PATRÍCIA SILVA NASCIMENTO BARROS

RECONHECIMENTO QUÂNTICO DE PADRÕES APLICADOS À SEQUÊNCIAS DE DNA

RECIFE

2011

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

RECONHECIMENTO QUÂNTICO DE PADRÕES APLICADOS À SEQUÊNCIAS DE DNA

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Modelagem Estatística e Computacional

Orientador: Prof. Dr. Wilson Rosa de Oliveira Junior

RECIFE

2011

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

Dissertação sob o título *RECONHECIMENTO QUÂNTICO DE PADRÕES APLICADOS À SEQUÊNCIAS DE DNA*, apresentada por Patrícia Silva Nascimento Barros e aprovada em 21 de fevereiro de 2011, em Recife, Estado de Pernambuco, pela banca examinadora constituída por:

Orientador:

Prof. Dr. Wilson Rosa de Oliveira Junior
Departamento de Estatística e Informática - UFRPE

Banca Examinadora:

Prof^a. Dr^a. Tatijana Stosic
Departamento de Estatística e Informática - UFRPE

Prof. Dr. Borko D. Stosic
Departamento de Estatística e Informática - UFRPE

Prof. Dr. Felipe Maia Galvão França
Centro de Tecnologia - COPPE - UFRJ

Dedico à meus pais Maria de Lourdes e Francisco Vitorino, marido Kleber e filho Khalel.

Agradecimentos

Agradeço, primeiramente, ao Programa de Pós-Graduação em Biometria e Estatística Aplicada pela estrutura física e o ambiente intelectual propício para estudos científicos diversos, o qual pude desfrutar durante os dois anos em que transcorreu a minha pós-graduação.

Agradeço também, aos professores Dr. Eufrázio de Souza Santos, Dr. Borko D. Stosic, Dr. José Antônio Aleixo da Silva, Dra. Tatijana Stosic, Dr. Cláudio Tadeu Cristino, por suas preciosas aulas e ensinamentos.

Ao meu orientador prof. Dr. Wilson Rosa de Oliveira Junior, por suas aulas e pela orientação que tornaram possíveis a realização deste trabalho científico.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro ao longo destes 1 ano e 6 meses, sem o qual não seria possível a conclusão deste trabalho.

À todos os estudantes de mestrado e doutorado em Biometria e Estatística Aplicada, com os quais convivi ao longo destes 2 anos.

Agradeço aos meus pais, com os quais tive contato efetivo e afetivo, e que sempre se orgulharam e incentivaram o meu crescimento.

Agradeço também, a meu marido Kleber pela força e incentivo à concluir este trabalho. E meu querido filho Khalel pela força para seguir em frente.

"*Qualquer um que não se choque com a teoria quântica é porque não a entendeu.*"

Niels Bohr

"*Como os sonhos, as estatísticas são uma forma de realização do desejo.*"

Jean Baudrillard

Resumo

A computação quântica é uma área de pesquisa recente que engloba três áreas conhecidas: matemática, física e computação. Com as pesquisas na área de algoritmos quânticos veio a necessidade de entender e expressar tais algoritmos do ponto de vista de programação. Diversas linguagens e modelos para programação quântica de alto nível têm sido propostas nos últimos anos. A Mecânica Quântica (MQ) é um conjunto de regras matemáticas que servem para a construção de teorias físicas, desde a sua criação até os dias de hoje ela tem sido aplicada em diversos ramos. Neste contexto se desenvolveu a Computação Quântica, talvez a mais espetacular proposta de aplicação prática da MQ. A dificuldade de se desenvolver algoritmos quânticos propicia o uso de técnicas alternativas à solução de problemas puramente algorítmica, como por exemplo o aprendizado de máquinas e algoritmos genéticos. Carlo Trugenberger propõe um modelo de memória quântica associativa onde os padrões binários de n bits são armazenados em superposição com um subconjunto apropriado da base computacional de n qubits. Este modelo resolve o problema de escassez de capacidade bem conhecida da memória clássica associativa, provendo uma melhoria grande em capacidade. A distribuição proposta por Trugenberger usa a distância de Hamming, em que as amplitudes tem um pico nos padrões armazenados, que tem menor distância em relação à entrada. A precisão do reconhecimento de padrões pode ser ajustado por um parâmetro b, isto é, aumentando b aumenta a probabilidade de reconhecimento. Este trabalho analisa a diversidade genética das abelhas sem ferrão *Melipona quinquefasciata*, obtidas de várias colônias silvestres, em localidades distintas da Chapada do Araripe-CE, Chapada da Ibiapaba-CE, cidade do Canto do Buriti-PI e Luziânia-GO. As sequências de DNA foram transformados substituindo A por 00, G por 01, C por 10 e T por 11. Os resultados mostram que essa probabilidade é muito eficiente para reconhecer os padrões de sequências de DNA das abelhas sem ferrão *Melipona quinquefasciata* das regiões 18S e ITS1 parcial. O algoritmo não é computacionalmente eficiente em um computador clássico, mas será extremamente eficiente em um computador quântico. Concluiu-se que este método de reconhecimento quântico de padrões é melhor que o método clássico utilizado por Pereira.

Palavras-chave: Computação Quântica; Reconhecimento Quântico de Padrões; Sequências de DNA; Abelhas sem ferrão *Melipona quinquefasciata*.

Abstract

Quantum computing is a recent area of research that encompasses three known areas: mathematics, physics and computing. With the research in quantum algorithms came the need to understand and express such algorithms in terms of programming. Several languages and programming models for high-level quantum have been proposed in recent years. Quantum mechanics (QM) is a set of mathematical rules that serve for the construction of physical theories, from its inception until the present day it has been applied in various branches. In this context we developed the Quantum Computation, perhaps the most spectacular proposal for practical implementation of QM. The difficulty in developing quantum algorithms provides the use of alternative techniques to the solution of purely algorithmic problems, such as machine learning and genetic algorithms. Carlo Trugenberger proposes a model of quantum associative memory which binary patterns of n bits are stored in a quantum superposition of an appropriate subset of the computational basis of n qubits. This model solves the problem of insufficient capacity of the well known classical associative memory, providing a large improvement in capacity. The distribution proposed by Trugenberger uses the Hamming distance, where the amplitudes have a peak in the stored patterns, which has smaller distance from the entrance. The accuracy of pattern recognition can be adjusted by the parameter b, in other words increasing b increases the probability of recognition. This study examines the genetic diversity of stingless bees *Melipona quinquefasciata*, obtained from several wild colonies in different localities of the Chapada do Araripe-CE, Chapada da Ibiapaba-CE, city's Canto do Buriti-PI and Luziânia-GO. DNA sequences were processed by replacing A by 00, G by 01, C by 10 and T by 11. The results show that this probability is very efficient to recognize the patterns of DNA sequences of the stingless bees *Melipona quinquefasciata* regions 18S and ITS1 partial. The algorithm is not computationally efficient on a classical computer, but is extremely efficient on a quantum computer. It was concluded that this method of recognition of quantum standards is better than the classic method used by Pereira.

Key words: Quantum Computation, Quantum Pattern Recognition; DNA sequences; stingless bees *Melipona quinquefasciata*.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Computação quântica (PRESKILL, 1998) é normalmente associada com novas classes de complexidade que são inacessíveis para máquinas de Turing clássica. Algoritmos quânticos podem acelerar a solução de tarefas com relação aos seus análogos clássicos, como exemplos o algoritmo de fatoração de Shor (SHOR, 1995) e o algoritmo de procura de Grover (GROVER, 1996). Há, no entanto, outro aspecto da computação quântica que representa uma grande melhora em relação a clássica. Nos computadores tradicionais o armazenamento de informações exige a criação de uma tabela (RAM). A principal desvantagem deste sistema de endereçamento de memória orientada reside na sua rigidez. Recuperação de informação requer um conhecimento preciso do endereço de memória e, portanto, insumos incompletos ou ruidosos não são permitidos.

O emaranhamento quântico fornece um mecanismo natural para melhorar a capacidade de armazenamento de memórias associativas e recuperar ruído ou informações incompletas. Na verdade, o número de padrões binários que podem ser armazenados em tal memória quântica é exponencial no número de n qubits, $p_{max} = 2^n$, ou seja, é ideal no sentido de que todos os padrões binários que podem ser formados com os n bits podem ser armazenados (TRUGENBERGER, 2002).

Trugenberger revisa e amplia o modelo de memória quântica associativa que propôs recentemente. Neste modelo os padrões binários de n bits são armazenados na superposição quântica do subconjunto apropriado da base computacional de n qubits. A informação pode ser recuperada executando uma rotação entrada-dependente da memória do estado quântico dentro deste subconjunto e medindo o estado resultante. A precisão do padrão recordado pode ser afinado ajustando um parâmetro que representa o papel de uma temperatura efetiva. Este modelo resolve o problema de escassez de capacidade bem conhecida da memória clássica associativa, provendo uma melhoria grande em capacidade (TRUGENBERGER, 2003). A eficiência de recuperação de informações no mecanismo depende da distribuição dos padrões armazenados. A eficiência do reconhecimento é melhor quando o número de padrões armazenados é muito grande ou quando

os padrões isolados são muito diferentes dos outros, duas características muito intuitivas.

Barros et al. utilizaram este método de reconhecimento quântico de padrões para analisar os primeiros 15 mil nucleotídeos do DNA mitocondrial de 16 espécies de animais obtidos a partir do Centro Nacional de Biotecnologia da Informação (NCBI), sendo publicado nos congressos International Biometric Conference (IBC) e Workshop Escola de Computação e Informação Quântica (WECIQ). Os resultados mostraram que essa probabilidade é eficiente para reconhecer os padrões das sequências de DNA das espécies investigadas (BARROS, 2010a; BARROS, 2010b). Em outro artigo Barros et al analisaram os primeiros 15 mil nucleotídeos do DNA mitocondrial de 7 plantas obtidos a partir do NCBI, os resultados também mostraram eficiência no reconhecimento dos padrões das sequências de DNA das plantas analisadas (BARROS, 2010c), o mesmo foi publicado no congresso Região Brasileira da Sociedade Internacional de Biometria (RBRAS). Neste trabalho aplicaremos esta técnica para reconhecimento quântico de padrões de DNA de abelhas sem ferrão *Melipona quinquefasciata* para comparar com o método clássico utilizado por Pereira (PEREIRA, 2006).

2 Justificativa e Objetivos

2.1 Justificativa

A dificuldade de se desenvolver algoritmos quânticos propicia o uso de técnicas alternativas à solução de problemas puramente algorítmica, como por exemplo o aprendizado de máquina (BANG, 2008), (BEHRMAN, 2008), (HUNZIKER, 2003) e algoritmos genéticos (VENTURA, 2000), (THESS, 2003), (GIRALDI, 2004), (TORONTO, 2006).

2.2 Objetivo Geral

Esta dissertação tem como objetivo científico geral aplicar o método de reconhecimento quântico de padrões para verificar a diversidade genética das abelhas sem ferrão *Melipona quinquefasciata* obtidas de várias colônias silvestres, em localidades distintas da Chapada do Araripe-CE, Chapada da Ibiapaba-CE, cidade do Canto do Buriti-PI e Luziânia-GO.

2.3 Objetivos Específicos

Especificamente queremos:

1. Investigar propriedades e potencialidades dos métodos de reconhecimento quântico de padrões (TRUGENBERGER, 2001) para verificar a diversidade genética das abelhas sem ferrão *Melipona quinquefasciata*;
2. Comparar os resultados obtidos com os obtidos pelos métodos clássicos por Pereira.

Entre as principais contribuições esperadas temos:

- Implementações eficientes de algoritmos de reconhecimento quântico de padrões;

- Novas aplicações da computação quântica em reconhecimento de padrões biométricos e agrários com vistas a novos produtos comerciais.

3 Revisão de Literatura

3.1 Mecânica Quântica

A mecânica quântica é a teoria física que obtém sucesso no estudo dos sistemas físicos cujas dimensões são próximas ou abaixo da escala atômica, tais como moléculas, átomos, elétrons, prótons e de outras partículas subatômicas, muito embora também possa descrever fenômenos macroscópicos em diversos casos. A Mecânica Quântica é um ramo fundamental da física com vasta aplicação, podendo afirmar que é a mais bem sucedida teoria em física. A palavra "quântica" (do Latim, *quantum*) quer dizer quantidade. Na mecânica quântica, esta palavra refere-se a uma unidade discreta que a teoria quântica atribui a certas quantidades físicas, como a energia de um elétron contido num átomo em repouso. A mecânica quântica é a base teórica e experimental de vários campos da Física e da Química, incluindo a física da matéria condensada, física do estado sólido, física atômica, física molecular, química computacional, química quântica, física de partículas, e física nuclear. Os alicerces da mecânica quântica foram estabelecidos durante a primeira metade do século XX por Albert Einstein, Werner Heisenberg, Max Planck, Louis de Broglie, Niels Bohr, Erwin Schrödinger, Max Born, John von Neumann, Paul Dirac, Wolfgang Pauli, Richard Feynman entre outros.

A partir dos anos 80 começaram a surgir ideias que misturavam conceitos de mecânica quântica com teoria da computação, dando origem a uma área de pesquisa física chamada de computação quântica. Um computador quântico é um dispositivo que executa cálculos fazendo uso direto de propriedades da mecânica quântica, tais como sobreposição e emaranhamento. O princípio básico da computação quântica é que as propriedades quânticas podem ser usadas para representar os dados e realizar operações sobre esses dados. Na mecânica quântica, a informação quântica é a informação física que é realizada no "estado" de um sistema quântico. A unidade mais popular de informação quântica é o qubit, um sistema quântico de dois níveis. A informação quântica difere da informação clássica em vários aspectos, dentre os quais temos: não pode ser lido sem que o estado

torne o valor medido, um estado arbitrário não pode ser clonado, o estado pode estar em uma superposição de valores da base. No entanto, apesar disso, a quantidade de informações que podem ser recuperados em um único qubit é igual a um bit. É no processamento de informação (computação quântica), que ocorre uma diferença. Para fins didáticos denomina-se Informação Quântica a identificação e o estudo dos recursos quânticos utilizáveis na área da informação, e Computação Quântica a aplicação direta desses recursos em chaves lógicas, algoritmos, etc (SHANKAR, 1994).

A mecânica quântica é construída sobre quatro postulados (NIELSEN, 2000):

1. O estado de um sistema: o estado de um sistema quântico é um vetor $|\psi(t)\rangle$ no espaço de Hilbert. Este estado contém toda informação que pode ser obtida pelo sistema. Trabalha-se com estados normalizados tal que $\langle\psi|\psi\rangle = 1$, o qual chama-se de vetores de estado.

2. Quantidades observáveis representadas por operadores: para toda variável dinâmica A que é uma quantidade fisicamente mensurável, corresponde um operador A. O operador A é Hermitiano e seus autovetores formam uma base ortonormal completa do espaço vetorial.

3. Medidas: são representadas por um conjunto de operadores de medidas $\{M_m\}$, onde o índice m refere-se aos possíveis resultados da medida. A probabilidade de que o resultado m seja encontrado em uma medida feita em um sistema quântico preparado no estado $|\psi\rangle$ é dada por:

$$p(m) = \left\langle \psi | M_m^\dagger M_m | \psi \right\rangle \tag{3.1}$$

4. A evolução temporal de um sistema quântico isolado: não interage com sua vizinhança, se dá através de transformações unitárias:

$$|\psi(t)\rangle = U(t)\,|\psi(0)\rangle \tag{3.2}$$

em que $U^\dagger U = I$, sendo I a matriz identidade.

3.2 Bits e qubits

A unidade de informação clássica é o bit. Um bit pode ter os valores lógicos “0” ou “1”. Analogamente, a unidade de informação quântica é o bit quântico, ou qubit. Um qubit pode ter os valores lógicos “0”, “1” ou qualquer superposição deles. Fisicamente, qubits são representados por qualquer objeto quântico que possua dois autoestados bem distintos (CARDONHA, 2004).

Os autoestados de um qubit são representados pelos seguintes *kets* (*estados, vetores*) (NIELSEN, 2000):

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{3.3}$$

O conjunto $\{|0\rangle, |1\rangle\}$ forma uma base no espaco de Hilbert de duas dimensões, chamada de uma base computacional. No caso de um spin 1/2 representar o qubit, identificamos $|0\rangle \equiv |\uparrow\rangle$ e $|1\rangle \equiv |\downarrow\rangle$

O estado genérico de um qubit é representado por (NIELSEN, 2000):

$$|\psi\rangle = a|0\rangle + b|1\rangle \tag{3.4}$$

em que a e b são números complexos, $|a|^2 + |b|^2 = 1$. Este estado pode ser parametrizado por ângulos θ e ϕ fazendo-se $a \equiv cos\theta/2$ e $b \equiv exp(i\phi)sin\theta/2$:

$$|\psi\rangle = cos\frac{\theta}{2}|0\rangle + e^{i\phi}sin\frac{\theta}{2}|1\rangle \tag{3.5}$$

Esta representação permite que o estado de um qubit seja vizualizado como um ponto sobre a superfície de uma esfera. Tal esfera é chamada de esfera de Bloch (Figura 3.1). Pontos especiais sobre a esfera de Bloch são mostrados na Tabela 3.1 (NIELSEN, 2000).

Tabela 3.1: Pontos especiais sobre a esfera de Bloch

θ	ϕ	$\|\psi\rangle$	Observação
0	0	$\|0\rangle$	pólo norte da esfera de Bloch
π	0	$\|1\rangle$	pólo sul da esfera de Bloch
$\pi/2$	0	$(\|0\rangle + \|1\rangle)/\sqrt{2}$	equador sobre o eixo x
$\pi/2$	$\pi/2$	$(\|0\rangle + i\|1\rangle)/\sqrt{2}$	equador sobre o eixo y

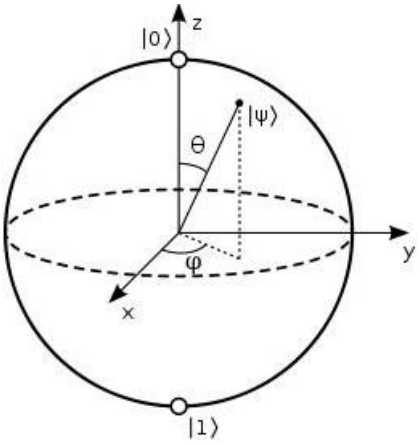

Figura 3.1: Representação da esfera de Bloch de um qubit

3.3 Matrizes e operadores

Um operador é uma regra matemática que pode ser aplicada em uma função para transformá-la em outra. Os operadores são frequentementes indicados pela colocação de um caractere acima do símbolo para denotar o operador. Por exemplo, podemos definir o operador derivada por (NIELSEN, 2000):

$$\hat{D} = \frac{d}{dx}$$

Esta idéia pode ser estendida para espaços vetoriais. Neste caso, um operador $\hat{A}$ é uma regra matemática que transforma um *ket* em outro *ket* que chamamos de $|\phi\rangle$:

$$\hat{A}\,|\psi\rangle = |\phi\rangle$$

3.3.1 Operadores Pauli

Os operadores Pauli são um conjunto de operadores de fundamental importância em computação quântica. Existem quatro operadores Pauli (NIELSEN, 2000):

- O operador identidade (σ_0 ou *I*),

$$\sigma_0\,|0\rangle = |0\rangle\,, \qquad \sigma_0\,|1\rangle = |1\rangle$$

- O operador *CNOT* (σ_1 ou σ_x ou X),

$$\sigma_1 |0\rangle = |1\rangle, \qquad \sigma_1 |1\rangle = |0\rangle$$

- O operador (σ_2 ou σ_y ou Y),

$$\sigma_2 |0\rangle = -i|1\rangle, \qquad \sigma_2 |1\rangle = i|0\rangle$$

- O operador (σ_3 ou σ_z ou Z),

$$\sigma_3 |0\rangle = |0\rangle, \qquad \sigma_3 |1\rangle = |-1\rangle$$

A representação matricial dos operadores Pauli são dados por:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

3.3.2 Operador Hadamard e rotação

Matriz Hadamard:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

temos, $H|0\rangle = \frac{1}{\sqrt{2}}[|0\rangle + |1\rangle]$ e $H|1\rangle = \frac{1}{\sqrt{2}}[|0\rangle - |1\rangle]$.

Matriz de Rotação:

$$R(\phi) = \begin{pmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{pmatrix}$$

temos, $R(\phi)|0\rangle = cos(\phi)|0\rangle + sin(\phi)|1\rangle$ e $R(\phi)|1\rangle = cos(\phi)|1\rangle - sin(\phi)|0\rangle$.

3.3.3 Operadores Hermitiano, Unitário e Normal

Um operador A é dito hermitiano se (NIELSEN, 2000),

$$\hat{A} = \hat{A}^{\dagger}$$

em que $A^{\dagger}$ é a transposta conjugada de A. Os operadores hermitianos possuem as seguintes propriedades: os autovalores são reais e os autovetores ortogonais.

O operador A é dito unitário se,

$$A^{\dagger}A = AA^{\dagger} = I$$

em que I é a matriz identidade. O operador unitário preserva seu comprimento e ângulo entre os vetores, e pode ser considerado como um tipo de operador rotação no espaço vetorial abstrato. Como os operadores hermitianos, seus autovetores são ortogonais. No entanto, seus autovalores não são necessariamente reais.

Um operador A é dito normal se,

$$A^{\dagger}A = AA^{\dagger}$$

Os operadores normais são diagonalizáveis, em outras palavras a comutatividade é preservada.

3.4 Misturas Estatísticas e Matriz Densidade

Em Computação Quântica e Informação Quântica frequentemente temos de lidar com situações em que o vetor de estado do sistema não é conhecido, mas apenas um conjunto possível de vetores $\{|\psi_i\rangle\}$, que ocorrem com probabilidades $\{p_i\}$. O conjunto $\{p_i, |\psi_i\rangle\}$ é chamado de *ensemble* estatístico. A ferramenta matemática adequada para tratar desses casos é a matriz densidade, ρ, definida como (NIELSEN, 2000):

$$\rho \equiv \sum_i p_i |\psi_i\rangle \langle\psi_i| \tag{3.6}$$

em que $p_i > 0$ e $\sum_i p_i = 1$. Algumas propriedades deste operador:

1. A matriz densidade é um operador positivo, ou seja, possui autovalores reais não-

negativos.

2. O traço de ρ é igual a 1.

3. O estado será puro se, e somente se, $Tr(\rho) = 1$ (traço de ρ).

3.5 Chaves Lógicas e Circuitos Quânticos

Uma operação lógica ou até mesmo um algoritmo pode ser descrito por um diagrama denominado circuito quântico. Estes determinam quais e em que ordem as chaves lógicas são aplicadas a um ou mais qubits de um sistema. Os circuitos quânticos são compostos de linhas e símbolos. As linhas representam os qubits (uma linha para cada qubit), necessários para realizar uma determinada operação e os símbolos representam as chaves lógicas. Por sua vez, as chaves lógicas (descritas por caixas, com letras) descrevem um conjunto de operações quânticas aplicadas a um ou mais qubits (NIELSEN, 2000).

A operacao $CNOT_a$ (não controlado) pode ser descrita pelo circuito quântico ilustrado na Figura 3.2, ($0 \oplus 0 = 0; 0 \oplus 1 = 1; 1 \oplus 0 = 1; 1 \oplus 1 = 0$). Esta é uma porta lógica controlada, que inverterá o qubit b se o $a = 1$.

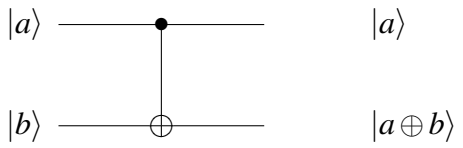

Figura 3.2: Circuito quântico que descreve a operação lógica $CNOT_a$

A matriz que representa a porta CNOT com controle no primeiro qubit (qubit A) é dada por:

$$CNOT_A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{3.7}$$

É fácil verificar que $CNOT_A |00\rangle = |00\rangle$, $CNOT_A |01\rangle = |01\rangle$, $CNOT_A |10\rangle = |11\rangle$, $CNOT_A |11\rangle = |10\rangle$.

As operações de um qubit são representadas por caixas, com o nome da chave lógica

que deve ser implementada, como encontra-se ilustrado na Figura 3.3, onde os símbolos representam as matrizes de Pauli (X, Y e Z), a porta de fase (S) e a porta $\pi/8(T)$.

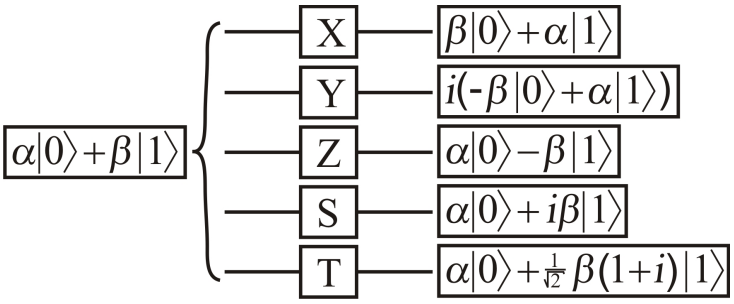

Figura 3.3: Circuitos quânticos que descrevem as operações lógicas de um qubit, X, Y, Z, S (porta de fase) e T (porta $\pi/8$)

Uma operação importante é a operação de troca (SWAP), onde os estados dos qubits são trocados, ou seja o estado de a passa para b e vice versa, como descrito pela equação:

$$SWAP\,|ab\rangle = |ba\rangle \tag{3.8}$$

Esta operação pode ser descrita pelo circuito quântico ilustrado na Figura 3.4 e, como pode ser visto, esta consiste da aplicação de três portas lógicas CNOT, como pode ser visto na sequência descrita pelas equações:

$$CNOT_a\,|a,b\rangle = |a,b\ominus a\rangle$$

$$CNOT_b\,|a,a\ominus b\rangle = |a\oplus a\ominus b,a\oplus b\rangle = |b,a\ominus b\rangle \tag{3.9}$$

$$CNOT_a\,|b,a\ominus b\rangle = |b,a\oplus b\ominus b\rangle = |b,a\rangle$$

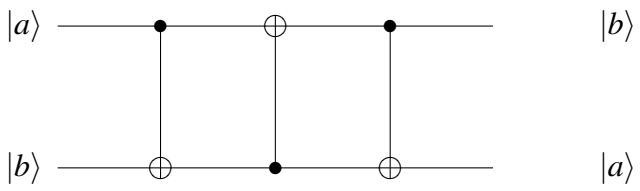

Figura 3.4: Circuito quântico que descreve a operação lógica U_{SWAP}, que troca os estados de dois qubits

3.6 Paralelismo Quântico

Considere um registrador com n qubits no estado:

$$|\psi\rangle := \sum_{i=0}^{2^n-1} a_i |i\rangle \tag{3.10}$$

A aplicação de uma matriz unitária U de dimensão 2^n sobre este registrador produz:

$$|\psi^r\rangle = U|\psi\rangle = U\left(\sum_{i=0}^{2^n-1} a_i |i\rangle\right) = \sum_{i=0}^{2^n-1} a_i U|i\rangle \tag{3.11}$$

isto é, uma única aplicação de U realiza um número exponencial de operações em estados básicos. Este fenômeno é chamado de paralelismo quântico. Usualmente, a matriz de Hadamard, é utilizada para gerar num registrador um estado quântico contendo a superposição de todos os estados básicos, todos com a mesma amplitude.

Seja $H_n := \bigotimes_{j=1}^{n} H$, onde H é a matriz de Hadamard. Chame H_n de matriz de Hadamard de ordem 2^n. Note que a aplicação de H_n a um registrador é feita simplesmente pela aplicação de H a cada qubit individual do registrador.

3.7 Algoritmo de Grover

Dado um conjunto desordenado de m elementos, encontrar um elemento em particular, classicamente, na pior das hipóteses leva m consultas. Em média, vamos encontrar o desejado elemento em $m/2$ consultas. O algoritmo de busca de Grover encontra em $\sqrt{m}$ consultas, tem muitas aplicações à teoria de banco de dados e outras áreas. Dado uma função $f : \{0,1\}^n \to \{0,1\}$ e que existe exatamente uma sequência binária x_0 tal que (YANOFSKY, 2008)

$$f(x) = \begin{cases} 1, & x = x_0 \\ 0, & x \neq x_0 \end{cases} \tag{3.12}$$

pede-se para encontrar x_0. Classicamente, na pior das hipóteses, teríamos que avaliar todas as 2^n cadeias binárias para encontrar o x_0 desejado. O algoritmo de Grover, exigirá apenas $\sqrt{2^n}$ avaliações.

Em termos de matrizes o circuito da Figura 3.5 pode ser escrito como $U_f(H^{\otimes n} \otimes I)|\mathbf{0},0\rangle$. Os estados são:

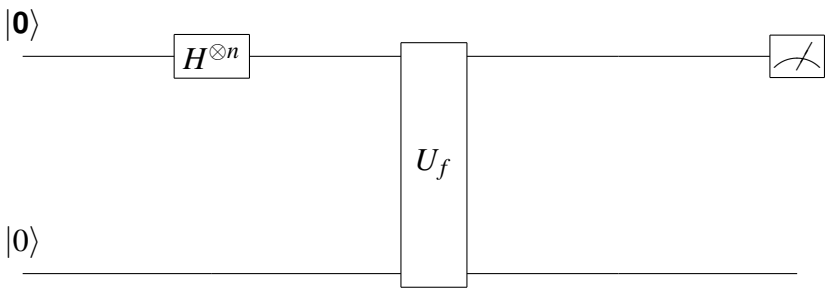

Figura 3.5: Circuito quântico que descreve a primeira tentativa do algoritmo de Grover

$$|\phi_0\rangle = |\mathbf{0}, 0\rangle \tag{3.13}$$

$$|\phi_1\rangle = \left[\frac{\sum_{x\in\{0,1\}^n} |\mathbf{x}\rangle}{\sqrt{2^n}}\right] |0\rangle \tag{3.14}$$

$$|\phi_2\rangle = \frac{\sum_{x\in\{0,1\}^n} |\mathbf{x}, f(\mathbf{x})\rangle}{\sqrt{2^n}} \tag{3.15}$$

Medindo os qubits superiores, com igual probabilidade, encontramos um dos 2^n binários. Medindo o qubit inferior teremos $|0\rangle$ com probabilidade $\frac{2^n-1}{2^n}$ e $|1\rangle$ com probabilidade 2^{-n}. O algoritmo de busca de Grover usa dois truques. O primeiro, chama-se *inversão de fase* (Figura 3.6), altera a fase do estado desejado. Ele trabalha como segue. Tome U_f e coloque o qubit inferior na superposição do estado (YANOFSKY, 2008)

$$\frac{|0\rangle - |1\rangle}{\sqrt{2}}. \tag{3.16}$$

Para um *x* arbitrário, temos o circuito:

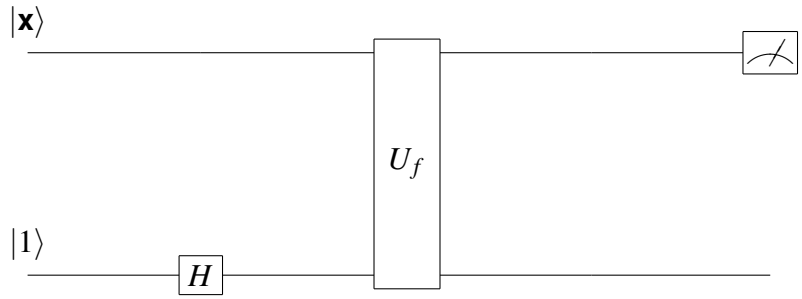

Figura 3.6: Circuito quântico que descreve a inversão de fase no algoritmo de Grover.

Em termos de matrizes, $U_f(I_n \otimes H)\,|x, 1\rangle$. Porque U_f e H são operações unitárias, é óbvio que a inversão de fase é uma operação unitária. Os estados são (YANOFSKY, 2008)

$$|\phi_0\rangle = |\mathbf{x}, 1\rangle \tag{3.17}$$

$$|\phi_1\rangle = |\mathbf{x}\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right] = \left[\frac{|x,0\rangle - |x,1\rangle}{\sqrt{2}}\right] \tag{3.18}$$

$$|\phi_2\rangle = |\mathbf{x}\rangle \left[\frac{|f(x)\oplus 0\rangle - |f(x)\oplus 1\rangle}{\sqrt{2}}\right] = |\mathbf{x}\rangle \left[\frac{|f(x)\rangle - \overline{|f(x)\rangle}}{\sqrt{2}}\right] \tag{3.19}$$

Lembrando que $a-b=(-1)(b-a)$ pode-se reescrever (YANOFSKY, 2008):

$$|\phi_2\rangle = (-1)^{f(x)} |\mathbf{x}\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right] = \begin{cases} -1\,|\mathbf{x}\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right], & x = x_0 \\ +1\,|\mathbf{x}\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right], & x \neq x_0 \end{cases} \tag{3.20}$$

O segundo truque é chamado de *inversão sobre a média* ou *inversão em torno da média*. Esta é uma maneira de impulsionar a separação das fases. Para isto utilizamos a seguinte operação: $-I+2A$, em que A é a matriz das médias.

Etapas do algoritmo de Grover (Figura 3.7) (YANOFSKY, 2008):

Algoritmo 1: Algoritmo de Grover
1: Inicia com o estado $\|\mathbf{0}\rangle$.
2: Aplica $H^{\otimes n}$.
3: Repete $\sqrt{2^n}$ vezes:
3a: Aplica a operação de inversão de fase: $U_f(I \otimes H)$.
3b: Aplica a operação de inversão sobre a média: $-I+2A$.
4: Mede os qubits.

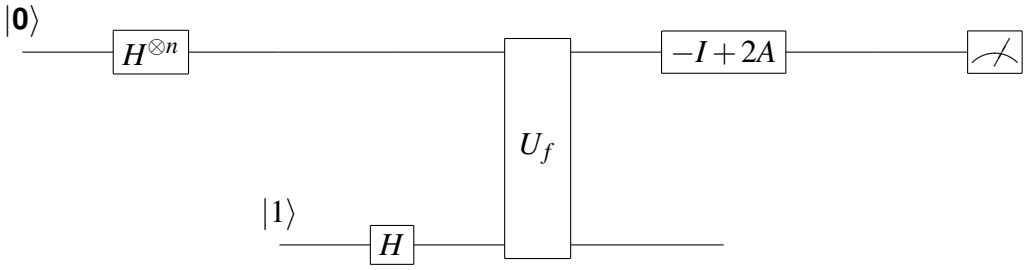

Figura 3.7: Circuito quântico que descreve o algoritmo de Grover

4 Materiais e Métodos

4.1 Banco de dados utilizados

4.1.1 *Melipona quinquefasciata*

As abelhas sem ferrão do gênero *Melipona* é exclusivo da região neotropical, onde mais de 40 espécies são encontradas na região equatorial, regiões tropicais e subtropicais do continente americano. O Nordeste do Brasil, a partir do estado do Maranhão para o estado da Bahia, conta com cerca de 23% das espécies conhecidas de *Melipona* - *M. asilvai, M. compressipes fasciculata, M. mandaçaia, M. marginata, M. quadrifasciata anthidioides, M. rufiventris, M. scutellaris, M. seminigra pernigra* e *M. subnitida* (LIMA-VERDE, 2002).

Melipona quinquefasciata (Figura 4.1), é uma espécie pertencente ao subgênero *Melikerria* e descrito por Lepeletier em 1836 (MOURE, 1975), que não está incluída entre as espécies relatadas para ocorrer no Nordeste do Brasil. A distribuição geográfica de *M. quinquefasciata* mostra sua ocorrência apenas em estados do sul do Brasil, do Sul do Espírito Santo ao Rio Grande do Sul, incluindo áreas de Minas Gerais, Goiás, Mato Grosso, Mato Grosso do Sul, Bolívia, na sua parte do Sul; Paraguai e Norte-Nordeste de Argentina (MARIANO-FILHO, 1911; SCHWARZ, 1932; KERR, 1948; MOURE, 1948, 1975; VIANA, 1987).

Até 2002, o Nordeste brasileiro, que representa 18,3% do território brasileiro, contava, com 23% das espécies de *Melipona* conhecidas (*M. asilvai, M. compressipes fasciculata, M. mandaçaia, M. marginata, M. quadrifasciata anthidioides, M. rufiventris, M. scutellaris, M. seminigra pernigra* e *M. subnitida*). Entretanto, a partir de relatos de agricultores e extrativistas de mel do Estado do Ceará, ocorreram as primeiras informações sobre a presença da “uruçú do chão” *M. quinquefasciata*, nas chapadas do Araripe, no sul do estado, bem como na chapada da Ibiapaba, localizada no oeste do Ceará (LIMA-VERDE, 2002).

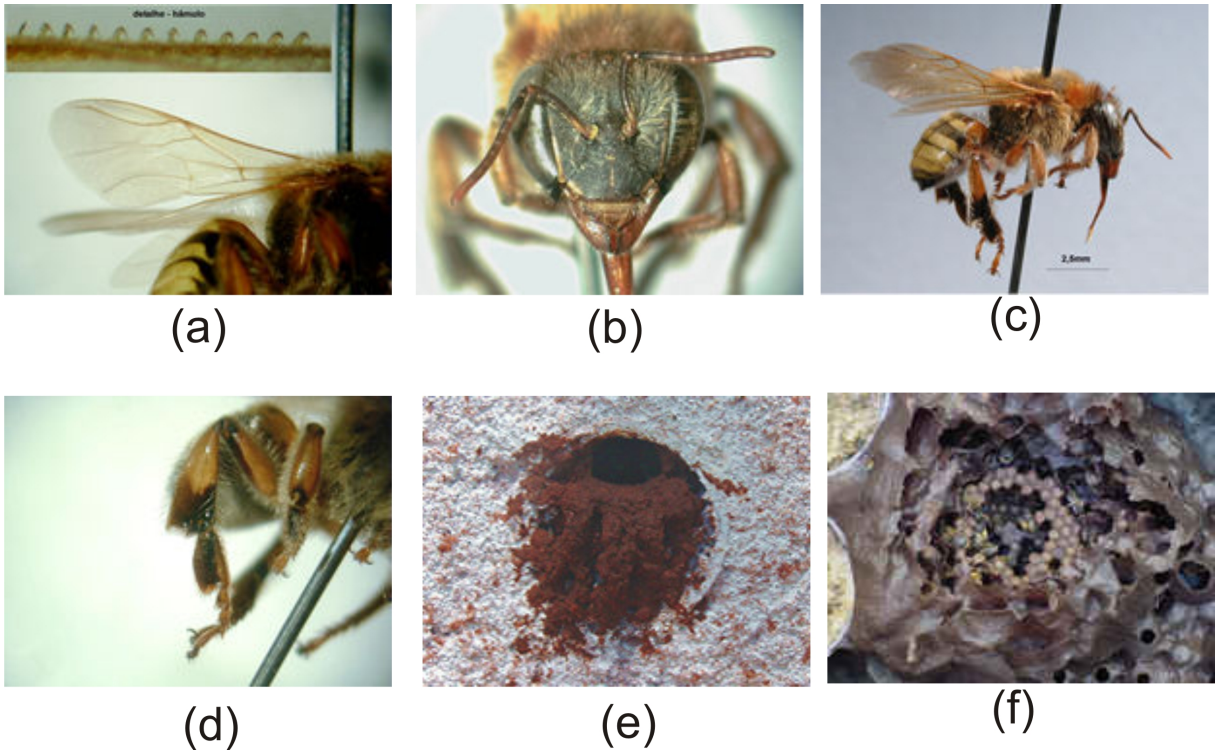

Figura 4.1: Melipona quinquefasciata: (a) asa; (b) vista frontal; (c) vista lateral; (d) pata posterior; (e) entrada do ninho; (f) interior do ninho

Esta espécie de *Melipona* possui a particular característica de nidificar sob o solo, podendo suas câmaras chegarem a uma profundidade de até 4 metros (KERR et.al, 2001; NOGUEIRA-NETO, 1997), utilizando-se de formigueiros ou cupinzeiros abandonados. O mel dessa abelha é bastante saboroso, e embora de difícil coleta, tem comércio garantido, sendo então cobiçado pelos meleiros do local. A extração do mel leva ainda à morte da colônia, já que nesta época a rainha está em seu período fértil, estando com sobrepeso, não permitindo seu voo de fuga, para uma futura formação de colônia, como acontece com outras espécies (KERR et.al, 2001). Possui um comprimento de 9 a 10,5mm, com cinco faixas de coloração amarelada no abdômen. Infelizmente, a *M. quinquefasciata* já consta em listas de espécies em extinção, como mencionado no Livro Vermelho da Fauna Ameaçada no Estado do Paraná (www.maternatura.org.br).

Todas as informações existentes sobre essa abelha em cativeiro foram obtidas nas regiões Sul e Sudeste (NOGUEIRA-NETO, 1997), não se sabendo se podem ser aplicadas às populações nordestinas, considerando as diversidades ecológicas das regiões e o isolamento das duas populações dessa espécie de abelha.

Estudos realizados com espécies de abelhas sem ferrão no Brasil tem sido capaz de separar em diferentes populações as espécies que se acreditava pertencer a uma única espécie, como é o caso de *M. rufiventris* e *M. mondury*. No entanto, as características

morfológicas nem sempre são suficientes para a distinção de confiança entre as espécies. Felizmente, a genética molecular surgiu como uma ferramenta útil. O sequenciamento completo do primeiro espaçador interno transcrito (ITS1) do DNA ribossômico nuclear de três espécies *Melipona* foi determinada por Fernandes-Salomão et. al (2005). As relações entre as oito espécies deste gênero foram inferidas a partir de sequências ITS1 parcial, demonstrando o potencial utilidade filogenética desta região (FERNANDES-SALOMÃO et. al, 2005). Mais recentemente, a utilidade deste espaçador para estudos intra-específica em Melipona também foi avaliada em um estudo utilizando amostras de *M. subnitida* de diferentes localidades do nordeste do Brasil (CRUZ et. al, 2006). Pereira et al. sequenciaram e compararam um segmento de ITS1 de *M. quinquefasciata* coletadas nos estados do Ceará, Piauí e Goiás, a fim de ajudar a determinar se as abelhas encontradas recentemente no nordeste do Brasil são, na verdade uma população de uma espécie descontínua, uma sub-espécie ou mesmo uma nova espécie. A análise filogenética demonstrou que todas as sequências ITS1 a partir de amostras do Nordeste juntamente com a amostra de Goiás pertencem a espécie de *M. quinquefasciata* (PEREIRA, 2009).

4.1.2 O DNA ribossômico nuclear

O ácido desoxirribonucleico (ADN, em português, ou DNA, em inglês: deoxyribonucleic acid), é um composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus. Os segmentos de DNA que são responsáveis por carregar a informação genética são denominados genes. O DNA é responsável pela transmissão das características hereditárias de cada ser vivo. A estrutura da molécula de DNA foi descoberta conjuntamente pelo estadunidense James Watson e pelo britânico Francis Crick em 7 de Março de 1953, o que lhes valeu o Prêmio Nobel de Fisiologia/Medicina em 1962, juntamente com Maurice Wilkins (BRUCE, 2008).

Os nucleotídeos são moléculas que, quando juntas, constituem a unidades estruturais do RNA e DNA. Há quatro bases de nucleotídeos de uma fita de DNA - adenina (A), citosina (C), guanina (G), timina (T). As bases purina são a guanina e a adenina, e as bases pirimidina são a citosina e a timina. Devido a isto, temos que G e A são bioquimicamente semelhantes. O mesmo ocorre com as bases C e T. (Figura 4.2). As formas e estrutura química das bases permitem ligações de hidrogênio para formar com eficiência apenas entre A e T e entre G e C (Figura 4.3) (BRUCE, 2008).

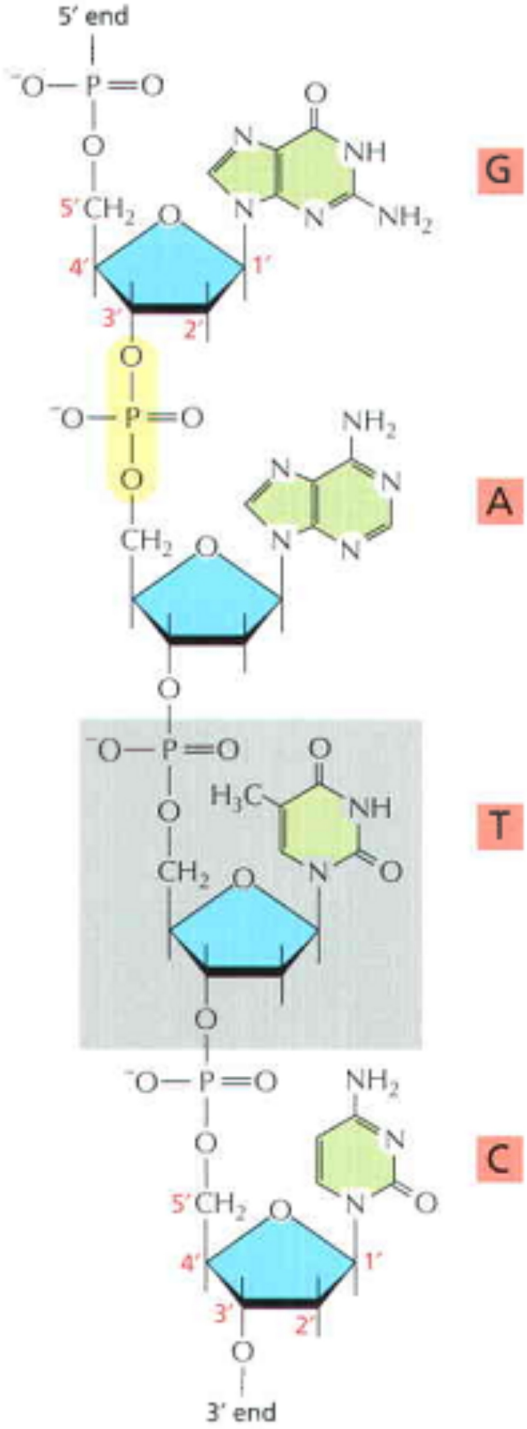

Figura 4.2: Representação de uma pequena parte de uma cadeia de uma molécula de DNA

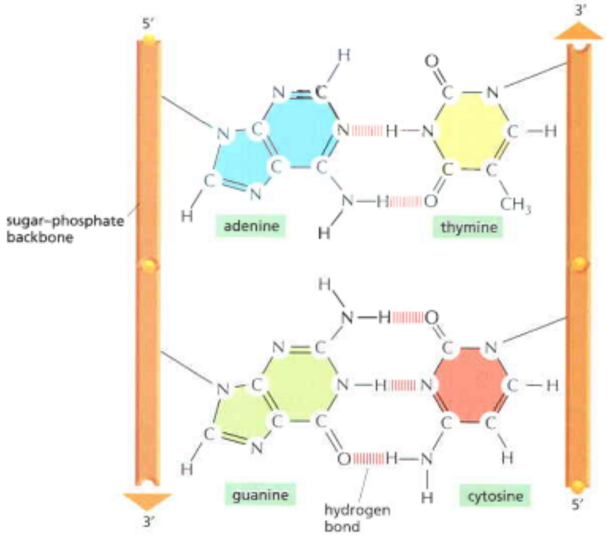

Figura 4.3: Pares de bases complementares na dupla hélice do DNA

O DNA ribossômico (DNAr) é uma sequência de DNA contida nos cromossomos do nucléolo que codifica RNA ribossômico. Estas sequências regulam a transcrição e iniciação da amplificação e contêm segmentos espaçadores transcritos e não-transcritos. O baixo nível de polimorfismo na unidade de transcrição de DNAr permite a caracterização de cada espécie usando só uns poucos exemplares e faz que este DNA seja útil para a comparação inter-específica.

As sequências de DNA evoluem de maneira mais regular do que caracteres fenotípicos (morfológicos, fisiológicos e embriológicos) (VARELA, 2003). Dentre as vantagens dos marcadores moleculares em relação aos caracteres morfológicos, podemos citar: possibilidade de serem classificados objetivamente eliminando o problema da subjetividade, e os marcadores serem obtidos em grandes quantidades e poderem ser utilizados em estudos de relações filogenéticas em diversos níveis (AMORIM, 2002). A construção de filogenias requer a escolha da molécula apropriada para o sequenciamento e a aplicação de uma metodologia para inferir a informação filogenética por meio da comparação das sequências (MATIOLI, 2001).

A sequência do gene nuclear que codifica para o rRNA da subunidade pequena do complexo ribossômico tem sido muito utilizada como locus apropriado para inferir relações entre diversos organismos (LIMA, 2003). Estudando sequências moleculares para inferências, Broccheri (2001), identificou que o gene que codifica para a subunidade pequena do RNA ribossômico (SSU rRNA , "*small subunit rRNA*") tem baixa taxa de mutação, é resistente à transferência lateral de genes, funcionalmente conservado e amplamente distribuído entre os organismos, portanto, possuindo características imprescindíveis para a

sua utilização em filogenia molecular (BROCCHERI, 2001).

Os genes codificadores da região do rRNA são essenciais para o organismo, por isso evoluem de maneira mais lenta. Sendo assim modificações pontuais em sítios específicos podem determinar a inativação da função da subunidade comprometendo a síntese protéica. O 18S RNA ribossomal (18S rRNA) é uma parte do RNA ribossomal. O S em 18S representa unidades Svedberg. A subunidade menor (SSU) do gene 18S rRNA é um dos genes mais frequentemente utilizados em estudos filogenéticos, pois sofre menores taxas de mutação.

ITS refere-se a um pedaço de RNA não-funcional situado entre estruturas RNAs ribossomais (rRNA) em uma transcrição precursor comum. As regiões ITS apresentam-se muito variáveis entre os diferentes organismos. Portanto, as sequências que codificam para essa região possuem grande heterogeneidade, sendo usadas dessa forma para resolver problemas filogenéticos.

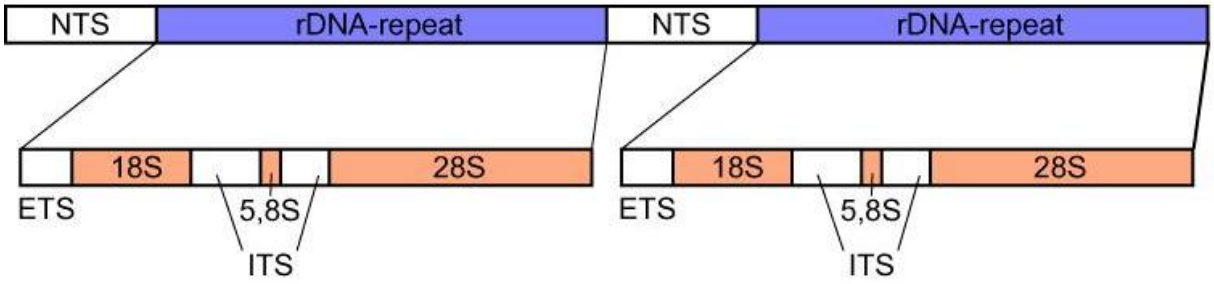

Figura 4.4: Disposição do DNAr 18S e ITS (espaçador transcrito interno)

Os estudos que tratam da biologia molecular de Melipona são muito escassos quando comparados com outros gêneros. As relações filogenéticas e o entendimento de vários aspectos biológicos e evolutivos de uma grande variedade de abelhas, são os principais objetivos desses estudos que desde o final do século passado, têm sido baseados na molécula do DNA. A aplicação do DNA em estudos populacionais e evolutivos se vale devido ao fato deste possuir em algumas regiões específicas, uma alta taxa de substituições de bases, apresentar alterações no tamanho da região estudada, devido à inserções e deleções, quando do confrontamento entre espécimes. Várias metodologias têm sido aplicadas para caracterização do genoma das abelhas para encontrar variabilidade genética entre populações e espécies (MELO, 1999).

4.1.3 Localização do experimento

As abelhas para a análise foram obtidas de várias colônias silvestres, em localidades distintas da Chapada do Araripe-CE, Chapada da Ibiapaba-CE, cidade do Canto do Buriti-PI e Luziânia-GO retiradas da tese de doutorado de Pereira (PEREIRA, 2006). Para a região 18S do nrDNA utilizou-se 6 amostras de tamanho 1823: Flona 2 (Crato-CE), Chapada 3 (Crato-CE), Jardim 1 (Jardim-CE), Goiás (Luziânia-GO), Ubajara (Ubajara-CE) e Araripe 3 (Araripe-CE). Para a região ITS1 parcial utilizou-se 9 amostras de tamanho 491: Flona 1 (Crato-CE), Flona 2 (Crato-CE), Flona 3 (Crato-CE), Jardim 1 (Jardim-CE), Araripe 1 (Araripe-CE), Araripe 3 (Araripe-CE), São Benedito (São Benedito-CE), Goiás (Luziânia-GO) e Piauí 1 (Canto do Buriti-PI).

4.2 Teste de Comparação de médias

Os experimentos inteiramente ao acaso, é o delineamento em que os tratamentos são designados às unidades sem qualquer restrição (de forma aleatória), e este experimento só pode ser conduzido quando as unidades são similares. A similaridade não significa igualdade, mas sim que as unidades respondam aos tratamentos da mesma forma.

A análise de variância (ANOVA) é uma extensão do teste t de Student que compara duas médias. A ANOVA permite a comparação de mais de duas médias. Na Tabela 4.1 está o procedimento para fazer uma ANOVA (VIEIRA, 2006).

Tabela 4.1: Tabela de Análise de variância

Fonte de Variação	gl	SQ	QM	Valor F	p-valor
Tratamenos	$k-1$	$\frac{\sum T^2}{r} - \frac{(\sum y)^2}{n}$	$\frac{SQTr}{k-1}$	$\frac{QMTr}{QMR}$	p-valor
Resíduos	$n-k$	$\sum y^2 - \frac{\sum T^2}{r}$	$\frac{SQM}{n-k}$		

em que y = valores observados; r = número de repetições; k = número de tratamentos e $n = k \times r$. Compara-se o valor calculado de F com o valor crítico de F (Tabela F), ao nível de significância estabelecido. Se o valor de F calculado for maior que o crítico de F, rejeita-se a hipótese de que as médias são iguais. Com base no p-valor, rejeita-se a hipótese de igualdade se o p-valor for menor que o nível de significância estabelecido.

Quando a análise de variância mostra que as médias diferem entre si, utiliza-se o Teste de Tukey para verificar quais as médias que diferem. Para proceder o teste, calcula-se a diferença mínima significativa. No entanto, Tukey chamou esse valor de diferença honestamente significativa (*honestly significant difference*). O valor dessa diferença é calculada

da seguinte forma:

$$\Delta = q\sqrt{\frac{QMR}{r}} \tag{4.1}$$

O valor de q é a amplitude estudentizada encontrado na Tabela de q ao nível de significância estabelecido, associado a quantidade de tratamentos e graus de liberdade do resíduo; QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições de cada tratamento. De acordo com este teste, duas médias são estatisticamente diferentes toda vez que o valor absoluto da diferença entre elas for igual ou maior do que a diferença mínima significativa. Utilizou-se o *software R* para os cálculos da ANOVA e teste de Tukey (VIEIRA, 2006).

4.3 Reconhecimento de Padrões

Reconhecimento de padrões (RP) tem o objetivo de classificar informações (padrões) baseado ou em conhecimento a priori ou em informações estatísticas extraídas dos padrões (entidade à qual se pode dar um nome, por exemplo: sinal de voz; rosto humano; imagem). Um sistema completo de reconhecimento de padrões consiste de um sensor que obtém observações a serem classificadas ou descritas; um mecanismo de extração de características que computa informações numéricas ou simbólicas das observações; e um esquema de classificação das observações, que depende das características extraídas (BISHOP, 2006).

O esquema de classificação é geralmente baseado na disponibilidade de um conjunto de padrões que foram anteriormente classificados, o “conjunto de treinamento”; o resultado do aprendizado é caracterizado como um aprendizado supervisionado. O aprendizado pode também ser não supervisionado, de forma que o sistema não recebe informações a priori dos padrões, estabelecendo então as classes dos padrões através de análise de padrões estatísticos.

Aplicações típicas do reconhecimento de padrões incluem reconhecimento de fala, classificação de documentos em categorias (por exemplo, mensagens de correio eletrônico que são spam ou não), reconhecimento de escrita, reconhecimento de faces, classificação de células de sangue, entre outras.

Categorização das áreas de aplicações:

- **Processamento de documentos**. Reconhecimento de caracteres impressos ou escritos; máquinas de leitura para cegos; leitores de códigos de barras; introdução automática de texto em documentos de processamento de texto; análise de documentos financeiros; compreensão de linguagem natural;

- **Automação industrial**. Inspeção e montagem/configuração de objetos complexos; inspeção de circuitos impressos; inspeção de partes de máquinas; processamento de imagem; visão por computador;

- **Detecção remota** (*Remote sensing*). Observação do planeta através de sensores em satélites ou aviões; previsão da evolução de culturas; planeamento de uso de terras; monitorização ambiental; meteorologia; exploração mineira; mapas topográficos;

- **Medicina e biologia**. Processamento de diversos sinais e imagens médicas; contagem de células no sangue; detecção de tumores em imagens de Raios-X; caracterização de tecidos usando ultra-sons; análise de imagens de cromossomas; interpretação de electrocardiogramas; diagnóstico médico;

- **Identificação de pessoas**. Restrição de acesso em instalações de segurança; reconhecimento de voz; identificação de impressões digitais; reconhecimento de caras;

- **Aplicações científicas**. Interpretação de ondas sísmicas para previsão de terramotos; análise de composição molecular através de imagens de microscópio eletrônico;

- **Aplicações na agricultura**. Direcionamento de equipamento; inspeção de produtos; ordenação e empacotamento de produtos.

A Análise hierárquica de agrupamentos (HCA - Hierarchical Cluster Analysis) é uma ferramenta exploratória que objetiva revelar agrupamentos naturais (*clusters*) dentro de conjuntos que aparentemente não apresenta grupos. Esta técnica é mais útil quando se quer agrupar um pequeno número (algumas centenas) de casos. Os objetos podem ser casos ou variáveis (BISHOP, 2006).

Existem diversas abordagens de *clustering*, tais como: probabilística, otimização, *clumping* e hierárquica. Cada uma utiliza uma maneira diferente para a identificação e representação dos *clusters*. Nesta dissertação, os agrupamentos são obtidos por meio de um algoritmo de *clustering* hierárquico, o qual representa os *clusters* em uma estrutura conhecida como dendograma, que consiste de um tipo especial de árvore na qual os nós pais agrupam os exemplos representados pelos nós filhos. Dessa maneira, um agrupamento hierárquico agrupa os dados de modo que se dois exemplos são agrupados em algum nível, nos níveis mais acima eles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de *clusters*. Essa técnica permite analisar os *clusters* em diferentes níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos. Os objetos que estão próximos um do outro pertencem ao mesmo grupo e se estão distantes pertencem a grupos diferentes. O critério básico para qualquer agrupamento é a distância. Alguns tipos de distâncias:

- **Distância de Hamming**: é assim chamada em homenagem a Richard Hamming, que introduziu o conceito em um artigo fundamental sobre códigos de Hamming *Error detecting and error correcting codes* em 1950. Na teoria da informação, a distância de Hamming entre duas strings de mesmo comprimento é o número de posições nas quais elas diferem entre si. Vista de outra forma, ela corresponde ao menor número de substituições necessárias para transformar uma string na outra, ou o número de erros que transformaram uma na outra.

- **Distância de Jaccard**: O índice de Jaccard, também conhecido como o coeficiente de similaridade de Jaccard (originalmente desenvolvido por Paul Jaccard), é uma estatística usada para comparar a similaridade e diversidade da amostra de conjuntos. O coeficiente de Jaccard mede a similaridade entre conjunto de amostras, e é definido como o tamanho da interseção dividido pelo tamanho da união dos conjuntos da amostra.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{4.2}$$

A distância de Jaccard, que mede a similaridade entre os conjuntos das amostras, é complementar ao coeficiente de Jaccard e é obtido subtraindo-se o coeficiente de Jaccard a partir de 1:

$$J_\delta(A,B) = 1 - J(A,B). \tag{4.3}$$

- **Distância de Canberra**: introduzida em 1966 (LANCE, 1966), é uma função métrica usada frequentemente para dados dispersos em torno de uma origem. A métrica de Canberra é semelhante à distância de Manhattan (que em si é uma forma especial de distância Minkowski). O que difere é que a diferença absoluta entre as variáveis dos dois objetos é dividido pela soma dos valores absolutos a variável antes de somar. A equação generalizada é dada na forma:

$$d_C(i,j) = \sum_{k=0}^{n-1} \frac{|y_{i,k} - y_{j,k}|}{|y_{i,k}| + |y_{j,k}|} \tag{4.4}$$

  Esta é uma forma modificada ligeiramente em relação à forma original dado por Lance (1966) e foi sugerido por Adkins (LANCE, 1967). Na equação d é a distância de Canberra entre dois objetos i e j, k é o índice de uma variável e n é o número total de variáveis y.

- **Distância de Jukes-Cantor**: No modelo de Jukes-Cantor desenvolvido em 1969, todas as bases tem igual probabilidade de ocorrência e a probabilidade de transição entre bases diferentes é igual em todos os pares possíveis. O calculo das distâncias é realizado a partir da seguinte fórmula:

$$\hat{d_T} = -\frac{3}{4} ln \left(1 - \frac{4}{3} p\right) \tag{4.5}$$

O método de agrupamento utilizado foi a vizinho mais próximo ou ligação simples, as conecções entre objetos e grupos ou entre grupos são feitas por ligações simples entre pares de objetos, ou seja, a distância entre os grupos é definida como sendo aquela entre os objetos mais parecidos entre esses grupos. Este método leva a grupos longos se comparados aos grupos formados por outros métodos de agrupamento.

4.4 Reconhecimento Quântico de Padrões

Trugenberger (2002) mostra que o emaranhamento quântico fornece um mecanismo natural para melhorar a capacidade de armazenamento de memórias associativas. Na verdade, o número de padrões binários que podem ser armazenados em tal memória quântica é exponencial no número n de qubits, $p_{max} = 2^n$, ou seja, é ideal no sentido de que todos os padrões binários que podem ser formados com os n bits podem ser armazenados. Dado p_i padrões binários p de tamanho n, não é difícil imaginar como uma memória quântica pode armazená-las. Na verdade, essa memória é fornecida pela seguinte superposição de n qubits emaranhado (TRUGENBERGER, 2002):

$$|m\rangle = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} |p\rangle^i \tag{4.6}$$

A única questão é como gerar este estado unitariamente a partir de um simples estado inicial de n qubits. Para isto, pode-se usar o algoritmo proposto em (VENTURA, 1999). Aqui, porém, propõe-se uma versão simplificada. Na construção $|m\rangle$ usa-se três registros: um primeiro registro p de n qubits em que alimenta subsequentemente os padrões p_i a ser armazenados; um registro u de dois qubits preparado no estado $|01\rangle$; e outro registro m de n qubits para manter a memória. Este último é preparado inicialmente no estado de $|0_1,\ldots,0_n\rangle$. O estado quântico total inicial é,

$$|\psi_0^1\rangle = |p_1^1,\ldots,p_n^1;01;0_1,\ldots,0_n\rangle \tag{4.7}$$

A idéia do algoritmo de armazenamento é separar esse estado em dois termos, uma correspondente aos padrões já armazenados e outro pronto para processar um novo padrão. Estas duas partes se distinguem pelo estado do utilitário do segundo qubit u_2: $|0\rangle$ para os padrões armazenados e $|1\rangle$ durante o período de processamento. Para cada padrão p_i a ser armazenado um tem de realizar as operações descritas abaixo (TRUGENBERGER, 2001):

$$|\psi_1^i\rangle = \prod_{j=1}^{n} 2XOR_{p_j^i u_2 m_j} |\psi_0^i\rangle \tag{4.8}$$

Isto simplesmente copia o padrão p_i para o registro da memória do processamento, identificadas por $|u_2\rangle = |1\rangle$.

$$|\psi_2^i\rangle = \prod_{j=1}^{n} NOT_{m_j} XOR_{p_j^i m_j} |\psi_1^i\rangle \tag{4.9}$$

$$|\psi_3^i\rangle = nXOR_{m_1 \ldots m_n u_1} |\psi_2^i\rangle \tag{4.10}$$

A primeira dessas operações faz com que todos os qubits da memória registre $|1\rangle$'s quando o conteúdo do padrão e do registro da memória são idênticos, o que é exatamente o caso apenas para o período de processamento. Juntas, essas duas operações muda o primeiro qubit utilitário u_1 do período de processamento para $|1\rangle$, deixando inalterado os padrões armazenados.

$$|\psi_4^i\rangle = CS_{u_1 u_2}^{p+1-i} |\psi_3^i\rangle \tag{4.11}$$

Esta é a operação central do algoritmo de armazenamento. Separa o novo padrão a ser armazenado, já com o fator de normalização correto.

$$|\psi_5^i\rangle = nXOR_{m_1 \ldots m_n u_1} |\psi_4^i\rangle \tag{4.12}$$

$$|\psi_6^i\rangle = \prod_{j=n}^{1} XOR_{p_j^i m_j} NOT_{m_j} |\psi_5^i\rangle \tag{4.13}$$

Estas duas operações são o inverso das Equações (4.9 e 4.10) e restaura qubit u_1 e o registro de memória m dos valores originais. Depois destas operações têm-se

$$|\psi_6^i\rangle = \frac{1}{\sqrt{p}} |p^i; 00; p^k\rangle + \frac{\sqrt{p-i}}{p} |p^i; 01; p^i\rangle \tag{4.14}$$

Com a última operação

$$|\psi_7^i\rangle = \prod_{j=n}^{1} 2XOR_{p_j^i u_2 m_j} |\psi_6^i\rangle \tag{4.15}$$

Restaura-se o terceiro registro m do termo de processamento, o segundo termo na Equação (4.14) acima, ao seu valor inicial $|0_1, \ldots, 0_n\rangle$. Neste ponto, pode-se carregar um novo padrão no registro p e passar a mesma rotina apenas como descrito. No final de todo o processo, o registro m está exatamente no estado $|m\rangle$, Equação (4.6).

Suponha agora que é dada uma entrada binária i, que pode ser, por exemplo, uma versão corrompida de um dos padrões armazenados na memória. O primeiro passo é fazer uma cópia da memória de $|m\rangle$ para ser usado no algoritmo de recuperação descrito abaixo. Devido ao Teorema de não clonagem (bits quânticos não podem ser copiados)(WOOTTERS, 1982), isto não pode ser feito de forma determinística (ou seja, utilizando apenas operações unitárias), uma cópia fiel de $|m\rangle$ só pode ser obtida com uma máquina de clonagem probabilística (DUAN, 1998). Assim, assume-se disponibilidade de uma máquina de clonagem probabilística para as quais $|m\rangle$ é um conjunto de estados linearmente independentes que podem ser copiados.

O algoritmo de recuperação requer também três registros. O primeiro registro i de n qubits contém o padrão de entrada, o segundo registro m, também de n qubits, contém a memória de $|m\rangle$; Finalmente, há um registro de controle c com b qubits todos inicializando no estado $|0\rangle$. O estado quântico inicial é assim

$$|\psi_0\rangle = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \left|i; p^k; 0_1, \ldots, 0_b\right\rangle \tag{4.16}$$

em que $|i\rangle = |i_1, \ldots, i_n\rangle$ denota os qubits de entrada, o segundo registro, m, contém a memória e todos os b qubits de controle no estado $|0\rangle$. Aplicando a porta Hadamard para o primeiro qubit de controle obtemos

$$|\psi_0\rangle = \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} \left|i_1, \ldots, i_n; p_1^k, \ldots, p_n^k; 0_1, \ldots, 0_b\right\rangle + \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} \left|i_1, \ldots, i_n; p_1^k, \ldots, p_n^k; 1_1, \ldots, 0_b\right\rangle \tag{4.17}$$

Agora, aplica-se a seguinte combinação de portas quânticas:

$$|\psi_1\rangle = \prod_{k=1}^{n} NOT_{m_k} XOR_{i_k m_k} |\psi_1\rangle \tag{4.18}$$

em que, como antes, os subscritos referem-se às portas do qubits em que são aplicados. Como resultado disto, os qubits do registro de memória estão no estado $|1\rangle$ se i_j e p_j^k são idênticos e $|0\rangle$ caso contrário:

$$|\psi_1\rangle = \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} \left|i_1, \ldots, i_n; d_1^k, \ldots, d_n^k; 0_1, \ldots, 0_b\right\rangle + \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} \left|i_1, \ldots, i_n; d_1^k, \ldots, d_n^k; 1_1, \ldots, 0_b\right\rangle \tag{4.19}$$

em que $d_j^k = 1$ se e somente se $i_j = p_j^k$ e $d_j^k = 0$ caso contrário. Considere agora o seguinte

Hamiltoniano:

$$\mathscr{H} = (d_H)_m \otimes (\sigma_3)_{c_1} \tag{4.20}$$

$$(d_H)_m = \sum_{k=1}^{n} \left(\frac{\sigma_3 + 1}{2} \right)_{m_k} \tag{4.21}$$

em que σ^3 é a terceira matriz de Pauli. $\mathscr{H}$ mede o número de 0's no registro m, com um sinal maior se c_1 está no estado $|0\rangle$ e um sinal menor se c_1 está em estado de $|1\rangle$. Dado o estado $|\psi_2\rangle$, isto não é nada mais do que o número de qubits que são diferentes na entrada e no registro de memória i e m. Esta quantidade é chamada de distância de Hamming e representa a distância (ao quadrado) euclidiana entre dois padrões binários. Cada termo na sobreposição (4.19) é um dos autoestados $\mathscr{H}$ com autovalor diferente. Aplicando, assim, o operador unitário $exp\left(\frac{i\pi}{2n}\mathscr{H}\right)$ para $|\psi_2\rangle$ obtém-se

$$|\psi_2\rangle = exp\left(\frac{i\pi}{2n}\mathscr{H}\right)|\psi_1\rangle \tag{4.22}$$

$$|\psi_3\rangle = \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} exp\left(\frac{i\pi}{2n} d_H(i,p^k)\right) \left|i_1,\ldots,i_n;d_1^k,\ldots,d_n^k;0_1,\ldots,0_b\right\rangle \tag{4.23}$$

$$+ \frac{1}{\sqrt{2p}} exp\left(-\frac{i\pi}{2n} d_H(i,p^k)\right) \sum_{k=1}^{p} \left|i_1,\ldots,i_n;d_1^k,\ldots,d_n^k;1_1,\ldots,0_b\right\rangle \tag{4.24}$$

$$\tag{4.25}$$

em que $d_H(i,p^k)$ indica a distância de Hamming entre a entrada i e o padrão armazenado p_k. Na etapa final restaura-se a porta de memória para o estado $|m\rangle$, aplicando a transformação inversa Equação (4.18) e aplica-se a porta Hadamard para a qubit controle, obtendo-se

$$|\psi_4\rangle = H_{c_1} \prod_{k=n}^{1} XOR_{i_k n_k} NOT_{m_k} |\psi_3\rangle \tag{4.26}$$

$$|\psi_4\rangle = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} cos\left(\frac{\pi}{2n} d_H(i,p^k)\right) |i_1,\ldots,i_n;p_1^k,\ldots,p_n^k;0_1,\ldots,0_b\rangle \tag{4.27}$$

$$+ \frac{1}{\sqrt{p}} sin\left(\frac{\pi}{2n} d_H(i,p^k)\right) \sum_{k=1}^{p} |i_1,\ldots,i_n;p_1^k,\ldots,p_n^k;1_1,\ldots,0_b\rangle \tag{4.28}$$

$$\tag{4.29}$$

A idéia agora é repetir estas operações sequencialmente para todos os b qubits de controle c_1 a c_b. Então temos:

$$|\psi_{final}\rangle = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \sum_{l=0}^{b} cos^{b-l}\left(\frac{\pi}{2n} d_H(i,p^k)\right) sin^{l}\left(\frac{\pi}{2n} d_H(i,p^k)\right) \sum_{\{J^l\}} |i;p^k;J^l\rangle \tag{4.30}$$

$$\tag{4.31}$$

em que $\{J^l\}$ denota o conjunto de todos os números binários de b bits com exatamente l bits 1 e $(b-l)$ bits 0.

Desde que o número esperado de repetições necessárias para medir o estado de registro de controle pretendido seja $\frac{1}{p_b^{rec}}$, com

$$p_b^{rec} = \frac{1}{p} \sum_{k=1}^{p} cos^{2b}\left(\frac{\pi}{2n} d_H(i,p^k)\right) \tag{4.32}$$

a probabilidade de medir $|c_1,\ldots,c_n\rangle = |0_1,\ldots,0_n\rangle$. Uma vez que o padrão i é reconhecido, a medição do registro de memória produz o padrão p^k armazenado com probabilidade:

$$P(p^k|i) = \frac{1}{Z} cos^{2b}\left(\frac{\pi}{2n} d_H(i,p^k)\right) \tag{4.33}$$

em que $Z = p \cdot p_b^{rec} = \sum_{k=1}^{p} cos^{2b}\left(\frac{\pi}{2n} d_H(i,p^k)\right)$.

Esta probabilidade tem pico em torno dos padrões que tem menor distância de Hamming para a entrada. A maior probabilidade de recuperação é portanto realizada pelo padrão que é mais próximo da entrada. Isto é sempre verdade, independentemente do número de padrões armazenados. Contrariamente à versão mais simples do modelo apresentado em (TRUGENBERGER, 2001), aqui há um segundo parâmetro ajustável, ou seja, o número b de qubits de controle. Este novo parâmetro b controla a eficiência de

identificação da memória quântica, pois aumentando b, a distribuição de probabilidade $P(p^k|i)$ torna-se mais e mais um pico onde a distância de Hamming é menor, ou seja $lim_{b\to\infty}P(p^k|i) = \delta_{kk_{min}}$, em que k_{min} é o padrão com a menor distância de Hamming para a entrada (TRUGENBERGER, 2002).

A distribuição quântica desta probabilidade é equivalente a uma distribuição de Boltzman canônica com temperatura $1/b$ e níveis de energia:

$$E^k = -2 \cdot log\left[cos\left(\frac{\pi}{2n} d_H(i,p^k)\right)\right] \tag{4.34}$$

com Z sendo a função partição.

Com essas informações podemos escrever a distribuição de probabilidade $P(p^k|i)$ da seguinte forma:

$$P(p^k|i) = \frac{1}{Z} e^{(-2\cdot log\left[cos\left(\frac{\pi}{2n} d_H(i,p^k)\right)\right]\cdot b)} \tag{4.35}$$

em que $Z = \sum_{k=1}^{p} e^{(-2\cdot log\left[cos\left(\frac{\pi}{2n} d_H(i,p^k)\right)\right]\cdot b)}$.

4.5 Comparação entre os métodos clássico e quântico de reconhecimento de padrões

A correlação cofenética mede o grau de ajuste entre a matriz de similaridade original (matriz S) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Tal correlação foi calculada usando-se (BUSSAB et. al, 1990)

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-(\bar{c}))(s_{ij}-(\bar{s}))}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-(\bar{c}))^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-(\bar{s}))^2}},$$

em que c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética; s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade; $\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}$; $\bar{s} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}$.

Nota-se que essa correlação equivale à correlação de Pearson entre a matriz de similaridade original e aquela obtida após a construção do dendrograma. Assim, quanto mais próxima de 1, menor será a distorção provocada pelo agrupamento dos indivíduos.

O grau da distorção $(1-\alpha)$ foi calculado por (KRUSKAL, 1964):

$$\alpha = \frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n} c_{ij}}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} s_{ij}}$$

em que c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética; s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade. Esse parâmetro mede a distorção entre a matriz original e aquela obtida após a construção do dendrograma. Quanto menor esse valor melhor.

O valor do estresse (S), foi calculado por (KRUSKAL, 1964):

$$S = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(s_{ij}-c_{ij})^2}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} s_{ij}}}$$

em que c_{ij}: valor de similaridade entre os indivíduos i e j , obtidos a partir da matriz cofenética; s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade.

Esta representação estatística do estresse (soma de quadrados de resíduos padronizados), proposta por Kruskal (1964), é um parâmetro que determina a precisão de ajuste da projeção gráfica. Neste caso, foi usada para determinar a precisão do ajuste obtido com a da projeção da matriz de similaridade no dendrograma. O estresse foi classificado de acordo com os critérios: nível de estresse 40% o ajuste é Insatisfatório; nível de estresse 20% o ajuste é Regular; nível de estresse 10% o ajuste é bom; nível de estresse 5% o ajuste é Excelente e nível de estresse 0% o ajuste é Perfeito.

5 Resultados e Discussão

5.1 Análise estatística dos dados

Nesta seção fazemos uma análise estatística das sequências de DNA das abelhas do gênero *Melipona quinquefasciata*. Na Tabela 5.1 podemos observar a descrição das sequências de DNA da região 18S de seis abelhas do gênero *Melipona quinquefasciata*, localizadas em Ubajara, Jardim1, Flona2, Chapada3, Araripe3 e Luziânia. Nota-se que as frequências de A e T são próximas de 25% cada, C aparece 23% das vezes e G 27% em todas sequências.

Tabela 5.1: Análise descritiva das sequências de DNA da região 18S

Localidade	T	A	C	G
Araripe 3	461 (0,2529)	453 (0,2485)	416 (0,2282)	493 (0,2704)
Chapada 3	455 (0,2496)	450 (0,2468)	420 (0,2304)	498 (0,2732)
Flona 2	459 (0,2518)	453 (0,2485)	417 (0,2287)	494 (0,2710)
Jardim 1	453 (0,2485)	456 (0,2501)	422 (0,2315)	492 (0,2699)
Luziânia	460 (0,2523)	456 (0,2501)	413 (0,2265)	494 (0,2710)
Ubajara	461 (0,2529)	457 (0,2507)	413 (0,2265)	492 (0,2699)

Em relação à análise de variância, observa-se que as médias das localidades são iguais com isso a soma de quadrados é igual a zero, portanto não existe diferença significativa entre elas. Assim, utilizamos um experimento inteiramente ao acaso para verificar se existe diferença entre as médias das bases. Observamos, na Tabela 5.2 com o teste F, que existe diferença significativa entre as médias das bases com p-valor de 2,2e-16. Dessa forma, fez-se o teste de Tukey (Tabela 5.3) para verificar quais das bases diferiam entre si. Obteve-se os valores de $q = 3,96$ e o $DMS = 4,8947$, com isso nota-se que entre as médias das bases apenas as bases A e T não diferem entre si, ao nível de 5% de significância, pois a diferença de médias entre elas é igual a 4 que é menor que o DMS.

Tabela 5.2: Tabela de Análise de variância para as sequências de DNA da região 18S

Fonte de Variação	gl	SQ	QM	Valor F	p-valor
Base	3	17839,2	5946,4	648,7	2,2e-16
Resíduos	20	183,3	9,2		

Tabela 5.3: Teste de Tukey para as sequências de DNA da região 18S, referente a base, com $DMS = 4,8947$

	T	A	C	G
T	–			
A	4,0000	–		
C	41,3333*	37,3333*	–	
G	35,6667*	39,6667*	77,0000*	–

*: existe diferença significativa entre as médias.

Na Tabela 5.4 apresentamos uma descrição das sequências de DNA da região ITS1 parcial de 9 abelhas do gênero *Melipona quinquefasciata*, localizadas em Flona 1, Flona 2, Flona 3, Araripe 1, Araripe 3, Jardim 1, São Benedito, Piauí 1 e Goiás. Verifica-se que a base T aparece com frequências próximas do intervalo 29% e 31%, em seguida C e G próximos de 28% cada e A entre 15% e 17% das vezes.

Tabela 5.4: Análise descritiva das sequências de DNA da região ITS1 parcial

Localidade	T	A	C	G
Araripe 1	142 (0,2892)	82 (0,1670)	136 (0,2770)	131 (0,2668)
Araripe 3	145 (0,2953)	79 (0,1609)	134 (0,2729)	133 (0,2709)
Flona 1	141 (0,2872)	82 (0,1670)	136 (0,2770)	132 (0,2688)
Flona 2	142 (0,2892)	82 (0,1670)	135 (0,2749)	132 (0,2688)
Flona 3	148 (0,3014)	76 (0,1548)	134 (0,2729)	133 (0,2709)
Goiás	146 (0,2974)	77 (0,1568)	133 (0,2709)	135 (0,2749)
Jardim 1	152 (0,3096)	77 (0,1568)	130(0,2648)	132 (0,2688)
Piauí 1	141 (0,2872)	83 (0,1690)	136 (0,2770)	131 (0,2668)
São Benedito	141 (0,2872)	83 (0,1690)	137 (0,2790)	130 (0,2648)

Para a análise de variância, utilizamos um experimento inteiramente ao acaso para verificar se existe diferença entre as médias das bases, pois observa-se que as médias das localidades são iguais (soma de quadrados igual a zero), não existindo diferença significativa entre elas. É possível verificar, na Tabela 5.5, a existência de diferença significativa entre as médias das bases com p-valor de 2,2e-16. Portanto, fez-se o teste de Tukey (Tabela 5.6) para verificar quais das bases diferem entre si. O valor de $q = 3,85$ e o $DMS = 3,4966$. Nota-se que entre as médias das bases apenas as bases C e G não diferem significativamente entre si sendo a diferença (2,4444) menor que o DMS.

Tabela 5.5: Tabela de Análise de variância para as sequências de DNA da região ITS1 parcial

Fonte de Variação	gl	SQ	QM	Valor F	p-valor
Base	3	22555,2	7518,4	1012,8	2,2e-16
Resíduos	32	237,6	7,4		

Tabela 5.6: Teste de Tukey para as sequências de DNA da região ITS1 parcial, referente a base, com $DMS = 3,4966$

	T	A	C	G
T	–			
A	64,1111*	–		
C	9,6667*	54,4444*	–	
G	12,1111*	52,0000*	2,4444	–

*: existe diferença significativa entre as médias.

5.2 Reconhecimento Quântico de Padrões

Esta seção analisa a probabilidade de reconhecimento quântico de padrões das sequências de DNA das regiões 18S e ITS1 parcial das abelhas do gênero *Melipona quinquefasciata*. Estes resultados foram publicados na X Jornada de Ensino, Pesquisa e Extensão (JEPEX) (BARROS, 2010d).

5.2.1 Região 18S

Nesta seção se analisa as sequências de DNA da região 18S de seis abelhas do gênero *Melipona quinquefasciata*. Nas Figuras 5.1, 5.2 e 5.3, tem-se o gráfico da probabilidade $P(p^k|i)$ com ruídos de 10% a 40%, para $b = 1$, 10 e 100, respectivamente. Verifica-se que esta probabilidade é eficiente para reconhecer padrões de sequências de DNA da região 18S. Nota-se que, aumentando o valor de b a probabilidade do padrão ser reconhecido aumenta mesmo com ruídos. Isto segue que o parâmetro b controla a eficiência de identificação, pois aumentando b, a distribuição de probabilidade $P(p^k|i)$ torna-se um pico maior onde a distância de Hamming é menor.

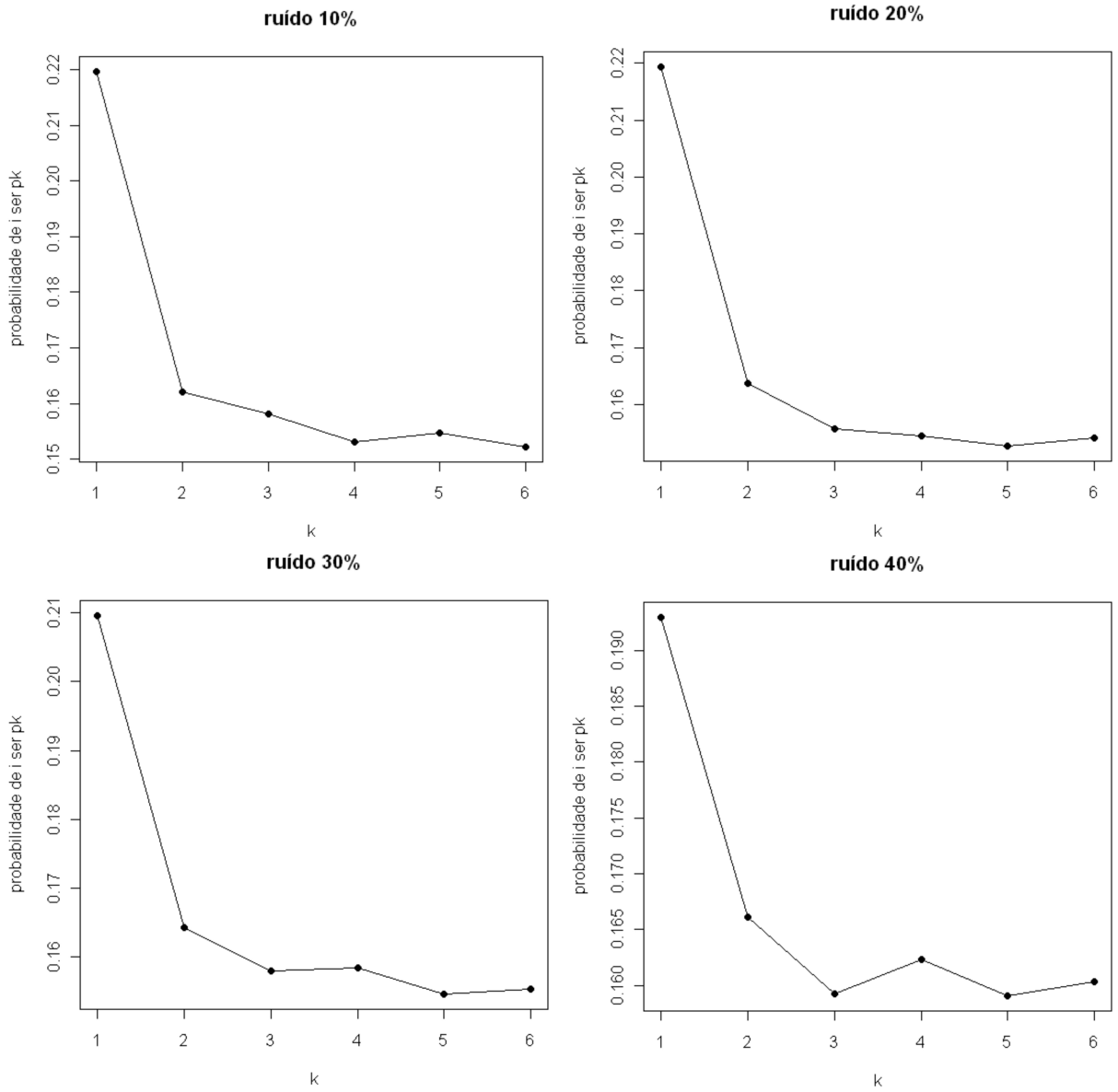

Figura 5.1: Probabilidade $P(p^k|i)$ para região 18S, com $b = 1$, com ruídos de 10% a 40%

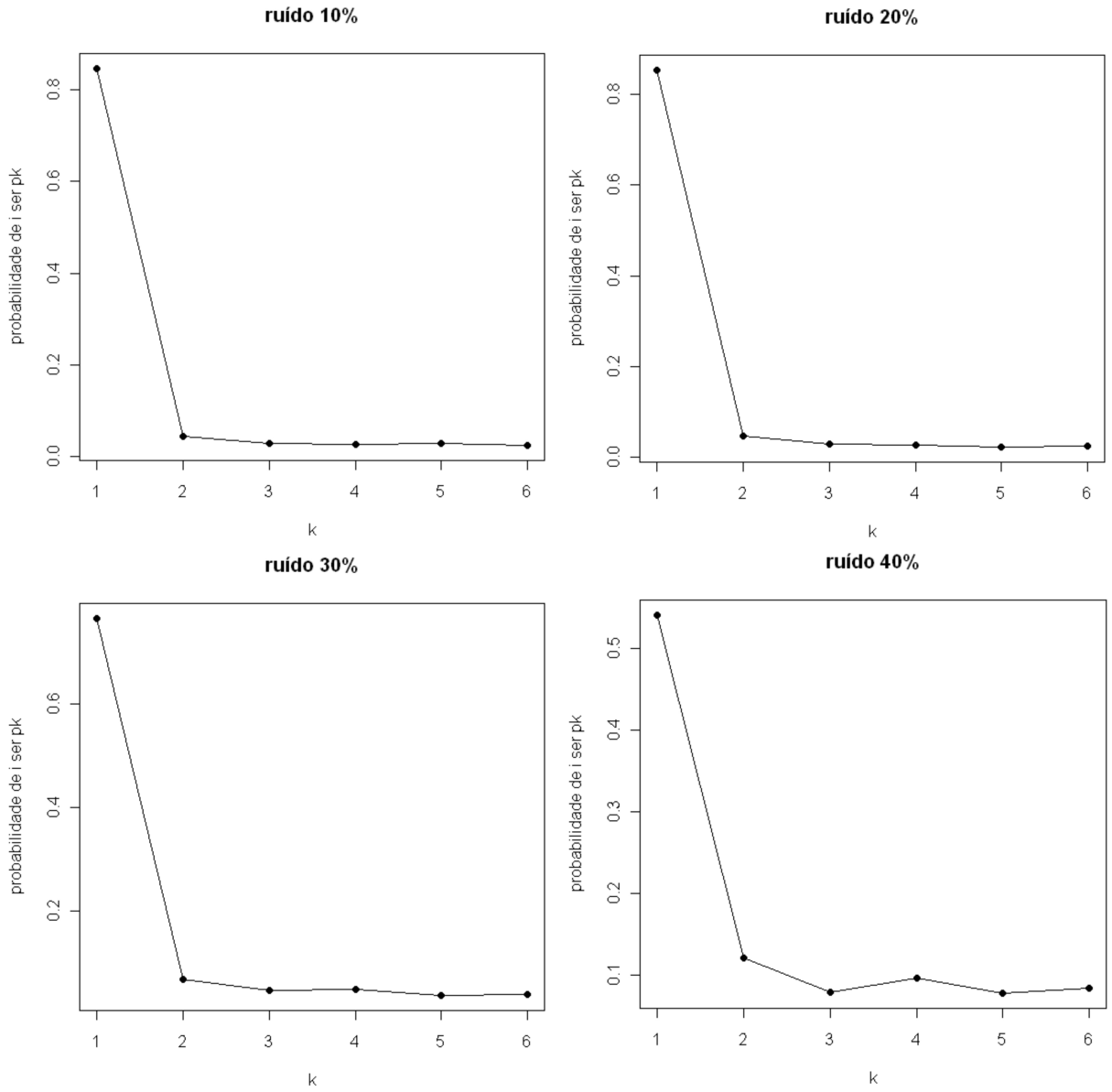

Figura 5.2: Probabilidade $P(p^k|i)$ para região 18S, com $b = 10$, com ruídos de 10% a 40%

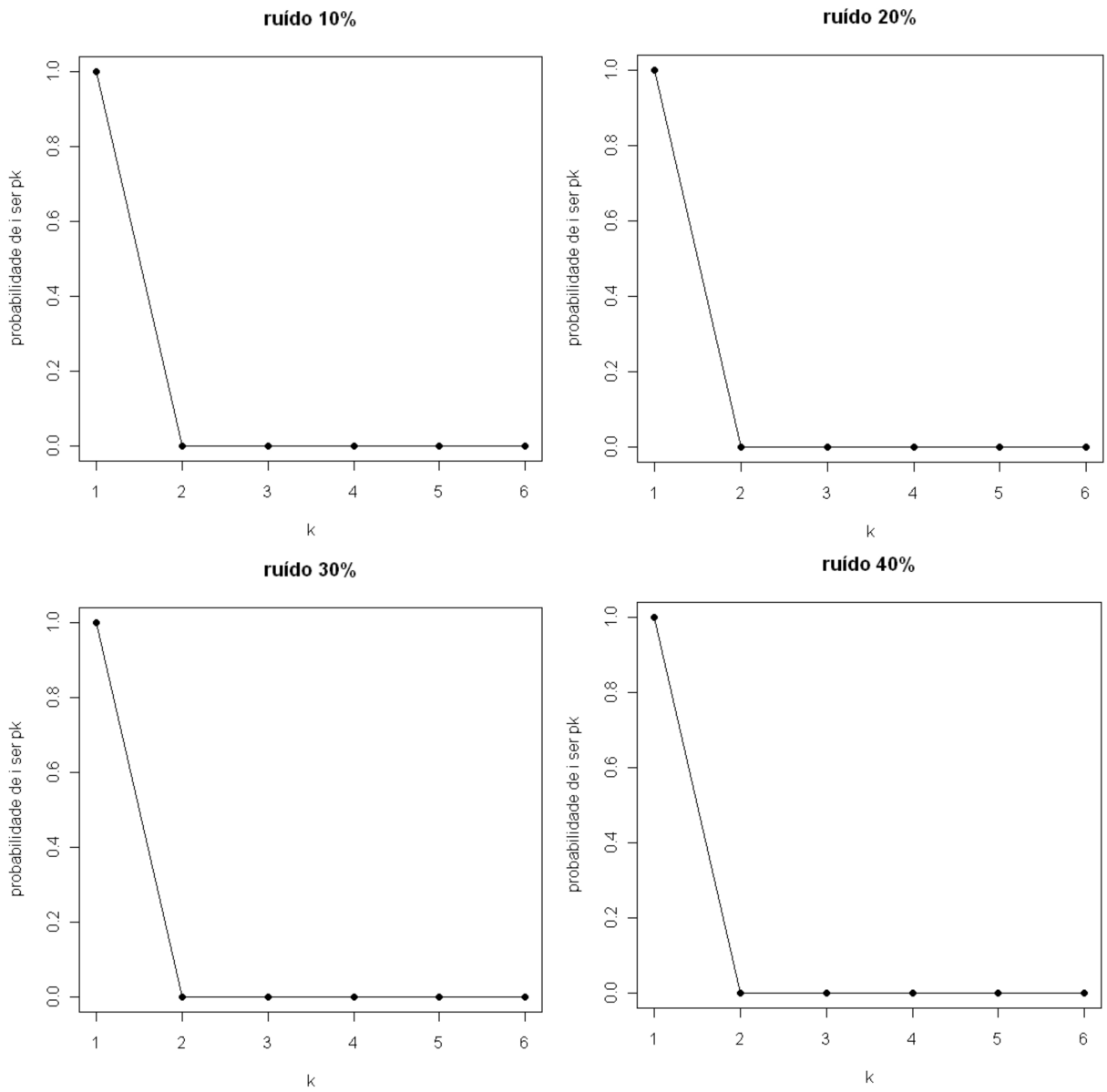

Figura 5.3: Probabilidade $P(p^k|i)$ para região 18S, com $b = 100$, com ruídos de 10% a 40%

Observamos, na Figura 5.4, que para $b = 1$ e $b = 10$ a probabilidade de reconhecimento das sequências de DNA vai diminuindo a medida que aumenta o ruído, sendo que para $b = 10$ é mais próximo de 1. Nota-se também que para $b = 100$, esta probabilidade é 1 para todos os ruídos, com exceção do ruído de 40%.

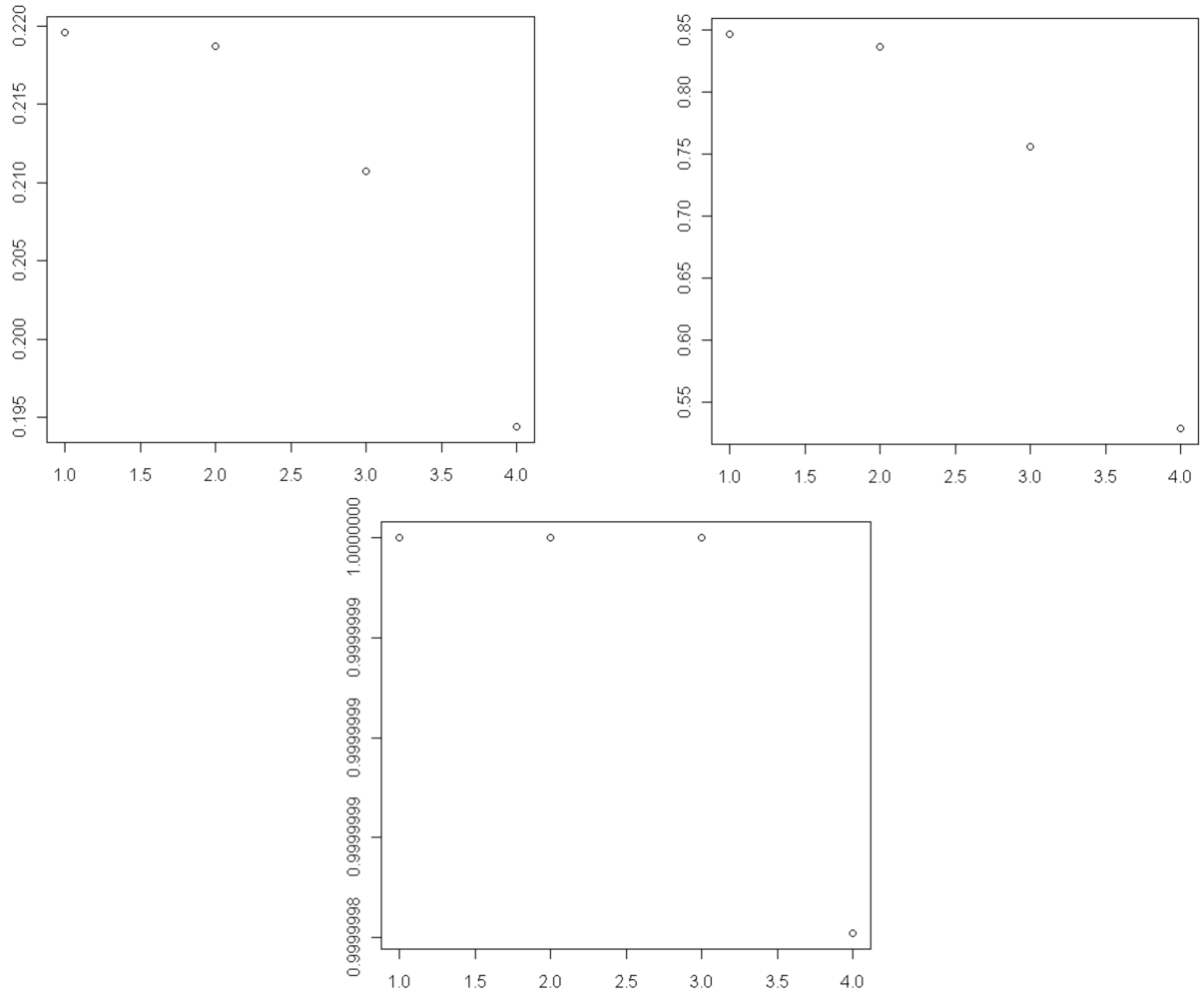

Figura 5.4: Probabilidade do padrão i ser da classe p_k para região 18S, com ruídos de 10% a 40%, para $b = 1$, 10 e 100

5.2.2 Região ITS1 parcial

Nesta seção analisa-se as sequências de DNA da região ITS1 parcial de nove abelhas do gênero *Melipona quinquefasciata*. Nas Figuras 5.5, 5.6 e 5.7, tem-se o gráfico da probabilidade $P(p^k|i)$ com ruídos de 10% a 40%, para $b = 1$, 10 e 100, respectivamente. Observa-se a eficiência desta probabilidade no reconhecimento de padrões de sequências de DNA da região ITS1 parcial. Da mesma forma que na região 18S, nota-se que aumentando o valor de b para 10 e para 100, a probabilidade do padrão ser reconhecido aumenta.

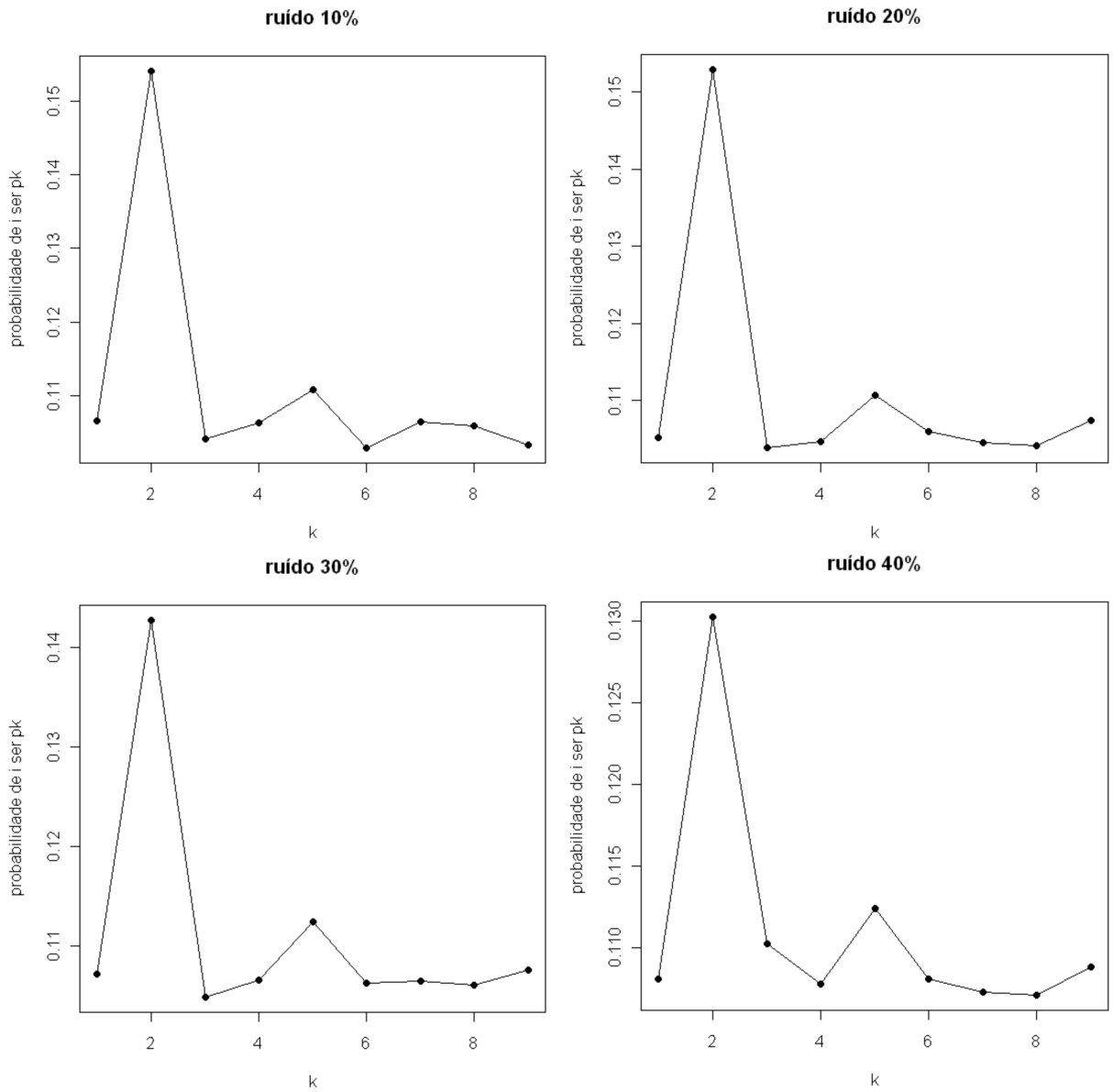

Figura 5.5: Probabilidade $P(p^k|i)$ para região ITS1 parcial, com $b = 1$, com ruídos de 10% a 40%

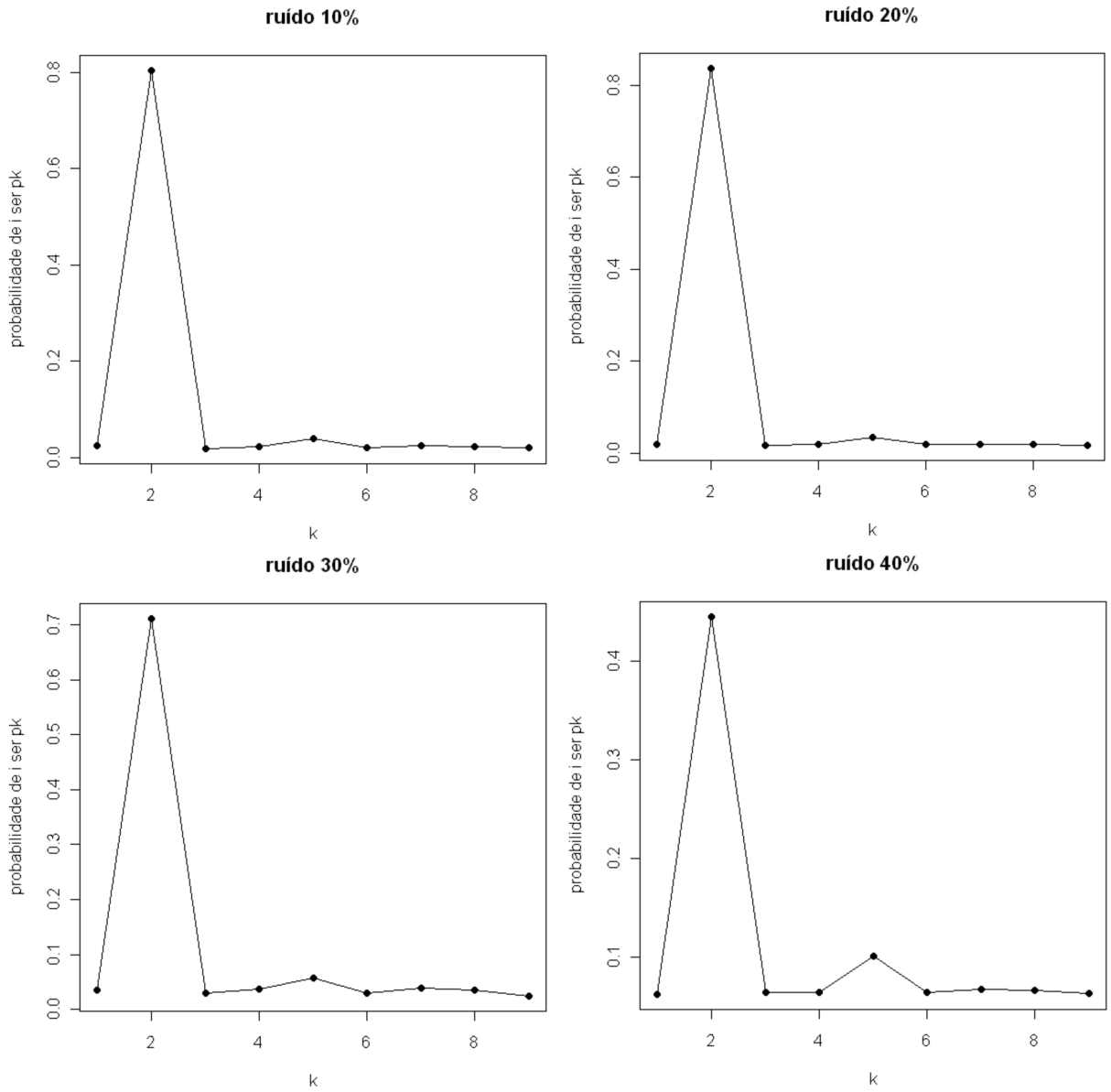

Figura 5.6: Probabilidade $P(p^k|i)$ para região ITS1 parcial, com $b = 10$, com ruídos de 10% a 40%

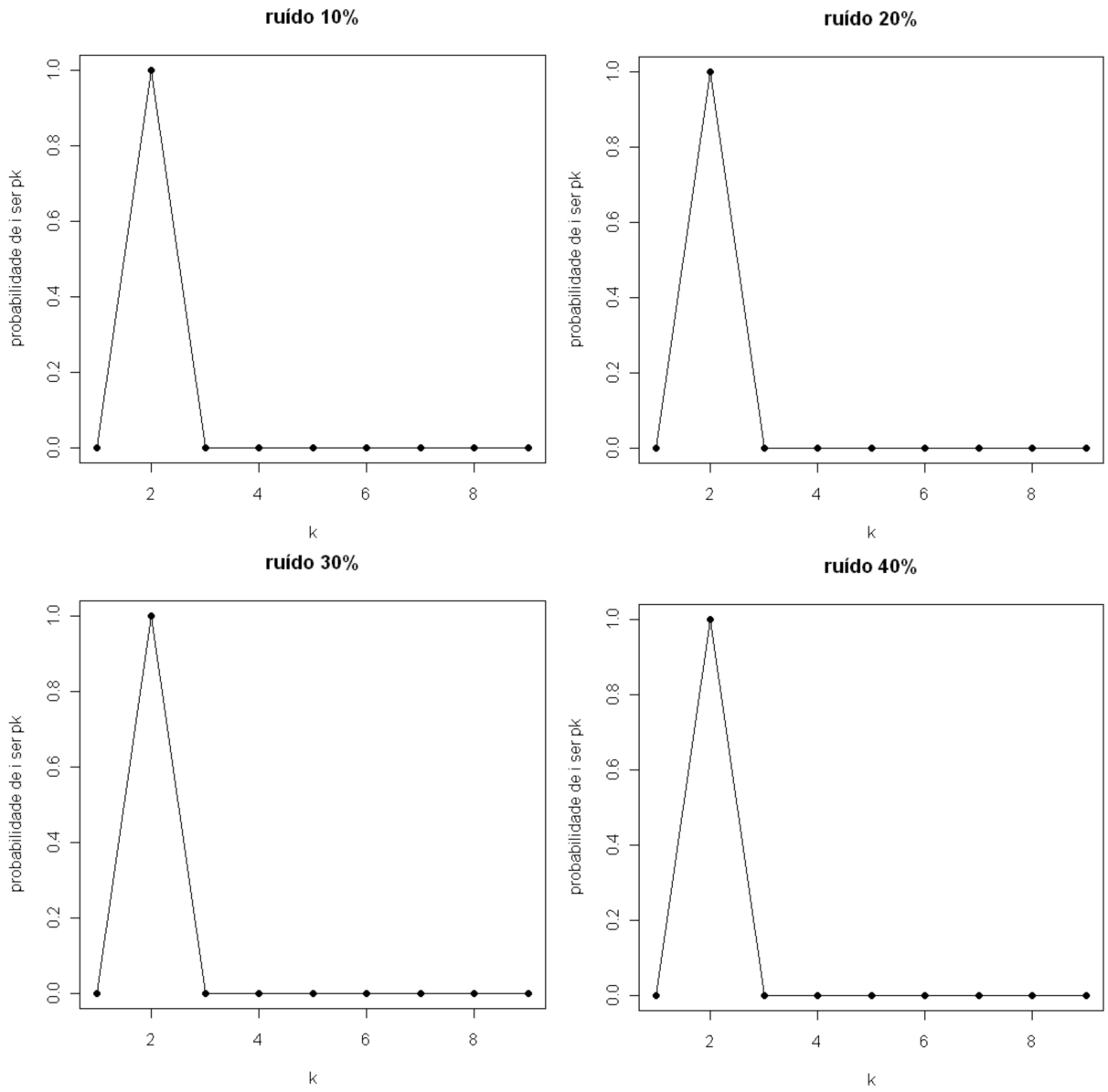

Figura 5.7: Probabilidade $P(p^k|i)$ para região ITS1 parcial, com $b = 100$, com ruídos de 10% a 40%

Observamos, na Figura 5.8, que a probabilidade de reconhecimento das sequências de DNA vai diminuindo a medida que aumenta o ruído. Também verifica-se que para aumentando o valor de b esta probabilidade se aproxima de 1 para todos os ruídos, com exceção do ruído de 40%.

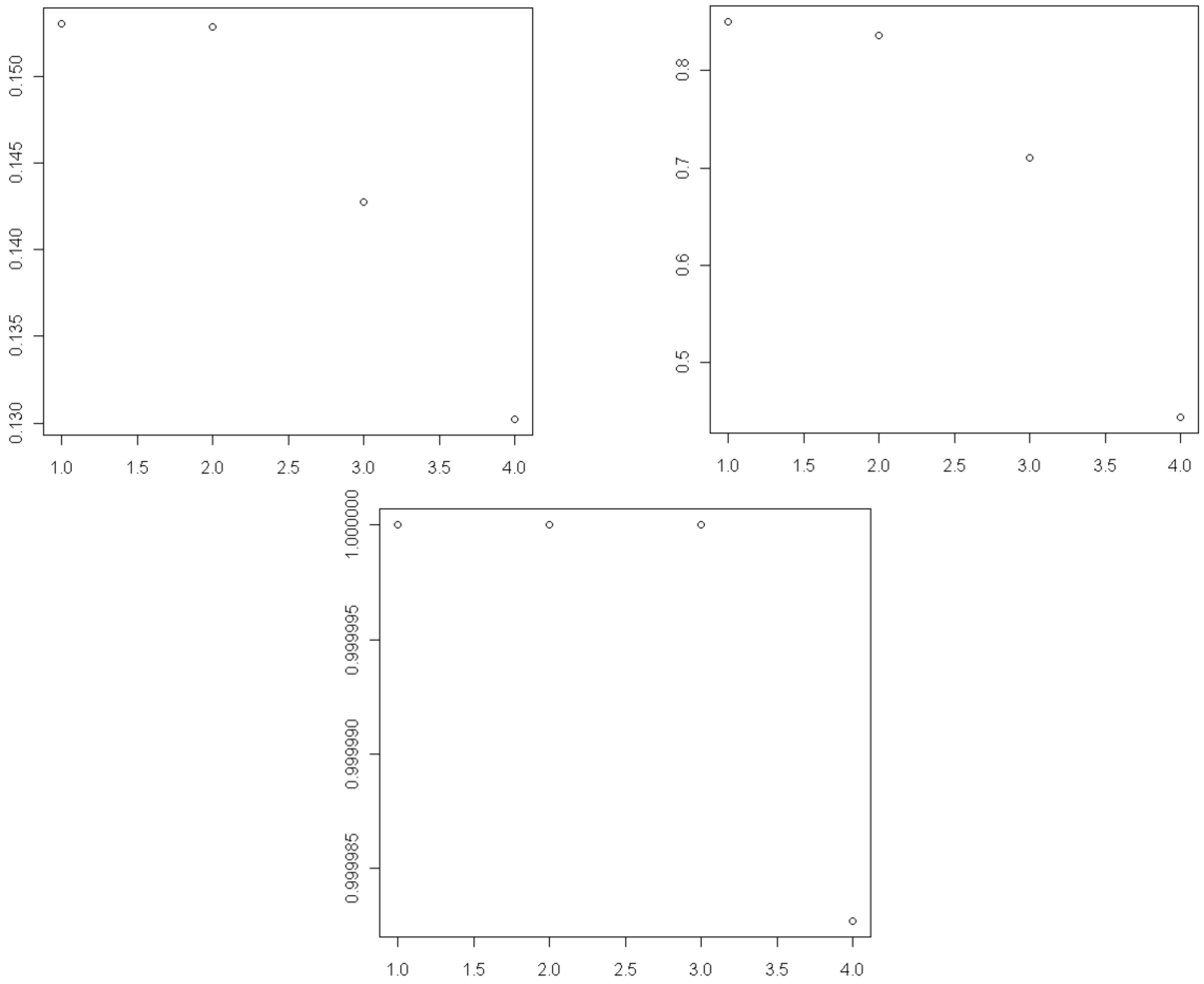

Figura 5.8: Probabilidade do padrão i ser da classe p_k para região ITS1 parcial, com ruídos de 10% a 40%, para $b = 1$, 10 e 100

5.3 Comparação entre os métodos clássico e quântico de reconhecimento de padrões

Nesta seção faz-se uma comparação entre o método de Reconhecimento de Padrões clássico utilizado por Pereira e Reconhecimento Quântico de Padrões aplicados ao mesmo conjunto de dados. Para isto, utiliza-se o dendograma e índices.

5.3.1 Região 18S

Na Figura 5.9, tem-se a comparação dos dendogramas utilizando a probabilidade de reconhecimento $P(p^k|i)$ (com as distâncias de Hamming, Jaccard e Canberra) e a distância de Jukes-Cantor para a região 18S. Observa-se que em todos os casos se separam em dois grupos, sendo no caso da probabilidade o primeiro grupo formado por Chapada 3, Araripe 3 e Luziânia; e o segundo grupo por Ubajara, Jardim 1 e Flona 2. No caso da distância de Jukes-Cantor o primeiro grupo formado por Chapada 3 e Flona 2; e o segundo grupo por Ubajara, Jardim 1, Araripe 3 e Luziânia. Podemos notar que as distâncias entre os grupos formados é mínima, indicando que pertencem ao mesmo grupo.

Na Tabela 5.7 podemos notar os valores dos índices para comparar nossos resultados com o de Pereira (2006). Observa-se que a correlação entre a matriz de similaridade original e a matriz resultante da simplificação proporcionada pelo método de agrupamento é mais próximo de 1 no método quântico 0,9999 (distância de Canberra), indicando que esse ajuste é melhor. A distorção e o estresse é mais próximo de 0 no método quântico (distância de Canberra), sendo 0,0026 e 0,0012 respectivamente, indicando que a precisão desse ajuste é maior que o método utilizado por Pereira (2006).

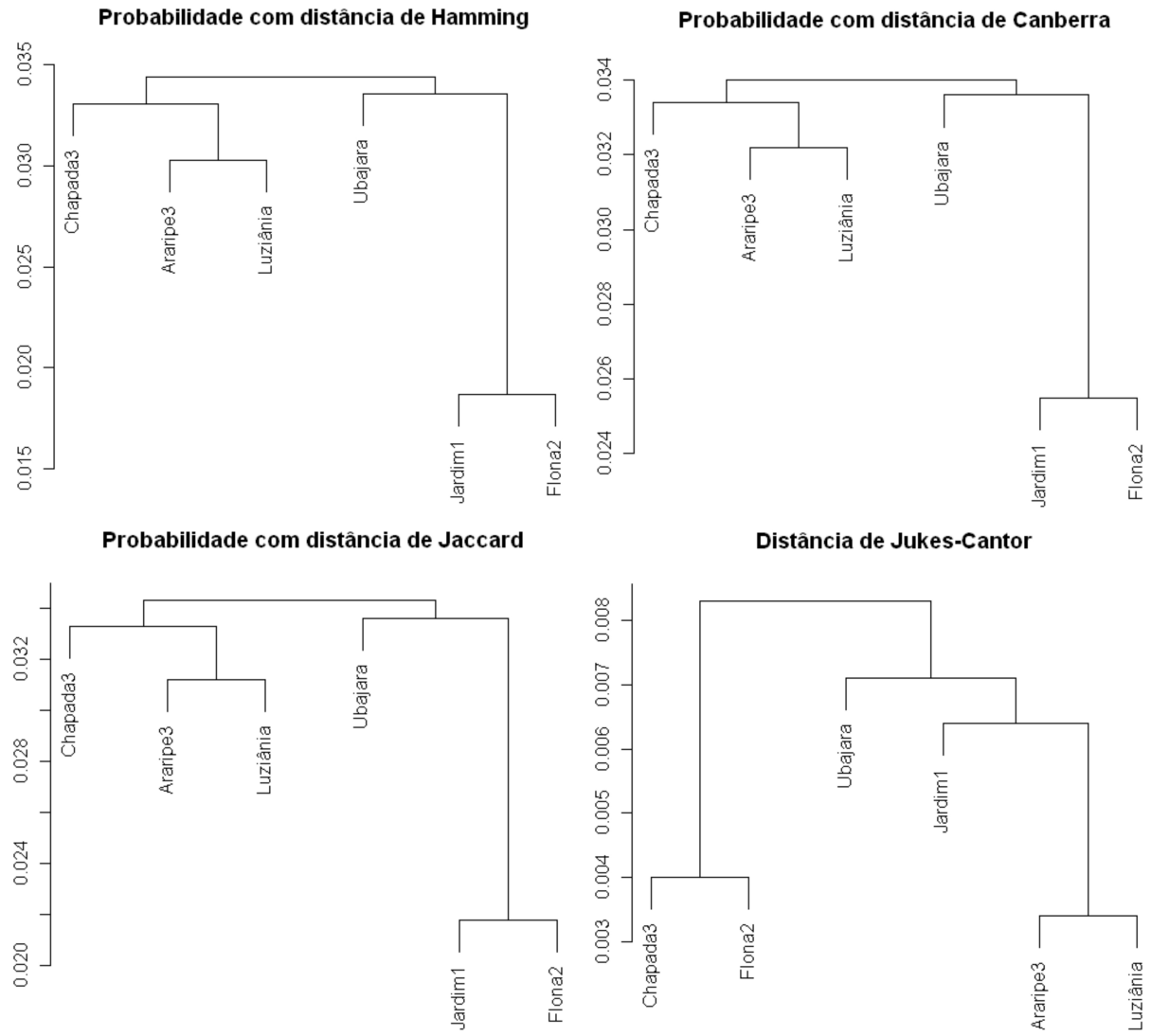

Figura 5.9: Comparação dos dendogramas utilizando a probabilidade de reconhecimento $P(p^k|i)$ com as distâncias de Hamming, Canberra e Jaccard, e a distância de Jukes-Cantor (Pereira, 2006) para a região 18S

Tabela 5.7: Índices para comparação de métodos de reconhecimento para a região 18S

Método	Correlação	Distorção	Estresse
Probabilidade com distância de Hamming	0,9994	0,0112	0,0035
Probabilidade com distância de Jaccard	0,9997	0,0065	0,0023
Probabilidade com distância de Canberra	0,9999	0,0026	0,0012
Distância de Jukes-Cantor (Pereira, 2006)	0,9618	0,1176	0,0167

5.3.2 Região ITS1 parcial

Na Figura 5.10 podemos comparar os dendogramas utilizando a probabilidade de reconhecimento $P(p^k|i)$ (com as distâncias de Hamming, Jaccard e Canberra) e a distância de Jukes-Cantor para a região ITS1 parcial. Verifica-se que em todos os casos, os objetos estão bem próximos um dos outros indicando que pertencem ao mesmo grupo.

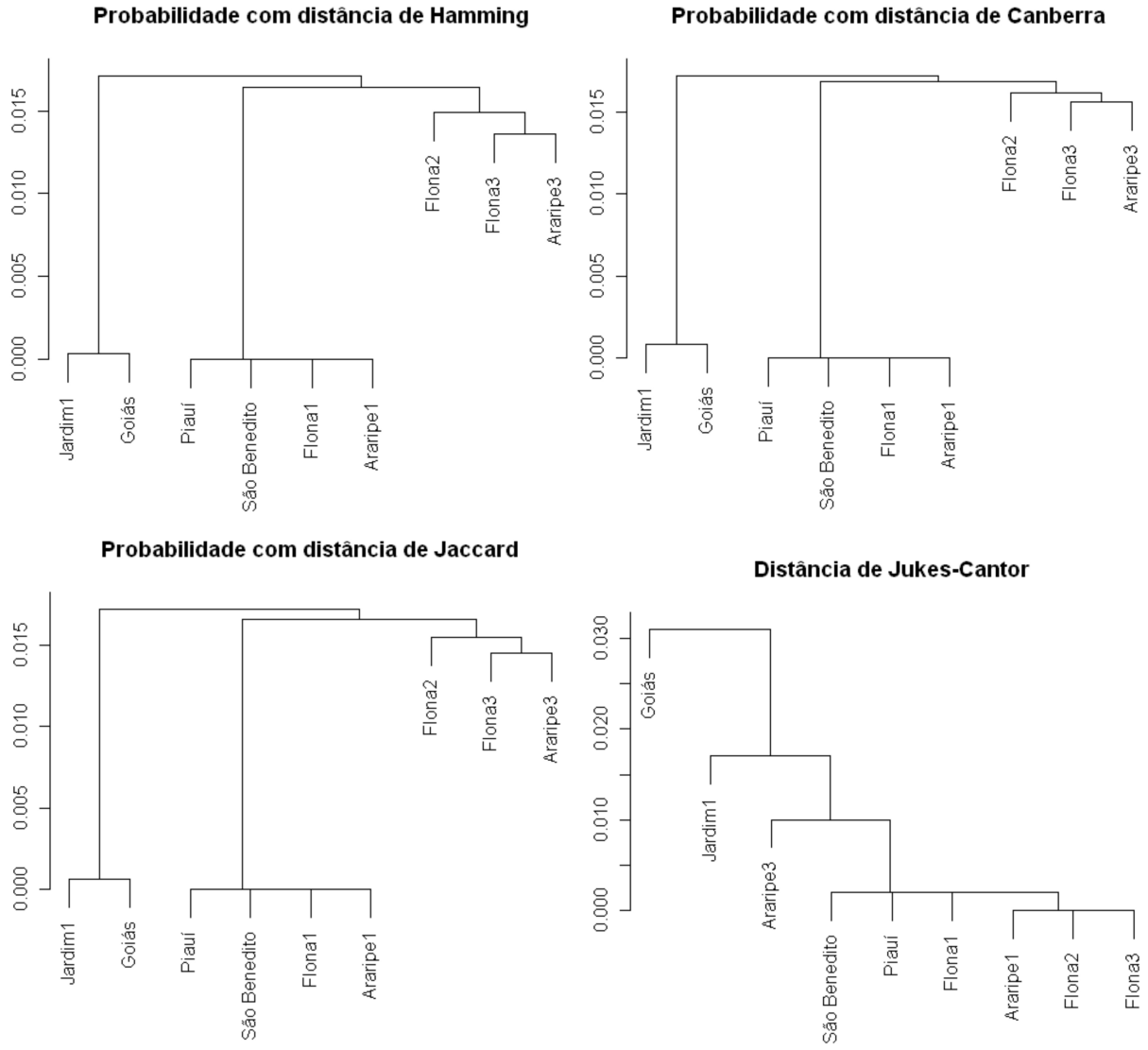

Figura 5.10: Comparação dos dendogramas utilizando a probabilidade de reconhecimento $P(p^k|i)$ com as distâncias de Hamming, Canberra e Jaccard, e a distância de Jukes-Cantor (Pereira, 2006) para a região ITS1 parcial

Na Tabela 5.8 encontramos os valores dos índices para comparar nossos resultados com o de Pereira (2006). Observa-se que a correlação entre a matriz de similaridade original e a matriz resultante da simplificação proporcionada pelo método de agrupamento é mais próximo de 1 no método quântico 0,9993 (distância de Canberra), indicando que esse ajuste é melhor. A distorção (0,0135) e o estresse (0,0032) é mais próximo de 0 no método quântico (distância de Canberra), indicando que a precisão desse ajuste é maior que o método utilizado por Pereira (2006).

Tabela 5.8: Índices para comparação de métodos de reconhecimento para a região ITS1 parcial

Método	Correlação	Distorção	Estresse
Probabilidade com distância de Hamming	0,9955	0,0362	0,0082
Probabilidade com distância de Jaccard	0,8791	0,0424	0,0337
Probabilidade com distância de Canberra	0,9993	0,0135	0,0032
Distância de Jukes-Cantor (Pereira, 2006)	0,9706	0,1065	0,0341

6 Considerações Finais

Esta dissertação teve como objetivo científico geral aplicar o método de reconhecimento quântico de padrões para verificar a diversidade genética das abelhas sem ferrão *Melipona quinquefasciata* obtidas de várias colônias silvestres, em localidades distintas da Chapada do Araripe-CE, Chapada da Ibiapaba-CE, cidade do Canto do Buriti-PI e Luziânia-GO. Com a análise estatística dos dados notou-se que as frequências das bases A e T não diferem entre si no caso da região 18S em todas sequências das abelhas sem ferrão *Melipona quinquefasciata*. Verificou-se, para a região ITS1 parcial, que as bases C e G não diferem significativamente entre si. Esses resultados foram comprovados utilizando o Teste de Tukey.

Na análise de reconhecimento quântico de padrões, verificou-se que nosso método é eficiente para reconhecer padrões de sequências de DNA das abelhas sem ferrão *Melipona quinquefasciata* das regiões 18S e ITS1 parcial. Notou-se que, aumentando o valor do parâmetro de eficiência b a probabilidade do padrão ser reconhecido aumentou, confirmando que este parâmetro controla a eficiência de identificação.

Na comparação entre o método de reconhecimento de padrões clássico utilizado por Pereira (distância de Jukes-Cantor) e Reconhecimento Quântico de Padrões observou-se para as regiões 18S e ITS1 parcial que a correlação entre a matriz de similaridade original e a matriz resultante da simplificação proporcionada pelo método de agrupamento é mais próximo de 1 no método quântico (distância de Canberra), indicando que esse ajuste é melhor. A distorção e o estresse é mais próximo de 0 no método quântico (distância de Canberra), indicando que a precisão desse ajuste é maior que o método utilizado por Pereira (2006). Verificamos com o dendograma que em todos os casos, os objetos estão bem próximos um dos outros indicando que pertencem ao mesmo grupo, isto também foi mostrado com a análise de variância em que concluiu-se que as médias das localidades não diferiam entre si.

Como trabalho futuro, pretende-se analisar a diversidade genética das abelhas sem ferrão *Melipona quinquefasciata* utilizando a técnica de Redes Neurais Quânticas sem peso, assim como também, fazer uma comparação com as Redes Neurais Clássicas.

Referências Bibliográficas

AMORIM, D. S. **Fundamentos de sistemática filogenética**. Ribeirão Preto: Holos, 2002, 156 p.

BANG, J. et al. **Quantum learning machine**, Cornell University Library, 2008. Disponível em:<http://arxiv.org/abs/0803.2976>. Acesso em 10 jan. 2010.

BARROS, P. S. N. et al. Quantum Pattern Recognition of Mitochondrial DNA sequences. In: INTERNATIONAL BIOMETRIC CONFERENCE, 25, 2010, Florianópolis. Anais do **International Biometric Conference**, Florianópolis: UFSC, 2010a. CD-ROM.

BARROS, P. S. N.; SILVA, A. J.; OLIVEIRA, W. R. Reconhecimento Quântico de Padrões Aplicados à Sequências de DNA. In: WORKSHOP ESCOLA DE COMPUTAÇÃO E INFORMAÇÃO QUÂNTICA, 3, 2010, Petrópolis. Anais do **Workshop Escola de Computação e Informação Quântica**, Petrópolis: LNCC, 2010b. CD-ROM.

BARROS, P. S. N.; OLIVEIRA, W. R.; BARROS, K. N. N. O. Reconhecimento Quântico de Padrões Aplicados à Sequências de DNA de Plantas. In: REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 55, 2010, Florianópolis. Anais do **Região Brasileira da Sociedade Internacional de Biometria**, Florianópolis: UFSC, 2010c. CD-ROM.

BARROS, P. S. N.; BARROS, K. N. N. O.; OLIVEIRA, W. R. Reconhecimento Quântico de Padrões aplicados a sequências de DNA da abelha *Melipona quinquefasciata*. In: JORNADA DE ENSINO, PESQUISA E EXTENSÃO, 10, 2010, Recife. Anais do **Jornada de Ensino, Pesquisa e Extensão**, Recife: UFRPE, 2010d. CD-ROM.

BEHRMAN, E. C. et al. Quantum algorithm design using dynamic learning. **Quantum Information and Computation**, Cornell University Library, v. 8, n. 1-2, p. 12-29, 2008.

BISHOP, C. M. **Pattern recognition and machine learning**, Singapura: Springer, 2006.

BUSSAB, W. de O.; MIAZAKI, E. S.; ANDRADE, D. F. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística. 1990. 105 p.

BROCCHIERI, L. Phylogenetics inference from molecular sequences: review and critique. **Theoretical Population Biology**. New York, v. 59, n. 1, p. 27-40, 2001.

BRUCE, A. et al. **Molecular biology of the cell**. 5. ed. New York: Garland Science, 2008.

CARDONHA, C. H.; SILVA, M. K. C.; FERNANDES, C. G. **Computação quântica: complexidade e algoritmos**, 2004. 34 f. Monografia (Ciência da Computação), Instituto de Matemática e Estatística. Universidade de São Paulo. São Paulo.

CHAPLINE, G. Quantum mechanics and pattern recognition. **International Journal of Quantum Information**, Hackensack, v. 2, n. 3, p. 295-303, 2004.

CRUZ, D. O. et al. Intraspecific variation in the first internal transcribed spacer of the nuclear ribosomal DNA in *Melipona subnitida* (Hymenoptera, Apidae), an endemic stingless bee from northeastem Brazil. **Apidologie**, França, v. 37, n. 3, p. 376-386, 2006.

DUAN, L. M.; GUO, G. C. Probabilistic cloning and identification of linearly independent quantum states. **Physical Review Letters**, New York, v. 80, p. 4999-5002, 1998.

FABER, J.; GIRALDI, G.; THESS, R. **Learning linear operators by genetic algorithms**. National Laboratory for Scientific Computing, Petrópolis, 2003, Disponível em: <http://qubit.lncc.br/files/jfaber_Learning-GA.pdf>. Acesso em 20 jan. 2010.

FERNANDES-SALOMÃO, T. M. et al. The first internal transcribed spacer (ITS-1) of *Melipona* species (Hymenoptera, Apidae, Meliponini): characterization and phylogenetic analysis. **Insectes Sociaux**. Paris, v. 52, n. 1, p. 11-18, 2005.

GIRALDI, G.; PORTUGAL, R.; THESS, R. **Genetic algorithms and quantum computation**. Cornell University Library, 2004. Disponível em: <http://arxiv.org/abs/cs/0403003>. Acesso em 10 jan. 2010.

GROVER, L. K. **A Fast Quantum Mechanical Algorithm For Database Search**. In: ANNUAL ACM SYMPOSIUM ON THE THEORY OF COMPUTATION, 96, 1996, New York. Anais do **Annual ACM Symposium on the Theory of Computation**, New York: ACM, 1996, 8 p.

HUNZIKER, M. et al. The geometry of quantum learning, **Quantum Information Processing**, v. 9, n. 3, p. 321-341, 2003.

KERR, W. E. **Estudos sobre a genética de *Melipona***. 1948. 276f. Tese (Doutorado em Genética). Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

KERR, W. E. et al. Aspectos poucos mencionados da biodiversidade amazônica - Biodiversidade, pesquisa e desenvolvimento na Amazônica. **Parcerias estratégicas**. Brasília, n. 12, p. 20-41, 2001.

KRUSKAL, J.B. Multidimensional scaling by optimizing goodness of fit to a nommetric hypothesis. **Psychometrika**, Williamsburg, v. 29, n. 1, p. 1-27, 1964.

LANCE G. N., WILLIAMS W. T. Computer programs for hierarchical polythetic classification (similarity analysis). **Computer Journal**, v. 9, p. 60-64, 1966.

LANCE G. N., WILLIAMS W. T. Mixed-data classificatory programs, I. Agglomerative Systems. **Australian Computer Journal**, v. 1, p. 15-20, 1967.

LIMA, J. P. M. S. **Estudos taxonômicos moleculares no táxon Phaseoleae DC. (Leguminosae, Papilionoideae) utilizando sequências de DNA ribossômico (rDNA)**. 2003. 136 f. Dissertação (Mestrado em Bioquímica Vegetal). Universidade Federal do Ceará, Ceará.

LIMA-VERDE, L. W.; FREITAS, B. M. Occurence and biogeographic aspects of *Melipona quinquefasciata* in NE Brazil (Hymenoptera, Apidae). **Brazilian Journal Biology**. v. 62, n. 3, p. 479-486, 2002.

MARIANO-FILHO, J. **Ensaios sobre os meliponidas do Brasil**. 1911, 140 f. Tese (Doutorado em Zoologia). Faculdade de Medicina do Rio de Janeiro, Rio de Janeiro.

MATIOLI, S. R. **Biologia Molecular e evolução**. Ribeirão Preto: Holos, 2001. cap. 10, p. 108-116.

MELO, G. A. R. Phylogenetic relationships and classification of the major lineages of Apoidea (Hymenoptera), with emphasis on the crabronid wasps. **Scientific Papers of the Natural History Museum of University of Kansas**. v. 14, p. 1-55, 1999.

MOURE, J. S. Notas sobre a espécie de *Melipona* descritas por Lepeletier em 1836. **Revista Brasileira de Biologia**. Rio de Janeiro. v. 35, p. 615-623, 1975.

MOURE, J. S. Estudando as abelhas do Brasil (pareceres de sistemática). **Chacaras e Quintais**, São Paulo, v. 77, p. 339-341, 1948.

NIELSEN, M. A.; CHUANG, I. L. **Quantum computation and quantum information**. Cambridge: University Press, 2000.

NOGUEIRA-NETO, P. **Vida e criação de abelhas indígenas sem ferrão**. São Paulo: Nogueirapis, 1997. 445 p.

PEREIRA, J. O. P. **Diversidade Genética da abelha sem ferrão *Melipona quinquefasciata* baseada no sequenciamento das regiões ITS1 parcial e 18S do DNA ribossômico nuclear**. 2006, 142 f. Tese (Doutorado em Zootecnia). Universidade Federal do Ceará, Ceará.

PEREIRA, J. O. P. et al. Genetic variability in *Melipona quinquefasciata* (Hymenoptera, Apidae, Meliponini) from northeastern Brazil determined using the first internal transcribed spacer (ITS1). **Genetics and Molecular Research**, v. 8, n. 2, p. 641-648, 2009.

PRESKILL, J. Making Weirdness Work: Quantum information and computation. In: IEEE AEROSPACE CONFERENCE, 1998. Anais do **IEEE Aerospace Conference**, Florianópolis: IEEE, 1998.

SCHUTZHOLD, R. Pattern recognition on a quantum computer. **Physical Review A**, v. 67, n. 6, 2003.

SCHUWARZ, H. F. The genus *Melipona*. The type genus of Meliponidae or stingless bees. **Bulletin of the American Museum Natural History**, v. 63, p. 231-459, 1932.

SHANKAR, A. **Principles of quantum mechanics**. 2. ed. New York: Plenum, 1994.

SHOR, P. W. Algorithms for Quantum Computation: Discrete Logarithm and Factoring. In: SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE, 35, 1995. Anais do **Symposium on Foundations of Computer Science**, p. 124-134, 1995.

TORONTO, N.; VENTURA, D. **Learning quantum operators from quantum state pairs**. In IEEE WORLD CONGRESS ON COMPUTATIONAL INTELIGENCE, 2006. Anais do **IEEE World Congress on Computational Intelligence**, p. 2607-2612, 2006.

TRUGENBERGER, C. A. Probabilistic quantum memories. **Physical Review Letters**, v. 87, n. 6, p. 067901-(1-4), 2001.

TRUGENBERGER, C. Phase transitions in quantum pattern recognition. **Physical Review Letters**, v. 89, n. 27, p. 277903-(1-4), 2002.

TRUGENBERGER, C. A. Quantum Pattern Recognition. **Quantum Information Processing**, v. 1, n. 6, p. 471-493, 2003.

VARELA, E. S. **Filogenia Molecular da subtribo diocleinae benth. (Leguminosa, Papilonoideae, Phaseoleae) baseada em sequencias dos espaçadores transcritos internos (ITS 1 e ITS 2) do DNA ribossômico nuclear (nrDNA)**. 2003. Monografia (Bacharelado em Ciências Biológicas). Universidade Federal do Ceará, Ceará.

VENTURA, D.; MARTINEZ, T. Initializing the Amplitude Distribution of a Quantum State. **Foundations of Physics Letters**, v. 12 n. 6, p. 547-559, 1999.

VENTURA, D. Learning quantum operators. In JOINT CONFERENCE ON INFORMATION SCIENCES, 2000. Anais do **Joint Conference on Information Sciences**, p. 750-752, 2000.

VIANA, L. S.; MELO, G. A. R. Conservação de abelhas. **Informativo agropecuário**. Aracaju, v. 13, n. 149, p. 23-26, 1987.

VIEIRA, S. **Análise de variância**. São Paulo: Atlas, 2006.

YANOFSKY, N. S.; MANNUCCI, M. A. **Quantum computing for computer scientists**. Cambridge: University Press, 2008.

WOOTTERS, W.; ZUREK, W. A Single Quantum Cannot be Cloned, **Nature**, v. 299, p. 802-803, 1982.

Apêndice

Abaixo, são apresentados os algoritmos utilizados nesta dissertação.

****

*PROGRAMA 1:

*UTILIZADO PARA ANOVA e Teste de Tukey

*SOFTWARE: R

****

Ler dados

S=matrix(c(461, 453, 416, 493, 455, 450, 420, 498, 459, 453, 417, 494, 453, 456, 422, 492, 460, 456, 413, 494, 461, 457, 413, 492),byrow=T,ncol=4)

Cálculos para ANOVA

C=SQT=SQTr=SQR=NULL

I=dim(S)[1];J=dim(S)[2]

n=I*J

C=sum(S)^2/n

aux1=matrix(0,I,J)

for(i in 1:I)

for(j in 1:J)

aux1[i,j]=S[i,j]^2

SQT=sum(aux1)-C

aux2=NULL

for(j in 1:J)

aux2[j]=sum(S[,j])^2

SQTr=sum(aux2)/I-C

```
SQR=SQT-SQTr

QMTr=SQTr/(J-1)

QMR=SQR/(n-j)

F=QMTr/QMR
```

Cálculos para o Teste de Tukey

```
q=3.96

dms=q*sqrt(QMR/I)

aux2=NULL

for(j in 1:J)

aux2[j]=sum(S[,j])

med=aux2/I

DIF=DIFF=matrix(0,J,J)

for(i in 1:J)

for(j in 1:J)

if(i>j)

DIF[i,j]=abs(med[i]-med[j])

if(DIF[i,j]<dms) DIFF[i,j]="i"

else DIFF[i,j]="d"

for(i in 1:J)

for(j in 1:J)

if(i>j)

DIFF[i,j]=DIFF[i,j]

else DIFF[i,j]=0
```

```
**********************************************************************************************

*PROGRAMA 2:
*UTILIZADO PARA RECONHECIMENTO QUÂNTICO DE PADRÕES
*SOFTWARE: R

**********************************************************************************************
```

Ler dados

dados=read.table(“geralbin.txt”)

dados=as.matrix(dados)

dados

k=7292

n=6

Probabilidade Pb(pk|q)

library(e1071)

q1=dados[1,]

b=100

Ruído de 10%

uni=runif(length(q1),0,1)

q1.10=NULL

for(i in 1:length(q1))

if(uni[i]<=.10) q1.10[i]=1-q1[i] else q1.10[i]=q1[i]

p.10=matrix(0,ncol=1,n)

for(w in 1:n)

p.10[w,]= cos((pi/(2*k))*(hamming.distance(q1.10, dados[w,])))$^{(2*b)}$

p.10

p.pad.10=matrix(0,ncol=1,n)

for(w in 1:n)

p.pad.10[w,]=p.10[w,]/(sum(p.10))

p.pad.10

Gráfico

par(mfrow=c(2,2))

plot(p.pad.10,type="o",xlab="k",ylab="probabilidade de i ser p^k",main="ruído 10%",pch=16)

Ruído de 20%

q1.20=NULL

for(i in 1:length(q1))

if(uni[i]<=.20) q1.20[i]=1-q1[i] else q1.20[i]=q1[i]

p.20=matrix(0,ncol=1,n)

for(w in 1:n)

p.20[w,]= cos((pi/(2*k))*(hamming.distance(q1.20, dados[w,])))$^{(2*b)}$

p.20

p.pad.20=matrix(0,ncol=1,n) for(w in 1:n) p.pad.20[w,]=p.20[w,]/(sum(p.20)) p.pad.20

Gráfico

plot(p.pad.20,type="o",xlab="k",ylab="probabilidade de i ser p^k",main="ruído 20%",pch=16)

Ruído de 30%

q1.30=NULL

for(i in 1:length(q1))

if(uni[i]<=.30) q1.30[i]=1-q1[i] else q1.30[i]=q1[i]

p.30=matrix(0,ncol=1,n)

for(w in 1:n)

p.30[w,]= cos((pi/(2*k))*(hamming.distance(q1.30, dados[w,])))$^{(2*b)}$

p.30

p.pad.30=matrix(0,ncol=1,n)

for(w in 1:n)

p.pad.30[w,]=p.30[w,]/(sum(p.30))

p.pad.30

Gráfico

plot(p.pad.30,type="o",xlab="k",ylab="probabilidade de i ser p^k",main="ruído 30%",pch=16)

Ruído de 40%

q1.40=NULL

for(i in 1:length(q1))

if(uni[i]<=.40) q1.40[i]=1-q1[i] else q1.40[i]=q1[i]

p.40=matrix(0,ncol=1,n)

for(w in 1:n)

p.40[w,]= cos((pi/(2*k))*(hamming.distance(q1.40, dados[w,])))$^{(2*b)}$

p.40

p.pad.40=matrix(0,ncol=1,n)

for(w in 1:n)

p.pad.40[w,]=p.40[w,]/(sum(p.40))

p.pad.40

Gráfico

plot(p.pad.40,type="o",xlab="k",ylab="probabilidade de i ser p^k",main="ruído 40%",pch=16)

Gráfico do ruído

par(mfrow=c(1,2))

ruido=c(p.pad.10[1,],p.pad.20[1,],p.pad.30[1,],p.pad.40[1,])

plot(ruido)

```
#*****************************************************************************************
#PROGRAMA 3:
#Algoritmo utilizado para criação dos dendogramas
#SOFTWARE: R
#*****************************************************************************************
```

Probabilidade Pb(pk|q)

library(e1071)

library(ape)

library(ade4)

require(prabclus)

library(apTreeshape)

dh=hamming.distance(dados)

dc=as.matrix(dist(dados,method="canberra"))

dj=jaccard(t(dados))

b=1

Canberra

pc=matrix(0,ncol=6,n)

for(w in 1:n)

pc[w,]= sin((pi/(2*k))*(dc[w,]))$^{(2*b)}$

pc

p.padc=matrix(0,ncol=6,n)

for(w in 1:n)

p.padc[w,]=pc[w,]/(sum(pc))

p.padc

Jaccard

pj=matrix(0,ncol=6,n)

for(w in 1:n)

pj[w,]= sin((pi/(2*k))*(dj[w,]))$^{(2*b)}$

pj

p.padj=matrix(0,ncol=6,n)

for(w in 1:n)

p.padj[w,]=pj[w,]/(sum(pj))

p.padj

Hamming

ph=matrix(0,ncol=6,n)

for(w in 1:n)

ph[w,]= sin((pi/(2*k))*(dh[w,]))$^{(2*b)}$

ph

p.padh=matrix(0,ncol=6,n)

for(w in 1:n)

p.padh[w,]=ph[w,]/(sum(ph))

p.padh

Criação dos dendogramas

par(mfrow=c(2,2))

Dendograma com a distância de hamming

prob=as.matrix(read.table(“probh.txt”,h=F))

rownames(prob)=c(“Araripe3”,“Luziania”,“Chapada3”,“Ubajara”,“Jardim1”,“Flona2”)

labels=rownames(prob)

p.dist=as.dist(round(prob,4))

hphylo<-hclust(p.dist,method=“single”)

plot(hphylo,labels=labels,main=“Probabilidade com distância de Hamming”)

Dendograma com a distância de Canberra

prob=as.matrix(read.table(“probc.txt”,h=F))

rownames(prob)=c(“Araripe3”,“Luziania”,“Chapada3”,“Ubajara”,“Jardim1”,“Flona2”)

labels=rownames(prob)

p.dist=as.dist(round(prob,4))

hphylo<-hclust(p.dist,method=“single”)

plot(hphylo,labels=labels,main=“Probabilidade com distância de Canberra”)

Dendograma com a distância de Jaccard

prob=as.matrix(read.table(“probj.txt”,h=F))

rownames(prob)=c(“Araripe3”,“Luziania”,“Chapada3”,“Ubajara”,“Jardim1”,“Flona2”)

labels=rownames(prob)

p.dist=as.dist(round(prob,4))

hphylo<-hclust(p.dist,method=“single”)

plot(hphylo,labels=labels,main=“Probabilidade com distância de Jaccard”)

Dendograma com a distância de Jukes-Cantor

prob=as.matrix(read.table(“probjc.txt”,h=F))

rownames(prob)=c(“Araripe3”,“Luziania”,“Chapada3”,“Ubajara”,“Jardim1”,“Flona2”)

labels=rownames(prob)

p.dist=as.dist(round(prob,4))

hphylo<-hclust(p.dist,method=“single”)

plot(hphylo,labels=labels,main=“Distância de Jukes-Cantor”)

```
#*****************************************************************************************
#PROGRAMA 4:
#Algoritmo utilizado para comparação dos métodos
#SOFTWARE: R
#*****************************************************************************************
```

Criação da matriz cofenética

D=read.table(“probh.txt”)

plclust(hclust(as.dist(D) ,method=“single”))

k=dim(D)[1]

M=matrix(0,k,k)

for(j in 1:k)for(i in 1:k)

if(i>j)

a=min(D);b=max(D)

lmax=200

x=round(seq(a,b,l=lmax),30)

M=cutree(hclust(as.dist(D),method=“single”) ,h=x)

c=seq(0,0,l=lmax)

I=dim(M)[1];J=dim(M)[2]

for(j in 2:J)

if(sum(M[,j])!=sum(M[,j-1])) c[j]=1

x[c==1]

Índices para comparação das matrizes de dissimilaridade e cofenética

Distância de Jukes-Cantor

Correlação

prob=as.matrix(read.table(“probjc.txt”,h=F))

prob1=as.matrix(read.table(“probjcd.txt”,h=F))

cor.test(prob,prob1)

Distorção

alfa= sum(prob1)/sum(prob)

1-alfa

Stress

stress=sqrt(sum((prob-prob1)^2)/sum(prob))

stress

Probabilidade com a Distância de Hamming

Correlação

prob=as.matrix(read.table(“probh.txt”,h=F))

prob1=as.matrix(read.table(“probhd.txt”,h=F))

cor.test(prob,prob1)

Distorção

alfa= sum(prob1)/sum(prob)

1-alfa

Stress

stress=sqrt(sum((prob-prob1)^2)/sum(prob))

stress

Probabilidade com a Distância de Canberra

Correlação

prob=as.matrix(read.table(“probc.txt”,h=F))

prob1=as.matrix(read.table(“probcd.txt”,h=F))

cor.test(prob,prob1)

Distorção

alfa= sum(prob1)/sum(prob)

1-alfa

Stress

stress=sqrt(sum((prob-prob1)^2)/sum(prob))

stress

Probabilidade com a Distância de Jaccard

Correlação

prob=as.matrix(read.table(“probj.txt”,h=F))

prob1=as.matrix(read.table(“probjd.txt”,h=F))

cor.test(prob,prob1)

Distorção

alfa= sum(prob1)/sum(prob)

1-alfa

Stress

stress=sqrt(sum((prob-prob1)^2)/sum(prob))

stress