Um Algoritmo Adaptativivo de Unificação de Redes Haplotípicas

R. H. G. Guiraldelli, R.L A. Rocha

***Resumo*— Para representação da realidade evolutiva das espécies, a rede haplotípica é uma das modelagens gráficas de referência em filogenética e filogeografia, condensando grande quantidade de informação mutacional simplesmente em vértices e arestas. Essa, no entanto, não costuma ser única para um mesmo conjunto de dados geradores de entrada e o resultado apresenta-se dependente do modelo de construção escolhido. Este artigo propõe uma representação unificada das diversas redes produzidas — fornecendo novos dados antes inexistentes — utilizando-se de tecnologia adaptativa para sua construção.**

Palavras-chave— **Rede haplotípica, grafos, adaptatividade, NCPA.**

I. Introdução

O estudo da genética de populações, nomeado filogenética, assim como a filogeografia, que relaciona o estudo filogeográfico com as influências ambientais, processa grande quantidade de dados para a extração de informações relevantes sobre a evolução de determinada espécie. Para esse objetivo, no entanto, faz-se necessário o uso de ferramentas de suporte eficiente para análise desses dados; a experiência na pesquisa biológica, portanto, guiou a solução para a representação gráfica através de grafos.

Essa representação, chamada na biologia de rede haplotípica, é de amplo uso e parte fundamental do principal algoritmo utilizado na filogeografia, o *nested clade phylogeographic analysis*; este algoritmo, porém, admite apenas uma única rede haplotípica de entrada para análise quando, na verdade, várias hipóteses de redes existem para o mesmo conjunto de dados.

Para uma melhor representação da realidade evolutiva das espécies, independentemente do modelo proposto para formação da rede haplotípica, propõe-se o uso de adaptatividade para unificação das diversas instâncias de redes existentes para o mesmo conjunto de dados em uma única rede, adicionando, ainda, informações de probabilidade para melhores escolhas de caminhos evolutivos.

Assim, este artigo se organiza da seguinte forma: primeiramente, este seção introdutória; a seção 2 descreve a rede haplotípica e as técnicas de sintetização existentes; a próxima seção explica, brevemente, o conceito de adaptatividade; então, a seção explicativa sobre a técnica proposta para unificação das redes; por fim, na seção 5, as conclusões, propondo, ainda, trabalhos futuros.

II. Redes Haplotípicas

No estudo da genética de populações, o sequenciamento do genoma de diversos indivíduos continua uma ferramenta básica para a realização de análises, mas não mais a única ferramenta. Para tanto, existe a necessidade de uma ferramenta de análise que relacione os indivíduos amostrados através de características comuns e que sintetize, em uma simples representação, grande densidade de informação.

Buscando satisfazer esses requisitos, através de notação por grafos, a rede haplotípica relaciona os haplótipos da população (de mesma espécie) amostrada através das diferenças mutacionais representadas pelas arestas, sendo os próprios haplótipos os vértices. É importante notar que um rede haplotípica faz sentido, apenas, com dados coletados de populações da mesma espécie, avaliando diferenças sutis no material genético sequenciado entre os indivíduos da espécie em estudo.

Contudo, há mais de um método para a construção das redes haplotípicas, com a particularidade de que cada um desses produz um modelo de grafo diferente do outro. Tais modelos, descritos brevemente em [1], são: (a) minimum spanning tree; (b) statistical parsimony e; (c) full median. A literatura destaca os métodos (a) e (b) com maior enfoque, especialmente o último por ser a metodologia de escolha do aplicativo TCS [2].

A **Erro! Fonte de referência não encontrada.** abaixo epresenta uma rede haplotípica extraída de [3] sintetizada utilizando-se o aplicativo supracitado.

Ricardo Henrique Gracini Guiraldelli é aluno de mestrado, tendo como o orientador o Prof. Dr. Ricardo Luis de Azevedo da Rocha, do Departamento de Sistemas Digitais da Escola Politécnica da Universidade de São Paulo. E-mails: rguira@usp.br; rlarocha@usp.br.

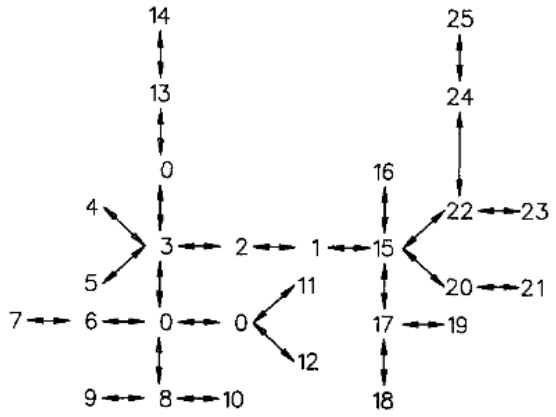

Fig. 1. Rede haplotípica de uma população de *Drosophila melanogaster* [3].

III. Adaptatividade

Os dispositivos computacionais tradicionais, como os autômatos ou a máquina de Turing, são formalismos de grande utilidade para a representação e reprodução de algoritmos. Em particular, os autômatos são dispositivos muito simples e de fácil entendimento até mesmo ao leigo, porém têm poder de expressividade limitada — limitam-se à representatividade de linguagens livres-de-contexto [4]; a máquina de Turing, por outro lado, apresenta maior complexidade para a representação de algoritmos, porém com a capacidade de aceitar, até mesmo, as linguagens recursivamente enumeráveis.

Contudo, embora possuam grande capacidade de representação de algoritmos e decisão sobre linguagens, os dispositivos de computação tradicionais são inflexíveis quanto a mudanças em tempo de execução[10]. Esta limitação, embora aparentemente desprezível, mantém forte relação com aprendizado [5] e com o poder computacional do dispositivo [6].

Buscando preencher essa lacuna, o conceito de adaptatividade emergiu, sendo matematicamente formalizado para aplicação em dispositivos computacionais [7]. Neste, um dispositivo adaptativo $AD = (ND_0, AM)$ com:

- ND_0 um dispositivo computacional não-adaptativo;
- AM um mecanismo adaptativo tal que $AM \subseteq BA \times NR \times AA$ e NR seja o conjunto de regras de ND_0;
- BA e AA são conjuntos de funções adaptativas, de forma que ambos contenham a ação vazia ($e \in BA \cup AA$).

Descritivamente, a adaptatividade é uma caminhada sobre um possivel espaço de (todos os) dispositivos de uma deteminada classe (e.g., autômatos) onde as funções adaptativas escolhem uma instância particular deste espaço para a execução do passo computacional não-adaptativo. É importante ressaltar que as funções adaptativas podem ser classificadas em *anterior* ($F_B \in BA$) e *posterior* ($F_A \in AA$), representando a ordem de execução destas em relação à transição não-adaptativa do dispostivo ND_k.

IV. Proposta de Algoritmo Adaptativo

Atualmente o uso de redes haplotípicas, especialmente aqueles gerados por sistemas computacionais como o aplicativo TCS, é de grande utilidade para os estudos de filogenética e fiologeografia, automatizando o processo não-imediato de construção da rede utilizando técnicas como a "parsimônia estatística" [2].

No entanto, o uso dessas ferramentas produz como resultado uma única rede haplotípica traduzindo a representação de apenas um modelo (como visto na seção II) e, assim, a inferência dos processos de evolução e mutação naturais segundo um único modelo de análise dos dados. No entanto, o uso de informações enganosas na análise filogeográfica guia a pesquisa a conclusões falsas; esse efeito ocorre, ainda, quando essas informações provêm da rede haplotípica [1].

Buscando auxiliar os estudos biológicos de filogeografia, propõe-se a unificação dos modelos geradores da rede haplotípica com o objetivo de, independentemente da modelagem escolhida, retratar os eventos naturais da melhor maneira observada e mais coerentemente possível. No entanto, para a utilização do método nested clade phylogeographic analysis, um único grafo se faz necessário e, assim, tal unificação vai além daquela de conceitos e inclui a dos próprios grafos.

Para a ocorrência deste, o uso de tecnologia adaptativa se mostra extremamente eficaz e adere-se naturalmente à representação gráfica da rede haplotíca, uma vez que esta mantém morfismo com autômatos [8], dispostivos computacionais imediatos para aplicação de adaptatividade.

A estratégia se traduz na busca, em todos os elementos do conjunto de redes haplotípicas para determinada população, por vértices do tipo zero, i.e. vértices (ou haplótipos) não mapeados na população mas necessários para manter a conectividade do grafo, ou pelo vértice representativo do centro de massa do grafo; esses vértices representam aspectos centrais do grafo, contendo alta densidade de informação por serem elementos-chave para a formação dos diversos ramos do grafo.

Após encontrado os nós com as propriedades acima enunciadas, inicia-se a busca por semelhanças entre os diversos grafos baseados no nó em análise, formando um novo grafo (nomeado grafo unificador) compartilhando as propriedades entre os diversos grafos e adicionando probabilidades às arestas. Formalmente, para um vértice v_i, o conjunto dos vértices vizinhos de v_i no grafo unificador será $\bigcup_k \{ v_{j_k} | v_{j_k}$ seja vizinha de v_i no grafo $k \}$; o valor de contagem da quantidade de arestas unindo o vértice v_i com um vértice v_j qualquer é determinado pela função

[10] A máquina de Turing universal, no entanto, é uma exceção: por possuir em sua fita de entrada a definição da máquina de Turing que irá interpretar, pode, portanto, alterar as regras daquela.

$\text{Count}(v_i, v_j) = \sum_{k=0}^{N-1} \text{Id}(k)$, onde $\text{Id}(k)$ é a função característica do conjunto gerado pela união em relação a aresta e N representa a quantidade de redes haplotípicas; se o grafo k contém v_j.

Por construção, o grafo unificador $G_\cup = (V_\cup, A_\cup)$ inicialmente definem-se os conjuntos $V_\cup = A_\cup = \text{Æ}$. Durante a busca por arestas semelhantes nas N redes haplotípicas, para cada vértice v_i em estudo, caso encontrada um elemento $(v_i, v_j) \in A$ da relação, então uma função adaptativa prévia F_B é disparada adicionando o vértice v_j em $V_\cup$ e criando uma aresta (v_i, v_j) no conjunto $A_\cup$ (relação simétrica).

Após a inserção do nó v_j, uma função adaptativa posterior F_A remove a aresta (v_i, v_j) e adiciona a aresta rotulada por ℓ, representada da forma $((v_i, v_j), \ell)$, onde ℓ é o rótulo da transição atribuído pela função de contagem $\text{Count}(v_i, v_j)$; $G_k \mid 0 \leq k < N$ no qual exista a relação (v_i, v_j), sendo assim representado por um número de maneira que $\ell \in \{0,1,2,\ldots\}$.

O grafo unificador, por fim, conterá todos os nós (ou haplótipos) dos N grafos concorrentes gerados pelos modelos iniciais indicando, através dos rótulos, as conexões com maior probabilidade de ocorrência após a unificação dos modelos. Espera-se, dessa maneira, diagnosticar a rede haplotípica que mais se assemelhe com a realidade encontrada na natureza e que melhor contribua com as inferências do método nested clade phylogeographic analysis ou outros métodos de estudo filogenéticos e filogeográficos.

Exemplificando a aplicação da técnica acima descrita, utiliza-se como entrada os três grafos representados pelas **Erro! Fonte de referência não encontrada.**, **Erro! Fonte de ferência não encontrada.** e **Erro! Fonte de referência não encontrada.**, interpretados como três redes haplotípicas diferentes (de até quatro haplótipos) para um conjunto de dados filogenéticos hipotético. É imediata a observação da diferença entre essas redes, expressando relações (de mutação) não-equivalentes entre os haplótipos nas diferentes figuras.

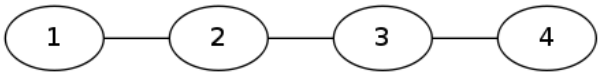

Fig. 2. Rede haplotípica de quatro haplótipos (ou grafo G_1), com mutações ocorrendo sequencialmente.

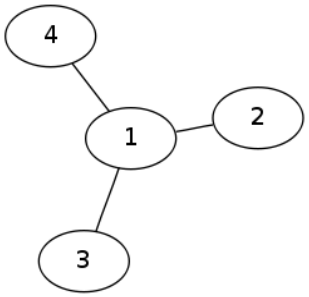

Fig. 3. Rede haplotípica de quatro haplótipos (ou grafo G_2), com todas as mutações diretamente derivadas do haplótipo **1**.

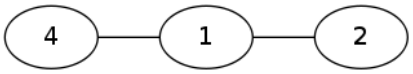

Fig. 4. Rede haplotípica de três haplótipos (ou grafo G_3), com mutações ocorrendo sequencialmente.

Das figuras, elege-se o *haplótipo 1* como o centro de massa e inicia-se, a partir deste, a formação do grafo unificador $G_\cup$, adicionando-se ao conjunto $V_\cup$ o *vértice 1*. Então, percorre-se os grafos G_1 a G_3 buscando todos os nós que mantém a relação simétrica $(1, k)$. E.g., de G_1 verifica-se a relação $(1,2)$, fazendo $V_\cup = \{1,2\}$ e $A_\cup = \{1,2\}$. A função adaptativa posterior F_A é então executada, removendo $(1,2)$ de $A_\cup$ e adicionando, $((1,2),1)$, fazendo-no $A_\cup = \{((1,2),1)\}$; este último nada mais é que a aresta $(1,2)$ nomeada pelo rótulo **1**, simbolizando o número de vezes que a relação foi encontra nos conjuntos *A* dos grafos analizados até então.

Esse procedimento algorítmico é repetido até o momento em que $V_\cup = \bigcup_{k=1}^{3} V_k$, $A_\cup = \bigcup_{k=1}^{3} A_k$ e todos os as arestas estejam rotuladas com o valor da frequência que aparecem. Nesta configuração, portanto, temos o grafo $G_\cup$ conforme indicado na **Erro! Fonte de referência não encontrada.**.

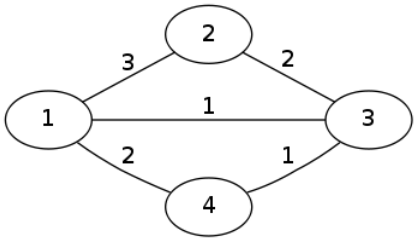

Fig. 5. Grafo unificador $G_\cup$.

Como é possível notar, $G_\cup$ mantém as relações das três

redes haplotípicas de entrada com a informação adicional de frequência em que as relações (mutacionais) ocorrem. Desta nova representação, é possível uma análise mais abrangente da evolução de populações, mesmo para processos como o nested clade phylogeographic analysis.

A figura a seguir (**Erro! Fonte de referência não ncontrada.**) destaca, em $G_{\cup}$, as relações com maior frequência, definindo um subgrafo com as mutações mais prováveis segundo os dados de entrada, podendo servir como entrada, por exemplo, para um processamento onde apenas uma rede haplotípica é admitida.

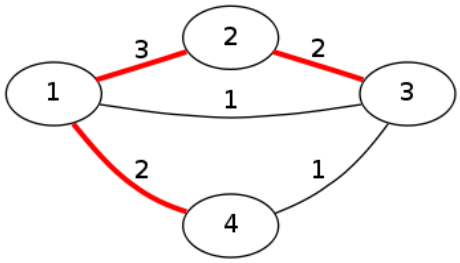

Fig. 6. Grafo unificador $G_{\cup}$ com as conexões mais prováveis em destaque (proximidade entre os vértices e a coloração das arestas).

V. Conclusão

Embora a notação gráfica da rede haplotípica seja uma ferramenta que agregue grande diferencial à pesquisa filogenética e filogeográfica, a não uniformidade nas representações para um mesmo conjunto de dados (devido aos diferentes modelos de construção das redes) torna-na objeto entrópico nas análises das populações [1]; a seleção de um único modelo, também, leva a perda de informações como a frequência com o qual algumas mutações ocorrem nas diversas modelagens dos dados.

A geração de um grafo unificador, por sua vez, unifica todos os benefícios fornecidos por cada um dos métodos de sintetização de redes haplotípicas além de adicionar novas informações, possibilitando ao utilizador extrair novas relações desta nova representação, como o exemplo do subgrafo da **Erro! Fonte de referência não encontrada.** e, inda, se ater ao antigo modelo de sua preferência.

Por fim, a adaptatividade trás benefícios na construção do grafo unificador, possibilitando não apenas a construção do mesmo em tempo de execução, mas também como a sua modificação para a inserção dos rótulos numéricos nas arestas representando a função de contagem das relações entre vértices.

A. Trabalhos Futuros

Tem-se em vista a complementação teórica do método proposto neste artigo, projetando-se uma modelagem de inferência indutiva para a extração de um conjunto de subgrafos ótimos segundo a ótica bio-evolutiva; neste, ainda, é possível realizar a extração do grafo unificador diretamente dos dados de entrada, independentemente dos métodos de síntese de rede haplotípica existentes.

Sugere-se, ainda, modificações no aplicativo TCS [2] para que gere como saída uma rede haplotípica baseado no grafo unificador para as múltiplas possíveis redes concebidas pela aplicação, ao invés de única rede através do método *statistical parsimony*.

Finalmente, também há a possibilidade de alterar o aplicativo GeoDis [9] para que suporte um grafo unificador como entrada de rede haplotípica e, então, realizar múltiplos processamentos paralelos do nested clade phylogeographic analysis.

Referências

[1] S. Joly, M. Stevens, and B. J. van Vuuren, "Haplotype networks can be misleading in the presence of missing data," *Systematic Biology*, vol. 56, no. 5, pp. 857–862, 2007.

[2] M. Clement, D. Posada, and K. A. Crandall, "Tcs: a computer program to estimate gene genealogies," *Molecular Ecology*, vol. 9, no. 10, pp. 1657–1660, 2000.

[3] A. R. Templeton, E. Boerwinkle, and C. F. Sing, "A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. i. basic theory and an analysis of alcohol dehydrogenase activity in drosophila," *Genetics*, vol. 117, pp. 343–351, October 1987.

[4] C. Papadimitriou and H. Lewis, *Elements of Theory of Computation*, Prentice-Hall, Ed. Prentice-Hall, 1998.

[5] H. Pistori and J. J. Neto, "Decision tree induction using adaptive fsa," *CLEI Eletronic Journal*, vol. 6, 2003.

[6] R. L. A. Rocha and J. J. Neto, "Autômato adaptativo, limites e complexidade em comparação com máquina de Turing," in *Proceedings of the second Congress of Logic Applied to Technology - LAPTEC 2000*. São Paulo: Faculdade SENAC de Ciências Exatas e Tecnologia, 2000, pp. 33–48.

[7] J. J. Neto, *Lecture Notes on Computer Science*. Springer-Verlag, 2001, ch. Adaptive Rule-Driven Devices - General Formulation and Case Study, pp. 234–250.

[8] B. C. Pierce, *Basic Category Theory for Computer Scientists*. MIT Press, 1991.

[9] D. Posada, K. A. Crandall, and A. R. Templeton, "Geodis: A program for the cladistic nested analysis of the geographical distribution of genetic haplotypes," *Molecular Ecology*, vol. 9, no. 4, pp. 487–488, 2000.

Ricardo Henrique Gracini Guiraldelli nasceu em Sorocaba, Brasil, em 10 de março de 1986 e gradou-se em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo. Desde 2009 é mestrando na EPUSP em Engenharia de Computação e membro da ACM (Association for Computing Machinery), atuando na área de Bioinformática e Fundamentos de Computação.

Ricardo Luis A. Rocha é natural do Rio de Janeiro-RJ e nasceu em 29/05/1960. Graduou-se em Engenharia Elétrica modalidade Eletrônica na

PUC-RJ, em 1982. É Mestre e Doutor em Engenharia de Computação pela EPUSP (1995 e 2000, respectivamente). Suas áreas de atuação incluem Tecnologias Adaptativas, Fundamentos de Computação e Modelos Computacionais.

Dr. Rocha é membro da ACM (Association for Computing Machinery) e da SBC (Sociedade Brasileira de Computação).