Universidade Federal de Alagoas
Instituto de Computação

Dissertação de Mestrado

Um Sistema Evolutivo para a Construção da Taxonomia dos Seres Vivos (SETAX)

LUIS HENRIQUE LEME PACHECO
rickpacheco@gmail.com

Maceió, 18 de Março de 2011

LUIS HENRIQUE LEME PACHECO

Um Sistema Evolutivo para a Construção da Taxonomia dos Seres Vivos (SETAX)

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:

Orientadora: Profa. Dra. Roberta Vilhena Vieira Lopes

Co-Orientador: Prof. Dr. Evandro de Barros Costa

Maceió, 18 de Março de 2011

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

P116u Pacheco, Luis Henrique Leme.
Um sistema evolutivo para a construção da taxonomia dos seres vivos (SETAX) / Luis Henrique Leme Pacheco. – 2011.
60 f. : il.

Orientadora: Roberta Vilhena Vieira Lopes.
Co-Orientador: Evandro de Barros Costa.
Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2011.

Bibliografia: f. 56-60.

1. Algoritmos evolucionários. 2. Otimização combinatória – Multiobjetivo. 4. Problema do caixeiro viajante. I. Título.

CDU: 004.421

Dissertação apresentada pelo aluno Luis Henrique Leme Pacheco como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional de Conhecimento pelo Programa de Pós-Graduação em Modelagem Computacional de Conhecimento, da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina:

Profa. Dra. Roberta Vilhena Vieira Lopes
UFAL – Instituto de Computação
Orientadora

Prof. Dr. Evandro de Barros Costa
UFAL – Instituto de Computação
Co-orientador

Prof. Dr. Manoel Agamemnon Lopes
UFAL – Centro de Ciências Agrárias
Examinador

Prof. Dr. Gildemberg Amorim Leal Junior
UFAL – Centro de Ciências Agrárias
Examinador

Prof. Dr. Guilherme de Alencar Barreto
UFC – Departamento de Engenharia de Teleinformática
Examinador

Maceió, março de 2011.

Agradecimentos

Eu tenho muito que agradecer, a começar pelo fato de estar vivo, e de poder compartilhar este trabalho com outros seres, iguais a mim, que gostam de se apoderarem dos conhecimentos herdados de nossos antepassados.

Eles são meus heróis e merecem meus agradecimentos, todos eles, aos seres humanos que viveram antes de mim, o meu muito obrigado!

Aos que participaram de minha vida, de forma direta, quero agradecer muito a minha mãe, que nas horas difíceis, manteve firmemente a vontade de escolarizar os filhos, ao meu pai cabe o agradecimento de neste ponto obedecer a minha mãe e ajudar na formação do meu carater.

A minha irmã Estelita e minha namorada Elisângela, alem de parentes próximos, todos estiveram presentes na longa caminhada deste Mestrado. Meu obrigado e meu carinho a todos.

Aos professores que passaram por minha vida, que foram muitos, quero agradecer especialmente a primeira, Prof[a]. Dr. Roberta, excelente professora e orientadora. Sem a senhora seria impossível e ao Prof. Agamemnon , muito obrigado!. Os Professores Dr. IG, Rafael Dr. Paraguaçu, também ao professor Dr. Evandro. Excelente professores.

Na UFAL quero agradecer ao secretario Vitor, que alem de ser um funcionário muito competente é um amigo e esta sempre pronto para atender qualquer solicitação.

Na sala, no convívio de cada dia, Tarsis e Carol, meus eternos amigos, que esteve presente durante todo o curso possui minha admiração e meu carinho.

Quero muito agradecer também a todos os colegas de turma e amigos, especialmente a Elvys, Higor, Jean, Marcos, Gustavo, Angela, Jarbas, PC e Fabrício. Foi bom conviver com vocês, Muito obrigado!

E finalmente à FAPEAL, pelo apoio financeiro em parte deste trabalho.

Resumo

Este trabalho apresenta um sistema que utiliza uma abordagem evolutiva para construção de árvores Filogenéticas denominado SETAX. Neste sistema o algoritmo genético baseado em tipos abstratos de dados (Genetic Algorithm Based on Abstract Data Types -GAADT) foi instanciado para encontrar a menor distância entre os grupos taxonômicos investigados, gerando um conjunto de árvores filogenéticas com o mesmo valor para o somatório das distância taxonômica entre as sub-árvores binárias que a compõe. Também é aqui apresentado, um estudo de árvores filogenéticas, os principais algoritmos para construção de árvores inspirados em métodos de inteligência artificial; um resumo biológico para o entendimento da construção de uma árvore filogenética, a instanciação do GAADT para construção de árvores filogenéticas de seres vivos bem como alguns resultados obtidos com o sistema proposto e comparações com os demais métodos.

Palavras-chave: Inteligência artificial, Filogenia, Árvores Filogenéticas, Biologia Computacional, Otimização Combinatória, Algoritmos Evolucionários e Algoritmos Genéticos.

Abstract

This work presents a system that use evolutionary approach to Phylogenetic tree construction called SETAX. In this system, the genetic algorithm based on abstract types of data (Genetic Algorithm Based on Abstract Data Types -GAADT)was instantiated to found the minimum distance between taxonomic groups investigated, generating a set of Phylogenetic trees with the same value for the sum of taxonomic distance between binary subtrees which composes it. Also report here, a study of Phylogenetic trees, the main algorithm for construction trees inspired by methods artificial intelligence; a biological summary for the knowledge of Phylogenetic trees contruction, the instanciation of GAADT to Phylogenetic tress of living beings as well as some results acquired by the proposed system and comparison with other methods.

Keywords: Artificial intelligence, Phylogeny, Phylogenetic trees, Computational Biology, Combinatorial Optimization, Evolutionary Algorithms and Genetic Algorithms.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Computação evolutiva ou evolucionária

Computação evolutiva ou evolucionária é uma área da ciência da computação que propõe um novo paradigma de processamento de dados. Este paradigma, diferentemente do convencional, não exige o conhecimento prévio dos elementos do domínio do problema para encontrar um resultado[20].

A computação evolucionária é baseada em mecanismos evolutivos encontrados na natureza, tais como a auto-organização e o comportamento adaptativo [19], [22]. Esses mecanismos foram descobertos e descritos por DARWIN (1859) em sua teoria da evolução das espécies. Nela, ele afirma que a vida na terra é o resultado de um processo de seleção, pelo meio ambiente, das espécies ou indivíduos mais aptos e adaptados, e por este motivo eles têm mais chances de reproduzir-se. Desta forma, a diversidade da vida, associada ao fato de que todos os seres vivos compartilham uma bagagem genética comum, pelo menos em termos de seus componentes básicos, é um exemplo eloquente das possibilidades do mecanismo de evolução natural.

Outras abordagens computacionais inspiradas na teoria da evolução das espécies surgiram como: programação genética [29] e estratégia evolucionária [41]. Na programação genética(PG), os indivíduos da população não são cadeias de "bits", mas sim programas de computador armazenados na forma de árvores sintáticas ou regras de semântica. Desta forma, os programas é que são os candidatos à solução do problema proposto, enquanto na Estratégia Evolucionária(EE) regras semântica, é que são candidatas a uma classificação.

As técnicas de Algoritmos Genéticos(AG), Estratégias Evolucionária(EE) e Programação evolucionária (PE) possuem algo fundamental em comum, cada uma delas trata reprodução, variação aleatória, competição e seleção de indivíduos de uma população. Estes quatro elementos formam a essência da Computação Evolutiva [4], [7].

Diante de uma perspectiva histórica, as primeiras iniciativas na área de computação evolucionária foram de biólogos e geneticistas interessados em simular os processos vi-

tais no computador, o que recebeu na época o nome de "Processos Genéticos"(PG)'[24]. Alguns desses cientistas, são: [8], [9], [23], Martin e [13]. O biólogo [39], simulou uma população de seres unicelulares, estrutura genética clássica (um gene, uma enzima), com estrutura diplóide, com cromossomos de 20 genes e 16 alelos permitidos em cada um.

Na década de 60, [27] e outros pesquisadores começaram a estudar os sistemas adaptativos complexos, que foram modelados como sistemas de aprendizagem de máquina. Esses modelos, conhecidos como algoritmos genéticos , implementavam populações de indivíduos contendo um genótipo, formado por cromossomos (que neste modelo eram representados por cadeias de "bits") aos quais se aplicavam operadores de seleção, recombinação, mutação e a inversão, este último operador não chegou a ser largamente utilizado segundo [44].

Uma das primeiras aplicações propostas para os algoritmos genéticos(cujo uso do termo cabe a Bagley na sua dissertação de 1967), seguindo o enfoque de Holland[27], foram os sistemas classificadores, que são sistemas de produção e usam os algoritmos genéticos como funções em uma parte do algoritmo global.

Desta forma, segue a Tabela 1.1 de [6] com um resumo das propriedades que caracterizam cada uma das abordagem referenciadas aqui de computação evolucionária.

	AG	**EE**	**PE**	**PG**
Representação	Cadeias binárias	Vetores reais	Vetores reais	Árvores
Auto-adaptação	Nenhuma	Desvio padrão e co-variâncias	Desvio padrão e coeficiente de co-relação	Nenhuma
O fitness é	Valor escalonado da função objetivo	Valor da função objetivo	Valor (escalonado) da função objetivo	Valor escalonado da função objetivo
Mutação	Operador secundário	Principal operador	Único operador	Um dos operadores
Recombinação	Principal operador	Diferentes variações, importante para a auto-adaptação	Nenhuma	Um dos operadores
Seleção	Probabilística	Determinística	Probabilística	Probabilística

Figura 1.1: Resumo adaptado de [6] sobre computação evolucionária.

Os AG possuem uma larga aplicação em muitas áreas científicas, dentre as quais podem ser destacadas:

- **Síntese de circuitos analógicos:** para uma certa entrada e uma saída desejada, por exemplo tensão, o AG gera a topologia, o tipo e o valor dos componentes do circuito;
- **Síntese de protocolos de comunicação:** determinação de quais funções do protocolo devem ser implementadas em hardware e quais devem ser implementadas em

software para que um desempenho aceitável seja alcançado;

- **Gerenciamento de redes:** supervisão do tráfego nos links e das filas nos "buffers"de roteadores para descobrir rotas ótimas e para reconfigurar as rotas existentes no caso de falha de algum link;
- **Otimização evolutiva multicritério:** otimização de funções com múltiplos objetivos que sejam conflitantes;
- **Problemas de otimização complexos:** problemas com espaços de soluções de dimensões elevadas. Exemplo: problema do caixeiro viajante, gerenciamento de carteiras de fundos de investimento;
- **Ciência biológicas:** modela processos biológicos para o entendimento do comportamento de estruturas genéticas.

Algoritmos Genéticos também podem ser aplicados em diversas áreas de desenvolvimento. Neste trabalho, destacamos as seguintes áreas: Setor de Petróleo e Gás, Musical, Telecomunicações, Saúde e na área de Biologia e Agronomia.

Petróleo e Gás-Inversão Sísmica: Conforme LINDEN (2008), o problema da inversão sísmica, que é um importante campo da geologia, consiste na determinação da estrutura dos dados de subsolo a partir da prospecção geológica, tendo como objetivo primário obter uma seção geológica ou um modelo 3D. Este problema é suscetível à aplicação de algoritmos genéticos, pois sua função objetivo é muito irregular, sendo altamente não linear, possuindo muitos mínimos e máximos locais e podendo apresentar descontinuidades.

Podemos concluir, que o uso de AGs torna-se bastante útil no campo de Petróleo e Gás, principalmente para o caso acima, pois o mesmo apresenta dados totalmente inconsistentes e incertos. A partir de sua implementação e com estudos detalhados, podemos obter informações valiosas para a solução desses problemas.

Música: foi apresentado em 1999 no evento CEC99 (IEEE - International Conference on Evolutionary Computation) um ambiente interativo, utilizando Algoritmos Genéticos, para a avaliação de sequências de acordes tocadas em arquivos MIDI. No caso, os indivíduos da população foram definidos em grupos de quatro vozes (soprano, contralto, tenor e baixo) ou coros. Cada um é avaliado segundo três critérios: melodia, harmonia e oitavas. A composição destes três critérios definia a aptidão (fitness) definida pela função de seleção, que retorna o melhor indivíduo ou melhor coro. Um ciclo genético é operacionalizado, criando novos indivíduos dos anteriores e procurando sempre pelo melhor. Quando um novo grupo é selecionado, ele é tocado em MIDI. A duração do ciclo genético determina o ritmo da evolução. O sistema criado foi denominado Vox Populi (Fukushima, 1999).

Telecomunicações: de acordo com (Blanchard,1994), no evento WCCI'94 (World Congress on Computational Intelligence), mostrou uma série de soluções promissoras a situações reais utilizando Algoritmos Genéticos. Blanchard mostrou-se o caso da US West, uma companhia regional de telecomunicações do estado do Colorado, que vem usando um sistema baseado em AGs que possibilita projetar, em duas horas, redes óticas especializadas, trabalho que levaria seis meses utilizando especialistas humanos. O sistema produz resultados ainda dez por cento melhores que os realizados pelo homem. A companhia, afirmou que o sistema foi capaz de trazer uma economia de 100 milhões de dólares até o final do século 20.

Saúde: técnicas de realidade virtual são usadas na construção de software para auxiliar o treinamento, planejamento e realização de cirurgias. Em [32] e [40] são descritos alguns software de planejamento cirúrgico. A entrada destes software são informações extraídas de exames, tais como tomografia computadorizada e ressonância magnética da parte do organismo a ser operada. Com base na entrada fornecida, o software constrói um organismo virtual para que o médico possa, dentro do ambiente virtual, testar vários procedimentos cirúrgicos e assim planejar a sua estratégia de trabalho. Já em [34], são encontradas descrições de um software de telecirurgia que permite que um médico na Alemanha opere um paciente no Brasil. As entradas deste software são informações captadas por sensores instalados sobre o organismo do paciente, cujas informações são enviadas para o servidor remoto, convertidas em um organismo virtual, sobre o qual o médico realiza a operação, e suas ações são enviadas de volta a sala cirúrgica, onde braços mecânicos realizam a operação com a ajuda de uma equipe médica.

Biologia: um dos principais problemas dentro da biologia molecular é a inferência das características de um ser vivo. Esse problema consiste em determinar a cadeia de DNA responsável pelo fenótipo de um conjunto de espécies da mesma família. Os algoritmos que realizam esta tarefa são identificados como algoritmos de alinhamento. Um algoritmo de alinhamento recebe como entrada um conjunto de DNA de espécies e tenta arrumar estes DNA's em linhas de tal modo que o caracter que ocupa a mesma posição em todas as linhas seja o mesmo. Os algoritmos de alinhamento mais conhecidos são o BLAST [1] e o FAST [11].

Agronomia: o melhoramento genético da cana-de-açúcar busca desenvolver variedades adaptadas as diversas condições edafoclimaticas [1] de cada região do Brasil. Basicamente as novas variedades devem ter ganhos na área agroindustrial, tais como: maior produtividade e com maior tolerância ao estresse hídrico, maior resistência às pragas e doenças e melhor adaptação à colheita mecanizada. Os órgãos de pesquisa que desenvolvem programas de melhoramento genético da cana utilizam, geralmente, conhecimentos das

[1] relação planta-solo-clima para plantio. Os fatores edafoclimáticos são referidos como os mais importantes não só para o desenvolvimento das culturas, como também para a definição de sistemas de produção.

áreas de biotecnologia vide Figura 1.2 , ciências do solo, nutrição de plantas, climatologia, fisiologia, fitopatologia, entomologia, economia e outras.

Figura 1.2: Produção de mudas em laboratório, fonte site da Embrapa [2].

A maioria das características da cana-de-açúcar é herdada de forma aditiva. Por exemplo, o cruzamento de duas variedades altas deve resultar numa variedade ainda mais alta. Porém, existe uma importante exceção que é a característica para a produtividade, em que as variações genéticas aditivas e não-aditiva parecem estar em igual grau de importância. Isso vem sendo o principal desafio nas pesquisas de melhoramento genético da cana-de-açúcar.

Outra questão bastante relevante em relação ao melhoramento genético da cana é a avaliação de novas variedades quanto à adaptação a diferentes ambientes. Isso é importante para a recomendação das melhores variedades para as regiões mais aptas.

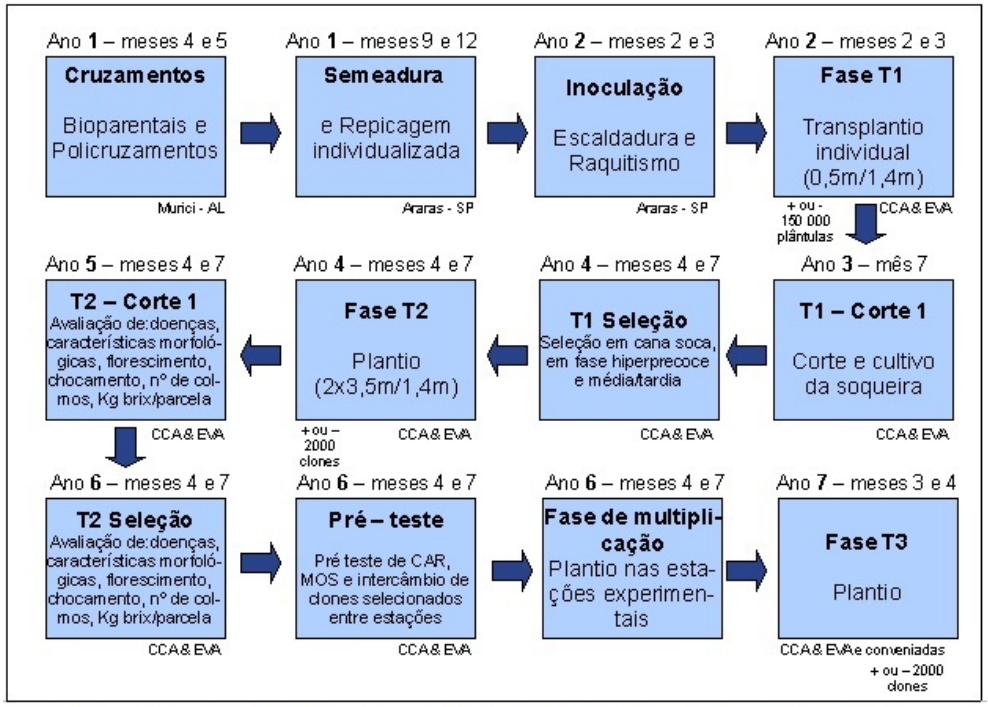
A Rede Interuniversitária para o Desenvolvimento do Setor Sucroalcooleiro (Ridesa) possui a estação de pesquisa Serra do Ouro no município de Murici, AL. A unidade está na latitude 9º 13 S, a 500 metros de altitude, onde a pluviosidade média é de pelo menos dois mil milímetros anuais e as temperaturas médias, de 19,5 a 26,5º C (Celsius).

Essas condições tornam o local muito propício para o florescimento da maioria das variedades de cana-de-açúcar e, também, para a boa fertilidade do pólen. As mudas de cana são plantadas em campos experimentais das unidades de pesquisa.

Por esses motivos este domínio será o estudo de caso utilizado nesta dissertação.

1.1.1 Motivação da Dissertação

Nesta dissertação existem duas motivações: uma biológica e outra computacional. A motivação biológica consiste em determinar as relações evolutivas de um conjunto de espécies. O número de relações evolutivas cresce exponencialmente à medida que novas espécies são consideradas. Na tentativa de gerar a história evolutiva das espécies manualmente,



Legenda: T1 = primeira fase de seleção; T2 = segunda fase de seleção; T3 = terceira fase de seleção; FE = fase experimental; FM = fase de multiplicação; CM = curva de maturação; CCA = Centro de Ciências Agrárias; EVA = Estação Experimental de Valparaíso; CAR = Carvão; CBP = complexo broca-podridão; ESC = escaldadura-das-folhas; ESV = estrias-vermelhas; FER = ferrugem; MOS = mosaico; NEM = Nematóides;RAQ = raquitismo-da-soqueira.

Fonte: Universidade Federal de São Carlos. Centro de Ciências Agrárias (2008).

Figura 1.3: Fluxograma de programa de melhoramento genético da cana-de-açúcar.

alguns pesquisadores passaram a representá-la por estruturas gráficas denominada árvore filogenética. Isso inviabiliza a construção da história evolutiva das espécies manualmente. Desta forma surgem os primeiros programas para construção de árvore filogenética.

Apesar de as árvores filogenéticas serem de fácil manipulação, a construção de árvores filogenéticas por processos manuais foi se tornando cada vez mais complexa. Além disso, esses processos requerem muito tempo do filogeneticista e propiciam a introdução de erros humanos na análise [2].

A construção de árvores filogenéticas é um problema N-P completo, pois o número de possíveis árvores para N espécies é N!. Desta forma os algoritmos que garantem a obtenção da solução ótima são infactíveis para problemas NP-completos como o de construção de árvore filogenéticas. Então como resolvê-los? Abrindo mão da garantia de obtenção da solução ótima em prol da factibilidade.

Por esta razão, os métodos de inferência filogenéticas normalmente se baseiam em algoritmos de agrupamento ou estratégias heurísticas de busca que permitem minimizar a quantidade de tempo gasto analisando as árvores candidatas [31]. A idéia por trás destes algoritmos é que uma solução boa, não necessariamente a ótima, é melhor que nenhuma solução ou que uma solução tomada aleatoriamente do espaço de busca. Com isso, propor

algoritmos que explorem o espaço de busca de forma eficaz, visando localizar soluções de boa qualidade, e que eventualmente podem corresponder à solução ótima.

Desta forma, os algoritmos genéticos, estão estabelecidos sobre os seguintes princípios:

- para a maioria dos problemas existe uma função de codificação que mapeia os possíveis resultados do problema na representação adotada para o cromossomo;
- para todo cromossomo existe uma função de adaptação, a qual informa a qualidade do resultado representado por este cromossomo para o problema;
- os cromossomos mais adaptados geram descendentes (pela ação dos operadores genéticos), os quais tendem a herdar as características adaptadas do cromossomo ancestral e as características dos cromossomos menos adaptados tendem a desaparecer.

Recentemente, estratégias envolvendo algoritmos evolutivos têm sido investigadas nesta área, mostrando resultados relevantes ([12]; [37]; [47]).

Nesse sentido, propõem-se usar o modelo de Algoritmo Genético Baseado em Tipos Abstratos de Dados (GAADT). Trata-se de um paradigma novo, utilizado neste trabalho, originalmente proposto por [45] e [46] e trabalha sobre um ambiente definido como uma estrutura na qual um dos componentes é a população. Segundo este modelo, as mudanças ambientais são vistas como o marco do início de um novo período de evolução durante o qual os cromossomos da população atual irão sofrer a ação dos operadores genéticos com o intuito de construir uma nova população formada somente por cromossomos que satisfazem aos requisitos do ambiente atual. Após o período de evolução vem o período de estagnação, durante o qual a população não evolui. O período de estagnação é finalizado quando uma nova alteração ambiental ocorre, dando início a um novo ciclo de um período de evolução seguido por um período de estagnação.

O resultado do problema para o ambiente atual é o cromossomo mais adaptado na população de estagnação atingida. A cada período de evolução existe uma população de cromossomos extintos associada. Os cromossomos desta população são provenientes de uma das populações já trabalhadas pelo algoritmo genético, os quais foram avaliados e descartados devido ao seu baixo grau de adaptação às configurações atuais do ambiente.

A transmissão das características adaptadas ao ambiente dos cromossomos pais que devem ser passadas para os cromossomos filhos.

Realizado pelo operador de cruzamento apresentado no GAADT constrói novos cromossomos somente com as características responsáveis pela adaptação dos cromossomos pais ao ambiente, as quais serão denominadas de genes dominantes. Uma função denominada grau de adaptação do gene é definida para informar o quanto uma dada característica pode contribuir com a adaptação do cromossomo ao ambiente.

A possibilidade dos cromossomos não adaptados ao ambiente evoluírem e até conduzirem à geração de cromossomos mais adaptados é trabalhada pelo operador de mutação do GAADT, que é aplicado somente sobre os cromossomos não adaptados, como uma forma de garantir a presença das características adaptadas ao ambiente destes cromossomos nas próximas gerações.

Também faz parte do ambiente do GAADT o registro da história genealógica (táxon) dos cromossomos da população atual, para que uma explicação sobre o resultado encontrado para um dado problema possa ser gerada sempre que necessário.

1.1.2 Objetivos da Dissertação

O objetivo desta dissertação, é propor um sistema evolutivo para a construção de árvore filogenética que usa o GAADT para construir a inferência filogenética. Tal sistema deve:

- manipular dados biológicos com diferentes representações;
- conter a explosão combinatorial de todas as possíveis soluções geradas pela presença de dados ambíguos e incompletos com o GAADT;
- permitir ao filogeneticista acompanhar o raciocínio desenvolvido pelo sistema através da reconstituição da história taxonômica dos cromossomos da população de estagnação;
- permitir a construção de árvores filogenéticas pela análise de pequenos e grandes volumes de dados;
- fornecer uma interface gráfica amigável.

1.1.3 Organização da Dissertação

Para cumprir os objetivos definidos nas seções anteriores deste capítulo, organizou-se este trabalho em cinco capítulos, incluindo esta introdução.

No capítulo 2, será demostrado um histórico sobre Árvore Filogenética e definiremos a Descrição Formal da Árvore Filogenética e analisaremos as matrizes distância utilizadas pelos filogeneticistas para armazenar os dados disponíveis sobre a evolução das espécies investigadas. Em seguida, serão apresentados alguns problemas gerados pela ocorrência de dados incompletos e ambíguos, e os critérios utilizados para solucionar esses problemas e demostraremos alguns algoritmos de construção de árvores filogenéticas que são: A Regra de Inclusão e Exclusão, O Algoritmo de Wagner e O Algoritmo das Médias (UPGMA).

No Capítulo 3, será apresentada a especificação do Algoritmo Genético Baseado em Tipos Abstratos de Dados(GAADT) ao mesmo tempo em que seus tipos básicos e opera-

dores vão sendo instanciados para construir árvores filogenéticas. Após a descrição dessa implementação, será demostrada uma aplicações prática do sistema proposto.

No capítulo 4, será mostrados o sistema proposto para construção de Arvore Filogenética(SETAX) e os resultados experimentais do sistema desenvolvido para construir árvores filogenéticas pela análise de pequenos e grandes volumes de dados com e sem ambiguidades e comparados com outros algoritmos para construção de árvore filogenética baseado em métodos de inteligencia artificial.

No capítulo 5, será mostrada a conclusão dessa dissertação e as perspectivas de trabalhos futuros a serem desenvolvidos.

Capítulo 2

Árvore Filogenética

2.1 Introdução

Para que o estudo dos seres vivos fosse mais preciso, cientistas adotaram uma sistemática de classificação que se baseia na semelhança entre as espécies, onde cada reino se divide em filos, cada filo se divide em classes, cada classe em ordens e assim por diante, até atingir o nível "espécie"[1]. Espécies com características comuns formam um gênero. Gêneros com características em comum formam uma família. As famílias se agrupam em ordens e estas em classes. As classes se agrupam em filos. A reunião de filos forma o reino. Em alguns casos existem subdivisões como: subfilo, subclasse, subgênero, etc. Os principais reinos são: Reino Animalia, Reino Plantae, Reino Fungi, Reino Protista e Reino Monera. Há novos reinos sendo criados.

Epidendrum foi um dos primeiros gêneros de orquídeas tropicais a serem publicados. Ele foi criado pelo primeiro sistema de classificação dos seres vivos, Lineu (1759), com a descrição da espécie Epidendrum nodosum L. coletada em Porto Rico, hoje conhecida como Brassavola nodosa. O Epidendrum nocturnum Jacquin foi selecionado posteriormente como sua nova espécie para evitar a alteração do nome de mais de mil espécies quando a antiga planta tipo de Epidendrum foi removida para o gênero Brassavola. O nome do gênero é uma referência ao fato das primeiras orquídeas atribuídas a este gênero serem epífitas [2].

Sua circunscrição exata, ou seja, que espécies pertencem a este gênero e quais devem ser atribuídas a outros é discutível. Epidendrum já incluiu muitas espécies hoje pertencentes a outros gêneros, como Encyclia e Prosthechea. Quando Lineu propôs a criação do gênero Epidendrum, havia somente oito gêneros de orquídeas. Assim, praticamente todas as primeiras orquídeas epífitas descobertas eram classificadas como Epidendrum. Ao longo

[1] Denomina-se espécie o conjunto de indivíduos que, por terem a mesma constituição genética possam gerar descendência semelhante.

[2] são aqueles que vivem sobre vegetais. Podem ser animais que vivem associados à um vegetal ou plantas que vivem sobre outra planta, neste caso, são chamadas de epífitas.

dos anos, conforme outras espécies foram sendo descobertas e agrupadas, muitos outros grandes gêneros foram criados a partir dele. Sua delimitação foi se estreitando mas, no entanto, o gênero sempre tem servido para agrupar espécies de Laeliinae que não encaixam bem em outros gêneros. Isso acontece também com diversos outros gêneros da família Orchidaceae, os principais dentre eles são Pleurothallis, 'Bulbophyllum, Oncidium, Maxillaria e Dendrobium. Todos estes gêneros são grandes e por demais heterogêneos.

A longo das décadas, a classificação de Epidendrum passou por diversas mudanças conforme os conceitos adotados pelos taxonomistas. Houve época em que se preferia dividir mais e então o gênero diminuía de tamanho, depois surgiam novos taxonomistas com outros conceitos e juntavam tudo novamente. No entanto, a mudança do conceito de classificação das espécies hoje passa por grandes mudanças com a incorporação dos conceitos oriundos da filogenia, ou seja, acrescentando informações sobre a evolução das espécies a partir de pesquisas genéticas. Atualmente, muitos taxonomistas consideram que a maior divisão no nível de gênero é mais didática que divisões em níveis de subgênero ou seções.

Com isso, a primeira representação diagramática da história ou genealogia de seres vivos foi possivelmente a de Lamark, em 1809. Um diagrama desse gênero é o único desenho que se encontra na Origem das Espécies (1859) de Darwin. Em 1866, E. Haeckel fez o que Darwin não realizara: uma tentativa de filogenia geral, abrangendo todos os grupos de seres vivos, microorganismos, plantas e animais; representou-a por outro diagrama denominado de árvore filogenética.

Para [38], filogenia pode ser melhor explicada como sendo uma estrutura em forma de árvore que define certos relacionamentos ancestrais entre conjuntos relacionados de objetos (proteínas, espécies, etc.). O nó na base da árvore, denominado raiz, representa o ancestral em comum a todos os objetos relacionados.

2.2 Descrição Formal da Árvore Filogenética

Uma árvore filogenética é uma estrutura F = (X, Υ, Φ, Γ), onde:

- X é um conjunto finito de espécies diferente do conjunto vazio,
- Υ é um conjunto finito das características apresentadas pelos elementos de X,
- Φ é uma relação de ordem parcial sobre os elementos do conjunto X, tal que:

  I - $\forall$ x $\in$ X, x Φ x; (reflexiva)

  II - $\forall$ x_1, $x_2 \in$ X, se x_1 Φ x_2 e x_2 Φ x_1 então $x_1 = x_2$; (anti-simétrica)

  III - $\forall$ x_1, $x_2 \in$ X, se x_1 Φ x_2, então x_2 Φ x_1; (simetrica)

  IV - $\forall$ x_1, x_2, $x_3 \in$ X, se x_1 Φ x_2 e x_2 Φ x_3, então x_1 Φ x_3. (transitiva)

A qual é representada pelo gráfico $G = (X,\Phi\ )comcadaarcoa_k = (x_{k,1}, x_{k,2})$ de G rotulado pela seqüência dos elemento do conjunto $k = xk, 2xk, 1$

- Γ : X $\rightarrow \mathbb{P}(\ \Upsilon\ )$, a qual é representada pelo gráfico G = (X, Φ) formado pelo arco $a_k = (x_{k,1}\ ,\ x_{k,2}\ )$ tal que $\forall x_{k,1}x_{k,2} \in X(\cap$ C gerou $(x_{k,2}) = X_{k,2})\ C$ gerou $(x_{k,1}) = \{x|x \in X(x_{k,1}$ gerou $x)$ tal que $\sum_x |x \in Xex_1$ gerou x_2 de G rotulado pela sequência dos elemento do conjunto $\Sigma_k = \Gamma\ (\ x_{k,2}\ )$ - $\Gamma\ (\ x_{k,1}\ )$.

Exemplo, seja F = (X, Y, gerou, Γ), onde:
X = {esp_1 , esp_2, esp_3, esp_4, esp_5, esp_6, esp_7 };
Υ = {c_1 , c_2, c_3, $c_{1'}$, $c_{2'}$, $c_{3'}$, $c_{2''}$ };
esp_1 gerou esp_1;
esp_1 gerou esp_2;
esp_1 gerou esp_3;
esp_3 gerou esp_3;
esp_3 gerou esp_4;
esp_3 gerou esp_5;
esp_5 gerou esp_5;
esp_5 gerou esp_6;
esp_5 gerou esp_7;
$\Gamma\ (esp_1) =\ c_1$, c_2, c_3;
$\Gamma\ (esp_2) =\ c_{1'}$, c_2, c_3;
$\Gamma\ (esp_3) =\ c_1$, $c_{2'}$, c_3;
$\Gamma\ (esp_4) =\ c_{1'}$, $c_{2''}$, c_3;
$\Gamma\ (esp_5) =\ c_1$, $c_{2'}$, $c_{3'}$;
$\Gamma(esp_6) =\ c_{1'}$, $c_{2'}$, $c_{3'}$;
$\Gamma(esp_7) =\ c_1$, $c_{2'}$, $c_{3'}$;
podendo ser representada graficamente pela Figura2.1 .

A análise da árvore filogenética da Figura 2.1 pode ser assim descrita: inicialmente a espécie esp_1 apresentava as características c_1, c_2 e c_3 a qual deu origem as espécies esp_2 e esp_3 que apresentam as características $c_{1'}$, c_2, c_3 e c_1, $c_{2'}$, c_3. A espécie esp_3 deu origem as espécies esp_4 e esp_5 que contem as características $c_{1'}, c_2$

, $c_3 ec_1, c_{2'}, c_{3'}$. A espécie esp_5 deu origem as espécies esp_6 e esp_7 contendo as características $c_{1'}$, $c_{2'}$, $c_{3'}$ e c_1, $c_{2'}$, $c_{3'}$.

Todas as características principais analisadas pelos processos filogenéticos correspondem às modificações ocorridas em qualquer expressão fenotípica de um conjunto de espécie com base genética. Como exemplo, a expressão fenotípica "asa"inserida no conjunto de insetos holometabólicos da Figura 2.2 apresenta as características: asas posteriores semelhantes às asas anteriores (A,B) e halter [3] (C,D).

[3]Halter- Característica que corresponde à presença só das asas posteriores.

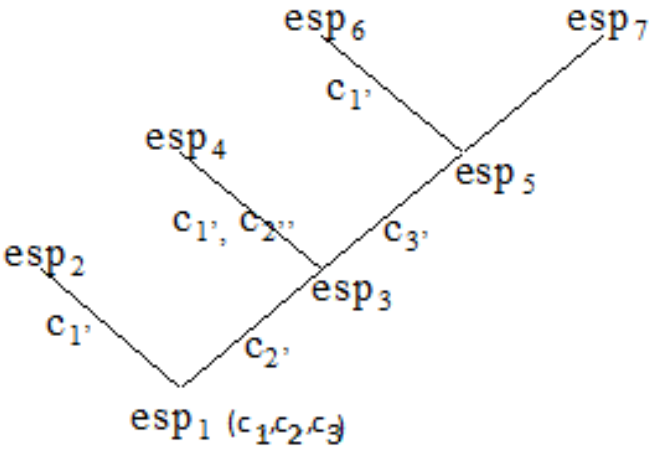


Figura 2.1: Exemplo de Árvore filogenética

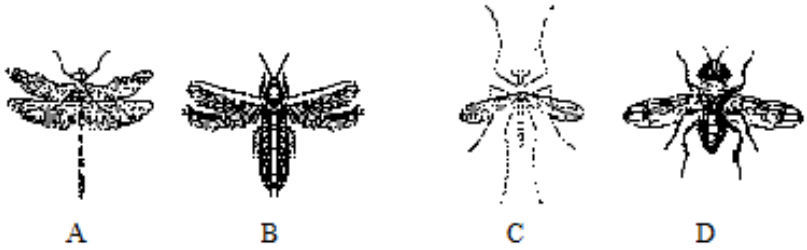


Figura 2.2: Insetos da família holometabólicos.

Já nos casos dos vírus, as expressões fenotípicas "presença de envoltório", "ausência de envoltório", fita "dupla", fita "simples", e natureza do ácido nucléico viral é "RNA"/ou "DNA"são analisadas na construção da árvore filogenética dessa espécie. A Figura 2.3 apresenta certos vírus que apresentam estas características.

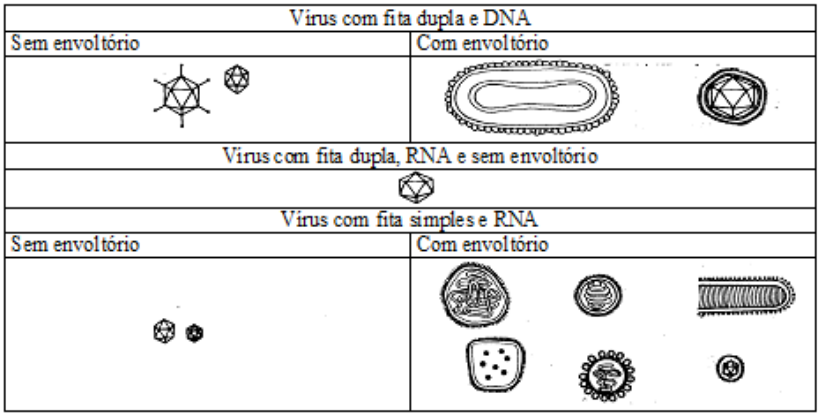


Figura 2.3: Representação esquemática de certos vírus.

Os processos da natureza responsáveis pela alteração das expressões fenotípicas das espécies durante a sua evolução recebem o nome de *processos filogenéticos fundamentais* [10] e compreendem os processos de:

- *anagênese*: faz a transmissão das expressões fenotípicas da espécie ancestral para a espécie descendente com algumas modificações;
- *cladogenêse*: responsável pela divisão de uma espécie em duas ou mais espécies;
- *estasigênese*:transmite as expressões fenotípicas da espécie ancestral para a espécie descendente sem modificações;
- *extinção*, responsável pela morte de uma espécie.

Na árvore filogenética as relações de parentesco entre as espécies são:

- relação ancestral direto, ocorre entre duas espécies esp_i e esp_j onde esp_i gerou esp_j, ou vice-versa (por exemplo, na Figura 2.1 , esp_1 é ancestral direto de esp_2 e esp_3);
- relação ancestral indireto, ocorre entre duas espécies esp_i e esp_j onde a espécie esp_i gerou a espécies esp_{i+1}, a espécie esp_{i+1} gerou a espécies esp_{i+2}, e assim por diante até a especie esp_{j-1} gerar a especie esp_j ou vice-versa (como na Figura 2.1, esp_1 é ancestral indireto de esp_4);
- relação ancestral comum, ocorre entre duas espécies esp_i e esp_j que não apresentam uma relação ancestral direto ou indireta entre si, mas possuem uma espécie esp_k que é ancestral direto ou indireto a ambos (por exemplo, na Figura 2.1 , esp_2 e esp_3 tem como ancestral comum esp_1).

2.3 Matriz Polarizada

Normalmente, as informações sobre a evolução das características apresentadas pelas espécies investigadas são armazenadas em uma matriz polarizada que relaciona a espécie à série de transformação polarizada (STP) de cada uma das expressões fenotípicas apresentadas pelas espécies investigadas [2] [10] [30] [3].

Essa série de transformação polarizada informa a ordem cronológica na qual as características dessa expressão fenotípica ocorreram ao longo do tempo. Como exemplo, sejam Θ ,ϵ e ς características de uma mesma expressão fenotípica. Dependendo da ordem na qual essas características surgiram ao longo da evolução das espécies na Terra, pode-se ter as seguintes séries de transformação polarizada, vide Figura 2.4.

As séries de transformação polarizada lineares do tipo, *a característica* Θ *gerou a característica* ϵ *e a característica* ϵ *gerou a característica* ς, mostra uma relação ancestral

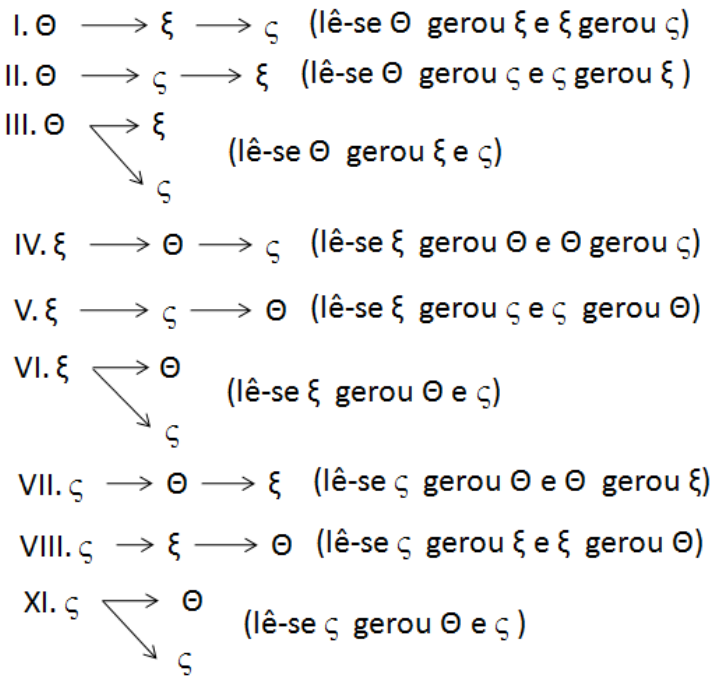


Figura 2.4: Séries de transformação polarizadas.

direto entre as espécies que apresentam a característica Θ e ϵ, uma relação ancestral direto entre as espécies que apresentam a característica ϵ e ς, e uma relação ancestral indireto entre as espécies que apresentam a característica Θ e ς.

Enquanto que as séries de transformação polarizadas paralelas do tipo, *a característica* Θ *gerou a característica* ϵ *e a característica* ς, expressam uma relação ancestral direto entre as espécies que apresentam a característica Θ e ϵ, uma relação ancestral direto entre as espécies que apresentam a característica Θ e ς, e uma relação ancestral comum entre as espécies que apresentam a característica ϵ e ς.

Porém, para que uma série de transformação polarizada possa ser armazenada em uma matriz, se faz necessário primeiro codificá-la de modo que, de posse desse código, qualquer pessoa possa reconstituir a série de transformação polarizada de uma expressão fenotípica a partir das informações contidas nessa matriz.

Se as características de uma expressão fenotípica, em uma série de transformação polarizada, são transmitidas de maneira linear como nas séries de transformação I, II, IV, V, VII e VIII mostradas anteriormente, a sua codificação será processada pela enumeração das características ocorridas nessa expressão fenotípica em ordem cronológica como na Tabela 2.1.

Porém, se as características de uma expressão fenotípica são transmitidas de maneira polarizada paralela, como nas séries de transformação III, VI e IX mostradas anteriormente, então deve-se primeiro decompor cada série desse tipo, em tantas séries de transformação polarizada lineares quanto for o número de características ocorridas simul-

Tabela 2.1: *Codificação em uma matriz polarizada de séries de transformação polarizadas linear.*

P	$\rightarrow \epsilon \rightarrow \varsigma$	$\rightarrow \varsigma \rightarrow \epsilon$	$\rightarrow \theta \rightarrow \varsigma$	$\rightarrow \varsigma \rightarrow \theta$	$\rightarrow \theta \rightarrow \epsilon$	$\varsigma \rightarrow \epsilon \rightarrow \theta$
Espécie(θ)	0	0	1	2	1	2
Espécie(ϵ)	1	2	0	0	2	1
Espécie(ς)	2	1	2	1	0	0

taneamente na série de transformação polarizada paralela em decomposição. Segundo, codificam-se separadamente cada uma dessas séries de transformação polarizada lineares vide Tabela 2.2.

Tabela 2.2: *Codificação em uma matriz polarizada de séries de transformação polarizadas via decomposição.*

Decomposição	$\theta \rightarrow \epsilon$	$\theta \rightarrow \varsigma$	$\epsilon \rightarrow \theta$	$\epsilon \rightarrow \varsigma$	$\varsigma \rightarrow \theta$	$\varsigma \rightarrow \epsilon$
Espécie(θ)	0	0	1	0	1	0
Espécie(ϵ)	1	0	0	0	0	1
Espécie(ς)	0	1	0	1	0	0

A utilização da matriz polarizada restringe o tipo de dados analisados aos dados de natureza morfológica, deixando de lado dados de outras naturezas, tais como genética, citológica, etc., que são utilizados para melhorar a compreensão do processo de evolução das espécies [28].

2.4 Matriz Característica

Algumas vezes as informações sobre a evolução das características apresentadas por um conjunto de espécies são representadas em uma matriz característica. Uma matriz característica é uma versão simplificada da matriz polarizada que relaciona cada espécie às características terminais do conjunto de espécies investigadas. Diz-se que, x é uma característica terminal do conjunto de espécies investigadas, se nenhuma das características apresentadas pelas espécies do conjunto forem geradas a partir de x. Por exemplo, sejam $\Theta, \omega, \epsilon, \rho, \tau$, κ, $\pi, \alpha, \beta, \sigma$ e δ características das expressões fenotípicas do conjunto de espécies $esp_1, esp_2, esp_3, esp_4$ que possuem as séries de transformação polarizada mostrada na Figura 2.5.

A espécie esp_1 apresenta as características $\Theta, \delta, \rho, \alpha$ e κ , a espécie esp_2 apresenta as características $\omega, \delta, \rho, \alpha$ e κ ; a espécie esp_3 apresenta as características $\omega, \epsilon, \tau, \alpha$ e π; e a espécie esp_4 apresenta as características $\omega, \epsilon, \tau, \beta$ e σ . A Tabela 2.3 apresenta a matriz polarizada para essas espécies.

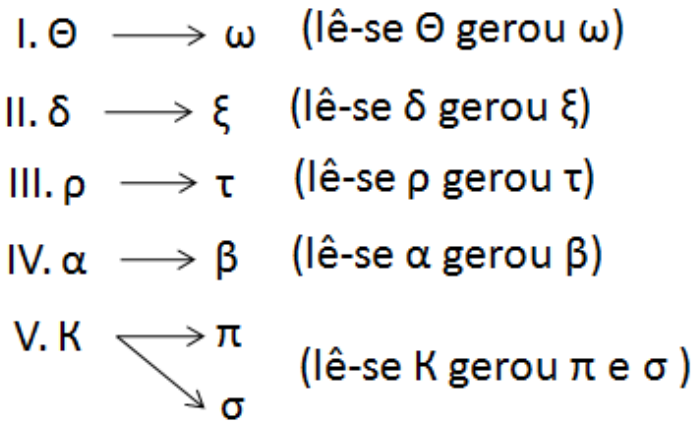


Figura 2.5: Séries de transformação polarizadas.

Tabela 2.3: *Matriz polarizada.*

Decomposição	$\theta \rightarrow \omega$	$\delta \rightarrow \epsilon$	$\rho \rightarrow \tau$	$\alpha \rightarrow \beta$	$\kappa \rightarrow \pi$	$\kappa \rightarrow \sigma$
esp_1	0	0	0	0	0	0
esp_2	1	0	0	0	0	0
esp_3	1	1	1	0	1	0
esp_4	1	1	1	1	0	1

A partir da Tabela 2.3, constrói-se a matriz característica das espécies investigadas na Tabela 2.4. Primeiro coloca-se na segunda linha e na primeira coluna a letra romana maiúscula A para representar a espécie esp_1, na terceira linha e na primeira coluna a letra romana maiúscula B para representar a espécie esp_2, e assim sucessivamente até que, todas as espécies da Tabela 2.3 estejam representados por uma letra romana maiúscula. Em seguida, coloca-se na primeira linha e na segunda coluna o número 1(um) para representar a característica terminal ω, na primeira linha e na terceira coluna o numero 2(dois) para representar a característica terminal ϵ, e assim sucessivamente até que, todas as características terminais das séries de transformação polarizadas da Tabela 2.3 estejam representados por um número.

Os elementos dessa matriz, das i linhas e j colunas, com 2(dois) $\leq i \leq$ número de linhas da matriz polarizada analisada e 2(dois) $\leq$ j $\leq$ número de colunas da matriz polarizada analisada, são representados por $c_{i,j}$ e tem o mesmo valor do elemento $p_{i+1,j}$ da matriz polarizada analisada [35].

Tabela 2.4: *Matriz característica.*

	1	2	3	4	5	6
A	0	0	0	0	0	0
B	1	0	0	0	0	0
C	1	1	1	0	1	0
D	1	1	1	1	0	1

2.5 Matriz Similaridade

A matriz similaridade é uma matriz quadrada de dimensão n que relaciona duas espécies a uma medida de similaridade, onde n é igual ao número de espécies investigadas mais 1(um).

Os elementos dessa matriz são obtidos a partir da análise dos elementos $c_{u,v}$ da matriz característica e são representados por $s_{i,j}$. Os elementos da primeira linha e coluna contêm uma letra romana maiúscula que corresponde ao nome das espécies investigadas.

O valor do elemento $s_{i,j}$ é igual ao número de dados $c_{i,k} = c_{j,k}$, com $2 \leq$ i,j $\leq$ número de linhas da matriz característica analisadas e $1 \leq$ k $\leq$ número de colunas da matriz característica analisadas.

Por exemplo, o valor dos elementos da segunda linha da matriz similaridade da Tabela 2.5 obtida a partir da tabela 2.4, é o seguinte:

$s_{2,2} = (c_{2,1} = c_{2,1}) + (c_{2,2} = c_{2,2}) + (c_{2,3} = c_{2,3}) + (c_{2,4} = c_{2,4}) + (c_{2,5} = c_{2,5}) + (c_{2,6} = c_{2,6})$
$= |(0 = 0)| + |(0 = 0)| + |(0 = 0)| + |(0 = 0)| + |(0 = 0)| + |(0 = 0)|$
$= 6$
$s_{2,3} = (c_{2,1} = c_{3,1}) + (c_{2,2} = c_{3,2}) + (c_{2,3} = c_{3,3}) + (c_{2,4} = c_{3,4}) + (c_{2,5} = c_{3,5}) + (c_{2,6} = c_{3,6})$
$= (0{=}1) + |(0 = 0)| + |(0 = 0)| + |(0 = 0)| + |(0 = 0)| + |(0 = 0)|$
$= 5$
$s_{2,4} = (c_{2,1} = c_{4,1}) + (c_{2,2} = c_{4,2}) + (c_{2,3} = c_{4,3}) + (c_{2,4} = c_{4,4}) + (c_{2,5} = c_{4,5}) + (c_{2,6} = c_{4,6})$
$= (0{=}1) + (0{=}1) + (0{=}1) + |(0 = 0)| + (0 = 1) + |(0 = 0)|$
$= 2$
$s_{2,5} = (c_{2,1} = c_{5,1}) + (c_{2,2} = c_{5,2}) + (c_{2,3} = c_{5,3}) + (c_{2,4} = c_{5,4}) + (c_{2,5} = c_{5,5}) + (c_{2,6} = c_{5,6})$
$= (0{=}1) + (0{=}1) + (0{=}1) + (0{=}1) + |(0 = 0)| + (0 = 1)$
$= 1$

2.6 Matriz Distância

A matriz distância é uma matriz quadrada de dimensão n que relaciona duas espécies a uma medida de distância, onde n é igual ao número de espécies investigadas mais 1(uma).

Tabela 2.5: *Matriz similaridade.*

	A	B	C	D
A	6	5	2	1
B	5	6	3	2
C	2	3	6	3
D	1	2	3	6

Os elementos dessa matriz são obtidos a partir da análise dos elementos $c_{u,v}$ da matriz característica e são representados por $d_{i,j}$.

Os elementos da primeira linha e coluna contém uma letra romana maiúscula que corresponde ao nome das espécies investigadas. O valor do elemento $d_{i,j}$ é igual ao número de dados $c_{i,k} \neq c_{j,k}$, com 2 $\leq$ i,j $\leq$ número de linhas da matriz característica analisada e 1 $\leq$ k $\leq$ número de colunas da matriz característica analisada. Por exemplo, o valor dos elementos da segunda linha da matriz distância da tabela 2.6 obtida a partir da tabela 2.4, é o seguinte:

$d_{2,2} = (c_{2,1} \neq c_{2,1}) + (c_{2,2} \neq c_{2,2}) + (c_{2,3} \neq c_{2,3}) + (c_{2,4} \neq c_{2,4}) + (c_{2,5} \neq c_{2,5}) + (c_{2,6} \neq c_{2,6})$
$= (0\ 0) + (0\ 0) + (0\ 0) + (0\ 0) + (0\ 0) + (0\ 0)$
$= 0$
$d_{2,3} = (c_{2,1} \neq c_{3,1}) + (c_{2,2} \neq c_{3,2}) + (c_{2,3} \neq c_{3,3}) + (c_{2,4} \neq c_{3,4}) + (c_{2,5} \neq c_{3,5}) + (c_{2,6} \neq c_{3,6})$
$= (0\ 1) + (0\ 0) + (0\ 0) + (0\ 0) + (0\ 0) + (0\ 0)$
$= 1$
$d_{2,4} = (c_{2,1} \neq c_{4,1}) + (c_{2,2} \neq c_{4,2}) + (c_{2,3} \neq c_{4,3}) + (c_{2,4} \neq c_{4,4}) + (c_{2,5} \neq c_{4,5}) + (c_{2,6} \neq c_{4,6})$
$= (0\ 1) + (0\ 1) + (0\ 1) + (0\ 0) + (0\ 1) + (0\ 0)$
$= 4$
$d_{2,5} = (c_{2,1} \neq c_{5,1}) + (c_{2,2} \neq c_{5,2}) + (c_{2,3} \neq c_{5,3}) + (c_{2,4} \neq c_{5,4}) + (c_{2,5} \neq c_{5,5}) + (c_{2,6} \neq c_{5,6})$
$= (0\ 1) + (0\ 1) + (0\ 1) + (0\ 1) + (0\ 0) + (0\ 1)$
$= 5$

Tabela 2.6: *Matriz distancia.*

	A	B	C	D
A	0	1	4	5
B	1	0	3	5
C	4	3	0	3
D	5	4	3	0

2.7 Alguns Algoritmos de Construção de Árvores Filogenéticas

Nesta secção iremos apresentar três algoritmos usados para criação de árvores filogenéticas.

2.7.1 A Regra de Inclusão e Exclusão

A regra de inclusão e exclusão é um algoritmo de construção da árvore filogenética de um conjunto de espécies, que recebe a matriz característica dessas espécies e devolve o conjunto de todas as possíveis árvores filogenéticas dessas espécies. A seguir serão apresentados os passos desse algoritmo [48].

1^o passo - Para cada característica terminal N considerada no estudo, será construída uma árvore filogenética AFi (Fig.2.6a, b, c, d, e e f), com $1 \leq i \leq$ número de características terminais consideradas;

2^o passo - Combine todas as relações de parentescos presentes nas AF_i construídas no passo anterior, de modo que, as relações de parentesco presentes na AF_i sejam preservadas pela adição das relações de parentesco presentes na AF_j com i $<$ j (Fig.2.6g).

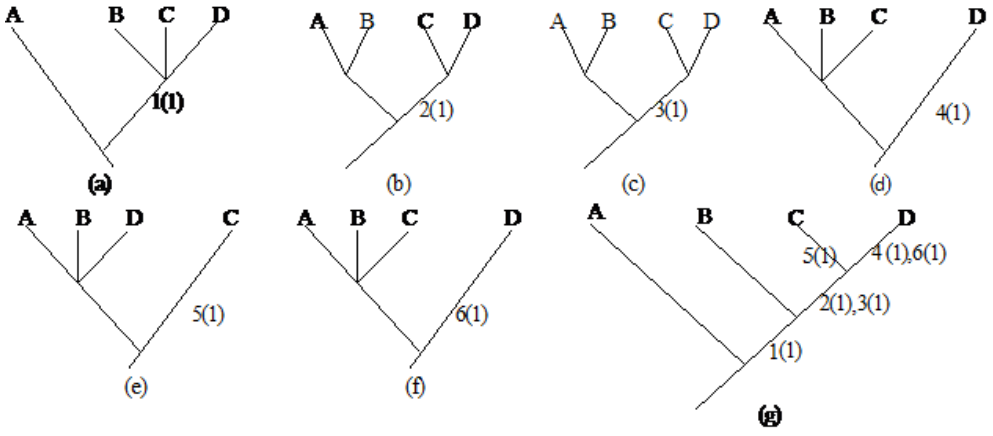


Figura 2.6: Árvore filogenética para as espécies da figura 4.6 obtida pela análise das características terminais: 1 (a), 2(b), 3(c), 4(d), 5(e) e 6(f). A combinação dessas árvores é mostrada em (g).

2.7.2 O Algoritmo de Wagner

O algoritmo de Wagner, é um algoritmo capaz de construir árvore filogenética de um conjunto de espécies, recebendo como entrada a matriz característica da espécies e retorna uma das árvores filogenéticas possíveis para esta espécies. Nesse sentido, iremos demonstrar os passos desse algoritmo segundo [2] e [48].

1^o passo - Especificar a espécie raiz;

2^o passo - Construir a matriz distância para as espécies da matriz característica fornecida;

3^o passo - Selecionar a espécie que tiver a menor distância para a espécie raiz. Esta será a espécie selecionada;

4^o passo - Criar um ramo ligando a espécie selecionada a espécie raiz com comprimento igual à distância entre essas duas espécies;

5^o passo - Selecionar a próxima espécie que apresente a menor distância para a espécie raiz, esta será a atual espécie selecionada;

6^o passo - Calcular a distância, usando a equação abaixo, entre a espécie selecionada e todas as outras espécies já selecionadas;

$$\text{Dist } (Esp_x, Esp_y) = \sum_{x=1}^{n} caract_x - \sum_y caract_y \quad (2.1)$$

onde $caract_i$ é o vetor característica ($c_{i,1}$,..., $c_{i,n}$) da espécie Esp_i, $c_{i,j}$ é o elemento da i linha e j coluna da matriz característica C de dimensão n.

7^o passo - Selecionar a espécie que apresenta a menor distância no passo anterior a atual espécie selecionada. Esta será a espécie irmã selecionada;

8^o passo - Criar um ramo ligando a atual espécie selecionada S ao meio do ramo que chega na espécie irmã selecionada I de comprimento Comp calculado por equação 2.7.2 (Fig. 2.7.2a,b,c);

$$\text{Comp(S,I)}=(Dist\,(S,I) + Dist\,(S, Ancestral\,(I))) - Dist\,(I, Ancestral\,(I))_{\frac{-}{2}} \quad (2.2)$$

9^o passo - Determine o vetor característica do ancestral comum A $Caract_A$ = ($c_{A,1}$,...,$c_{A,n}$) entre a atual espécie selecionada S e a espécie irmã selecionada I, onde $c_{A,i}$ = menor($c_{O,i}, c_{I,i}$) e $c_{x,y}$ é o elemento da x linha e y coluna da matriz característica C;

10^o passo- Enquanto existir uma espécie que ainda não foi selecionada volte ao passo 5.

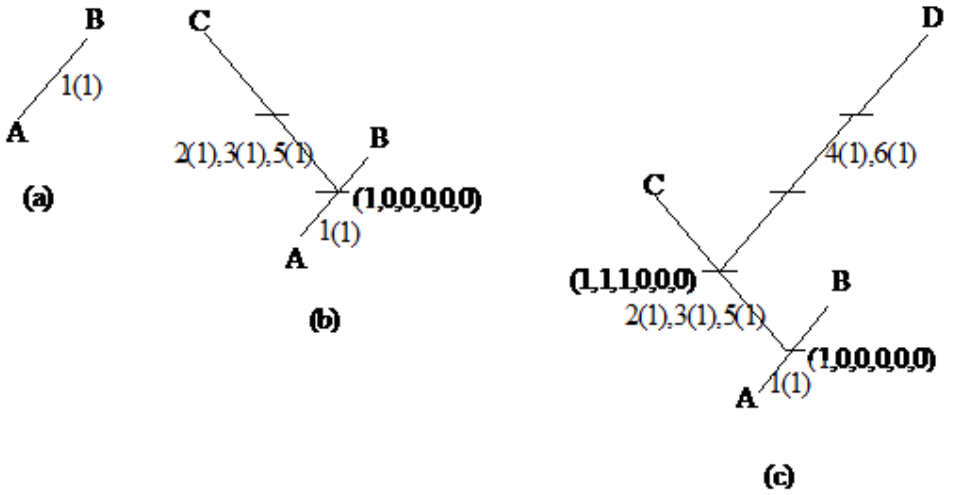


Figura 2.7: Árvores filogenéticas construídas pela primeira (a), segunda (b) e terceira (c) execução do 5o passo do algoritmo de Wagner para as espécies da figura 4.6.

2.7.3 O Algoritmo das Médias (UPGMA)

O algoritmo das médias é um algoritmo de construção da árvore filogenética de um conjunto de espécies que recebe a matriz distância dessas espécies e devolve uma das árvores filogenéticas possíveis para essas espécies. Note que os ramos da árvore filogenética construída por esse algoritmo, não contém rótulos. A seguir serão apresentado os passos deste algoritmo [36].

1^o passo - Tome o par de espécies com menor distância entre si e agrupe-os numa super-espécie. Este par de espécies terá um ancestral comum direto (Fig. 2.8a B e A, 2.8b C e S1 e 2.8c D e S2);

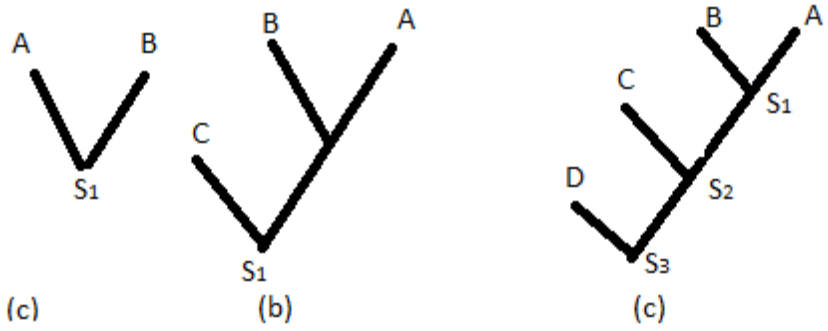


Figura 2.8: Árvores filogenéticas construídas pelo primeira (a), segunda (b) e terceira (c) execução do 5o passo do algoritmo das médias para as espécies da Tabela 2.7

2^o passo - Recalcule a distância de cada uma das demais espécies Si para a super-espécie recém criada como sendo a média das distâncias de Si para cada uma das espécies que constituem a super-espécie (Tabelas 2.7 e 2.8);

3^o passo - Repita os passos 1 e 2 enquanto houver duas ou mais (super-)espécies não

visitadas.

Tabela 2.7: *Matriz distância obtida a partir dos dados da Tabela 2.6 considerando que as espécies A e B tem a super-espécie S1 como ancestral direto.*

	S_1	C	D
S_1	0	2,5	5,5
C	2,5	0	3
D	5,5	3	0

Tabela 2.8: *Matriz distância obtida a partir dos dados da tabela 2.7 considerando que as espécies S_1 e C tem a super-espécie S_2 como ancestral direto*

	S_2	D
S_2	0	4,25
D	4,25	0

No próximo capítulo será apresentado o algoritmo genético baseado em tipos abstratos de dados para construção de árvore filogenética.

Capítulo 3

O Algoritmo Genético Baseado em Tipos Abstratos de Dados para Construção de Árvore Filogenética

3.1 Introdução

Algoritmos genéticos são uma família de modelos computacionais inspiradas na teoria de evolução das espécies de Charles [16]. Estes algoritmos modelam uma solução para um problema específico em uma estrutura de dados denominada de cromossomo e aplicam operadores que recombinam e selecionam essas estruturas preservando informações críticas.

O primeiro algoritmo genético foi proposto na década de setenta por um estudante de doutorado em ciência da computação da Universidade de Michigan chamado John H. Holland. Ele tentava desenvolver um método computacional voltado para a abordagem de fenômenos gerados por sistemas adaptativos complexos[1] e durante seu trabalho percebeu a evidente semelhança entre tais fenômenos e o processo de evolução das espécies; assim como a interação entre os agentes adaptativos determinava o resultado dos fenômenos por ele investigados, a interação entre os fatores ambientais determinava a próxima população de uma determinada região. Foi baseado nesta percepção que Holland propôs um modelo computacional para simular o processo de evolução das espécies, os algoritmos genéticos.

Algoritmos genéticos trabalham sobre uma população de possíveis soluções (cromossomos) para um problema, os quais são melhorados através de várias iterações com o intuito de gerar um cromossomo satisfatório para o problema. O mecanismo de transformação de uma população em outra obedece ao princípio de seleção natural, descrito por Darwin [16], segundo o qual, na natureza, os indivíduos mais adaptados têm maior chance de sobreviver.

[1] Fenômenos cujos resultados dependem das interações não lineares entre vários agentes adaptativos.

O algoritmo descrito na Figura 3.1 apresenta um pseudocódigo para algoritmos genéticos. Segundo este pseudocódigo, um algoritmo genético tem início com a geração da população inicial p composta por n cromossomos, que podem ser fornecidos pelo usuário, gerados aleatoriamente ou construídos por um algoritmo baseado no conhecimento existente sobre o problema a ser solucionado. Em seguida, é calculada a adaptação dos cromossomos da população, para verificar a qualidade destes cromossomos para o problema investigado.

Enquanto o critério de parada definido para o problema não for satisfeito, então um novo ciclo incrementa o contador de iterações t, seleciona cromossomos para gerar descendentes, gera descendentes, nascimento de alguns dos cromossomos gerados e morte dos cromossomos existentes na população dependendo do valor da adaptação desses cromossomos conforme explicitado no pseudocódigo da Figura 3.1.

Algoritmo 1 Pseudocódigo de um algoritmo genético

```
1: t ← 0
2: gera p
3: calcula o valor da adaptação dos cromossomos de p
4: while condição de parada não for satisfeita do
5:     t ← t + 1
6:     p̌ ← seleciona h cromossomos da população p
7:     ṗ ← aplica a operação de cruzamento e mutação sobre os cromossomos da população p̌
8:     p ← substitui l cromossomos da população p por l cromossomos da população ṗ
9:     calcula o valor da adaptação dos cromossomos de p
10: end while
```

Figura 3.1: Pseudocódigo de um algoritmo genético

A população dos cromossomos capazes de gerar descendente $\check{p}$ é construída pela seleção de h cromossomos da população p baseada na adaptação dos cromossomos ou em uma medida relativa ao mérito dos cromossomos dentro da população.

Os cromossomos da população $\check{p}$ são submetidos à ação dos operadores genéticos de cruzamento e mutação para gerar descendentes. Esses operadores genéticos, têm por objetivo, respectivamente: construir dois novos cromossomos a partir da combinação da cadeia de dois cromossomos da população $\check{p}$ e construir um novo cromossomo a partir da alteração da cadeia de um cromossomo da população $\check{p}$. A frequência com que os cromossomos da população $\check{p}$ geram descendentes é estabelecida por uma medida de probabilidade associada a cada operação, para facilitar o entendimento segue a Figura 3.2 que representa uma comparação entre os termos biológicos com os termos computacionais.

A nova população p é formada por l cromossomos da população $\dot{p}$ e por $n - l$ cromossomos da população p. O critério adotado para o nascimento dos l cromossomos da população $\dot{p}$ é o mesmo usado para a escolha dos cromossomos progenitores da população

Termos Biológicos	Termos Computacionais
Cromossomo	Indivíduo
Gene	Caractere ou atributo
Alelo	Valor do caractere
Lócus	Posição do caractere
Genótipo	Vetor de caracteres que representa o indivíduo
Fenótipo	Interpretação do vetor de caracteres

Figura 3.2: Representação dos Termo Biológico -> Termos computacionais [14]

p, o qual é aplicado no sentido inverso para determinar os l cromossomos da população p que devem morrer. Por fim, é calculado o valor da adaptação dos cromossomos da população p recém construída. A população é representada por um vetor, de tamanho fixo, de cromossomos, que por sua vez, são representados por vetores binários de tamanho fixo, ou seja, vetores de tamanho fixo de elementos pertencentes ao conjunto $\{0, 1\}$. Holland, chamou de *locu* as posições do cromossomo, *alelo* o valor que ocupa um *locu* no cromossomo e *gene* o conjunto de *alelos* que podem ocupar um *locu* no cromossomo.

O algoritmo proposto por Holland possui algumas limitações como a representação binária, convergência prematura,tamanho fixo dos cromossomos e das populações.

A questão da representação binária do cromossomo apresentada por Holland parece levar críticas sobre a pré-definição da estrutura de dados a ser adotada, no caso vetor de tamanho fixo, que está no sentido inverso do percurso natural adotado pelos engenheiros da computação na busca de uma solução computacional para um problema, onde primeiro se estuda o problema para só depois definir a estrutura de dados utilizada para modelar os objetos do domínio do problema [45, 46].

Outro problema relacionado com a representação proposta por Holland é a influência que o tamanho da população tem sobre a qualidade dos cromossomos da população retornada pelo algoritmo genético [19]. O primeiro registro deste problema foi reportado por Grefenstette (1986) ao usar um algoritmo genético para definir o valor ideal dos parâmetros do algoritmo de Holland na abordagem de um problema.

Apesar de todo o esforço em contornar os problemas apontados no trabalho de Holland, muitos destes problemas permanecem até o dia de hoje sem uma resposta, como o problema do tempo de convergência, da convergência prematura e da epistasis. O problema do tempo de convergência está relacionado ao tempo gastos por este algoritmo para produzir resultados de melhor qualidade do que outros métodos computacionais.

As soluções apresentadas para este problema objetivam explorar, ao máximo, as possibilidades de paralelizar e distribuir a execução do algoritmo, como nos algoritmos genéticos

paralelos [26], nos algoritmos genéticos de tempo contínuo [15] e nos hardware evolutivos [43]. Quando a complexidade está associada apenas ao cálculo da adaptação do cromossomo, uma versão simplificada da mesma pode ser a solução, como nos algoritmos genéticos baseados em técnicas de relaxamento [42].

A convergência prematura acontece quando as informações contidas nos cromossomos mais adaptados da população não são ótimas para o problema, os quais passam a predominar nas próximas populações geradas pelo algoritmo genético, direcionando assim este algoritmo para uma população de cromossomos ótimos locais.

A epistasis ocorre quando a adaptação de um cromossomo depende fortemente da interação entre as informações presentes em sua representação [19]. A dificuldade em tratar problemas com epistasis pode ser de natureza representacional e teórica. Dependendo do problema, o custo para desenvolver um novo projeto de representação para o cromossomo livre de epistasis pode ser muito grande [25]. Assim, uma alternativa para tratar este problema é adotar uma nova forma de representação para o cromossomo e desenvolver operadores genéticos de cruzamento e/ou mutação adequados [17].

Buscando solucionar os problemas do algoritmo de Holland foi criado o GAADT (algoritmo genético baseado em tipos abstratos de dados) [45, 46], denominado GAADT, do inglês *genetic algorithm based on abstract data types*. Os aspectos que o algoritmo contribui são principalmente a liberação desses algoritmos da restrição sobre a representação fixa do alfabeto 0,1 e do tamanho fixo da população, bem como a redução das operações genéticas a somente duas: cruzamento e mutação.

A adoção da abordagem de TAD, para construção de algoritmos genéticos permite, entre outras coisas, o desenvolvimento de uma sistemática de especificação do mesmo a partir do estudo dos componentes básicos do problema, para o qual o algoritmo genético a ser desenvolvido provê a solução. Outro importante ganho potencial com a visão acima referida é a portabilidade das soluções via algoritmo genético, uma vez que o mesmo esquema de solução pode resolver uma família de problemas, dependendo da natureza e do ambiente dos mesmos.

Neste capítulo será apresentado o GAADT, os tipos abstratos manipulados por este algoritmo, os operadores genéticos utilizados para promover a adaptação da população, o ambiente ao qual os cromossomos do GAADT devem se adaptar, além da definição do algoritmo propriamente dito e a especificação do GAADT para o problema.

3.2 Tipos Básicos do GAADT

No GAADT [45, 46], os cromossomos são representados por seu material genético, o qual têm nas bases suas unidades elementares de formação.

Definição 3.2.1 (Base) *Uma base B é o conjunto de todas as unidades genéticas elementares que podem ser usadas na formação do material genético dos cromossomos de uma população.*

Os elementos de uma base se agrupam em sequências para formar as características (genes) dos cromossomos, entretanto, nem toda sequência de bases representa uma característica para o cromossomo, portanto, deve existir uma *lei de formação* para indicar como as bases devem ser agrupadas para formar uma dada característica. O GAADT representa a *lei de formação* de características pelo conjunto de *Axiomas de Formação de Genes* (AFG), o qual deverá ser definido para cada caso de acordo com a semântica atribuída ao gene.

Definição 3.2.2 (Gene) *Um gene g é uma sequência formada pelos elementos da base que pertence ao conjunto AFG.*

Os genes são agrupados em conjuntos para formar os cromossomos da população. O conjunto de genes $\{g_1, g_2, \ldots, g_n\}$ que compõe um dado cromossomo serve para identificar este cromossomo dentro da população. A identidade dos cromossomos será usada para impedir que várias cópias de um cromossomo possam coexistir ou renascer na população em qualquer tempo durante o processo de evolução da mesma na busca por um cromossomo mais adaptado.

As características apresentadas pelos cromossomos de uma população servem também para classificá-los em grupos taxonômicos (espécies e famílias) em função do grau de similaridade das características compartilhadas pelos mesmos. O início e o fim da formação de uma característica em determinada espécie, no GAADT, é representado pelo conjunto de *Axiomas de Formação de Cromossomos* (AFC) que deve também garantir que o cruzamento de dois cromossomos de uma mesma espécie resulte em um cromossomo da mesma espécie. Assim como o AFG, o AFC também deverá ser definido para cada situação de acordo com a semântica adotada para o cromossomo.

Definição 3.2.3 (Cromossomo) *Um cromossomo c é um conjunto de genes que obedece às condições estabelecidas pelo* AFC.

Cromossomos são agrupados em conjuntos para formar uma população. Esta representação para a população irá garantir a imparcialidade na avaliação dos cromossomos que compõem a população, já que cada cromossomo só poderá ocorrer uma vez na população.

Definição 3.2.4 (População) *Uma população P é um conjunto de cromossomos construídos conforme descrito na Definição 3.2.3*

O tipo população é o conjunto formado por todos os conjuntos formados por objetos do tipo cromossomo que são possíveis resultados para o problema em foco, segundo a interpretação adotada para os tipos C, G e B. A geração da população vazia pelo algoritmo indica que a interpretação adotada para o problema está errada.

3.3 Operadores Genéticos do GAADT

O GAADT trabalha com dois tipos de operadores genéticos: o de cruzamento e o de mutação. O operador genético de cruzamento caracteriza-se por combinar os genes de dois cromossomos (cromossomos - pai) para formar outros cromossomos (cromossomos - filho), enquanto que o operador genético de mutação caracteriza-se por alterar a identidade de um cromossomo para formar um outro cromossomo (cromossomo-mutante).

O gene dos cromossomos - pai para uma dada característica que fará parte dos cromossomos - filho é aquele que melhor satisfaz as restrições do problema sobre a característica expressa por este gene, e recebe o nome de gene-dominante.

Dados dois genes g_1 e g_2, que expressem uma mesma característica com diferentes fenótipos[2], diz-se que um gene g_1 melhor satisfaz as restrições do problema do que o gene g_2, se o grau de adaptação do gene g_1 for superior ou igual ao grau de adaptação do gene g_2. No GAADT, o grau de adaptação de um gene é dado por uma função *grau* e é considerada a existência de um gene g_λ, denominado de gene-inócuo, que é usado para representar um gene que não expressa qualquer característica, de forma que a sua presença ou ausência não altera a identidade do cromossomo, o qual satisfaz as restrições impostas pelo conjunto de axiomas de formação de genes. Seu grau de adaptação é menor que o grau de adaptação de qualquer outro elemento do tipo gene.

Definição 3.3.1 (Grau) *O grau de adaptação de um gene é uma função* grau *do seguinte tipo:*
grau : $G \rightarrow K$ *tal que, a cada gene* g, $g \in G$, *é associado um único número* k, $k \in K$ *(*K *é um corpo ordenado*[3]*), chamado de grau*(g) *e que reflete, segundo a interpretação adotada para o problema, uma estratificação comparativa entre a adaptação dos genes.*

O gene inócuo [4] é uma constante do sistema, cujo valor deverá ser definido no momento da instanciação do algoritmo a um dado problema, pertence ao tipo gene, tem grau de adaptação igual ao elemento neutro do corpo K para a operação de adição, e para todo gene g com $g \neq g_\lambda$ $grau(g)$ é maior do que $grau(g_\lambda)$.

Por uma questão de nomenclatura, é definida uma relação de equivalência cromossômica, denotada por $\equiv_C$, definida da seguinte maneira: $\equiv_C$: $C \leftrightarrow C$, tal que $c_1 \equiv_C c_2$, se e somente se, $c_1 - \{g_\lambda\} = c_2 - \{g_\lambda\}$.

O gene dominante é identificado pela função *domi* que recebe um par de genes, um de cada um dos cromossomos - pai, e retorna o gene de maior grau de adaptação se os genes fornecidos expressarem uma mesma característica. Caso os genes fornecidos não expressem

[2] Características observáveis de um organismo que resultam da interação entre o genótipo e o ambiente.
[3] Estrutura algébrica, com duas operações, sem divisores próprios de zero e munido de uma ordem. Ex: $\langle \mathbb{R}, \leq, +, \times, 0, 1 \rangle$.
[4] O grau de adaptação é menor que o grau de adaptação de qualquer outro elemento do tipo gene. Tal gene será denominado de gene-inócuo.

uma mesma característica, então a função *domi* retornará g_λ. Dados dois genes, diz-se que eles expressam uma mesma característica se existe um atributo relevante para o problema em foco que seja satisfeito pelos genes fornecidos. No GAADT, o atributo relevante para o problema é especificado pelo conjunto *atributoRelevante*, que é uma constante do sistema cujo conteúdo dependerá da interpretação adotada para o problema. A comparação entre esses dois genes para constatar se expressam uma mesma característica, ou seja, se existe um *atributoRelevante* para o problema em questão que seja satisfeito pelos dois genes, é feita através da relação *mesma*, que obedece aos seguintes preceitos:

Lema 3.3.1 $\forall g : G \bullet (g, g) \in mesma.$

Lema 3.3.2 $\forall g_1, g_2 : G | (g_1, g_2) \in mesma \bullet (g_2, g_1) \in mesma.$

Lema 3.3.3 $\forall g_1, g_2, g_3 : G | (g_1, g_2) \in mesma \land (g_2, g_3) \in mesma \bullet (g_1, g_3) \in mesma.$

O gene dominante, portanto, é definido como:

Definição 3.3.2 (Dominante) *O gene dominante é uma função domi do seguinte tipo:*
$domi : G \times G \rightarrow G$

$$domi(g_1, g_2) = \begin{cases} g_\lambda & se\ (g_1, g_2) \notin mesma, \\ g_1 & se\ (g_1, g_2) \in mesma \land grau(g_1) \geq grau(g_2), \\ g_2 & se\ (g_1, g_2) \in mesma \land grau(g_1) < grau(g_2). \end{cases}$$

A produção de novos cromossomos durante o processo evolutivo de uma população serve para direcionar a busca por cromossomos mais adaptados através da transmissão das características de maior grau de adaptação presentes nos cromossomos da população atual. A adaptação de um cromossomo é dada pela função *adapt*.

Definição 3.3.3 (Adaptação) *A adaptação de um cromossomo é uma função adapt do seguinte tipo:*
$adapt : C \rightarrow K$
$adapt(c) = \sum_{g \in c} \Theta_{c,g} \times grau(g)$
onde $\Theta_{c,g}$ *é o peso com o qual o gene g contribui para a adaptação do cromossomo c.*

O valor do parâmetro peso usado no cálculo da adaptação de um cromossomo é definido de acordo com o problema em que o GAADT será aplicado, já que este cálculo pode levar em consideração a presença ou ausência de um dado gene no cromossomo.

A operação de cruzamento recebe dois cromossomos - pai, aptos a cruzar, e retorna uma população cujos cromossomos são formados somente pelos genes dominantes dos cromossomos fornecidos. Logo, para se definir esta função precisa-se antes definir uma função para selecionar os cromossomos aptos a cruzarem (seleção) e uma função para retornar o

conjunto de genes dominantes para todas as características existentes nos cromossomos - pai (fecundação).

A função de seleção recebe uma população P_1 e retorna a subpopulação de P_1 formada pelos cromossomos que satisfazem um requisito do problema r, descrito por uma fórmula em lógica de primeira ordem, o qual indica quando um dado cromossomo é considerado apto a cruzar.

Definição 3.3.4 (Seleção) *A seleção dos cromossomos que satisfazem um predicado r é uma função sel do seguinte tipo:*
$sel : \mathbb{P}(P) \times \mathbb{P}(P) \rightarrow \mathbb{P}(P)$
$sel(P_1, r) = P_1 \cap r.$

A função fecundação recebe dois cromossomos e retorna o conjunto de genes dominantes entre todos os genes dos cromossomos fornecidos.

Definição 3.3.5 (Fecundação) *A fecundação é uma função fec do seguinte tipo:*
$fec : C \times C \rightarrow \mathbb{P}(G)$
$fec(c_1, c_2) = \{g | \forall g_1 \in c_1 \, \forall g_2 \in c_2 (g = domi(g_1, g_2))\}$

Lema 3.3.4 $\forall c : C \bullet (fec(c, c), c) \in \equiv_C$.

Os cromossomos - pai aptos a cruzarem são representados pelo conjunto $MACHO$ e $FEMEA$, formados da seguinte forma: $MACHO = sel(P_1, M)$ e $FEMEA = sel(P_1, F)$, onde P_1 é uma subpopulação da população atual formada por cromossomos adaptados ao ambiente e, M e F são dois predicados sobre o tipo população pertencentes ao conjunto de requisitos do ambiente Rq, escritos em uma linguagem de primeira ordem.

Dependendo da especificação dos requisitos do ambiente M e F, a reprodução gerada pode ser sexuada, assexuada ou mista. A reprodução sexuada ocorre quando $M \cap F = \emptyset$, a assexuada quando $M = F$, e a mista quando $M \cap F \neq \emptyset$ e $M \neq F$.

Definição 3.3.6 (Cruzamento) *O cruzamento é uma função* cruz *do seguinte tipo:*
$cruz : MACHO \times FEMEA \rightarrow \mathbb{P}$
$cruz(c_1, c_2) = \{c | c \subseteq fec(c_1, c_2)\}$

O operador genético de mutação, definido para o GAADT, é composto pelas funções de inserção, supressão e troca, tal que os cromossomos resultantes da ação destes operadores apresentarão parte dos genes contidos no cromossomo que lhe deu origem. A operação de inserção *ins* adiciona um conjunto de genes ao cromossomo de origem.

Definição 3.3.7 (Inserção) *A inserção é uma função ins do seguinte tipo:*
$ins : C \times \mathbb{P}(G) \rightarrow C$
$$ins(c, G_1) = \begin{cases} c \cup G_1 & \text{se } c \cup G_1 \in AFC, \\ c & \text{caso contrário.} \end{cases}$$

A operação de supressão *del* remove um conjunto de genes do cromossomo de origem.

Definição 3.3.8 (Supressão) *A supressão é uma função del do seguinte tipo:*
$del : C \times \mathbb{P}(G) \rightarrow C$
$$del(c, G_1) = \begin{cases} c - G_1 & se\ c - G_1 \in AFC, \\ c & caso\ contrário. \end{cases}$$

A operação de troca *troc* remove um conjunto de genes do cromossomo de origem e lhe adiciona outro conjunto de genes.

Definição 3.3.9 (Troca) *A troca é uma função troc do seguinte tipo:*
$troc : C \times \mathbb{P}(G) \times \mathbb{P}(G) \rightarrow C$
$$troc(c, G_1, G_2) = \begin{cases} (c \cup G_1) - G_2 & se\ (c \cup G_1) \in AFC \wedge (c \cup G_1) - G_2 \in AFC, \\ c \cup G_1 & se\ (c \cup G_1) \in AFC \wedge (c \cup G_1) - G_2 \notin AFC, \\ c - G_2 & se\ (c \cup G_1) \notin AFC \wedge c - G_2 \in AFC, \\ c & se\ (c \cup G_1) \notin AFC \wedge c - G_2 \notin AFC. \end{cases}$$

Ações da função de inserção e supressão podem ser vistas como casos particulares da ação da função de troca. Conforme estabelece o seguinte corolário:

Corolário 3.3.1 $\forall c : C; G_1, G_2 : \mathbb{P}(G) \bullet troc(c, G_1, G_2) = del(ins(c, G_1), G_2)$

Definição 3.3.10 (Mutação) *A mutação é um predicado* $mut \subseteq \mathbb{P}(P)$*, tal que:*
$mut(c_1) = \{c_2 | \exists G_1, G_2 : \mathbb{P}(G)\ ((\#G_1 \leq \#c_1 \div 2) \wedge (\#G_2 \leq \#c_1 \div 2) \wedge (troc(c_1, G_1, G_2) = c_2)$
$\wedge (adapt(c_2), adapt(c_1))\}$

A restrição $(adapt(c_2), adapt(c_1))$ garante que todo cromossomo-mutante é mais adaptado do que o cromossomo que lhe deu origem. A alteração do número de genes do cromossomo no cromossomo-mutante é limitada em cinquenta por cento do tamanho do cromossomo fornecido devido ao fato de que se as mutações ocorridas em um cromossomo de uma dada espécie forem muito grandes, então este cromossomo seria repelido pelos cromossomos da sua espécie, por não ser considerado mais um igual a estes.

3.4 Ambiente

O GAADT opera sobre populações de cromossomos que evoluem de acordo com as características de um ambiente A.
Um ambiente A é uma 8-tupla $\langle P, \mathbb{P}(P), Rq, AFG, AFC, Tx, \Sigma, P_0 \rangle$, onde:

- P é a população,
- $\mathbb{P}(P)$ é o conjunto potência de P,

- Rq é o conjunto dos requisitos (características expressas através de fórmulas numa linguagem de primeira ordem) do problema que influenciam a genealogia da população P,
- AFG é o conjunto de axiomas de formação dos genes dos cromossomos da população P,
- AFC é o conjunto de axiomas de formação dos cromossomos da população P e
- Tx é o conjunto de pares de cromossomos (x, y), onde x é um cromossomo construído a partir do cromossomo y, pela ação da operação de cruzamento ou mutação, registrando desta forma a genealogia dos cromossomos pertencentes às populações geradas pelo GAADT durante a sua execução,
- Σ é o conjunto de operadores genealógicos que atuam sobre a população P,
- P_0 é uma sub-população pertencente a $\mathbb{P}(P)$, chamada de população inicial, com no mínimo um cromossomo.

O processo de evolução darwinista, segundo o qual todas as espécies desenvolveram-se a partir de outras espécies, pela transmissão hereditária de pequenas variações, em sucessivas gerações, resultando na sobrevivência das espécies que melhor adaptaram-se ao ambiente [16], é induzido pelas alterações ambientais produzidas pela natureza. Este papel desempenhado pela natureza, na visão de evolução darwinista, aqui será representado pelo GAADT, que é quem submete os cromossomos de uma população à ação dos requisitos do problema Rq, resultando assim na geração de novos cromossomos a partir daqueles já existentes [45, 46].

3.5 Algoritmo

O GAADT é uma função que recebe a população P_0 e, depois de submetê-la à simulação de um processo evolutivo, devolve uma população P_t. Os cromossomos da população P_t são os cromossomos das populações P_0,P_1,...,P_{t-1} que ainda satisfazem os requisitos do problema Rq, ou então são novos cromossomos resultantes da ação genealógica das operações de cruzamento e mutação sobre os cromossomos da população P_{t-1} que apresentam adaptação maior do que a adaptação dos cromossomos que lhes deram origem. Diz-se então que a população P_t evoluiu da população P_0.

Os cromossomos das populações P_0,P_1,...,P_{t-1} que não mais satisfaçam os requisitos do problema Rq não participarão da construção da população P_t, podendo ser assim entendidos como fazendo parte da população de cromossomos "mortos", que não figurarão entre os cromossomos da população P_t e das populações seguintes manipuladas pela função

$GAADT$. Não obstante, tais cromossomos serão recuperados pela análise da taxonomia Tx dos cromossomos da população atual para evitar que eles apareçam novamente nas próximas iterações da função $GAADT$. Esta restrição atende ao entendimento do processo de evolução darwinista, que não contempla a possibilidade de uma espécie extinta voltar a aparecer num outro momento futuro.

Na definição da função $GAADT$, um critério de preservação sobre a população atual P_t é estabelecido para orientar o corte dos cromossomos que não devem figurar nas populações P_{t+1},P_{t+2},.... Este ponto de corte é representado por um predicado unário p_{corte} pertencente ao conjunto de requisitos do problema Rq, que atua sobre os cromossomos de P_t.

Os critérios de parada adotados pela função $GAADT$ são: o número máximo de iterações desejadas; e o valor da adaptação dos cromossomos considerado satisfatório para a solução do problema em análise. Estes critérios também fazem parte do conjunto de requisitosRq do problema.

Definição 3.5.1 (GAADT) *O $GAADT$ é uma função $GAADT$ do seguinte tipo:*
$GAADT : A \rightarrow A$

$$GAADT(P_t) = \begin{cases} P_{otm} & se\ P_{otm} = \{c | \forall c : P_t\,(adapt(c) \geq k)\} \neq \emptyset, \\ P_{t+1} & se\ t+2 = T, \\ GAADT(P_{t+1}) & caso\ contrário. \end{cases}$$

onde $P_{t+1} = cruz(a,b) \cup mut(c) \cup p_{corte}(P_t)$ com e $a,b,c \in P_t$, P_0 é a população inicial considerada, $k \in K$ é um valor imposto pelo ambiente A, como critério de aceitação de cromossomos em P_t que satisfazem o problema e $T \in \mathbb{N}$ é um número dado como critério de satisfação do número de iterações.

A segunda condição de parada ocorrerá garantidamente se eventualmente a primeira não ocorrer. Para qualquer entrada P_t, o processo $GAADT$ dá uma saída bem determinada, o que implica dizer que o $GAADT$ é um algoritmo e desta forma um procedimento correto.

3.6 Especificação do GAADT para o Problema

O GAADT foi proposto com o objetivo de ser aplicado a diferentes problemas, a qualidade dos resultado encontrado depende da instanciação dos tipos abstratos manipulado por este algoritmo, da adequação dos operadores genéticos ticos e das restrições impostas pelo ambiente. Portanto, o GAADT, quando instanciado para um problema particular, deve preservar todas as suas propriedades fundamentais. Em Vieira [45, 46] apresentada uma metodologia que auxilie a modelar os requisitos do problema, baseada em 7 passos:

1. Examinar o problema, com o objetivo de determinar qual a natureza da solução

requerida no sentido de orientar a escolha metafórica dos tipos básicos: base, gene e cromossomo;

2. Identificar o gene inócuo g_λ e a população inicial P_0;

3. Definir as funções grau de adaptação do gene grau e peso do gene no cromossomo $\Theta, earelaoatributoRelevante; Definiroselementosde Rq(F, Mer) usados pela funo cruz de acordo$

4. Definir os elementos de Rq (pcorte, t e k) utilizados pela função GAADT;

5. Construir o algoritmo;

6. Verificar os eventuais ajustes relativos para verificação dos tipos manipulados pelas funções, relações e operações manipuladas pela função do GAADT, e a modelagem do problema através da definição dos componentes do ambiente.

O restante desta seção se destina a descrever como o GAADT pode ser aplicado e refinado ao problema do melhoramento genético da cana de açúcarda.

Segue abaixo os elementos básicos do GAADT aplicado ao problema de interesse:

3.6.1 Tipos Básicos

Neste trabalho, os cromossomos serão representados por seu material genético, o qual têm nas bases suas unidades elementares de formação. A única exigência para o tipo abstrato base do GAADT é que ele tenha pelo menos uma base b_λ, denominado de base-inócuo, que será usado na construção do gene-inócuo g_λ e do cromossomo-inócuo c_λ. As constantes do GAADT base-inócua, gene-inócuo e cromossomo-inócuo deverão ter seu valor definido no momento da instanciação do GAADT a um dado problema.

Os indivíduos serão representados por seu material genético (cromossomos), o qual tem nas bases nucléicas suas unidades elementares de formação. O tipo abstrato bases para o problema construção de árvores filogenéticas é o conjunto formado pelo nome das espécies investigadas acrescido de uma espécie denominada de inócua, que será usada para formar o gene inócuo.

Definição 3.6.1 (Base) *O conjunto base $B = X \cup \{e_\lambda\}$, onde e_λ é denominado de gene-inócuo e X é o conjunto de espécies analisado.*

Onde cada instância representa uma especie e suas característica, a partir desta *base* é gerada uma **matriz característica**. Além disso, o sistema gera um ***vetor de bases*** com o grau de similaridade entre as instâncias da base.

Definição 3.6.2 (Gene) *O conjunto gene $G = \{< e_y, e_x, e_z > / e_y, e_x, e_z \in B\}$ que satisfaz o seguinte conjunto de axiomas de formações de genes:*

- $\forall_x \forall_y \forall_z$ raiz $< e_y, e_x, e_z >= e_x \Leftrightarrow e_x \neq e_\lambda$ para todo gene se a base da raiz for a base e_λ , então a base nas folhas também será a base e_λ;
- $\forall_x \forall_y \forall_z$ folha$_L$<e$_y, e_x, e_z >= e_y \Leftrightarrow e_y \neq e_\lambda$ para todo gene se a base da raiz for diferente da base e_λ , então a base nas folhas deve ser diferente da base da raiz e_x;
- $\forall_y \forall_x \forall_z$ folha$_R$<e$_y, e_x, e_z >= e_z \Leftrightarrow e_z \neq e_\lambda$ para todo gene se a base da raiz e folhas forem diferentes da base e_λ , então elas devem ser todas diferentes.
- $\forall_x G(e_x) = \{g/g \in G\}$

A interpretação dada para o tipo abstrato gene $g =< e_y, e_x, e_z >$ é que ele representa por uma de árvores simples (Figura 4.7), com raiz em e_x, folha à esquerda e_y e folha à direita e_z, os quais satisfazem o seguinte conjunto de axiomas de formação de genes:

O gene inócuo neste caso é o gene $<e_\lambda$,e_λ ,$e_\lambda >$. Por exemplo, os genes da árvore filogenética da figura 4.7 são (e_6, e_5, e_7), (e_4, e_3, e_5) e (e_2, e_1, e_3).

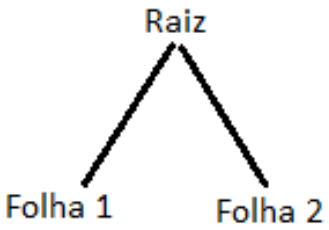


Figura 3.3: Representação de um Gene

Definição 3.6.3 (Cromossomo) *O conjunto cromossomo* $C = \{c = \{g_1, g_2, ..., g_j\} / \{g_1, g_2, ..., g_j\} \subseteq G$ *que satisfaz os seguintes axiomas de formação de cromossomo:*

- $acf_1 \equiv (\forall c \in C(\{x/\forall g_i \in C(x = raiz(g_i))\} = (\{x \in X/\forall g_i \in C(x = folha_L(g_i))\} \cup \{x \in X/\forall g_i \in C(x = folha_R(g_i))\}) + 1)$;
- $acf_2 \equiv ((\{x \in X/\forall g_i \in C(x = folha_L(g_i))\} \cap \{x \in X/\forall g_i \in C(x = folha_R(g_i))\}) = \oslash)$;
- $acf_3 \equiv ((\{x \in X/\forall g_i \in C(x = raiz(g_i))\} \cap \{x \in X/\forall g_i \in C(x = folha_L(g_i))\} \cup \{x \in X/\forall g_i \in C(x = folha_R(g_i))\}) = 1)$.

O cromossomo aqui é visto como o conjunto das árvores elementares, o qual atende as exigências imposta pelos axiomas de formação de cromossomos listados abaixo:

- para todo cromossomo o número máximo de ocorrências dos elementos do tipo base como raiz é 1;
- para todo cromossomo o número máximo de ocorrências dos elementos do tipo base como folha é 1; e
- para todo cromossomo o número de bases cuja sua única ocorrência é como raiz é 1.

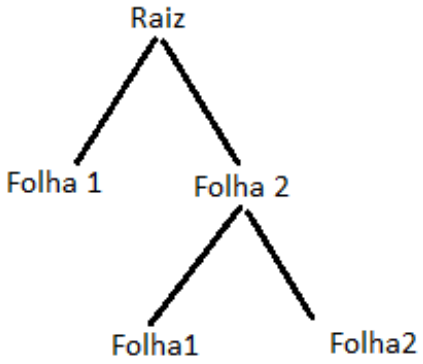


Figura 3.4: Representação de um Cromossomo

Um *conjunto de genes* contendo todos os possíveis genes que poderão compor os cromossomos (o cromossomo é um conjunto de genes, que não se repetem) de determinada população.

3.6.2 Operadores Genéticos

Assim como a instanciação dos tipos abstratos dados: base, gene, cromossomo e população, preservaram as exigências contidas na definição do GAADT. A instanciação da funções e relações necessárias para o calculo da função GAADT devem atende a todos as pré-condições de sua definição original. Nesta seção serão apresentado a definição somente das funções e relações cuja instanciação para o presente problema mais concreta do que a definição original, ficando subtendido que as funções e relações que forem re-definidas nesta seção e preservada sua definição original.

Definição 3.6.4 (Grau) *O grau de adaptação de um gene é uma função grau do seguinte tipo:*

grau: $P \rightarrow K$ tal que grau$(\langle e_y, e_x, e_z \rangle) = \sum |m_{e_y,i} + m_{e_z i} - 2m_{e_x i}|$ onde K é um corpo ordenado [5].

[5]Um corpo ordenado é uma estrutura algébrica, com duas operações, sem divisores próprio de zero e munido de uma ordem. $Ex :< R, \leq, +, \mathrm{x}, 0, 1 >$

Por exemplo, o grau de adaptação do gene (e_2, e_1, e_3) é:

$grau\,(\rangle\, e_2, e_1, e_3\,\langle) = |(m_{e_2,1} + m_{e_3,1} - 2m_{e_1,1})| + ... + |(m_{e_2,7} + m_{e_3,7-2}m_{e_1,7})| = |(1 + 0 - 2\text{x}1)| + |(1 + 1 - 2\text{x}0)| + |(1 + 1 - 2\text{x}1)| + |(0 + 1 - 2\text{x}0)| + |(0 + 0 - 2\text{x}1)| + |(0 + 0 - 2\text{x}0)| + |(0 + 0 - 2\text{x}0)| = 6$

Definição 3.6.5 (Dominante) *O gene-dominante, ou simplesmente domi, é uma função do seguinte tipo:*

$domi : G \times G \rightarrow G$

$$domi(g, \widehat{g}) = \begin{cases} g_\lambda & \text{se } \neg(raiz\,(g) = raiz\,(\widehat{g})) \\ g & \text{se } (raiz\,(g) = raiz\,(\widehat{g}) \wedge (grau((g) \leq grau(\widehat{g})) \\ \widehat{g} & \text{se } (raiz(g) = raiz(\widehat{g}) \wedge (grau(\widehat{g}) \leq grau(g)). \end{cases}$$

Por convenção, o grau de adaptação do gene-inócuo igual ao elemento neutro do corpo K para a operação de adição, logo o grau de adaptação do gene-inócuo menor do que o grau de adaptação dos outros genes pertencentes ao tipo abstrato G.

O gene dominante para uma dada característica do gene que corresponde a expressão fenotípica desta característica mais adaptada a configuração atual do ambiente. A função domi foi definida para informar o gene dominante de um par de gene, se os genes fornecidos não expressem uma mesma característica a função domi retornando o g_λ, caso contrário a função domi retorna o gene fornecido mais adaptado o ambiente atual.

A produção de novos cromossomos durante o processo evolutivo de uma população serve para direcionar a busca por cromossomos mais adaptados através da transmissão das características de maior grau de adaptação presentes nos cromossomos da população atual. A adaptação de um cromossomo é dada pela função adapt.

Definição 3.6.6 (Adaptação) *O grau de adaptação de um cromossomo é uma função adpt do seguinte tipo:*

$adapt : P \rightarrow K$

$adapt(c) = \sum_{g \in c} grau(g)$

Capítulo 4

O sistema proposto para construção de Árvore Filogenética(SETAX)

4.1 Arquitetura

Normalmente a construção de uma árvore filogenética é realizada por um taxonomista e um geneticista que a constroem manualmente. O SETAX auxilia nesta tarefa evitando possíveis erros na hora da construção e no tratamento de dados ambíguos e incompletos. Ele também permite manipular dados com diferentes representações, gerenciar explosões combinatórias de todas as possíveis soluções geradas pela presença parcimônica[1].

Desta forma, este trabalho apresenta um sistema computacional para criação de árvores filogenéticas a partir da entrada de uma matriz característica das espécies, esse sistema foi desenvolvido utilizando a linguagem de programa JAVA com a plataforma de desenvolvimento NetBeans IDE 6.7.1. O reconhecimento dessa base de dados séra feita através da interface do sistema e a criação das possíveis árvores candidatas séra feita através do GAADT (Vieira, 2003), conforme ilustra a Figura 4.1.

- **Usuário:** O usuário escolhe qual a base de dados com as características das especies e a serem avaliada e o sistema ler essa base e verifica se a base está consistente.
- **Interface:** A interface verifica a base de características e gera a matriz similaridade das espécies investigada, logo após ela cria os tipos e operadores genéticos utilizados no GAADT.
- **GAADT:** Utilizado para manipular e arranjar as árvores criadas na busca da melhor árvore gerada que atenda a determinada requisição de acordo com critérios

[1] Conceito utilizado na sistemática moderna que estabelece que ao construir e selecionar árvores filogenéticas, ou seja, os clados, o melhor critério é baseado em seus princípios: normalmente é correto o relacionamento mais simples encontrado entre dois indivíduos, aquele que apresente o menor números de passos intermediários ou mudanças evolucionárias.

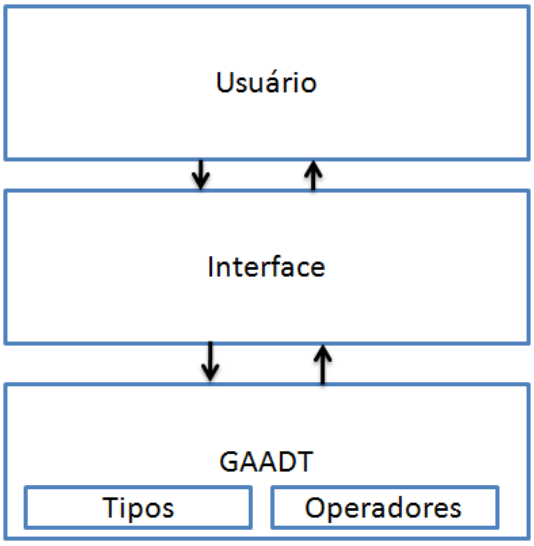


Figura 4.1: Arquitetura utilizada na solução implementada.

pré-estabelecidos. Seus tipos e operadores específicos atuam nos serviços da camada anterior de forma a orquestrar os mesmos para criação das árvores filogenéticas. Para isso ele necessita de seus elementos básicos adaptado ao o problema.

Além disso, é necessário ressaltar que este sistema foi projetado para auxiliar e orientar o usuário durante o processo de execução e sempre que um dado for necessário ou um parâmetro for indispensável e não estiver nos valores iniciais, o usuário será solicitado a fornecer este dado ou informação faltante. Por exemplo, o nome do arquivo que contém o conjunto de espécies a serem analisadas não pode fazer parte do conjunto de valores iniciais, razão pela qual ele é solicitado sempre que o usuário ordenar o início de execução. Uma tela específica de busca de arquivos no padrão Windows será exibida para que o usuário aponte para o arquivo desejado. Desta forma, os recursos e as facilidades disponíveis nesta ferramenta ajudaram muito o desenvolvimento deste trabalho e poderão ser úteis também para novas pesquisas sobre Filogenia e sobre Computação Evolutiva.

As seguintes características não foram implementadas nesta versão do SETAX:

- escolher algum outro método de busca;
- melhorar a perfomace;
- comparar o resultados mais detalhadamente;
- oferecer explicações de como o sistema chegou ao resultado.

4.2 Funcionalidades do SETAX

O SETAX auxilia na tarefa de construção de árvores filogenéticas evitando possíveis erros na hora da construção, pois fornece explicações para decisões tomadas pelo sistema, no

tratamento de dados ambíguos e incompletos. Ele também permite manipular dados com diferentes representações, gerenciar explosões combinatórias de todas as possíveis soluções geradas pela presença parcimônica.

O sistema possui interface gráfica que foi projetada para facilitar sua execução pelos usuários, mesmo que estes não dominem completamente as técnicas de computação e permite ajuste dos principais parâmetros relativos ao algoritmo genético utilizado, tais como tamanho da população inicial, porcentagem de mutações, porcentagem de cruzamento e taxa de seleção.

Inicialmente o usuário executará a instrução de inicialização do SETAX. Após a execução dessa instrução, o módulo simbólico do SETAX apresentará a tela principal mostrada na Fig.4.2. Acessando as opções desta tela, o usuário poderá fornecer os dados sobre as espécies investigadas ("Seleciona Base") Figura 4.3, gerar uma árvore filogenética para essas espécies ("Gerar Árvore"), mostrar quem criou o sistema ("Sobre") e pedir ajuda para entender o funcionamento do sistema ("Ajuda"). Inicialmente, usando a opção "Seleciona Base", o usuário fornecerá a base que ele pretende gerar a árvore, acessando a opção "Gerar Árvore"do menu pull-down da opção(Figura. 4.3) e exibe a melhor solução.

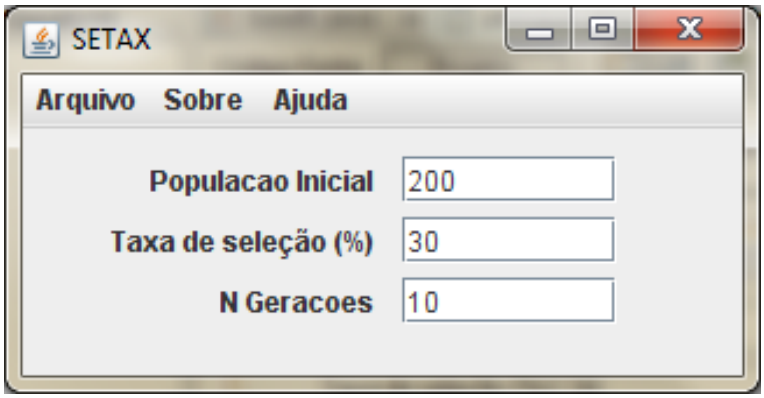


Figura 4.2: Tela principal do sistema.

Para abrir o arquivo que contem a base fornecida pelo usuário o SETAX solicita o arquivo que conterá essa base como na Figura 4.3 .

Se o arquivo não for selecionado ou não for compatível, SETAX mostrará uma janela de alerta informando para o usuário que o arquivo não existe e que o sistema não pode abri-lo vide Fig.4.4.

De posse da base como na Figura 4.5, o usuário pode solicitar ao sistema que gere a árvore filogenética vide Fig.4.3 "Gera Árvore", neste momento, os dados fornecido são preparados para serem enviados ao algoritmo descrito no capítulo 3 e o sistema gera a árvore como na Figura 4.6.

Conforme a demostração o GAADT é responsável pela predição das possíveis árvores filogenéticas, para tal, recebe como entrada uma base de dados representada pelas característica da espécie e fornece como saída as árvores.

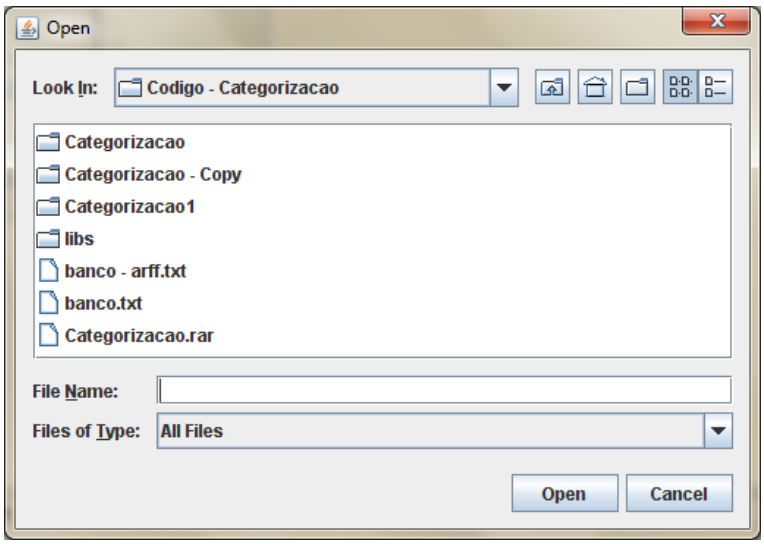


Figura 4.3: Tela principal com o menu do sistema.

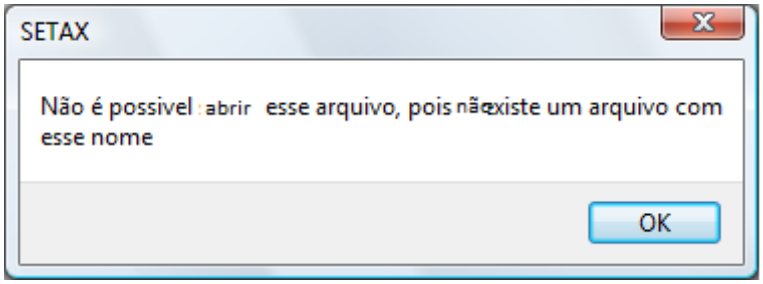


Figura 4.4: Tela de alerta.

	Genótipo	Prog. Fem	Prog. Masc.	BRI	MAT	TCH	FIB	BRO	PER	DES	HAB	DIA	FLO	CHO	CHO F
1	RB70141	Co331	?	3	1	2	2	1	1	1	2	2	1	1	4
2	RB72199	NCo334	?	3	1	1	2	2	1	1	3	1	1	1	1
3	RB72454	CP53-76	?	3	3	3	2	1	1	2	2	2	1	1	4
4	RB72910	NI		1	3	3	3	3	3	3	3	1	1	1	1
5	RB75126	C278	?	3	3	3	2	2	1	2	3	3	4	4	4
6	RB751194	CB40-49	?	1	3	3	2	2	2	2	3	2	3	4	4
7	RB8060	F147	?	3	1	2	1	1	2	1	2	2	2	2	4
8	RB80457	IAC48/65	RB705096	3	2	2	2	2	2	1	3	2	4	4	4
9	RB81145	NA56-79	Co331	3	2	3	2	2	1	2	3	2	1	2	4
10	RB83342	NA56-79	SP70-1143	3	1	2	2	2	2	1	3	2	3	2	4
11	RB83102	NA56-79	SP70-1143	3	2	3	2	2	2	2	3	2	1	2	4
12	RB83160	NA56-79	SP70-1143	3	1	2	2	3	2	2	3	1	1	3	4
13	RB83594	RB72454	B3337	3	3	3	2	1	2	2	2	2	4	4	4
14	RB842021	B3337	RB72454	3	3	2	2	1	1	2	2	2	4	4	4
15	RB91514	RB7893	?	3	3	2	2	1	1	2	3	2	2	4	4
16	RB91524	Co6806	RB72454	3	1	2	2	1	1	2	3	2	2	3	4
17	RB91539	H59-9018	?	2	3	2	2	1	1	3	2	3	2	4	4
18	RB92508	RB7831	?	3	2	2	2	1	1	2	3	2	2	4	4
19	RB92533	RB72454	?	3	1	3	2	2	1	1	2	1	1	1	4
20	RB92579	RB75126	RB72199	2	2	3	2	3	3	2	3	2	3	3	4
21	RB92606	Q107	RB72454	3	2	2	1	1	1	2	3	3	2	4	4
22	RB9350	Q107	RB72199	3	1	3	2	2	2	1	3	2	4	4	4
23	RB9364	RB72454	RB83102	2	3	3	2	1	1	2	3	2	4	4	4
24	RB9367	RB72454	RB83102	3	1	2	2	1	1	1	3	2	4	4	4

Figura 4.5: Amostra de uma base (Cana de açúcar) do estudo de caso.

A sequência de dados biológicos são representadas no computador por *instâncias da base de dados*. Ao utilizar o sistema proposto neste trabalho, o usuário deve informar como entrada o arquivo da base formatado (banco.csv), ou seja, as instâncias da base, representando a característica da espécie, para a qual deseja obter suas possíveis árvores.

O GAADT recebe como entrada a base, ou *instâncias da base de dados*, onde cada

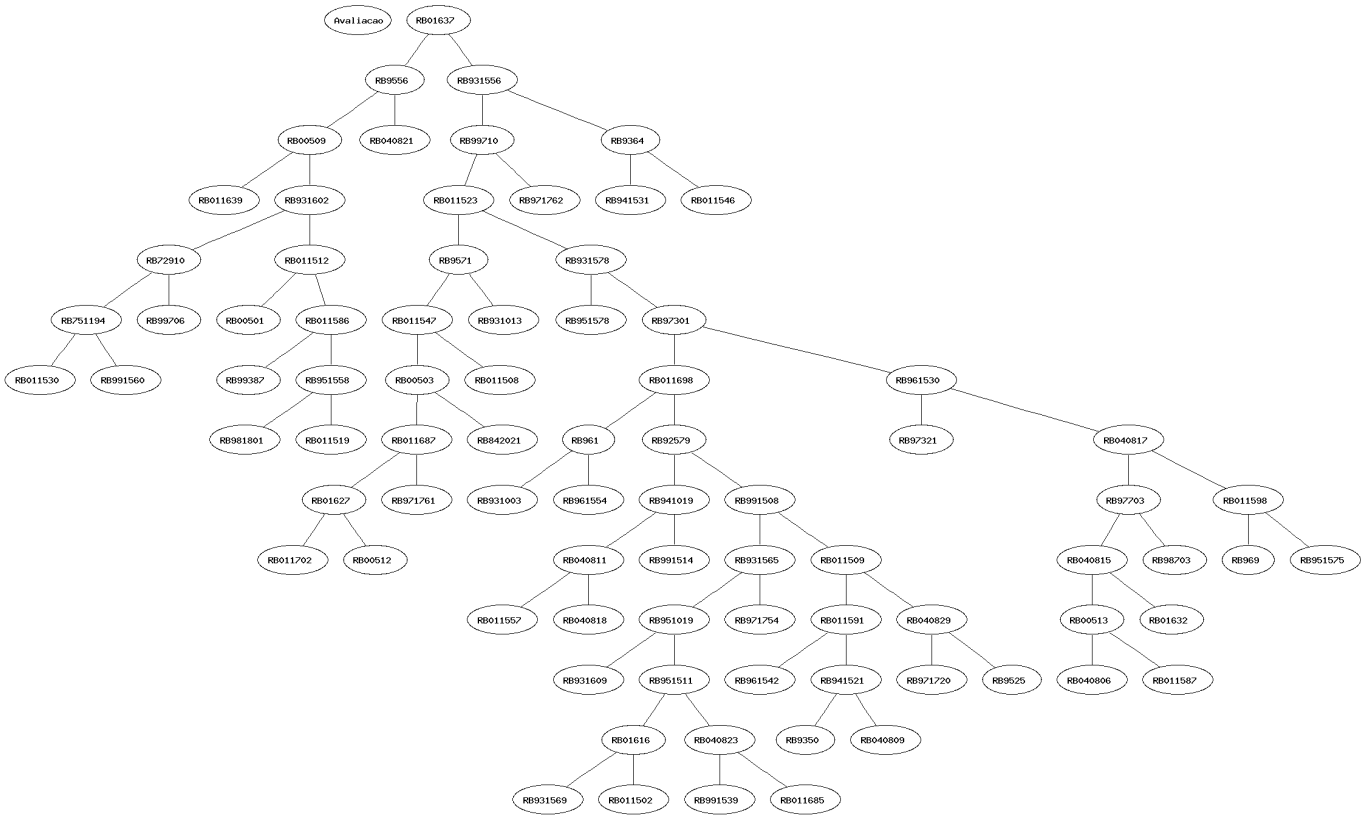


Figura 4.6: Exemplo de árvore gerada pelo sistema através da base de dados (Cana de açúcar) do estudo de caso.

instância representa uma especie e suas característica, a partir desta *base de dados* de entrada o SETAX gera uma **matriz característica** como na Figura 4.5. Além disso, o sistema gera um ***vetor de bases*** com o grau de similaridade entre as instâncias da base. Feito isso, o sistema gera a *população inicial* que partir do vetor de base um números de genes são gerados aleatoriamente respeitando o axioma de formação dos genes como demonstrado no capítulo 3.

Cada gene pode assumir uma possíveis estruturas definidas neste trabalho, conforme ilustra a Figura 4.7. Junto à geração da *população inicial*, é gerado um *conjunto de genes* contendo todos os possíveis genes que poderão compor os cromossomos (o cromossomo é um conjunto de genes, que não se repetem) de determinada população. Todos os genes dos cromossomos de qualquer população fazem parte do referido *conjunto de genes*. Na *população inicial*, cada cromossomo é constituído de genes do *conjunto de genes*.

Nesta *população inicial*, a lista de genes gerados aleatoriamente passa pelo operador de fecundação que gera a população inicial. Nessa lista de fecundação um gene é escolhido aleatoriamente e as folhas desse gene são combinadas com a raiz de um outro gene que combine, formando uma árvore conforme ilustra a Figura 4.8. O cromossomo é um conjunto de genes, que respeita formação de cromossomos que o mesmo cromossomo não apareça mais de uma vez em uma determinada população.

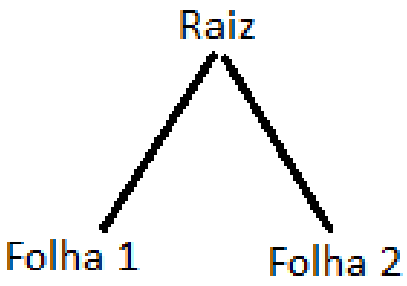


Figura 4.7: Representação do gene no SETAX.

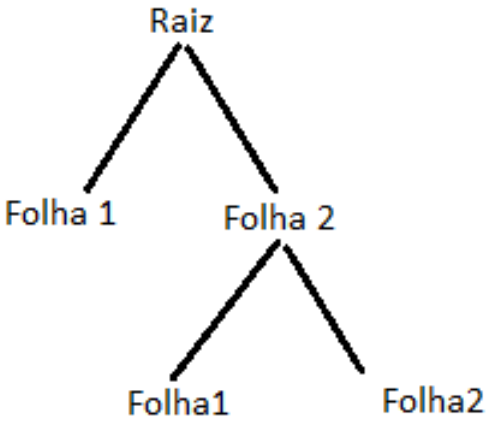


Figura 4.8: Fecundação entre dois genes no SETAX

Após a criação da população inicial, os operadores são aplicados à população, o primeiro operador aplicado é o de seleção. No entanto, antes do processo de seleção, é aplicada a seleção elitista, onde é selecionado os melhores indivíduos da população e os adicionam na próxima geração sem ter que passar pelos outros operadores.

No operador genético de seleção o *ponto de corte* estabelecido é a *adaptação média* dos cromossomos da população. Esta *adaptação média* é obtida através da média aritmética da população, o somatório das *adaptações* de cada cromossomo da população é dividido pelo número de indivíduos da população. A *adaptação* de cada cromossomo é obtida pelo somatório do *grau de adaptação* de cada gene que compõe o cromossomo, onde a *adaptação* de cada gene é a soma da características comuns entre a raiz e as folhas. Assim, os indivíduos que estiverem acima desse ponto de corte são selecionados. Após a seleção, o operador de cruzamento é aplicado.

No operador genético de cruzamento, os indivíduos escolhidos na seleção são selecio-

nados aleatoriamente, os compara, os genes dominantes de cada cromossomo é escolhido e essa lista de genes dominantes passa pelo processo de fecundação. Assim, todos os indivíduos passam pelo processo de cruzamento.

Após o cruzamento, o operado genético de mutação é aplicado. O operador genético de mutação ao qual a *população* é submetida utiliza apenas a função de remoção, que remove a folha de um gene dos cromossomos aleatoriamente.

Após a mutação ser aplicada o ciclo se repete. A condição de parada do GAADT é o valor da *adaptação média* da população resultante do operador genético de mutação ser igual ao valor da *adaptação média* da população de origem, o que implica que a operação genética de mutação não adicionou genes a nenhum cromossomo da população de origem e contém, portanto, os cromossomos que representam as possíveis árvores em questão. Caso isso não aconteça, o ponto de parada poderá ser o número de gerações pré-determinado pelo usuário.

4.3 Um estudo de caso real: Baseado no melhoramento genético da cana de açúcar

O ponto de partida do programa de melhoramento genético da Cana-de-Açúcar da RIDESA é o banco de germoplasma localizado na estação de floração e cruzamento da Serra do Ouro (UFAL), no município de Murici, Estado de Alagoas. Nesse banco estão reunidos mais de 2000 genótipos, entre cultivares utilizados no país, clones, outras espécies relacionadas ao gênero Saccharum e cultivares importadas das diferentes regiões canavieiras do mundo, que estar localizado na unidade acadêmica centro de ciências agrárias U.A.CECA /UFAL (Figura 4.9) [2].

Após a obtenção das sementes em cruzamentos pré-estabelecidos pelas equipes das universidades que compõem a RIDESA, as mesmas são enviadas aos respectivos Estados, onde são produzidas as plântulas que, uma vez transplantadas para o campo, definem a primeira fase de seleção ($T1$). A RIDESA tem produzido anualmente mais de 1.500.000 plântulas para a fase T1. Em algumas universidades a seleção é feita em duas épocas, abril e julho, de forma a se buscar, naquela primeira época, clones que apresentem a característica importante de precocidade. Este ramo de seleção leva a sigla HP (Hiper precoce). Mais de 5 mil novos clones têm sido gerados anualmente pelas universidades, incluindo também os HP. Estes clones são avaliados posteriormente nas estações experimentais na fase denominada T2. Os clones são avaliados experimentalmente em parcelas de um sulco de 5 a 8 metros de comprimento fonte [3].

Na segunda fase de seleção ($T2$) seleciona-se em planta e soca os clones superiores

[2] http://www.ceca.ufal.br/

[3] http://www.ridesa.com.br/mgenetico.htm

que por sua vez são avaliados na fase $T3$. A RIDESA tem selecionado para a fase $T3$ mais de 3000 novos clones por ano. A partir desta fase os clones selecionados em cada universidade são intercambiados entre elas. Nesta etapa os novos clones são multiplicados e introduzidos nas usinas e destilarias conveniadas com as respectivas universidades que atuam nas diferentes regiões canavieiras do Brasil. Nas terras das usinas e destilarias tem sido avaliados por meio de experimentos por três anos consecutivos os clones promissores. Esta fase é denominada Fase Experimental FE.

O programa de desenvolvimento de novos cultivares de cana-de-açúcar é por natureza essencialmente de longa duração. Logo, a persistência é uma virtude das pessoas envolvidas neste processo. Normalmente o lançamento de novos cultivares tem ocorrido após cerca de 13 anos de inúmeras avaliações dos clones por meio de experimentos observando-se a reação dos clones às doenças e pragas e a produtividade dos mesmos em diferentes ambientes de produção.

Buscando diminuir esse tempo, iniciamos um estudo sobre a base de dados existente na RIDESA, este estudo de caso refere-se a um banco de caracterização pertence ao Programa de Melhoramento genético da cana-de-açúcar - PMGCA.

Figura 4.9: Imagem da Serra do Ouro localizada em Murici-Al Fonte: Ceca

4.3.1 Descrição da Base do estudo de caso.

Em busca de um melhor entendimento da base de dados iremos descrever as características existente na base.

Segue abaixo uma leve explicação sobre cada atributo da espécies de cana-de-açúcar existente na base:

- **BRIX** - é uma escala numérica que mede a quantidade de sólidos solúveis em uma solução de sacarose;
- **MAT** - Maturação (ou amadurecimento) é o processo de desenvolvimento dos seres vivos ou suas partes no sentido de tornar o organismo apto para a reprodução. Em relação aos animais, diz-se que, nessa altura, eles atingem a maturidade. Os seres que ainda não atingiram a maturidade designam-se como "imaturos";
- **TCH** - Tonelada de Cana por hectare;
- **FIB** - Fibra - Quanto mais fibra a cana fica mais resistente a praga;
- **BRO** - Brotação de socaria- Quantificação do aparecimento de brotos na soca (2º corte) que darão origem a novas plantas;
- **PER** - Perfilho - são brotações que já estão quase todas as características da uma planta adulta;
- **DES** - Desenvolvimento - o tipo de crescimento;
- **HAB** - Hábito de crescimento de uma planta, ou seja, sua arquitetura de crescimento.
- **DIA** - Diâmentro do colmo - a espessura da cana;
- **FLO** - Florescimento - Aparecimento da inflorescência da cana, o que é prejudicial para o setor industrial da cana, pois tem-se uma perda na produtividade;
- **CHO** - Chochamento - O chochamento (ou isoporização) tem início com a ocorrência do processo de florescimento e se caracteriza pelo secamento do interior do colmo e perda de peso final, oriunda da redução do volume de caldo;
- **CHO F** - Chochamento Fisiológico - esta presente no gene da cana, ou seja, a variedade apresenta-o tendo o aparecimento da inflorescência ou não.

4.4 Comparações entre os algoritmos

Nesta seção iremos comparar duas matrizes características, uma sem parcimônia e outra com parcimônia. Para isso será gerada uma árvore para cada tipo de algoritmo descrito neste trabalho usando essas matrizes e também serão geradas árvores para representar o resultado de cada um desses algoritmos mencionados. Todos os algoritmos irão usar as

mesmas bases de teste que é uma matriz característica sem parcimônia vide Tabela 4.1 e outra base com parcimônia vide Tabela4.5.

Tabela 4.1: *Matriz característica sem parcimônia usada para testar os algoritmos.*

	C1	C2	C3	C4	C5	C6
A	0	0	0	0	0	0
B	1	0	0	0	0	0
C	1	1	1	0	1	0
D	1	1	1	1	0	1

Tabela 4.2: *Matriz distância sem parcimônia.*

	A	B	C	D
A		0	0	0
B	0		1	1
C	0	1		3
D	0	1	3	

Tabela 4.3: *Matriz característica com parcimônia usada para testar os algoritmos.*

	C1	C2	C3	C4	C5	C6	C7	C8
A	1	1	0	0	0	0	1	1
B	1	1	1	1	0	0	0	0
C	1	1	0	0	1	1	0	0
D	0	0	1	1	1	1	1	1

Tabela 4.4: *Matriz Distância com parcimônia.*

	A	B	C	D
A		2	2	2
B	2		2	2
C	2	2		2
D	2	2	2	

Seguem os resultados e as árvores gerada com os algoritmos mencionados anteriormente.

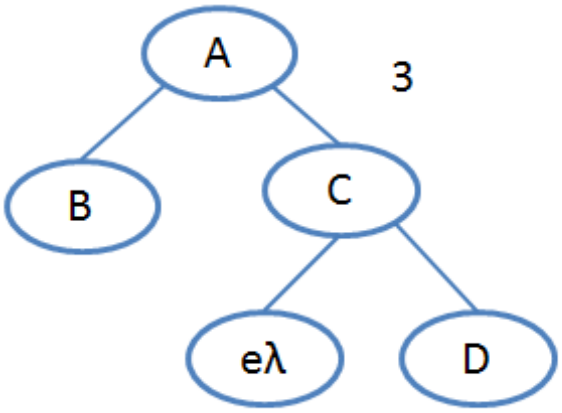


Figura 4.10: Árvore gerada pela regra de Inclusão e Exclusão na matriz sem parcimônia

O primeiro algoritmo mencionado neste trabalho foi o de *regra de Inclusão e Exclusão*. Este por sua vez gerou a árvore com a matriz característica sem parcimônia vide Tabela 4.1, gerou a árvore mostrada na Figura 4.10.

Notamos que o algoritmo obteve um escore máximo de 3(três), a primeira raiz gerada foi a A, notamos que na matriz a espécie A tem os valores de suas características 0(zero). Isto significa que esta espécie não tem nenhuma característica similar a outra, portanto, deveria ser gerada por último.

Já na matriz característica com parcimônia vide Tabela 4.5 esse algoritmo não consegue gerar a árvore, pois todas as espécies têm a mesma cardinalidade.

O segundo algoritmo (*Algoritmo de Wagner*) usando a matriz característica sem parcimônia gerou a seguinte árvore, conforme mostrado na Figura 4.11.

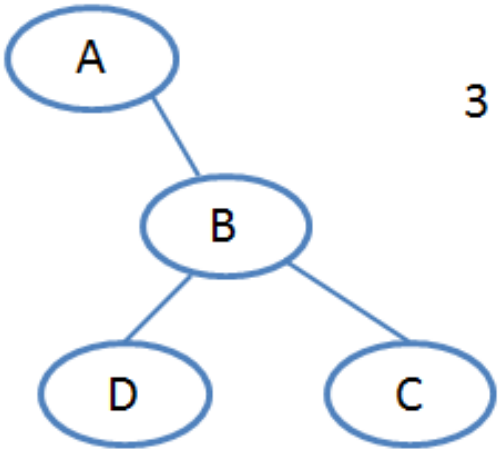


Figura 4.11: Árvore gerada pelo Algoritmo de Wagner

A árvore gerada vai depender da ordem na qual as espécies estão dispostas na matriz característica. Para se obter uma árvore filogenética ótima nesse caso, é necessário alterar as posições das espécies na matriz analisada.

No caso do algoritmo de Wagner aplicado à matriz característica com parcimônia, não conseguimos criar a árvore, mesmo alterando a ordem das espécies, pois as espécies têm as mesmas cardinalidades.

Já o caso do *Algoritmo das Médias* conseguimos gerar a seguinte árvore vide Figura 4.12 para a matriz característica sem parcimônia.

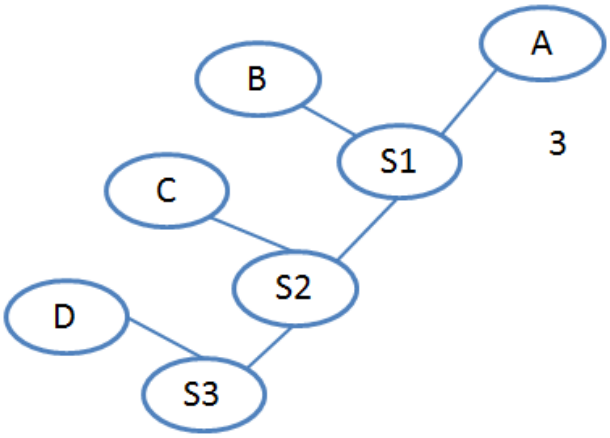


Figura 4.12: Árvore gerada pelo Algoritmo de Médias

Na matriz característica sem parcimônia, o algoritmo de Médias constrói a árvore ótima como monstrado na Figura 4.12, pois o mesma não tem parcimônia. Caso tenha, esse algoritmo irá depender da ordem na qual as espécies estão dispostas na matriz distância. O algoritmo neste caso funciona igual ao o de Wagner.

Por último, foram analisadas as árvores geradas pelo SETAX para as mesmas matrizes, conforme a Figura 4.13.

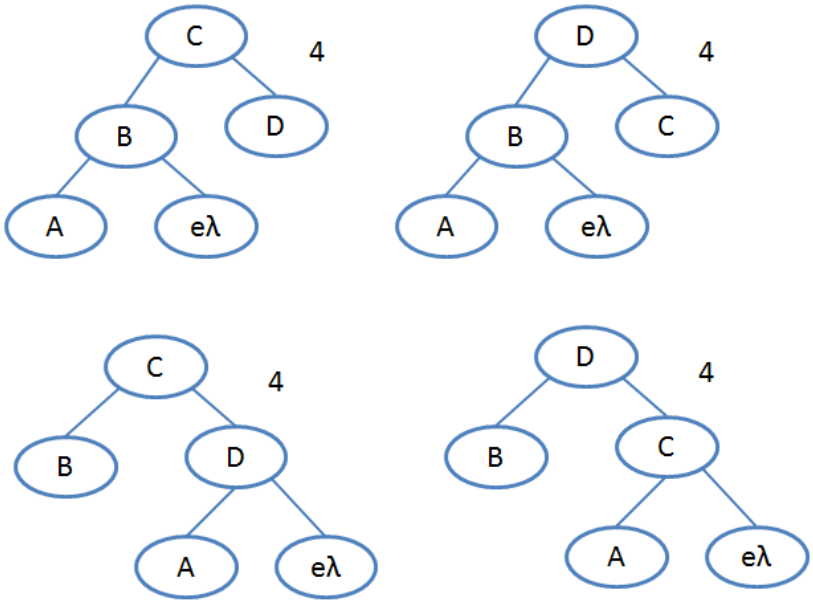


Figura 4.13: Árvores geradas pelo SETAX.

Conforme as árvores geradas pelo SETAX, podemos concluir que o mesmo não apre-

senta os problemas dos outros algoritmos, pois o mesmo gera árvores ótimas para as duas matrizes, percebe-se também que ele gera todas as árvores ótimas, deixando para o usuário o critério da escolha.

Tabela 4.5: *Score de cada algoritmo.*

Algoritmos	Tab. Com Parcimônia	Tab. sem Parcimônia
Inclusão e Exclusão	1	3
Wagner	1	4
Médias (UPGMA)	1	4
SETAX	4	4

A árvore filogenética construída pelo SETAX está respaldada sob as premissas Biológicas, já que a árvore filogenética construída é resultante da análise das características apresentadas na matriz características dos seres investigado e nos fatos fornecidos pelo usuário [2] , [37], [47], [33]. Deve-se ressaltar que a árvore filogenética construída pelo SETAX não será dependente da ordem na qual os organismos estão dispostos na matriz característica.

4.5 Problemas Encontrados

O algoritmo da regra de inclusão e exclusão e o algoritmo de Wagner servem muito bem para construir a árvore filogenética ótima quando os dados sobre a evolução das características apresentadas pelas espécies não apresentam conflitos [4]. Pois, se esses dados forem conflitantes, então a matriz característica de dimensão nxm terá duas colunas i e j, tal que:

$\forall$x $\in 1, ..., m, \exists S_x =$ n$|\ c_{n,x} \neq 0$ com $S_i \cap S_j \neq \oslash, Si \subset Sj$ e vice-versa.

Por exemplo, considere a matriz característica da Tabela 4.6.

Tabela 4.6: *Matriz característica com conflito*

	1	2	3	4	5	6
A	0	0	0	0	0	0
B	1	1	0	0	1	1
C	1	1	1	1	1	1
D	1	1	1	1	0	0

Inicialmente constrói-se os conjunto $S_1 = \{B, C, D\}$,S$_2 = \{B, C, D\}$,S$_3 = \{C, D\}$,S$_4$

[4]Conflito - O conflito ocorre quando as relações de parentesco expressas por uma característica contradizem as relações de parentesco expressas por uma outra característica.

$= \{C, D\}$,$S_5 = \{B, C\}$,$S_6 = \{B, C\}$, $note que S_1 = S_2, S_3 = S_4, e S_5 = S_6$, então tem-se a seguinte análise dos dados:

- $S_1 \cap S_3 = S_3$ e $S_1 \cap S_5 = S5$, logo as características 1 e 2 não apresentam conflitos com as características 3, 4, 5 e 6;
- $S_3 \cap S_5 = C, S_3 S_5$ e $S_5 S_3$, logo as características 3 e 4 apresentam conflitos com as características 5 e 6.

Nesse caso, o algoritmo da regra de inclusão e exclusão irá construir um conjunto de árvores filogenéticas para as espécies analisadas com cardinalidade maior que 1(um) vide Figura 4.14. A árvore filogenética ótima, para as espécies de uma matriz característica com conflito construída pelo algoritmo da regra de inclusão e exclusão, será a árvore filogenética obtida pela aplicação de um critério de otimização ao conjunto de árvores filogenéticas construído. Enquanto que o algoritmo de Wagner irá construir uma árvore filogenética que será dependente da ordem na qual as espécies estão dispostas na matriz característica com conflito analisada (Figura 4.14a). Para se obter a árvore filogenética ótima com o algoritmo de Wagner, deve-se primeiro alternar as posições das espécies na matriz analisada, para se obter o conjunto de todas as possíveis árvores filogenéticas para essas espécies. E depois, aplica-se sobre esse conjunto um critério de otimização.

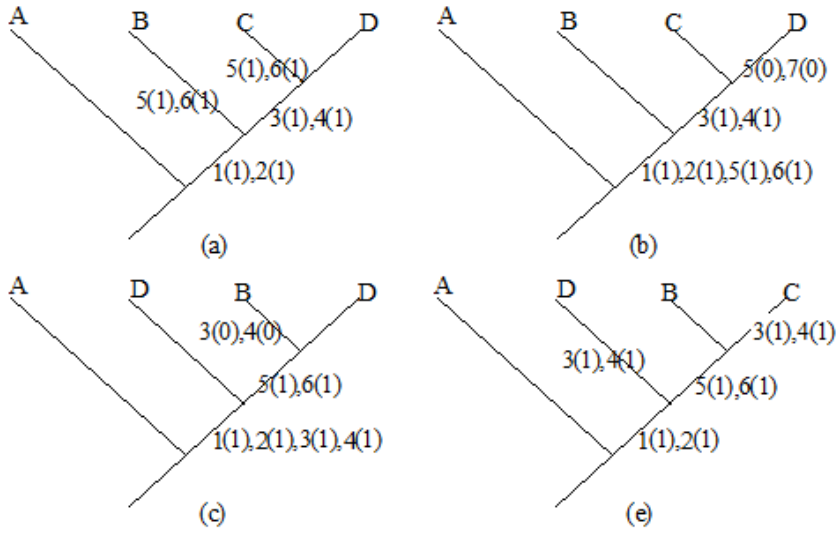


Figura 4.14: Árvores filogenéticas construídas pelo algoritmo de inclusão e exclusão a partir da Tabela 4.6

O algoritmo das médias constrói a árvore filogenética ótima, quando os dados sobre a evolução das características apresentadas pelas espécies não apresenta conflito. Pois, se os dados sobre a evolução das características apresentadas pelas espécies forem conflitantes, então a matriz distância de dimensão nxn, com θ igual ao menor valor armazenado, terá

dois ou mais elementos $d_{i,j} = d_{u,v} = \theta$, tal que: i, j $\neq$ u, v e i, j $\neq$ v, u. Por exemplo, considere a matriz distância da Tabela 4.7 construída a partir da Tabela 4.6.

Tabela 4.7: *Matriz distância com conflito*

	A	B	C	D
A	0	4	6	4
B	4	0	2	4
C	6	2	0	2
D	4	4	2	0

O menor valor não-nulo armazenado nesta tabela é 2(dois) pois o valor zero significa que a espécie é a mesma . O valor 2(dois) corresponde a distância entre as espécies B e C, e entre as espécies C e D, o que é um conflito. Quando a matriz distância analisada pelo algoritmo das médias apresenta conflito, a árvore filogenética construída irá depender da ordem na qual as espécies estão dispostas na matriz distância (Fig. 4.15). A árvore filogenética ótima, construída pelo algoritmo das médias a partir de uma matriz distância com conflito, será obtida da mesma forma que a árvore filogenética ótima construída pelo algoritmo de Wagner a partir de uma matriz característica com conflito.

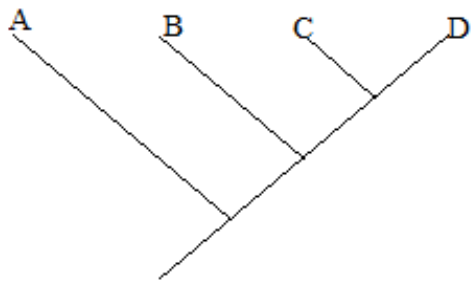


Figura 4.15: Árvores filogenéticas construídas pelo algoritmo das médias a partir da Tabela 4.7.

No próximo capítulo será apresentada a conclusão e trabalhos futuros.

Capítulo 5

Conclusões

Filogenia não é um assunto recente, desde os tempos de Darwin [16] se procura montar a árvore das espécies. Como o espaço de soluções deste problema é gigantesco (apresenta custo fatorial), o emprego de algoritmos genéticos se apresenta como sendo uma boa alternativa para sua exploração.

O uso de Computação Evolutiva se mostrou muito eficiente e eficaz em sua exploração ao longo desta pesquisa. Em todas as simulações, sempre foi obtido um bom resultado dentro de um período de tempo razoável(de 30 minutos a uma hora). E seus resultados gerados neste trabalho foram avaliados por um pesquisador da área de Agronomia(Bruno Nascimento [1]).

Uma grande vantagem de se usar computação evolutiva é a possibilidade de se trabalhar com muitos pontos de partida simultaneamente, além da disponibilidade de mecanismos de recombinação das soluções existentes. Com isso, aumentam-se as chances de escapar de máximos locais ruins e atingir o máximo global ou pelo menos um bom máximo local [21] e [18].

Outro ponto que merece ser destacado é o uso de técnicas de Computação Evolutiva, que se mostrou uma poderosa ferramenta para ajudar no problema de encontrar uma boa árvore filogenética, a partir de uma sequência de bases analisadas, devido ao seu elevado poder de exploração de um vastos espaços de solução conforme Cancino[12].

O sistema (SETAX) que foi desenvolvido e utilizado ao longo deste mestrado. A interface gráfica deste sistema foi criada para facilitar o seu uso por parte do usuário, mesmo que este não domine a computação, nele é possível alterar os parâmetros iniciais sobre o algoritmo genético utilizado internamente, tais como tamanho da população inicial (padrão 200), porcentagem de mutações(padrão 30%), quantidade de iterações (padrão 10) que limita o tempo máximo de busca. Não é necessária a digitação de todos estes parâmetros, pois o SETAX possui um conjunto de valores iniciais que se mostraram bastante eficientes na maioria dos testes realizados.

[1] http://www.ceca.ufal.br/.

Além disso, é necessário ressaltar que este sistema foi projetado para auxiliar e orientar o usuário durante o processo de execução e sempre que um dado for necessário ou um parâmetro for indispensável e não estiver nos valores iniciais, o usuário será solicitado a fornecer este dado ou informação faltante. Por exemplo, o nome do arquivo que contém o conjunto de espécies a serem analisadas não pode fazer parte do conjunto de valores iniciais, razão pela qual ele é solicitado sempre que o usuário ordenar o início de execução. Uma tela específica de busca de arquivos no padrão Windows será exibida para que o usuário aponte para o arquivo desejado. Desta forma, os recursos e as facilidades disponíveis nesta ferramenta ajudaram muito o desenvolvimento deste trabalho e poderão ser úteis também para novas pesquisas sobre Filogenia e sobre Computação Evolutiva.

As seguintes características não foram implementadas nesta versão do SETAX:

- escolher algum outro método de busca;
- melhorar a perfomace;
- comparar o resultados mais detalhadamente;
- oferecer explicações de como o sistema chegou ao resultado.

Apoiado por este sistema, foi possível verificar o poder da Computação Evolutiva para resolver este complexo problema, cujo espaço de solução cresce em função do número de bases analisadas. Problemas complexos como este impedem o uso de técnicas convencionais para sua solução. É praticamente impossível analisar todas as árvores candidatas (busca exaustiva) quando o número de sequências de bases for superior a umas poucas dezenas ([31]).Esta foi a principal motivação que levou ao uso de Computação Evolutiva.

Nesse sentido, as técnicas de Computação Evolutiva permitem explorar de maneira muito eficiente o espaço de solução e, geralmente, uma boa resposta pode ser encontrada em tempo razoável ([5]). Por outro lado, é preciso ressaltar que não existem garantias de ser encontrada a melhor solução para o problema: o que se garante é uma boa solução, caso se proceda a uma escolha adequada de parâmetros estabelecidos para a busca. As técnicas relativas à Computação Evolutiva foram discutidas no Capítulo 1.

Durante todo o processo de construção do SETAX, desde proposições de idéias até desenvolvimento de aplicações, foram submetidas a congressos e periódicos nacionais e internacionais objetivando uma melhor avaliação/validação da proposta. Dentre eles podemos destacar 13th IEEE Joint International Computer Science and Information Technology Conference(JICSIT 2011), onde tivemos o artigo aceito, o Journal of Engineering and Computer Innovations também tivemos o artigo aceito no Simpósio de Computação Aplicada SCA, onde o trabalho também foi publicado.

Com isso, espera-se ter contribuído nas áreas de *i)* Inteligência Artificial,*ii)*Agronomia e *iii)* Bioinformática.

Nesta dissertação procurou-se lidar com a explosão combinatória e com a necessidade de gerar uma justificativa para a solução de um problema do tipo NP-completo resolvido com um algoritmo genético. O domínio de aplicação restringiu-se à Agronomia. A idéia inicial era construir um sistema que gerasse a árvore filogenética para um conjunto de seres vivos a partir de dados de sua matriz característica fornecida pelo usuário.

Trabalhos Futuros

Quanto aos trabalhos futuros, podemos destacar:

- Utilização e análise de desempenho entre as diversas estratégias, a fim de obter dados mais precisos quanto aos tempos de resposta (eficiência das estratégias) e a qualidade das soluções encontradas;
- Testes comparativos com implementações de outras abordagens para verificar desempenho e qualidade das soluções encontradas;
- Refatoramento do código criado com fins de incremento de performance, uma vez que o objetivo da concepção do mesmo foi o de prova de conceito;
- Comparar com outras técnicas de Inteligencia Artificial como redes neurais;
- Utilizar outro estudo de caso como o da saúde.

Referências

[1] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997), 'Gapped blast and psiblast: a new generation of protein database search programs', *NUCLEIC ACIDS RESEARCH* **25**(17), 3389–3402.

[2] Amorim, D. (1994), 'Elementos básicos de sistemática filogênica', *Sociedade Brasileira de Entomologia, São Paulo-SP, 314p.*

[3] Andreatta, A. A. & Ribeiro, C. C. (2002), 'Heuristics for the phylogeny problem', *Journal of Heuristics* **8**, 429–447. URL `http://portal.acm.org/citation.cfm?id=594956.595096`.

[4] Atmar, W. (1994), 'Notes on the simulation of evolution', *IEEE Transactions on Neural Networks, Vol. 5:1, pp. 130–147.*

[5] Bäck, T. (1996), *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*, Oxford University Press, Oxford, UK.

[6] Bäck, T. & Schwefel, H.-P. (1993), 'An overview of evolutionary algorithms for parameter optimization', *Evol. Comput.* **1**, 1–23. URL `http://dx.doi.org/10.1162/evco.1993.1.1.1`.

[7] Back, T., Fogel, D. B. & Michalewicz, Z., eds (1999), *Basic Algorithms and Operators*, IOP Publishing Ltd., Bristol, UK, UK.

[8] Barricelli, N. A. (1957), 'Symbiogenetic evolution processes realized by artificial methods', *Methodos 9:35-36.*

[9] Barricelli, N. A. (1962), 'Numerical testing of evolution theories', *Part I. Theroetical Introduction and Basic Tests. Acta Biotheoretzca.*

[10] Bernardi, N. (1999), 'Resumo das aulas sobre: Sistemática filogenética, i curso especial de sistemática zoológica do departamento de ciências biológicas', *São Carlos - SP.*

[11] Boukerche, A., de Melo, A. C. M. A., Ayala-Rincón, M. & Walter, M. E. M. T. (2007), 'Parallel strategies for the local biological sequence alignment in a cluster of workstations', *J. Parallel Distrib. Comput.* **67**, 170–185. URL `http://dx.doi.org/10.1016/j.jpdc.2006.11.001`.

[12] Cancino, W. & Delbem, A. C. B. (2007), A multi-objective evolutionary approach for phylogenetic inference, *in* 'Proceedings of the 4th international conference on Evolutionary multi-criterion optimization', EMO'07, Springer-Verlag, Berlin, Heidelberg, pp. 428–442. URL `http://portal.acm.org/citation.cfm?id=1762545.1762584`.

[13] Cockerham, C. (1963), 'Estimation of genetic variances, in: W.d. hanson, h.f. robinson (eds.)', *Statistical Genetics and Plant Breeding, National Academy of Sciences, Washington, DC*.

[14] Coelho, G. P., da Silva, A. E. A. & Von Zuben, F. J. (2007), Evolving phylogenetic trees: a multiobjective approach, *in* 'Proceedings of the 2nd Brazilian conference on Advances in bioinformatics and computational biology', BSB'07, Springer-Verlag, Berlin, Heidelberg, pp. 113–125. URL `http://portal.acm.org/citation.cfm?id=1776474.1776486`.

[15] Cohoon, J. P., Hegde, S. U., Martin, W. N. & Richards, D. (1987), Punctuated equilibria: a parallel genetic algorithm, *in* 'Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application', L. Erlbaum Associates Inc., Hillsdale, NJ, USA, pp. 148–154. URL `http://portal.acm.org/citation.cfm?id=42512.42532`.

[16] Darwin, C. (1859), *A Origem das Espécies e a Seleção Natural*, 5ª ed., Editora Hemus.

[17] Davis, L. (1985), Job shop scheduling with genetic algorithms, *in* 'Proceedings of the 1st International Conference on Genetic Algorithms', L. Erlbaum Associates Inc., Hillsdale, NJ, USA, pp. 136–140. URL `http://portal.acm.org/citation.cfm?id=645511.657084`.

[18] de Castro, L. N. (2006), *Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications (Chapman & Hall/Crc Computer and Information Sciences)*, Chapman & Hall/CRC. URL `http://www.worldcat.org/isbn/1584886439`.

[19] De Jong, K. A. (2002), *Evolutionary Computation: A Unified Approach*, 1st ed., The MIT Press. URL `http://www.worldcat.org/isbn/0262041944`.

[20] Deb, K. & Kalyanmoy, D. (2001), *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, Inc., New York, NY, USA.

[21] Fogel, D. (1994), 'An introduction to simulated evolutionary computation', *IEEE Transactions on Neural Networks, vol. 5, no. 1, pp. 3- 14.*

[22] Fogel, D. (1999), 'Evolutionary computation toward a new philosophy of machine intelligence', *2nd edition, IEEE Press.*

[23] Fraser, W. & Hart, J. F. (1964), 'Remark on algorithm 162: Near-minimax polynomial approximations and partitioning of intervals', *Commun. ACM* **7**, 486–489. URL `http://doi.acm.org/10.1145/355586.364820`.

[24] Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[25] Greene, W. A. (2005), Schema disruption in tree-structured chromosomes, *in* 'Proceedings of the 2005 conference on Genetic and evolutionary computation', GECCO '05, ACM, New York, NY, USA, pp. 1401–1408. URL `http://doi.acm.org/10.1145/1068009.1068233`.

[26] Hauser, R. & Männer, R. (1994), 'Implementation of standard genetic algorithm on mind machines', *Parallel Problem Solving from Nature III*.

[27] Holland, J. H. (1975), 'Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence', *MIT Press*.

[28] K., B. (1995), 'Progress and pitfalls in systematics: cladistics, dna and morphology', *Acta Bot. Neerl.*

[29] Koza, J. R. (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)*, 1 ed., The MIT Press. URL `http://www.worldcat.org/isbn/0262111705`.

[30] Lawrence, G. H. M. (1951), 'Taxonomy of vascular plants', *Macmillan*.

[31] Lewis, D. D. (1998), Naive (bayes) at forty: The independence assumption in information retrieval, Springer Verlag, pp. 4–15.

[32] Marsh, A., Simistira, F. & Robb, R. (1998), 'Vr in medicine: Virtual colonoscopy', *Future Gener. Comput. Syst.* **14**, 253–264. URL `http://dx.doi.org/10.1016/S0167-739X(98)80025-3`.

[33] Maruyama, O., Matsuda, A. & Kuhara, S. (2005), 'Reconstructing phylogenetic trees of prokaryote genomes by randomly sampling oligopeptides', *Int. J. Bioinformatics Res. Appl.* **1**, 429–446. URL `http://portal.acm.org/citation.cfm?id=1356495.1356501`.

[34] Mayer, H., Nagy, I., Knoll, A., Braun, E. U., Bauernschmitt, R. & Lange, R. (2007), 'Haptic feedback in a telepresence system for endoscopic heart surgery', *Presence: Teleoper. Virtual Environ.* **16**, 459–470. URL `http://portal.acm.org/citation.cfm?id=1297017.1297019`.

[35] Mckay, R. I., Hoai, N. X., Whigham, P. A., Shan, Y. & O'Neill, M. (2010), 'Grammar-based genetic programming: a survey', *Genetic Programming and Evolvable Machines* **11**, 365–396. URL `http://dx.doi.org/10.1007/s10710-010-9109-y`.

[36] Meidanis J., S. J. (1994), 'Uma introdução à biologia computacional', *Recife - PE.*

[37] Poladian, L. & Jermiin, L. S. (2006), 'Multi-objective evolutionary algorithms and phylogenetic inference with multiple data sets', *Soft Comput.* **10**, 359–368. URL http://portal.acm.org/citation.cfm?id=1102886.1102895.

[38] Reijmers, T. (1999), 'Using genetic algorithms for the construction of phylogenetic trees: application to g-protein coupled receptor sequences.', *BioSystems - Netherlands, 49:3143.*

[39] Rosenberg, A. L. (1967), 'Real-time definable languages', *J. ACM* **14**, 645–662. URL http://doi.acm.org/10.1145/321420.321423.

[40] Ross, J. C., Estépar, R. S., Díaz, A., Westin, C.-F., Kikinis, R., Silverman, E. K. & Washko, G. R. (2009), Lung extraction, lobe segmentation and hierarchical region assessment for quantitative analysis on high resolution computed tomography images, *in* 'Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II', MICCAI '09, Springer-Verlag, Berlin, Heidelberg, pp. 690–698. URL http://dx.doi.org/10.1007/978-3-642-04271-3_84.

[41] Rudolph, G. (1993), 'Massively parallel simulated annealing and its relation to evolutionary algorithms', *Evol. Comput.* **1**, 361–383. URL http://dx.doi.org/10.1162/evco.1993.1.4.361.

[42] Sastry, K. (2001), 'Evaluation-relaxation schemes for genetic and evolutionary algorithms. master's thesis, university of illinois at urbana-champaign, general engineering department', *General Engineering Department, Urbana, IL.*

[43] Scott, S. D., Samal, A. & Seth, S. (1995), Hga: a hardware-based genetic algorithm, *in* 'Proceedings of the 1995 ACM third international symposium on Field-programmable gate arrays', FPGA '95, ACM, New York, NY, USA, pp. 53–59. URL http://doi.acm.org/10.1145/201310.201319.

[44] Searls, D. B. (1999), 'Formal language theory and biological macromolecules'.

[45] Vieira, R. V. (2003), Um Algoritmo Genético Baseado em Tipos Abstratos de Dados e sua Especificação em Z, (tese de doutorado), Centro de Informática(CIn) Universidade Federal de Pernambuco, Recife - PE.

[46] Vieira, R. V. & Lopes, M. A. (1999), 'Uma abordagem teórica inicial para os algoritmos genéticos através de tipos abstratos de dados', *Relatório Técnico do Departamento de Informática da Universidade Federal de Pernambuco.*

[47] Wiens, J. J. (2006), 'Missing data and the design of phylogenetic analyses', *J. of Biomedical Informatics* **39**, 34–42. URL `http://dx.doi.org/10.1016/j.jbi.2005.04.001`.

[48] Wiley E.O., Siegel-Causey D., B. D. (1991), *Funk V.A.,The compleast cladist: A primer of philogenetic procedures*, University of Kansas Museum of Natural Históry, Kansas, USA.