Aspectos de Física Estatística na Evolução e no Crescimento Molecular

Wenderson Alexandre de Sousa Silva

Dissertação apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, para obtenção do Título de Mestre em Biotecnologia.

São Paulo
2009

Aspectos de Física Estatística na Evolução e no Crescimento Molecular

Wenderson Alexandre de Sousa Silva

Dissertação apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, para obtenção do Título de Mestre em Biotecnologia.

Área de concentração: Biotecnologia

Orientador: Prof. Dr. Airton Deppman

São Paulo
2009

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

© reprodução total

Sousa Silva, Wenderson Alexandre.

Aspectos de física estatística na evolução e no crescimento
molecular / Wenderson Alexandre Sousa Silva. -- São Paulo, 2009.

Orientador: Airton Deppman.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de
Ciências Biomédicas. Programa de Pós-Graduação Interunidades em
Biotecnologia USP/IPT/Instituto Butantan. Área de concentração:
Biotecnologia. Linha de pesquisa: Evolução molecular.

Versão do título para o inglês: Aspects of statistical physics on
evolution and in molecular growth.

Descritores: 1. Evolução molecular 2. Física estatística 3.
Crescimento genômico 4. Modelagem computacional 5. Teoria de
informação 6. Sistemas complexos I. Deppman, Airton II.
Universidade de São Paulo. Instituto de Ciências Biomédicas.
Programa de Pós Graduação em Biotecnologia III. Título.

ICB/SBIB61/2009

UNIVERSIDADE DE SÃO PAULO
Programa de Pós-Graduação Interunidades em Biotecnologia
Universidade de São Paulo, Instituto Butantan, Instituto de Pesquisas Tecnológicas

Candidato(a): Wenderson Alexandre Sousa Silva.

Título da Dissertação: Aspectos de física estatística na evolução e no crescimento molecular.

Orientador(a): Ailton Deppman.

A Comissão Julgadora dos trabalhos de Defesa da **Dissertação de Mestrado**,
em sessão pública realizada a/...../....., considerou

☐ **Aprovado(a)**

☐ **Reprovado(a)**

Examinador(a): Assinatura:
Nome:
Instituição:

Examinador(a): Assinatura:
Nome:
Instituição:

Presidente: Assinatura:
Nome:
Instituição:

Às minhas famílias formadas em Campo Grande e Piracicaba,
que tantos ombros e braços me emprestaram. À minha esposa
que me apresentou Jesus e me deu duas preciosidades.

AGRADECIMENTOS

A vida é um quebra-cabeça que não sabemos e não podemos montar sozinhos. Algumas pessoas surgem e colocam apenas uma peça e não são mais vistas. Outras contribuem com tantas peças que não sabemos se o quebra-cabeça é nosso ou se é desses ajudantes. Todas têm sua importância, seja apenas por ouvir ou apenas por falar. Alguns desses “jogadores” de quebra-cabeça merecem destaques na conclusão dessa etapa de minha vida e um agradecimento especial:

Ao prof. Chuck e aos integrantes do colégio Avant Garde, em Campo Grande-MS, por concretizar o meu sonho de estudar em uma grande universidade.

Aos meus **amigos** Rodrigo Bombardi, Vlamir da Silva Jr. e Gian Barbosa por tantas noites e madrugadas divididas entre livros, listas de exercícios, trabalhos e gargalhadas.

Aos meus filhos João Gabriel e Leticia por serem muito mais que uma motivação. Papai ama vocês.

A minha esposa Eliane por tanta paciência e apoio durante todo o mestrado. O tempo passado longe de vocês está acabando. Te amo!

A minha mãe que, mesmo com simplicidade e dificuldade, me incentivou a estudar. A minha irmã Lidiane que sempre sorridente me faz sentir especial, mesmo que remotamente.

Ao meu orientador Airton Deppman, que, para um pizzaiolo, é um ótimo físico. Apesar de o trabalho nunca terminar, teve paciência de esperar o “filho pródigo” voltar de tantos sumiços.

À Sileine por tanto auxílio acadêmico, profissional e pessoal. Valeu Si!

Ao meu padrinho Sr. Baudijão, o primeiro a acreditar em meu potencial.

“O Ótimo é inimigo do Bom”
Voltaire

“O Ótimo é inimigo do Bom. A Seleção
Natural prefere o Bom”
W.A.S.S. Perfe

RESUMO

PERFE, W.A.S.S. **Aspectos de Física Estatística na Evolução e no Crescimento Molecular.** Dissertação (Mestrado)- Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2008.

A evolução molecular, impulsionada pela Teoria Sintética da Evolução, tornou um assunto indispensável na compreensão da evolução da vida. O crescimento genômico, etapa responsável pelo maior potencial de armazenamento de informação e estabilidade, também foi submetido à indelével ação da seleção natural. Utilizando a metodologia dos ciclos de amplificação-mutação-seleção, tal como o SELEX (systematic evolution of ligands by exponential enrichment), que mimetizam a seleção natural, e ferramentas da Teoria de Informação, foram desenvolvidos e implementados programas para simular a evolução, considerando, além de outros, um parâmetro pouco explorado na literatura: a variação do tamanho do genoma. Foram estudados dois cenários distintos; no primeiro a seleção era dependente da busca exata de uma sequência pré-determinada (o filtro). Além disso, a entropia de Shannon considerada era referente ao alinhamento da molécula toda. Avaliando configurações simples desse modelo, foi possível desenvolver uma equação analítica que descreveu bem os resultados (para tamanho de genoma constante). No segundo cenário, foram exploradas a seleção não específica de uma sequência, o número máximo de bases constante, a entropia apenas das regiões de interesse e a presença de até cinco filtros de seleção. A entropia da molécula toda se mostrou pouco significativa (primeiro cenário), diferentemente da avaliada em apenas uma região. Foi possível observar que o crescimento do genoma foi pouco acentuado, predominando as moléculas menores, mesmo com grande quantidade de filtros, o que indica que o sistema está sob “seleção por compressão”, sendo, pois, necessário atribuir explicitamente vantagens às moléculas mais complexas para poder haver aumento no crescimento médio.

Palavras-chave: Evolução Molecular; Teoria de Informação; SELEX; Física Estatística; Modelagem Computacional; Evolução Darwiniana; Evolução do tamanho do genoma, Sistemas Complexos.

ABSTRACT

PERFE, W.A.S.S. **Aspects of Statistical Physics on Evolution and Genomical Growth.** Dissertation (Master) - Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2008.

Molecular Evolution, stimulated and supported by Evolution Synthetic Theory, became essential to understand evolution of life. Genomic growth was responsible to increase the capacity of storage information and to stability of the molecule; besides, it was also submitted to natural selection. Using amplification-mutation-selection methodology, such as SELEX (*systematic evolution of ligands by exponential enrichment*), and tools of Information Theory, it was developed computer program to simulate macromolecule evolution taking into account, besides other, a less study parameter in specialized journal: the genome size variation. It was studied two different scenarios. In the first one, selection was dependent of a specific sequence to be searched; moreover, Shannon entropy took into account all nucleotides of all molecules and it was studied with until two sequence target. It was possible develop an analytical equation to well describe simple settings of this model. In the second scenario, in another hand, selection depends on a specific sequence, but is not required to match the whole sequence. Also, to compute Shannon entropy it was taking into account only the least Hamming distance sequence of each molecule. It was studied until five sequence target in this scenario. Entropy was not significant in first scenario as it was in the second one. Size genome evolution shows the system were under compression selection, being necessary to get other advantage to become possible an increase in genome size.

Keywords: Molecular Evolution; Information Theory; SELEX, Statistical Physics; Computational Modelling; Darwinian Evolution; Genome Size Evolution; Complex Systems.

LISTA DE ABREVIACÕES

AMS	Amplificação-Mutação-Seleção
MC	Método de Monte Carlo
A	Adenina
G	Guanina
U	Uracila
T	Timina
C	Citosina
SELEX	Systematic evolution of ligands by exponential enrichment
PCR	Polymerase Chain Reaction
BLAST	Basic Local Alignment Search Tool
DNA	Ácido desoxirribonucléico
RNA	Ácido ribonucléico

LISTA DE FIGURAS

Figura 1: Representação esquemática simplificada de um ciclo (round) do SELEX.....	14
Figura 2: Fluxograma do algoritmo de amplificação, mutação e seleção do cenário I. .	19
Figura 3: Fluxograma do algoritmo de simulação do cenário II.	20
Figura 4: Fluxograma do programa principal.....	21
Figura 5: Cenário I - Taxa de mutação 0%.....	25
Figura 6: Cenário I -Taxa de mutação 5%.....	26
Figura 7: Cenário I -Taxa de mutação 10%.....	26
Figura 8 : Cenário I -Taxa de mutação 20%.....	27
Figura 9: Gráfico representando a relação entre afinidade e taxa de mutação.	28
Figura 10: Cenário I -Taxa de mutação 0% [com quebra e junções].	28
Figura 11: Cenário I -Taxa de mutação 5% [com quebra e junções].	29
Figura 12: Cenário I -Taxa de mutação 10% [com quebra e junções].	29
Figura 13: Cenário I -Eficiência 10%.....	30
Figura 14 : Cenário I -Eficiência 20%.....	31
Figura 15: Cenário I -Eficiência 80%.....	31
Figura 16: Cenário I -Eficiência 100%.....	32
Figura 17 : Cenário I -Eficiência do filtro 20% [com quebra e junções].	33
Figura 18: Cenário I -Eficiência do filtro 40% [com quebra e junções].	33
Figura 19: Cenário I -Eficiência do filtro 90% [com quebra e junções].	34
Figura 20: Gráfico da quantidade de moléculas, obtido nos resultados da simulação utilizando as expressões (4), (5) e (6) com parâmetros eficiência de 60% e taxa de mutação de 10%.....	37
Figura 21: Gráfico da afinidade, obtido nos resultados da simulação utilizando as expressões (4), (5) e (6) com parâmetros eficiência de 60% e taxa de mutação de 10%.	37
Figura 22.: Cenário II -Taxa de mutação 1%	39
Figura 23: Cenário II -Taxa de mutação 2%	40
Figura 24: Cenário II -Taxa de mutação 5%	40
Figura 25: Cenário II -Taxa de mutação 10%	41
Figura 26: Cenário II -Taxa de mutação 40%	41
Figura 27: Comprimento médio das moléculas sob seleção por compressão.	42
Figura 28.: Cenário II -Eficiência do filtro 20%	43
Figura 29: Cenário II -Eficiência do filtro 80%	44

SUMÁRIO

1 INTRODUÇÃO	8
2 OBJETIVOS E MOTIVAÇÃO.....	11
3 CONSIDERAÇÕES TEÓRICAS	12
3.1 Teoria de Informação e o crescimento genômico	12
3.2 SELEX e a evolução <i>in silico</i>	13
3.3 Um pouco de Teoria de Informação	15
4 METODOLOGIA.....	17
5 ANÁLISE DE RESULTADOS.....	24
5.1 Cenário I.....	24
5.1.1 Presença de 1 filtro	24
5.1.2 Equação do crescimento.....	35
5.1.3 Presença de 2 filtros.....	38
5.2 Cenário II	38
5.2.1 Variando a taxa de mutação	39
5.2.2 Variando a eficiência do filtro	43
6 SUGESTÕES E PERSPECTIVAS	46
7 CONSIDERAÇÕES FINAIS E CONCLUSÃO	46
REFERÊNCIAS BIBLIOGRÁFICAS	49
APÊNDICE A	52
APÊNDICE B.....	54

1 INTRODUÇÃO

O surgimento da vida na Terra é ainda uma questão aberta no meio científico. Muitas teorias e hipóteses foram formuladas, em diferentes áreas, para tentar esclarecer esta dúvida. Panspermia cósmica [1], o desenvolvimento de moléculas em gotas [2] e a “sopa primordial” [3-5] são algumas das possibilidades.

Os fatos que sucederam a singularidade do surgimento, por outro lado, são provavelmente tão ou mais ricos que o próprio “início de tudo”. Registros fósseis, por exemplo, mostram vestígios de vida já há 3,8 bilhões de anos atrás [6]; há 3,5 bilhões outros fósseis apresentam estruturas mais parecidas com as nossas atuais bactérias [7]. Este período de cerca de 300 milhões de anos reserva um mundo de acontecimentos sem registros, tais como o surgimento de moléculas com propriedades catalíticas e autocatalíticas, uma maior capacidade de armazenamento de informação, o surgimento dos envoltórios que protegem os ácidos nucleicos e as organelas.

Para entender a evolução molecular, teorias, modelos biológicos e físico-matemáticos foram desenvolvidos. O modelo matemático mais conhecido é, provavelmente, o Modelo de Eigen [8] e suas extensões [9]. O Modelo de Eigen descreve, originalmente, a evolução do fago Q β , por meio de equações diferenciais. Novos conceitos surgiram com esse modelo ou tornaram populares, tais como “quasi-espécie” e “catástrofe do erro”. Na Biologia, a hipótese do “Mundo de RNA” [10] tornou-se muito popular com a descoberta do RNA catalítico [11,12]. Após a maior popularidade de métodos da Biologia Molecular, tais como o sequenciamento e o PCR (*polymerase chain reaction*) [13], modelos baseados nos ciclos de amplificação-mutação-seleção (AMS) [14] têm sido fonte de grandes descobertas para a biotecnologia e pesquisa aplicada [15,16,]. O modelo de AMS mais conhecido é o SELEX (*systematic evolution of ligands by exponential enrichment*) [15,16]. O SELEX é responsável por diversas descobertas e desenvolvimentos de tratamentos ou detecção de doenças [17,18,19].

Outros modelos de evolução molecular levam em conta a estrutura secundária das moléculas [20,21], a catálise mútua (coevolução) ao invés da autocatálise [22,23].

Introdução

Há também abordagem mais Matemática [24], mais Física [25,26] ou mais Bioquímica [27,28].

Juntamente à Biologia, a Matemática, a Física e a Computação fornecem teorias e diversas ferramentas para analisar os resultados dos modelos. Por exemplo, há uso da *informação* de Shannon [29] (ou, alternativamente, a entropia); a informação pode ser entendida como a medida da diminuição de aleatoriedade das moléculas no sistema; isto é, quanto maior a informação do sistema, menor a “variedade” das regiões controladoras do metabolismo ou dos processos vitais da vida, o que permite garantir a permanência das moléculas mais adaptadas na população. Portanto, para verificar se alguma sequência ou molécula está evoluindo (em um dado ambiente e condição) e sendo selecionada, precisa-se verificar, por exemplo, se há aumento da “informação de Shannon” [30,31] ou complexidade [32] ou ainda a diminuição da entropia.

A forma mais comum de avaliar se alguma sequência de nucleotídeos ou região de uma molécula está sendo conservada é por meio do alinhamento de um conjunto de moléculas provindas de vários momentos/etapas diferentes [33]. Assim, pode-se analisar e verificar as regiões das moléculas que permanecem imutáveis após várias etapas. Na bioinformática, por exemplo, há busca de regiões conservadas (oligo ou polinucleotídeos) e/ou genes observando sua existência e significados em bancos de dados, tal como o BLAST [34], por exemplo. Estas buscas e comparações permitem a construção de dendogramas (árvores filogenéticas – não usadas neste trabalho) e também permite calcular a separação temporal entre espécies – o conhecido relógio molecular.

A evolução molecular precisa de mais do que seqüências conservadas para ocorrer. O aumento no número de nucleotídeos das moléculas também deve ter sido uma etapa fundamental. Como a informação é armazenada primeiramente nas seqüências, então uma seqüência maior tem possibilidade de armazenar mais informação [35]. No entanto, no período de surgimento da vida o ambiente ainda estava inóspito, o que deveria favorecer a quebra das seqüências que eventualmente cresciam. Assim, possivelmente houve uma “competição” entre quebras e crescimento. Por haver poucos registros biológicos, poucos experimentos surgiram para comprovar esta possibilidade¹. Toda a

¹ ADAMI, C. Comunicação pessoal.

Introdução

análise teórica sobre vantagens de haver crescimento ou não pode ser realizada sob a luz da Teoria da Informação de Shannon [36]. Vale ressaltar que não é linear a relação entre o crescimento genômico e a complexidade do organismo, fato esse conhecido por “paradoxo do valor C”.

A análise do fluxo de informação genética e da variação de tamanho (número de nucleotídeos) das moléculas através dos ciclos de amplificação-mutação-seleção será o foco deste trabalho.

2 OBJETIVOS E MOTIVAÇÃO

Os processos que envolvem a amplificação-mutação-seleção (AMS) mimetizam o processo de evolução darwiniana e podem representar o fluxo de informação entre as várias gerações. Experimentos *in vitro* utilizando AMS mostram também o efeito da seleção natural em todo o processo.

O intuito deste trabalho é avaliar, por meio de simulações computacionais, a importância de algumas variáveis e parâmetros relacionados à evolução de macromoléculas biológicas, focando, em especial, a entropia do *ensemble* das moléculas, o crescimento genômico médio e a adaptação das moléculas às condições ambientais. A escolha destes parâmetros foi devida sua atual importância na análise da evolução contida na literatura; muitos outros parâmetros e variáveis diferentes das aqui utilizadas podem ser testadas e avaliadas (e.g., estruturas secundárias e terciárias do RNA, introdução de elementos transponíveis como fonte de crescimento, mudança de cenário/condição ambiental durante a evolução, interações intragênicas etc.). As simulações são realizadas utilizando como base o protocolo AMS. O desenvolvimento e a implementação dos programas para a simulação fazem parte do objetivo deste trabalho.

3 CONSIDERAÇÕES TEÓRICAS

3.1 Teoria de Informação e o crescimento genômico

A Teoria de Informação (TI) [37] é a área de conhecimento que estuda a transmissão de símbolos e mensagens através de canais. A TI foi desenvolvida por Shannon em 1948, ao estudar Engenharia de Telecomunicações, mas seu uso é feito nas mais diversas áreas do conhecimento, tais como Física, Ciência da Computação, Biologia e Linguística [37]. Na engenharia de telecomunicações, estuda o envio e a recepção de mensagens, onde os símbolos poderiam ser as palavras ditas em um telefone, o canal seria o fio ou ondas eletromagnéticas, e as mensagens seriam passadas ao receptor. A mensagem que chega ao receptor pode sofrer alguma forma de alteração devido a falhas em etapas do envio, como uma interferência, por exemplo; essas falhas são chamadas de ruídos. Em Biologia, os símbolos que são passados são os nucleotídeos ou os aminoácidos, o canal é a replicação, a transcrição ou a tradução e a mensagem que deve ser passada é, em última instância, os genes. O ruído, em Biologia, é representado pelos erros ocorridos nos canais de transmissão (mutações).

A evolução molecular e a dinâmica de crescimento do genoma estão intrinsecamente relacionadas. No entanto, sob a luz da Teoria de Informação, não se pode afirmar que é sempre uma vantagem evolutiva haver o crescimento do genoma. Há pelo menos três diferentes tipos de pressão seletiva que um genoma pode sofrer: I) seleção por compressão II) seleção por transmissão III) seleção neutral [36].

A *seleção por compressão* está relacionada, utilizando os jargões da Teoria de Informação, ao custo de transmitir cada símbolo através de um canal. Na Biologia, pode-se entender como a energia despendida para se incluir na replicação cada nucleotídeo e também o tempo necessário para essa inclusão. Incluir muitos nucleotídeos ou incluí-los muito rapidamente teria um custo alto. Sendo assim, genomas muito grandes seriam desfavorecidos em detrimento aos menores genomas. Portanto, a seleção por compressão seleciona moléculas menores e predomina em ambientes onde as fontes de nucleotídeos ou de energia são pequenas, ou ainda em um cenário muito hostil, onde não houvesse tempo suficiente para as moléculas se duplicarem.

Considerações Teóricas

Ambientes onde houvesse fonte de energia ou não fosse tão hostil não favoreceriam moléculas pequenas necessariamente. Os genomas teriam mais vantagens se “aprendessem” a aproveitar outros recursos do ambiente ou ainda refinar o processo de replicação. É nesse cenário que predomina a *seleção por transmissão*, onde a largura da banda do canal de transmissão seria maior. Para aumentar a quantidade de informação armazenada, os genomas precisariam crescer, visto que as informações são armazenadas nas sequências (e também nas estruturas secundárias, terciárias e quaternárias) favorecendo, pois, a existência e o predomínio de genomas maiores.

A pressão por compressão e a pressão por seleção não consideram explicitamente os efeitos da mutação. Na Teoria de Informação, a mutação pode ser entendida como a existência de ruídos no canal de transmissão das mensagens. A mutação, em muitos casos, pode ser letal para a molécula. Considerando moléculas de DNA, por exemplo, mutação na região dos genes pode tornar a proteína correspondente completamente sem função; caso a proteína seja de uma etapa vital para a molécula (ou célula), a mutação pode ocasionar o seu fim.

As mutações, no entanto, não são sempre deletérias. Para a Teoria de Informação, se a *mensagem* que passa pelo canal de transmissão não for alterada, apesar de alguns símbolos terem sido alterados, então a mensagem que não sofreu mudança é tão boa quanto a que foi alterada. Pensando em moléculas, uma molécula que não tem a região dos genes (no caso do DNA) alterada depois de sofrer mutações ou sofreu uma mutação sinônima será tão boa quanto àquelas que não sofreram mutação. Para a molécula pode existir, inclusive, vantagens evolutivas. Muitas dessas mudanças não deletérias podem conferir mais robustez contra outras mutações. Esse cenário compõe a terceira classe de seleção: a *seleção neutral*. Vale notar que a seleção neutral age significativamente considerando seus efeitos na população, não somente em uma molécula individualmente [36,38].

3.2 SELEX e a evolução *in silico*

A evolução molecular realizada em laboratório é também chamada “evolução *in vitro*”. Baseado no estudo da evolução do RNA catalítico em laboratório, Joyce desenvolveu um protocolo para a seleção rápida de RNA com atividade catalítica: os ciclos de amplificação, mutação e seleção[14]. No ano seguinte, Ellington e Szostak[15]

Considerações Teóricas

e Tuerk e Gold[16] desenvolveram independentemente uma técnica para a seleção de aptâmeros (ácidos nucleicos com afinidade a uma sequência ou estrutura especial): o SELEX (*systematic evolution of ligands by exponential enrichment*). Esta técnica também envolve três etapas: amplificação, mutação e seleção. Veja a ilustração a abaixo:

Figura 1: Representação esquemática simplificada de um ciclo (round) do SELEX.

O ciclo começa no cDNA (DNA complementar) da figura. Primeiramente, o conjunto de DNA sofre amplificação por PCR error prone (cada molécula será copiada → replicada). O total de moléculas após o PCR é da ordem de 10^{15} moléculas, chegando a 10^{18} . A cópia, no entanto, não é precisa, havendo, pois, alguns erros. O novo conjunto (cópia com erro + original) é submetido ao processo de transcrição. A transcrição é a síntese de RNA a partir do DNA. O conjunto de RNA resultante da transcrição é

Considerações Teóricas

submetido à seleção (no caso do SELEX, o RNA se liga a alguma proteína de interesse). A maioria dos RNA's não selecionados é eliminada. Os selecionados (pouquíssimos inicialmente) sofrerão o processo de transcrição reversa (síntese de cDNA a partir de RNA). O novo grupo de cDNA reiniciará o ciclo.

Repetindo o ciclo do SELEX diversas vezes pode-se selecionar, por exemplo, moléculas de RNA com propriedades catalíticas[15,16] ou moléculas que se ligam à regiões ativas de proteínas causadoras de doença, inativando esta proteína [17]. A avaliação de quanto a população está “enriquecida” pode ser feita, por exemplo, por cromatografia [38].

3.3 Um pouco de Teoria de Informação

Para avaliar, *in silico*, a quantidade de moléculas com as características desejadas, pode-se analisar a razão entre as moléculas com a propriedade desejada e a quantidade total de moléculas. Esta razão é uma medida equivalente à constante de equilíbrio de uma equação química e será denominada *afinidade*. A análise da afinidade indicará a adaptação do grupo de moléculas ao sistema.

Outra análise que será feita será a da *entropia*. Shannon, em seu artigo de 1948 [29], desenvolveu uma das grandes ferramentas de análise de informação em diversas áreas [37]. A entropia de Shannon (H) é dada pela seguinte expressão:

$$H(X) = -\sum_{i=1}^n p(x_i) \log p(x_i) \quad (1)$$

onde $p(x_i)$ é a probabilidade de ocorrer x_i no conjunto X. No caso de sistemas biológicos, $p(x_i)$ representa a probabilidade de se encontrar uma base nitrogenada (A (adenina), C(citosina), G(guanina) ou U(uracila)/T(timina)) no *ensemble*. Para analisar a entropia de um alinhamento² de moléculas, é necessário calcular a entropia por posição. A expressão utilizada foi a seguinte:

² *Grosso modo*, é o processo de colocar as moléculas empilhadas com a primeira base de todas coincidindo. Pode-se, pois, analisar e verificar as regiões da molécula que são mais conservadas. Possivelmente, essas regiões conservadas têm a mesma origem (são homólogas).

Considerações Teóricas

$$H = \sum_{i=1}^L [-\sum_{b=A}^T p(b,i) \log_2 p(b,i)] \quad (2)$$

onde L é o tamanho da sequência (número de nucleotídeos), i é a posição de cada nucleotídeo, $b = \{A, C, G, T/U\}$ e $p(b,i)$ é a probabilidade de encontrar a base “b” na posição “i”.

Como foram avaliadas duas situações, a saber, com e sem quebra de moléculas (número de bases variado e constante, respectivamente), o valor de L poderia variar. Como as probabilidades provêm de um alinhamento então o L adotado foi a da maior sequência. A base do logaritmo determina a unidade que será dada a entropia H. Caso a base seja 2, a unidade será o *bit*. Se for logaritmo neperiano, a unidade será *nat*; se for logaritmo na base 10, a unidade será o *mer* [32].

Como as funções (1) e (2) podem aumentar com o acréscimo no número de moléculas ou no número de bases, então a medida H torna-se dependente da quantidade total de bases, o que dificulta a comparação da entropia de dois sistemas ou condições distintas. Para contornar este problema calcula-se a entropia por base (entropia por quantidade total de bases). Esse método foi utilizado no cenário I.

A simulação será feita utilizando o método de Monte Carlo (MC). Este método permite mimetizar a natureza por meio de seleção estatística. O MC, neste trabalho, permite escolher ao acaso elementos da população que será retirado da população para mantê-la constante ou para determinar a permanência de moléculas pouco afins (detalhes abaixo na seção Metodologia).

O ciclo de AMS pode ser visto como um simulador da seleção natural descrita por Darwin em *Origem das espécies*. Os mais “aptos” são os que sobrevivem ao “filtro” (ver Metodologia) e ao MC. Por isso, as moléculas mais adaptadas são denominadas *aptâmeros* [15].

4 METODOLOGIA

A metodologia empregada neste trabalho foi análoga a dos ciclos de Amplificação-Seleção-Mutação (AMS). Foram desenvolvidos e implementados programas na Linguagem Java que simulam uma versão adaptada desses ciclos. Em sua forma original, os ciclos de AMS não incluem a possibilidade de as moléculas crescerem ou quebrarem. Nas simulações, ambas as possibilidades foram avaliadas.

As seções que fazem *amplificação* e *mutação* foram feitas em um passo só, por meio do PCR³ com erro (PCR error prone). Além de realizar as replicações este PCR permite que haja erro na molécula-filha. No caso das simulações, esse erro era determinado *a priori* e foi um dos parâmetros avaliados. Como na evolução *in vitro* não há mecanismos de reparo⁴, o erro ocorrido nas replicações pode ser grande. Por isso e por se tratar de uma população pequena, a amplitude dos valores dos erros considerados foi maior do que aqueles que ocorrem na evolução biológica da natureza.

A *seleção* implementada mimetiza a ocorrida em métodos como o próprio SELEX. Foram selecionadas arbitrariamente seqüências pequenas de bases que serviram de referências para a adaptabilidade – aqui denominadas *seqüências conservadas, alvos ou filtros*. As moléculas consideradas “aptas” são aquelas que têm “afinidade”, complementaridade com as seqüências conservadas. Essas seqüências seriam análogas aos genes presentes nos DNA's .

Para o estudo da evolução, dois cenários foram considerados: I) as seqüências ditas afins deveriam ser exatamente como a seqüência conservada. II) as seqüências afins deveriam ter pelo menos 50% de similaridade com relação à seqüência conservada: quanto mais similar, maior a chance de a molécula permanecer no sistema. O similar se refere à distância de Hamming⁵ que deve ser no máximo igual à metade da seqüência conservada. Se a distância de Hamming fosse maior, então a molécula é

³ *Polymerase chain reaction*. É o método utilizado para fazer muitas replicações (cópias) de um mesmo RNA ou DNA.

⁴ Mecanismos de correção de bases realizado por proteínas para permitir maior fidelidade entre moléculas mães e filhas.

⁵ A distância de Hamming é o número de diferenças encontradas ao se comparar, posição por posição, duas seqüências de caracteres de mesmo tamanho. A distância de Hamming entre as palavras “CASA” e “CAAS” é dois.

Metodologia

eliminada. No cenário I), a entropia calculada considerou a molécula toda e não apenas uma região específica como foi feito no cenário II).

O modelo desenvolvido permite que as moléculas mudem de tamanho. Pode haver quebras ou inserções (no caso, junções) de outras moléculas, o que propicia uma diminuição ou aumento no comprimento de cada molécula, respectivamente. O critério de quebra adotado foi um tanto arbitrário: onde existisse a sequência UUU haveria quebra e este códon (sequência de três bases) seria os últimos da molécula. Esta sequência de uracilas funcionaria como um *códon de parada* (stop condon), uma sequência de três bases que indicam onde é o fim da tradução (processo de síntese de proteína) ou uma região onde enzimas de restrição cortariam a molécula. Naturalmente, caso o UUU fosse encontrado, há uma *probabilidade* de haver quebra, o que permite uma maior flexibilidade ao modelo (no cenário II a probabilidade é 1). Não foi escolhido um dos três códons de parada (UAG, UGA e UAA) por que a sequência UUU é mais fácil de ser observada visualmente (para testes visuais).

O critério de junção foi o que mais emanou esforço e tentativas. Muitos critérios avaliados resultavam em um crescimento muito rápido e indefinido ou um excesso de quebras que inviabilizavam a continuidade das moléculas no sistema (resultados não mostrados). Por fim, para haver junção, foram considerados dois casos, um para cada cenário. No cenário I, foi determinada uma sequência de cinco bases consecutivas para mimetizar o “sinal” de junção. Caso uma molécula tivesse as cinco bases e outra molécula tivesse cinco bases complementares⁶, então havia uma probabilidade delas se juntarem. No cenário II, as moléculas deveriam ter uma sequência específica de quatro bases, sendo que essas sequências juntas formam um palíndromo. Nesse caso, não foi considerada nenhuma probabilidade de não haver junção. Como a junção deve resultar em vantagem para as moléculas, as sequências escolhidas, no cenário II, continham parte da sequência conservada.

O estudo do cenário I considerava as moléculas com tamanho sempre constante no primeiro momento e depois com tamanhos variados, sendo constante a quantidade máxima de moléculas, o que gerou moléculas com tamanho muito grande. O cenário II

⁶ As bases complementares são A-T, T-A, A-U, U-A, C-G e G-C. Se tivéssemos a sequência AUUCG, a sua sequência complementar seria UAAGC.

Metodologia

foi estudado considerando o número de nucleotídeos máximo constante e igual à quantidade inicial.

O fluxo do *core* do programa que simula o cenário I (um ciclo ou round) está representado na figura abaixo:

Figura 2: Fluxograma do algoritmo de amplificação, mutação e seleção do cenário I.

Figura 3: Fluxograma do algoritmo de simulação do cenário II.

Metodologia

Figura 4: Fluxograma do programa principal.

O programa foi implementado na linguagem Java e é constituído de 8 classes. São elas: a classe `Arquivo.java`, `Entropia.java`, `RNA.java`, `TextoPL.java`, `Distancia.java`, `ScoreDistance.java`, `Selex.java` e `Selex_const.java`. A classe `Arquivo.java` é utilizada para a manipulação de arquivos (abertura, fechamento, leitura e escrita). A classe `TextoPL` monta todo o script para a elaboração dos gráficos. Os gráficos são “plotados” utilizando o software livre GNUPLOT, versão 4.2.

A classe `RNA` é a responsável por fazer a criação das moléculas utilizadas. A sequência de base é escolhida aleatoriamente. A classe `RNA` também faz a busca das sequências conservadas, diz onde que estão essas sequências, além de conter diversos métodos para manipulação de Strings (sequências de caracteres).

A classe `Entropia` faz o cálculo da entropia dada pelas equações (1) e (2) acima, além de calcular a *informação* utilizando as equações adaptadas de Schneider [24] (a *informação* não foi utilizada explicitamente nesta etapa). O cálculo da entropia no cenário I é realizado após avaliar o alinhamento da molécula toda. Assim, não houve região privilegiada na análise. No cenário II, no entanto, apenas uma região de interesse foi levada em conta no cálculo. Essa região era que apresentou menor distância de

Metodologia

Hamming com relação à primeira região conservada, dentro de cada molécula considerada afim.

A classe `Distancia.java` é utilizada para calcular a distância de Hamming e a distância de Levenshtein (análogo à distância de Hamming, mas usada para calcular a distância de sequências com diferentes quantidades de caracteres). Foi usada, no entanto, somente a distância de Hamming.

A classe `ScoreDistance.java` é responsável pela busca da sequência com menor distância de Hamming dentro de cada molécula do sistema. Essa menor distância é utilizada. Além disso, essa classe armazena o local e a sequência com menor distância para o posterior cálculo da entropia.

As classes principais são a `Selex.java` e a `Selex_const.java`, utilizadas no cenário I e II, respectivamente. As figuras 2 e 3 representam os fluxogramas que representam o trecho principal dessas classes. A figura 4 representa a classe principal que recebe todos os parâmetros para executar o programa. Além disso, foram implementados métodos para calcular o tamanho médio das moléculas, a variância, o desvio padrão, além, é claro, de calcular a afinidade e realizar a amplificação e seleção.

O cenário II considerava constante a quantidade máxima de nucleotídeos. Se uma molécula fosse eliminada por não ser afim ou por não ter distância de Hamming zero, então os nucleotídeos dessa molécula iriam para um reservatório. Para uma molécula se replicar era necessário que houvesse nucleotídeos suficientes no reservatório. Se não houvesse, não haveria replicação. Assim, nem sempre o sistema apresentava todos os nucleotídeos no mesmo ciclo.

A afinidade pode ser calculada pela seguinte expressão:

$$Aff = \frac{NA}{NA + NB} \quad (3)$$

Onde: NA é a quantidade de moléculas com afinidade ao(s) filtro(s) e NB representa as moléculas sem afinidade. O intuito de deixar moléculas sem afinidade na população é a possibilidade dessas moléculas sofrerem mutação e se tornarem afim,

Metodologia

além, é claro, possibilitar a sobrevivência em uma eventual mudança de condições ambientais (seleção natural).

O critério de parada do programa dependia da afinidade e do número de ciclos; a afinidade deveria ser igual a 1 (isto é, ter somente seqüências conservadas na população) ou o número de ciclos ser igual ao um dado limite. Este limite foi imposto para evitar que o programa entre em um loop infinito (e.g., quando afinidade é sempre menor que 1) ou nenhuma mudança significativa estiver ocorrendo. O software GNUPLOT permite que seja “plotado” gráficos durante a execução do programa, o que facilitou o monitoramento dos ciclos *real time*.

5 ANÁLISE DE RESULTADOS

Abaixo são mostrados os resultados utilizando um e dois filtros para o cenário I e de um até cinco filtros para o cenário II. Para a análise de cada parâmetro foi fixado o valor dos outros parâmetros. Os valores fixos utilizados no cenário I foram os seguintes: 50 bases, 500 moléculas inicialmente, 80% de eficiência do filtro (porcentagem de moléculas que permanecem mesmo sem conter alvo ou sequência conservada) e 5% de taxa de mutação. A população máxima, quando utilizada, foi fixada em 500 moléculas.

O cenário II foi teve os seguintes valores fixados: 100 bases, 500 moléculas iniciais, 50% de eficiência do filtro, 5% de taxa de mutação. A escolha dos valores dos parâmetros foi baseada em diversos critérios, tais como a literatura e alguns testes pilotos.

As figuras apresentadas contêm 4 gráficos cada, sendo gráficos da afinidade, da entropia por base, o tamanho médio das moléculas e a quantidade total de moléculas. O título dos gráficos do cenário I se refere à variável que está mudando e sendo avaliada. A figura que contém a quantidade moléculas, no primeiro cenário, há duas curvas. Uma delas é intitulada “Quantidade de moléculas” e se refere à quantidade após o método que deixa a população constante. A outra curva é intitulada “Número de moléculas após o Selex” e indica a quantidade de moléculas que permanecem após a “seleção”. A primeira curva quase não é vista nas figuras que não há quebras e junções, o mesmo ocorrendo para o tamanho médio das moléculas.

5.1 Cenário I

5.1.1 Presença de 1 filtro

Os resultados utilizando somente 1 filtro (ou alvo) são detalhadas nesta seção. Foram avaliados os impactos dos seguintes parâmetros: taxa de mutação, tamanho inicial das moléculas e eficiência do filtro. No entanto, serão discutidos e mostrado o impacto na variação da taxa de mutação e também da eficiência do filtro, pois apresentaram resultados mais significativos.

Análise de Resultados

Há muitas outras variáveis que não foram contempladas nesta etapa do projeto. Para cada novo parâmetro, a quantidade de gráficos aumenta pelo produto de partições de seu domínio (número de valores avaliados) pela quantidade de gráficos anterior. Portanto, mostrar todas as combinações de variáveis é impraticável e pouco pedagógico. Assim, apesar de terem sido estudadas todas as combinações, serão mostrados apenas os resultados mais representativos.

Os valores de mutação utilizados foram 0%, 5%, 10% e 20%, com e sem quebra e junção. Para o caso sem quebra, será apresentado um gráfico resumo de Afinidade *versus* Erro. No entanto, os gráficos de 30% a 100% de erro não foram mostrados. Os resultados estão apresentados abaixo:

5.1.1.1 Variando a taxa de mutação

a) Sem quebra

Figura 5: Cenário I - Taxa de mutação 0%.

Análise de Resultados

Figura 6: Cenário I -Taxa de mutação 5%.

Figura 7: Cenário I -Taxa de mutação 10%.

Análise de Resultados

Figura 8: Cenário I -Taxa de mutação 20%.

Pode-se perceber, ao analisar as figuras 5-8, que há a diminuição na **variação** da entropia, apesar do comportamento tão distinto. Uma maior taxa de mutação faz gerar moléculas-filha cada vez mais distintas das moléculas-mãe, o que pode justificar tal diminuição. Há também a diminuição da afinidade com o aumento da taxa de mutação. Acima de 10% de mutação pode-se considerar que a afinidade apresentada é meramente estatística. O programa foi executado utilizando a taxa de mutação múltipla de 10 até 100%. A figura abaixo ilustra o comportamento afinidade com a variação da mutação:

Análise de Resultados

Figura 9: Gráfico representando a relação entre afinidade e taxa de mutação.

A figura acima não veio de nenhuma função, apesar da suavidade. Os dados vieram da simulação (dados acima de 20% não mostrados). Baseado no gráfico acima e em estudos pilotos, o valor de 20% de mutação foi tomado como máximo nas simulações.

b) Presença de quebra e inserções

Figura 10: Cenário I -Taxa de mutação 0% [com quebra e junções].

Figura 11: Cenário I -Taxa de mutação 5% [com quebra e junções].

Figura 12: Cenário I -Taxa de mutação 10% [com quebra e junções].

Variando a mutação, mas incluindo agora quebras e junções. A afinidade varia rapidamente, mostrando grande sensibilidade à mudança de mutação. A entropia para a

Análise de Resultados

taxa de mutação de 20% mostra que as (pequenas) moléculas são praticamente aleatórias e a afinidade indica que não há adaptação dessas moléculas ao meio (resultado não apresentado). Houve grande variação no tamanho médio das moléculas, mostrando uma “competição” entre quebras e junções para os gráficos de 5% e 10%.

5.1.1.2 Variando a eficiência do filtro

A eficiência do filtro é o número que indica o quanto de moléculas não afins (sem afinidade) permanecerão na população no próximo *round*. Abaixo estão os resultados.

a) Sem quebra

Figura 13: Cenário I -Eficiência 10%.

Análise de Resultados

Figura 14: Cenário I -Eficiência 20%.

Figura 15: Cenário I -Eficiência 80%.

Análise de Resultados

Figura 16: Cenário I -Eficiência 100%.

O parâmetro que está sendo avaliado agora é a eficiência do filtro. Pode-se notar que a afinidade é insensível para a maioria dos valores apresentados; só houve mudança significativa para valores acima de 80%.

O último gráfico mostra o que aconteceria se não houvesse filtro: não haveria seleção, pois a geração e a eliminação de moléculas afins e não afins seria da mesma forma (mesma intensidade).

A entropia também é insensível, na média, à mudança de eficiência do filtro, apesar da pequeníssima variação (baixa quantidade de informação armazenada). Sendo assim, para o caso de um filtro apenas, a utilização de filtros abaixo de 80% é indiferente, podendo-se, assim, usar o mais conveniente.

b) Com quebra e junções

Figura 17 : Cenário I -Eficiência do filtro 20% [com quebra e junções].

Figura 18: Cenário I -Eficiência do filtro 40% [com quebra e junções].

Análise de Resultados

Figura 19: Cenário I -Eficiência do filtro 90% [com quebra e junções].

Ao analisar os gráficos acima, pode-se destacar três características. 1) A entropia assume os menores valores avaliados até o momento. 2) A afinidade é basicamente insensível à mudança de eficiência do filtro (como o caso sem quebra) 3) O tamanho médio molecular “explode”.

Aqui, podem-se notar algumas limitações do modelo (as limitações estarão sumarizadas no final). A primeira limitação é que o modelo não calcula bem a entropia se a diferença entre os tamanhos for muito grande. Para entender este fato, veja o seguinte exemplo. Sejam duas moléculas de tamanhos diferentes alinhadas:

Molécula 1: CCUCACUGGG

Molécula 2: GCGGUUCGUCGAACCAAGUUGAACC

A entropia até a décima posição pode ser diferente de zero (é em 80% das bases do exemplo). No entanto, a partir da décima primeira, a entropia é zero sempre, não contribuindo para o cálculo da entropia do sistema. Como a entropia tem como fator de normalização a quantidade média de nucleotídeos, aqueles nucleotídeos que não contribuem vão favorecer ou forçar a diminuição da entropia, o que pode levar a um valor fictício. Quanto maior for a diferença de tamanho, maior será a diferença entre a

Análise de Resultados

entropia real e a simulada. Para o cenário II foi resolvida essa limitação avaliando apenas um trecho da molécula de comprimento fixado.

Para os parâmetros aqui utilizados, temos a segunda limitação (na verdade, uma idiossincrasia do modelo desenvolvido). A mesma seqüência da seleção é a utilizada para realizar a junção. Essa dupla capacidade aumenta enormemente a chance de haver junção, visto que uma das seqüências necessárias para haver a junção está sempre presente no sistema (basta observar a alta afinidade).

5.1.2 Equação do crescimento

Com intuito de entender alguns aspectos, resultados e problemas que surgiram, foi desenvolvida uma função matemática que descrevesse o aumento no número de moléculas e a dinâmica da variação da afinidade. Isto é, a função descreve o programa quando é utilizado 1 filtro e não há quebras. A expressão que modela a simulação está apresentada abaixo e seu desenvolvimento está apresentado no Apêndice A:

$$Q(t) = 2^t \prod_{i=1}^t (1 - diff_i) [(1 - \alpha)^t N_A + \beta^t N_B + \alpha N_A \sum_{i=1}^t (1 - \alpha)^{t-1} \beta^{i+1}] \quad (4)$$

Onde:

$$diff_t = \begin{cases} 0, & \text{se } popMax < Q_{t+\frac{1}{2}} \\ \frac{popMax - Q_{t+\frac{1}{2}}}{Q_{t+\frac{1}{2}}}, & \text{se } popMax \geq Q_{t+\frac{1}{2}} \end{cases} \quad (5)$$

e

$$Q_{t+\frac{1}{2}} = 2(1 - \alpha)NA + 2\alpha NA + 2NB \quad (6)$$

$Q(t)$ é a quantidade de moléculas no “tempo” t (ciclo ou round);

α é o erro na replicação; β é a eficiência do filtro; N_A , N_B são as quantidades iniciais de moléculas afins e não afins; $popMax$ é a população máxima (aqui tem-se $popMax=500$ moléculas); NA , NB são as quantidades de moléculas afins e não afins no “tempo” t qualquer.

Análise de Resultados

Vale notar que a função $Q(t)$ é explicitamente dependente das quantidades iniciais de moléculas afins; caso não haja nenhuma no início elas não serão geradas por replicação com mutação. Na natureza, a probabilidade de haver uma mutação benéfica também é irrisória. Para levar em conta a baixíssima mas existente probabilidade de surgir uma molécula afim, bastaria incluir outro termo na equação (6), substituindo o $2NB$ por $2(1-\epsilon)NB+2\epsilon NB$, onde ϵ ($\epsilon \ll 1$) representa a probabilidade de surgir uma molécula afim durante a replicação.

Resumamos as três equações. A equação (6) descreve a replicação (duplicação) com um erro alfa; a parcela $2(1-\alpha)NA$ é a população com afinidade, enquanto as outras duas indicam a população não afim (a parcela $2\alpha NA$ indicam as moléculas-filha que perderam a afinidade na replicação).

A equação (5) é o fator de diminuição da população. Caso a população, ao se duplicar, ultrapasse o valor máximo (popMax), então o diff é calculado e será responsável por eliminar aleatoriamente as moléculas excedentes.

A equação (4) executa os cálculos dos passos subsequentes, resultando na quantidade de moléculas após a seleção. A afinidade é calculada conforme a equação (3)

$$(Aff = \frac{NA}{NA + NB}).$$

Para ilustrar, comparamos o resultado da simulação com o calculado pelas expressões (4), (5) e (6) acima. Foi adotado eficiência de 60% e taxa de mutação de 10%:

Análise de Resultados

Figura 20: Gráfico da quantidade de moléculas, obtido nos resultados da simulação utilizando as expressões (4), (5) e (6) com parâmetros eficiência de 60 % e taxa de mutação de 10 %.

Figura 21: Gráfico da afinidade, obtido nos resultados da simulação utilizando as expressões (4), (5) e (6) com parâmetros eficiência de 60 % e taxa de mutação de 10 %.

Análise de Resultados

5.1.3 Presença de 2 filtros

Foi realizado o estudo detalhado com dois filtros. O comportamento, no entanto, foi similar ao caso com apenas um filtro. O valor da afinidade foi sistematicamente maior com a presença de outro filtro. Isto pode ser, em parte, entendido avaliando a função $Q(t)$ e o impacto no aumento de um filtro. Se incluíssemos outro filtro com eficiência γ , então bastaria substituir o β por $\gamma\beta$. Assim, por exemplo, a afinidade com um filtro apenas é calculada com a seguinte expressão, logo após a replicação e seleção:

$$Aff_{1_{filtro}} = \frac{(1-\alpha)NA}{\alpha\beta NA + \beta NB}. \text{ Ao acrescentar outro filtro teremos a seguinte expressão:}$$

$$Aff_{2_{filtros}} = \frac{(1-\alpha)NA}{\alpha\beta\gamma NA + \beta\gamma NB} = \frac{Aff_{1_{filtro}}}{\gamma}, \text{ resultado este que está de acordo com os}$$

resultados simulados ($Aff_{2_{filtros}} \geq Aff_{1_{filtro}}$, pois $\gamma \leq 1$).

5.2 Cenário II

O segundo cenário possui muitas características distintas do primeiro. O cálculo da entropia e a seleção das sequências conservadas são as principais. Os resultados abaixo serão apresentados em figuras que contêm quatro gráficos cada. Em cada gráfico há cinco curvas, cada uma representando a inserção de mais um filtro. Assim, onde tiver “entropia3” ou “afinidade5” significa que a entropia foi computada com três filtros, enquanto a afinidade foi calculada utilizando cinco.

Apesar de haver diversos parâmetros que podem ser avaliados, será feita a análise dos mesmos parâmetros do cenário I: a taxa de mutação e a eficiência do filtro. Outros parâmetros foram estudados, tais como comprimento inicial das moléculas no sistema e quantidade inicial de moléculas.

5.2.1 Variando a taxa de mutação

Abaixo seguem os resultados com a taxa de mutação variando de gráfico para gráfico.

Figura 22.: Cenário II -Taxa de mutação 1%.

Análise de Resultados

Figura 23: Cenário II -Taxa de mutação 2%.

Figura 24: Cenário II -Taxa de mutação 5%.

Análise de Resultados

Figura 25: Cenário II -Taxa de mutação 10%

Figura 26: Cenário II -Taxa de mutação 40%.

Análise de Resultados

A afinidade, observadas nos quatro gráficos acima, diminui sistematicamente com o aumento da taxa de mutação. O comportamento da afinidade foi o mesmo, não dependendo da quantidade de filtros, o que indica ser a afinidade uma característica do sistema. Para uma taxa de 40%, diferentemente das outras taxas, tem-se a afinidade diminuindo com o passar dos rounds para todas as configurações de filtros.

A entropia diminuiu com o aumento da taxa de mutação, mostrando a importância de mecanismos mais precisos para replicação como fonte de moléculas mais adaptadas. Para taxa de 40%, a pequena entropia indica quase não haver informação sendo armazenada. Ao contrário da entropia do cenário I, a entropia do cenário II foi significativa, mostrando o comportamento esperado (maior mutação implicando menor variação na entropia).

O tamanho médio do genoma diminuiu em todos os experimentos numéricos realizados. Pode-se entender esse fato utilizando as ferramentas da Teoria de Informação descritas na seção 3.1. O sistema permitia haver quebras ou junções de moléculas. Dependendo das pressões do ambiente pode-se ter seleção por compressão ou seleção por transmissão. Na primeira pressão, o comprimento médio também diminui, enquanto na segunda o tamanho aumenta [36]. O gráfico a seguir ilustra um sistema sob seleção por compressão:

Figura 27: Comprimento médio das moléculas sob seleção por compressão.

Fonte: Adami, Ofria, Collier (2003), com permissão.

Análise de Resultados

Analizando a figura 27 acima e o tamanho médio obtido avaliando o cenário II, pode-se concluir que o sistema avaliado está de fato sob seleção por compressão. A quantidade de moléculas e o tamanho médio são quase complementares. Tem-se a seguinte relação entre eles:

$$(\text{Total_Nucleotídeos}) = (\text{tamanho médio}) * (\text{quantidade de moléculas}) + \text{Reservatório}$$

5.2.2 Variando a eficiência do filtro

A seguir, são apresentados os gráficos com eficiência do filtro diferentes.

Figura 28: Cenário II -Eficiência do filtro 20%.

Figura 29: Cenário II -Eficiência do filtro 80%.

Os gráficos 24, 28 e 29 correspondem aos mesmos parâmetros, apenas variando a eficiência do filtro. A afinidade também é dependente da eficiência do filtro, apesar de grandes variações na eficiência não corresponder a grandes mudanças na afinidade. A entropia, por outro lado, mostrou-se bastante sensível à alteração da eficiência, indicando que a permanência de moléculas não afins dificulta o armazenamento de informação no sistema. Vale notar que as moléculas não afins não são contabilizadas no cálculo da entropia. O tamanho médio e, por conseguinte, a quantidade de moléculas não sofreram alterações com a variação na eficiência do filtro.

A comparação entre os cenários permite observar uma medida robusta no sistema. A afinidade obteve os mesmos valores para as configurações correspondentes em cada cenário. O tamanho médio, no entanto, não aumentou em nenhum caso avaliado no segundo cenário, enquanto a variação no filtro permitiu um crescimento indeterminado no primeiro cenário.

Análise de Resultados

Outro resultado esperado foi o aumento do genoma. Acreditou-se que o aumento de filtro pudesse favorecer o aumento do número de nucleotídeos, retirando do sistema moléculas com poucos ou sem filtros. No entanto, esse resultado não foi encontrado no cenário II. Uma justificativa seria a não vantagem explícita em se ter mais de um filtro. Uma molécula que contenha dois filtros e se for submetido ao terceiro, então há uma probabilidade de a molécula ser eliminada caso não tenha a terceira região conservada, mesmo com dois filtros.

6 SUGESTÕES E PERSPECTIVAS

Esta dissertação finaliza apenas poucas etapas do projeto e uma ínfima parte do assunto. Assim, há muito que pode ser pesquisado. Passos seguintes que serão realizados pelo grupo e outros que ficam como sugestão de continuação (o grupo poderá realizar as sugestões *a posteriori* também).

A) Passos que meu grupo de pesquisa seguirá:

- 1) Incluir vantagens (menor chance de quebra, maior número de descendentes, menor chance de ser eliminada do sistema) para as moléculas que tiverem mais de uma sequência conservada.
- 2) Estudar a variação da pressão seletiva.
- 3) Estudar a difusão das moléculas e a interação de curto e longo alcance.
- 4) Fazer uso explícito dos diversos *fitness landscape* presentes na literatura.
- 5) Reproduzir alguns resultados clássicos, tal como a catástrofe do erro.

B) Sugestões

- 1) Estudar estrutura secundária
- 2) Estudar as sequências e os dendogramas relacionados.
- 3) Estudar outras formas de aumento de tamanho do genoma, tais como a inserção de elementos transponíveis.
- 4) Incluir explicitamente parâmetros utilizados na simulação biológica, tais como temperatura e pH.

7 CONSIDERAÇÕES FINAIS E CONCLUSÃO

Neste trabalho foram desenvolvidos programas para simular a evolução de macromoléculas. Esses programas trabalhavam com as moléculas propriamente ditas e não com simulação de equações. A vantagem em se simular dessa maneira é a flexibilidade e facilidade de explorar aspectos que as equações não contemplam. Como senão, o uso das moléculas exige um custo computacional muito grande.

Foi analisado o impacto de três parâmetros: a taxa de mutação, a eficiência do filtro e a quantidade de filtro. As variáveis analisadas apresentadas foram a Afinidade, a Entropia, a Quantidade de moléculas e o tamanho médio. Além disso, dois cenários foram explorados.

No primeiro cenário, a afinidade máxima média mostrou-se dependente da taxa de mutação e da eficiência do filtro. O aumento no número de filtro aumentou sistematicamente a afinidade por aumentar a retirada de moléculas não afins da população. O tamanho médio foi variável para os diferentes valores de mutação, mas aumentou indefinidamente para diferentes valores de eficiência do filtro. Isto pode ser entendido lembrando que, no cenário I, a sequência de seleção é também utilizada como fator de união das moléculas. A entropia mostrou-se pouco significativa para o cenário I porque foi considerada a molécula toda para o cálculo e, quando havia quebras, essas moléculas eram de tamanhos diferentes. Pôde-se concluir que quantidade diferente de filtro muda os valores finais, mas mantém o mesmo comportamento. Não houve combinação de parâmetros que tornasse o tamanho médio maior que o valor inicial sem crescer indeterminadamente. Os casos em que o tamanho médio diminuiu, pode-se considerar o sistema sob pressão por compressão. Para o caso de aumento indeterminado de genoma, não foi possível classificá-lo em nenhuma das três pressões citadas (compressão, seleção e neutral).

O cenário II não possui os problemas encontrados no primeiro cenário. Assim, a entropia foi calculada utilizando-se apenas a primeira sequência conservada (primeiro alvo), mesmo quando havia mais alvos. A entropia e a afinidade são bastante sensíveis à variação da taxa de mutação. Variando a mutação de 1% até 10% houve uma perda de

Considerações Finais e Conclusão

cerca de 12bits. Para taxa de 40%, há pouquíssima informação armazenada e apenas metade da população é afim. Para o cenário II também não houve crescimento do genoma. Assim, pode-se concluir que, sem haver outras vantagens evolutivas, tais como maior taxa de replicação ou menor chance de a molécula ser eliminada, o sistema não deixa de sofrer a seleção por compressão. Portanto, as novas etapas deverão contemplar explicitamente a aquisição de vantagens e incluir baixas taxas de mutação (maior quantidade de informação armazenada, mais afinidade).

REFERÊNCIAS BIBLIOGRÁFICAS*

- [1] ORÓ, J. Comets and formation of biochemical compounds on primitive Earth. **Nature**, v. 190, p. 380, 1961.
- [2] WOESE, C.R. A proposal concerning the origin of life on the planet Earth. **Journal of Molecular Evolution**, v. 13, pp. 95-101, 1979.
- [3] OPARIN, A.I. The origin and development of life. **NASA TTF**, v. 488, 1936.
- [4] OPARIN, A.I. The origin of life in space. **Space Science Reviews**, v. 3, p. 5-26, 1964.
- [5] HALDANE, J.B.S. A mathematical theory of natural and artificial selection. **Transactions Cambridge Philosophical Society**, v. 23, p. 19-41, 1929.
- [6] MOJZSIS, S.J.; ARRHENIUS, G., et al. Evidence for life on Earth before 3,800 million years ago. **Nature**, v. 384, p. 55-59.
- [7] SCHOPF, J. W. Microorganisms from late Precambrian of central Australia. **Science**, v. 150, n.394, p. 337-339, 1965.
- [8] EIGEN, M. Self organization of matter and the evolution of biological macromolecules. **Naturwissenschaften**, v. 56: 465, 1971.
- [9] EIGEN, M.; SCHUSTER, P. **The Hypercycle, a Principle of Natural Self-Organization**. Berlin, Germany: Springer-Verlag, 1979.
- [10] GILBERT, W. The RNA world. **Nature**, v. 319, p. 618, 1986.
- [11] CECH, T.R. Self-splicing RNA: implications for evolution. **International Review of Cytology**, v. 93, p. 3-22, 1985.
- [12] CECH, T.R et al. Biological catalysis of RNA. **Annual Review of Biochemistry**, v. 55, p. 559-629, 1986.
- [13] MÜLLIS, K.B. "The unusual origin of the polymerase chain reaction". **Scientific American**, v. 262, n. 4, p. 56-61, 1990.
- [14] JOYCE, G. F. Amplification, mutation and selection of catalytic RNA. **Gene**, v.82, p. 83-87, 1989.
- [15] ELLINGTON, A. D.; SZOSTAK, J. W. In vitro selection of RNA molecules that bind specific ligands. **Nature**, v. 346, p. 818-822, 1990.

* De acordo com:

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6023**: Informação e documentação: referências: elaboração. Rio de Janeiro, 2002.

Referências Bibliográficas

- [16] TUERK, C.; GOLD, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. **Science**, v. 3, p. 505–510, 1990.
- [17] ULRICH, H.; MAGDESIAN, M. H.; ALVES, M. J.; COLLI, W.; HENNING, U. In vitro selection of RNA aptamers that bind to cell adhesion receptors of *Trypanosoma cruzi* and inhibit cell invasion. **Journal of Biological Chemistry**, v. 277, p. 20756–20762, 2002.
- [18] TANG, Z. et al. Selection of Aptamers for Molecular Recognition and Characterization of Cancer Cells. **Analytica Chemistry**, v. 79, p. 4900–4907, 2007.
- [19] ZHANG, Z. Advances in SELEX and application of aptamers in the central nervous system. **Biomolecular Engineering**, v. 24, p. 583–592, 2007.
- [20] ZUKER M.; SANKOFF, D. RNA secondary structures and their prediction. **Bulletin Mathematical Biology**, v. 46, p. 591–621, 1984.
- [21] FLAMM, C.; FONTANA, W.; HOFACKER, I. L.; SCHUSTER, P. Elementary step dynamics of RNA folding. **RNA**, v. 6, p. 325–338, 1999.
- [22] KAUFFMAN, S.A. Autocatalytic sets of proteins. **Journal of Theoretical Biology**, v. 119, p. 1–24, 1986.
- [23] KAUFFMAN, S. A. **The Origins of Order**. New York: Oxford University Press, 1993.
- [24] LEVINE, H.; NILSEN-HAMILTON, M. A. A mathematical analysis of SELEX. **Computational Biological Chemistry**, v. 31, p. 11–35, 2007.
- [25] WALLACE, R.; WALLACE, R. Information Theory, Scaling Laws and the Thermodynamics of Evolution. **Journal of Theoretical Biology**, v.192, p. 545–559, 1998.
- [26] LUQUE, B. An Introduction to Physical Theory of Molecular Evolution. **Central European Journal of Physics**, v. 3, p. 516–555, 2003.
- [27] JOYCE, G.F. RNA Evolution and origins of life. **Nature**, v. 338, p. 217–224, 1989.
- [28] ORGEL, L.E. Prebiotic Chemistry and the Origin of the RNA World. **Critical Reviews in Biochemistry and Molecular Evolution**, v. 33, p. 99–123, 2004.
- [29] SHANNON, C. E. “A mathematical theory of communication”, **Bell System Technical Journal**, v. 27, p. 379–423, 623–656, 1948.
- [30] SCHNEIDER, T. D. Evolution of biological information. **Nucleic Acids Reserch**, v. 28, p. 2794–2799, 2000.

Referências Bibliográficas

- [31] NUSSENZVEIG, H. M. (Org.). **Complexidade e Caos**. Rio de Janeiro: Editora UFRJ/COPEA, 2003.
- [32] ADAMI, C. “Information Theory in Molecular Evolution”. **Physics of Life Reviews**, v. 1, 2004.
- [33] AYNECHI, T.; KUNTZ, I. D. An information theoretic approach to macromolecular modeling: I. Sequence alignments. **Biophysics Journal**, v. 89, p. 2998-3007, 2005.
- [34] BLAST. Disponível em: <<http://blast.ncbi.nlm.nih.gov/Blast.cgi>>. Acesso em: 22 mar. 2009.
- [35] CAVALIER-SMITH, T. **The evolution of genome size**. Chichester: Wiley, 1995.
- [36] OFRIA, C.; ADAMI, C.; COLLIER, T. Selective Pressures on Genomes in Molecular Evolution. **Journal of Theoretical Biology**, v. 222, p. 477-483, 2003.
- [37] COVER, T. et al. **Elements of Information Theory**. New York: John Wiley, 2001. (Wiley Series).
- [38] BAILON, P. et al. **An Overview of Affinity Chromatography**. Totowa, NJ: Humana Press, 2000.

APÊNDICE A

Desenvolvimento da equação do crescimento

Abaixo está apresentada a forma como foi determinada a equação do crescimento exposta na seção 5.1.2. Os símbolos utilizados são os seguintes:

$Q(t)$.: Quantidade de moléculas após t ciclos (rounds)

t .: Número de ciclos (rounds)

α ..erro na replicação

β .:Eficiência do filtro (fração das moléculas não-afins que permanecem no sistema)

NA .:Quantidade de moléculas afins na população inicial

NB .:Quantidade de moléculas não-afins na população inicial

1-diff.: fator de diminuição da população, para manter a população máxima constante.

K .: constantes auxiliares.

Serão descritos três ciclos completos e, então, será feita a generalização. Será desenvolvida a equação para a configuração com apenas um filtro e sem quebras. Vale ressaltar que a inclusão de novos filtros ou da parcela de transformação de moléculas não-afins em afins é trivial, como fora mostrado. Onde tiver indicado “PCR” será feita a duplicação com erro α . O “FILTRO” eliminará parte das moléculas não afins, enquanto o “CP” manterá a população constante.

Ciclo I

1) $Q(0) = NA + NB$ (quantidade inicial)

2) PCR

3) $Q\left(\frac{1}{2}\right) = 2(1 - \alpha)NA + \underbrace{2\alpha NA + 2NB}_{N\tilde{A}O-AFINS}$ (moléculas após duplicação)

4) CP

5) $diff_1 = \frac{Q\left(\frac{1}{2}\right) - popMax}{Q\left(\frac{1}{2}\right)}$ (fator de diminuição da população)

5) $Q\left(\frac{3}{4}\right) = 2(1 - diff_1)(1 - \alpha)NA + 2(1 - diff_1)(\alpha NA + NB)$ (quantidade menor ou igual à população máximo.)

6) FILTRO

7) $Q(1) = 2(1 - diff_1)(1 - \alpha)NA + 2(1 - diff_1)(\beta\alpha NA + \beta NB)$ (fim do ciclo I)

Apêndice A

Ciclo II

$$1) Q(1) = \underbrace{2(1 - diff_1)}_{K_1} [(1 - \alpha) NA + (\beta \alpha NA + \beta NB)] \quad (\text{moléculas. No}$$

início do ciclo II).

2) PCR

$$3) Q(1 + \frac{1}{2}) = K_1 [2(1 - \alpha)^2 NA + 2\alpha(1 - \alpha)NA + 2(\beta \alpha NA + \beta NB)]$$

4) CP

$$5) diff_2 = \frac{Q(1 + \frac{1}{2}) - popMax}{Q(1 + \frac{1}{2})} \quad (\text{Fator de diminuição da população})$$

6)

$$Q(1 + \frac{3}{4}) = K_1 \underbrace{2(1 - diff_2)}_{K_2} [(1 - \alpha)^2 NA + \alpha(1 - \alpha)NA + \beta \alpha NA + \beta NB]$$

7) FILTRO

$$8) Q(2) = K_1 K_2 [(1 - \alpha)^2 NA + \alpha(1 - \alpha)\beta NA + \alpha\beta^2 NA + \beta^2 NB] \quad (\text{fim do ciclo II})$$

Ciclo III

$$1) Q(2) = K_1 K_2 [(1 - \alpha)^2 NA + \alpha(1 - \alpha)\beta NA + \alpha\beta^2 NA + \beta^2 NB]$$

2) PCR

3)

$$Q(2 + \frac{1}{2}) = K_1 K_2 [2(1 - \alpha)^3 NA + 2\alpha(1 - \alpha)^2 NA + 2\alpha(1 - \alpha)\beta NA + 2\alpha\beta^2 NA + 2\beta^2 NB]$$

4) CP

5)

$$Q(2 + \frac{3}{4}) = K_1 K_2 \underbrace{2(1 - diff_3)}_{K_3} [(1 - \alpha)^3 NA + \alpha(1 - \alpha)\beta NA + \alpha(1 - \alpha)^2 NA + \alpha\beta^2 NA + \beta^2 NB]$$

6) FILTRO

7)

$$Q(3) = K_1 K_2 K_3 [(1 - \alpha)^3 NA + \alpha(1 - \alpha)\beta^2 NA + \alpha(1 - \alpha)^2 \beta NA + \alpha\beta^3 NA + \beta^3 NB]$$

Ciclo t

Observando o padrão nos três ciclos anteriores, pode-se notar que a expressão geral será:

$$Q(t) = \prod_{j=0}^t K_j \left[(1 - \alpha)^t NA + \beta^t NB + \alpha NA \sum_{i=1}^t (1 - \alpha)^{t-1} \beta^{t+1} \right]$$

Substituindo os valores de K e notando que $\prod_0^t 2 = 2^t$, tem-se:

$$Q(t) = 2^t \prod_{j=0}^t (1 - diff_j) \left[(1 - \alpha)^t NA + \beta^t NB + \alpha NA \sum_{i=1}^t (1 - \alpha)^{t-1} \beta^{t+1} \right]$$

APÊNDICE B

Exemplo de gráficos auxiliares para análise do cenário II

Para análise dos resultados do cenário II (seção 5.2) foram utilizados diversos gráficos, além dos apresentados. Abaixo, serão apresentados os gráficos para a configuração com 5% de mutação e 50% de eficiência do filtro. O gráfico dos histogramas é tridimensional e sua visualização e análise são melhores quando feitas diretamente no software GNUPLOT, pois é possível girar o gráfico em todas as direções e sentidos, além de poder aproximar ou afastar os gráficos (fazer “zoom”). O intuito de mostrar estes gráficos é demonstrar outras maneiras e ferramentas possíveis e úteis na análise dos resultados.

Os gráficos apresentados serão o do Histograma do tamanho do genoma, do Coeficiente de Desvio dos tamanhos, da Distância Média de Hamming e o da Quantidade de Nucleotídeos no Reservatório. Durante o estudo, foi avaliado gráfico por gráfico e também o conjunto.

Figura 1: Coeficiente de desvio do tamanho do genoma. Taxa de mutação 5%

Figura 2. Histograma do tamanho do genoma. Para saber o tamanho limite de cada classe de tamanho basta multiplicar o valor do gráfico por 10.

Figura 3. Quantidade de nucleotídeos no reservatório. Esses são os nucleotídeos que podem ser usados na replicação.

Figura 4.: Distância de Hamming média. Refere-se somente às sequências com a mínima distância de Hamming de cada molécula com afinidade. Essas foram as moléculas utilizadas no cálculo da entropia.