Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Curso de Ciência da Computação

THIAGO HENRIQUE LEMOS FONSECA

ESTUDO DA DERIVA GENÉTICA UTILIZANDO O ALGORITMO PAGERANK

São Luís
2015

THIAGO HENRIQUE LEMOS FONSECA

ESTUDO DA DERIVA GENÉTICA UTILIZANDO O ALGORITMO PAGERANK

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Profº Valeska Martins de Souza

São Luís

2015

Fonseca, Thiago Henrique Lemos

Estudo da deriva genética utilizando o algoritmo PageRank/ Thiago Henrique Lemos Fonseca. – 2015.

50 f.

Orientador: Valeska Martins de Souza

Monografia (Graduação) – Universidade Federal do Maranhão, Curso de Ciência da Computação, 2015.

1. Cadeias de Markov 2. Pagerank 3. Modelo de Wright-Fisher 4. Deriva Genética I. Estudo da deriva genética utilizando o algoritmo PageRank

CDU 004.056.53

THIAGO HENRIQUE LEMOS FONSECA

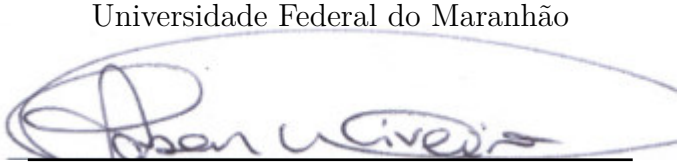
ESTUDO DA DERIVA GENÉTICA UTILIZANDO O ALGORITMO PAGERANK

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 2 de julho de 2015.

BANCA EXAMINADORA

Profª Dra Valeska Martins de Souza
(Orientador)
Universidade Federal do Maranhão



Profº Dr Alexandre César Muniz de Oliveira
Universidade Federal do Maranhão

Profº Msc Carlos Eduardo Portela Serra de Castro
Universidade Federal do Maranhão

Aos meus pais e irmãos,
que são meu esteio...

Agradecimentos

Agradeço primeiramente a Deus por ter me dado força em cada dia da minha vida para que eu pudesse caminhar e estudar, assim chegando hoje onde estou.

Agradeço aos meus familiares. Às minhas tias Marta, Rosimar, Isabel e minha avó Terezinha, por terem cuidado de mim desde pequeno e me ensinado a ser um homem íntegro, preservando verdadeiros valores. À minha avó Nazeth e minha tia Rosane, por serem sempre presentes e terem dado tanto amor e carinho a mim e meu irmão. Ao meu avô Olavo que mesmo distante sempre me apoiou nas decisões difíceis. À minha tia Socorro, pelos conselhos e confiança no meu sucesso. Ao meu tio João, pelas brincadeiras, conselhos e pela amizade de irmão mais velho.

Agradeço aos meus pais, em especial a meu pai, por ser um exemplo de homem trabalhador a ser seguido, que mesmo diante das adversidades, nunca desistiu e deu a volta por cima.

Agradeço aos meus irmãos, por serem sempre atenciosos e prestativos, especial ao meu irmão Victor que esteve sempre comigo nessa longa caminhada rumo a excelência.

Agradeço a minha namorada, Andressa Rozental, por estar sempre presente desde o início da minha jornada acadêmica, juntos superamos crises e brindamos conquistas.

Agradeço aos meus amigos de longa data, Katyane e Vanessa, pois estiveram comigo desde o começo, quando ainda estávamos no ensino fundamental, e provaram ao longo dos anos serem amigas leais e verdadeiras.

Agradeço a minha orientadora, Valeska Martins de Souza, pelas conversas, orientações e principalmente por ter me acolhido e confiado no meu potencial, serei eternamente grato por tudo.

Agradeço aos professores da Universidade Federal do Maranhão, pois sem eles, não estaria aqui sentindo-me mais confortável em redigir assuntos e temas da área de Computação.

"Pois estou convencido de que nem a morte, nem a vida,
nem os anjos, nem os principados, nem as coisas do presente,
nem do porvir, nem os poderes, nem a altura, nem a profundidade,
nem qualquer outra criatura poderá separar-nos do amor de Deus,
que está em Cristo Jesus, nosso Senhor."
(Bíblia Sagrada, Romanos 8:38,39)

RESUMO

Nesta monografia, é proposta uma nova metodologia para o mapeamento de frequências alélicas sob o efeito da deriva genética baseadas no software de ranking de páginas do Google, o Pagerank. Inicialmente obtém-se uma implementação do algoritmo Pagerank utilizando Cadeias de Markov em tempo discreto aliado à equação de Chapman-Kolmogorov como medida de otimização para a análise de páginas em rede. A partir do algoritmo construído, o problema é modificado para a análise de frequências alélicas com a pertubação da matriz de transição e o modelo de previsão biológico de Wright-Fisher. Uma das principais vantagens desta metodologia é a capacidade de prever a configuração alélica de uma população futura afetada pela deriva, auxiliando medidas preventivas contra possíveis perdas de variabilidade genética ou extinção de espécies em ambientes de seleção neutra por parte de pesquisadores na área de fenômenos biológicos e otimização.

Palavras-chaves: Cadeias de Markov, Pagerank, Modelo de Wright-Fisher, Deriva Genética.

ABSTRACT

In this work, a new methodology for mapping allele frequencies under the effect of genetic drift based on software Google page ranking, Pagerank, is proposed. Initially, we obtain an implementation of the Pagerank algorithm using Markov chains in discrete-time coupled with the Chapman-Kolmogorov equation as optimization measure for network pages of analysis. From the algorithm built, the problem is modified for analysis of allele frequencies with disturbance of transition matrix and the biological forecasting model of Wright-Fisher. One of the main advantages of this methodology it is the ability to predict the allelic configuration of a future population affected by drift, helping preventive measures against possible loss of genetic variability or extinction of species in the neutral selection of environments by researchers in the phenomena area biological and optimization.

Keywords: Markov Chains, Pagerank, Wright-Fisher Model, Genetic Drift.

Lista de ilustrações

Lista de códigos

Lista de tabelas

Sumário

1 Introdução

A deriva genética ou alélica é um mecanismo microevolutivo que ocasiona mudanças nas frequências dos alelos devido a erros de amostragem, sua influência está relacionada com a redução da diversidade genética e extinção de alelos em ambientes de seleção neutra. A deriva é um processo estocástico, pois não é possível prever a direção da mudança na frequência de um alelo causado pela mesma, por isso possui capacidade de modificar as frequências alélicas em uma única geração, mas seu poder como mecanismo evolutivo aumenta significativamente quando seus efeitos se compõem ao longo de várias gerações.

A importância da deriva encontra-se em sua capacidade de alterar o fluxo gênico por meio de eventos aleatórios, sua influência é afetada por fatores como tamanho da população, distribuição geográfica, desastres ecológicos entre outros. O geneticista Peter Buri estudou este fenômeno empiricamente em pequenas populações de moscas de frutas e seu artigo (BURI, 1956) tornou-se referência padrão de prova experimental na deriva genética.

1.1 Objetivos

Esta Monografia terá por objetivo propor uma nova metodologia para a análise de frequências alélicas sobre o efeito da deriva utilizando Cadeias de Markov e o modelo binomial de Wright-Fisher amplamente utilizado em genética de populações, com o intuito de permitir o acesso rápido aos dados pertinentes ao estudo, bem como promover um método de baixo custo para futuros pesquisadores na área de modelagem de fenômenos biológicos. Para mapear as frequências ao longo das gerações, será desenvolvida uma aplicação denominada Biopagerank, um algoritmo baseado no software de análise de páginas web do mecanismo de busca Google (PAGE, 1998) .

1.2 Organização da Monografia

Esta Monografia está organizada em 5 capítulos, a saber:

Capítulo 1: Introdução

Neste capítulo, é definido o problema principal e seu grau de importância, bem como o método proposto que será aplicado e implementado para sua resolução.

Capítulo 2: Fundamentação Teórica

Neste capítulo, é feita uma revisão de alguns fundamentos teóricos relacionados com cadeias de Markov e genética, fornecendo base matemática para o entendimento dos

capítulos seguintes.

Capítulo 3: Modelagem do algoritmo Pagerank

Este capítulo apresenta a modelagem e implementa a versão inicial do Pagerank com o objetivo de evidenciar a utilização das cadeias de Markov em sua estrutura e entender o mecanismo de análise de páginas visando a modificação de seu domínio de dados.

Capítulo 4: Modelagem do Pagerank Genético

Este capítulo mpresenta a modelagem e implementa o software baseado no algoritmo de ranking de páginas, chamado Biopagerank, com objetivo de formalizar o modelo proposto em uma aplicação executável e eficiente.

Capítulo 5: Conclusão

Este capítulo resume as contribuições desta monografia e indica direções para pesquisas futuras.

2 Fundamentação Teórica

2.1 A Deriva Genética

Esta seção introduz conceitos preliminares sobre a deriva genética, sua influência e importância para o processo de evolução dos seres vivos.

2.1.1 Equilíbrio de Hardy-Weinberg

A base genética de populações é definida pelo *princípio de Hardy-Weinberg*, demonstrado pelo matemático inglês G. H. Hardy e pelo fisiologista alemão W. Weinberg, em 1908. Este princípio é abordado na definição seguinte.

Definição 1. *Para qualquer lócus gênico, as frequências relativas dos genótipos, em populações de cruzamentos ao acaso (panmíticas), permanecem constantes, de geração a geração, a menos que certos fatores perturbem esse equilíbrio.*

Esses fatores são *mutação*, *seleção natural*, *migração* e *deriva genética*. Logo, alguns requisitos devem ser atendidos para que uma população seja dita em equilibrio de Hardy-Weinberg. Estes requisitos são (BORGES-OSORIO, 2013):

a) A população deve estar em ambiente neutro;

b) A população deve ser panmítica (os cruzamentos entre indivíduos de diferentes genótipos devem ocorrer ao acaso, sem qualquer preferência);

c) A relação entre indivíduos do sexo masculino e feminino deve ser 1:1;

d) A população deve ser mendeliana, ou seja, formada por organismos da mesma espécie, com limites geográficos bem definidos;

O estudo deste princípio tem grande importância na avaliação de fatores que ocasionam desvios reais opostos às condições idealizadas (BORGES-OSORIO, 2013).

Dado um conjunto gênico hipotético em que qualquer gameta masculino tem igual probabilidade de se unir a um feminino, as frequências esperadas nos zigotos da geração seguinte podem ser calculadas se a frequência dos alelos A e a considerados for conhecida. Sendo p a frequência do alelo A e q a frequência do alelo a, caso ocorra o cruzamento entre dois heterozigotos para o mesmo lócus ($Aa\ X\ Aa$), a distribuição genotípica é dada pela Tabela 1 (BORGES-OSORIO, 2013).

Tabela 1 – Aa X Aa.

Masc|Fem	A(p)	a(q)
A(p)	AA(p^2)	Aa(pq)
a(q)	aa(pq)	aa(q^2)

Fonte: (BORGES-OSORIO, 2013)

Essas frequências se manterão contantes desde que os requisitos impostos pelo equilíbrio sejam satisfeitos. Se as frequências continuarem contantes não haverá evolução, logo os fatores que perturbam o equilíbrio de Hardy-Weinberg são também responsáveis pela evolução.

2.1.2 Introdução à deriva

A evolução consiste no desenvolvimento de adaptações através da seleção natural e muitos outros fatores, incluindo o acaso. Mesmo que restrinjamos nossa atenção à evolução através da seleção natural, existem numerosos fatores seletivos além daqueles impostos pelo mundo ecológico externo. Dentre eles podemos citar: relações internas entre caminhos bioquímicos e do desenvolvimento, e as relações internas entre diferentes órgãos. O ambiente também possui sua influência no processo evolutivo. Aspectos como salinidade, disponibilidade de água, membros com os quais o indivíduo irá interagir em vários contextos sociais e outras características físicas e químicas podem interferir de maneira positiva ou negativa na adaptação de uma determinada espécie (FUTUYAMA, 2002) .

Basicamente, existem dois tipos de mecanismos evolutivos, os que criam variações e os responsáveis pela distribuição dessas variações. Os que criam variações são *mutação* e *recombinação gênica*. Os responsáveis pela distribuição dessas variações são *seleção natural*, *migração* e *deriva genética*.

O processo de variação genética inicia-se na *mutação*. Uma mutação é qualquer variação no DNA de um organismo. De característica aparentemente aleatória, sua influência geralmente é danosa ou neutra aos seus portadores, permitindo a criação e manutenção da variação genética das populações. A taxa de ocorrência da mutação é geralmente baixa, porém suficiente para modicar a frequência dos alelos, uma vez que um grande número de genes podem ser afetados (SADAVA et al., 2009).

O *fluxo gênico* ou *migração* ocorre quando indivíduos ou gametas deslocam-se para outras localidades e se reproduzem, adicionando alelos ao *pool gênico* da população local ou modificando suas frequências alélicas (SADAVA et al., 2009).

Dentre os mecanismos evolutivos, a *deriva genética* pode ser considerada um dos mais interessantes por sua característica matemática singular, exercendo função de mecanismo micro evolutivo que modifica aleatoriamente as frequências alélicas ao longo do

tempo. A deriva é um processo estocástico, como consequência de sua influência podemos destacar a perda de variação genética e a fixação de alelos em diferentes loci. Seu estudo é de suma importância em programas de melhoramento genético, coleta e regeneração de germoplasma, estudos evolutivos, entre outros.

> Deriva, em qualquer de suas formas, é uma propriedade estatística de um conjunto de ensaios ou eventos: deriva é erro estatístico. Uma série de nascimentos, sobrevivências, mortes, e reproduções manifestam seus resultados - mensurados como mudanças nas frequências-divergem do que o previsto por diferenças de aptidão. (WALSH, 2002)

A deriva genética pode ser matematicamente representada como um processo de mudanças aleatórias nas frequências dos alelos ao longo das gerações. A deriva é modelada quantitativamente utilizando o *modelo binomial de Wright-Fisher*.

Conforme Ewens (2004), dado um locus contendo dois alelos neutros a e A em uma população diploide de tamanho N. Temos, assim, o tamanho da população haploide (número de cópias do gene) igual a $2N$. O número de alelos A na população na geração t é denotado como $n(t)$ e sua frequência como $p(t) = n(t)/2N$. A probabilidade de transição do estado $n(t) = i$ para o estado $n(t+1) = j, 0 \leq i, j \leq 2N$ é dado pela seguinte probabilidade condicional:

$$P_{ij} = Pr[n(t+1) = j | n(t) = i)] \tag{2.1}$$

A Equação 2.1 pode ser modelada utilizando o *Wright-Fisher*, como definido na próxima subseção.

2.1.3 Modelo de Wright-Fisher

Esta subseção apresenta uma introdução acerca de modelo de distribuição de Wright-Fisher utilizando as definições de Durrett (2008).

Quando considera-se um locus genético com dois alelos que possuem o mesmo *fitness* em uma população diploide de tamanho constante N com ausência de sobreposição, alguns conceitos devem ser explicados.

Um *locus genético* é um local específico no genoma do organismo, uma sequência de nucleotídeos que compõe um gene.

Alelos são diferentes versões da informação genética codificada em um locus.

O *fitness* de um indivíduo é uma mensuração da habilidade individual de sobrevivência e produção de descendentes. Nesta monografia são considerados casos de *evolução neutra*, ou seja, a mutação muda a sequência de DNA, mas não interfere no fitness.

Indivíduos diploides possuem duas cópias de seu material genético em cada célula, ou seja, N indivíduos possuem $2N$ cópias.

O modelo de Wright-Fisher é uma distribuição binomial representada por:

$$\begin{pmatrix} 2N \\ j \end{pmatrix} \left(\frac{i}{2N}\right)^{j} \left(1 - \frac{i}{2N}\right)^{2N-j} \tag{2.2}$$

em que i/ 2N é a probabilidade do alelo A se fixar na geração seguinte, e 1-(i/2N) é a probabilidade do alelo a se fixar. O coeficiente binomial

$$\begin{pmatrix} 2N \\ j \end{pmatrix} = \frac{(2N)!}{j!(2N-j)!}, \tag{2.3}$$

é o número de modos de escolha de j e 2N é o número total de elementos.

Probabilidade de fixação

Eventualmente, o número de alelos em uma população X_n pode torna-se 0 (indicando a perda total do alelo) ou 2N (indicando a fixação total do alelo). Quando um alelo se perde ou se fixa em uma população, o mesmo nunca mais retorna ou se extingue respectivamente, ou seja, os estados 0 e 2N são denominados *estados absorventes* para X_n esta característica só é possível se o ambiente utilizado for evolutivamente neutro.

$$\tau = min[n : X_n = 0 \quad ou \quad X_n = 2N] \tag{2.4}$$

A Equação (2.4) representa o tempo de fixação, ou seja, o momento exato na história da evolução em que a população consiste apenas de A ou a.

Usa-se P_i para denotar a probabilidade de distribuição do processo X_n, quando inicia-se $X_0 = i$ e E_i para o valor esperado com respeito a P_i.

Teorema 1. *No modelo de Wright-Fisher, a probabilidade de fixação do alelo A é dada por:*

$$P_i(X_t = 2N) = \frac{i}{2N} \tag{2.5}$$

Demonstração. Segundo Xavier Didelot (2014), se o número de indivíduos é finito é possível que eventualmente uns dos alelos estudados se tornará fixo. Dado X_n ser o número de *A's* no tempo n, então a Equação (2.2) é $2Np$ (sendo p a probabilidade de um alelo se fixar e q a probabilidade do outro alelo que compõe o gene se fixar).

$$E(X_{n+1}|X_n = i) = 2N\left(\frac{i}{2N}\right) = i = X_n \tag{2.6}$$

O valor esperado $EX_{n+1} = EX_n$, ou seja, a média X_n mantem-se constante no tempo. Intuitivamente, temos que:

$$i = E_i X_\tau = 2NP_i(X_\tau = 2N) \tag{2.7}$$

Para provar isto, tem-se que se $X_n = X_\tau$ quando $n > \tau$,

$$i = E_i X_n = E_i(X_\tau; \tau \leq n) + E_i(X_n; \tau > n) \tag{2.8}$$

em que $E(X, A)$ é pequeno para o valor esperado de X sobre o conjunto A. Então dado $n \to \infty$ e usando o fato que $|X_n \leq 2N|$, conclui-se que o primeiro termo converge para $E_i X_\tau$ e o segundo para 0.

□

Heterozigosidade

Para medir o quão longe a variabilidade genética pode ir, deve-se analisar a *heterozogosidade*, definida como a probabilidade das duas cópias do locus escolhidas no tempo n serem diferentes:

$$H_n^0 = \frac{2X_n(2N - X_n)}{2N(2N - 1)} \tag{2.9}$$

Teorema 2. *Dado $h(n) = EH_n^0$ como o valor médio de heterozogosidade no tempo n. No modelo de Wright-Fisher:*

$$h(n) = \left(1 - \frac{1}{2N}\right)^n h(0) \tag{2.10}$$

Demonstração. Para facilitar a prova é conveniente referir-se as 2N cópias do locus como indivíduos. Suponha que escolhamos dois indivíduos denominados $x_1(0)$ e $x_2(0)$ no tempo *n*. Cada indivíduo $x_i(0)$ é um descendente de algum indivíduo $x_i(1)$ no tempo *n-1*, que é descendente de $x_i(2)$ no tempo *n-2* e assim sucessivamente. $x_i(m), 0 \leq m \leq n$ descreve a linhagem de $x_i(0)$.

Se $x_1(m) = x_2(m)$, então teremos $x_1(l) = x_2(l)$ para $m < l \leq n$. Se $x_1(m) \neq x_2(m)$, então as duas escolhas de parentes são feitas independentemente, logo $x_1(m+1) \neq x_2(m+1)$ com probabilidade $1-(1/2N)$. Para $x_1(n) \neq x_2(n)$, diferentes parentes podem ser escolhidos em todos os tempos $1 \leq m \leq n$, em um evento com probabilidade $(1-1/2N)^n$. Quando duas linhagens evitam uma a outra, $x_1(n)$ e $x_2(n)$ são dois indivíduos escolhidos aleatoriamente na população no tempo *0*, logo a probabilidade de eles serem diferentes é H_0^o=h(0). (Xavier Didelot, 2014) □

2.2 O Pagerank

O Pagerank é uma técnica utilizada para identificar a importância de páginas web através da modelagem da rota tomada pelas pessoas enquanto navegam entre as páginas.

O algoritmo Pagerank é visto como a principal razão para o sucesso do motor de busca *Google* (LANGVILLE, 2012). Para entender como o Pagerank age, primeiro deve-se analisar como um mecanismo de busca funciona. Referências para mais informações são encontradas em Witten et al. (1999) e Manning et al. (2008).

Crawling

Antes de um mecanismo de busca retornar os resultados da pesquisa feita pelo usuário, deve-se ter um conjunto de resultados possíveis. Ao processo de identificar essas páginas web dá-se o nome de *crawling*. Na rede web, hiperlinks conectam as páginas e crawlers descobrem novas páginas para esses links, este processo pode continuar indefinidamente, uma vez que novas páginas são continuamente inseridas na web.

Análise de texto

Para que o resultado da pesquisa seja eficaz, é necessário que o mecanismo de busca consiga entender o texto nas páginas. A análise textual é feita através de um processo de indexação em que é gerado uma base de dados contendo todas as páginas com o termo procurado. Frases também são indexadas, melhorando a eficiência do processo.

Análise de links

Um mecanismo de busca deve analisar os links entre as páginas indexadas em seus servidores para determinar a *qualidade* da página. O Pagerank é uma das medidas utilizadas para este fim. O algoritmo identifica páginas que possuem um grande número de votos (páginas que são referenciadas em grande quantidade) e atribui um valor de relevância para estas páginas baseando-se nesse e outros aspectos. Um analisador de links também deve ser capaz de diferenciar páginas originais de possíveis fraudes (spam).

Regressão de ranking

As análise de texto e link são utilizadas para determinar uma classificação com a ordem final dos resultados. A função responsável por gerar esta classificação é o principal segredo dos mecanismos de busca e muitas vezes são baseadas em técnicas de aprendizado de máquina.

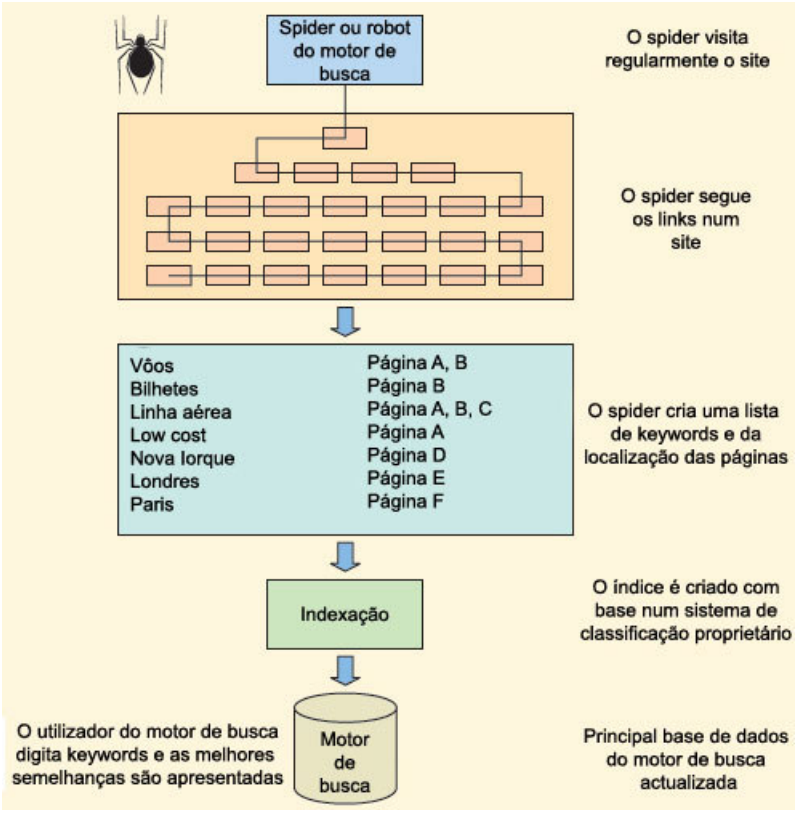


Figura 1 – *Crawling*: funcionamento

Fonte: Ascensão (2014)

Produção do ranking

O último passo é integrar todas as análises anteriores sobre um conjunto de documentos produzindo uma lista de páginas que será exibida para o usuário em milisegundos.

A Figura 1 ilustra o processo de análise realizado pelo Crawler.

O Pagerank, então, é um algoritmo de análise de link utilizado por um mecanismo de busca para ajudar na classificação de páginas web.

O algoritmo Pagerank pode ser interpretado como um grafo direcionado em que as páginas são representadas por *nós* e os links entre as páginas são representados por *arestas*.

A Figura 2 mostra o relacionamo entre páginas aleatórias da Wikepedia. A abstração do cálculo do Pagerank como um grafo facilita na implementação e modelagem, uma vez que grafos são constantemente tratados em fenômenos computacionais.

O modelo de grafo do Pagerank pode ser generalizado para qualquer grafo arbitrário permitindo várias interpretações. Como exemplo de usos para o algoritmo pode-se citar:

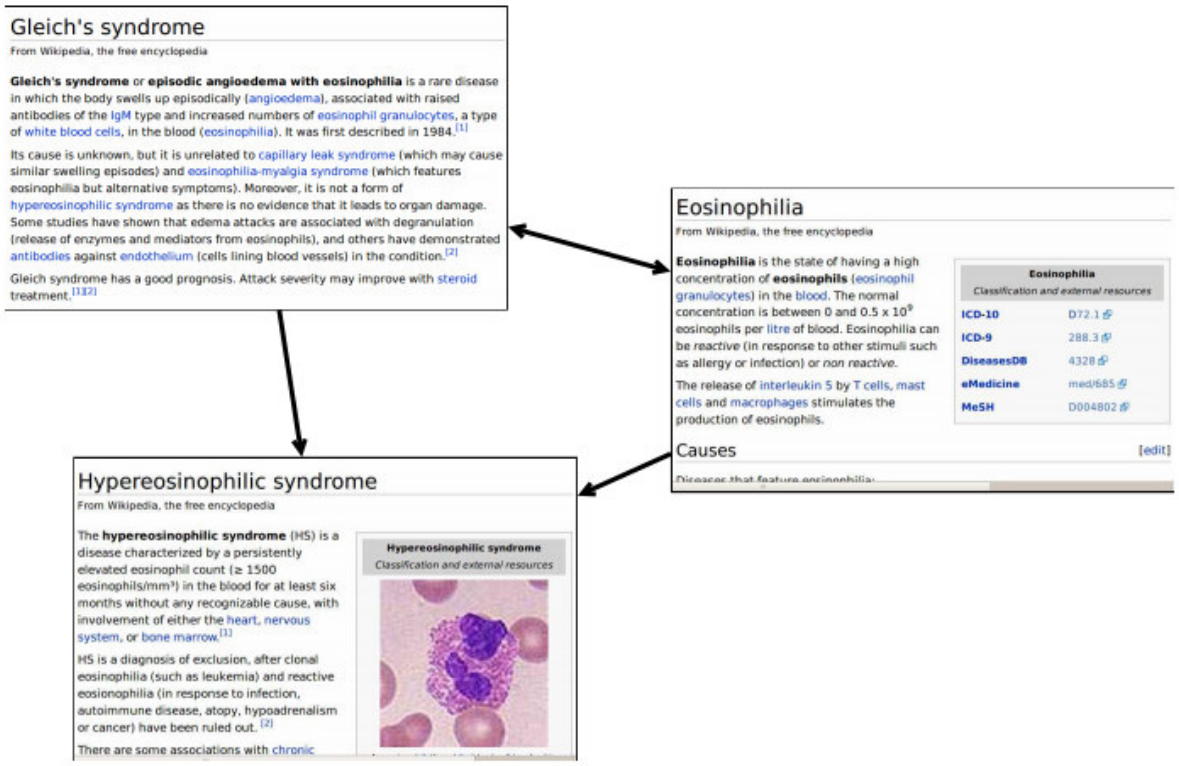


Figura 2 – *Páginas Web e Grafos*

Fonte: Gleich (2009)

Clustering

O problema de clusterização concentra-se em encontrar formas de dividir um grafo separando os nós em grupos fortemente relacionados, pode-se utilizar o Pagerank para calcular esses grupos de nós de forma extremamente rápida (ANDERSEN et al., 2006).

Ranks esportivos

O Pagerank pode ser usado para computar rank's esportivos substituindo um usuário navegando aleatoriamente pela internet por times e conjuntos de estatísticas sobre os mesmos. A partir desses dados, o modelo pode inferir qual time possui maior chance de passar para a fase seguinte (GOVAN et al., 2008).

Bioinformática

O Generank é uma modificação no algoritmo Pagerank que produz uma lista de genes com caracteristicas relevantes para os experimentos para os quais está sendo usado. Para isso o Generank analisa as relações entre os genes para inferir níveis de ativação e outros aspectos necessários para os pesquisadores (MORRISON et al., 2005).

Partindo-se da ideia do algoritmo Pagerank como um grafo direcional G com a

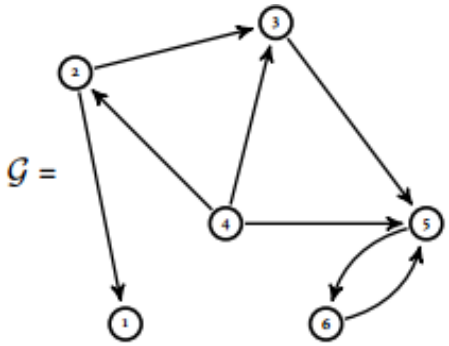
matriz adjacente A ($A_{i,j} = 1$ se o nó i tem ligação com o nó j e $A_{i,j} = 0$ se não há ligação). O vetor pagerank é definido aplicando um algoritmo estocástico para a matriz, o qual representa o comportamento a longo prazo do sistema discreto.

$$P = A^T D^+, \tag{2.11}$$

em que D^+ é uma matriz diagonal com entradas D_{ii} = grau de saída do nó i e D^+ definida por Golub e Loan (1996) da seguinte maneira:

$$(D^+)_{ii} = \begin{cases} 1/D_{ii} & D_{ii} \neq 0 \\ 0 & D_{ii} = 0 \end{cases} \tag{2.12}$$

Cada página web é representada por um nó no grafo, o nó u é conectado com o nó v se, e somante se, existe um link da página u para a página v.



$$\mathbf{P} = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 1/2 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Figura 3 – *Representação do grafo como matriz estocástica*

Fonte: Gleich (2009)

Baseado na Figura 3, podemos definir um vetor V que representa a importância das páginas no momento inicial. Suponha que $V = [1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6]$, ou seja, todas as páginas tem inicialmente o mesmo grau de importância. Usando o modelo do PageRank, chega-se ao vetor estacionário

$$x = [0.049 \quad 0.041 \quad 0.059 \quad 0.032 \quad 0.425 \quad 0.394] \tag{2.13}$$

A Equação (2.13) representa o vetor Pagerank com o grau de relevância de cada página.

2.3 Cadeias de Markov

As propriedades hoje conhecidas como processos de Markov foram investigadas pela primeira vez em 1907 por Andrey Markov. Os sistemas cuja evolução é descrita por esses processos tem a característica de que conhecendo o estado atual do sistema,

os estados passados não possuem influência sobre os estados futuros, ou seja, o processo possui ausência de memória, pois o estado futuro depende apenas da informação corrente (CUNHA; VELHO, 2003). Os conceitos apresentados neste tópico estão baseados em Maia (2008) .

2.3.1 Cadeias de Markov em Tempo Discreto

Definição 2.3.1. *Uma sequência de variáveis aleatórias X_1, X_2,... é uma cadeia de Markov em tempo discreto se para todo t ($t = 0, 1, 2, ...$) e para todo o espaço de estados do sistema tem-se*

$$P[X_t = x_t | X_0 = x_0, X_1 = x_1, ..., X_{t-1} = x_{t-1}] = P[X_t = x_t | X_{t-1} = x_{t-1}] \tag{2.14}$$

Podemos pensar em uma cadeia de Markov como um grafo com estados ligados entre si através de transições possíveis, a cada intervalo de tempo t um estado é visitado.

Utilizando o conceito de probabilidade condicional e propriedades markovianas, temos que a probabilidade $\pi_i^{(t)} = P(X_t = i)$ da cadeia estar no estado i no instante t é determinada pela matriz de transição e pela distribuição de probabilidade inicial. Dado um espaço de estados $S = 0, 1, 2, ..., N$ em que N é um número inteiro positivo definido pelo vetor $\pi^{(t)} = (\pi_0^{(t)}, ..., \pi_n^{(t)})$ infere-se que:

$$\begin{aligned} &\pi_j^{(t)} = P(X_t = j) = \sum_{i \in X} P(X_t = j, X_{t-1} = i) \\ &= \sum_{i \in X} P(X_t = j | X_{t-1} = i) P(X_{t-1} = i) \\ &= \sum_{i \in X} p_{ij}^{(t-1)} \pi_i^{(t-1)} \end{aligned} \tag{2.15}$$

$\pi^{(t)}$ pode ser interpretado como uma matriz linha definida pela equação matricial

$$\pi^{(t)} = \pi^{(t-1)} P^{(t-1)}, \tag{2.16}$$

cuja iteração leva à recursão

$$\pi^{(t)} = \pi^{(0)} P^{(0)} P^{(1)} ... P^{(t-1)} \tag{2.17}$$

Se as probabilidades de transição são *estacionárias*, ou seja, a probabilidade da cadeia passar de um estado a outro em um intervalo de tempo não depende de t e S, podemos determinar uma *matriz de transição* $P = (p_{ij})$ em que $p_{ij} = \mathbb{P}(X_t = j | X_{t-1} = i)$ com $\sum_{j \in S} p_{ij} = 1$ pra todo $i \in S$. Cadeias com essa caracteristicas são denominadas *homogêneas*. A Equação (2.17) pode ser reescrita na forma

$$\pi^{(t)} = \pi^{(0)} P^t \tag{2.18}$$

em que P^t é a t-ésima potência de P.

Definimos, assim, a *matriz de transição de n passos* como sendo $P^{(n)} = (p_{ij}^{(n)})$ tal que $p_{ij}^{(n)} = (X_{t+n} = j | X_t = i)$.

Teorema 2.3.1. *Seja X_n uma cadeia de Markov homogênea. Então a matriz de transição de n passos é igual a n-ésima potência da matriz de transição, isto é $P^{(n)} = P^n$.*

Demonstração. Temos que:

$$P(X_{n+1} = i_{n+1}, ..., X_{n+m} = i_{n+m} | X_0 = i_0, ..., X_n = i_n) = \frac{P(X_0 = i_0, ..., X_{n+m} = i_{n+m})}{P(X_0 = i_0, ..., X_n = i_n)} \tag{2.19}$$

O lado direito da igualdade pode ser reescrito como: (Para melhorar a visualização foi adotado $p_{i_n, i_{n+1}} = P(i_n, i_{n+1})$)

$$\frac{\pi_{i0}^0 P(i_0, i_1), ..., P(i_{n+m-1}, i_{n+m})}{\pi_{i0}^0 P(i_0, i_1), ..., P(i_{n-1}, i_n)} \tag{2.20}$$

Logo,

$$\begin{aligned} &P(X_{n+1} = i_{n+1}, ..., X_{n+m} = i_{n+m} | X_0 = i_0, ..., X_n = i_n) \\ &= P(i_n, i_{n+1}), ..., P(i_{n+m-1}, i_{n+m}) \end{aligned} \tag{2.21}$$

para um tratamento mais eficiente, é conveniente reescrever a Equação (2.21) da seguinte forma

$$\begin{aligned} &P(X_{n+1} = j_1, ..., X_{n+m} = j_m | X_0 = i_0, ..., X_n = i) \\ &= P(i, j_1), ..., P(j_{m-1}, j_m) \end{aligned} \tag{2.22}$$

Sejam $A_0, ..., A_{n1}$ subconjuntos de S. Segue a partir da Equação (2.22) que

$$\begin{aligned} &P(X_{n+1} = j_1, ..., X_{n+m} = j_m | X_0 \in A_0, ..., X_{n-1} \in A_{n-1}, X_n = i) \\ &= P(i, j_1), ..., P(j_{m-1}, j_m) \end{aligned} \tag{2.23}$$

Sejam $B_1, ..., B_m$ subconjuntos de S. Segue a partir da Equação (2.23)

$$\begin{aligned} &P(X_{n+1} \in B_1, ..., X_{n+m} \in B_m | X_0 \in A_0, ..., X_{n-1} \in A_{n-1}, X_n = i) \\ &= \sum_{y_1 \in B_1} \cdots \sum_{y_m \in B_m} P(i, j_1), ..., P(j_{m-1}, j_m) \end{aligned} \tag{2.24}$$

As probabilidades de transição $P^m(i,j)$ são definidas por

$$P^m(i,j) = \sum_{y_1 \in B_1} \cdots \sum_{y_m \in B_m} P(i,j_1), ..., P(j_{m-1}, j_m), para \quad m \geq 2$$

$$P^1(i,j) = P(i,j) \tag{2.25}$$

$$P^0(i,j) = \begin{cases} 1 & se \quad i = j \\ 0 & caso \quad contrário \end{cases}$$

A Equação (2.25) fornece a probabilidade do sistema ir do estado i para o estado j em m etapas. Dados $B_1 = ...B_{m-1} = S$ e $B_m = j$, tem-se em (2.24)

$$P(X_{n+m} = j | X_0 \in A_0, ..., X_{n-1} \in A_{n-1}, X_n = i) = P^m(i,j) \tag{2.26}$$

Dados $A_0 = ...A_{n_1} = S$, tem-se

$$P(X_{n+m} = j | X_n = i) = P^m(i,j) \tag{2.27}$$

$$P(X_{n+m} = j | X_0 = i, X_n = k) = P^m(k,j) \tag{2.28}$$

Utilizando a fórmula sequencial de Bayes (MIGON, 2006) tem-se,

$$\begin{gathered} P^{n+m} = P(X_{n+m} = j | X_0 = i) \\ = \sum_k P(X_n = k | X_0 = i) P(X_{n+m} = j | X_0 = i, X_n = k) \\ = \sum_k P^n(i,k) P(X_{n+m} = j | X_0 = i, X_n = k) \end{gathered} \tag{2.29}$$

A partir da Equação (2.28), obtém-se

$$P^{(n+m)} = \sum_k P^n(i,k) P^m(k,j) \tag{2.30}$$

Tomando $S = 0, 1, ..., N$, tem-se

$$p_{i,j}^{n+m} = \sum_{k=0}^{N} p_{i,k}^{(n)} P_{k,j}^{(m)} \tag{2.31}$$

A Equação (2.31) é denominada *Equação de Chapman-Kolmogorov.*

□

2.3.2 Equação de Chapman-Kolmogorov

A equação de Chapman-Kolmogorov fornece um método simples para obter a probabilidade de ordem superior de uma cadeia de Markov em termos das probabilidades de ordem inferior, ou seja, a fórmula descreve como calcular, por um método simples e explícito, todas as probabilidades de transições de ordem superior em termos da probabilidade de transição de uma etapa (DASGUPTA, 2011).

Teorema 2.3.1. *Seja X_n uma cadeia de Markov estacionária com espaço de estados S. Seja $n, m \geq 1$. Então,*

$$p_{ij}(m+n) = P(X_{m+n} = j | X_0 = i) = \sum_{k \in S} p_{ik}(m) p_{kj}(n). \tag{2.32}$$

Demonstração. Para chegar ao estado j a partir do estado i em $m+n$ etapas, a cadeia deve ir para algum estado $k \in S$ em m etapas e posteriormente ir do estado k para o estado j nas n etapas restantes. Adicionando todos os estados $k \in S$ possíveis, tem-se a equação de Chapman-Kolmogorov (DASGUPTA, 2011).

□

Corolário 2.3.1. *Seja $P^{(n)}$ a matriz de transição das probabilidades em n etapas S. Então, para todo $n \geq 2$, $P^{(n)} = P^n$, onde P^n.*

Demonstração. Utilizando a definição de produto de matrizes, para todo $m, n \geq 1$, $P^{(m+n)} = P^{(m)} P^{(n)} \Rightarrow P^{(2)} = PP = P^2$. Por indução, suponha que $P^{(n)} = P^n \quad \forall n \leq k$. Então, $P^{(k+1)} = P^{(k)} P = P^k P = P^{k+1}$.

□

Proposição 2.3.1. *Seja X_n uma cadeia de Markov estacionária com espaço de estados S e P uma matriz de transição, fixando $n \geq 1$. Então, $\lambda_n(i) = P(X_n = i) = \sum_{k \in S} p_{ki}(n) P(X_0 = k)$. Em notação matricial, se $\lambda = (\lambda_1, \lambda_2, ...)$ representa o vetor de probabilidades iniciais $P(X_0 = k), k = 1, 2, ...,$ e se λ_n representa o vetor de linha de probabilidades $P(X_n = i), i = 1, 2, ...,$ então $\lambda_n = \lambda P^n$.*

Demonstração. Ver demonstração em DasGupta (2011).

□

Esta é uma importante fórmula por que explica de maneira explícita como encontrar a distribuição de X_n a partir da distribuição inicial λ e da matriz de transição P, permitindo sua fácil implementação para o uso computacional.

3 Modelagem do algoritmo Pagerank

Neste capítulo, é abordado o desenvolvimento de uma versão do Pagerank, algoritmo de ranking de páginas do Google. Para tal, a modelagem e codificação do algoritmo foram baseados em Page (1998).

3.1 Contextualização

O primeiro passo para obter o vetor Pagerank é perturbar a matriz de adjacência A para obter uma matriz coluna estocástica S com $S_{i,j} > 0, \forall i, j$ e $\sum S_{:,j} = 1, \forall j$.

A perturbação Pagerank aplica-se à matriz de graus D, usada para obter as perturbações estocásticas, em que $D_{i,j}$ é igual ao grau do vértice v_i.

$$D_{ij} = \begin{cases} 0 & 1 \neq j \\ \sum A_{i,:} = grau(v_i) = grau(v_j) = \sum A_{:,j} & i = j \end{cases} \tag{3.1}$$

Se A é uma matriz de adjacência de um grafo conectado,as entradas diagonais de D são estritamente não nulas. Logo, a inversa D^{-1} é dada pela alternativa diagonal de entrada $D^{-1}_{i,i} = 1/D_{i,i}, \forall i$.

Example 3.1.1. *Considere matriz de adjacência representada na Tabela 2 com a inversa das entradas diagonais* $\boldsymbol{D}$ *representada na Tabela 3 . As entradas diagonais* $\boldsymbol{D}$ *correspondem aos graus dos vértices do grafo.*

Tabela 2 – Matriz de adjacência A

	a	b	c	d	e
a	0	1	0	0	1
b	1	0	1	0	1
c	0	1	0	1	0
d	0	0	1	0	1
e	1	1	0	1	0

Uma matriz linha estocástica, S, é uma matriz não negativa que a soma das linhas é igual a 1, i.e., $\sum S_{i,:} = 1$. Dada uma matriz de adjacências A e sua respectiva matriz de grau D, a matriz linha estocástica, S pode ser obtida por

$$S = D^{-1}A. \tag{3.2}$$

Tabela 3 – Matriz de Grau D^{-1}

	a	b	c	d	e
a	1/2	0	0	0	0
b	0	1/3	0	0	0
c	0	0	1/2	0	0
d	0	0	0	1/2	0
e	0	0	0	0	1/3

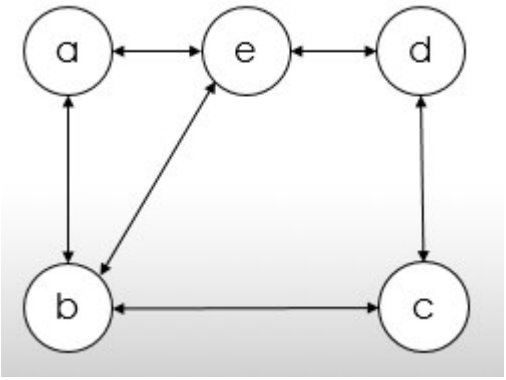


Figura 4 – Representação em grafo da matriz de adjacência A

Fonte: Fornecida pelo autor.

Uma matriz coluna estocástica S é uma matriz não negativa que a soma das colunas é igual a 1, isto é, $\sum S_{:,j} = 1$. Dada uma matriz de adjacências A e sua respectiva matriz de grau D, a matriz coluna estocástica, S pode ser obtida invertendo a ordem da Equação (3.2)

$$S = AD^{-1}. \tag{3.3}$$

Example 3.1.1. *A Tabela 4 representa um exemplo de uma matriz linha estocástica.*

Tabela 4 – Matriz estocástica linha

					$\sum$
0.0	0.5	0.0	0.0	0.5	1
0.3	0.0	0.3	0.0	0.3	1
0.0	0.5	0.0	0.5	0.0	1
0.0	0.0	0.5	0.0	0.5	1
0.3	0.3	0.0	0.3	0.0	1

Uma matriz Pagerank estocástica S é uma matriz que reflete a probabilidade de transição entre pares de estados em uma cadeia de Markov, um processo aleatório com

ausência de memória onde a probabilidade de ir para o próximo estado depende apenas do estado atual.

Dada uma matriz Pagerank S, a probabilidade de distribuição x_{i+1} é dada por

$$x_{i+1} = Sx_i \tag{3.4}$$

O vetor x representa a distribuição estacionária do processo de ausência de memória, ou seja, a probabilidade de se visitar cada vértice do grafo de maneira aleatória.

Baseado nesta premissa, pode-se afirmar que um vértice v_i é associado ao seu respectivo valor Pagerank x_i, com respeito a alguma matriz de transição P em que

$$P \cdot A \cdot P^T \leftrightarrow P \cdot S \cdot P^T \leftrightarrow P \cdot x. \tag{3.5}$$

Associando-se cada vértice do grafo à uma página, a importância da mesma será definida em termos do número de links que está página recebe em toda a rede. Como exemplo, dada a rede hipotética definida na Figura 5.

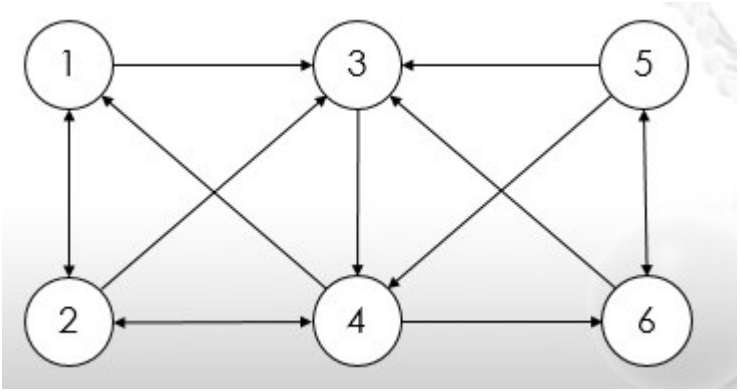


Figura 5 – Grafo de uma rede hipotética com 6 páginas

Fonte: Fornecida pelo autor.

Considerando $\pi(k)$ como o número de links que a página k recebe, temos que $\pi(1) = 2, \pi(2) = 2, \pi(3) = 4, \pi(4) = 2, \pi(5) = 1, \pi(6) = 2$, porém, é necessário um modo de avaliar o peso de uma página pelo número de links que apontam para ela, e com isso ajustar sua influência para as páginas para as quais ela referencia. esse ajuste é dado por:

$$\pi(k) = \sum_{j \in L_k} \frac{\pi(j)}{n_j} \tag{3.6}$$

tal que n_j é o número de links da página j e $L_k \in \mathbb{N}$.

Remodelando o problema de forma matemática, tem-se uma matriz $H = [H_{ij}]$ dada por:

$$H_{ij} = \begin{cases} 1/n_j & se \quad j \in L_i \\ 0 & se \quad j \notin L_i \end{cases} \tag{3.7}$$

Para o exemplo acima, a matriz H é definida como

$$H = \begin{pmatrix} 0 & 1/2 & 0 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/3 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 1/3 & 1/2 \\ 0 & 0 & 1 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 \end{pmatrix} \tag{3.8}$$

Por 3.6, temos que :

$$\pi = H\pi. \tag{3.9}$$

Calculando π obtemos o resultado: $\pi(1) = 0,1646$, $\pi(2) = 0,1646$, $\pi(3) = 0,2351$, $\pi(4) = 0,2458$, $\pi(5) = 0,0742$, $\pi(6) = 0,1157$.

A partir do resultado, conclui-se que a página $\pi(4)$ é que a possui o maior Pagerank, portanto, a que é mais relevante dentro da rede estudada.

A matriz H é uma matriz de transição de uma cadeia de Markov em tempo discreto, portanto, o vetor Pagerank é o vetor estacionário desta cadeia.

3.2 Codificação

O processo de implementação de uma versão do Pagerank consiste de duas partes: a construção de uma interface gráfica que permita inicializar a rede estudada e o cálculo e análise do vetor estacionário Pagerank.

Nesta monografia será abordada apenas a implementação dos algoritmos mais relevantes.

Perturbação da matriz de adjacências

Para construir a matriz de perturbação, o sistema manualmente identifica o número de páginas da rede e o links entre elas.

A modelagem da rede é feita através de uma matriz de ordem igual ao número total de páginas, em que a diagonal principal representa o relacionamento entre as páginas que possuem links.

O Código 1 implementa a função responsável pela pertubação da matriz de adjacência.

```
1 /**
2 *  Autor: Thiago Lemos
3 */
4 public float[][] Perturbacao_Matriz
5 {
6   int _total_paginas= func_numero_de_paginas_rede();
7     float _matrix_paginas[_total_paginas][_total_paginas];
8     int vetor_relacionamento[_total_paginas];
9
10    // verifica quantos links cada página na rede possui
11
12    for(int i=0;i<_total_paginas)
13    {
14      vetor_relacionamento[i]=func_total_links(i);
15    }
16
17    // Constroi a matriz D
18    for(int i=0;i<_total_paginas)
19    {
20      _matrix_paginas[i][i]=1/(func_total_links(i));
21    }
22
23    return _matrix_paginas;
24
25
26 }
```

Código 1 – Implementação da perturbação da matriz de adjacência

Equação de Chapman-Kolmogorov

Para que o software consiga achar o grau de relevância da página atual é necessário um algoritmo que permita analisar a importância das outras páginas associadas, porém sem calcular todos os dados da matriz.

A equação de Chapman-Kolmogorov, representada em (2.31) possibilita que este requisito seja efetivamente obedecido.

O Código 2 implementa a classe "*ChapmanKolmogorov*" que modela a função recursiva capaz de calcular a importância da página web sem a necessidade de calcular todos os elementos da matriz de transição (que em alguns casos pode ser da ordem dos milhões).

```
1 /**
2 *  Autor: Thiago Lemos
3 */
4 public class ChapmanKolmogorov
5 {
6     int u;
7     double fator=0;
8     public double probabilidade(double v[][],int t,int o,int i,
    int j){
9         if(t==1) return v[i][j];
10        else
11                {
12          for(u=0;u<o;u++)
13                    {
14            fator+=v[i][u]*probabilidade(v,t-1,o,u,j);
15          }
16          return fator;
17        }
18      }
19 }
```

Código 2 – Implementação da equação de Chapman-Kolmogorov

3.3 Análise do algoritmo

O sistema foi idealizado em contexto no qual uma rede de páginas, juntamente com seus relacionamentos (links) são inseridos e posteriormente analisados. O software devolve para o usuário o vetor Pagerank contendo a relevância de cada página em ordem decrescente, a matriz de transição que modela a rede e as páginas organizadas por seus respectivos Pageranks.

O sistema gera um grafo orientado representando a rede estudada, a análise sobre um grafo permite observar o fluxo de dados e assim, tomar medidas paliativas para redirecionar o fluxo em áreas de grande congestionamento.

A Figura 6 representa o protótipo desenvolvido para um exemplo específico de rede com 30 páginas, o relacionamento entre as páginas pode ser observado no grafo representado na Figura 7.

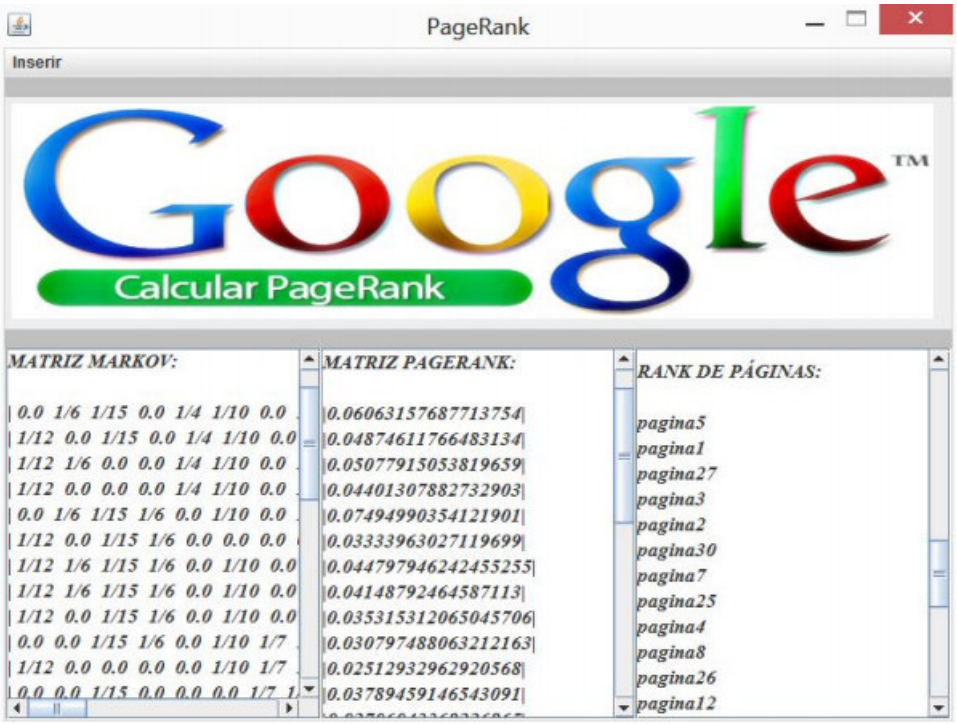


Figura 6 – *Pagerank Software*: funcionamento da versão implementada para uma rede hipotética de 30 páginas inseridas em ordem numérica

Fonte: Fornecida pelo autor.

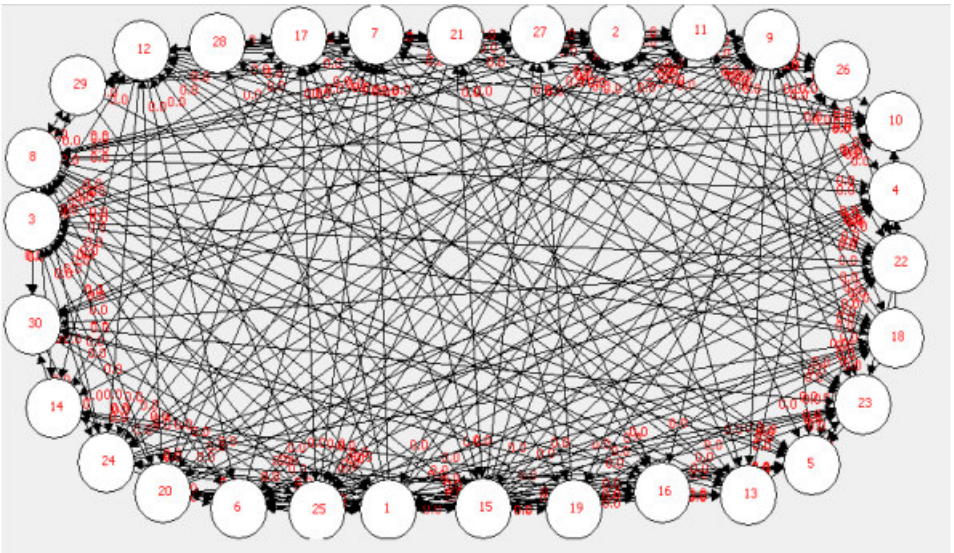


Figura 7 – *Pagerank Software*: grafo da rede representada na Figura 6

Fonte: Fornecida pelo autor.

4 Modelagem do Pagerank genético

Neste capítulo, será abordado o desenvolvimento de um software que mapeia as frequências alélicas sob o efeito da deriva genética denominado *Biopagerank*, para tanto, foram utilizando os conceitos de Durrett (2008), Page (1998), e Maia (2008) e a teoria do *Pagerank* abordada no capítulo anterior.

4.1 Contextualização

Para evidenciar os efeitos da deriva, a modelagem do Biopagerank considera populações com as seguintes características:

- População finita;
- Organismos diploides;
- Número infinito de gametas;
- Gerações sem sobreposição;
- Ambiente de seleção neutra;

Utilizando o modelo de Wright-Fisher para dois alelos hipotéticos, A e a, é possível representar a probabilidade condicional do número de genes A no tempo $t+1$, dado o número de genes A no tempo t e o tamanho das populações. Se $n(t)$ for o número de genes A na população nesse tempo t, tem-se a seguinte distribuição:

$$P(n(t+1)=j|n(t)=i,N)=\begin{pmatrix}2N\\ j\end{pmatrix}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{2N-j} \tag{4.1}$$

Analisando a Equação (4.1) percebe-se que a ocorrência de j sucessos em N eventos, na qual só pode resultar em sucesso ou fracasso, quando a probabilidade de sucesso é constante, é conhecido como Distribuição Binomial, logo o Modelo de Wright-Fisher pode ser modelado como uma cadeia de Markov em tempo discreto cuja matriz de transição é determinada pela Equação (4.1).

A matriz de transição correspondente para uma população de dois indivíduos é dada pela matriz de ordem $2N+1$:

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ P_{10} & P_{11} & P_{12} & P_{13} & ... & P_{1n} \\ P_{20} & P_{21} & P_{22} & P_{23} & ... & P_{2n} \\ P_{30} & P_{31} & P_{32} & P_{33} & ... & P_{3n} \\ : & : & : & : & ... & : \\ P_{n0} & P_{n1} & P_{n2} & P_{n3} & ... & P_{nn} \\ 0 & 0 & 0 & 0 & ... & 1 \end{pmatrix} \tag{4.2}$$

A Equação (4.1) pode ainda ser interpretada como a probabilidade de termos j cópias do alelo A no tempo $t+1$, dado que em t temos i cópias do alelo A.

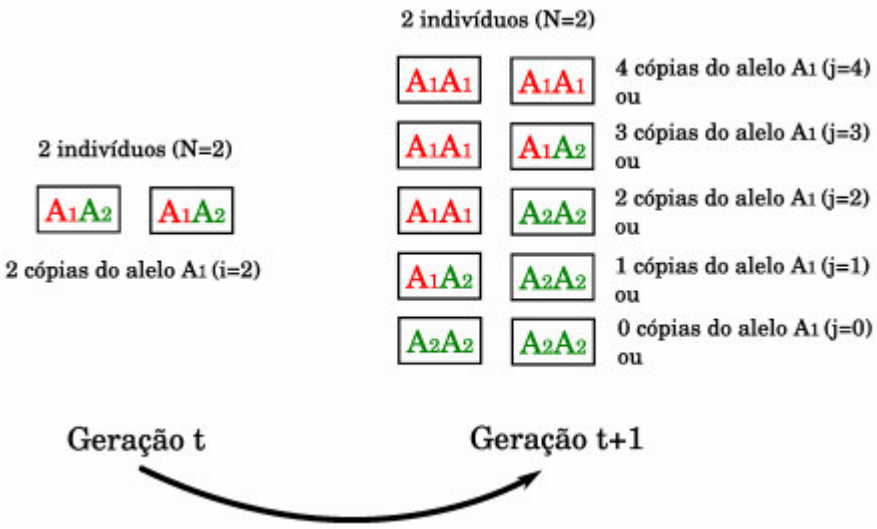


Figura 8 – Configuração genética para 2 indivíduos

Fonte: Fornecida pelo autor.

A modelagem baseada em Cadeias de Markov permite que uma quantidade de dados elevada seja tratada computacionalmente de maneira mais rápida aplicando Chapman-Kolmogorov (2.32) para calcular o numero de populações em cada frequência alélica ao longo do tempo.

Seja P a matriz de transição, o elemento $P_{ij}^{(n)}$ da matriz P que representa a probabilidade de que o processo, iniciado no estado s_i, esteja no estado s_j depois de n passos é dado por

$$p_{ij}^{(n)} = \sum_{r=1}^{k} p_{ir}^{(u)} p_{rj}^{(n-u)}, 0 < u < n. \tag{4.3}$$

A característica aleatória da deriva permite que o software utilize um processo estocástico para mapear as frequências dos alelos do seguinte modo: a matriz de transição

P terá múltiplas iterações com um vetor inicial para proporcionar o estado alélico de cada população em determinados intervalos de tempo.

A equação de Chapman-Kolmogorov facilita esse processo, uma vez que não é necessário calcular todas iterações para achar um valor específico $P_{ij}^{(n)}$.

Os valores calculados são inseridos em uma matriz linha, chamada de matriz BiopageRank, que contêm a frequência alélica das populações estudadas.

4.2 Codificação

A implementação do Biopagerank possui uma interface que permite inicializar o software com :

- Número de gerações;
- Frequência dos alelos observada na geração inicial;
- Tamanho da população;
- Número de populações no estudo;

Algoritmo de Wright-Fisher

Utilizando a método implementado de Wright-Fisher, pode-se calcular a frequência exata do alelo estudado.

A Tabela 5 representa os resultados do algoritmo implementado para uma série de casos de testes com tamanhos de populações variáveis.

Tabela 5 – Tabela de teste para Wright-Fisher

Tamanho da população	2	5	10	50	100	500	1000	10.000
Cópias do gene(2N)	4	10	20	100	200	1000	2000	20.000
Wright-Fisher	0.375	0.246	0.176	0.080	0.056	0.025	0.018	0.006

Os dados da Tabela 5 ratificam que quanto maior o número de indivíduos na populaçao, menor o efeito da deriva genética, efetivando sua atuação mais influente em populações pequenas.

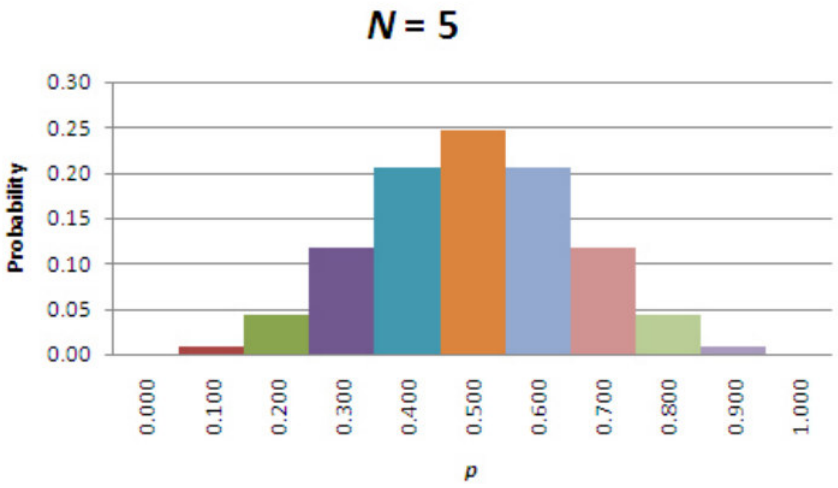


Figura 9 – Probabilidade de frequência alélica na próxima geração em populações de cinco indivíduos (N=5)

Fonte: Kliman et al. (2008)

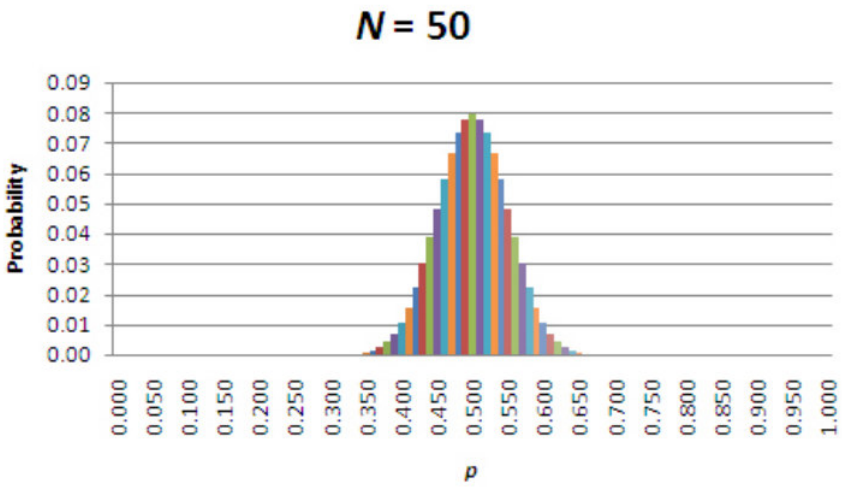


Figura 10 – Probabilidade de frequência alélica na próxima geração em populações de cinquenta indivíduos (N=50)

Fonte: Kliman et al. (2008)

Com o aumento significativo do número de indivíduos na população os efeitos da deriva tornam-se menos visíveis, isto é, devido a uma diminuição do erro de amostragem. Apesar da direção da mudança da frequência ser aleatória, é possível prever o caminho mais provável baseando-se da frequência da população anterior.

```
/**
*  Autor: Thiago Lemos
*/
public double Wright_Fisher(double i,double j,double n1)
{
            double fator_Num,fator_Den,fator_Den2;
            double fatorial,gene_first, gene_second=0;
            double u;n=0;
            n=n1;
            fator_Num=2*n;
            fator_Den=j;
            fator_Den2=fator_Num-fator_Den;
            for (u=fator_Num-1;u>0;u--) fator_Num=fator_Num*u;
            if((fator_Den==0.0)||(fator_Den==1.0)) fator_Den=1;
            else{
              for (u=fator_Den-1;u>0;u--) fator_Den=fator_Den*u;
            }
            if((fator_Den2==0)||(fator_Den2==1)) fator_Den2=1;
            else{
              for (u=fator_Den2-1;u>0;u--)fator_Den2=fator_Den2*u;
            }

            fator_Den=fator_Den*fator_Den2;
            fatorial=fator_Num/fator_Den;
            gene_first=Math.pow(i/(2*n),j);
            gene_second=Math.pow((1-(i/(2*n))),2*n-j);
            return (fat*gene_first*gene_second);
        }
}
```

Código 3 – Implementação do modelo de Wrigh-Fisher

Algoritmo da matriz de transição

O algoritmo de Matriz de Transição calcula a probabilidade de termos j cópias do alelo A no tempo $t+1$, dado que em t temos i cópias do alelo A para cada estado possível no Pool gênico utilizando o algoritmo de Wright-Fisher. Apenas os estados absorventes não são calculados.

A equação de Chapman-Kolmogorov otimiza o processo de calculo das probabilidades, pois permite que uma dada probabilidade seja calculada sem a necessidade de se contruir toda a Matriz de Transição.

```
1 /**
2 *  Autor: Thiago Lemos
3 */
4 public double[][] CalculaTransicao(int n){
5
6       int n1=(2*n)+1;
7       int i,j;
8       double vetorTransic[][]=new double[n1][n1];
9       for(i=0;i<n1;i++){
10          for (j=0;j<n1;j++){
11            if( (i==0)&&(j==0)) vetorTransic[i][j]=1;
12            else if((i==0)&&(j!=0)) vetorTransic[i][j]=0;
13            else if ((i==n1-1)&&(j==n1-1)) vetorTransic[i][j]=1;
14            else if ((i==n1-1)&&(j!=n1-1)) vetorTransic[i][j]=0;
15            else vetorTransic[i][j]=al.WrightFisher(i, j, n);
16          }
17      }
18        return vetorTransic;
19  }
```

Código 4 – Implementação da matriz de transição

4.3 Resultados

O software BiopageRank mapeia as frequências alélicas sob o efeito da deriva genética utilizando o modelo proposto. A Tabela 6 resume um conjunto de populações com características hipotéticas, bem como o resultado do processamento realizado pelo Biopagerank.

Tabela 6 – Mapeamento do alelo estudado ao longo de 100 gerações

Características	Quantidade
Gerações	100
Frequência inicial do alelo	50
Tamanho da população	100
Número de populações	50
Populações extintas	**12**
Populações sobreviventes	**38**
Populações com alelo fixado	**13**

A Figura 11 representa o fluxo alélico das populações estudadas ao longo das gerações.

Devido ao número reduzido de indivíduos, a influência da deriva torna-se evidente na maneira como as populações se comportam.

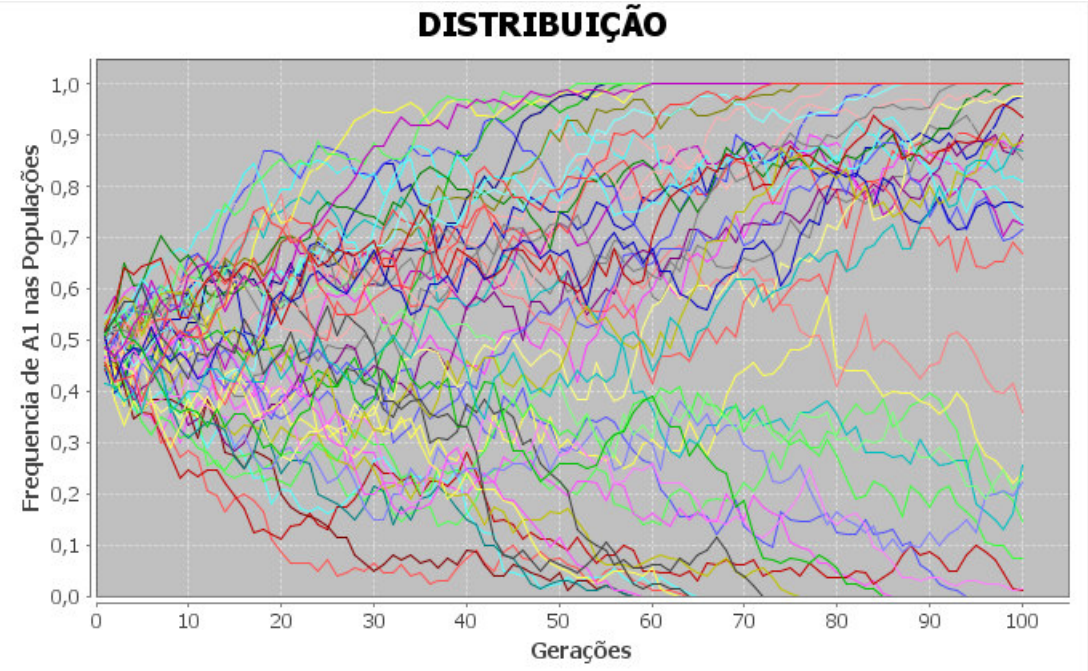


Figura 11 – Efeito da deriva genética durante 100 gerações (teste 1)

Fonte: Fornecida pelo autor.

A simulação do Biopagerank para populações menores demonstra que o erro de amostragem aumenta à medida que o número de indivíduos na população diminui, este processo pode modificar a frequência de um alelo inapto, permitindo que o mesmo sobreviva mesmo em condições pouco favoráveis.

A Tabela 7 resume um conjunto de pequenas populações, com o intuito de evidenciar os efeitos da deriva sobre o mesmo gene em populações diferentes.

Tabela 7 – Mapeamento do alelo estudado ao longo de 100 gerações

Características	Quantidade
Gerações	100
Frequência inicial do alelo	30
Tamanho da população	50
Número de populações	2
Populações extintas	**1**
Populações sobreviventes	**1**
Populações com alelo fixado	**1**

A Figura 12 simula uma população hipotética na qual o alelo estudado tem frequência inicial de 30% em duas populações distintas.

A distribuição da Figura 12 analisa o mesmo alelo em duas populações diferentes sob condições de seleção neutra com apenas a deriva genética atuando. Percebe-se que mesmo o alelo estando em desvantagem com relação ao alelo correspondente do mesmo lócus, (A1: 30% A2: 70%), o mesmo consegue se fixar na população vermelha (A1: 100% A2: 0%). Este tipo de acontecimento ratifica o efeito da deriva genética em populações pequenas.

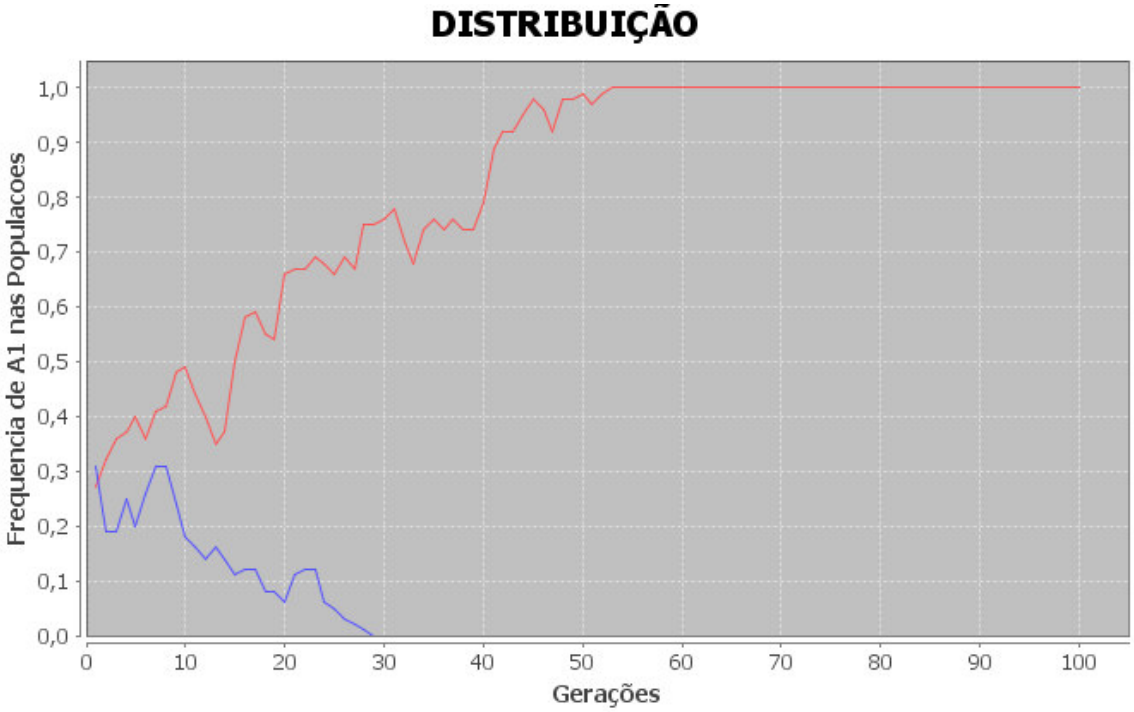


Figura 12 – Efeito da deriva genética durante 100 gerações (teste 2)

Fonte: Fornecida pelo autor.

A Tabela 8 resume o teste realizado por Buri (1956) com moscas do tipo Drosophila.

Tabela 8 – Mapeamento Realizado por Buri

Características	Quantidade
Gerações	19
Frequência inicial do alelo	50
Tamanho da população	16
Número de populações	107
Populações extintas	**30**
Populações sobreviventes	**77**
Populações com alelo fixado	**28**

Para validar o modelo proposto, a Tabela 9 apresenta os dados do experimento utilizando o Biopagerank.

Segundo os resultados de Buri (1956), quando a deriva genética está atuando a relação entre extinções e fixações devem ficar próximas de 1.0. A taxa observada 30:28 satisfaz a condição esperada para a deriva genética.

Tabela 9 – Mapeamento do teste Buri pelo Biopagerank

Características	Quantidade
Gerações	19
Frequência inicial do alelo	50
Tamanho da população	16
Número de populações	107
Populações extintas	**35**
Populações sobreviventes	**72**
Populações com alelo fixado	**33**

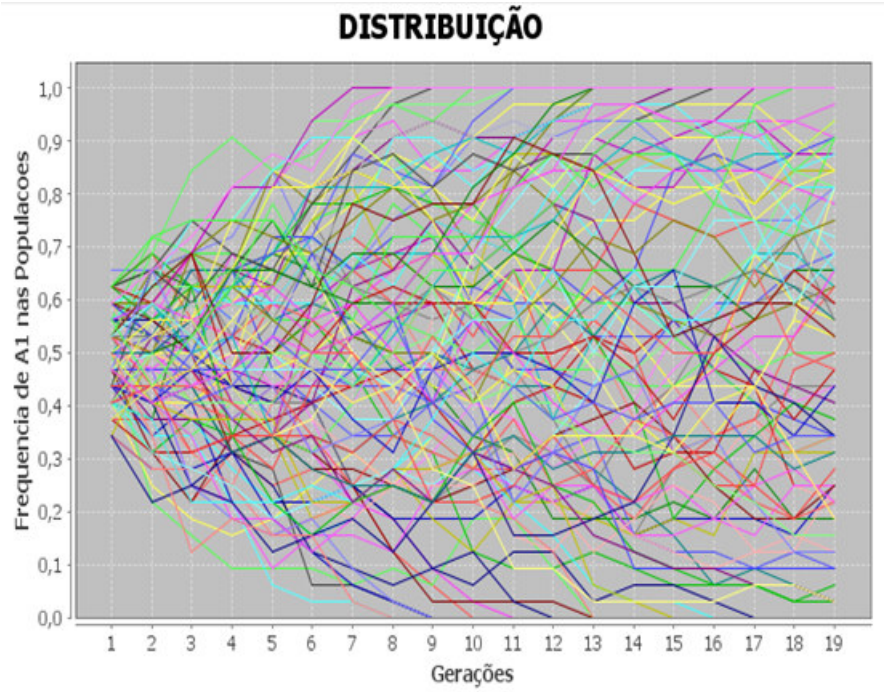


Figura 13 – Efeito da deriva genética durante 19 gerações utilizando o Biopagerank

Fonte: Fornecida pelo autor.

Na Figura 13 percebemos que quase metade das populações está sofrendo evolução, mas a seleção natural não é a responsável por esse processo. A relação 35:33 encontrada pela computação do Biopagerank também satisfaz as condições necessárias para a deriva genética.

A variabilidade Genética sob o efeito da deriva cai haja vista que o número de heterozigotos nas populações também diminui. A Figura 14 ilustra o mapeamento da frequência de heterozigotos de acordo com o experimento de Buri e com o software baseado no modelo proposto.

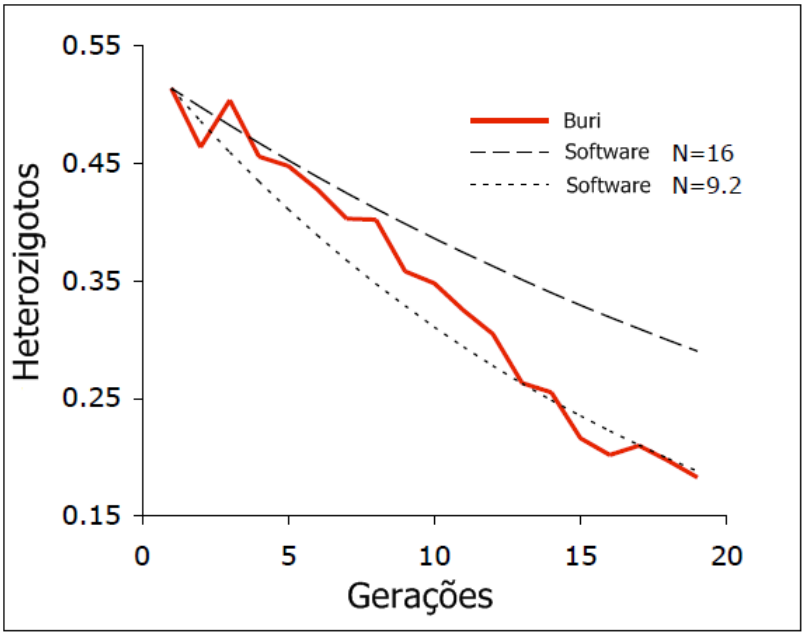


Figura 14 – Mapeamento de Heterozigotos

Fonte: Fornecida pelo autor.

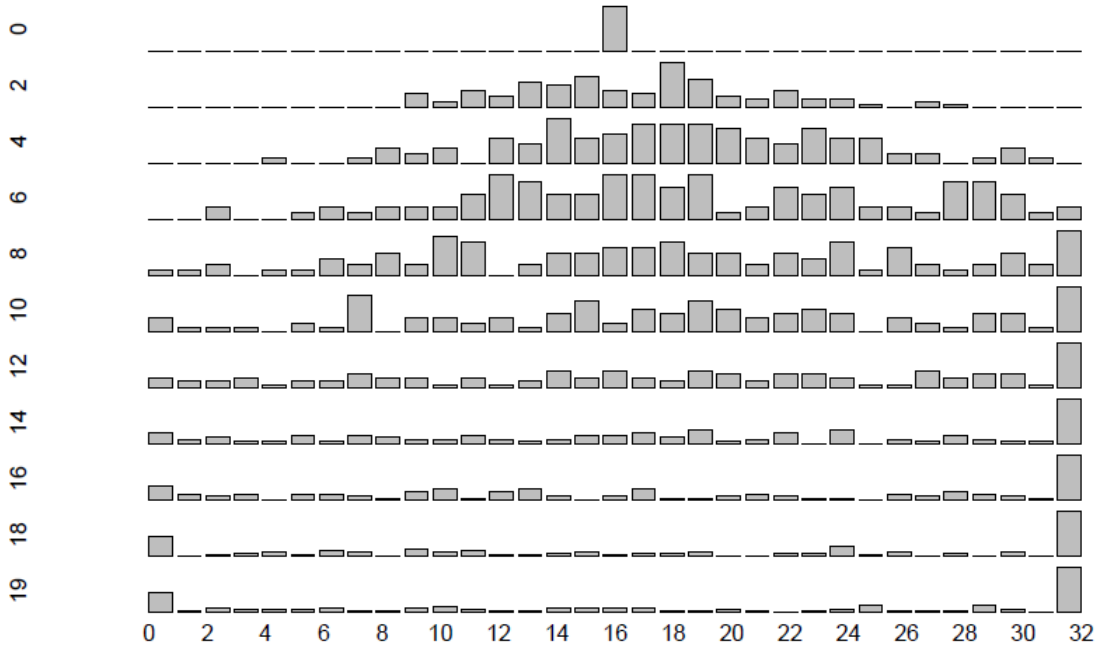


Figura 15 – Distribuição de Heterozigotos

Fonte: Buri (1956)

Percebe-se que a frequência de heterozigotos é aproximadamente igual ao resultado encontrado por Buri. A distribuição em cada população é dada pela Figura 15

Com relação a distribuição binomial, uma probabilidade **k** pode calculada por

$$P(k) = \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k} \tag{4.4}$$

A variância da variável aleatória do tipo binomial é calculada usando-se

$$Var(K) = Np(1-p) \tag{4.5}$$

Substituindo $k = p$ em (4.5), tem-se

$$Var(p) = Var(K/N) = \frac{1}{N^2} Var(K) = \frac{p(1-p)}{N} \tag{4.6}$$

Aplicando os dados do experimento de Buri utilizando o Biopagerank, obtem-se:

$$Var(p_{t+1}) = \frac{p_t(1-p_t)}{2N} \tag{4.7}$$

A variância mensura o grau de incerteza sobre as frequências dos alelos nas próximas gerações dados as atuais frequências. Quanto maior o grau de incerteza, maior será o efeito da deriva. Esta afirmativa pode ser provada analisando o desvio padrão da deriva genética, pois quanto menor o valor de N (número de indivíduos), maior será o desvio padrão.

O Biopagerank modela o seguinte sistema:

$$p_{k+1} = p_k + / - 2\sqrt{\frac{p_k q_k}{2N}}, q_{k+1} = q_k + / - 2\sqrt{\frac{p_k q_k}{2N}}, \quad \forall p, q \in \mathbb{R} \tag{4.8}$$

A aplicabilidade do modelo computacional projetado torna mais clara a forma como os alelos se distribuem durante as gerações permitindo a implementação de outros métodos em sua estrutura.

5 Conclusão

Este trabalho apresentou um modelo preditivo para análise de frequências alélicas sob o efeito da deriva via Cadeias de Markov e o modelo de Wright-Fisher, tornando-se uma alternativa para futuros pesquisadores que queiram resultados rápidos e eficientes para suas pesquisas.

Inicialmente foi modelado uma versão simplicada do algoritmo de Rankeamento de páginas do Google, o Pagerank, utilizando a equação de Chapman-Kolmogorov como medida de otimizaçã. Baseando-se no softare desenvolvido, utilizou-se o modelo binomial de Wright-Fisher para inicializar a matriz de transição da cadeia de Markov do Pagerank e modificou-se o conjunto de entradas para analisar configurações gênicas.

Essas modificações permitiram o software mapear as frequências dos alelos inseridos no seu conjunto de entradas sob o efeito da deriva. O software pode ser facilmente estendido para englobar outros mecanismos evolutivos, bem como modificar as condições do próprio Biopagerank, como exemplo pode-se alterar as populações de finitas para infinitas, considerar gerações com sobreposição, etc.

5.1 Trabalhos Futuros

Como perspectivas futuras é possível destacar:

- Incluir no software Biopagerank a modelagem de outros fenômenos biológicos, como mutação, migração, etc;
- Substituir as Cadeias de Markov por um formalismo inteligente baseado em algoritmos Genéticos ou Redes Neurais;
- Considerar populações infinitas, sobreposição, e comportamento dinâmico dos indivíduos ao longo do tempo em ambientes não-neutros;
- Analisar o método apresentado na área de Sistemas Dinâmicos não Lineares ;
- Estruturar o ambiente gráfico e usabilidade do software Biopagerank;

Referências

ANDERSEN, R. et al. Local graph partitioning using pagerank vectors. *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006. Citado na página 22.

ASCENSãO, C. P. *Como funciona um motor de busca?* 2014. Website. Disponível em: <http://www.portalwebmarketing.com/MotoresdeBusca/Comofuncionaummotordebusca/tabid/435/Default.aspx>. Acesso em: 04/02/2015. Citado na página 21.

BORGES-OSORIO, W. M. R. M. R. *Genética Humana*. 3. ed. São Paulo: Artmed Editora, 2013. ISBN 9788565852906. Citado 2 vezes nas páginas 15 e 16.

BURI, P. Gene frequency in small populations of mutant drosophila. *Evolution*, 1956. Citado 5 vezes nas páginas 13, 42, 43, 44 e 50.

CUNHA, A. M. da; VELHO, L. Metodos probabilisticos para reconhecimento de voz. *IMPA - VISGRAF Laboratory*, 2003. Citado na página 24.

DASGUPTA, A. *Probability for Statistics and Machine Learning: Fundamentals and Advanced*. New York: Springer, 2011. ISBN 978-1-4419-9633-6. Citado na página 27.

DURRETT, R. *Probability Models for DNA Sequence Evolution*. 2. ed. New York: Springuer, 2008. ISBN 978-0-387-78169-3. Citado 2 vezes nas páginas 17 e 35.

EWENS warren J. *Mathematical Population Genetics*. 2. ed. New York: Springuer, 2004. ISBN 0-38720191-2. Citado na página 17.

FUTUYAMA, D. J. *Biologia Evolutiva*. 2. ed. São Paulo: FUNPEC, 2002. ISBN 0-87893-188-0. Citado na página 16.

GLEICH, D. F. *Models and algorithms for Pagerank sensitivity*. 198 p. thesis, 2009. Citado 2 vezes nas páginas 22 e 23.

GOLUB, G. H.; LOAN, C. F. V. *Matrix Computations*. [S.l.]: The Johns Hopkins University Press, 1996. ISBN 9780801854149. Citado na página 23.

GOVAN, A. Y. et al. Generalizing google's pagerank to rank national football league teams. *SAS Global Forum*, 2008. Citado na página 22.

KLIMAN, B. S. R. et al. *Genetic Drift and Effective Population Size*. 2008. Website. Disponível em: <http://www.nature.com/scitable/topicpage/genetic-drift-and-effective-population-size-772523>. Acesso em: 12/05/2015. Citado na página 38.

LANGVILLE, C. D. M. A. N. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. 2. ed. New Jersey: Princeton, 2012. ISBN 978-0691152660. Citado na página 20.

MAIA, L. P. *Uma introdução a dinâmina estocástica de populações*. São Paulo: SBMAC, 2008. ISBN 978-85-86883-41-5. Citado 2 vezes nas páginas 24 e 35.

MANNING, C. D. et al. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. ISBN 0521865719. Citado na página 20.

MIGON, H. S. *Introdução a inferência Bayesiana*. Rio de Janeiro: UFRJ, 2006. Citado na página 26.

MORRISON, R. B. J. L. et al. Generank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, Septemberl 2005. Citado na página 22.

PAGE, L. The pagerank citation ranking: Bringing order to the web. *Stanford Paper*, 1998. Citado 3 vezes nas páginas 13, 28 e 35.

SADAVA, D. et al. *Vida: A Ciência da Biologia*. 8. ed. São Paulo: Artmed Editora, 2009. ISBN 9788536320595. Citado na página 16.

WALSH, T. L. D. M. Cotação. In: _____. *The Trials of Life: Natural Selection and Random Drift**. Denis M. Walsh, 2002. Disponível em: <http://faculty.arts.ubc.ca/jbeatty/WalshEtAl2002.pdf>. Acesso em: 28 jan. 2015. Citado na página 17.

WITTEN, I. H. et al. *Managing Gigabytes: Compressing and Indexing Documents and Images*. 1. ed. Massachusetts: Morgan Kaufmann, 1999. ISBN 978-1558605701. Citado na página 20.

Xavier Didelot. *Statistical population genetics*. 2014. Website. Disponível em: <http://www.stats.ox.ac.uk/~didelot/popgen/lecture2.pdf>. Acesso em: 28 jan. 2014. Citado 2 vezes nas páginas 18 e 19.

APÊNDICE A – Experimento de Peter Buri

A.1 Experimento preliminar

Nesta experiência utilizando moscas de fruta do tipo Drosophila, Buri tentou determinar se os genótipos no locus *bw* para cor dos olhos diferiam na aptidão. Ele realizou uma série de cruzamentos para determinar se as freqüências de genótipos na prole diferia das expectativas mendelianas, o que seria uma indicação de seleção neste lócus no ambiente de laboratório.

Buri realizou três séries de cruzamentos.

Série I envolveu fêmeas de um único genótipo e machos de um único genótipo. Testadas as diferenças entre genótipos de viabilidade.

Série II envolveu machos de um único genótipo e fêmeas de dois genótipos e testado para diferenças entre genótipos em uma combinação de sobrevivência e fecundidade feminina.

Série III envolveu fêmeas de um único genótipo e machos de dois genótipos e testado para diferenças entre genótipos em uma combinação de sobrevivência e fertilidade masculina.

Em seu trabalho, apenas três ensaios (denominados II-Ca, III-Ca, e III-Cb) fizeram as freqüências observadas dos genótipos na prole adulta diferirem significativamente. Quando estes três ensaios foram repetidos, os resultados não mostraram diferenças de expectativas sob nenhuma seleção. Por conseguinte, esta série de experiências indica a ausência de selecção natural detectável neste locus sob condições de laboratório.

A.2 Experimento principal

Na experiência principal, Buri iniciou 107 populações com a frequência de cada alelo inicial de 0,5. Cada geração possuia oito machos e oito fêmeas escolhidos aleatoriamente a partir da prole adulta emergente da geração anterior e transferidos para uma nova gaiola para fundar a próxima geração. Os genótipos destes 16 indivíduos também foram marcados para determinar a freqüência do gene. O experimento foi executado por 20 gerações.

A Figura 16 mostra o histograma de frequência genética do alelo *bw75* para as 107 populações para cada uma das gerações. Note que a variância do histograma aumenta gradualmente e que o número de populações fixo para um alelo ou outro também aumenta.

O número de populações fixas para o alelo *bw75* é aproximadamente a mesma que

para o alelo *bw*, como esperado a partir das frequências iniciais dos dois alelos.

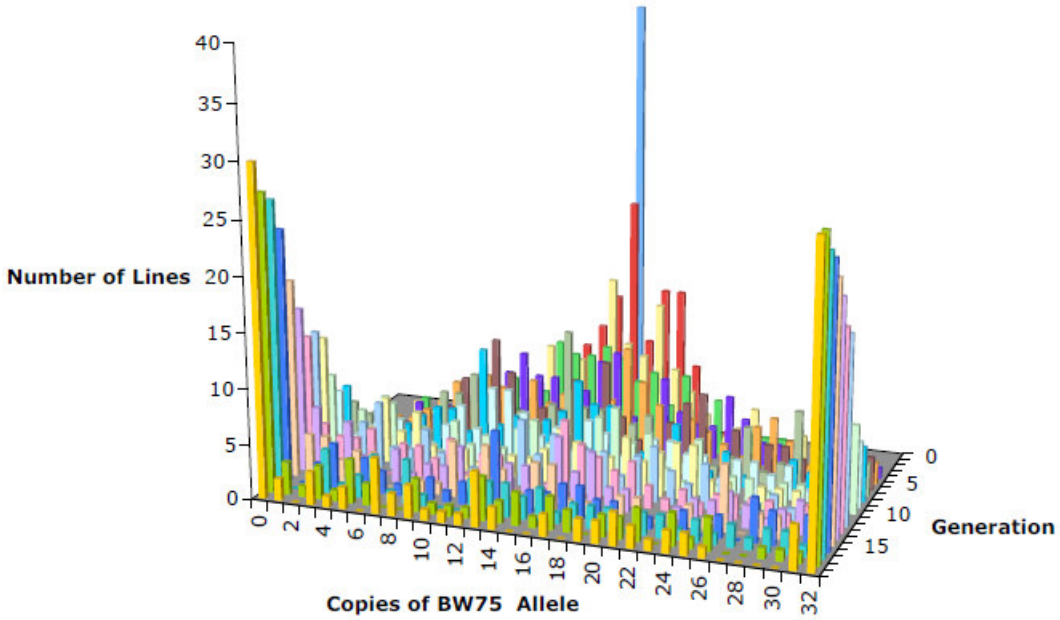


Figura 16 – Histograma para o genótipo *bw75*

Fonte: Buri (1956)