_l}^{Q_u} \binom{R}{Q} \binom{L-R}{P-Q} (1-\mu)^{L-P-R+2Q} \mu^{P+R-2Q}, \qquad (2.4)$$

onde $Q_l = \max(0, P+R-L)$ e $Q_u = \min(P, R)$. A população é descrita de uma maneira mais conveniente pelo conjunto $\{n''_{PR}\}$ do que pelo vetor $\mathbf{n}''$. De fato, com essa notação, dado n'_R a distribuição de probabilidade condicional de $\{n''_{PR}\}$ é mais uma vez multinomial,

$$\mathcal{P}_m(n''_{0R}, n''_{1R}, \ldots, n''_{LR} \mid n'_R) = \frac{n'_R!}{n''_{0R}!\, n''_{1R}! \ldots n''_{LR}!} M_{0R}^{n''_{0R}} M_{1R}^{n''_{1R}} \ldots M_{LR}^{n''_{LR}}, \qquad (2.5)$$

para $R = 0, \ldots, L$.

Portanto, nesse modelo a freqüência de moléculas do tipo P na próxima geração $\Pi_P(t+1)$, após o sistema ter sofrido as transições $\mathbf{n} \longrightarrow \mathbf{n}' \longrightarrow \mathbf{n}''$, é dada simplesmente por $(1/N)\sum_R n''_{PR}$. Esta freqüência é então usada como a nova freqüência da sopa de seqüências que será utilizada para gerar a nova população de N moléculas de comprimento L de acordo com a distribuição (2.1) (conforme esquematizado na figura (2.3)). Assim, durante o tempo de incubação do experimento, estão ocorrendo as transições que definem uma geração neste modelo. Este procedimento é então repetido a cada geração, simulando as passagens entre os tubos de ensaio.

2.3 Resultados analíticos

No modelo construído na seção anterior, cada transição de estado da população é caracterizada pela probabilidade de atingirmos esse estado e, portanto, para descrevermos analiticamente o modelo, devemos apresentar uma estatística do comportamento do sistema. Fazemos isso através do número de ocupação

médio $\overline{\mathbf{n}} = (\overline{n}_0, \overline{n}_1, ..., \overline{n}_L)$, calculando as médias de suas componentes em várias gerações após atingido o estado estacionário e para várias amostras da população.

Vamos a seguir derivar uma equação de recorrência para a freqüência média de moléculas do tipo P na população. Naturalmente, isso só será possível se fizermos algum tipo de aproximação na dinâmica do modelo. Primeiramente, dadas as freqüências Π_P para $P = 0, ..., L$ na geração t, vamos calcular o valor esperado do número de moléculas do tipo P na próxima geração. Essa grandeza é dada por

$$\overline{n}_P(t+1) = \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} \sum_{R} n''_{PR}(t) \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}_{\Pi}\left(\mathbf{n}\right), \qquad (2.6)$$

onde os somatórios levam em conta todas as possibilidades para os números de ocupação gerados pelas freqüências $\Pi_P(t)$ como também as subseqüentes transições devidas aos procedimentos de reprodução e mutação. Esse processo de média é, obviamente, análogo a média sobre diferentes *rodadas* das simulações ou sobre diferentes populações, exceto que aqui esta média é realizada a cada geração, enquanto que nas simulações do modelo de amostragens ela é efetuada somente depois que o estado estacionário foi atingido. Esse modo de tomar médias sobre todas as possíveis realizações do sistema estocástico a cada geração é semelhante a aproximação *annealed* da Mecânica Estatística de sistemas desordenados e, naturalmente não leva em conta as flutuações de n_P em diferentes populações. De fato, estuda-se a evolução de uma *população média* que pode ser totalmente diferente de qualquer uma das populações específicas cujas evoluções são simuladas.

Neste estágio já podemos efetuar alguns dos somatórios em (2.6). Em particular, usando as distribuições (2.3) e (2.5) obtemos (ver Apêndice B)

$$\overline{n}''_{PR}(t) = \sum_{\{n''_{PR}\}} n''_{PR} \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) = n'_R(t) M_{PR} \qquad (2.7)$$

e

$$\overline{n}'_{PR}(t) = \sum_{n'_R} n'_R \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) = N W_R\left(\mathbf{n}\right), \tag{2.8}$$

que permite-nos reescrever a equação (2.6) como

$$\overline{n}_P(t+1) = N \sum_{\mathbf{n}} \sum_{R} M_{PR} W_R\left(\mathbf{n}\right) \mathcal{P}_{\Pi}\left(\mathbf{n}\right). \tag{2.9}$$

Lembrando que a freqüência de moléculas do tipo P na geração $t+1$ é simplesmente $\overline{\Pi}_P(t) = \overline{n}_P(t)/N$, poderíamos obter uma relação de recorrência *fechada* se substituíssemos Π_P pela sua média em $\mathcal{P}_{\Pi}\left(\mathbf{n}\right)$. Essa substituição, que despreza as flutuações dessa freqüência, caracteriza a aproximação de *campo médio* utilizada por diversos autores [19, 20, 21]. Entretanto, essa aproximação é exata para o modelo de amostragens, uma vez que, conforme discutido na seção (2.2), a própria definição do modelo incorpora esse tipo de substituição e a conseqüente quebra de correlações. Daí, a equação de recorrência fundamental, na qual o restante dessa tese será baseado, é a seguinte

$$\overline{\Pi}_P(t+1) = \sum_{\mathbf{n}} \sum_{R} M_{PR} W_R\left(\mathbf{n}\right) \mathcal{P}_{\overline{\Pi}}\left(\mathbf{n}\right). \tag{2.10}$$

É fácil verificar a condição de normalização $\sum_P \overline{\Pi}_P(t+1) = 1$, uma vez que as distribuições (2.3) e (2.5) levam a $\sum_R W_R\left(\mathbf{n}\right) = 1$ e $\sum_P M_{PR} = 1\ \forall R$, respectivamente. É de fundamental importância que se enfatize o único tipo de aproximação utilizado na derivação de (2.10), a saber, a desconsideração das flutuações do número de ocupação n_P para diferentes populações a cada geração. Conforme mencionado, isso nos leva a estudar a evolução de uma população média, que pode não ser representativa das populações simuladas. De fato, como veremos na seção (2.6), essa aproximação falha quando a dinâmica estocástica for tal que a mesma população inicial possa alcançar dois estados estacionários completamente distintos.

Portanto, a equação de recorrência (2.10) descreve a dinâmica do modelo de amostragens de uma maneira geral, no sentido de levar em conta os efeitos de

uma população finita. Dessa maneira, ela é uma versão estocástica do modelo de quase-espécies, que mantém as principais características de um modelo micro-evolucionário. Nas próximas seções deste capítulo vamos iterá-la, restringindo-nos ao relevo de um pico, mostrando suas propriedades no equilíbrio, tanto no limite de populações infinitas como no de populações de tamanho finito, permitindo-nos assim estudar como a estrutura da população é alterada e como aparece o limiar de erro nesses limites.

Entretanto, antes de procurarmos soluções para a equação (2.10) vale a pena fazer uma breve observação sobre suas propriedades mais gerais que serão detalhadas no próximo capítulo. Para irmos mais adiante com nossa análise, devemos especificar a taxa de replicação A_P para cada indivíduo na população através de um relevo de replicação particular. Para fazermos isso no caso de um relevo arbitrário, podemos substituir a equação (2.2) em (2.10) obtendo

$$\overline{n}_P(t+1) = N \sum_{\mathbf{n}} \sum_{R} \left(\frac{n_R A_R}{\sum_K n_K A_K} \right) M_{PR} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \,. \tag{2.11}$$

Como pode ser notado desta equação, o *fitness* médio da população, definido como

$$\overline{W} = \sum_{K=0}^{L} n_K A_K \,, \tag{2.12}$$

depende de $\mathbf{n}$ o que dificulta o cálculo no somatório em $\mathbf{n}$ na equação em (2.11) para um relevo de replicação arbitrário definido pelo conjunto de valores A_P. Diante disso, lançamos mão de um truque matemático simples de maneira a contornar tal situação, que com efeito será utilizado com sucesso no próximo capítulo.

Neste capítulo a proposta é nos mantermos dentro dos parâmetros mais utilizados na literatura especializada no modelo de quase-espécies e, portanto, especificarmos na equação (2.11) os dados relativos ao relevo de replicação de um

pico, isto é, atribuirmos $A_L = a$ e $A_P = 1$, para $P \neq L$. Isso nos leva a equação

$$\overline{n}_P(t+1) = N \sum_{\mathbf{n}} \left(\frac{\sum\limits_{R \neq L} n_R M_{PR} + n_L a M_{PL}}{N + (a-1) n_L} \right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) , \tag{2.13}$$

onde lembremos que a dependência em t no lado direito desta equação está contida em $\overline{\Pi}$. Notemos que agora o *fitness* médio da população a cada geração

$$\overline{W} = N + (a-1) n_L \tag{2.14}$$

depende somente da variável n_L, que é o número de seqüências com L 1's (mestras), o que nos permite efetuar facilmente as somatórias sobre as variáveis $n_0, n_1, ..., n_{L-1}$ (os detalhes dos cálculos são apresentados no Apêndice C). O resultado final é a equação de recorrência para as freqüências de seqüências do tipo P,

$$\overline{\Pi}_P(t+1) = M_{PL} \left[\overline{\Pi}_L(t) \right]^N + \sum_{n_L=0}^{N-1} B_{n_L} \frac{\sum_{R=0}^{L-1} \overline{\Pi}_R(t) \left[M_{PR} + a \frac{r}{1-r} M_{PL} \right]}{1 + r\,(a-1)} \tag{2.15}$$

para $P = 0, \ldots, L$. Aqui introduzimos a notação

$$B_{n_L} = \binom{N-1}{n_L} \left[\overline{\Pi}_L(t) \right]^{n_L} \left[1 - \overline{\Pi}_L(t) \right]^{N-1-n_L} , \tag{2.16}$$

e $r = n_L/N$. Então, dada a freqüência molecular inicial média $\overline{\Pi}_P(t=0)$ para $P = 0, ..., L$, a equação (2.15) é iterada até o regime estacionário ser alcançado.

2.4 Comparação com a formulação de Eigen

Para melhor apreciarmos a relevância desta formulação, vamos comparar a equação (2.15) no regime determinístico ($N \to \infty$) com a solução exata da equação cinética (1.12) para L finito. Nesse caso, a soma sobre n_L na equação (2.15) é dominada pelo inteiro mais próximo de $(N-1)\overline{\Pi}_L(t)$, de maneira que nesse regime $r \to \overline{\Pi}_L(t)$, e a equação de recorrência reduz-se a (ver Apêndice C para detalhes)

$$\overline{\Pi}_P(t+1) = \frac{\sum\limits_{R=0}^{L-1} M_{PR} \overline{\Pi}_R(t) + a M_{PL} \overline{\Pi}_L(t)}{1 + (a-1)\,\overline{\Pi}_L(t)} . \tag{2.17}$$

Esta equação apresenta o mesmo estado estacionário da equação (1.12) do modelo original de quase-espécies [9] quando nos restringimos ao relevo de replicação de um pico.

Assim, a equação (2.15) (baseada no *algoritmo de amostragens*) sintetiza um processo evolucionário que no regime determinístico reproduz a mesma fenomenologia observada na formulação de Eigen da teoria de quase-espécies. É interessante notar que isso indica que no regime estacionário do modelo de quase-espécies não há nenhum *desequilíbrio de ligação* ao nível da população, isto é, a população é uma combinação aleatória das moléculas constituintes. De fato, esse resultado é válido para qualquer escolha de relevo de replicação A_P, como veremos no próximo capítulo com maior detalhe. Todavia, para o regime determinístico, poderíamos nos adiantar e verificar isso facilmente tomando o limite $N \to \infty$ na equação (2.10).

Para ilustrar como o *modelo de amostragens* mantém as principais características do modelo de quase-espécies, iteramos a equação de recorrência determinística (2.17) até alcançar o estado estacionário e apresentamos os resultados na figura (2.4). Essa figura ilustra como a distribuição de quase-espécies modifica-se à medida que mudamos a taxa de mutação. É mostrada a freqüência de seqüências em função da distância de Hamming em relação a mestra ($d = 0$ ou $P = L$) para vários valores da taxa de mutação. Em particular, na parte (a) da figura apresentamos essa distribuição para $\mu = 0$, para ser tomada como referência para outros valores de μ. A partir deste valor, à medida que a taxa de mutação aumenta, a largura da distribuição também aumenta até um valor máximo (comparar com os resultados da figura (1.4) para cada valor de μ apresentados aqui). Nas partes (e) e (f) da figura, essa largura volta a diminuir e permanece aproximadamente constante. Nesses dois casos, a distribuição é aproximadamente a binomial $\Pi_d = \binom{L}{d}\frac{1}{2^L}$, indicando que todas as seqüências (não as

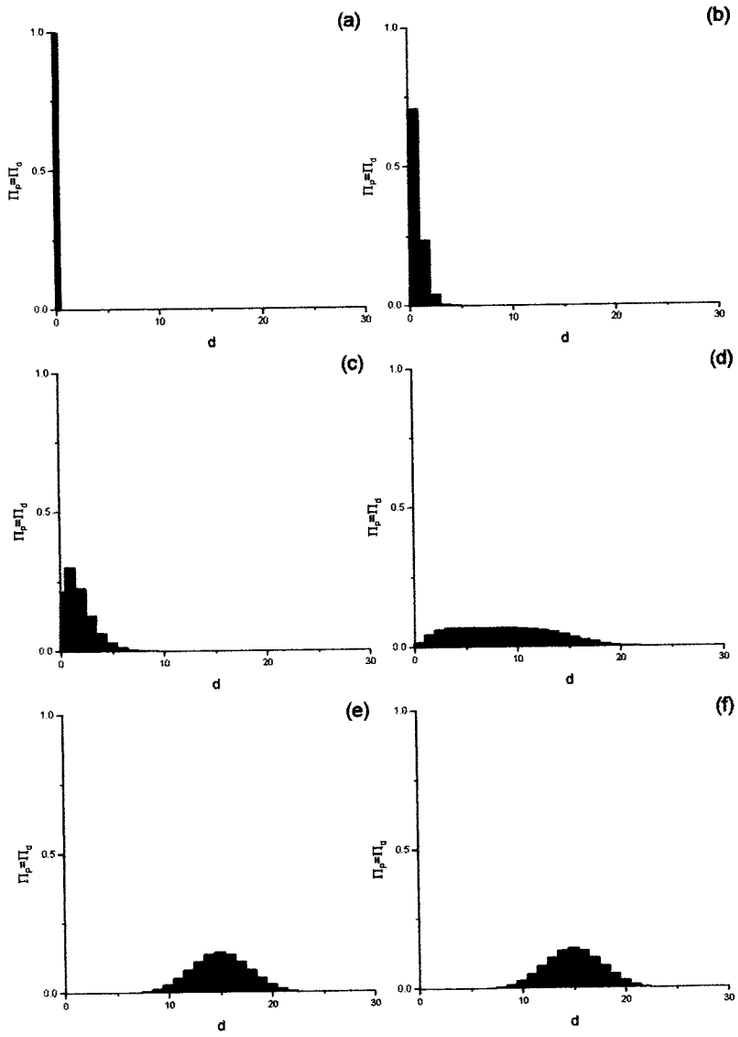

Figura 2.4: Distribuição das freqüências de seqüências $\Pi_P = \Pi_d$ (Pé o número de 1's e dé o número de zeros) em função da distância de Hamming de cada classe de seqüências à mestra no equilíbrio de seleção. (a) Destacamos a distribuição em que $\mu = 0$, para ser tomada como referência; (b) $\mu = 0.01$; (c) $\mu = 0.04$; (d) $\mu = 0.07$; (e) $\mu = 0.10$; e (f) $\mu = 0.13$. Note que em $\mu = 0.07$ o sistema está bem perto, porém abaixo do limiar de erro, e em $\mu = 0.10$ o sistema já ultrapassou esse valor crítico. É interessante comparar esse resultado com a figura (1.4), feita para ilustrar o modelo de quase-espécies original, e notar que ambas figuras reproduzem o mesmo estado estacionário. Aqui fizemos $a = 10$ e $L = 30$ na equação (2.16) e também tomamos as freqüências iniciais como $\Pi_L(0) = 1$ e $\Pi_P(0) = 0$; $P = 0, ..., L-1$.

classes) têm probabilidades idênticas $(1/2^L)$ de serem encontradas na população.

Vamos agora discutir o problema da localização do limiar de erro. Primeiramente, vamos relembrar o resultado obtido do modelo cinético, onde o *limiar de erro* é caracterizado pela equação (1.16), que pode ser facilmente reescrita como

$$-\ln q_c = \frac{1}{L}\ln a, \tag{2.18}$$

sem a necessidade da suposição $q_c \approx 1$. Esta definição, correntemente aceita para o *limiar de erro*, nos dá a taxa de erro na qual a freqüência da seqüência mestra Π_L anula-se [4, 8]. O problema com essa definição é que, mesmo no regime determinístico $N \to \infty$, Π_L nunca vai a zero para L *finito*. De fato, conforme mencionado acima temos $\Pi_L = 1/2^L$ no regime uniforme. O desaparecimento da seqüência mestra da população é um resultado da desconsideração das mutações reversas [4], que pode ser justificado somente no limite $L \to \infty$. Devemos enfatizar, portanto, que a equação (2.18) é somente uma aproximação para valores finitos de L.

Existe uma definição alternativa para o limiar de erro [28], que pode ser utilizada tanto no caso de populações finitas como infinitas. Esta caracterização do *limiar de erro* é obtida considerando-se as propriedades estatísticas de toda a população de moléculas. Particularmente, para lidar com seqüências de tamanho finito, a análise é feita considerando-se o efeito da taxa de mutação sobre a *distância de Hamming média* normalizada de toda a população no regime estacionário, definida por

$$\overline{d} = \frac{1}{L}\sum_{P=0}^{L}(L-P)\,\overline{\Pi}_P. \tag{2.19}$$

Essa grandeza mede a distância de Hamming média da seqüência mestra à toda a população, ou seja, o número médio de monômeros do tipo 0 na população. Com essa grandeza, a localização do limiar de erro é determinada pela taxa de erro na qual a variância ou a dispersão de d em torno de $\overline{d}$ é máxima [28]. Assim,

conhecidas as freqüências $\overline{\Pi}_P$ no equilíbrio em função de μ procuramos o valor $\mu_c = 1 - q_c$ que maximiza

$$\sigma_d^2 = \frac{1}{L^2}\left[\sum_{P=0}^{L}(L-P)^2\,\overline{\Pi}_P - \left(\sum_{P=0}^{L}(L-P)\,\overline{\Pi}_P\right)^2\right]. \tag{2.20}$$

A idéia aqui é simples: essa grandeza mede o quão espalhada a população está em relação à sequência mestra sem levar em conta o efeito das flutuações em Π_P, que é substituída pela média $\overline{\Pi}_P$ (isso é irrelevante no caso determinístico para o qual $\Pi_P = \overline{\Pi}_P$). Essa dispersão, por sua vez, aumenta à medida que aumentamos a taxa de mutação, atinge um valor máximo e então diminui até alcançar o valor $1/4L$ em $\mu = 0.5$. O máximo de σ_d^2 define a linha de transição entre o regime da quase-espécie e o uniforme.

Daí, a localização do limiar de erro calculada dessa maneira pode ser confrontada com o resultado da relação (2.18). Realmente, na figura (2.5) apresentamos o logaritmo da precisão da replicação no *limiar de erro* (q_c) como função do logaritmo da vantagem seletiva (a) para vários valores de L, no caso determinístico. Como pode ser observado, há uma excelente concordância entre os resultados das duas definições de q_c para valores pequenos de $\ln a$. Naturalmente, quando a for da ordem de 2^L (isto é, $\ln a = L \ln 2$) a equação (2.18) perde seu significado: se $\Pi_L = 1/2^L$ em uma certa geração então Π_L será de ordem 1 na próxima geração, de modo que a aproximação $1/2^L \approx 0$ falha completamente neste caso.

Apesar da concordância para $\ln a$ pequeno, como já havíamos comentado antes, não existe uma definição geral para o *limiar de erro*. Na próxima seção lidaremos com a arbitrariedade dessa definição para podermos explorar a possibilidade de definir o *limiar de erro* para uma população finita. Devemos frisar, no entanto, que o máximo de σ_d^2 diverge nos limites $L \to \infty$ e $q \to 1$ (tomados de forma que $q^L \equiv Q$ é finito) como $(Q_L^c - 1/a)^{-2}$, onde Q_L^c dá a posição do pico de

σ_d^2 para L finito. Do ponto de vista físico, esse resultado motiva a utilização de

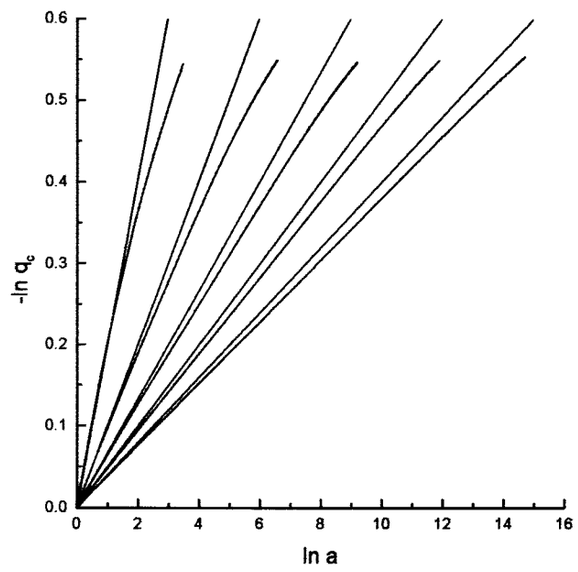

Figura 2.5: Precisão de replicação no limiar de erro no regime determinístico em função da vantagem seletiva. As curvas em vermelho são obtidas usando a equação (2.18) e as em azul a partir da maximização de (2.20). De cima para baixo essas linhas representam: L =5, 10, 15, 20 e 25.

$Q_L^c = q_c^L$ como o limiar de erro para L finito. A discussão detalhada dos efeitos de comprimento finito na transição do limiar de erro é apresentada na referência [28].

Finalmente, é importante lembrar que o modelo desenvolvido neste capítulo é baseado em um sistema modelo (o método de transferência em série) diferente do adotado para derivar o modelo original de quase-espécies (o reator de fluxo). A diferença principal reside na forma de implementar a diluição: descontínua no experimento com transferências em série e contínua no reator de fluxo. Esta diferença é que faz com que o modelo descrito aqui seja discreto e estocástico como desejável para um modelo mais realístico, no sentido de que leva em conta os efeitos de população finita. Mesmo assim, é surprendente encontrarmos na lite-

ratura uma elaboração teórica de um aparato chamado de *máquina de evolução*, por Y. Husimi e colaboradores [26], que consiste do método de passagens em série, sendo *continuamente* amostrado através de um reator de fluxo. A característica relevante para esses autores era a de ter um controle exato da população para poder lidar com concentrações de macro-moléculas e assim evitar flutuações na população.

Naturalmente, muitos trabalhos teóricos têm tentado generalizar a formulação determinística do modelo de quase-espécies de modo a levar em conta os efeitos de N finito. A primeira proposta de uma teoria estocástica de quase-espécies foi feita por Ebeling e Feistel [27], que todavia é válida somente para relevos de replicação suaves. Uma outra formulação [11], empregando um processo de nascimento e morte simples, não consegue reproduzir a distribuição do estado estacionário no regime determinístico. Estas e outras formulações serão discutidas e comparadas com o modelo de amostragens nos próximos capítulos.

2.5 O limiar de erro em populações finitas

Na seção anterior mostramos que o formalismo baseado no algoritmo de amostragens descreve exatamente as propriedades de equilíbrio da formulação determinística original da teoria de quase-espécies, quando tomamos o limite $N \to \infty$ em (2.15). Para caracterizarmos o *limiar de erro* em populações finitas através do formalismo de amostragem, voltamos a equação (2.15) que, como comentamos, é a versão estocástica proposta para o modelo de quase-espécies. Dessa forma, dada a freqüência molecular média inicial $\overline{\Pi}_P(t=0)$, a equação (2.15) é iterada até o regime estacionário ser alcançado. Feito isso, podemos calcular a distância de Hamming média da população à mestra bem como o seu desvio padrão através de (2.19) e (2.20), respectivamente.

Na figura (2.6) (a) comparamos os resultados das simulações usando o algo-

ritmo de amostragem com a predição teórica para dependência de $\overline{d}$ com a taxa de mutação $\mu = 1 - q$, dada pela solução no estado estacionário da equação (2.15). Acrescentamos, a título de comparação, os resultados da simulação de um algoritmo genético simples [18], que não utiliza o procedimento de quebra de correlação das freqüências a cada geração. Em ambos os casos, a população inicial foi gerada com $\Pi_L = 1$ e $\Pi_P = 0$ para $P \neq L$ e deixamos os algoritmos evoluírem por 2×10^3 gerações. Não notamos nenhuma diferença significativa quando permitimos que o sistema evoluísse por mais gerações ou para escolhas diferentes das freqüências moleculares iniciais. De fato, mesmo para populações idênticas na geração inicial, o caráter aleatório das transições $\mathbf{n} \rightarrow \mathbf{n}' \rightarrow \mathbf{n}''$ as torna distintas já na próxima geração. Cada ponto nas figuras envolve dois tipos de médias: para cada simulação tomamos a média sobre $\overline{d}$ e σ_d em 100 gerações depois que o equilíbrio de seleção foi alcançado; esses valores são então mediados sobre 200 simulações.

Como esperado, os efeitos das flutuações em d para diferentes simulações são acentuados para valores pequenos de N e, portanto, nossa aproximação analítica não produz bons resultados, embora ela reproduza bastante bem o padrão de comportamento qualitativo dessas grandezas. Contudo, já para $N = 100$ há uma excelente concordância entre os resultados teóricos e as simulações, desde que a taxa de mutação $\mu = (1 - q)$ não esteja muito próxima do limiar de erro. É surpreendente que as simulações usando o algoritmo genético concordem melhor com os resultados analíticos.

Na figura (2.6) (b) mostramos mais uma vez uma comparação entre as simulações e os resultados analíticos, agora para a dependência do desvio padrão médio, obtido tomando-se a raíz quadrada de (2.20), com a taxa de mutação $\mu = (1 - q)$. Aqui, como para a distância de Hamming média mostrada na parte (a) da figura, iteramos (2.15) até o equilíbrio ser alcançado. Neste caso, a predição teórica nos

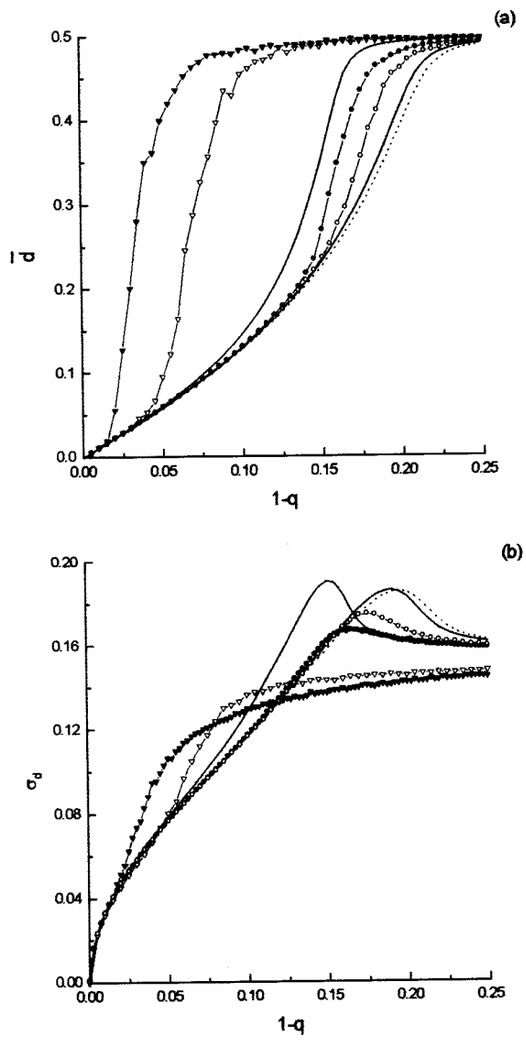

Figura 2.6: Comparação entre os resultados analíticos (curvas sólidas) e as simulações usando dois tipos de algoritmos: o algoritmo de amostragem que usa a hipótese de equilíbrio de ligação (símbolos sólidos) e o algoritmo genético (símbolos vazios). As figuras mostram, no estado estacionário, (a) a distância de Hamming média normalizada $\overline{d}$ entre a seqüência mestra e toda a população e (b) o desvio padrão σ_d de $\overline{d}$. Ambos resultados são apresentados em função da taxa de erro por dígito $\mu = 1 - q$. As curvas em vermelho são para o caso $N = 10$ e as em azul para o caso $N = 100$. A linha pontilhada em verde é a predição para $N \to \infty$. Os parâmetros restantes são $L = 10$ e $a = 10$.

fornece a informação de que o efeito do tamanho da população sobre o limiar de erro, medido pela posição do máximo de σ_d, é o de deslocá-lo para valores cada vez menores à medida que N diminui. Em particular, nota-se o aumento abrupto de σ_d à medida que nos aproximamos do limiar de erro a partir de valores pequenos de μ, e um decaimento mais suave depois que utrapassamos a região de máximo. Todavia, como esperado, a nossa predição teórica para o desvio padrão é bastante pobre para valores de N pequenos, desde que nosso esquema de aproximação desconsidera as flutuações em Π_P entre diferentes rodadas. A concordância entre teoria e simulações torna-se melhor à medida que N aumenta mas, mesmo para $N = 10$, a aproximação de campo médio dá resultados excelentes para μ pequeno.

O que é notável e surpreendente do discutido até aqui, é a completa falta de correlação preditiva entre os resultados analíticos e as simulações do modelo para o desvio padrão no caso $N = 10$, apesar da concordância qualitativa entre teoria e simulações para a distância de Hamming média, mostrada na parte (a) da figura. Como pode ser observado, para $N = 10$ as simulações não produzem um máximo para σ_d em valor algum de μ, enquanto que o a aproximação analítica localiza o limiar de erro, o máximo do desvio padrão, em torno de $\mu \approx 0.15$. Esta aparente falha de nossos resultados esconde, na verdade, um outro fenômeno, associado ao tamanho da população e que *compete* com o limiar de erro para destruir a estrutura da população, conforme discutiremos na próxima seção.

Para continuarmos com esta análise, entretanto, é importante enfatizar o significado das grandezas utilizadas nesta seção para caracterizar o limiar de erro. A medida de dispersão σ_d, definida pela equação (2.20), é uma medida do espalhamento da população no espaço de seqüências a partir da mestra que, quando tomada em função da taxa de mutação, nos permite calcular a localização do limiar de erro através da dispersão máxima. Isso pode ser feito parametricamente, tanto no caso em que a população é infinita e desejamos caracterizar o limiar de

erro para tamanhos de seqüência finitos [28], como no caso desenvolvido nesta seção em que fixado o tamanho da seqüência estamos interessados em calcular esse limiar para diferentes tamanhos da população [9]. Todavia, essa medida de dispersão não leva totalmente em conta as flutuações em Π_P, uma vez que em sua definição aparece apenas o valor médio $\overline{\Pi}_P$. Para lidarmos com as flutuações em Π_P, somos levados a definir uma nova dispersão:

$$\begin{aligned}\sigma_N^2 &= \frac{1}{L^2}\left[\overline{\left(\sum_{P=0}^{L} P\Pi_P\right)^2} - \overline{\left(\sum_{P=0}^{L} P\Pi_P\right)}^2\right] \\ &= \frac{1}{L^2}\sum_{P,Q} PQ\left(\overline{\Pi_P\Pi_Q} - \overline{\Pi}_P\overline{\Pi}_Q\right). \end{aligned} \tag{2.21}$$

Naturalmente, no limite determinístico Π_P deixa de ser uma variável aleatória e, portanto, as covariâncias em (2.21) vão a zero levando a $\sigma_N^2 = 0$. Assim, embora essa grandeza não seja útil para a determinação do limiar de erro, pois não leva aos resultados do limite determinístico, ela é de grande interesse já que mede verdadeiramente a dispersão da variável aleatória d em torno do valor médio $\overline{d}$ mostrado na figura (2.6) (a). Dentro do formalismo de amostragens, σ_N^2 pode ser calculada tomando-se o cuidado de definir corretamente o segundo momento de n_P na distribuição de probabilidades para o número de ocupação $\mathbf{n}$. A questão de natureza técnica é como calcular $\overline{\Pi_P^2} = \overline{n_P^2}/N^2$ e $\overline{\Pi_P\Pi_Q} = \overline{n_P n_Q}/N^2$ no regime de equilíbrio de seleção. Ora, definido o valor médio $\overline{n}_P$ através de (2.6), segue similarmente que,

$$\overline{n^2}_P(t+1) = \sum_{\mathbf{n}}\sum_{\mathbf{n}'}\sum_{\{n''_{PR}\}}\left(\sum_R n''_{PR}(t)\right)^2 \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right)\mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right)\mathcal{P}_\Pi\left(\mathbf{n}\right), \tag{2.22}$$

a qual simplifica em (ver Apêndice B para detalhes)

$$\overline{n^2}_P = \overline{n}_P + N\left(N-1\right)\left[\sum_{\mathbf{n}}\left(\sum_R M_{PR}W_R\right)^2 \mathcal{P}_\Pi\left(\mathbf{n}\right)\right], \tag{2.23}$$

onde fizemos $\overline{n_P^2}(t+1) = \overline{n_P^2}(t) = \overline{n_P^2}$, já que estamos interessados no regime estacionário. Procedimento análogo leva a equações similares para as correlações

$\overline{\Pi_P \Pi_Q}$. O segundo momento $\overline{\Pi_P^2}$ será de particular importância na próxima seção pois desempenha um papel fundamental na determinação de limites que caracterizam os efeitos estocásticos da finitude de N na estrutura da população. A figura

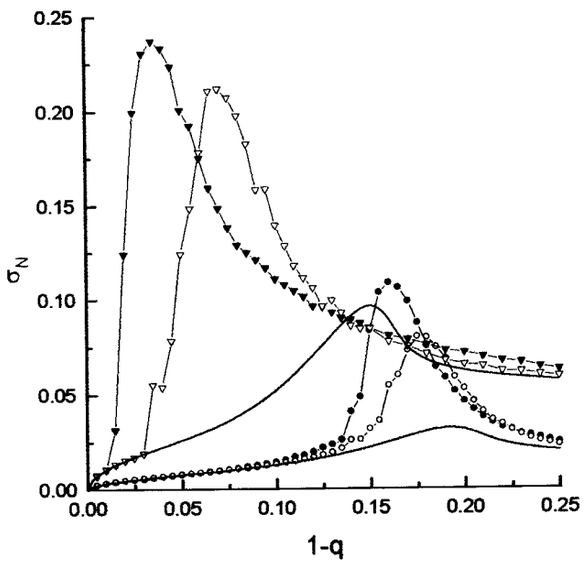

Figura 2.7: Comparação entre resultados analíticos (curvas sólidas) e as simulações usando os dois tipos de algoritmos: o algoritmo de amostragem (símbolos sólidos) e o algoritmo genético (símbolos vazios), para o desvio padrão calculado a partir da equação (2.19). Aqui foram mantidos os mesmos parâmetros e a mesma notação de cores da figura (2.6) (b).

(2.7) mostra o desvio padrão definido em (2.21) comparado com os resultados das simulações utilizando o algoritmo de amostragens e o algoritmo genético. Os resultados analíticos descrevem qualitativamente bem o padrão observado nas simulações desde que não estejamos muito perto do limiar de erro. É interessante observar que a posição do máximo de σ_N^2 permanece a mesma com relação à obtida na figura (2.6) (b), apesar de, como já mencionado, σ_N não ser apropriada para caracterizar o limiar de erro, já que ela vai a zero para $N \to \infty$. Mais uma vez, quanto menor a população, maior é a discrepância dos resultados.

De uma maneira geral, o efeito da aproximação analítica é superestimar o limiar de erro, acarretando um deslocamento da posição do máximo da variância para valores maiores de $\mu = (1-q)$. A questão que permanece é por que as simulações para $N = 10$ não produzem um máximo quando estamos considerando a variância σ_d^2 e este reaparece quando levamos em conta as flutuações nos números de ocupação médios, σ_N^2. O formalismo desenvolvido até aqui não consegue explicar esse fato, todavia, vamos argumentar a seguir que essa aparente contradição reside em um efeito, denominado *escape estocástico,* no qual as flutuações devido a finitude da população dominam todas as outras pressões evolucionárias.

2.6 Deriva genética e escape estocástico

A *deriva genética* em uma população é entendida como a fixação de algum indivíduo em particular em detrimento dos outros da população, devido às flutuações causadas por *amostragens* naturais como, por exemplo, ocorre no processo de evolução *in vivo* de vírus e bactérias à medida que são transmitidos entre suas células hospedeiras. Temos argumentado que o formalismo desenvolvido nas seções precedentes caracteriza-se por levar em conta este fenômeno através da amostragem inicial, representada pela distribuição multinomial (2.1), como também através das flutuações inerentes ao processo de seleção e de mutação. Todavia, se quisermos manter a descrição de uma população evoluindo em um relevo de replicação de um pico, devemos considerar que essas flutuações poderão destruir a estrutura da população, mesmo que tomemos como população inicial o caso mais otimista em que todas as seqüências são as mais aptas do relevo. Esse é o caso, por exemplo, da figura (2.6) (b) para uma população fixa de $N = 10$ seqüências em que as equações do modelo previam uma dispersão máxima da população em torno da seqüência mestra, enquanto que nos *experimentos* descritos pelas simulações essa dispersão aumentou até ficar assintoticamente constante,

sem atingir um máximo, apesar da distância de Hamming média ser qualitativamente a mesma em ambos os casos (ver figura (2.6) (a)). Esse resultado não era esperado e, com efeito, pudemos recuperar o pico do desvio padrão quando correlacionamos a distância de Hamming em cada rodada de replicação (ver figura 2.7) e conjecturamos um efeito do tamanho da população que não era descrito pela aproximação analítica.

Para tirarmos conclusões dessa discussão, devemos ser bastante cuidadosos quando lidarmos com o relevo de replicação de um pico em populações finitas. A questão que se coloca no caso desse relevo é o quão freqüente é a seqüência mestra quando amostramos uma população, já que a sua ausência em uma amostra leva a um caso de evolução neutra no qual todos os indivíduos têm a mesma capacidade de se reproduzir. Intuitivamente, esta freqüência deve ser menor à medida que amostramos um número menor de seqüências, de maneira que, em média, poderíamos ter diferenças entre as simulações e as predições teóricas do modelo.

Para ilustrar melhor essa situação, vamos discutir um outro fenômeno interessante, denominado *escape estocástico*, recentemente introduzido na literatura [9, 20] e que é característico de populações finitas. A idéia geral é determinar qual é a probabilidade de que uma seqüência em uma população seja perdida devido às flutuações inerentes aos processos estocásticos (amostragem, reprodução e mutação) que governam a evolução da população. Esse conceito é uma conseqüência de introduzirmos a deriva genética no problema.

Em particular, no caso do relevo de replicação que estamos discutindo neste capítulo, podemos estudar esse fenômeno estimando a probabilidade de que a seqüência mestra, e mesmo alguns mutantes, sejam perdidos devido à flutuações na população. Desde já pode-se adiantar que no limite $L \to \infty$ a perda da seqüência mestra é irreversível, uma vez que nenhuma mutação reversa conseguirá

restabelecer essa seqüência.

Para caracterizar o escape estocástico para tamanhos de seqüências finitos podemos facilmente derivar um limitante inferior para a probabilidade de que a seqüência mestra esteja ausente da população. Ora, desde que o número de seqüências na população com L 1's obedece a desigualdade $n_L \geq 0$, segue que

$$\overline{n}_L = \sum_{n_L=1}^{N} n_L \Pr(n_L) \geq \sum_{n_L=1}^{N} \Pr(n_L) = 1 - \Pr(n_L = 0). \tag{2.24}$$

Isso nos leva a desigualdade

$$1 - \overline{n}_L \leq \Pr(n_L = 0). \tag{2.25}$$

Com essa relação e usando $\overline{n}_L = N\overline{\Pi}_L$, podemos encontrar um valor para a precisão na replicação q_{inf} tal que a condição $\overline{\Pi}_L = 1/N$ seja satisfeita para L e a fixos. Daí podemos garantir que para $q < q_{\text{inf}}$ a probabilidade de que a seqüência mestra esteja ausente da população é não nula. Assim, podemos determinar o valor de q para o qual a equação (2.15) iguala a $1/N$ no equilíbrio de seleção.

Na figura (2.8) apresentamos esse limitante inferior q_{inf} (curvas em vermelho) juntamente com a precisão de replicação no limiar de erro q_c (curvas em azul) em função do inverso do tamanho da população para vários valores de L e $a = 10$. Adiantamos que no limite $N \to \infty$ temos $q_{\text{inf}} \to 0$ para L finito, desde que no regime determinístico $\overline{\Pi}_L$ é limitado por $1/2^L$ e, portanto, a seqüência mestra está sempre presente na população. Os valores de q_c foram obtidos calculando-se a posição do máximo de (2.20) para cada valor de N. É interessante observar dessa figura que para cada L fixo existe um valor de N em que $q_{\text{inf}} = q_c$. Isso indica que, para valores de N menores que esse valor, a precisão de replicação para que o escape estocástico da mestra ocorra é maior do que a precisão exigida no limiar de erro. A conseqüência disso é que para valores de taxa de mutação menores que o limiar de erro $\mu < \mu_c$ existe uma probabilidade não nula de que

a seqüência mestra esteja ausente da população ou, posto de outra maneira, é provável que a seqüência mestra esteja ausente da quase-espécie para taxas de mutação no intervalo $\mu_{\text{inf}} < \mu < \mu_c$. No caso do relevo de replicação *abstrato*

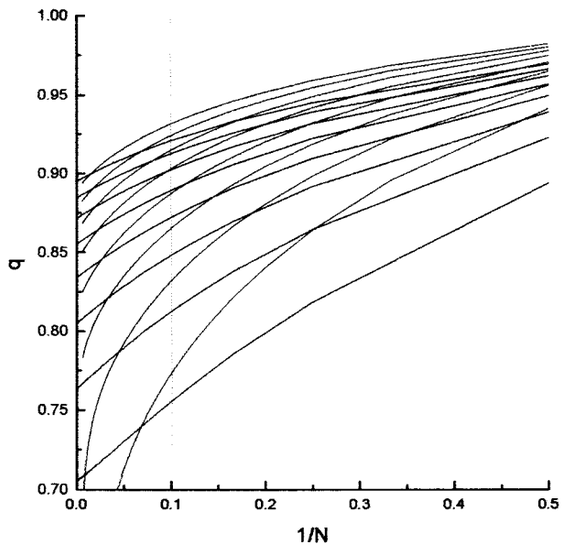

Figura 2.8: Precisão de replicação no limiar de erro q_c (curvas em azul) e o limite inferior para a precisão de replicação abaixo da qual o fenômeno do escape estocástico ocorre q_{inf} (curvas em vermelho) em função do inverso do tamanho da população. Os parâmetros são: $a = 10$ e (de baixo para cima) $L = 6, 8, ..., 20$. A linha destacada em verde serve para comparar esses valores para $N = 10$.

de um pico, essa possibilidade é catastrófica para a estrutura de população. Na figura, destacamos ainda a linha passando por $N = 10$ que é o caso que vínhamos discutindo. Note que para todos os valores de L indicados, esse tamanho da população produz $q_{\text{inf}} > q_c$ (ou $\mu_{\text{inf}} < \mu_c$).

Contudo, desde que a probabilidade de que a mestra esteja ausente da população pode ser diferente de zero também para $q > q_{\text{inf}}$, q_{inf} nos dá somente um limitante inferior para a precisão de replicação, abaixo da qual o fenômeno do escape estocástico realmente acontece. Podemos calcular também o limitante superior para essa probabilidade através de outra desigualdade, que pode ser

derivada utilizando-se de teoremas limites da teoria de probabilidade clássica. Especificamente, uma forma forte da desigualdade de Chebyshev produz [30]

$$\Pr(n_L = 0) \leq \frac{\overline{n_L^2} - \overline{n}_L^2}{\overline{n_L^2}}, \tag{2.26}$$

que nos dá, portanto, um limitante superior para a probabilidade da seqüência mestra estar ausente da população, que pode ser facilmente calculado com o auxílio das equações (2.6) e (2.23). A figura (2.9) a seguir mostra os valores dos limitantes inferior (em vermelho) e superior (em azul) da probabilidade da seqüência mestra *escapar* da população. Como vem sendo adotado, as linhas cheias representam a predição teórica e as linhas com símbolos sólidos os resultados das simulações para o algoritmo de amostragens. Com o intuito de fazer comparações, manteremos os mesmos parâmetros das figuras (2.6) e (2.7). Na parte (a) da figura mostramos o caso em que $N = 10$ e destacamos a linha passando pela taxa de mutação $\mu_c \approx 0.151$, em que ocorre o limiar de erro (linha em verde) e que foi obtida calculando-se a posição do máximo de σ_d. A figura (2.9) (a) confirma os resultados da figura (2.8) de uma maneira contundente. Para esse valor da taxa de mutação na qual o limiar de erro ocorre, a probabilidade de que a mestra não esteja presente na população (mesmo começando com uma população inicial que só contém mestras) satisfaz as desigualdades $0.6 \leq \Pr(n_L = 0) \leq 0.8$, indicando que o efeito do escape estocástico da mestra neste caso deve ser maior do que o da mestra deteriorar-se devido ao limiar de erro. Na parte (b) da figura descrevemos o que acontece com essa probabilidade no caso $N = 100$. O limiar de erro nesse caso ocorre em $\mu_c \approx 0.189$ e os limitantes não produzem informação útil sobre $\Pr(n_L = 0)$ uma vez que dão $0 \leq \Pr(n_L = 0) \leq 1$. Desde que o limitante inferior assume o seu valor mínimo e o superior, o seu valor máximo, podemos supor que a ausência da mestra na população devido às flutuações de N finito não desempenha papel importante na desestruturação da população, que

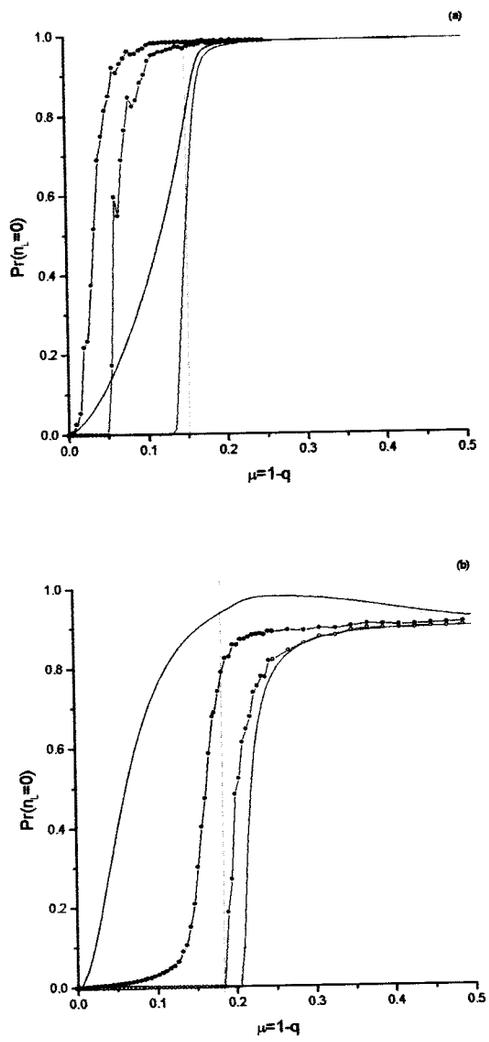

Figura 2.9: Limite inferior (em vermelho) e limite superior (em azul) para dois valores de tamanho da população: (a) $N = 10$ e (b) $N = 100$. Em ambos os gráficos destacamos o valor da taxa de mutação $\mu_c = 1 - q_c$ para o limiar de erro de replicação (linha em verde). As curvas sólidas representam os resultados analíticos. Os símbolos representam o resultado das simulações utilizando o algoritmo de amostragens.

se deve então apenas ao fenômeno do limiar de erro.

Do discutido até aqui nesta seção podemos concluir que o limiar de erro em populações finitas é um conceito relativo e, mesmo no relevo de replicação de um pico onde seu efeito é mais pronunciado, desempenha um papel menos fundamental em evolução molecular do que lhe vem sendo atribuído. De fato, fixado o comprimento da seqüência L e o valor seletivo a da mestra em relação aos mutantes, existe um valor *mínimo* do tamanho da população, $N_{\min}$, abaixo do qual o limiar de erro não ocorre, ou seja, não é a competição entre seleção e mutação que determina os parâmetros em que ocorre a perda da informação genética na população, mas sim as flutuações inerentes a finitude da população que acabam por levar ao desaparecimento da mestra da população. Esse valor mínimo pode ser determinado para qualquer valor de L calculando-se o valor de N tal que $q_c = q_{\inf}$ na figura (2.8). Os resultados são apresentados na figura (2.10) onde

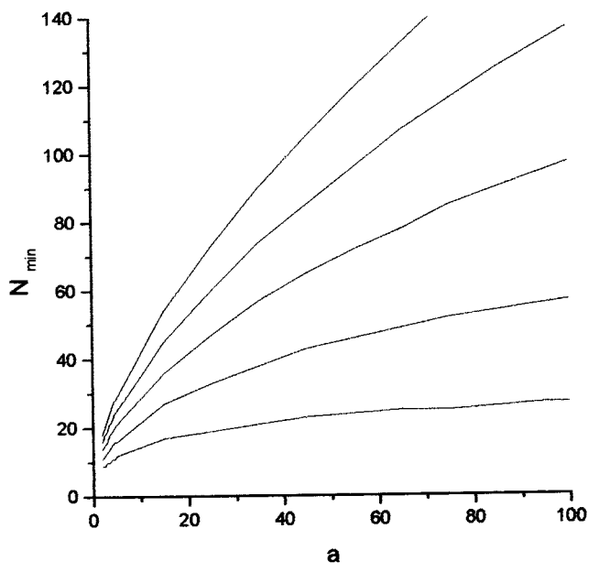

Figura 2.10: Tamanho mínimo da população $N_{\min}$ para que ocorra o limiar de erro, em função do valor seletivo a da seqüência mestra. Apresentamos as curvas para alguns valores de tamanho de seqüência (de baixo para cima), $L = 6, 8, ..., 14$.

mostramos $N_{\min}$ como função do valor seletivo da mestra a para vários tamanhos de seqüência. Note que $N_{\min}$ aumenta com L, como esperado, pois quanto maior o tamanho da seqüência menos prováveis são as mutações reversas e, portanto, mais severo é o efeito da perda da mestra.

Podemos colocar esses resultados ainda de uma outra maneira, desde que o limiar de erro é entendido como um ponto abstrato no qual a seleção não é mais capaz de sobrepor-se aos efeitos da mutação, o que estamos afirmando é que abaixo de $N_{\min}$ os efeitos da deriva genética se sobrepõem aos da seleção antes que os da mutação. Isto deve gerar, perto do limiar de erro e no equilíbrio de seleção, dois tipos distintos de populações: um caracterizado pela existência da mestra e seus vizinhos próximos (quase-espécie), e outro formado somente por mutantes que perderam a mestra em gerações passadas. Deve-se frisar que não há coexistência entre esses dois componentes, ou seja, em uma dada população apenas um deles ocorre. A figura (2.11) ilustra esse fato mostrando os resultados das simulações utilizando o algoritmo de amostragens no caso em que $N = 10$, $L = 10$ e $a = 10$. Nesta figura mostramos a freqüência da mestra e dos mutantes possíveis em função da distância de Hamming. Os resultados são obtidos após 2×10^3 gerações, através de médias nas útimas 100 gerações e em 200 rodadas do algoritmo. Como pode ser observado nas partes (b) e (c) onde as taxas de mutação são $\mu = 0.03$ e $\mu = 0.05$, respectivamente, as duas distribuições podem aparecer no equilíbrio: a primeira representando populações que mantem a mestra, e a outra populações uniformes de mutantes que perderam a mestra devido ao escape estocástico. A partir de $\mu = 0.07$ (parte (d) da figura), entretanto, os efeitos da deriva genética levam a uma deteriorização permanente da população. Observe que a região onde a mudança de comportamento ilustrado nas partes (c) e (d) ocorre coincide com o pico da variância σ_N mostrado na figura (2.7). A largura da distribuição a partir desse valor de μ vai aumentando até ficar assintoticamente

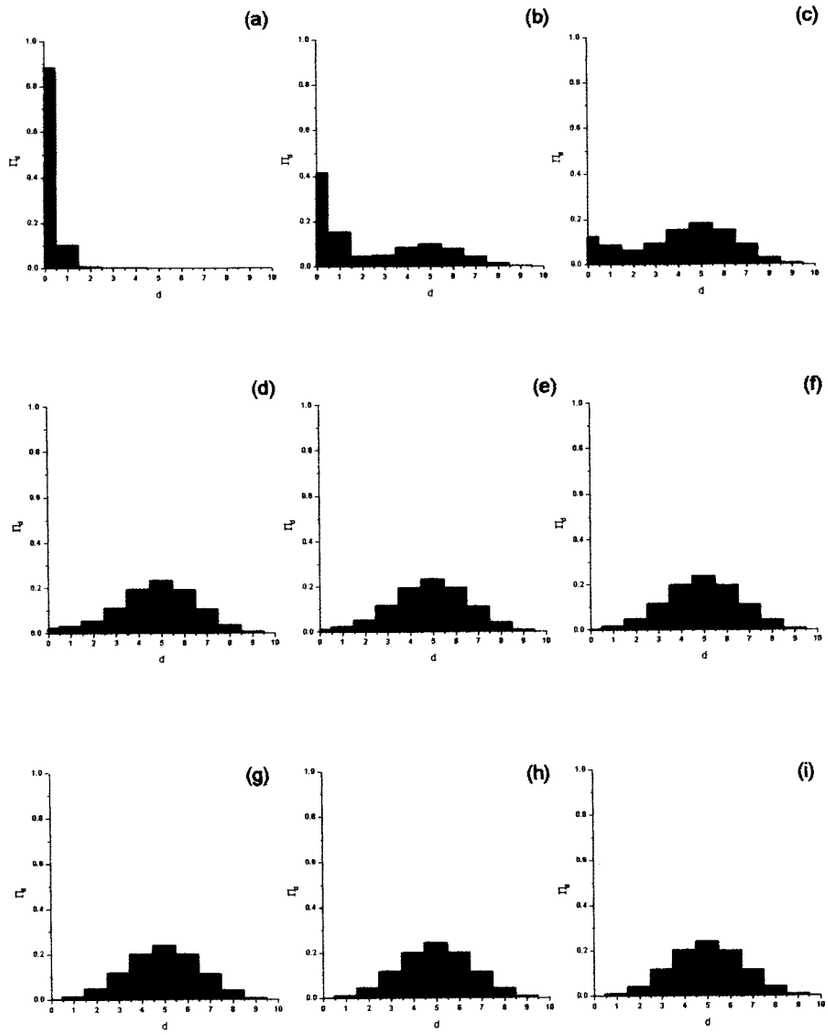

Figura 2.11: Distribuição das freqüências Π_d em função da distância de Hamming de cada classe de seqüências à mestra no equilíbrio de seleção para o algoritmo de amostragens no caso $N = 10$. Aqui fizemos $a = 10$ e $L = 10$ e também tomamos as freqüências iniciais como $\Pi_L(0) = 1$ e $\Pi_i(0) = 0$, $i = 0, ..., L - 1$. Apresentamos os resultados para vários valores de taxa de mutação: (a) $\mu = 0.01$, (b) $\mu = 0.03$, (c) $\mu = 0.05$, (d) $\mu = 0.07$, (e) $\mu = 0.09$, (f) $\mu = 0.11$, (g) $\mu = 0.13$ (h) $\mu = 0.19$ e (i) $\mu = 0.25$.

constante, seguindo o mesmo padrão observado na figura (2.6) (b). Para fazer um

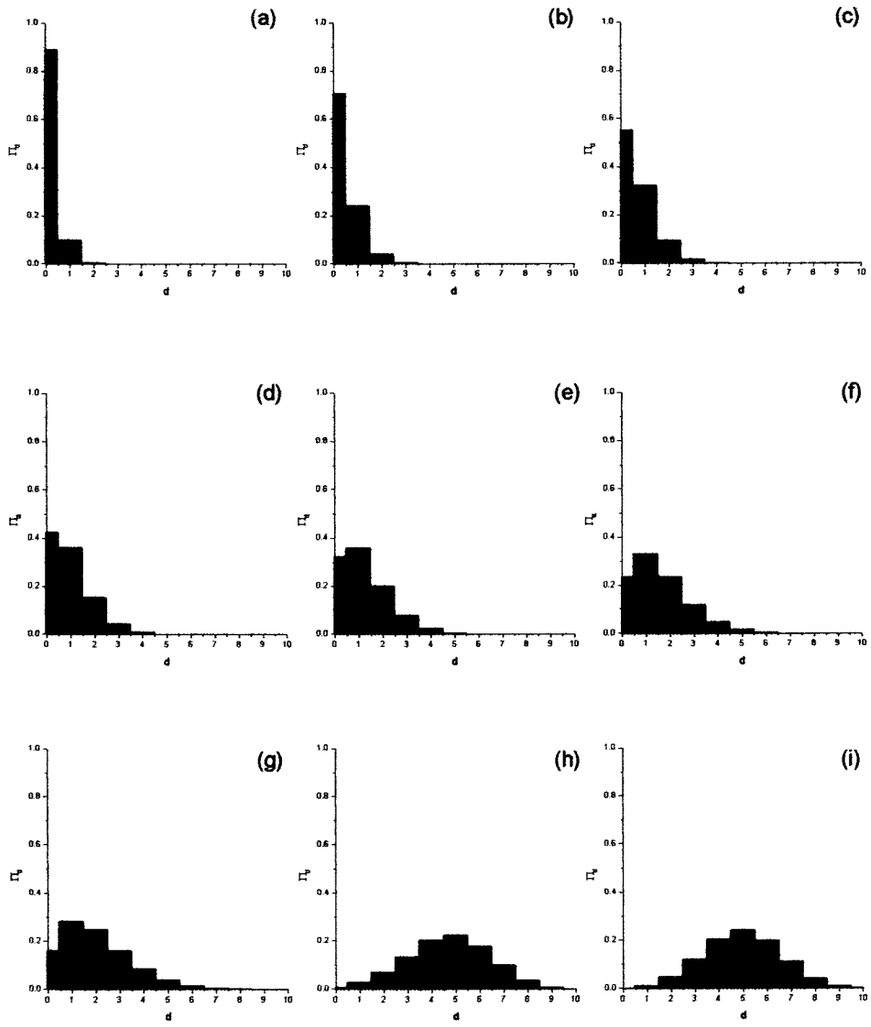

Figura 2.12: Mesmo que a figura (2.11) para o caso em que a população é constituída de $N = 100$ seqüências.

contraste com esse comportamento apresentamos na figura (2.12) os resultados de uma simulação similar usando os mesmos parâmetros para o caso $N = 100$. Fica claro que neste caso a perda da mestra se dá pelo efeito da mutação. De $\mu = 0.01$ (parte (a) da figura) até um pouco antes do limiar de erro (parte (g) da figura

em que $\mu = 0.13$), a seleção é capaz de manter a mestra na população e a deriva genética tem pouca importância no estado final de equilíbrio. A parte (h) (onde $\mu = 0.19$) representa o caso em que a taxa de mutação está muito próxima ao limiar de erro e, portanto, a distribuição é aproximadamente uma binomial com $\overline{d} = L/2$. Além de assinalar a importância relativa dos efeitos da deriva genética e da mutação sobre a seleção, essa figura mostra a diferença entre o limiar de erro em população finita e infinita: no primeiro caso a transição ocorre de maneira mais suave que no segundo.

A idéia proposta aqui de que o efeito do escape estocástico, conseqüência da deriva genética, se sobrepõe ao da seleção levando a um novo padrão de comportamento da quase-espécie é um passo audacioso em evolução molecular. Existe um debate (reminiscente da genética de organismos autônomos) sobre a importância desse efeito ser maior que o da seleção (ver [29] para uma revisão) no processo de evolução *in vivo*. Essa disputa, todavia, continua em aberto desde que não existem dados experimentais conclusivos.

Para finalizar, resta-nos comentar o modelo desenvolvido neste capítulo a luz de modelos alternativos que também descrevem o efeito do tamanho da população na teoria de quase-espécies. Há na literatura pelo menos três trabalhos estudando populações finitas que evoluem no relevo de um pico. O primeiro deles, de autoria de Nowak e Schuster [11], propõe um modelo de nascimento e morte, cujas simulações baseadas no algoritmo de Gillespie [31] reproduzem o mesmo padrão observado na figura (2.7). Todavia, o formalismo desenvolvido ali é apenas uma caricatura dos resultados das simulações (e algumas vezes contraditória) já que suas equações são aproximações válidas apenas para N suficientemente grande. De qualquer maneira, aqueles autores não conseguem recuperar os resultados do modelo original no regime determinístico e ainda subestimam a importância das mutações reversas para calcular a magnitude do limiar de erro. Um outro modelo

bastante similar ao nosso, proposto por Bonnaz e Koch [21], é obtido desconsiderando a possibilidade de erros mútiplos na replicação, sendo a diferença mais drástica a não utilização da amostragem inicial das N moléculas da sopa de genomas representada pela hipótese (2.1). Esta escolha, em nossa opinião, reduz a capacidade preditiva do modelo, apesar de que seus resultados sejam bastante similares aos apresentados na figura (2.6). Entretanto, a localização do limiar de erro é calculada estimando-se a taxa de mutação onde a freqüência da seqüência mestra vai a zero na população, o que os leva a fazer algumas simplificações de maneira que seus resultados sejam válidos apenas para valores de N suficientemente grandes. Curiosamente, seus gráficos são apresentados para um único tamanho de população, a saber, $N = 200$. Por fim, há um último trabalho que utiliza a técnica de *finite-size scaling* para determinar localização do limiar de erro em populações finitas, restrito ao relevo de um pico e $L \to \infty$ [32]. A vantagem dessa técnica é que ela não depende de qualquer definição do limiar de erro que, como discutimos, é sempre arbitrária. Entretanto, os resultados são obtidos para tamanhos de seqüências infinitos, $L \to \infty$, caracterizando o limiar de erro como uma transição de fase de primeira ordem. Um fato interessante é que em todos esses modelos encontra-se que o deslocamento do limiar de erro devido à finitude da população varia com $1/\sqrt{N}$ para N grande. Na figura (2.8) verificamos a dependência do limiar de erro com $1/N$ e $1/\sqrt{N}$ e os melhores resultados foram para a dependência com $1/N$, sendo que esse resultado não é conclusivo.

Capítulo 3

Evolução e estrutura da população em outros modelos de relevos adaptativos

Para apresentarmos as questões a serem discutidas neste capítulo vamos lembrar alguns pontos importantes desenvolvidos até aqui. Introduzidas as idéias de espaço de seqüências e de relevo de replicação, a motivação para desenvolver o modelo de amostragens era o fato de que a teoria de quase-espécies não leva em conta que a população é de tamanho finito. A principal crítica era a de que não podemos supor que haja uma concentração finita de cópias de cada seqüência, desde que o número de seqüências possíveis de um dado comprimento aumenta exponencialmente com o comprimento, e pode ser muito maior que o número total de indivíduos na população.

Ao considerarmos que a população consiste de um número finito de indivíduos com seqüências geneticamente relacionadas, introduzimos fenômenos novos no sistema, como a deriva genética competindo com os mecanismos de seleção e mutação. Particularmente, observamos que o efeito de população finita sobre o fenômeno do limiar de erro era deslocar esse ponto para valores menores de taxa de mutação. Todavia, esse efeito é pequeno se comparado com o que surge sobre a estrutura da população, desde que é importante que dentro dela exista um grupo de seqüências distintas com capacidades seletivas distintas, do contrário a seleção natural não tem como atuar.

Toda a fenomenologia estudada até aqui esteve restrita ao relevo de replicação

de um pico no equilíbrio de seleção. Esta escolha, entretanto, não é acidental e, realmente, como já tivemos oportunidade de comentar, este relevo de replicação é freqüentemente adotado na discussão do limiar de erro, tanto em modelos de população infinita [4, 13, 8, 42, 43], como em de população finita [9, 11, 21, 32]. Com efeito, como colocado por Wiehe [28], ao se estudar a teoria de quase-espécies é muito fácil ficar com a impressão de que qualquer topografia do relevo de replicação que contenha uma *mestra com valor seletivo superior* pode produzir um limiar de erro, cujo valor é inversamente proporcional ao tamanho da seqüência ([34], pag. 222). Outro aspecto importante ressaltado em seu trabalho [28] é que o estudo do limiar de erro e sua dependência com o relevo de replicação está em geral restrito ao limite $L \to \infty$, perdendo a característica essencial desse fenômeno que é o de limitar o tamanho das seqüências [35].

Portanto, neste capítulo estaremos preocupados em estudar a estrutura da população em uma outra classe bastante geral de relevos de replicação em que não é levado em consideração a interação entre as seqüências da população. Esta análise será realizada considerando-se populações finitas, tomando-se o limite determinístico como caso particular, em todo o espaço de parâmetros relevantes do sistema. Em particular, estaremos interessados em distingüir os estados estacionários desses dois regimes. Para essa empreitada precisamos generalizar o modelo de amostragens para um relevo de replicação arbitrário, o que é feito a seguir.

3.1 Generalização do formalismo de amostragens

Como já havíamos discutido, a equação de recorrência (2.10) é a versão do modelo de quase-espécies para populações de tamanho finito que, para o relevo de replicação de um pico, resulta em uma expressão simples em função dos

parâmetros do sistema. Também havíamos adiantado que para generalizar esta equação para um relevo de replicação arbitrário necessitamos especificar a taxa de replicação relativa de cada indivíduo dada por (2.2), o que nos levou a equação (2.10)

$$\overline{n}_P(t+1) = N \sum_{\mathbf{n}} \sum_{R} \left(\frac{n_R A_R}{\sum_K n_K A_K} \right) M_{PR} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}), \tag{3.1}$$

e que, por sua vez, não podia ser diretamente iterada para um conjunto A_P qualquer devido ao custo computacional de se efetuar o somatório em $\mathbf{n}$ que envolve $(N+1)^{L+1}$ termos.

Para resolver esse incoveniente, vamos utilizar a seguinte relação, válida para uma variável aleatória contínua X distribuída exponencialmente com parâmetro λ [25]:

$$E(X) = \int_0^\infty x \left(\lambda e^{-\lambda x} \right) dx = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}, \tag{3.2}$$

onde $E(X)$ é o valor esperado (média) de X. Fazendo-se $\lambda = \sum_K n_K A_K$ e voltando-se com essa forma integral em (2.10) ou (3.1), ficamos com

$$\overline{n}_P(t+1) = N \sum_{\mathbf{n}} \sum_{R} \left(n_R A_R \int_0^\infty e^{-(\sum_{K=0}^{L} n_K A_K)x} dx \right) M_{PR} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}). \tag{3.3}$$

Finalmente, com a definição (2.1), podemos efetuar facilmente o somatório sobre $\mathbf{n}$ (ver Apêndice D para detalhes dos cálculos) até chegarmos em

$$\overline{\Pi}_P(t+1) = N \int_0^\infty e^{-x} \sum_{R=0}^{L} M_{PR} \overline{\Pi}_R(t) \left(\sum_{K=0}^{L} e^{-x A_K / A_R} \overline{\Pi}_K(t) \right)^{N-1} dx. \tag{3.4}$$

Esta equação é facilmente iterada para qualquer tipo de relevo replicativo e taxa de mutação, dentro das prescrições estabelecidas [23], ou seja, dadas as freqüências iniciais de seqüências ela descreve os números de ocupação médios da população a cada geração, bastando especificar as taxas de replicação A_P de cada seqüência e os elementos da matriz de mutação M_{PR}.

Uma outra grandeza que pode ser derivada nessa forma mais geral é o segundo momento de Π_P, que será importante nas próximas seções. Assim, usando um

procedimento similar ao utilizado para derivar a equação (2.23), obtemos (ver Apêndice D)

$$\overline{\Pi^2}_P(t+1) = \frac{\overline{\Pi}_P(t+1)}{N} +$$

$$+ (N-1)\int_0^\infty x e^{-x} \sum_{R=0}^{L} e^{-A_R x} M_{PR}^2 A_R^2 \overline{\Pi}_R(t) \left(\sum_{K=0}^{L} e^{-xA_K} \overline{\Pi}_K(t) \right)^{N-1} dx$$

$$+ (N-1)^2 \int_0^\infty x \left(\sum_{R=0}^{L} e^{-A_R x} M_{PR} A_R \overline{\Pi}_R(t) \right)^2 \left(\sum_{K=0}^{L} e^{-xA_K} \overline{\Pi}_K(t) \right)^{N-2} dx. \quad (3.5)$$

No regime estacionário fazemos $\overline{\Pi^2}_P(t+1) = \overline{\Pi^2}_P(t) = \overline{\Pi^2}_P$.

O modelo analítico, refletido pela equação (3.4), é de uma generalidade sem precedentes na literatura, no sentido de ser válido para qualquer relevo de replicação em que podemos desprezar as interações internas entre as seqüências (mesmo assim, na seção (3.3) estudaremos como este modelo comporta-se quando consideramos um tipo de *interação média* entre seqüências chamada de *epistase*). Realmente, nas seções seguintes vamos fazer uma análise detalhada desta equação para outros relevos, onde estaremos interessados nas propriedades de equilíbrio da população.

Antes disso, todavia, é interessante derivarmos o caso determinístico dessa equação, que é obtido tomando-se o limite $N \to \infty$ na eq. (3.4). Nesse limite podemos efetuar a integral utilizando o método de integração por ponto de sela. Observando que a equação (3.4) é da forma

$$\overline{n}_P(t+1) = \sum_{R=0}^{L} M_{PR} \overline{\Pi}_R(t) I_P(t), \quad (3.6)$$

onde

$$\begin{aligned} I_P(t) &= N^2 \int_0^\infty \frac{e^{-xA_R} \overline{\Pi}_R(t)}{\sum_{K=0}^{L} e^{-xA_K} \overline{\Pi}_K(t)} e^{N \ln\left(\sum_{K=0}^{L} e^{-xA_K} \overline{\Pi}_K(t)\right)} dx \\ &= N^2 \int_0^\infty g(x) e^{N f(x)} dx, \end{aligned} \quad (3.7)$$

e desde que o máximo do argumento da exponencial, ocorre em $x = 0$, podemos expandir $f(x)$ e $g(x)$ em torno desse valor. Depois de algum cálculo, chegamos a equação de recorrência para a freqüência de cada seqüência na população, dado que N é suficientemente grande,

$$\overline{\Pi}_P(t+1) = \frac{\sum_{R=0}^{L} M_{PR}\overline{\Pi}_R(t)A_R}{\overline{w}(t)}\left[1 - \frac{1}{N\overline{w}(t)}\left(B_R - \frac{C_2}{\overline{w}(t)}\right) - O\left(N^{-2}\right)\right], \quad (3.8)$$

onde $\overline{w}(t) = \sum_{K=0}^{L}\overline{\Pi}_K(t)A_K$ é o *fitness* médio da população normalizado. Os termos de primeira ordem em $1/N$ da expansão são

$$B_R = A_R - \overline{w}(t) \quad (3.9)$$

e

$$C_2 = \overline{w^2}(t) - \overline{w}(t)^2, \quad (3.10)$$

onde $\overline{w^2}(t) = \sum_{K=0}^{L}\overline{\Pi}_K(t)A_K^2$. É interessante comentar que ao verificarmos a normalização em (3.8), $\sum_{P=0}^{L}\overline{\Pi}_P(t) = 1$, cada termo da expansão se anula exceto o independente de N, desde que a matriz de mutação é uma matriz de transição e portanto vale a relação $\sum_{P=0}^{L} M_{PR} = 1$.

Assim, podemos tomar o limite $N \to \infty$ em (3.8) que finalmente reduz-se a

$$\overline{\Pi}_P(t+1) = \frac{\sum_{R=0}^{L} M_{PR}\overline{\Pi}_R(t)A_R}{\overline{w}(t)}, \quad (3.11)$$

generalizando então a equação (2.17) obtida no capítulo anterior. Como comentado naquele capítulo, essa equação indica que para o modelo mais geral de quase-espécies a população é uma combinação aleatória das moléculas constituintes; este resultado, portanto, sendo válido para qualquer relevo de replicação. Obviamente, especificando o relevo de replicação de um pico em (3.4) ou (3.11), reproduzimos os resultados obtidos no capítulo anterior.

Neste capítulo, vamos estar estudando a aplicação deste formalismo a vários relevos de replicação, a maioria deles encontrados na literatura especializada.

Neste ponto é importante introduzirmos alguns elementos da linguagem da genética de populações, já que a grande maioria dos artigos dedicados a evolução de populações em relevos de replicação que não o de um pico utilizam-se dessa linguagem. Assim, as macro-moléculas ou simplemente seqüências binárias serão interpretadas como *indivíduos* de uma população de *genomas* de comprimento L. Cada posição do *genoma* será entendida como um *gene* (ou *locus*) que pode assumir diferentes formas ou *alelos*, representados pelos dígitos 0 e 1, no caso das seqüências binárias que vem sendo estudado até aqui. Portanto, deste ponto em diante, utilizaremos indistintamente os termos seqüência ou genoma e monômero ou alelo. Essa forma de expressão poderá auxiliar o leitor interessado em se aprofundar no tema da evolução biológica.

Por último, aproveitamos o contexto para introduzir o processo conhecido como *catraca* ou *cremalheira* de Muller (do inglês Muller's ratchet) [36] que é bastante estudado em genética de populações [17, 20, 28, 37, 38, 39, 40]. Em uma população de genomas de tamanho infinito a perda dos indivíduos mais aptos é irreversível. Isso acontece se a população for finita, pois a deriva genética certamente levará à extinção sucessiva do indivíduo mais apto, do segundo mais apto, e assim por diante, acarretando um decréscimo do *fitness* médio da população como em uma cremalheira de relógio. Esse processo é descrito no modelo de Muller [36] que, por sua vez, tem uma estrutura matemática idêntica a do modelo de Eigen. Ambos modelos, baseados em equações para o balanço entre seleção e mutação, predizem um limite superior para a taxa de mutação acima da qual a evolução não pode ser controlada pela seleção natural, entretanto através de processos diferentes. Portanto, além da generalização do formalismo de amostragem, neste capítulo iremos lidar com relevos de replicação que são utilizados no estudo da *catraca* de Muller, o que nos permitirá estudar como esse fenômeno afeta a evolução da quase-espécie.

3.2 Relevo de replicação multiplicativo

A primeira aplicação dessa forma mais geral do modelo de amostragem que vamos estudar é a evolução de uma população finita de N indivíduos em um relevo de replicação *suave* chamado de relevo *multiplicativo*, devido à forma como é construído. Sua importância reside no fato de ser bastante utilizado no estudo do processo da *catraca* Muller.

Como tem sido feito até agora, vamos supor que cada indivíduo da população seja representado por um genoma de L genes e que cada gene possa assumir um entre dois possíveis alelos, 0 ou 1. Para introduzir o relevo de replicação multiplicativo, vamos imaginar que cada 1 no genoma represente um alelo favorável e vamos atribuir-lhe um valor seletivo relativo igual a 1. Da mesma maneira, interpretamos cada 0 na seqüência como sendo um alelo desfavorável e tendo um valor seletivo relativo igual a $(1-s)$, com $0 \leq s \leq 1$. Assim, mantendo-se a notação utilizada até aqui, se P for o número de $1's$ que uma seqüência particular contém, o valor seletivo de um indivíduo com $L-P$ alelos desfavoráveis será

$$A_P = (1-s)^{L-P}. \tag{3.12}$$

Utilizando também a definição de distância de Hamming de um indivíduo diferindo de d alelos da seqüência com L $1's$, podemos reescrever o valor seletivo de um indivíduo com Ld alelos desfavoráveis como $A_d = (1-s)^{Ld}$. Obviamente, as duas notações são equivalentes e são apresentadas apenas como ilustração. O importante de se destacar é que, na forma da função que representa o valor seletivo de um indivíduo A_P ou A_d, estamos supondo que a contribuição dos diferentes loci seja independente, originando daí o caráter multiplicativo do relevo. O relevo de replicação nesse caso tem um genoma *ótimo*, no sentido de possuir o maior valor seletivo possível, sendo que para qualquer outro genoma na população, este valor depende da distância de Hamming ao genoma *ótimo*. Qualquer relevo com

essa propriedade também pode ser chamado de relevo de um pico, todavia para distingüir do caso estudado no capítulo 2, vamos chamá-lo simplesmente de relevo multiplicativo ou ainda do tipo Fujiama devido a semelhança com o monte que leva esse nome [33]. Como no caso do relevo de um pico, o multiplicativo representa um caso extremo, ou seja, uma aproximação local de um relevo mais geral com muitos picos suaves, no sentido de que o valor seletivo aumenta regularmente até o pico em todas as direções (na próxima seção vamos fazer uma comparação mais detalhada entre esses dois relevos).

Introduzido o relevo de replicação, para estudarmos como as mutações se acumulam na população, iremos nos concentrar na distância de Hamming média ($\overline{d}$) da população ao *genoma ótimo*, já que ela mede o número médio de genes desfavoráveis por indivíduo ($L\overline{d}$). No modelo de amostragens, esta grandeza é facilmente obtida para o relevo multiplicativo, sendo dada por

$$\overline{d} = \frac{1}{L}\sum_{P}(L-P)\,\overline{\Pi}_P, \tag{3.13}$$

onde as freqüências médias $\overline{\Pi}_P$ são calculadas através da equação (3.4) no equilíbrio de seleção. Particularmente, especificando o relevo multiplicativo em (3.4) ficamos com

$$\overline{\Pi}_P(t+1) = N\int_0^\infty e^{-x}\sum_{R=0}^{L} M_{PR}\overline{\Pi}_R(t)\left(\sum_{K=0}^{L} e^{-x(1-s)^{R-K}}\overline{\Pi}_K(t)\right)^{N-1} dx. \tag{3.14}$$

Esta equação pode então ser iterada de maneira a nos dar o comportamento da população para quaisquer valores dos parâmetros do sistema.

Como primeiro passo em nossa análise, podemos resolver (3.14) exatamente para dois limites em que o valor de $\overline{d}$ é conhecido [20, 33, 37]. O primeiro é o chamado limite de evolução neutra, ou seja, quando $s = 0$. Isto significa que o relevo de replicação é *plano*, ou seja, como todos os genomas são seletivamente equivalentes, a seleção natural não atua. A equação (3.14) neste limite nos leva

a relação simples

$$\overline{\Pi}_P(t+1) = \sum_{R=0}^{L} M_{PR}\overline{\Pi}_R(t),$$

e portanto a

$$\overline{d}(t+1) = \mu + (1-2\mu)\,\overline{d}(t), \tag{3.15}$$

de onde concluímos que, no equilíbrio de seleção, o ponto fixo é $\overline{d} = 0.5$ qualquer que seja a população inicial e para qualquer valor de $\mu > 0$. À medida que aumentamos μ, esse ponto fixo é alcançado cada vez mais rapidamente até isso acontecer em apenas uma geração, quando $\mu = 0.5$. Para este valor de taxa de mutação ($\mu = 0.5$) e s arbitrário, temos o outro limite conhecido, que em nosso modelo é uma solução degenerada de (3.14), no sentido de que mais uma vez $\overline{d} = 0.5$, independente dos outros parâmetros do sistema. É importante frisar que esse é um comportamento médio e, no caso neutro, nos dá pouca informação sobre o comportamento real de uma população. De fato, para $\mu = 0$ e $s = 0$ qualquer um dos 2^L tipos de indivíduos poderá fixar-se na população com a mesma probabilidade (supondo uma distribuição inicial uniforme). Essa fixação é devida somente ao fenômeno da deriva genética. Assim, se mediarmos sobre um número infinito de populações obteremos $\overline{d} = 0.5$ simplesmente porque o número de populações em que um indivíduo com $P = L/2$ (a classe mais numerosa) fixou-se é muito maior do que qualquer outra possibilidade de fixação.

Em geral, sabe-se também que $\overline{d}$ é uma função crescente de μ e uma função decrescente de s, todavia esses resultados têm sido apresentados apenas como aproximações das simulações dos modelos propostos [20, 28]. Obviamente, uma solução analítica *fechada* da equação (3.14) não é possível, entretanto, sua integração numérica é extremamente estável e simples, além de concordar muito bem com as simulações do modelo. Assim, iteramos a equação (3.14) para vários valores de $s > 0$ para estudar a estrutura da população em função da taxa de

mutação μ.

Começamos com uma população composta apenas por indivíduos idênticos à seqüência ótima evoluindo no relevo multiplicativo discutido acima. É claro que inicialmente todas as mutações são desfavoráveis e, portanto, a população afasta-se da seqüência ótima, aumentando o número de alelos do tipo 0, $L\bar{d}$. À medida que d aumenta, a probabilidade de que uma mutação favorável ocorra também aumenta. Depois de um certo tempo, a população alcança um estado estacionário no qual a ocorrência de mutações desfavoráveis é balanceada pela ação da seleção mais a ocorrência de mutações favoráveis. Na figura (3.1) apresentamos a distância de Hamming média normalizada da população à seqüência ótima em função da taxa de mutação μ, no estado estacionário. Ali comparamos os resultados das simulações do modelo de amostragem (linhas com símbolos) com a predição teórica da equação (3.14) (linhas cheias) para vários valores de $s > 0$. A população descrita contém $N = 10$ seqüências, todas elas com tamanho fixo $L = 10$ e, como pode ser observado, já para esse tamanho de população os resultados das simulações concordam bastante bem com os resultados analíticos. As linhas pontilhadas coloridas representam o regime determinístico e as duas linhas pontilhadas em preto são os dois casos extremos $s = 0$ (em (a)) e $s = 1$ (em (b)). As simulações são feitas deixando-se o algoritmo de amostragens evoluir por 2×10^3 gerações e mediando-se sobre as últimas 100 gerações. Cada ponto representa uma média sobre 100 rodadas independentes de simulação. Como no capítulo 2, não foi observada nenhuma diferença significativa para escolhas diferentes das freqüências iniciais dos genomas.

Como já comentamos, $L\bar{d}$ representa o número médio de genes desfavoráveis por indivíduo e, portanto, observamos que, de uma maneira geral, para $\mu \to 0$ somente a seqüência ótima permanece na população no estado estacionário, ou seja, $L\bar{d} \to 0$. À medida que vamos aumentando a taxa de mutação, o comportamento

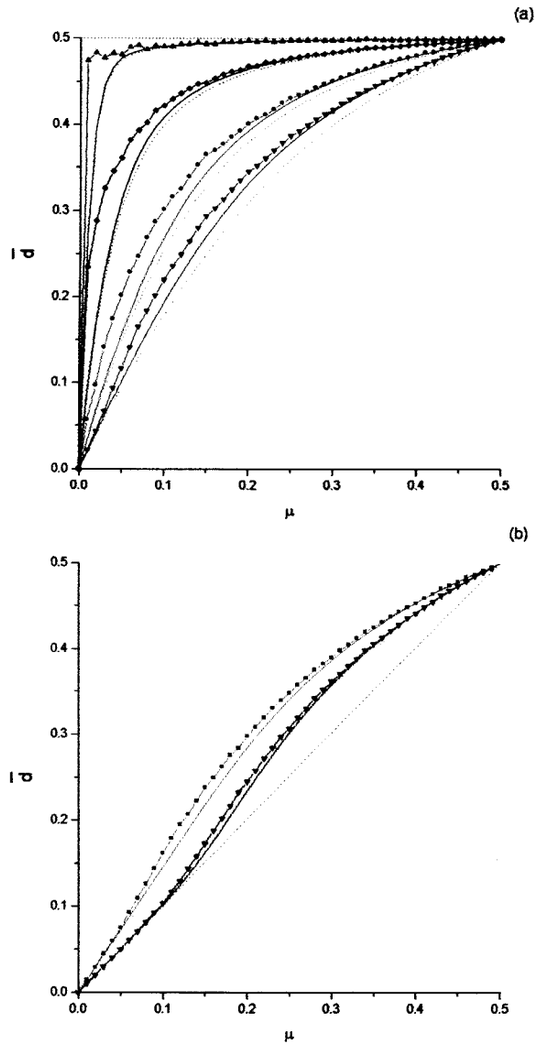

Figura 3.1: Distância de Hamming média em função da taxa de mutação para uma população de $N = 10$ seqüências, cada uma de tamanho $L = 10$, evoluindo em um relevo de replicação multiplicativo. As linhas com símbolos são os resultados das simulações e as linhas cheias mostram os resultados da equação (3.15) para vários valores de $s > 0$. As cores distingüem cada valor de s apresentados de forma que de cima para baixo tenhamos: (a) $s = 0.01$, $s = 0.1$, $s = 0.3$ e $s = 0.5$; (b) $s = 0.7$ e $s = 0.99$. As linhas pontilhadas coloridas são o limite determinístico para cada valor de s representado. As linhas pontilhadas em preto representam em (a) o limite neutro $s = 0$ e em (b) o caso $s = 1$.

da população para vários valores de s é similar, e quando $\mu \to 0.5$ os genomas dos filhos não têm nenhuma correlação com os genomas dos pais. Apesar desse comportamento geral para s, observa-se uma mudança na estrutura da população para $s \to 1$. Isso justifica a separação da figura (3.1) em duas partes: (a) para $0 \leq s \leq 0.5$ e (b) para $0.5 < s \leq 1$.

Assim, na parte (a), em vermelho, destacamos a curva para $s = 0.01$, mostrando que neste caso a seleção natural não é suficientemente forte para manter a população ordenada e para valores pequenos da taxa de mutação a população já está bastante próxima do regime neutro (esse resultado é mais contundente nas simulações). É importante destacar que o efeito de população finita aqui é pequeno em comparação com a devastação mutacional, o que justifica a proximidade dos resultados para uma população tão pequena com a curva que representa o regime determinístico (curva pontilhada em vermelho). Como pode ser notado da figura, com o aumento de s as curvas para população finita afastam-se cada vez mais das curvas que representam uma população infinita, indicando que o efeito de aumentar s (ou seja, diminuir o valor seletivo dos mutantes na população) é aumentar a importância da deriva genética frente à mutação. Assim, fixados s e μ, o número médio de genes deletérios acumulados na população, $L\overline{d}$, aumenta à medida que diminuímos N. Este fenômeno é a *catraca de Muller* para populações com seqüências de tamanho finito L [41] que *pára de rodar* mantendo a quase-espécie a uma distância $\overline{d}$ do pico do relevo. Esta distância depende da taxa de mutação a que a população está sujeita. No caso conhecido desse processo, em que $L \to \infty$ [36], a perda da seqüência ótima é irreversível e a *catraca roda* até a população não ser mais viável, ou seja, $\overline{d} = 0.5$ no equilíbrio.

Para valores seletivos $s > 0.5$, observamos que existem dois tipos de comportamentos da população no relevo multiplicativo para N finito. Isso pode ser visto na parte (b) da figura (3.1) onde destacamos dois valores de s em que a

pressão de seleção é muito grande: $s = 0.7$ (em azul claro) e $s = 0.99$ (em azul marinho), um valor bastante próximo do caso limite $s = 1$. Para valores pequenos de μ até um valor crítico μ_c, as curvas para população finita e infinita para cada valor de s (as linhas cheias e as pontilhadas) dão o mesmo resultado para a distância de Hamming média normalizada. A partir desse valor crítico de taxa de mutação, o distanciamento das curvas é similar ao observado para outros valores de s mostrados na parte (a) da figura.

Antes de tentarmos entender esse fenômeno, é importante frisar que a tendência a um *descolamento* das curvas para população finita em relação às curvas para população infinita, observado para algum valor $\mu_c > 0$, ocorre já para valores de $s > 0.5$, diferentemente do que fica subentendido no artigo de Woodcock e Higgs [20], onde aparentemente esse é um efeito exclusivo do caso extremo em que $s \longrightarrow 1$. Para ressaltar esse fato apresentamos na figura (3.2) a grandeza $\sigma_N^2(L)$ em função da taxa de mutação. Essa grandeza mede o espalhamento da seqüência ótima na população e, conforme discutido no capítulo anterior, é escrita de maneira simples como

$$\sigma_N^2(L) = \overline{\Pi_L^2} - \overline{\Pi}_L^2. \tag{3.16}$$

A figura mostra o comportamento dessa grandeza para alguns valores de s (em vermelho, $s = 0.01$, em azul, $s = 0.3$, em verde, $s = 0.7$ e em rosa, $s = 0.99$). As linhas com símbolos cheios são os resultados das simulações e as curvas sólidas representam os resultados analíticos obtidos iterando-se as equações (3.4) e (3.5) até o equilíbrio de seleção ser alcançado. O resultado interessante que pode ser observado na figura é o crescimento do máximo da variância do genoma ótimo e o subseqüente decréscimo à medida que aumentamos s a partir de $s = 0.01$, até $s = 0.99$. Ora, era de se esperar um decréscimo constante do espalhamento da seqüência ótima à medida que melhoramos seletivamente a população. Entretanto, neste relevo de replicação temos que considerar dois efeitos agindo con-

comitantemente, e daí a importância relativa da distância de Hamming média para descrever o comportamento da população. À medida que s aumenta a partir de zero, a diferença entre o *fitness* dos indivíduos de classes diferentes torna-se maior, e o indivíduo mais adaptado (que possui o genoma ótimo) começa a gerar mais descendentes às expensas dos menos adaptados. Nesse caso o espalhamento

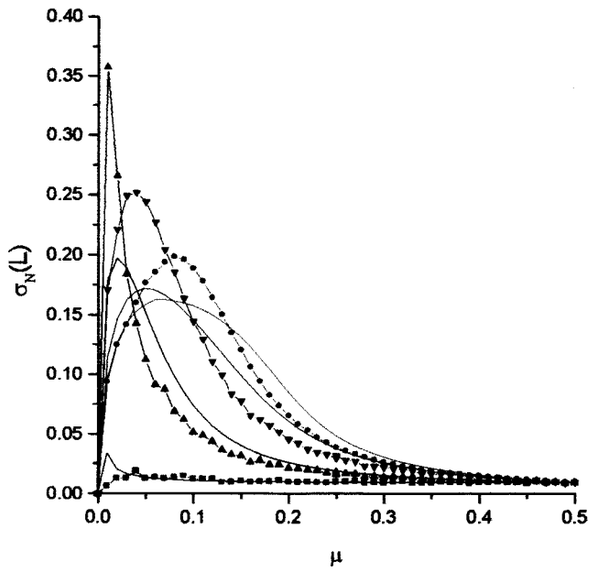

Figura 3.2: Variância da seqüência ótima em função da taxa de mutação. As cores assinalam os valores seletivos utilizados: $s = 0.01$ (em vermelho), $s = 0.3$ (em azul), $s = 0.7$ (em verde), e $s = 0.99$ (em rosa). As linhas com símbolos representam as simulações utilizando o algoritmo de amostragens e as linhas sólidas a aproximação analítica. Os resultados mostrados são para $N = 10$, $L = 10$ e $a = 10$.

da seqüência ótima na população tende a aumentar rapidamente para valores de taxa de mutação pequenos. Se s é muito próximo a 1, o espalhamento de valores seletivos possíveis na população torna-se pequeno e a variância do genoma ótimo torna-se menor indicando que, nesse limite, para remover o genoma ótimo da população necessita-se de uma taxa de mutação muito maior do que para remover o mutante distante uma mutação do ótimo. No limite em que a seleção é muito grande ($s \to 1$) uma grande parte da população é forçada para a classe mais apta.

Podemos então entender o comportamento do sistema no caso em que $s > 0.5$ se compararmos os resultados da figura (3.1) (b) com os mostrados na figura (3.2). Fazendo isso, devemos concluir que para valores pequenos de μ até um valor crítico μ_c, a seqüência ótima está quase sempre presente na população, enquanto que para valores da taxa de mutação maiores que esse valor crítico, a seqüência com maior *fitness* varia de uma geração a outra e, freqüentemente, não é mais a ótima. Esse raciocínio também deve nos levar à conclusão de que, mesmo para valores extremos de seleção, a deriva genética faz com que a população não seja capaz de manter a seqüência ótima.

Entretanto, a partir desse valor μ_c da taxa de mutação a perda da seqüência ótima nesse relevo não é tão drástica como era no caso do relevo de um pico. No relevo multiplicativo, à medida que aumentamos a taxa de mutação a quase-espécie vai se afastando lentamente do pico, enquanto que no relevo de um pico isso ocorre mais rapidamente, mesmo para populações finitas. Assim, nos dois relevos, o de um pico e o multiplicativo, podemos observar um comportamento similar com relação a manutenção na quase-espécie da seqüência com o maior *fitness* possível do relevo (a mestra no relevo de um pico e a ótima no multiplicativo). Para populações finitas, isso é possível até um valor crítico μ_c da taxa de mutação, a partir do qual essa seqüência tem uma freqüência muito pequena na população. Em ambos relevos, a causa da perda desse genoma, ou melhor ainda, da diminuição acentuada de sua freqüência, é o fato da seleção natural, aqui *enfraquecida* pela deriva genética, não ser capaz de reparar o efeito mutacional. A diferença principal entre os dois relevos é o modo como essa perda afeta a estrutura da quase-espécie. No relevo de um pico a perda da mestra (o limiar de erro) é desastroso, já que a população é *jogada* para uma distância $L/2$ do pico do relevo. Entretanto, no relevo multiplicativo a conseqüência da perda da seqüência ótima é deslocar a quase-espécie para uma distância $L-1$ do pico do relevo (não há um

limiar de erro) e à medida que aumentamos o valor seletivo, cada vez maior tem de ser a taxa de mutação μ_c para que isso ocorra. Para valores seletivos grandes, a perda da seqüência ótima é postergada e o tamanho da população irá definir o valor de μ_c através do escape estocástico. Como discutimos no capítulo anterior, o escape estocástico só é efetivo no relevo de um pico, onde torna-se o responsável pela perda da mestra para uma população muito pequena (menor que $N_{\min}$).

Para ressaltar essa interpretação da tendência ao descolamento das curvas observada na figura (3.1) (b), apresentamos na figura (3.3) uma análise do escape estocático da seqüência ótima similar ao estudo feito no capítulo anterior (ver figura (2.9)), no caso extremo $s = 0.99$ que apresenta mais pronunciadamente esse duplo comportamento. As linhas com símbolos são os resultados das simulações usando o algoritmo de amostragens e as linhas cheias representam a aproximação analítica. Em azul destacamos o limitante superior e em vermelho o limitante inferior para a probabilidade de que a seqüência ótima não esteja presente na população. Os resultados analíticos foram obtidos iterando (3.14) e substituindo a solução no estado estacionário em (3.5) particularizada para o relevo multiplicativo. Com os dados obtidos, utilizamos as relações (2.25) para obter o limitante inferior e (2.26) para obter o limitante superior. Na parte (a) mostramos o caso $N = 10$ e na parte (b) o caso $N = 100$. Em ambos os casos as curvas (azul e vermelha) definem o intervalo do gráfico em que o escape estocástico da seqüência ótima ocorre e confirma o que tinha sido discutido anteriormente. Se nos atermos as linhas que representam as simulações, podemos notar que a probabilidade de que a seqüência ótima não esteja presente na população aumenta suavemente, satisfazendo $0 \leq \Pr(n_L = 0) \leq \Pr_{\text{sup}}$, à medida que aumentamos μ ($\Pr_{\text{sup}}$ sendo o ponto da curva em azul para cada μ), até atingirmos um certo valor de taxa de mutação ($\mu \approx 0.14$, para $N = 10$ e $\mu \approx 0.34$, para $N = 100$, destacados com a linha verde). A partir desses valores de μ, temos $0 < \Pr_{\text{inf}} \leq \Pr(n_L = 0) \leq \Pr_{\text{sup}}$

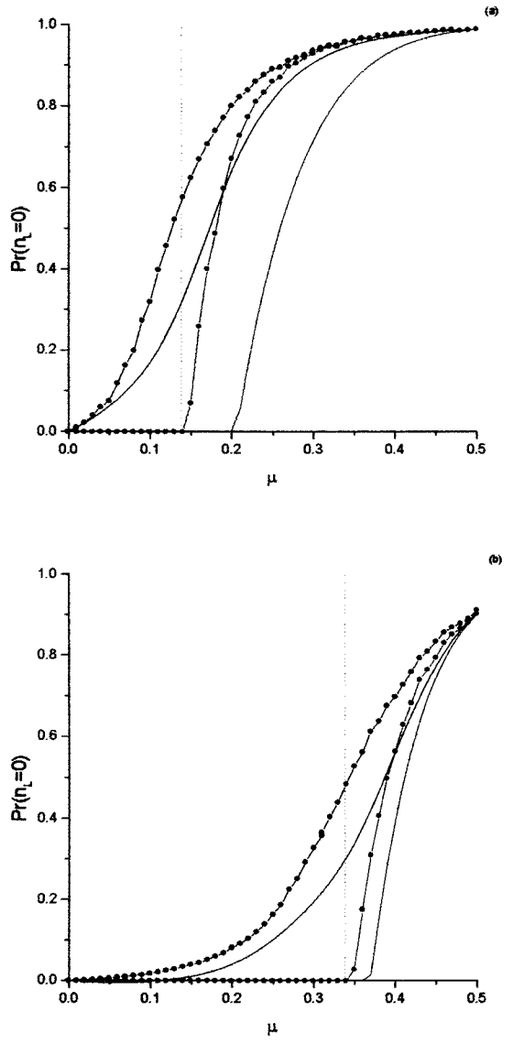

Figura 3.3: Escape estocástico da seqüência ótima de uma população evoluindo em um relevo multiplicativo, no caso em que a pressão de seleção é muito forte ($s = 0.99$). Limitante inferior (em vermelho) e limitante superior (em azul) para dois valores de tamanho de população (a) $N = 10$ e (b) $N = 100$. Em ambos os gráficos destacamos um valor aproximado da taxa de mutação μ_c no qual a catraca de Muller começa a rodar (linha em verde).

nos dois casos, ou seja, estes valores da taxa de mutação são aproximadamente os pontos nos quais o limitante inferior passa a ser maior que zero. É interessante que esses valores de μ destacados em verde na figura concordam muito bem com os valores de μ_c obtidos pela análise da figura (3.1). Para finalizar, devemos mencionar a excelente concordância qualitativa entre os padrões de comportamento dos limitantes obtidos através das simulações e da aproximação analítica.

3.3 Seleção com interação epistática

Até aqui estudamos o balanço entre os mecanismos de seleção e mutação supondo que a presença de um dado alelo em um *locus* não dependa dos outros alelos presentes em outros *loci*, ou seja, desprezamos as interações internas entre os diferentes *loci* do genoma. Na linguagem da genética de populações, isso quer dizer que não levamos em conta as *interações epistáticas*, ou simplesmente *epistase*. Realmente, da maneira como construímos o nosso modelo, este tipo de interação não pode ser trabalhada de maneira realística, já que em nossa simplificação supomos a independência entre os diferentes genes. Entretanto, mesmo nesse tipo de modelo é possível introduzir uma *interação média* entre os genes. Com efeito, este é um tópico extensivamente estudado em genética de populações desde sua introdução por Kimura e Muruyama [44] para investigar o efeito da epistase sobre a perda mutacional da população. Esses autores argumentaram que uma epistase média entre os genes pode ser definida sobre o relevo de replicação no qual uma população evolui. Assim, a partir do relevo multiplicativo, dado pela equação (3.12), podemos definir um relevo multiplicativo com epistase através da generalização

$$A_P = (1-s)^{(L-P)^\alpha}, \tag{3.17}$$

onde α é um parâmetro positivo que caracteriza a ordem da interação. Seguindo a notação comumente empregada [28, 40], essa equação gera três tipos de relevos,

a saber, relevo com epistase atenuante ($0 < \alpha < 1$), relevo multiplicativo ou com ausência de epistase ($\alpha = 1$), ou ainda relevo com epistase sinergística ($\alpha > 1$). Nosso objetivo nesta seção é tentar entender como esse tipo de interação afeta a estrutura da quase-espécie, atendo-nos, em particular, a dependência do limiar de erro com relação a intensidade e o tipo de epistase.

Para podermos comparar as propriedades dos relevos a serem discutidos aqui, primeiramente devemos fazer uma mudança de escala no relevo de um pico, utilizando a propriedade de que as equações de evolução são invariantes para uma transformação linear de escala no relevo [17]. Assim, sem perda de generalidade, escrevemos $A_L = 1$, para a mestra e $A_P = (1 - s)$, para $P = 0, ..., L - 1$, com $0 < s < 1$. É fácil ver que recuperamos os mesmos resultados obtidos na análise do capítulo 2 (onde $A_L = 10$ e $A_P = 1$) fazendo-se $s = 0.9$. É instrutivo neste ponto revermos comparativamente as características dos dois relevos simples estudados até agora, a saber, o de um pico e o multiplicativo. Na figura (3.4)

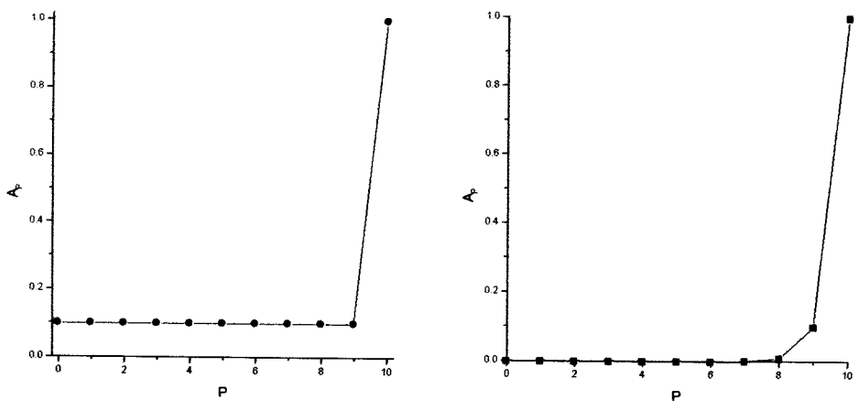

Figura 3.4: Desenho esquemático dos dois relevos de replicação com um pico simples. (a) Um pico, isolado dos outros genótipos com valores seletivos menores. (b) Multiplicativo que, apesar de também ter um pico, é mais suave. Aqui tomamos $s = 0.1$ e $L = 10$. As linhas foram adicionadas entre os pontos para uma melhor visualização.

mostramos esquematicamente esses dois relevos replicativos para ilustrar suas diferenças. Como já comentamos, os dois relevos representam casos extremos contendo um pico, produzindo aproximações locais de relevos mais gerais com muitos picos. O caso (a) corresponde a um relevo rugoso onde há picos isolados rodeados por um *mar* de genomas seletivamente fracos e equivalentes; e o caso representado em (b) corresponde a um relevo mais suave, no sentido de que o valor seletivo aumenta regularmente até os picos, em todas as direções. Neste último caso, a evolução sobre esse relevo é como a escalada de uma *montanha* suave em que podemos divisar o topo estando em qualquer ponto do relevo, enquanto que no primeiro caso não é possível saber onde o *topo da montanha* está, a não ser que estejamos exatamente sobre ele. É interessante observar ainda que estes dois relevos convergem a um relevo *plano* quando $s = 0$, que é o caso degenerado onde todas as seqüências têm o mesmo valor seletivo.

Antes de continuarmos com nossa análise, é importante fazer uma breve observação sobre os valores da taxa de mutação. É comum encontrar na literatura especializada no modelo de quase-espécies estudos do comportamento da população para $\mu > 0.5$ [4, 8, 28, 34, 42, 43, 45]. Realmente, Eigen e Schuster interpretam o intervalo $\mu \in [0.5, 1]$ como contendo, em geral, o regime de replicação estocástica e um outro regime, denominado de regime *complementar*, devido às oscilações das freqüências das seqüências complementares que aparecem na população, proporcionais à intensidade da taxa de mutação [4, 8]. Até aqui, entretanto, temos suposto implicitamente que a taxa de mutação não pode exceder o valor $\mu = 0.5$ para evitar preciosismos matemáticos em detrimento de uma análise mais realística. Esta observação seria desnecessária, não fosse o trabalho recente de Wiehe [28] que permite a taxa de mutação exceder esse limite, levando-o à conclusões matematicamente atraentes (como pontos de bifurcação, por exemplo) mas, no nosso entender, vazias de sentido biológico.

Voltando a análise das interações epistáticas, há diversos aspectos que devem ser frisados na generalização proposta em (3.17). Primeiro, o *fitness* da seqüência ótima é $A_L = 1$ para qualquer α, enquanto que o dos L primeiros mutantes é $A_{L-1} = 1 - s$ também para qualquer α. Assim, a epistase afeta apenas o *fitness* dos segundos mutantes em diante. Segundo, o limite $\alpha = 0$ corresponde ao relevo de um pico estudado no capítulo 2, no qual $A_L = 1$ e $A_{P \neq L} = 1 - s$. Terceiro, o caso $\alpha = 1$ corresponde ao relevo multiplicativo, no qual cada mutação contribui igualmente para o decréscimo do *fitness* do indivíduo. À medida que α decresce partindo de 1, o efeito do acúmulo das mutações passa a ser menos pronunciado (daí a denominação epistase atenuada) até alcançar o caso limite $\alpha = 0$, no qual tanto faz se ocorrer uma ou L mutações, já que o *fitness* de qualquer mutante será sempre $A_{P \neq L} = 1 - s$. Por outro lado, à medida que α cresce partindo de 1, o efeito acumulado das mutações contribui cada vez mais fortemente para o decréscimo do *fitness* dos mutantes (daí a denominação epistase *sinergística*, do grego *synergia* que significa cooperação). Finalmente, com o aumento de α, o *fitness* dos primeiros mutantes começa a se distanciar do fitness dos mutantes de ordens mais altas. Em particular, para $\alpha >> 1$ apenas a seqüência ótima e os primeiros mutantes são capazes de gerar descendentes; os outros indivíduos são gerados apenas através da replicação errônea das seqüências pertencentes a essas duas classes. Portanto, neste limite devemos obter $\overline{d} \approx \mu$ e $\sigma_d^2 \approx \mu(1-\mu)/L$ (esses resultados são obtidos supondo-se que $\Pi_P = M_{PL}$ para $P \neq L$).

Para se ter uma visão geral de como cada tipo de epistase afeta a estrutura da quase-espécie, mostramos na figura (3.5) o comportamento de duas grandezas relevantes do sistema, no caso em que a população é infinita. Na parte (a) da figura comparamos a variação da distância de Hamming média normalizada com a taxa de mutação para vários valores do parâmetro α, mantendo-se $L = 10$ e $s = 0.5$ fixos. As curvas foram obtidas iterando-se a equação (3.11) até alcançar

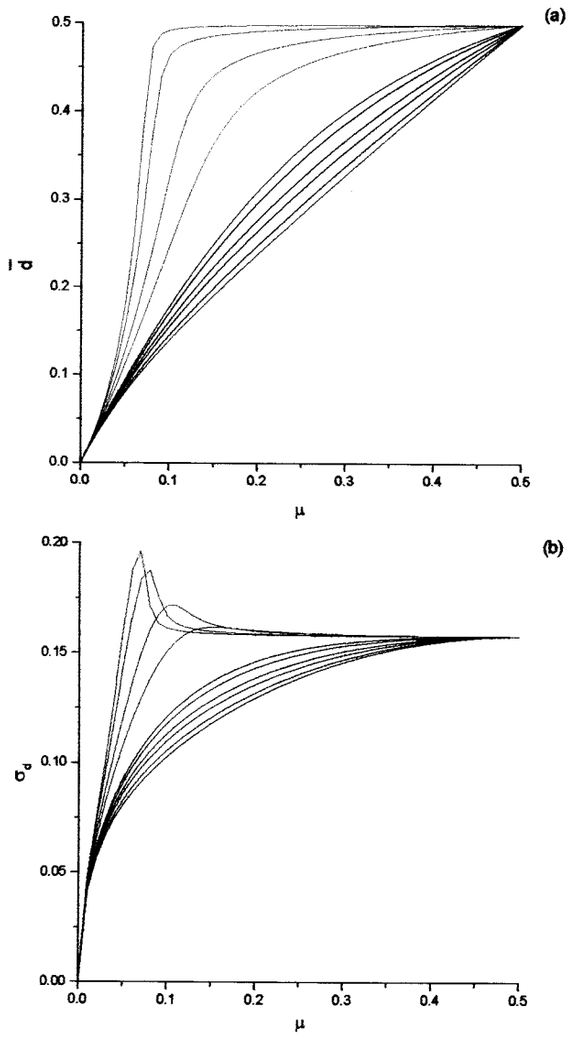

Figura 3.5: Relevo replicativo com epistase para vários valores de α. As figuras mostram, no estado estacionário, os resultados analíticos para $N \to \infty$, (a) a distância de Hamming média normalizada $\overline{d}$, (b) o desvio padrão σ_d de $\overline{d}$. Ambos os resultados são apresentados em função da taxa de mutação μ. Em cada gráfico temos, de cima para baixo: $\alpha = 0.01, 0.1, 0.3, 0.5$ (representando a epistase atenuante), $\alpha = 1$ (relevo sem epistase), $\alpha = 1.1, 1.3, 1.5, 2.0$ (para epistase sinergística). As cores distinguem os tipos de epistase, e em verde destacamos o caso em que $\alpha = 100$. Os parâmetros restantes são $L = 10$ e $s = 0.5$.

o equilíbrio de seleção. A distinção das cores das curvas ajuda a visualizar o conjunto de valores de α para cada tipo de interação. Assim, as curvas em azul claro representam os casos com epistase atenuante ($\alpha < 1$), a curva em vermelho é o caso em que não há epistase ($\alpha = 1$), e as em azul escuro representam os casos com epistase sinergística ($\alpha > 1$). No primeiro caso observamos o mesmo padrão de comportamento obtido na análise do relevo de um pico realizada no capítulo 2. A distância de Hamming tende rapidamente ao valor $\overline{d} = 0.5$ em $\alpha = 0.01$. Esse crescimento rápido é postergado para valores de μ cada vez maiores à medida que α aumenta. À medida que α cresce a partir de $\alpha = 1$ (que é o caso de relevo multiplicativo), as curvas afastam-se lentamente, convergindo assintoticamente para a curva em verde que representa um caso extremo em que $\alpha = 100$.

Os resultados para a distância de Hamming média discutidos acima são confirmados e melhor explicados na parte (b) da figura onde apresentamos o desvio padrão σ_d de $\overline{d}$. As curvas foram obtidas utilizando-se a relação (2.20) depois que o equilíbrio foi alcançado. As curvas em rosa corroboram a análise feita para a distância de Hamming média: para $\alpha < 1$ a população evoluindo nesse relevo tem que lidar com um limiar de erro de replicação, já que, pelo critério discutido no capítulo 2, a existência desse limiar é sinalizada pelo pico no desvio padrão σ_d. À medida que α cresce até $\alpha = 1$, o desvio padrão diminui e o seu máximo ocorre em valores da taxa de mutação cada vez maiores, indicando que o aumento da epistase protege a população de um limiar de erro precoce. A partir de $\alpha = 1$ não há limiar de erro, pois o máximo da variância ocorre apenas em $\mu = 0.5$, onde de qualquer maneira a população já está distribuída aleatóriamente.

Este estudo ainda pode ser estendido para qualquer valor seletivo s, permitindo-nos caracterizar em que condições aparece o limiar de erro na população. Com efeito, mostramos na figura (3.6) um estudo geral do fenômeno do limiar de erro em populações infinitas. Nessa figura apresentamos o limiar de erro (μ_c)

em função da intensidade da interação epistática (α). As curvas mostram como essa dependência modifica-se para vários valores do valor seletivo s. Assim, as curvas em azul marinho representam esse comportamento para $0.1 \leq s \leq 0.9$. Todos os resultados foram obtidos mantendo-se o tamanho da seqüência $L = 10$ constante. O primeiro resultado importante a se destacar é o valor do limiar de erro para $\alpha = 0$. Este é o caso do relevo de um pico estudado no capítulo

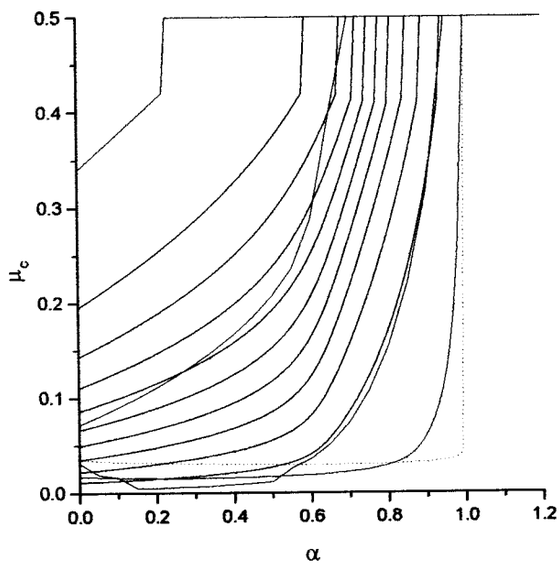

Figura 3.6: Limiar de erro μ_c em função da intensidade da epistase α em uma população infinita de genomas para vários valores do valor seletivo s. Em azul marinho temos (de baixo para cima) $s = 0.1, 0.2, ..., 0.9$. Em vermelho destaca-se dois limites para s: o de seleção muito grande, $s = 0.99$ (curva mais acima) e dois valores para indicar a tendência ao regime neutro (curvas mais abaixo) onde $s = 0.01$ (linha cheia) e $s = 0.001$ (linha pontilhada).

2, que com a mudança de escala nos permite descrever como o limiar de erro é afetado pela diferença entre as capacidades de produzir descendentes da mestra e dos mutantes que constituem o resto da população. O efeito de modificar L está ilustrado pelas linhas em verde que representam os casos extremos $s = 0.9$ e $s = 0.1$ para $L = 100$, indicando então que a variação de s produz menos efeitos

no caso de seqüências longas.

De uma maneira geral, à medida que aumentamos a epistase, mantido o valor de s fixo, o limiar de erro aumenta. Isso pode ser comprovado se acompanharmos o exemplo da figura (3.5) (b) para $s = 0.5$, onde podemos observar que ao aumentarmos α o pico de σ_d diminui e ocorre em valores de taxa de mutação cada vez maiores. É interessante notar que, em cada curva na faixa de valores seletivos $0.1 < s < 0.9$ apresentada na figura (3.6), podemos encontrar um valor de interação epistática atenuante, $\alpha = \alpha_{\max} < 1$, em que não existe limiar de erro. Em outras palavras, mesmo nesse caso em que a interação entre os genes é fraca (epistase atenuante), é possível prevenir o limiar de erro na população. Ainda, como esperado, observamos que para $\alpha > 1$ não há limiar de erro para $N \to \infty$ e este resultado é independente de s. Como a transição para $\alpha < \alpha_{\max}$ é de primeira ordem [15], pode-se pensar em $\alpha_{\max}$ como um ponto crítico convencional, no sentido de que sinaliza o término da transição de primeira ordem.

O mesmo estudo pode ser realizado para uma população finita, tomando-se o cuidado de lembrar que, dependendo do tamanho da população, o escape estocástico da seqüência ótima pode ocorrer para valores de taxa de mutação menores que o limiar de erro. Com efeito, na figura (3.7) apresentamos resultados similares aos da figura (3.6) para uma população composta de $N = 100$ genomas. Cada ponto das linhas em azul é a predição teórica para o máximo de σ_d em cada valor de α, obtida usando o nosso esquema de aproximação, onde iteramos a equação (3.4) até o equilíbrio, e posteriormente calculamos o máximo de σ_d definido em (2.20). Fazemos isso para cada curva representando um dado valor seletivo s. Analogamente, as curvas para as taxas de mutação onde ocorre o escape estocástico da seqüência ótima (em vermelho) são calculadas utilizando-se a relação $\overline{\Pi}_L = 1/N = 0.01$. Lembremos que esta condição dá apenas um limitante inferior para $\Pr(n_L = 0)$, entretanto, na discussão que segue vamos supor

que esse limitante seja o valor correto dessa probabilidade, esperando apenas que essa suposição resulte em resultados qualitativamente corretos. Assim, se acompanharmos as duas curvas para $s = 0.5$, por exemplo, podemos concluir que à medida que aumentamos α, o limiar de erro prevalece sobre o escape estocástico e a estrutura da quase-espécie é destruída devido aos efeitos da mutação. A partir

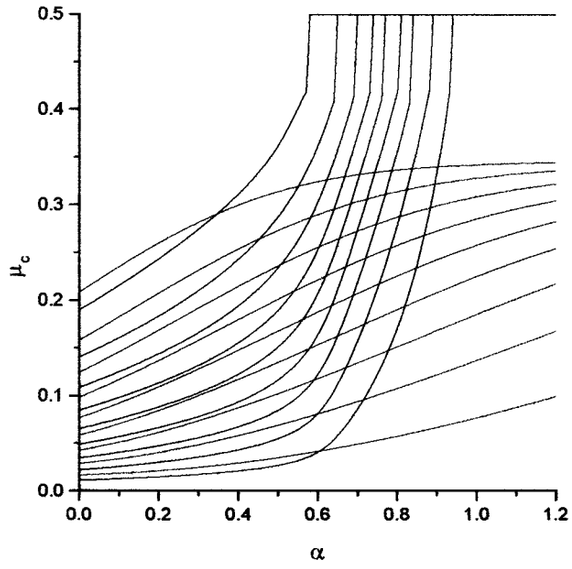

Figura 3.7: Estudo do limiar de erro (curvas em azul) e do escape estocástico da seqüência ótima (curvas em vermelho), em função da epistase para o caso $N = 100$. As curvas, de baixo para cima, representam valores seletivos crescentes $s = 0.1$, 0.2, ..., 0.9.

de $\alpha \approx 0.561$, o cruzamento entre as curvas azul e vermelha para esse valor de s, deveríamos esperar que o efeito da deriva genética fosse o responsável pela perda da seqüência ótima, como interpretado no capítulo 2, no caso $N = 10$. Entretanto, para $\alpha > 0$ essa perda (ou mesmo uma diminuição acentuada da freqüência da seqüência ótima) não é relevante para a quase-espécie. Realmente, fizemos algumas simulações para os casos $N = 10$ e $N = 100$, tentando encontrar o mesmo padrão observado nos itens (b) e (c) da figura (2.11). Isso só foi possível

no caso $N = 10$ para valores muito pequenos de $\alpha > 0$. De uma maneira geral é bastante difícil atribuir a desestruturação da quase-espécie à deriva genética para $\alpha > 0$. Para se ter uma visão mais clara dos resultados obtidos nesta seção, seria interessante verificá-los em um modelo que considere uma interação direta entre os genes e não apenas uma interação média como no modelo estudado aqui.

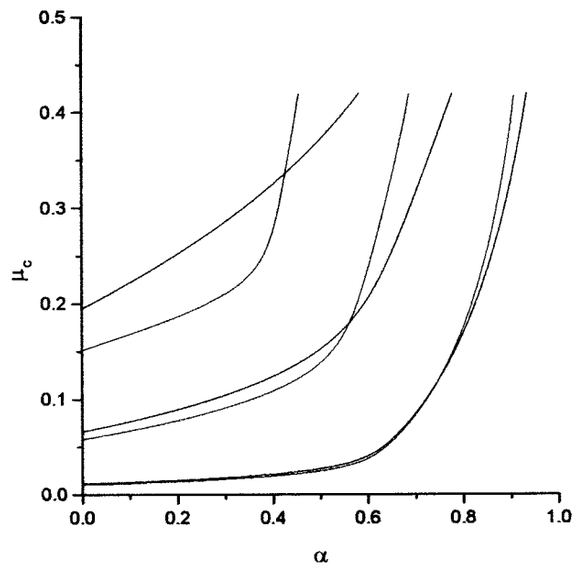

Figura 3.8: Comparação do limiar de erro para dois tamanhos de população $N = 10$ (em vermelho) e $N \to \infty$ (em azul), em função de α. De baixo para cima os pares de curvas são para $s = 0.1$, $s = 0.5$ e $s = 0.9$.

Para finalizar apresentamos na figura (3.8) uma comparação entre os limiares de erro para os casos $N \to \infty$ (em azul) e $N = 10$ (em vermelho) em função do parâmetro α, para três valores distintos de s. Na figura, as curvas terminam no ponto crítico $\alpha = \alpha_{\max}$, a partir do qual não há mais limiar de erro. Como havíamos concluído no capítulo 2 para o relevo de um pico, o efeito de população finita sobre a população é, em geral, antecipar o limiar de erro para valores de taxa de mutação menores. De fato, esse comportamento pode ser observado na figura para cada par de linhas, partindo-se de $\alpha = 0$ até o ponto onde as

linhas se cruzam. A partir desse ponto de cruzamento entretanto, a situação inverte-se, e uma população finita passa a postergar o aparecimento do limiar de erro em relação ao valor correspondente em populações infinitas. Esse resultado aparentemente desconcertante é outro sem prescedentes na literatura desde que era de se esperar que a deriva genética *sempre* atuasse no sentido de *enfraquecer* a seleção natural causando portanto uma deterioração precoce da população. Entretanto, esse aparente paradoxo, como também muitos outros na teoria de quase-espécies, deve-se à tentativa ingênua de conclusões obtidas no estudo do relevo de um pico para outros relevos de replicação. De fato, no relevo de um pico é natural esperar que μ_c diminua à medida que N aumenta, já que o efeito das flutuações torna-se mais importante para populações pequenas e a replicação deve então ser mais precisa para manter a mestra na população. Claramente, essa linha de raciocínio não se aplica no caso de um relevo suave, onde a ausência da seqüência ótima ou de seus primeiros mutantes simplesmente desloca o centro da quase-espécie no espaço de seqüências, sem destruir sua estrutura. Assim, para um relevo de replicação genérico não há como saber *a priori* se μ_c deve aumentar ou diminuir com o aumento da população. A figura (3.8) ilustra de forma inequívoca essa última afirmação.

Capítulo 4

Quase-espécies e genética de populações

O capítulo anterior encerrou a análise baseada no modelo de amostragens da evolução de organismos haplóides, representados por seqüências binárias, sujeitos estritamente à competição das forças evolucionárias da seleção natural e mutação. A generalização da teoria de quase-espécies bem como o controle de todos os parâmetros relevantes do sistema foi possível graças a robustez e simplicidade do modelo de amostragens.

Neste capítulo, estaremos interessados em entender outros aspectos da evolução molecular e mostrar como as idéias da teoria de quase-espécie são relevantes para modelos estudados em genética de populações, que têm sido resolvidos apenas em casos extremamente simples. Particularmente, vamos nos ater a dois problemas relacionados com modelos multi-locus, ou seja, genomas com várias posições possíveis para se alocar os alelos. Ora, a correspondência entre as formulações de quase-espécie e genética de populações é natural e realmente o modelo que tratamos até aqui para a evolução de indivíduos haplóides é, na linguagem da genética de populações, um modelo multi-locus (L-locus) com dois alelos [47]. Usualmente, os biólogos teóricos têm estudado modelos com um ou dois loci e que podem ser resolvidos em uma grande variedade de casos [7, 48]. Entretanto, a generalização para modelos multi-locus freqüentemente requer o uso de métodos numéricos [38, 44, 49, 50]. Vamos argumentar neste capítulo que o modelo de amostragens pode ser utilizado no ansatz da genética de populações como uma

ferramenta analítica poderosa, mesmo fazendo-se severas restrições no modelo descritas na seção seguinte.

Assim, primeiramente vamos considerar ainda o caso de organismos haplóides, todavia, permitindo que os genes assumam múltiplas formas em cada locus, ou seja, um modelo multi-locus com múltiplos alelos [43]. Ainda com relação a organismos haplóides, vamos estudar, no caso de dois alelos, a evolução dessa população em um relevo de replicação onde a seleção é truncada. Um outro estudo será realizado para investigar a evolução de uma população de indivíduos diplóides, que necessitam de duas seqüências binárias para se reproduzir [45]; os resultados para uma população haplóide sendo obtidos como um caso particular. Toda a análise neste capítulo estará restrita ao estudo de populações infinitas, já que estaremos interessados em obter o máximo de resultados possíveis que podem ser extraídos de nossa formulação, muitos deles não divisados no modelo original. No mais, estaremos atentos quando o efeito de população finita for relevante e, neste caso, daremos indicações de como se deveria proceder a análise.

4.1 A hipótese binomial

No âmbito da teoria da genética de populações devemos interpretar os resultados do capítulo 2 e 3 como sendo válidos para organismos caracterizados por seus genomas com L loci, cada locus podendo conter um alelo ótimo (o dígito 1, por exemplo) ou um alelo mutante (o dígito 0), que se reproduzem fazendo cópias de seus genomas [22]. Como vimos, o comportamento dessa população evoluindo em um relevo de replicação qualquer é descrito pela equação (3.4) que nos dá a freqüência de genótipos com P alelos ótimos a cada geração. Já que queremos tirar vantagem dos resultados da teoria de quase-espécie para avançar nessa área, vamos restringir inicialmente nossa análise ao relevo de replicação de um pico. Assim, a equação (3.4) pode ser reduzida a equação (2.15) reduzindo

a dependência da freqüência de genótipos ao valor seletivo do genótipo mestre, que é formado por L alelos do tipo 1. Em geral também, os estudos realizados em genética de populações concentram-se na freqüência de alelos de um tipo específico, o que em nosso modelo pode ser feito facilmente, escrevendo-se uma equação de recorrência para a freqüência de monômeros do tipo 1 na população, $\overline{p}(t) = 1/L \sum_P P\overline{\Pi}_P(t)$, a saber,

$$\overline{p}(t+1) = \mu + (1-2\mu)\left[\left[\overline{\Pi}_L(t)\right]^N + \sum_{n_L=0}^{N-1} B_{n_L} \frac{\overline{p}(t) - \overline{\Pi}_L(t) + a\frac{r}{1-r}\left(1-\overline{\Pi}_L(t)\right)}{1+r\,(a-1)}\right], \tag{4.1}$$

que pode ser facilmente obtida da equação (2.15). Como naquela equação B_{n_L} é dado por (2.16) e $r = n_L/N$. Claramente, esta equação não tem utilidade imediata, desde que devemos resolver a equação de recorrência (2.15) para achar $\overline{\Pi}_L(t)$. Entretanto, podemos obter uma equação fechada para $\overline{p}(t)$ fazendo o ansatz de que as freqüências de genomas sejam dadas por uma distribuição binomial da freqüência de alelos [43]

$$\overline{\Pi}_P(t) = \binom{L}{P} (p(t))^P \, (1-p(t))^{L-P}. \tag{4.2}$$

Dessa maneira a equação (4.1) irá envolver apenas a freqüência de alelos na geração t, já que de (4.2) $\overline{\Pi}_L(t) = [\overline{p}(t)]^L$. Assim, com essa hipótese simplificadora, a nossa formulação para a dinâmica de N indivíduos evoluindo no relevo de um pico passa a ser descrita por uma única equação de recorrência para a freqüência de alelos,

$$\overline{p}(t+1) = \mu + (1-2\mu)\left[[\overline{p}(t)]^{LN} + \sum_{n_L=0}^{N-1} B_{n_L} \frac{\overline{p}_t - [\overline{p}(t)]^L + a\frac{r}{1-r}\left(1-[\overline{p}(t)]^L\right)}{1+r\,(a-1)}\right], \tag{4.3}$$

onde agora ficamos com

$$B_{n_L} = \binom{N-1}{n_L} \left([\overline{p}(t)]^L\right)^{n_L} \left(1-[\overline{p}(t)]^L\right)^{N-1-n_L}. \tag{4.4}$$

Ainda, seguindo o mesmo procedimento utilizado para derivar a equação (2.17), é fácil escrever a versão determinística dessa equação. De fato, fazendo-se $N \to \infty$ em (4.3) chegamos a

$$\overline{p}(t+1) = \mu + (1-2\mu)\left[\frac{\overline{p}(t) + (a-1)\left[\overline{p}(t)\right]^L}{1 + (a-1)\left[\overline{p}(t)\right]^L}\right]. \tag{4.5}$$

Portanto, ao adotarmos a hipótese binomial, (4.2), estamos simplificando o modelo de amostragens, supondo que em cada geração a *sopa de genomas* dos pais seja decomposta por sua vez em uma sopa de alelos dos tipos 1 e 0 nas proporções $p(t)$ e $(1-p(t))$, respectivamente. Em outras palavras a população composta dos descendentes dos genomas presentes na geração $t-1$ é destruída e a freqüência de alelos presentes nessa população é então usada para criar uma nova população de acordo com (4.2).

Antes de prosseguirmos com nossa análise é importante discutir o efeito da hipótese binomial sobre as predições da distribuição de equilíbrio e a localização do limiar de erro. Na figura (4.1) (a) apresentamos a freqüência de genomas no equilíbrio de seleção, iterando-se a equação (4.5) para $L = 10$ e $a = 50$, como uma função da taxa de mutação μ. A freqüência inicial de alelos do tipo 1 na população é $p(0) = 1$. Como pode ser notado, utilizando-se a hipótese binomial obtemos um padrão de comportamento muito similar ao obtido com o modelo de amostragens, entretanto, a transição que caracteriza o limiar de erro é muito mais abrupta (na verdade descontínua), do que a prevista pelo modelo de amostragens original. Assim, no equilíbrio de seleção podemos caracterizar a população pelos pontos fixos da equação (4.5) $p(t+1) = p(t) = p^*$, que são as raízes de $f(p) = 0$, onde

$$f(p) = \mu\left[2p - 1 + (a-1)\,p^L\right] - (a-1)\,(1-p)\,p^L. \tag{4.6}$$

Para taxas de mutação pequenas (por exemplo, para $\mu < 0.239$ no caso da figura (4.1)) essa equação possui apenas uma raiz $p^* \approx 1$ que corresponde a um ponto

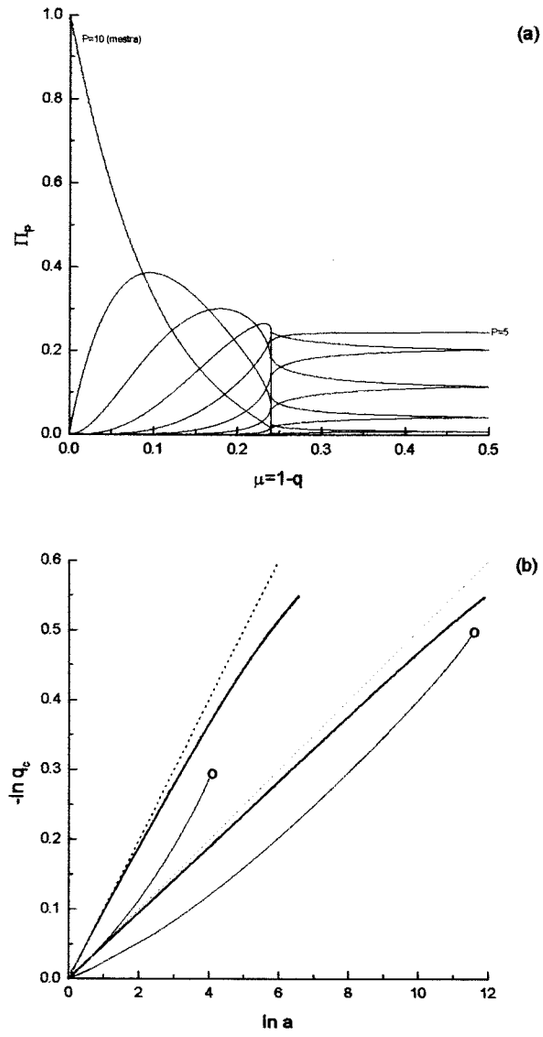

Figura 4.1: Efeito da hipótese binomial sobre a distribuição da população e sobre a localização do limiar de erro. (a) Freqüência de genomas no estado estacionário Π_P $(P = 0, ..., L)$ em função da taxa de mutação μ para $L = 10$ e $a = 50$. O limiar de erro ocorre para $\mu \approx 0.241$. (b) Precisão de replicação no limiar de erro no regime determinístico, em função da vantagem seletiva. Como na figura (2.5), as curvas em azul são obtidas usando a equação (2.16) e as em verde pontilhadas a equação (2.17). Acrescentamos a predição da equação (4.5) (curvas em vermelho). De cima para baixo essas linhas representam: $L = 10$ e 20.

fixo estável da equação de recorrência (4.4). A existência desse ponto fixo caracteriza o regime de quase-espécies. Acima desse valor, um segundo ponto fixo $p^* \approx 1/2$ surge e coexiste com o primeiro até que a taxa de mutação atinja o limiar de erro $\mu \approx 0.241$, quando então o ponto fixo $p^* \approx 1$ desaparece descontinuamente. A partir daí a população entra no regime de replicação estocástica ou uniforme caracterizado pelo ponto fixo $p^* \approx 1/2$. O limiar de erro nesse modelo, coincide com o aparecimento da raiz dupla de $f(p)$ e é facilmente determinado resolvendo-se $f(p) = 0$ e $df(p)/dp = 0$ para p e $\mu = \mu_c$ simultaneamente. Na região em que os dois pontos fixos estáveis coexistem, eles competem de tal modo que há uma *seleção* de apenas um deles. O ganhador, contudo, não é determinado pela taxa de replicação somente, mas também pela freqüência inicial de alelos $p(0)$. Vamos voltar a esse estudo na seção (4.3), onde estaremos analizando a evolução de uma população em um relevo de replicação com dois picos.

Verificamos também que a predição da equação (4.5) para a localização do limiar de erro produz resultados quantitativamente diferentes dos previamente obtidos. Isto pode ser observado na parte (b) da figura onde mostramos uma comparação entre as predições do modelo de amostragens (as curvas em azul são os mesmos resultados da figura (2.5)) e o resultado utilizando a hipótese binomial (curvas em vermelho) para dois valores distintos de L. Em particular, para um dado L o modelo com a hipótese binomial prevê um valor seletivo crítico no qual o limiar de erro deixa de ocorrer, ou seja, um ponto crítico (círculos) $a_{\max}$ sinalizando o final da transição. De uma maneira geral, observamos também que $\overline{\Pi}_P$ difere significativamente de uma distribuição binomial somente perto do limiar de erro. Um outro fato notável é que a hipótese binomial dá a solução exata para o regime determinístico no caso do relevo de replicação multiplicativo, sendo que no limite de seleção fraca as correções em $1/N$ da distância de Hamming entre o genoma ótimo e toda a população de mutantes podem ser calculadas

analiticamente [20]. No caso do relevo de um pico, o limiar de erro com essa simplificação, ocorre para valores de taxa de mutação muito menores do que as corretas, esse sendo o preço a pagar pela concisão dessa descrição. Mesmo assim, é importante frisar que essa versão simplificada do modelo de amostragens produz resultados qualitativos suficientemente confiáveis para podermos verificar, como uma primeira análise, o efeito de novos parâmetros e novos mecanismos no modelo de quase-espécies. Para ilustrar esse ponto, enquanto todas as análises prévias do modelo de quase-espécies lidaram exclusivamente com genomas binários, na próxima seção vamos tentar generalizar a equação (4.5) para estudar genomas mais complexos. Finalizando essa seção introdutória, deve-se mencionar que a hipótese binomial (4.2) é muito mais restritiva do que se necessita para fechar a equação de recorrência (4.1) no caso do relevo de um pico. De fato, precisamos apenas de uma relação entre $\overline{p}$ e $\overline{\Pi}_L$, sendo desnecessário o conhecimento de como $\overline{\Pi}_{P\neq L}$ depende de $\overline{p}$. Em princípio, qualquer relação satisfazendo $\overline{\Pi}_L = 0$ para $\overline{p} = 0$ e $\overline{\Pi}_L = 1$ para $\overline{p} = 1$ seria satisfatória e, sem dúvida, seria muito interessante estudar o comportamento estacionário das soluções da equação de recorrência para prescrições diferentes utilizadas nesta seção, a saber, $\overline{\Pi}_L = \overline{p}^L$.

4.2 Modelo multi-locus com múltiplos alelos

Vamos considerar os κ^L genomas que podem se formar quando estudamos a evolução de indivíduos com L loci; cada locus podendo assumir uma das κ formas diferentes de um certo gene (ou alelo). Assim, podemos supor que uma dada classe de genomas é caracterizada por um vetor $\mathbf{P} = (P_1, P_2, \ldots, P_\kappa)$, onde P_α é o número de alelos do tipo α em qualquer genoma dentro daquela classe. Desde que $\sum_\alpha^\kappa P_\alpha = L$, os κ^L genomas estão agrupados em $(L+\kappa-1)!/L!\,(\kappa-1)!$ classes, de acordo com o número de alelos que cada um possui, independente de sua posição no genoma. É fácil ver que para $\kappa = 2$ recuperamos o problema

tratado até aqui, no qual as 2^L seqüências binárias possíveis, representando os genomas, são agrupadas em $L+1$ classes. Como antes, vamos também supor que os genomas pertencentes a mesma classe sejam equivalentes, no sentido de que todos possuem o mesmo valor de *fitness*.

Dessa maneira, a hipótese simplificadora crucial dessa abordagem é a de que, conhecida a freqüência de alelos na geração t, $p_\alpha(t)$ (com $\sum_\alpha^\kappa p_\alpha(t) = 1$), as freqüências de genomas na classe $\mathbf{P}$ são dadas pela distribuição multinomial

$$\Pi_t(\mathbf{P}) = C_{\mathbf{P}}^L \left[p_1(t)\right]^{P_1} \left[p_2(t)\right]^{P_2} \ldots \left[p_\kappa(t)\right]^{P_\kappa} \tag{4.7}$$

onde $C_{\mathbf{P}}^L = L!/P_1!P_2!\ldots P_\kappa!$. Como antes, isto modifica o modelo de amostragens no sentido de que nossas conclusões sobre a freqüência de genomas são obtidas indiretamente a partir da freqüência de alelos.

Antes de incluirmos a hipótese (4.7) no modelo, vamos entender qual é o efeito isolado de passar de genomas binários para genomas com κ alelos. Com efeito, se nos restringirmos à evolução de uma população infinita, a equação de recorrência (3.10) caracteriza o modelo de amostragem em qualquer relevo de replicação, no caso em que os indivíduos têm L loci com dois alelos. Ela pode ser escrita em termos da frequência de monômeros como

$$\overline{p}(t+1) = \mu + (1-2\mu)\left[\frac{\sum\limits_{R=0}^{L} R\overline{\Pi}_R(t)A_R}{L\sum\limits_{K=0}^{L}\overline{\Pi}_K(t)A_K}\right]. \tag{4.8}$$

Devemos notar que ao escrevermos esta equação temos de calcular o número *médio* de mutações $\sum_{P=0}^{L} PM_{PR} = (1-\mu)R + \mu(L-R)$ (ver Apêndice A) de um genoma contendo R 1's e $(L-R)$ 0's, ou seja, o número médio de alelos do tipo 1 que replicam corretamente, $(1-\mu)R$, mais o número médio de alelos do tipo 0, que devido aos erros na replicação mutam a 1, $\mu(L-R)$. Ora, quando tratamos com κ tipos de alelos, se quisermos calcular a freqüência de alelos de um tipo particular α ou ainda a freqüência de seqüências com P_α alelos do tipo α, devemos levar

em consideração o aparecimento desse alelo devido a possibilidade de mutação de todos os outros tipos $\beta \neq \alpha$. É evidente que essa consideração introduz uma somatória em $\beta \neq \alpha$ para cada matriz de mutação na equação (3.11), tornando-a algebricamente incômoda, apesar de tratavel. Por outro lado, uma equação análoga a (4.8) pode ser trivialmente escrita para o caso em que temos κ tipos de alelos, já que o número *médio* de alelos α que replicam corretamente é $(1-\mu)P_\alpha$ e o número *médio* de alelos $\beta \neq \alpha$, que devido aos erros na replicação mutam a α, é simplesmente $[\mu/(\kappa-1)]\sum_{\beta\neq\alpha} P_\beta$. Assim, a equação para a frequência média de alelos do tipo α, utilizando o modelo de amostragem pode ser escrita como

$$\overline{p}_\alpha(t+1) = \frac{\mu}{\kappa-1} + \left(1 - \frac{\kappa\mu}{\kappa-1}\right)\left[\frac{\sum_{\mathbf{P}} P_\alpha \overline{\Pi}_{\mathbf{P}}(t) A_{\mathbf{P}}}{L\overline{w}(t)}\right] \tag{4.9}$$

onde $\overline{w}(t) = \sum_{\mathbf{P}} \overline{\Pi}_{\mathbf{P}}(t) A_{\mathbf{P}}$ é a taxa de replicação média de toda população na geração t. A notação $\sum_{\mathbf{P}}$ significa $\sum_{P_1=0}^{L} \cdots \sum_{P_\kappa=0}^{L} \delta(L, \sum_\alpha^\kappa P_\alpha)$, onde $\delta(i,j)$ é o delta de Kronecker. Mais uma vez, esta equação não tem utilidade imediata desde que necessitamos especificar $\overline{\Pi}_{\mathbf{P}}(t)$. Todavia, como antes, podemos obter uma equação fechada para a freqüência de alelos se incorporarmos a hipótese multinomial (4.7).

Então, o modelo descrito pela equação (4.9) é equivalente ao modelo clássico de genética de populações multi-loci com múltiplos alelos [22], exceto pelo mecanismo de mutação, que deve ser adaptado para satisfazer os vínculos impostos pela estrutura interna dos genomas. Mais ainda, como pode ser notado da seção anterior, esse formalismo pode ser estendido para tratar tanto populações finitas, como relevos de replicação, genéricos.

Daí, para obtermos um visão do comportamento do sistema, vamos especificar a taxa de replicação $A_{\mathbf{P}}$ de cada tipo de genoma, ou seja, vamos especificar o relevo de replicação. No caso do relevo de um pico devemos atribuir $A_{\mathbf{P}} = a$ se $\mathbf{P} = (L, 0, ..., 0)$ (o genoma mestre) e $A_{\mathbf{P}} = 1$ para todos os outros genomas.

Nesse caso, a equação (4.9) reduz-se a

$$\overline{p}_1(t+1) = \frac{\mu}{\kappa - 1} + \left(1 - \frac{\kappa\mu}{\kappa - 1}\right)\left[\frac{\overline{p}_1(t) + (a-1)\left[\overline{p}_1(t)\right]^L}{1 + (a-1)\left[\overline{p}_1(t)\right]^L}\right] \tag{4.10}$$

e

$$\overline{p}_\alpha(t+1) = \frac{\mu}{\kappa - 1} + \left(1 - \frac{\kappa\mu}{\kappa - 1}\right)\left[\frac{\overline{p}_\alpha(t)}{1 + (a-1)\left[\overline{p}_1(t)\right]^L}\right], \tag{4.11}$$

para $\alpha \neq 1$. Aqui nos aproveitamos da relação (4.7) para fazer $\overline{\Pi}_L(t) = \left[\overline{p}_1(t)\right]^L$ e simplificar a equação (4.9) analogamente ao realizado na seção anterior. Por simplicidade, manteremos a simetria entre os alelos do tipo $\alpha \neq 1$ fazendo suas freqüências iniciais iguais a $p_\alpha(0) = [1 - p_1(0)]/(\kappa - 1)$. Ainda, forçaremos a população inicial ser formada exclusivamente de genomas ótimos, $p_1(0) \approx 1$. Com essas simplificações podemos estudar a dependência do limiar de erro com o número de tipos de alelos κ, desde que esse parâmetro não introduz qualquer dificuldade analítica em nosso modelo. Realmente, na figura (4.2) apresentamos

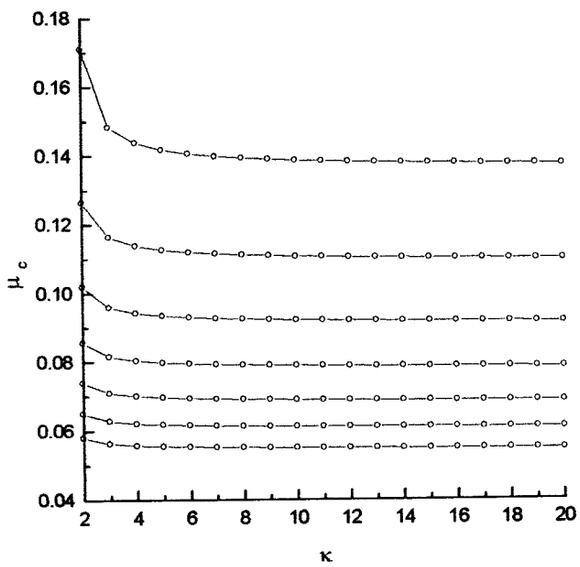

Figura 4.2: Limiar de erro μ_c em função do número de alelos do tipo κ para $a = 10$ e (de cima para baixo) $L = 8$, 10, 12, 14, 16, 18 e 20.

essa dependência do limiar de erro para $a = 10$ e vários valores de L. Para

calcular o limiar de erro nessa situação utilizamos a generalização natural da equação (4.6), para $\kappa \geq 2$

$$f(p) = \mu\left[\kappa p - 1 + (a-1)\,p^L\right] - (\kappa - 1)\,(a-1)\,(1-p)\,p^L, \tag{4.12}$$

e empregamos o mesmo procedimento para calcular μ_c. Assim, o padrão de comportamento das freqüências de alelos para $\kappa > 2$ é qualitativamente similar àquele discutido anteriormente, ou seja, os dois regimes da quase-espécie e o uniforme são caracterizados pelos pontos fixos $p_1^* \approx 1$ e $p_1^* \approx 1/\kappa$, respectivamente. Não notamos qualquer mudança significativa nessa figura ao variarmos a. É interessante notar que em torno de $\kappa \approx 4$ o limiar de erro é praticamente insensível ao aumento posterior de κ. Portanto, podemos interpretar esse resultado pensando que se desejarmos maximizar a quantidade de informação na quase-espécie, é vantajoso escolher κ o maior possível.

4.3 Outras aplicações da hipótese binomial

Nesta seção vamos mostrar os resultados de dois estudos realizados [43] utilizando o modelo de amostragem simplificado para a evolução das freqüências de alelos. O primeiro deles considera um relevo de replicação onde truncamos os valores seletivos para parte da população de maneira a tornarmos plana uma parte do relevo. Nesse caso estaremos interessados em entender o efeito dos parâmetros do relevo na evolução da população , similarmente ao estudo feito para relevos de um pico nos capítulos anteriores. No segundo estudo vamos investigar o comportamento dinâmico do modelo em um relevo de replicação com dois picos. No que segue vamos considerar o caso $\kappa = 2$ somente.

4.3.1 Seleção truncada

Até aqui temos estudado a evolução de uma população em um relevo de replicação com um pico pronunciado, no qual existe uma diferença abrupta entre

o valor seletivo atribuído ao genoma ótimo e o resto da população (relevo de um pico), ou ainda, em um relevo sem degenerescências (suave), em que atribuimos um valor seletivo crescente a cada mutante, dependendo de sua distância ao genoma ótimo (relevo multiplicativo). Passamos agora a discutir o comportamento da quase-espécie em um outro tipo de relevo de replicação que, além conter as características dos relevos mencionados, é uma generalização de um tipo de seleção truncada que foi recentemente introduzida na literatura [43]. Em particular, relevos com essa propriedade parecem desempenhar um papel importante na dinâmica evolucionária de seqüência repetitivas que aparecem nos eucariotas [28, 51], todavia vamos nos limitar somente à caracterização da quase-espécie e a localização do limiar de erro nesse relevo.

Assim, vamos supor que a taxa de replicação dos genomas aumente com o número de alelos do tipo 1 que ele possui. Mais especificamente,

$$A_P = a_0 + (a - a_0)\left(\frac{P-\rho}{L-\rho}\right)^{\gamma} \tag{4.13}$$

se os genomas tiverem $P \geq \rho$ alelos do tipo 1 e $A_P = a_0$ para os outros genomas da população. Aqui ρ é um inteiro que pode assumir os valores $0, 1, ..., L-1$ e γ é uma variável real positiva. Claramente, ρ mede o tamanho da região chata do relevo, ou seja, o ponto em que truncamos o relevo suave e forçamos todos os genomas terem o mesmo valor seletivo; enquanto γ determina o quão suave é o relevo. Notamos que, se fixarmos $\gamma = 1$, para $\rho = 0$ recuperamos o relevo multiplicativo sem a mudança de escala adotada no capítulo 3; à medida que aumentamos ρ a partir desse valor, o relevo vai ficando mais plano, até que em $\rho = L - 1$ recuperamos o relevo de um pico. Em particular, o parâmetro γ desempenha um papel similar ao da epistasis discutida no capítulo 3, no sentido de regular a suavidade do relevo. Para entender como esses dois parâmetros atuam, apresentamos na figura (4.3) os resultados de um estudo da evolução da

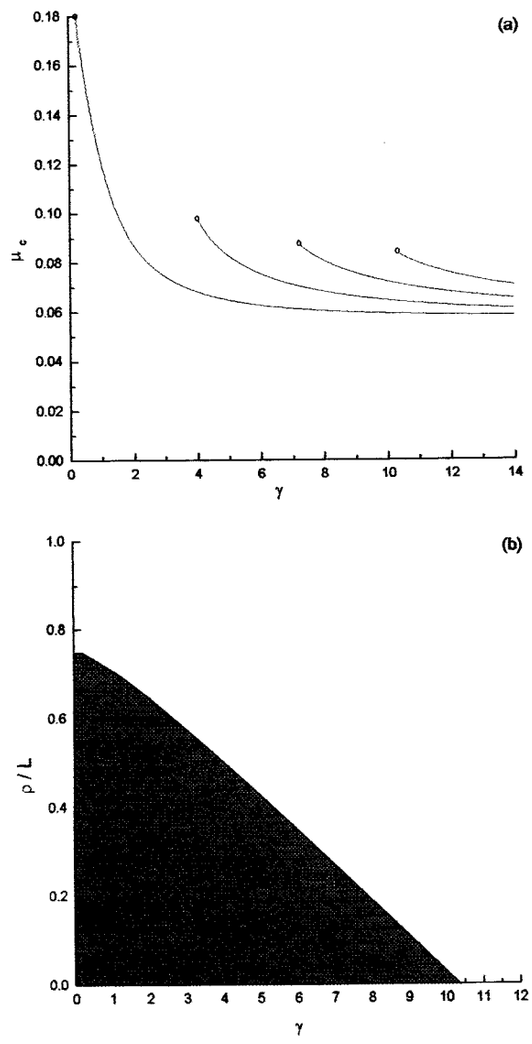

Figura 4.3: (a) Limiar de erro $\mu_c$em função do expoente γ (da esquerda para direita) $\rho/L = 0.75, 0.5, 0.25$ e 0. As curvas começam em $\gamma_{\max} = \gamma_{\max}(\mu)$. (b) Valor crítico do expoente γ em função da razão ρ/L (em azul escuro) e região onde o limiar de erro não ocorre (área em azul claro). Em ambos os gráficos fizemos $L = 20$ e $a = 10$.

população nesse relevo no caso em que $L = 20$, $a = 10$ e $a_0 = 1$. Na parte (a) da figura, mostramos o limiar de erro como uma função do expoente γ para vários valores de ρ. Obviamente, os pontos das curvas foram obtidos utilizando-se a equação (4.9) no equilíbrio de seleção, através da qual podemos derivar uma equação similar à (4.12) para os pontos fixos e usar a mesma prescrição para encontrar a localização do limiar de erro. Como pode ser notado, à medida que aumentamos γ, para um valor de ρ fixo, a taxa de mutação onde ocorre o limiar de erro diminui até ficar aproximadamente constante e insensível ao aumento de γ. Assim, enquanto o aumento da intensidade da interação epistática α protege a população do limiar de erro, o aumento de γ antecipa o aparecimento desse fenômeno. É interessante notar também da parte (a) da figura que existe um valor crítico $\gamma_{\max}$, abaixo do qual o limiar de erro não ocorre. Esse valor crítico do expoente é mostrado na parte (b) da figura em função da razão ρ/L. Fica claro dessas duas figuras que quanto menor o truncamento do relevo (ρ pequeno), mais a população pode resistir a catástrofe de erro ou até mesmo evitá-la, dependendo do valor do expoente γ (no caso $\gamma = 1$ e $\rho = 0$ recuperamos o relevo multiplicativo no qual não há limiar de erro). Valores grandes de γ, todavia, aumentam o tamanho e importância da região plana e portanto podem favorecer o aparecimento do limiar de erro. Fizemos um estudo análogo para diferentes valores de L e a e observamos que isso não altera qualitativamente os resultados. Observamos também que esse cenário é em geral o mesmo para populações finitas.

4.3.2 Dois picos estreitos

Vamos considerar aqui um relevo de replicação contendo dois picos, ou seja, dois genomas (no caso, distantes geneticamente um do outro) com valores seletivos maiores que os do resto da população. Dessa maneira podemos fazer $A_{P=0} = A_0$, $A_{P=L} = A_L$ e $A_{P\neq 0'L}$ e estudar qual é a influência da freqüência inicial de alelos

no equilíbrio de seleção da população de genomas. Isto é ilustrado na figura (4.4) onde mostramos a freqüência do alelo do tipo 1 em função do número de gerações t para $L = 20$, $A_0 = 200$, $A_{20} = 10$, começando com várias freqüências iniciais. A evolução para $\mu = 0$ é apresentada na parte (a) da figura, onde pode ser observada a existência de dois pontos fixos estáveis $p^* = 1$ e 0. Apesar da grande diferença entre as taxas de replicação dos genomas associados a esses pontos fixos, suas bacias de atração são praticamente do mesmo tamanho (elas devem ser estritamente iguais se $A_0 = A_{20}$). O principal efeito de uma taxa de replicação muito grande é aumentar a velocidade de convergência para o ponto fixo de maior *fitness*. Aumentando-se a taxa de mutação, um novo ponto fixo $p^* = 1/2$ aparece, associado ao regime de replicação estocástico. O efeito conjunto dos três pontos fixos estáveis é mostrado na parte (b) da figura para $\mu = 0.01$. Portanto, para um taxa de mutação diferente de zero, a bacia de atração do menor ponto fixo é consideravelmente maior que a do ponto fixo maior. Naturalmente, suas bacias de atração diminuíram se comparadas com o caso $\mu = 0$. Como pode ser notado, as duas quase-espécies não coexistem, ou seja, para uma dada população inicial, apenas um dos pontos fixos é selecionado. Finalmente, na parte (c) da figura mostramos a evolução para $\mu = 0.06$. Nesse caso, o maior ponto fixo, associado ao genoma com a menor taxa de replicação, desaparece e o ponto fixo associado com o regime de replicação estocástica toma conta de sua bacia de atração. Se continuarmos a aumentar a taxa de mutação, o ponto fixo menor irá eventualmente desaparecer também.

Utilizando o mesmo procedimento, nós podemos facilmente estudar a competição entre um pico estreito e outro largo [8]. Os resultados mostram as mesmas características qualitativas das que foram apresentadas acima. Em particular, desde que um pico largo possui um limiar de erro maior (ver figura 4.3 (a)) ele irá desempenhar um papel similar ao descrito para o pico com maior taxa de

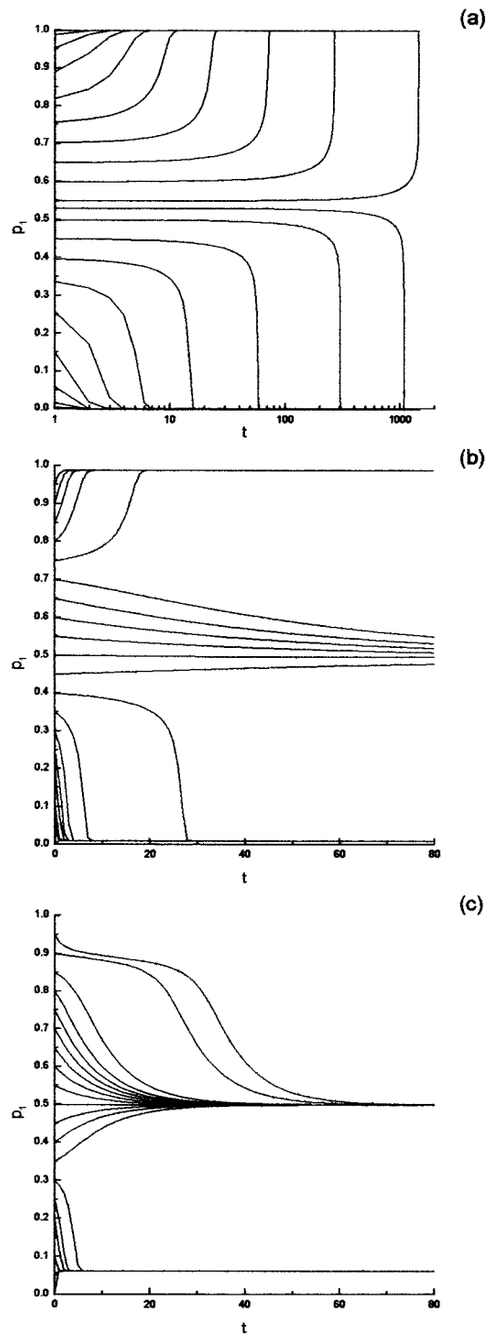

Figura 4.4: Freqüência de alelos do tipo 1 em função do número de gerações para um relevo de replicação com dois picos estreitos e várias freqüências iniciais $p(0)$. Os parâmetros são $L = 20$, $A_0 = 200$, $A_{20} = 10$, para três valores de taxa de mutação: (a) $\mu = 0$, (b) $\mu = 0.01$ e (c) $\mu = 0.06$.

replicação.

4.4 Quase-espécies e a evolução de organismos diplóides

Até agora temos restringido nossa análise a evolução de organismos haplóides, ou seja, organismos que *auto-replicam* seu único genoma, com ou sem erro, sendo que o número de cópia geradas dependende da capacidade reprodutiva do genoma *auto-replicador*. Nas seções seguintes vamos estender alguns dos resultados apresentados até aqui para organismos que necessitam de dois genomas para se reproduzir: organismos diplóides. Em particular, como já comentado, vamos explorar as possibilidades do modelo de amostragem simplificado para lidar com a evolução desses organismos. Antes disso, é importante esclarecer que tipo de reprodução será considerado nesta seção. Vamos estudar o tipo de reprodução conhecida na literatura como *partenogênese* (de parteno + gênese que significa aparecimento não fecundado), ou seja, estaremos estudando organismos diplóides que se reproduzem assexuadamente, simplesmente copiando todo o seu genoma diplóide [47]. Esse fenômeno também é conhecido como *apomixia* [52]. Os descendentes são portanto idênticos aos pais, a menos de alguma mutação que possa ter ocorrido durante o processo de reprodução.

Dessa maneira fica fácil estudar a evolução de organismos diplóides, utilizando o ansatz de que a freqüência de genomas é uma binomial da freqüência de alelos do tipo 1, dada por (4.2). Vamos por simplicidade considerar o caso $\kappa = 2$ apenas. Assim, se $A_{P^iP^j}$ denotar o valor seletivo dos genótipos P^iP^j, isto é, genótipos de um organismo diplóide compostos de qualquer par de seqüências com L loci P^i e P^j (os gametas haplóides), então a fração de alelos do tipo 1 que os genótipos P^iP^j contribuem para a geração $t+1$ pode ser calculada de maneira idêntica ao caso haplóide. Daí, se quisermos estudar a evolução de uma população infinita desses

indivíduos, a freqüência da alelos do tipo 1 a cada geração será proporcional a três fatores: (a) sua freqüência na população $\Pi_{P^iP^j}(t)$, (b) seu valor seletivo $A_{P^iP^j}$, e (c) o número médio de alelos 1 que se replicam corretamente, $(1-\mu)(P^i+P^j)$, mais o número médio de alelos que mutam a 1, $\mu[2-P^i-P^j]$. Como no caso de organismos haplóides, um cálculo simples produz a seguinte relação para a evolução temporal da freqüência de alelos 1,

$$\overline{p}(t+1) = \mu + (1-2\mu)\left[\frac{\sum_{P^i}\sum_{P^j}(P^i+P^j)\,\overline{\Pi}_{P^iP^j}(t)A_{P^iP^j}}{2L\overline{w}(t)}\right], \tag{4.14}$$

onde agora o *fitness* médio da população passa a ser dado por

$$\overline{w}(t) = \sum_{P^i}\sum_{P^j}\overline{\Pi}_{P^iP^j}(t)A_{P^iP^j}. \tag{4.15}$$

Podemos dizer então, que a relação (4.14) é a generalização do modelo de genética de populações [44] multi-locus com dois alelos para uma população diplóide, supondo partenogênese como tipo de reprodução.

Para simplificar nossa análise, vamos supor que os organismos diplóides sejam formados por encontros aleatórios entre os gametas haplóides, ou seja,

$$\overline{\Pi}_{P^iP^j}(t) = \overline{\Pi}_{P^i}(t)\overline{\Pi}_{P^j}(t). \tag{4.16}$$

Enquanto que essa relação de independência estatística é como hipótese *ad hoc* necessária em outros estudos similares [42, 44, 47], em nosso modelo ela pode ser facilmente modificada para levar em conta situações mais complicadas de reprodução.

Para continuar com nossa análise, devemos especificar o valor seletivo dos genótipos diplóides P^iP^j. Para isso, iremos adotar o seguinte relevo de replicação, proposto por Wiehe *et al* [42], que é o análogo do relevo de um pico estreito para organismos diplóides,

$$A_{P^iP^j} = \begin{cases} (1+a)^2 & \text{se } P^i = P^j = L \\ (1+a)^{2h} & \text{se } P^i = L \text{ e } P^j \neq L \\ 1 & \text{se } P^i \neq L \text{ e } P^j \neq L \end{cases} \tag{4.17}$$

onde $a > 0$ é o parâmetro que mede a vantagem seletiva do gameta mestre $P = L$, $-\infty < h < \infty$ é o coeficiente de dominância, ou simplesmente dominância. Daí, o gameta mestre é completamente dominante para $h = 1$ e completamente recessivo para $h = 0$. Para $h = 1/2$ achamos que $A_{P^iP^j} = A_{P^i}A_{P^j}$ e portanto não há dominância. Nesse caso, a equação (4.14) reduz-se a equação que governa a evolução de organismos haplóides reproduzindo-se assexuadamente, descrita pela equação (4.5). Então os intervalos $h \in [0, 1/2)$ e $h \in (1/2, 1]$ delimitam as regiões de recessividade e de dominância, respectivamente, do gameta mestre. Há outros casos de interesse também que são freqüentemente estudados em genética de populações: $h > 1$ modela o fenômeno da heterose ou vigor híbrido (vantagem do genótipo heterozigoto), enquanto que $h < 0$ modela o fenômeno que ocorre nos primeiros estágios da especiação, quando os híbridos são menos viáveis (desvantagem do heterozigoto). Podemos, portanto, inserir a equação (4.17) na equação de recorrência (4.14) e derivar a seguinte equação para a frequência média de alelos do tipo 1 na geração t,

$$\overline{p}(t+1) = \mu + (1-2\mu)\left[\frac{\Lambda_1\left[\overline{p}(t)\right]^{2L} + \Lambda_2\left(\overline{p}(t)+1\right)\left[\overline{p}(t)\right]^{L} + \overline{p}(t)}{\Lambda_1\left[\overline{p}(t)\right]^{2L} + 2\Lambda_2\left[\overline{p}(t)\right]^{L} + 1}\right] \tag{4.18}$$

onde

$$\Lambda_1 = (1+a)^2 - 2(1+a)^{2h} + 1 \tag{4.19}$$

e

$$\Lambda_2 = (1+a)^{2h} - 1. \tag{4.20}$$

Na figura (4.5) apresentamos a freqüência dos gametas haplóides no estado estacionário, obtida resolvendo-se a equação de recorrência com $p(0) \approx 1$, como função da taxa de mutação μ para $L = 10$ e $a = 2$, e diferentes valores do coeficiente de dominância. No caso de replicação sem erros ($\mu = 0$), o ponto fixo $p^* = 0$ é sempre instável, enquanto que $p^* = 1$ é estável para $h \leq 1$ somente. Para $h > 1$ um terceiro ponto fixo (estável) $1/2 < p^* \approx 1$ aparece, sinalizando a

emergência da heterose. Para $h \leq h_{\max} \approx 1.75$ há dois regimes distintos: o regime

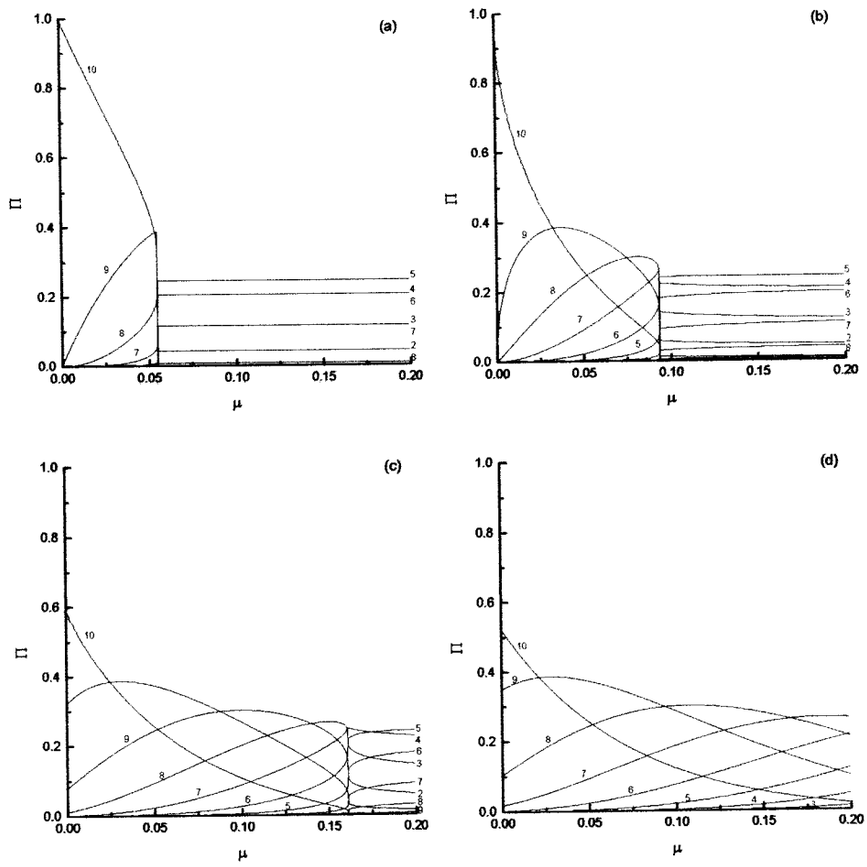

Figura 4.5: Freqüência de gametas haplóides no estado estacionário pertencentes às classes $P = 10$ (gameta mestre) a $P = 0$ em função da taxa de mutação por alelo μ para $L = 10$ e (a) $h = 0$, (b) $h = 1$, (c) $h = 1.5$, e (d) $h = 2$.

de quase-espécies caracterizado pela população dominada pelo gameta mestre e seus vizinhos geneticamente próximos, e o regime uniforme onde os 2^L gametas aparecem na mesma proporção (claramente, a classe $P = L/2$ é a mais favorecida nesse caso). Então, os mesmos fenômenos observados na teoria de quase-espécies para organismos haplóides também ocorrem na evolução de organismos diplóides. Em particular, à medida que h aumenta, o tamanho do salto da transição diminui até desaparecer no ponto crítico $h = h_{\max}$. Perto desse valor já não é possível

distinguir os dois regimes. Esse fenômeno do desaparecimento do limiar de erro em um valor crítico $h_{\max}$ é idêntico ao observado quando analisamos a epistase no capítulo 3. Realmente, o parâmetro h aqui desempenha o mesmo papel que a epistase α, que caracteriza a interação entre os alelos, já que o coeficiente de dominância determina o grau de interação dos gametas constituíntes do genoma, sendo que o efeito de ambos na população é o mesmo.

4.5 O efeito da dominância sobre o limiar de erro

Para melhor caracterizar a transição que define o limiar de erro vamos concentrar nossa análise na natureza dos pontos fixos $\overline{p}(t+1) = \overline{p}(t) = p^*$. Mais uma vez o procedimento é o mesmo do caso haplóide, no qual eles são dados pelas raízes reais de $f(p) = 0$, que neste caso fica

$$f(p) = \Lambda_1 \left(p + \mu - 1\right) p^{2L} + \Lambda_2 \left(p + 2p\mu - 1\right) p^L - \mu \left(1 - 2p\right). \qquad (4.21)$$

Assim, para valores pequenos de taxa de mutação esta equação tem somente uma raiz real que corresponde ao ponto fixo estável $p^* \approx 1$ associado ao regime de quase-espécies. À medida que μ aumenta, uma raíz dupla aparece, originando dois novos pontos fixos: um estável, $p^* \approx 1/2$, associado ao regime uniforme, e um instável que delimita a bacia de atração dos pontos fixos estáveis. Esses pontos fixos coexistem até a taxa de mutação atingir o limiar de erro μ_c, onde o ponto fixo que caracteriza a quase-espécie e o instável coalescem. Como antes, podemos determinar facilmente a transição do limiar de erro resolvendo-se $f(p) = df(p)/dp = 0$ simultaneamente para p e $\mu = \mu_c$. Podemos além disso, determinar o ponto crítico $h_{\max}$ analiticamente, fazendo o ajuste fino do valor de h de modo que as três raízes reais de (4.21) coincidam, ou seja, resolvemos as três equações $f(p) = df(p)/dp = d^2f(p)/dp^2 = 0$ simultaneamente para p, $\mu = \mu_c$ e $h = h_{\max}$.

Com essas prescrições apresentamos na figura (4.6) o limiar de erro em função de

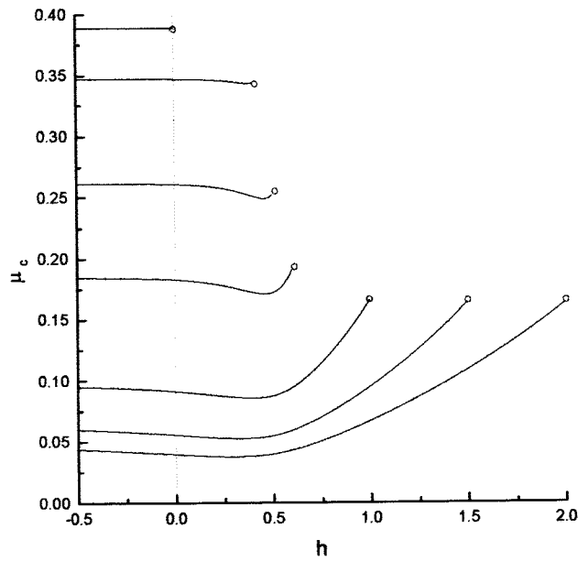

Figura 4.6: Limiar de erro μ_c como função do coeficiente de dominância h para $L = 10$ e (de cima para baixo) $a = 314.8$, 186.1, 57.0, 18.8, 4.4, 2.0 e 1.3. O parâmetro a foi escolhido de maneira que as linha de transição terminem nos pontos críticos localizados em $h = 0$, 0.4, 0.5, 0.6, 1.0, 1.5 e 2.0 respectivamente.

h para $L = 10$ e vários valores de a. O limiar de erro μ_c é praticamente constante para variações de h quando este parâmetro é negativo ou positivo mas pequeno, alcançando seu valor mínimo em torno de $h = 0.5$ (regime de não dominância) e então aumentando rapidamente á medida que o sistema entra na região de dominância, $h > 0.5$. Esse fato é interessante, já que indica que um gameta recessivo ($h < 0.5$) pode ser melhor que um não dominante ($h \approx 0.5$). Notamos também um comportamento reentrante dessas linha de transição, ou seja, para um certo valor de μ fixo, o sistema sofre duas transições descontínuas à medida que h cresce. Realmente, esse fenômeno, que também pode ser visto como uma bi-estabilidade da quase-espécie (ou ao menos de dois regimes ordenados para dois valores do coeficiente de dominância), tem sido observado na literatura [42, 53]

inclusive em estudos do efeito do mecanismo de recombinação em populações virais [54].

Portanto, o principal efeito da dominância é postergar o limiar de erro, a um preço de reduzir a concentração do gameta mestre na população. Entretanto, a dominância não afeta somente a localização do limiar de erro, mas também ajusta a variabilidade da população. Ambas análises, a localização do limiar de erro e a composição da população, indicam que a dominância permite ao gameta mestre resistir a taxas de mutação maiores que no caso de não dominância. Realmente, para h suficientemente grande pode-se evitar completamente a catástrofe de erro. Esse resultado tem sido proposto como uma possível explicação para o fato notável de que o gameta predominante na população (o melhor adaptado ao ambiente) é frequentemente dominante: os gametas dominantes podem ser os predominantes simplesmente porque eles podem tolerar taxas de mutação mais altas [42].

Capítulo 5

Conclusões e perspectivas

Reservamos este capítulo para as considerações finais e aproveitamos para introduzir de maneira resumida, algumas questões interessantes que, baseados em estudos preliminares, acreditamos possam ser investigadas utilizando o modelo de amostragens. Antes dessas colocações, entretanto, devemos lembrar alguns aspectos importantes e gerais aprendidos da competição entre as duas principais forças evolucionárias estudadas aqui. Em geral, a seleção tende a aumentar a freqüência do genótipo mais apto, enquanto que a mutação tende a diminuí-la. Existe um limite superior para a taxa de mutação que pode ser tolerada pela população, sendo que em torno desse valor, a seleção não pode compensar o influxo mutacional e a mutação passa a ser a força dominante sobre a população acarretando em uma deterioração genética. A seleção natural necessita de um número e uma variedade suficiente de seqüências para que possa operar, de outra maneira, essa *força* é por si muito fraca para atuar na evolução dos organismos. Assim, se a quase-espécie evolui em um ambiente extremamente especializado, como é o caso, por exemplo, do relevo de um pico, a seleção terá dificuldades de aprimorar os indivíduos da população, já que seus genomas estarão limitados em conteúdo de informação. O intuito de se colocar a competição entre seleção e mutação dessa maneira é banalizar a idéia de que a seleção natural depende dos vínculos impostos ao sistema como tamanho da população, tipo e forma de relevo adaptativo, etc..., e portanto qualquer conclusão geral que se queira inferir sobre

essa competição só é possível controlando-se tais vínculos.

Com efeito, nesta monografia ocupamo-nos em argumentar que o modelo de amostragens é suficientemente simples e geral para podermos avançar no entendimento de um sistema darwiniano. Realmente, para um sistema estritamente sujeito a seleção e mutação, fomos capazes de verificar o quadro apresentado acima, em regiões do espaço de parâmetros onde a formulação original de quase-espécies não tem acesso [1, 4, 2], como é o caso de populações finitas [9], onde nossa análise ressalta a importancia desse efeito sobre a estrutura da quase-espécie e e torna relativa a defendida generalidade do fenômeno do limiar de erro [34]. Mais ainda, a nossa generalização para diferentes relevos de replicação mostra que, a limitação à evolução molecular atribuida a esse fenômeno, aparece somente em casos particulares de relevos e, se somado ao efeito da deriva genética enfatiza, que ele é um mecanismo de catraca de Muller [36] iminente, possivelmente não uma crise de informação. Outro problema com o limiar de erro que tentamos ressaltar é a falta de uma definição geral desse fenômeno [28]. Até onde nos foi possível, acompanhamos e discutimos os resultados de outros modelos existentes na literatura, onde muitas dessas questões já haviam sido levantadas porém não satisfatoriamente respondidas, ou porque a análise era proibitiva no espaço de parâmetros completo como são os casos da formulação original de Eigen [1, 14, 15] e da sua primeira extensão para populações finitas [27]; ou porque a formulação, por construção, ficou restrita ao estudo de algum limite particular [21, 28, 32]; ou ainda, porque a aproximação era apenas uma caricatura do modelo original [11, 46].

Ainda, com relação ao espaço de parâmetros estudado nessa monografia, podemos avançar com a análise do modelo de amostragens permitindo que alguns parâmetros de controle variem. Assim, podemos permitir que o tamanho da população N varie, considerando inclusive a situação em que N obedece uma dinâmica própria. Quando isso acontece observa-se uma deterioração da popu-

lação que a leva a sua rápida extinção [55]. Outro parâmetro que foi suposto constante em nosso modelo é o comprimento dos genomas L. Poderíamos, por exemplo, permitir *inserções* e/ou *deleções* de um pequeno número de bases. Embora a maioria das vezes essas alterações sejam letais para o organismo há um número grande de exemplos na literatura envolvendo *inserções* e *deleções* em mútiplos de três nucleotídeos (em particular no HIV isso tem sido observado em virtualmente todos os estudos) [56, 57]. Particularmente, Ebeling [24] considera inserções em sua extensão do modelo de quase-espécies a populações finitas, ficando restrito a um relevo replicativo suave e aditivo.

Nossa análise também foi limitada a relevos de replicação simples e que permanecem constantes no tempo. O relevo de replicação de uma espécie biológica real é muito mais complexo do que os apresentados aqui, além de estarem constantemente mudando no tempo. No caso, por exemplo, de um relevo rugoso estático com muitos ótimos locais, se usarmos o algoritmo de amostragens é possível reproduzir resultados similares aos encontrados por Bonhoffer e Stadler [58] para esses tipos de relevos correlacionados, todavia ainda é necessário refinar essa análise para investigar esse tipo de relevo. Uma versão mais realística do modelo de amostragens, entretanto, pode ser obtida através do estudo da evolução do *fitness* médio da população em função do tempo. Essa dinâmica temporal está bem documentada em uma série de experimentos recentes [59] sobre a evolução de vírus de RNA controlada em laboratório, tendo sido modelada satisfatoriamente utilizando-se um formalismo de *campo médio* para descrever a evolução dessas populações virais [60]. Porém este modelo está restrito a um relevo de replicação suave e unidimensional, além de conter algumas simplificações drásticas na determinação das taxas de transições entre os genomas. Os resultados analíticos derivados com o modelo de amostragem [61] são bastante superiores aos da versão original do modelo [60] e indicam que é possível extrapolar os resultados para o

caso em que o relevo de replicação é flutuante, isto é, dependente do tempo.

Além de seleção e mutação, outros mecanismos contribuem ativamente para a diversidade genética de várias populações de organismos e que podem também ser incorporados no modelo [57, 56]. O primeiro deles é a *recombinação* que é o mecanismo pelo qual os organismos misturam seu material genético permitindo uma rápida *troca* de *características*. Este fenômeno é estudado em geral para modelar reprodução sexual [62] e organismos diplóides [45], mas tem sido demonstrado ocorrer na replicação *in vivo* para alguns grupos de vírus, tornando-se importante para a biologia desses organismos. Para lidar com esse mecanismo no contexto da teoria da quase-espécies onde a reprodução é assexuada, vamos nos basear em um modelo proposto recentemente [54] no qual cada partícula viral carrega duas cópias de seu genoma e o genoma recombinante é produzido misturando-se várias partes dessas cópias. Esse tipo de troca de material pode ser incorporado facilmente no algoritmo de amostragens, colocando-se entre os operadores de seleção e mutação um operador de troca de material (operador de *crossover*) comumente utilizado na literatura de algoritmos genéticos [18]. Uma diferença importante entre recombinação sexual e assexual, é que esta última depende da densidade da população, de maneira que o evento de recombinação é mais freqüente à medida que a população aumenta. Entretanto, a recombinação pode evitar a extinção de mutantes vantajosos devido a variações bruscas no relevo replicativo e também devido a deriva genética, desde que as populações são finitas e inicialmente pequenas. Este fato, é extremamente relevante a luz dos resultados obtidos nesta monografia para a dependência do limiar de erro com o tamanho da população.

Talvez o mecanismo mais interessante que pode ser incorporado no modelo de amostragens seja o de *hipermutação*. Para populações virais, por exemplo, este processo é definido como a capacidade desses organismos de aumentarem

a sua taxas de mutação em apenas um ciclo de replicação [56]. Este fenômeno tem sido observado para vários tipos de vírus [57] e, em particular, para o HIV, a *hipermutação* é utilizada para explicar a preponderância de substituições de bases do tipo G por bases do tipo A [56]. Este mecanismo pode ser visto como uma das estratégias possíveis que um organismo dispõe para otimizar sua *prospecção genética* lançando mão de *genes mutadores*. Aqui o *limiar de erro* deve desempenhar um papel fundamental, já que abaixo, mas perto desse valor, as condições para evolução são otimizadas: a seqüência mestra é mais estável estando presente um número máximo de mutantes na distribuição. Isto sugere que podemos trabalhar com um procedimento do tipo *annealing*, ultrapassando o limiar de erro por um período curto de tempo e assim aumentando a *velocidade* da evolução. Obviamente, violações constantes do limiar de erro podem causar deteriorização da quase-espécie. Assim, uma nova mutação pode aparecer aleatoriamente, cujos benefícios podem sobrepujar o custo total de todas as mudanças deletérias (é claro que essas medidas drásticas não garantem os efeitos desejados; já que as mudanças podem ser todas deletérias). Desde que é bem sabido que uma dinâmica com passos variáveis acelera o processo de otimização, especialmente em um meio não estacionário [63], é razoável conjecturar que a evolução biológica já tenha descoberto e incorporado este princípio de otimização.

As possibilidades de continuidade deste trabalho são muitas e, realmente, o encaramos apenas como um começo promissor de uma área de pesquisa mais abrangente. Com efeito, as questões delineadas acima têm uma motivação especial, a saber, viabilizar a formulação proposta nesta monografia para o estudo de populações virais.

Os vírus podem ser vistos como *programas genéticos* com uma única mensagem para a célula hospedeira: - *reproduza-me*! Todavia, outras características atraem a atenção nesses *quase-organismos*. Em particular, as populações são formadas

por uma distribuição de mutantes largamente dispersa no espaço de genomas, ao invés de uma população homogênea de sequências do tipo preponderante (*wild type*), mostrando que esse tipo de organização é possível com uma alta taxa de erro de replicação, quando comparadas com a de organismos autônomos. Suas taxas de mutação são adaptadas ao tamanho do seu genoma e, desde que a adaptação depende da distribuição de mutantes, vírus diferentes podem mostrar estruturas de população bastante diferentes, mesmo possuindo genomas do mesmo tamanho. Realmente, as taxas de erro de replicação têm sido determinadas experimentalmente para vários tipos de vírus, sendo que todos os resultados mostram uma correlação entre taxa de mutação e tamanho do genoma [57, 64].

Talvez o exemplo mais importante de uma quase-espécie viral seja o HIV. Pacientes infectados com este vírus abrigam uma população viral extremamente diversa, com muitos mutantes diferentes. As mutações nesse caso são geradas na codificação da enzima transcriptase reversa do vírus, que é produzida com uma taxa de erro da ordem de 10^{-4} a 10^{-1} por base, ou seja, durante cada replicação do genoma, ocorrem de 1 a 10 erros; número este suficientemente grande para que o HIV opere muito perto do *limiar de erro de replicação.* Como pode ser observado então, para os vírus de uma maneira geral, a substituição do termo *espécie* por *quase-espécie* não é simplesmente semântica: uma *quase-espécie viral* é uma população viral que se *auto-perpetua,* composta de entidades diversas mas relacionadas, atuando como um todo.

Todavia, tanto a reprodução quanto a interpretação de como a mutação opera para gerar as seqüências devem ser repensadas tomadas cuidadosamente, levando a algumas modificações no modelo de amostragens original. Particularmente, com relação ao mecanismo de replicação dentro da célula hospedeira, devemos introduzir uma pressão de seleção adicional devido ao processo de translação e empacotamento acontecendo em seus aparatos de replicação e reparo. Aqui torna-

se importante a distinção entre rodadas de replicação e ciclos de infecção, já que em cada ciclo ocorrem muitas rodadas de replicação onde a populacão viral dentro da célula hospedeira divide a mesma maquinaria de replicação e translação. Esta modificação é importante para modelar o fato que é observado depois da *explosão* da célula hospedeira devido a proliferação dos vírus: das milhares de partículas virais resultantes, apenas uma fração pequena é infecciosa.

Também, a freqüência das mutações, isto é, a probabilidade de que uma enzima erre no momento de incorporar uma base em uma certa posição, tem que ser distingüida da freqüência de mutantes, dada pela proporção de certos mutantes na população. Por exemplo, um *sítio quente* (hot spot) é uma região no genoma com uma freqüência de mutantes muito grande, podendo ser gerado de dois modos: ou a taxa de mutação é muito alta nessas posições, ou a seleção favorece (ou tolera) variações nessa região. Neste segundo caso, as freqüências de mutantes e de mutação são completamente diferentes. Portanto, contar mutantes em uma distribuição de quase-espécies não é o modo correto de determinar a taxa de mutação, pois esta última é muito mais uniforme. É particularmente importante ter isso em mente desde que frequentemente os dados de seqüenciamento são interpretados de diferentes modos, devido a falta de aceitação das definições acima.

Apêndice A: Matriz de mutação

Dada uma seqüência binária de comprimento L, com R dígitos do tipo 1 e $L-R$ dígitos do tipo 0, podemos calcular a probabilidade de que esta seqüência sofra mutação de maneira que a seqüência resultante contenha P dígitos do tipo 1 e, portanto, $L-P$ dígitos do tipo zero. Para isso, definimos primeiramente μ como a probabilidade de um dígito do tipo 1 mutar para um dígito do tipo 0 e vice-versa. Assim, a probabilidade de que Q dígitos do tipo 1, entre os R presentes na seqüência, não sofram mutação e que os $R-Q$ restantes mutem para o dígito 0 é dada por

$$\Pr(Q \mid R) = \binom{R}{Q} (1-\mu)^{Q} \, \mu^{R-Q}. \tag{A.1}$$

Devemos também calcular a probabilidade de que uma parte S, dos $L-R$ dígitos do tipo 0 restantes da seqüência, mutem para dígitos do tipo 1. A probabilidade de que isso ocorra é dada de maneira similar por

$$\Pr(S \mid L-R) = \binom{S}{L-R} \mu^{S} \, (1-\mu)^{L-R-S}. \tag{A.2}$$

Dessa maneira, se esses dois eventos ocorrerem simultaneamente na seqüência de L dígitos, o número de dígitos do tipo 1 na seqüência resultante será simplesmente $P = Q + S$ como pode ser visto no esquema abaixo. A probabilidade de que uma seqüência com P dígitos do tipo 1 seja formada é então

$$M_{PR} = \Pr(P \mid R) = \sum_{Q=0}^{R} \Pr(Q \mid R) \sum_{S=0}^{L-R} \Pr(S \mid L-R) \, \delta_{S,P-Q}$$

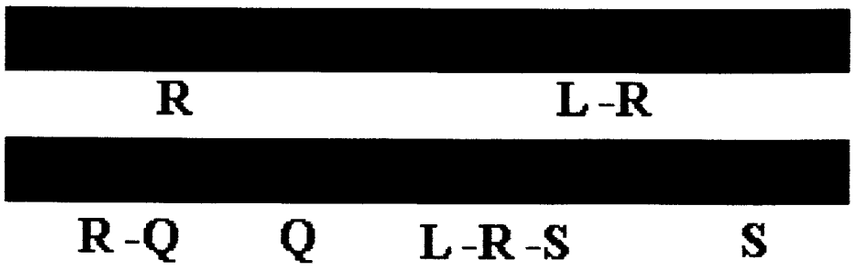

Figura A.1: Esquema da mutação sofrida por uma seqüência de comprimento L com R $1's$.

$$= \sum_{Q=0}^{R} \sum_{S=0}^{L-R} \delta_{S+Q,P} \binom{R}{Q} (1-\mu)^{Q} \mu^{R-Q} \binom{L-R}{S} \mu^{S} (1-\mu)^{L-R-S}, \qquad \text{(A.3)}$$

onde $\delta_{i,j}$ é o delta de Kronecker. O vínculo para S, $S = P - Q$, impõe restrições a Q, de modo a garantir a existência dos coeficientes binomiais. Assim, $Q_l \leq Q \leq Q_u$, onde $Q_l = \max(0, P + R - L)$, desde que $Q \leq R$ e $P - Q \leq L - R$, e $Q_u = \min(P, R)$, desde que $0 \leq Q$ e $0 \leq P - Q$. O resultado final é a matriz de transição

$$M_{PR} = \sum_{Q=Q_l}^{Q_u} \binom{R}{Q} \binom{L-R}{P-Q} (1-\mu)^{L-P-R+2Q} \mu^{P+R-2Q}, \qquad \text{(A.4)}$$

que é a matriz de mutação (2.4) utilizada no texto.

É intrutivo também verificarmos a normalização da matriz de mutação (A.4), $\sum_P M_{PR} = 1$. Para isso usamos (A.4), fazendo

$$\sum_P M_{PR} = \sum_P \sum_{Q=0}^{R} \sum_{S=0}^{L-R} \delta_{S+Q,P} \binom{R}{Q} (1-\mu)^{Q} \mu^{R-Q} \binom{L-R}{S} \mu^{S} (1-\mu)^{L-R-S}$$

$$= \sum_{Q=0}^{R} \sum_{S=0}^{L-R} \underbrace{\sum_P \delta_{S+Q,P}}_{=1} \binom{R}{Q} (1-\mu)^{Q} \mu^{R-Q} \binom{L-R}{S} \mu^{S} (1-\mu)^{L-R-S}$$

$$= \underbrace{\sum_{Q=0}^{R} \binom{R}{Q} (1-\mu)^{Q} \mu^{R-Q}}_{=1} \underbrace{\sum_{S=0}^{L-R} \binom{L-R}{S} \mu^{S} (1-\mu)^{L-R-S}}_{=1} = 1.$$

Outro cálculo importante é o número médio de mutações a partir de uma seqüência com R $1's$ definido como

$$\sum_P PM_{PR} = \sum_P \sum_{Q=0}^{R} \sum_{S=0}^{L-R} P\delta_{S+Q,P} \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} \binom{L-R}{S} \mu^S (1-\mu)$$

$$= \sum_{Q=0}^{R} \sum_{S=0}^{L-R} \underbrace{\sum_P P\delta_{S+Q,P}}_{=S+Q} \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} \binom{L-R}{S} \mu^S (1-\mu)^{L-R-S}$$

$$= \sum_{Q=0}^{R} Q \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} + \sum_{S=0}^{L-R} S \binom{L-R}{S} \mu^S (1-\mu)^{L-R-S}.$$

Usando o procedimento usual para calcular média sobre uma distribuição binomial, obtemos

$$\sum_P PM_{PR} = R(1-\mu) + (L-R)\mu = \mu L + (1-2\mu) R. \tag{A.5}$$

Podemos também calcular o segundo momento dessa grandeza, de maneira análoga, fazendo

$$\sum_P P^2 M_{PR} = \sum_P \sum_{Q=0}^{R} \sum_{S=0}^{L-R} P^2 \delta_{S+Q,P} \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} \binom{L-R}{S} \mu^S (1-\mu)^{L-R-S}$$

$$= \sum_{Q=0}^{R} \sum_{S=0}^{L-R} \underbrace{\sum_P P^2 \delta_{S+Q,P}}_{=(S+Q)^2} \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} \binom{L-R}{S} \mu^S (1-\mu)^{L-R-S}$$

$$= \sum_{Q=0}^{R} Q^2 \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} + \sum_{S=0}^{L-R} S^2 \binom{L-R}{S} \mu^S (1-\mu)^{L-R-S}$$

$$+2 \left[\sum_{Q=0}^{R} Q \binom{R}{Q} (1-\mu)^Q \mu^{R-Q} \right] \left[\sum_{S=0}^{L-R} S \binom{L-R}{S} \mu^S (1-\mu)^{L-R-S} \right].$$

Mais uma vez, utilizando os truques usuais para calcular a média e o segundo momento de uma distribuição binomial podemos simplificar a expressão acima, obtendo

$$\sum_P P^2 M_{PR} = \mu(1-\mu) L + [\mu L + (1-2\mu) R]^2. \tag{A.6}$$

Para completar essa seqüência de derivações, vamos escrever finalmente a variância do número médio de mutações a partir de uma seqüência com P $1's$ que dá uma medida da largura da distribuição. De fato, utilizando (A.5) e (A.6) escrevemos

$$\sigma_R^2 = \frac{1}{L^2}\left[\sum_P P^2 M_{PR} - \left(\sum_P P M_{PR}\right)^2\right] = \frac{\mu(1-\mu)}{L}, \tag{A.7}$$

que, curiosamente não depende de R.

Apêndice B: Derivação das equações do modelo de amostragem

A derivação da equação (2.10) é realmente bastante simples. Partindo da equação (2.6)

$$\begin{aligned}\overline{n}_P(t+1) &= \sum_{\mathbf{n}}\sum_{\mathbf{n}'}\sum_{\{n''_{PR}\}}\sum_{R} n''_{PR}\mathcal{P}_m\left(\{n''_{PR}\}\mid\mathbf{n}'\right)\mathcal{P}_s\left(\mathbf{n}'\mid\mathbf{n}\right)\mathcal{P}_{\Pi}\left(\mathbf{n}\right)\\ &= \sum_{\mathbf{n}}\sum_{\mathbf{n}'}\sum_{R}\underbrace{\sum_{\{n''_{PR}\}} n''_{PR}\mathcal{P}_m\left(\{n''_{PR}\}\mid\mathbf{n}'\right)}_{=\overline{n}''_{PR}}\mathcal{P}_s\left(\mathbf{n}'\mid\mathbf{n}\right)\mathcal{P}_{\Pi}\left(\mathbf{n}\right), \quad \text{(B.1)}\end{aligned}$$

calculando o somatório mais interno:

$$\begin{aligned}\overline{n}''_{PR} &= \sum_{\{n''_{PR}\}} n''_{PR}\mathcal{P}_m\left(\{n''_{PR}\}\mid\mathbf{n}'\right)\\ &= \sum_{\{n''_{PR}\}} n''_{PR}\frac{n'_R!}{n''_{0R}!\,n''_{1R}!\ldots n''_{LR}!}M_{0R}^{n''_{0R}}M_{1R}^{n''_{1R}}\ldots M_{LR}^{n''_{LR}}\\ &= M_{PR}\frac{\partial}{\partial M_{PR}}\left(M_{0R}+\ldots+M_{LR}\right)^{n'_R}\\ &= n'_R M_{PR}\left(\underbrace{M_{0R}+\ldots+M_{LR}}_{=1}\right)^{n'_R-1}\\ &= n'_R M_{PR}. \quad \text{(B.2)}\end{aligned}$$

Assim, ficamos com

$$\overline{n}_P(t+1) = \sum_{\mathbf{n}}\sum_{R} M_{PR}\underbrace{\sum_{\mathbf{n}'} n'_R\mathcal{P}_s\left(\mathbf{n}'\mid\mathbf{n}\right)}_{=\overline{n}'_R}\mathcal{P}_{\Pi}\left(\mathbf{n}\right). \quad \text{(B.3)}$$

Analogamente, calculamos a média

$$\overline{n}'_R = \sum_{\mathbf{n}'} n'_R\mathcal{P}_s\left(\mathbf{n}'\mid\mathbf{n}\right)$$

$$
\begin{aligned}
&= \sum_{\mathbf{n}'} n'_R \frac{N!}{n'_0!\, n'_1! \ldots n'_L!} \left[W_0(\mathbf{n})\right]^{n'_0} \left[W_1(\mathbf{n})\right]^{n'_1} \ldots \left[W_L(\mathbf{n})\right]^{n'_L} \\
&= W_R \frac{\partial}{\partial W_R} \left(W_0 + \ldots + W_L\right)^N \\
&= N W_R \left(\underbrace{W_0 + \ldots + W_L}_{=1}\right)^{N-1} \\
&= N W_R. \qquad \text{(B.4)}
\end{aligned}
$$

Com a aproximação de campo médio, substituimos Π_P por $\overline{\Pi}_P$. Dessa maneira obtemos a equação de recorrência (2.10),

$$
\overline{\Pi}_P(t+1) = \frac{\overline{n}_P(t+1)}{N} = \sum_{\mathbf{n}} \sum_{R} M_{PR} W_R(\mathbf{n}) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}). \qquad \text{(B.5)}
$$

Podemos também calcular o segundo momento das freqüências das moléculas, definido por

$$
\begin{aligned}
\overline{n_P^2}(t+1) &= \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} \left[n''_P\right]^2 \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\
&= \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} \left[\sum_R n''_{PR}\right]^2 \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\
&= \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} \left[\sum_R n''_{PR}\right] \left[\sum_S n''_{PS}\right] \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\
&= \sum_R \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} \left[n''_{PR}\right]^2 \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) + \\
&\quad \sum_R \sum_{S \neq R} \sum_{\mathbf{n}} \sum_{\mathbf{n}'} \sum_{\{n''_{PR}\}} n''_{PR} n''_{PS} \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) \mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}). \qquad \text{(B.6)}
\end{aligned}
$$

Vamos calcular primeiramente as médias em $\mathcal{P}_m$ das duas parcelas dessa equação. Na primeira fazemos

$$
\sum_{\{n''_{PR}\}} \left[n''_{PR}\right]^2 \mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) = \sum_{\{n''_{PR}\}} \left[n''_{PR}\right]^2 \frac{n'_R!}{n''_{0R}! \ldots n''_{LR}!} M_{0R}^{n''_{0R}} \ldots M_{LR}^{n''_{LR}}
$$

$$
\begin{aligned}
&= \sum_{n''_{P0}} \ldots \sum_{n''_{PL}} \left[n''_{PR}\right]^2 \begin{pmatrix} n'_0 \\ n''_{00} \ldots n''_{L0} \end{pmatrix} \prod_{P=0}^{L} M_{P0}^{n''_{P0}} \ldots \begin{pmatrix} n'_0 \\ n''_{0L} \ldots n''_{LL} \end{pmatrix} \prod_{P=0}^{L} M_{PL}^{n''_{PL}} \\
&= \underbrace{\left(\sum_{P=0}^{L} M_{P0}\right)^{n'_0}}_{=1} \ldots M_{PR} \frac{\partial}{\partial M_{PR}} \left(M_{PR} \frac{\partial}{\partial M_{PR}} \left(\sum_{P=0}^{L} M_{PR}\right)^{n'_R}\right) \ldots \underbrace{\left(\sum_{P=0}^{L} M_{PL}\right)^{n'_L}}_{=1}
\end{aligned}
$$

$$= M_{PR}\left[n'_R \underbrace{\left(\sum_{P=0}^{L} M_{PR}\right)^{n'_R-1}}_{=1} + M_{PR}n'_R\left(n'_R-1\right)\underbrace{\left(\sum_{P=0}^{L} M_{PR}\right)^{n'_R-2}}_{=1}\right]$$

$$= M_{PR}n'_R + M^2_{PR}\left[n'_R\right]^2 - M^2_{PR}n'_R. \tag{B.7}$$

Analogamente, calculamos

$$\sum_{\{n''_{PR}\}} n''_{PR}n''_{PS}\mathcal{P}_m\left(\{n''_{PR}\} \mid \mathbf{n}'\right) = \sum_{\{n''_{PR}\}} n''_{PR}n''_{PS}\frac{n'_R!}{n''_{0R}!\ldots n''_{LR}!}M_{0R}^{n''_{0R}}\ldots M_{LR}^{n''_{LR}}$$

$$= \sum_{n''_{P0}}\ldots\sum_{n''_{PL}} n''_{PR}n''_{PS}\begin{pmatrix} n'_0 \\ n''_{00}\ldots n''_{L0}\end{pmatrix}\prod_{P=0}^{L} M_{P0}^{n''_{P0}}\ldots\begin{pmatrix} n'_0 \\ n''_{0L}\ldots n''_{LL}\end{pmatrix}\prod_{P=0}^{L} M_{PL}^{n''_{PL}}$$

$$= \underbrace{\left(\sum_{P=0}^{L} M_{P0}\right)^{n'_0}}_{=1}\ldots M_{PR}\frac{\partial}{\partial M_{PR}}\left(\sum_{P=0}^{L} M_{PR}\right)^{n'_R}\ldots M_{PS}\frac{\partial}{\partial M_{PS}}\left(\sum_{P=0}^{L} M_{PS}\right)^{n'_S}\ldots\underbrace{\left(\sum_{P=0}^{L} M_{PL}\right)^{n'_L}}_{=1}$$

$$= M_{PR}n'_R M_{PS}n'_S. \tag{B.8}$$

Voltando-se com os resultados de (B.7) e (B.8) em (B.6), ficamos com

$$\begin{aligned}\overline{n^2_P}(t+1) &= \sum_R\sum_{\mathbf{n}}\sum_{\mathbf{n}'}\left(M_{PR}n'_R + M^2_{PR}\left[n'_R\right]^2 - M^2_{PR}n'_R\right)\mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right)\mathcal{P}_{\overline{\Pi}}\left(\mathbf{n}\right) \\ &\quad + \sum_R\sum_{S\neq R}\sum_{\mathbf{n}}\sum_{\mathbf{n}'}\left(M_{PR}M_{PS}n'_R n'_S\right)\mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right)\mathcal{P}_{\overline{\Pi}}\left(\mathbf{n}\right).\end{aligned} \tag{B.9}$$

Mais uma vez temos que efetuar as médias, uma a uma, das mudanças ocorridas no número de ocupação devido ao processo de seleção. Entretanto, de (B.2) e (B.4), as médias envolvendo n'_R são triviais e resta-nos fazer:

$$\begin{aligned}\sum_{\mathbf{n}'}\left[n'_R\right]^2\mathcal{P}_s\left(\mathbf{n}' \mid \mathbf{n}\right) &= \sum_{\mathbf{n}'}\left[n'_R\right]^2\frac{N!}{n'_0!\,n'_1!\ldots n'_L!}\left[W_0\left(\mathbf{n}\right)\right]^{n'_0}\ldots\left[W_L\left(\mathbf{n}\right)\right]^{n'_L} \\ &= W_R\frac{\partial}{\partial W_R}\left[W_R\frac{\partial}{\partial W_R}\left(W_0+\ldots+W_L\right)^N\right]\end{aligned}$$

$$= W_R\left[N\underbrace{\left(W_0+\ldots+W_L\right)^{N-1}}_{=1} + W_R N\left(N-1\right)\underbrace{\left(W_0+\ldots+W_L\right)^{N-2}}_{=1}\right]$$

$$= W_R N + W^2_R N\left(N-1\right), \tag{B.10}$$

e ainda

$$\begin{aligned}\sum_{\mathbf{n}'} n'_R n'_S \mathcal{P}_s(\mathbf{n}' \mid \mathbf{n}) &= \sum_{\mathbf{n}'} n'_R n'_S \frac{N!}{n'_0!\, n'_1! \ldots n'_L!} [W_0(\mathbf{n})]^{n'_0} \ldots [W_L(\mathbf{n})]^{n'_L} \\ &= W_R \frac{\partial}{\partial W_R}\left[W_S \frac{\partial}{\partial W_S}(W_0 + \ldots + W_L)^N\right] \\ &= W_R W_S N(N-1) \underbrace{(W_0 + \ldots + W_L)^{N-2}}_{=1} \\ &= W_R W_S N(N-1). \end{aligned} \tag{B.11}$$

Assim, juntando-se esses resultados em (B.9) e rearranjando termos, ficamos com

$$\begin{aligned}\overline{n_P^2}(t+1) &= \sum_{\mathbf{n}}\left[N\sum_R M_{PR}W_R\right]\mathcal{P}_{\overline{\Pi}}(\mathbf{n}) + \\ &\quad N(N-1)\sum_{\mathbf{n}}\left[\sum_R M_{PR}^2 W_R^2 + \sum_R \sum_{S\neq R} M_{PR}M_{PS}W_R W_S\right]\mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\ &= \overline{n}_P(t+1) + N(N-1)\sum_{\mathbf{n}}\left(\sum_R M_{PR}W_R\right)^2 \mathcal{P}_{\overline{\Pi}}(\mathbf{n}).\end{aligned}$$

Finalmente, podemos escrever a equação de recorrência para o segundo momento da frequência de sequência com P $1's$ como

$$\overline{\Pi_P^2}(t+1) = \frac{\overline{n_P^2}(t+1)}{N^2} = \frac{\overline{\Pi}_P(t+1)}{N} + (N-1)\sum_{\mathbf{n}}\left(\sum_R M_{PR}W_R\right)^2 \mathcal{P}_{\overline{\Pi}_P}(\mathbf{n}).$$

Apêndice C: Equações para o relevo de um pico

Partindo da equação (2.13),

$$\overline{\Pi}_P(t+1) = \sum_{\mathbf{n}} \left(\frac{\sum\limits_{R \neq L} n_R M_{PR} + n_L a M_{PL}}{N + (a-1)n_L} \right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}), \tag{C.1}$$

vamos agora calcular os somatórios sobre os números $n_0, ..., n_{L-1}$. Para fazer isso devemos escrever a somatória em $\mathbf{n}$ em termos da função probabilidade marginal da distribuição mutinomial para o conjunto n_L. Daí,

$$\begin{aligned}
\sum_{\mathbf{n}} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) &= \sum_{n_0+\ldots+n_L=N} \frac{N!}{n_0!\, n_1! \ldots n_L!} \overline{\Pi}_0^{n_0} \ldots \overline{\Pi}_L^{n_L} \\
&= \sum_{n_L=0}^{N} \left[\sum_{n_0+\ldots+n_{L-1}=N-n_L} \frac{N!}{n_0!\, n_1! \ldots n_L!} \overline{\Pi}_0^{n_0} \ldots \overline{\Pi}_{L-1}^{n_{L-1}} \right] \\
&= \sum_{n_L=0}^{N} \frac{N!}{n_L!\,(N-n_L)!} \overline{\Pi}_L^{n_L} \left[\sum_{n_0+\ldots+n_{L-1}=N-n_L} \frac{(N-n_L)!}{n_0!\, n_1! \ldots n_{L-1}!} \overline{\Pi}_0^{n_0} \ldots \overline{\Pi}_{L-1}^{n_{L-1}} \right] \\
&= \sum_{n_L=0}^{N} \frac{N!}{n_L!\,(N-n_L)!} \overline{\Pi}_L^{n_L} \left(\overline{\Pi}_0 + \ldots + \overline{\Pi}_{L-1} \right)^{N-n_L} \\
&= \sum_{n_L=0}^{N} \frac{N!}{n_L!\,(N-n_L)!} \overline{\Pi}_L^{n_L} \left(1 - \overline{\Pi}_L \right)^{N-n_L},
\end{aligned} \tag{C.2}$$

e podemos então calcular cada média que resulta dos termos entre parenteses da equação (2.13). Assim

$$\sum_{\mathbf{n}} n_R \mathcal{P}_{\overline{\Pi}_P}(\mathbf{n}) = \sum_{n_L=0}^{N} n_R \binom{N}{N-n_L} \overline{\Pi}_L^{n_L} \left(1 - \overline{\Pi}_L \right)^{N-n_L} \tag{C.3}$$

$$= \sum_{n_L=0}^{N} \binom{N}{N-n_L} \overline{\Pi}_L^{n_L} n_R \left(\overline{\Pi}_0 + \ldots + \overline{\Pi}_{L-1} \right)^{N-n_L}$$

$$
\begin{aligned}
&= \sum_{n_L=0}^{N} \binom{N}{N-n_L} \overline{\Pi}_L^{n_L} \overline{\Pi}_R \frac{\partial}{\partial \Pi_R} \left(\overline{\Pi}_0 + \ldots + \overline{\Pi}_{L-1}\right)^{N-n_L} \\
&= \sum_{n_L=0}^{N-1} \binom{N}{N-n_L} \overline{\Pi}_L^{n_L} \overline{\Pi}_R \left(N-n_L\right) \left(1-\overline{\Pi}_L\right)^{N-1-n_L} \\
&= N \sum_{n_L=0}^{N-1} B_{n_L} \overline{\Pi}_R(t) \qquad \text{(C.4)}
\end{aligned}
$$

onde

$$
B_{n_L} = \binom{N-1}{n_L} \left[\overline{\Pi}_L(t)\right]^{n_L} \left[1-\overline{\Pi}_L(t)\right]^{N-1-n_L}. \qquad \text{(C.5)}
$$

Da mesma maneira, podemos calcular

$$
\begin{aligned}
\sum_{\mathbf{n}} n_L \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) &= \sum_{n_L=0}^{N} n_L \binom{N}{N-n_L} \overline{\Pi}_L^{n_L} \left(1-\overline{\Pi}_L\right)^{N-n_L} \\
&= \left[\Pi_L(t)\right]^N + \sum_{n_L=0}^{N-1} \frac{n_L \left(1-\overline{\Pi}_L\right)}{\left(N-n_L(t)\right)} \binom{N-1}{n_L} \overline{\Pi}_L^{n_L} \left(1-\overline{\Pi}_L\right)^{N-n_L-1} \\
&= M_{PL} \left[\Pi_L(t)\right]^N + \sum_{n_L=0}^{N-1} B_{n_L} \frac{r}{1-r} \sum_{R \neq L} \overline{\Pi}_R(t), \qquad \text{(C.6)}
\end{aligned}
$$

com $r = n_L/N$. Juntando-se os dois resultados, obtemos a equação (2.15) do texto principal

$$
\overline{\Pi}_P(t+1) = M_{PL} \left[\Pi_L(t)\right]^N + \sum_{n_L=0}^{N-1} B_{n_L} \frac{\sum_{R=0}^{L-1} \Pi_R(t) \left[M_{PR} + a\frac{r}{1-r} M_{PL}\right]}{1+r\left(a-1\right)}. \qquad \text{(C.7)}
$$

Outro resultado bastante simples é o limite determinístico, $N \to \infty$ dessa equação. Nesse limite temos $r \to \overline{\Pi}_L(t)$ e

$$
\begin{aligned}
\overline{\Pi}_P(t+1) &= \underbrace{M_{PL} \left[\Pi_L(t)\right]^N}_{=0} + \sum_{n_L=0}^{N-1} B_{n_L} \frac{\sum_{R=0}^{L-1} \Pi_R(t) \left[M_{PR} + a\frac{\overline{\Pi}_L(t)}{1-\overline{\Pi}_L(t)} M_{PL}\right]}{1+\overline{\Pi}_L(t)\left(a-1\right)} \\
&= \left[\sum_{R=0}^{L-1} M_{PR} \overline{\Pi}_R(t) + a\frac{\overline{\Pi}_L(t)}{1-\overline{\Pi}_L(t)} M_{PL} \underbrace{\sum_{R=0}^{L-1} \overline{\Pi}_R(t)}_{=1-\overline{\Pi}_L(t)}\right] \underbrace{\sum_{n_L=0}^{N-1} B_{n_L}}_{=1} \\
&= \frac{\sum_{R=0}^{L-1} M_{PR} \overline{\Pi}_R(t) + a M_{PL} \overline{\Pi}_L(t)}{1+\left(a-1\right)\overline{\Pi}_L(t)}, \qquad \text{(C.8)}
\end{aligned}
$$

que é a equação (2.17).

Apêndice D: Generalização para qualquer relevo de replicação

Analogamente ao apêndice anterior partimos da equação (3.1),

$$\overline{n}_P(t+1) = N \sum_{\mathbf{n}} \sum_{R} \left(\frac{\overline{n}_R A_R}{\sum_K \overline{n}_K A_K} \right) M_{PR} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) . \tag{D.1}$$

Diferentemente do caso do relevo de um pico, esta equação não pode ser iterada para um conjunto A_P qualquer. Assim, utilizando a relação auxiliar (3.2), escrevemos

$$\overline{\Pi}_P(t+1) = \sum_{\mathbf{n}} \sum_{R} M_{PR} n_R A_R \left(\int_0^\infty e^{-(\sum_{K=0}^{L} n_K A_K)x} dx \right) \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) . \tag{D.2}$$

Vamos manipular esta equação trabalhando primeiramente os termos exponenciais

$$\begin{aligned} \overline{\Pi}_P(t+1) \;=\; & \int_0^\infty dx \sum_{\mathbf{n}} M_{P0} n_0 A_0 e^{-(n_0A_0+\ldots+n_LA_L)x} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\ & + \int_0^\infty dx \sum_{\mathbf{n}} M_{P1} n_1 A_1 e^{-(n_0A_0+\ldots+n_LA_L)x} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\ & + \ldots \\ & + \int_0^\infty dx \sum_{\mathbf{n}} M_{PL} n_L A_L e^{-(n_0A_0+\ldots+n_LA_L)x} \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) . \end{aligned} \tag{D.3}$$

Rearranjando os somatórios

$$\begin{aligned} \overline{\Pi}_P(t+1) = {} & M_{P0} \int_0^\infty dx \left[\sum_{n_0} n_0 A_0 e^{-n_0A_0x} \sum_{n_1} e^{-n_1A_1x} \ldots \sum_{n_L} e^{-n_LA_Lx} \right] \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\ & + M_{P1} \int_0^\infty dx \left[\sum_{n_0} e^{-n_0A_0x} \sum_{n_1} n_1 A_1 e^{-n_1A_1x} \ldots \sum_{n_L} e^{-n_LA_Lx} \right] \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \\ & + M_{PL} \int_0^\infty dx \left[\sum_{n_0} e^{-n_0A_0x} \sum_{n_1} e^{-n_1A_1x} \ldots \sum_{n_L} n_L A_L e^{-n_LA_Lx} \right] \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) \end{aligned} \tag{D.4}$$

ou de uma forma mais concisa

$$\overline{\Pi}_P(t+1) = \sum_{R=0}^{L} M_{PR} \int_0^\infty dx \left[\sum_{n_R=0}^{N} n_R A_R e^{-n_R A_R x} \prod_{S\neq R} \left(\sum_{n_S=0}^{N} e^{-n_S A_S x} \right) \right] \mathcal{P}_{\overline{\Pi}}(\mathbf{n})$$

$$= \sum_{R=0}^{L} M_{PR} \int_0^\infty dx \left[\sum_{n_R=0}^{N} n_R A_R e^{-n_R A_R x} \prod_{S\neq R} \left(\sum_{n_S=0}^{N} e^{-n_S A_S x} \right) \right] \frac{N!}{n_0! \ldots n_L!} \overline{\Pi}_0^{n_0} \ldots \overline{\Pi}_L^{n_L}$$

$$= \sum_{R=0}^{L} M_{PR} \int_0^\infty dx \sum_{n_R=0}^{N} n_R A_R \left(\overline{\Pi}_R e^{-A_R x} \right)^{n_R} \prod_{S\neq R} \left(\sum_{n_S=0}^{N} \frac{N!}{n_0! \ldots n_L!} \overline{\Pi}_S e^{-n_S A_S x} \right) \quad \text{(D.5)}$$

onde aproveitamos para explicitar a somatória em $\mathbf{n}$ de $\mathcal{P}_{\overline{\Pi}}(\mathbf{n})$ ou seja, deixamos a equação em termos de um relevo arbitrário. Fazendo-se

$$\Theta_R = \overline{\Pi}_R e^{-A_R x}, \quad \text{(D.6)}$$

a equação (D.5) é reescrita como

$$\overline{\Pi}_P(t+1) = \sum_{R=0}^{L} M_{PR} \int_0^\infty dx \sum_{n_R=0}^{N} n_R A_R \Theta_R^{n_R} \prod_{S\neq R} \left(\sum_{n_S=0}^{N} \frac{N!}{n_0! \ldots n_L!} \Theta_S^{n_S} \right). \quad \text{(D.7)}$$

Podemos agora escrever esta equação de uma maneira mais interessante utilizando a identidade

$$\Theta_R^{n_R} \prod_{S\neq R} \left(\sum_{n_S=0}^{N} \frac{N!}{n_0! \ldots n_L!} \Theta_S^{n_S} \right) = \frac{N!}{(N-n_R)! n_R!} \Theta_R^{n_R} \left(\sum_{S=0}^{L} \Theta_S - \Theta_R \right)^{N-n_R} \quad \text{(D.8)}$$

de forma que

$$\begin{aligned}
\overline{\Pi}_P(t+1) &= \sum_{R=0}^{L} M_{PR} A_R \int_0^\infty dx \sum_{n_R=0}^{N} n_R \frac{N!}{(N-n_R)! n_R!} \Theta_R^{n_R} \left(\sum_{S=0}^{L} \Theta_S - \Theta_R \right)^{N-n_R} \\
&= \sum_{R=0}^{L} M_{PR} A_R \int_0^\infty dx \Theta_R \frac{\partial}{\partial \Theta_R} \left(\sum_{K=0}^{L} \Theta_S \right)^N \\
&= N \int_0^\infty dx \left(\sum_{K=0}^{L} \Theta_K \right)^{N-1} \sum_{R=0}^{L} M_{PR} A_R \Theta_R \\
&= N \int_0^\infty dx \left(\sum_{K=0}^{L} \overline{\Pi}_K e^{-A_K x} \right)^{N-1} \sum_{R=0}^{L} M_{PR} A_R \overline{\Pi}_R e^{-A_R x}. \quad \text{(D.9)}
\end{aligned}$$

Finalmente fazendo a mudança da variável $x \rightarrow x A_R$, para facilitar a integração, chegamos a

$$\overline{\Pi}_P(t+1) = N \int_0^\infty e^{-x} \sum_{R=0}^{L} M_{PR} \overline{\Pi}_R(t) \left(\sum_{K=0}^{L} e^{-x A_K / A_R} \overline{\Pi}_K(t) \right)^{N-1} dx, \quad \text{(D.10)}$$

que é a equação (3.3) do texto.

A outra grandeza que pode ser derivada nessa forma mais geral é o segundo momento de Π_P. Do Apêndice B sabemos que

$$\overline{\Pi_P^2}(t+1) = \frac{\overline{n_P^2}(t+1)}{N^2} = \frac{\overline{\Pi}_P(t+1)}{N} + (N-1)\sum_{\mathbf{n}}\left(\sum_R M_{PR}W_R\right)^2 \mathcal{P}_{\overline{\Pi}}(\mathbf{n}). \quad \text{(D.11)}$$

Vamos primeiramente rearranjar o somatório em $\mathbf{n}$ de $\mathcal{P}_{\overline{\Pi}}(\mathbf{n})$ nessa equação:

$$\sum_{\mathbf{n}}\left(\sum_R M_{PR}W_R\right)^2 \mathcal{P}_{\overline{\Pi}}(\mathbf{n}) = \int_0^\infty x dx \sum_{\mathbf{n}}\left[\left(\sum_R M_{PR}n_R A_R\right)^2 e^{-(\sum_{K=0}^L n_K A_K)x}\right]\mathcal{P}_{\overline{\Pi}}(\mathbf{n})$$

$$= \int_0^\infty x dx \sum_R M_{PR}^2 A_R^2 \left(\sum_{\mathbf{n}} n_R^2 e^{-(\sum_{K=0}^L n_K A_K)x}\mathcal{P}_{\overline{\Pi}}(\mathbf{n})\right) +$$

$$\int_0^\infty x dx \sum_R \sum_{S\neq R} M_{PR}A_R M_{PS}A_S \left(\sum_{\mathbf{n}} n_R n_S e^{-(\sum_{K=0}^L n_K A_K)x}\mathcal{P}_{\overline{\Pi}}(\mathbf{n})\right). \quad \text{(D.12)}$$

Devemos agora calcular as duas somatórias entre parenteses separadamente. Assim

$$\sum_{\mathbf{n}} n_R^2 e^{-(\sum_{K=0}^L n_K A_K)x}\mathcal{P}_{\overline{\Pi}}(\mathbf{n}) = \sum_{\mathbf{n}} n_R^2 e^{-(\sum_{K=0}^L n_K A_K)x}\frac{N!}{n_0!\ldots n_L!}\overline{\Pi}_0^{n_0}\ldots\overline{\Pi}_L^{n_L}$$

$$= \sum_{\mathbf{n}} n_R^2 \frac{N!}{n_0!\ldots n_L!}\left(e^{-A_0 x}\overline{\Pi}_0\right)^{n_0}\ldots\left(e^{-A_L x}\overline{\Pi}_L\right)^{n_L}$$

$$= \sum_{\mathbf{n}} n_R^2 \frac{N!}{n_0!\ldots n_L!}\Theta_0^{n_0}\ldots\Theta_L^{n_L}$$

$$= \sum_{\mathbf{n}} \Theta_R \frac{\partial}{\partial\Theta_R}\left[\Theta_R\frac{\partial}{\partial\Theta_R}\left(\frac{N!}{n_0!\ldots n_L!}\Theta_0^{n_0}\ldots\Theta_L^{n_L}\right)\right]$$

$$= \Theta_R\frac{\partial}{\partial\Theta_R}\left[\Theta_R\frac{\partial}{\partial\Theta_R}\left(\sum_K\Theta_K\right)^N\right]$$

$$= N\Theta_R\left(\sum_K\Theta_K\right)^{N-1} + N(N-1)\Theta_R^2\left(\sum_K\Theta_K\right)^{N-2}$$

$$= N\overline{\Pi}_R e^{-xA_R}\left(\sum_K e^{-xA_K}\right)^{N-1} + N(N-1)\left(\overline{\Pi}_R e^{-xA_R}\right)^2\left(\sum_K e^{-xA_K}\right)^{N-2} \quad \text{(D.13)}$$

Um cálculo análogo nos leva a

$$\sum_{\mathbf{n}} n_R n_S e^{-(\sum_{K=0}^L n_K A_K)x}\mathcal{P}_{\overline{\Pi}}(\mathbf{n}) = N(N-1)\overline{\Pi}_R e^{-xA_R}\overline{\Pi}_S e^{-xA_S}\left(\sum_K e^{-xA_K}\right)^{N-2}. \quad \text{(D.14)}$$

Voltando com esses resultados em (D.12) chegamos a equação (3.5) do texto principal

$$\overline{\Pi^2}_P(t+1) = \frac{\overline{\Pi}_P(t+1)}{N} +$$

$$+ (N-1)\int_0^\infty x e^{-x} \sum_{R=0}^{L} e^{-A_R x} M_{PR}^2 A_R^2 \overline{\Pi}_R(t) \left(\sum_{K=0}^{L} e^{-x A_K} \overline{\Pi}_K(t) \right)^{N-1} dx$$

$$+ (N-1)^2 \int_0^\infty x \left(\sum_{R=0}^{L} e^{-A_R x} M_{PR} A_R \overline{\Pi}_R(t) \right)^2 \left(\sum_{K=0}^{L} e^{-x A_K} \overline{\Pi}_K(t) \right)^{N-2} dx. \quad \text{(D.15)}$$

Bibliografia

[1] M. Eigen. *Selforganization of matter and the evolution of biological macromolecules,* Naturwissenchaften, **58**, 465 (1971).

[2] B. Kuppers. *Molecular theory of evolution: outline of a physical-chemical theory of the origin of life* (Springer-Verlag, Berlin, 1983).

[3] F. Monteiro e F. Morán. *Biofísica - Procesos de autoorganización en biologia* (Eudema Universidad, Madrid, 1992).

[4] M. Eigen, J. McCaskill e P. Schuster. *The molecular quasi-species,* Adv. Chem. Phys., **75**, 149 (1989).

[5] C. J. Thompson e J. L. McBride. *On the Eigen's theory of the selforganization of molecules and the evolution of biological macromolecules,* Math. Biosci., **21**, 127 (1974).

[6] B. L. Jones, R. H. Enns e S. S. Rangnekar. *On the theory of selection of coupled macromolecular systems,* Bull. Math. Biol., **38**, 12 (1976).

[7] S. Wright. *Evolution and the genetics of populations* Vol. I (The University of Chicago Press, Chicago, 1968).

[8] J. Swetina e P. Schuster. *Self-replication with errors - a model for polynucleotide replication,* Biophysical Chemistry, **16**, 329 (1982).

[9] D. Alves e J. F. Fontanari. *Error threshold in finite populations,* Phys. Rev. E, **57**, 7008 (1998).

[10] D. R. Mills, F. R. Kramer e S. Spiegelman. *Complete Nucleotide Sequence of a Replicating RNA Molecule*, Science, **180**, 916 (1973).

[11] M. Nowak e P. Schuster. *Error thresholds of replication in finite populations mutation frequencies and the onset of Muller's ratchet,* J. Theor. Biol. **137**, 375 (1989).

[12] J. M. Smith. *Models of Evolution,* Proc. R. Soc. Lond. B, **219**, 315, (1983).

[13] P. R. A. Campos e J. F. Fontanari. *Finite size scaling of the quasispecies model,* Phys. Rev. E, **58**, 2664 (1998).

[14] I. Leuthäusser. *An exact correspondence between Eigen's evolution model and a two-dimensional Isisng system,* J. Chem. Phys. **84**, 1884 (1986). I. Leuthäusser, *Statistical Mechanics of Eigen's evolution model,* J. Stat. Phys. **48**, 343 (1987).

[15] S. Galluccio. *Exact solution of the quasispecies model in a sharply peaked fitness landscape*, Phys. Rev. E, **56**, 4526 (1997). S. Galluccio, R. Graber e Y. C. Zhang, *Diffusion on a hypercube lattice with pinning potential: exact results for the errror-catastrophe problem in biological evolution,* J. Phys. A: Math. Gen., **29**, L249 (1996).

[16] M. Eigen e P. Schuster. *The Hypercycle - A Principle of Natural Self-Organization* (Springer-Verlag, Berlin, 1979).

[17] G. P. Wagner e P. Krall. *What the difference between models of error thresholds and Muller's ratchet?*, J. of Math. Biol., **32**, 33 (1993).

[18] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Weley, Reading, MA, 1989).

[19] A. Prügel-Bennet. *Modelling evolving populations*, J. Theor. Biol., **185**, 81 (1997).

[20] G. Woodcock e P. G. Higgs. *Population evolution on a mutiplicative single-peak fitness landscape,* J. Theor. Biol., **179,** 61 (1996).

[21] D. Bonnaz e A. J. Koch. *Stochastic model of evolving populations*, J. Phys. A: Math. Gen., **31**, 417 (1998).

[22] D. L. Hartl e A. G. Clark. *Principles of Population Genetics* (Sinauer Associates, Sunderland, 1989).

[23] D. Alves e J. F. Fontanari. em preparação.

[24] R. Feistel e W. Ebeling. *Evolution of Complex Systems* (Kluwer Academic, 1989).

[25] S. Ghahramani. *Fundamentals of Probability* (Prentice Hall, New Jersey, 1996).

[26] Y. Husimi, K. Nishigaki, Y. Kinoshita e T. Tanaka. *Cellstat - a continuous culture system of a bacteriophage for the study of the mutation rate and the selection process at the DNA level*, Rev. Sci. Instrum., **53**, 517 (1981).

[27] W. Ebeling e R. Feistel. *Stochastic theory of molecular replication processes with selection character*, Ann. Phys., **34**, 81 (1977).

[28] T. Wiehe. *Model dependency of error thresholds: the role of fitness functions and contrasts beteen the finite and infinite sites models*, Gen. Res. Camb., **69**, 127 (1997).

[29] J. M. Smith. *Evolutionary Genetics* (Oxford University Press, Oxford, 1989).

[30] N. Karmarkar, R. Karp, G. Lueker e A. Odlyzko. *Probabilistic Analysis of Optimum Partitioning*, J. Appl. Prob. **23**, 626 (1986).

[31] D. Gillespie. J. Comp. Phys. 22, 403 (1976); J. Phys. Chem. 81, 2340 (1977).

[32] P. R. A. Campos e J. F. Fontanari. *Finite-size scaling of the error threshold transition in finite populations*, J. Phys. A: Math. Gen., **32**, L1 (1998).

[33] L. Peliti. *Introduction to the statistical theory of Darwinian evolution*, cond-mat/9712027 (1997).

[34] M. Eigen e Biebricher. *Sequence space and quasispecies distribuition.* In RNA genetics (ed. E. Domingo, J. J. Holland e P. Ahlquist), vol. III, 211 (Boca Raton: CRC Press, 1988).

[35] S. Nee e J. M. Smith. *The evolutionary biology of molecular parasites*, Parasitology, **100**, S5 (1990).

[36] H. J. Muller. *The relation of recombination to mutational advance*, Mutational Research, **1**, 2 (1964).

[37] P. G. Higgs e G. Woodcock. *The accumulation of mutations in asexual populations and the structure of genealogical trees in the presence of selection*, J. Math. Bio., **33**, 677 (1995).

[38] J. Haigh. *The accumulation of deleterious genes in a population: Muller's ratchet*, Theor. Pop. Bio., **14**, 251 (1978).

[39] D. Charlesworth, M. T. Morgan e B. Charlesworth. *Mutation accumulation in finite outbreeding and inbreeding populations*, Genet. Res., **61**, 39 (1993).

[40] A. S. Kondrashov. *Muller's ratchet under epistatic selection*, Genetics, **136**, 1469 (1994).

[41] G. P. Wagner e W. Gabriel. *Quantitative variation in finite parthenogenetic populations: what stops Muller's ratchet in the absence of recombination?*, Evolution, **44**, 715 (1990).

[42] T. Whiehe, E. Baake and P. Schuster. *Error propagation in reproduction of diploid organisms: a case study on peaked landscapes*, J. Theor. Biol., **177**, 1 (1995).

[43] D. Alves e J. F. Fontanari. *Population genetic approach to the quasispecies model*, Phys. Rev. E, **54**, 4048 (1996).

[44] M. Kimura e T. Muruyama. *The mutational load with epistatic gene interaction in fitness*, Genetics, **54**, 1337 (1966).

[45] D. Alves e J. F. Fontanari. *Error threshold in the evolution of diploid organisms*, J. Phys. A: Math. Gen., **30**, 2601 (1997).

[46] R. Malarz e D. Tiggemann. *Dynamics in Eigen quasispecies model*, Int. J. of Modern Phys. C, (1998).

[47] P. G. Higgs. *Error thresholds and stationary mutant distributions in mutilocus diploid genetics models*, Genet. Res., Camb., **63**, 63 (1994).

[48] J. F. Crow e M. Kimura. *An introduction to population genetics theory.* (Harper and Row, New York, 1970).

[49] A. S. Kondrashov. *Selection against harmful mutation in large sexual and asexual population.* Genet. Res., **40**, 325 (1982).

[50] B. Charlesworth. *Mutation-selection balance and evolutionary advantage of sex and recombination.* Genet. Res., **55**, 199 (1990).

[51] B. Charlesworth, P. Sniegowiski e W. Stephan. *The evolutionary dynamics of repetitive DNA in eukaryotes.* Nature, **371**, 215 (1994).

[52] J. M. Smith. *The evolution of sex.* (Cambridge University Press, 1978).

[53] E Baake e T Wiehe. *Bifurcations in haploid and diplod sequence space models.* J. Math. Biol., **35**, 321 (1997).

[54] M. C. Boerlijst, S. Bonhoeffer e M. A. Nowak. *Viral quasi-species and recombination.* Proc. R. Soc. Lond. B, **263**, 1577 (1996).

[55] M. Lynch, R. Bürger e W. Gabriel. *The mutational meltdown in asexual populations*, J. Hered., **84**, 339 (1993).

[56] R. S. Diaz. *Genetic diversity of HIV*, Revista de Microbiologia, **28**, 69 (1997).

[57] J. M. Coffin. *Genetic diversity and evolution of retroviruses.* Current Topics in Microbiology, **176**, 143 (1992).

[58] S. Bonhoffer e P. F. Stadler. *Error thresholds on correlated fitness landscapes.* Preprint Universitat Wien, (1992).

[59] I. S. Novella, E. A Duarte, S. F. Elena, A. Moya, E. Domingo e J.J. Holland. *Exponential increases od RNA virus fitness during large populations transmissions*, Proc. Natl. Acad. Sci. USA, **92**, 5841 (1995).

[60] L. Tsimring, H. Levine e D. Kessler. *RNA virus evolution via a fitness-space model.* Phys. Rev. Lett., **76**, 4440 (1996).

[61] D. Alves. *Monografia apresentada como exame de qualificação.* Instituto de Física de São Carlos -USP, (1998).

[62] M. W. Feldman, F. B. Christiansen e L. D. Brooks. *Evolution of recombination in a constant environment.* Proc. Nat. Acad. Sci. U. S. A., **77**, 4838 (1980).

[63] M. Vicente, O. Kinouchi e N. Caticha. preprint (1998).

[64] M. A. Novak, R. M. May e R. M. Anderson. *The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease.* AIDS. **4**, 1095 (1990).