



UNIVERSIDADE
ESTADUAL DE LONDRINA

FABRIZIO MALAGHINI ARANHA

**MEDIDAS DE DISTÂNCIA E ALGORITMOS DE
AGRUPAMENTO NA CARACTERIZAÇÃO E AVALIAÇÃO DE
GERMOPLASMA DE *Capsicum* spp.**

Londrina
2015

FABRIZIO MALAGHINI ARANHA

**MEDIDAS DE DISTÂNCIA E ALGORITMOS DE
AGRUPAMENTO NA CARACTERIZAÇÃO E AVALIAÇÃO DE
GERMOPLASMA DE *Capsicum* spp.**

Dissertação apresentado ao Programa de Pós-graduação em Agronomia - Curso de Mestrado da Universidade Estadual de Londrina.

Orientador: Prof. Dr Leandro Simões Azeredo Gonçalves.

Londrina
2015

**Catálogo elaborado pela Divisão de Processos Técnicos da Biblioteca
Central da Universidade Estadual de Londrina**

Dados Internacionais de Catalogação-na-Publicação (CIP)

A662m Aranha, Fabrizio Malaghini
Medidas de distância e algoritmos de agrupamento na caracterização e
avaliação de germoplasma de *Capsicum* spp. / Fabrizio Malaghini
Aranha. – Londrina, 2015.
61 f

Orientador: Leandro Simões Azeredo Gonçalves.
Dissertação (Mestrado em Agronomia) – Universidade Estadual de Londrina,
Centro de Ciências Agrárias, Programa de Pós-Graduação em Agronomia, 2015.
Inclui bibliografia.

1. Pimenta – Melhoramento genético – Teses. 2. Germoplasma vegetal –
Recursos – Utilização. 3. Genética vegetal – Teses. 4. Algoritmos – Teses.
I. Gonçalves, Leandro Simões Azeredo II. Universidade Estadual de
Londrina. Centro de Ciências Agrárias. Programa de Pós-Graduação em
Agronomia. III. Título.

CDU 631.52:633.841

FABRIZIO MALAGHINI ARANHA

**MEDIDAS DE DISTÂNCIA E ALGORITMOS DE AGRUPAMENTO NA
CARACTERIZAÇÃO E AVALIAÇÃO DE GERMOPLASMA DE
Capsicum spp.**

Dissertação apresentada ao Programa de Pós-Graduação em Agronomia da Universidade Estadual de Londrina.

BANCA EXAMINADORA

Orientador: Prof. Dr. Leandro Simões Azeredo
Gonçalves
Universidade Estadual de Londrina – UEL

Profa. Dra. Rosana Rodrigues Universidade
Estadual do Norte Fluminense Darcy Ribeiro –
UENF

Profa. Dra. Inês Cristina de Batista Fonseca
Universidade Estadual de Londrina – UEL

Profa. Dra. Lúcia Sadayo Assari Takahashi
Universidade Estadual de Londrina – UEL

Dra. Cláudia Pombo Sudré
Universidade Estadual do Norte Fluminense
Darcy Ribeiro – UENF

Londrina, 27 de fevereiro de 2015.

Aos meus exemplos de vida

Meus pais Ariovaldo Alves Aranha e

Maria Terezinha Malaghini Aranha

Pela educação e amor sem fim!

A minha tia Dalva Aranha Meneghel

Pelo amor e carinho

Dedico

Toda honra, toda glória

Todo louvor, a TiSenhor

Ofereço

AGRADECIMENTOS

À Deus, pela constante presença em minha vida, pelo auxílio nas minhas escolhas e conforto nas horas mais difíceis.

À Jesus Cristo pelo dom do entendimento, pela sabedoria e pela inspiração.

À Universidade Estadual de Londrina e ao Programa de Pós-Graduação em Agronomia pela oportunidade de realização do mestrado e a Capes pela concessão da bolsa de estudos.

Aos meus pais Terezinha e Ariovaldo e minha Tia Dalva, por todo amor, carinho, incentivo e pelo apoio e compreensão nos momentos difíceis, e principalmente por todos os sacrifícios que fizeram para que eu pudesse estar aqui hoje.

Ao meu orientador Leandro Simões Azeredo Gonçalves por todos os ensinamentos transmitidos durante o período do curso. A todos os professores e funcionários do programa de Pós-Graduação em Agronomia da Universidade Estadual de Londrina que de maneira ou de outra contribuíram para minha formação.

À Professora Rosana Rodrigues, pelos dados fornecidos, que foram utilizados na dissertação.

Aos amigos que fiz durante essa jornada, que espero levar para a vida toda, Fernando Tupick, Ronan Colombo, Julia Abati, Jaqueline Cazado Felix.

Muito obrigado a todos que de qualquer forma, foram responsáveis de uma forma ou de outra pela concretização desse sonho.

ARANHA, Fabrizio Malaghini. **Medidas de distância e algoritmos de agrupamento na caracterização e avaliação de germoplasma de *Capsicum* spp.** 2015. 61 f. Dissertação (Mestrado em Agronomia) – Universidade Estadual de Londrina, Londrina, 2015.

RESUMO

Com o aumento da perda da variabilidade genética, devido a queimadas, desmatamentos e a procura por genótipos mais adaptados e produtivos a caracterização e a avaliação dos genótipos conservados em um banco de germoplasma são de elevada importância. A caracterização e a avaliação podem ser obtidas de várias formas, gerando, eventualmente, dados quantitativos e qualitativos. Uma análise conjunta dessas variáveis pode ser considerada uma excelente estratégia para a avaliação do germoplasma. Dessa forma, esse trabalho foi desenvolvido com o objetivo de avaliar diferentes medidas de distância e agrupamentos para caracterização e avaliação de acessos de *Capsicum* spp.. Foram utilizados dados de caracterização de uma coleção de 54 acessos de *Capsicum* spp. do banco de germoplasma da Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF). As distâncias utilizadas foram seis quantitativas (A1 – média das diferenças absoluta dos rank-padronizados, A2 – correlação de Pearson, A3 – Kulczynski, A4 – Canberra, A5 – Bray-Curtis, e A6 – Morisita) combinadas com a distância para dados qualitativos – Coincidência Simples (B1). Os agrupamentos foram realizados pelos métodos hierárquicos aglomerativos (UPGMA, e Ward) e não-hierárquicos (K-médias e PCAmix). Todas as distâncias combinadas foram altamente correlacionadas. Pelo coeficiente de correlação cofenética entre as matrizes de agrupamentos hierárquicos e de distância combinada, o agrupamento UPGMA obteve os maiores valores, entretanto aplicando o coeficiente aglomerativo, o agrupamento Ward obteve os maiores valores em relação ao UPGMA. Quando foi utilizado como metodologia a representação gráfica da área sob a função da densidade acumulativa na determinação do número ótimo de grupos com diferentes distâncias combinadas, três grupos foram considerados como número ideal para maioria dos agrupamentos. Apenas para as distâncias A5B1 e A6B1 utilizando o Ward, e A5B1 utilizando K-médias não foi possível determinar o número ótimo de grupos. Comparando os agrupamentos obtidos pelo método Ward e UPGMA, verifica-se que o método Ward obteve maior eficiência em maximizar as dessemelhanças entre os complexos *C. annuum* e *C. baccatum* nas distâncias combinadas A2B1, A3B1, A5B1 e A6B1. A análise do PCAmix permitiu a separação dos acessos em relação a espécies, utilizando simultaneamente dados quantitativos e qualitativos, demonstrando ser uma alternativa para análise simultânea dos dados conjuntos, visando uma comparação entre diferentes agrupamentos.

Palavras-chave: Divergência genética. Métodos de agrupamento. Pimenta. K-Médias. Pcamix. Estrutura genética.

ARANHA, Fabrizio Malaghini. **Measures of distance and clustering algorithms in the characterization and evaluation of germplasm of Capsicum spp.** 2015. 61 p. Dissertation (Masters in Agronomy) – Universidade Estadual de Londrina, Londrina, 2015.

ABSTRACT

The increasing loss of genetic variability makes highly relevant the characterization and assessment of the genotypes stored in a germplasm bank. Such performance can be reached through several ways producing eventually quantitative and qualitative data. This way, a joint analysis of those variables may prove to be a significant strategy aiming at a precise characterization of the germplasm. This work had the objective to assess different distance measures and groupings for characterization and assessment of the *Capsicum* spp. accessions. For characterization, data were collected from a 54 accessions collection from the germplasm bank of the Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF). For this study six quantitative distances were used (A1 - absolute difference average of the standard rank, A2 – Pearson correlation, A3 – Kulczynski, A4 – Canberra, A5 – Bray-Curtis, and A6 – Morisita) and combined with distance for qualitative data – Simple Coincidence (B1). The groupings were carried out by using hierarchical agglomerative methods (UPGMA and Ward) and non-hierarchical (K-means and PCAMIX). All the distances combined were highly correlated. From the coefficient of correlation between the hierarchical grouping matrixes and combined distance, the UPGMA groupings reached the highest values, however when applying the agglomerative coefficient the Ward groupings reached the highest values compared to UPGMA. When using graphic area representation under the accumulative density function for determining the optimal group number with different distances combined, three groups were considered as the ideal number for the majority of groupings. However, the distances A5B1 and A6B1 using Ward, and A5B1 using K-means the optimal number of groups could not be determined. By comparing the groups reached by Ward and UPGMA methods it is possible to state that Ward method showed highest efficiency towards maximizing the dissimilarities between the *C. annuum* and *C. baccatum* complexes towards the distances combined A2B1, A3B1, A5B1 and A6B1. The analysis of the PCAmix allowed the separation of the accessions in relation to the species by using simultaneously both quantitative and qualitative data, proving to be an alternative for simultaneous analysis of mix data when seeking for a comparison between the different groupings.

Keywords: Multivariate analysis. Germplasm bank. Genetic diversity. Clustering methods. Pepper.

LISTA DE FIGURAS

- Figura 4.1** - Representação gráfica da área sob a função da densidade cumulativa (FDC) (ΔK) em relação ao número de grupos (k) para os agrupamentos hierárquicos (UPGMA e Ward) e K-médias em relação às distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples), Londrina, PR. 2015.....36
- Figura 4.2** - Representação gráfica do índice Calinski&Harabasz (1974) em relação ao número de grupos (k) para os agrupamentos hierárquicos (UPGMA e Ward) e K-médias em relação às distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples), para 56 acessos de *Capsicum spp.*, Londrina, PR.201537
- Figura 4.3** - Análise PCAmix para 56 acessos de *Capsicum spp.*, Londrina, PR.2015.....42

LISTA DE TABELAS

- Tabela 4.1** - Correlação entre matrizes de distâncias conjuntas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morista; e qualitativa B1: coincidência simples,Londrina,PR.201533
- Tabela 4.2** - Coeficiente de correlação cofenética entre as matrizes de agrupamento hierárquico (UPGMA e WARD) e de distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples, Londrina,PR.2015.....34
- Tabela 4.3** - Coeficiente aglomerativo entre as matrizes de agrupamento hierárquico (UPGMA e WARD) e de distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples,Londrina,PR.2015.....35
- Tabela 4.4** - Valor da robustez das médias dos grupos pelos agrupamentos UPGMA, Ward e K-médias utilizando as distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples,Londrina,PR.2015.....38
- Tabela 4.5** - Agrupamento hierárquico Ward usando distância conjunta (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples) para 56 acessos de Capsicum spp.,Londrina,PR.201539
- Tabela 4.6** - Agrupamento hierárquico UPGMA usando distância conjunta (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples) para 56 acessos de Capsicum spp.,Londrina,PR.201541

SUMÁRIO

1	INTRODUÇÃO	12
2	REVISÃO DE LITERATURA	13
2.1	CONSERVAÇÃO DE RECURSOS GENÉTICOS.....	13
2.2	BANCO DE GERMOPLASMA	14
2.2.1	Importância	14
2.2.2	Tipos de Banco de Germoplasma	15
2.2.3	Caracterização e Avaliação do Germoplasma	16
2.3	PROCEDIMENTOS ESTATÍSTICOS NA CARACTERIZAÇÃO E AVALIAÇÃO DO GERMOPLASMA	17
2.3.1	Tipos de Variáveis e Medidas de Distância.....	17
2.3.2	Agrupamento Hierárquico Versus Não Hierárquicos.....	19
2.3.3	Análise Conjunta (Variáveis Qualitativas e Quantitativas).....	22
2.4	<i>CAPSICUM</i> SPP	24
2.4.1	Origem e Botânica.....	24
2.4.2	Importância Econômica e Nutricional	25
2.4.3	Bancos de Germoplasma de <i>CAPSICUM</i> sua Utilização	26
3	MATERIAL E MÉTODOS	28
3.1	MATERIAL VEGETAL	28
3.2	CARACTERÍSTICAS AGRONÔMICAS E MORFOLÓGICAS AVALIADAS	28
3.3	ANÁLISE ESTATÍSTICA	30
4	RESULTADOS E DISCUSSÃO	3
	CONCLUSÕES	43
	REFERÊNCIAS	44
	APÊNDICES	51

APÊNDICE A – Figura 1 – Agrupamento hierárquico Ward utilizando a distância combinada de Gower para 56 acessos de <i>Capsicum spp</i>	52
APÊNDICE B – Figura 2 – Agrupamento hierárquico Ward utilizando a distância combinada (Pearson e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	53
APÊNDICE C – Figura 3 – Agrupamento hierárquico Ward utilizando a distância combinada (Kulczynski e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	54
APÊNDICE D – Figura 4 – Agrupamento hierárquico Ward utilizando a distância combinada (Canberra e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	55
APÊNDICE E – Figura 5 – Agrupamento hierárquico Ward utilizando a distância combinada (Bray-Curtis e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	56
APÊNDICE F – Figura 6 – Agrupamento hierárquico Ward utilizando a distância combinada (Morisita e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	57
APÊNDICE G – Figura 6 – Agrupamento hierárquico Ward utilizando a distância combinada (Morisita e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	58
APÊNDICE H – Figura 8 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Pearson e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	59
APÊNDICE I – Figura 9 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Kulczynski e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	60
APÊNDICE J – Figura 10 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Canberra e coincidência simples) para 56 acessos de <i>Capsicum spp</i>	61

APÊNDICE K – Figura 11 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Bray-Curtis e coincidência simples) para 56 acessos de <i>Capsicum</i> spp.....	62
---	----

1 INTRODUÇÃO

Com a conscientização crescente dos riscos de extinção e da perda de variabilidade genética, centros de conservação têm sido estabelecidos mundialmente. Além disso, há uma demanda crescente pelo aumento da produtividade agrícola, melhoramento de características nutricionais e ampliação da base genética para resistência/tolerância a condições de estresses bióticos e abióticos.

A caracterização e a avaliação dos genótipos conservados são de elevada importância, pois contribuem para maior conhecimento destes, viabilizando a detecção de possíveis genótipos a serem utilizados em programas de melhoramento de plantas e identificação de possíveis duplicatas nos bancos de germoplasma.

A caracterização e a avaliação de germoplasma podem ser obtidas por meio de informações agronômicas, morfológicas, citológicas, bioquímicas e moleculares, nas quais, os dados envolvem frequentemente medidas numéricas e categóricas, e em muitos casos, combinações de tipos de variáveis diferentes (quantitativas e qualitativas). Embora uma análise conjunta das variáveis quantitativas e qualitativas seja potencialmente um indicador mais completo da variabilidade existente nos bancos poucos trabalhos têm se utilizados dessa estratégia, devido principalmente à falta de conhecimento das técnicas estatísticas que permitem essa abordagem, a carência de softwares livres, juntamente com a tendência dos pesquisadores de darem mais importância àquelas variáveis que estejam diretamente relacionadas com os caracteres a serem melhorados em um programa de melhoramento.

A proposta central do presente trabalho foi avaliar diferentes técnicas de agrupamentos para caracterização e avaliação de acessos de *Capsicum* spp. utilizando variáveis quantitativas e qualitativas. Como objetivos específicos: *i*) comparar métodos de agrupamentos com mistura de variáveis; *ii*) identificar a combinação mais adequada para realizar estudo de diversidade genética em *Capsicum* spp. quando se têm mistura de variáveis; e *iii*) avaliar e comparar a eficiência dos agrupamentos obtidos por meio de combinação de variáveis quantitativas e qualitativas com os métodos hierárquicos (UPGMA, WARD) e não-hierárquicos (K-médias e PCAmix).

2 REVISÃO DE LITERATURA

2.1 CONSERVAÇÃO DE RECURSOS GENÉTICOS

Constata-se uma diminuição da diversidade intraespecífica das plantas cultivadas, consequência da incessante busca, por parte dos agricultores e melhoristas, por genótipos com maior homogeneidade e adaptabilidade (HAUSSMANN et al., 2004). Essa busca tem levado ao desaparecimento de variedades locais (*landraces*), sem que estas estejam devidamente caracterizadas e avaliadas para uma eventual utilização em programas de melhoramento (GOEDERT, 2006). Esse processo é conhecido como erosão genética. Com o intuito de evitar a perda dessa diversidade, observa-se o aumento da importância da conservação dos recursos genéticos, levando o surgimento de várias expedições para a coleta e armazenamento de germoplasma (NASS, 2001).

A conservação dos recursos genéticos, uma das questões mais eminentes entre melhoristas e cientistas de áreas afins. Isto porque a exploração da biodiversidade é válida não somente para programas de melhoramento genético de plantas, como também para programas de segurança alimentar, indústrias farmacêuticas, entre outros (CALLOW et al., 1997).

Lleras(1988) definiu quatro tipos de prioridades que devem ser consideradas na coleta de recursos genéticos: *i*) resgate de cultivares que estão sendo substituídas por outra de maior potencial, *ii*) resgate de variedades locais pela alta adaptabilidade as condições ambientais específicas, *iii*) áreas sujeitas a mudanças severas em função de alterações drásticas no ambiente. Sendo assim, a conservação dos recursos genéticos envolve as seguintes etapas: aquisição, multiplicação, armazenamento e manejo do germoplasma, sendo o manejo dividido em: caracterização e avaliação, regeneração e multiplicação, documentação, seleção e pré-melhoramento (ALVES et al., 2010).

Existem várias formas de conservação, dentre estas, a conservação *in situ*, a qual permite a continuidade da evolução, não só da espécie em questão, mas de todo o bioma em que está inserida. Outra forma de conservação é *ex situ*, que consiste na conservação forada local de ocorrência. Para tanto, são organizados bancos de germoplasma, na forma de sementes, plantas cultivadas *in vitro*, entre outros.

2.2 BANCO DE GERMOPLASMA

2.2.1 Importância

O termo germoplasma refere-se como a base física da herança genética, transmitida de uma geração para outra por meio de células reprodutivas (IBPGR, 1991). É considerado como material hereditário, toda estrutura ou propágulo capaz de regenerar um organismo vivo completo. Portanto, dependendo da espécie, são considerados germoplasmas: células, embriões, sementes, óvulos, espermatozoides, pólen, estacas e indivíduo adulto. (COSTA; SPEHAR, 2012).

Segundo Zimmerman e Teixeira (1996) bancos de germoplasma são locais onde são armazenados em condições adequadas, amostras de genótipos, variedades melhoradas, variedades locais, espécies silvestres e relacionadas a uma determinada espécie de interesse, todas genericamente denominadas de acessos. Possuem como objetivo principal o de preservar a variabilidade genética existente em determinada(s) espécie(s) em um ambiente controlado (NASS, 2001).

Uma justificativa para a conservação desses recursos genéticos se deve ao fato de que apenas 30% das espécies constituem 95% da nutrição humana oriunda dos vegetais, das quais se exige o máximo de eficiência produtiva para suprir a demanda da população. Porém, caso essas culturas apresentassem problemas com estresses bióticos e/ou abióticos, a alimentação humana, ou parte dela ficaria seriamente comprometida, se não dispusesse de acessos resistentes e/ou tolerantes, mantidos nos bancos de germoplasma, que pudessem ser utilizados no melhoramento genético (WALTER et al., 2007).

Em relação à América Latina, o Brasil possui o maior número de acessos conservados em bancos de germoplasma e a melhor estrutura para a conservação em longo prazo com uma capacidade para armazenar 250 mil acessos. No panorama mundial, o Brasil ocupa a sétima posição em relação ao número de acessos conservados em longo prazo (KNUDSEN, 2000).

A manutenção de Bancos de Germoplasmas no Brasil é efetuada principalmente pelo Embrapa Recursos Genéticos e Biotecnologia e instituições dos governos estaduais, como o Instituto Agrônomo de Campinas (IAC). Esses Bancos de germoplasmas normalmente são mantenedores de um grande número de

coleções de espécies diversas, sendo assim, são conservados no Cenargen cerca de 75 mil acessos, aproximadamente 400 espécies de plantas. Estima-se a existência de mais de 250 Bancos ativos de Germoplasma em todo território nacional (VEIGA, 2012).

Ainda segundo Veiga (2012) no mundo existem aproximadamente mais de 1300 Bancos ativos de Germoplasmas de plantas ou organismos econômicos. Incluem-se nesses dados, os Bancos ativos de Germoplasmas mantidos pela comunidade internacional, no Grupo Consultivo para a Pesquisa Agrícola Internacional (CIAR- www.warda.cgiar.org) com 15 centros internacionais- *International Maize and Wheat Improvement Center (Cimmyt)*, *International Center for Tropical Agriculture (CIAT)*, *International Potato Center (CIP)*, *International Institute of Tropical Agriculture (IITA)*, *International Center for Agricultural Research in the Dry Areas (ICARDA)*, *International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)*, *International Rice Research Institute (IRRI)* e *West Africa Rice Development Association (WARDA)*; além de outros que atuam apoiando áreas diversas em recursos genéticos – *Center for International Forestry Research (CIFOR)*, *World Agroforestry Centre (ICRAF)*, *International Livestock Research Institute (ILRI)*, *International Water Management Institute (IWMI)*, *International Food Policy Research Institute (IFPRI)* e *Bioversity International* (antigos IPGRI e IBPGR).

2.2.2 Tipos de Banco de Germoplasma

Bancos de Germoplasma funcionam como uma forma de conservação *ex situ*, em que amostras da variabilidade genética de determinada espécie é conservada em condições controladas (NASS, 2001). Segundo Neto (2004), os principais tipos de coleção *ex situ* são: ativa, trabalho, campo, *in vitro*, criopreservação, nuclear e genômico.

Na coleção base, para espécies com semente ortodoxa, os acessos são conservados a longo prazo, com grau de umidade entre 3 e 7% e temperatura entre -18°C e -20°C. A coleção ativa conserva os acessos a médio prazo, com temperaturas superiores a zero grau e umidades iguais a da coleção base. A estrutura física que conserva a coleção ativa é chamada de banco ativo de germoplasma (BAG), e suas principais funções são: multiplicação, caracterização e

avaliação e distribuição dos acessos. A coleção de trabalho, mantém o germoplasma armazenado a curto prazo, e está na maioria das vezes está ligada a um programa de melhoramento genético. As coleções de campo e *in vitro*, conservam espécies com sementes recalcitrantes ou com propagação vegetativa. Coleção em criopreservação utiliza a técnica que preserva o germoplasma a -196°C , em nitrogênio líquido, sem limite de tempo, sendo conservados na forma de pólen, sementes, calo, ápice caulinar, meristema e embrião (WITHERS, 1989).

2.2.3 Caracterização e Avaliação do Germoplasma

Devido a sua importância para programas de melhoramento e para a conservação da diversidade genética, a caracterização dos acessos possibilita a quantificação e utilização da variabilidade genética de modo eficiente (VALLS, 2007).

A caracterização e a avaliação de um germoplasma visam, basicamente, descrever um acesso pelas suas características morfológicas, fisiológicas, agronômicas, bioquímicas, citogenéticas ou moleculares. Essas características são úteis para a identificação de genes ou genótipos de interesse (RAMANATHA RAO, 2001). De acordo com Valls (2007), o processo de caracterização e avaliação de germoplasma pode ser classificado em cinco etapas subsequentes e complementares: a) identificação botânica; b) cadastro dos acessos (preenchimento dos dados de passaporte); c) caracterização; d) avaliação preliminar; e) avaliação.

A caracterização é feita com a utilização de descritores chamados qualitativos, que geralmente são controlados por poucos genes que sofrem pouca influência ambiental e com alta reprodutibilidade. Além da busca por novos genes, a caracterização busca detectar duplicatas dos acessos dentro de uma coleção, em função de sinonímias e diferentes doadores de mesmo germoplasma. A presença de duplicatas aumenta o custo das atividades e o uso de recursos humanos para conservar o mesmo germoplasma (SILVA et al. 2001). A caracterização de germoplasma pode ser realizada por diferentes métodos, incluindo desde práticas tradicionais, que envolvem o uso de lista de descritores morfológicos e agronômicos, até aplicações bioquímicas para detectar diferenças entre isoenzimas, proteínas e marcadores moleculares (VICENTE et al., 2005).

Já a avaliação do germoplasma consiste na descrição de uso por meio de descritores quantitativos das características agronômicas de interesse. Esses descritores possuem controle genético mais complexo, poligênico, são bastante afetados pelo ambiente e possuem baixa herdabilidade e reprodutibilidade. Para uma completa avaliação, propõem-se a implementação de experimentos em delineamentos estatístico, com repetições, casualização, e conduzidos em locais diferentes, obedecendo os princípios básicos da experimentação agrícola (SILVA et al., 2001).

2.3 PROCEDIMENTOS ESTATÍSTICOS NA CARACTERIZAÇÃO E AVALIAÇÃO DO GERMOPLASMA

2.3.1 Tipos de Variáveis e Medidas de Distância

Para evitar classificações equivocadas das categorias de variáveis empregadas em avaliações de recursos genéticos, a definição das conceituações e particularidades torna-se necessária.

As variáveis quantitativas são aquelas que medem relações quantitativas; por conseguinte, podem ser mensuradas, isto é, representadas pelo seu valor numérico. Há dois tipos de variáveis quantitativas: contínuas e descontínuas (discretas). As quantitativas contínuas são aquelas que assumem quaisquer valores entre certos limites (tais como, altura, diâmetro e produtividade), enquanto as discretas, por outro lado, assumem apenas valores específicos que não são observados em uma escala contínua por causa da existência de descontinuidades (lacunas) entre os valores. Geralmente expressa por números inteiros, exemplo de variável discreta é a observação do número de plantas mortas por uma doença em uma unidade experimental (AMARAL JÚNIOR et al., 2010).

As variáveis qualitativas são aquelas que expressam relações mutuamente exclusivas, portanto, não-mensuráveis, requerendo, assim, uma codificação. Podem ser binárias ou multicategóricas. As variáveis qualitativas binárias são aquelas que exibem apenas dois estados, como a presença ou ausência de uma dada característica. As variáveis qualitativas são multicategóricas quando existem mais de duas características mutuamente exclusivas (AMARAL JÚNIOR et al., 2010).

Uma variável é denominada ordinal quando pode ser distribuída em determinado número de categorias mutuamente exclusivas, porém, possuindo ordenação natural. Este é o caso de variáveis que podem ser ordenadas numa sequência lógica de magnitudes, como o estágio de desenvolvimento de uma doença (AMARAL JÚNIOR et al., 2010).

Diferentes medidas de distância são disponíveis para os tipos de variáveis; por conseguinte, o emprego da medida de distância mais adequada é de suma importância para que a variabilidade quantificada reflita, de forma mais fidedigna possível, a variação expressa pelos genótipos. Em se tratando de variáveis quantitativas, as distâncias mais comumente utilizadas têm sido a Euclidiana e a Mahalanobis (MOHAMMADI; PRASANNA, 2003).

Em relação aos dados qualitativos, diversas medidas envolvendo características binárias têm sido propostos, sobretudo relacionados a estudos de taxonomia numérica e na interpretação dos resultados obtidos por marcadores moleculares. Os coeficientes de similaridade (ou dissimilaridade) podem ser diferenciados por considerarem ou não a ausência conjunta. Dentre os diferentes coeficientes, o Jaccard é o mais empregado para estudo de recursos genéticos (MOHAMMADI; PRASANNA, 2003).

Embora uma variável que possua mais de dois níveis possa ser transformada em binária, com posterior emprego dos coeficientes de distância, esse procedimento tem sido evitado, por não possibilitar a adequada expressão da variabilidade genética presente nos genótipos avaliados. Nesse aspecto, a medida de dissimilaridade de Coincidência Simples é mais apropriada, sobretudo quando diversos descritores multicategóricos são utilizados, por possibilitar a análise combinada das características (AMARAL JÚNIOR et al., 2010).

Outra forma de estudar as variáveis conjuntamente é combinando medidas específicas para variáveis quantitativas e qualitativas por meio de um peso pré-determinado. Sarkaret al. (2014) propuseram uma mistura de seis diferentes medidas de distância combinadas, sendo três para dados quantitativos e duas para dados qualitativos. Para os dados quantitativos, foram consideradas as distâncias: a1) média da diferença absoluta padronizada, a2) correlação de Pearson, e a3) redimensionamento com base no score padronizado, para os dados qualitativos foram utilizados: b1) coincidência simples padronizado e b2) distância redimensionada com base na diferença absoluta média. Os autores verificaram que

a distância a1b2 usando k-médias proporcionou melhor alocação dos acessos de arroz avaliados no trabalho.

2.3.2 Agrupamento Hierárquico *Versus* Não-Hierárquicos

Informações sobre a estrutura genética de coleções de germoplasma é de suma importância para a conservação e utilização dos recursos genéticos (ODONG et al., 2011). A determinação da estrutura genética das coleções de germoplasma é realizado principalmente por métodos estatísticos multivariados tradicionais, tais como análise de agrupamento hierárquico aglomerativo, componentes principais e escala multidimensional, com base em dados genealógicos, agrônômicos, fisiológicos, bioquímicos e moleculares (AMARAL JÚNIOR et al., 2010).

As técnicas hierárquicas aglomerativas partem do princípio de que no início do processo de agrupamento tem-se n conglomerados, ou seja, cada conglomerado do conjunto de dados avaliado é considerado como sendo um grupo isolado. Em cada passo do algoritmo, os conglomerados vão sendo agrupados, formando novos grupos até o momento no qual todos os elementos considerados estão num único grupo (MINGOTI, 2007). Nesse sentido, devido a propriedade de hierarquia, é possível construir um gráfico chamado de dendrograma que representa a “árvore” ou a história do agrupamento.

Dentre as diferentes técnicas hierárquicas aglomerativas, as mais utilizadas na avaliação dos recursos genéticos têm sido: i) UPGMA (*Unweighted Pair Group Method Arithmetic Average*) (SOKAL; MICHENER, 1958), ii) WARD (WARD, 1963), e iii) Ligação Simples (*Single Linkage*) (SNEATH, 1957) (MOHAMMADI; PRASANNA, 2003). O método UPGMA (método de ligação simples) agrupa os acessos com base nas médias das distâncias entre estes, a partir do par mais semelhante. O procedimento de Ward consiste no agrupamento a partir das somas de quadrados dos desvios entre os acessos ou, alternativamente, a partir do quadrado da distância euclidiana. Por sua vez, o método do vizinho mais próximo fundamenta-se na identificação de acessos mais próximos a partir do par mais semelhante.

Milligan e Cooper (1985) simularam diferentes níveis de cortes em dendrogramas com base em diversas medidas de distâncias genéticas e concluíram

que o método de Ligação Simples revelou resultados menos consoantes com a estrutura genética dos materiais avaliados, ao passo que WARD e UPGMA permitiram os agrupamentos mais adequados, respectivamente, para tamanhos de grupos idênticos e diferentes.

Ao realizar a análise de agrupamento hierárquico aglomerativo, estamos interessados em responder alguns questionamentos: i) existe concordância entre a distância original e a distância entre indivíduos representados pelo dendrograma; ii) o que pode o dendrograma dizer sobre a estrutura do conjunto de dados; e iii) qual é o número ótimo de grupos para um determinado conjunto de dados?

Nesse contexto, uma medida bastante comum de concordância entre a distância original e a distância no dendrograma é o coeficiente de correlação cofenético, no qual correlaciona as duas matrizes (distâncias observadas e as distâncias recuperadas da análise de agrupamento) (SNEATH; SOKAL, 1973). Outra medida é a proposta por Kaufman e Rousseeuw (1990), coeficiente aglomerativo (CA) no qual se mede a qualidade de um agrupamento aglomerativo. Para cada indivíduo i , $d(i)$ é sua dissimilaridade em relação ao primeiro agrupamento em que foi inserido dividido pela dissimilaridade na etapa final do algoritmo. O coeficiente é então definido como:

$$CA = \frac{1}{n} \sum_i^n 1 - d(i)$$

onde n é o número total de indivíduos do conjunto de dados. Os valores do coeficiente variam entre 0 e 1, sendo valores baixos correspondem a estruturas ruins, no qual nenhum agrupamento foi encontrado. Por outro lado, valores próximos a 1 representam que estruturas muito claras foram identificadas.

Uma questão também importante é de como se deve proceder para escolher o número final de grupos que define a partição do conjunto de dados analisados ou de outra forma, em qual passo k o algoritmo de agrupamento deve ser interrompido (MINGOTI, 2007). Infelizmente, não existe qualquer procedimento de seleção padrão e objetivo. Existem inúmeras propostas conforme observado pelo pacote NbClust do programa R (<http://www.r-project.org>).

Milligan e Cooper (1985) apresentaram um estudo comparativo de 30 critérios de corte para determinação do número de agrupamentos, e utilizando

dados artificiais com número conhecido de agrupamentos, mostraram que critérios diferentes podem conduzir a resultados muito discrepantes.

Faria et al. (2012) compararam três métodos (Mojena, Tocher e RMSSTD) para determinar o número de grupos em estudo com aplicação de métodos hierárquicos com dados obtidos a partir da caracterização de acessos de *Capsicum*, sendo verificado que o método RMSSTD (Root Mean Square Standard Deviation) permitiu um maior poder de discriminação.

Diferentemente dos métodos hierárquicos, nos procedimentos não-hierárquicos (método divisivo) já se sabe, a priori o número k de grupos a serem formados antes mesmo de se iniciar a análise (FERREIRA, 2011). Em um exemplo simples, o primeiro passo é selecionar uma “semente de agrupamento” (centroides) como o centro inicial de um agregado, e todos os indivíduos dentro de uma distância de referência pré-especificada são incluídos no agrupamento resultante. Em seguida, outra semente de agrupamento é escolhida, e a designação continua até que todos os indivíduos tenham sido designados (HAIR et al., 2005).

Uma das principais vantagens dos métodos não-hierárquicos em relação aos métodos hierárquicos é a possibilidade de um padrão poder mudar de agrupamento com a evolução do algoritmo. Entretanto, a desvantagem está no fato do número de agrupamentos ter que ser escolhido *a priori*, o que pode inferir em interpretações errôneas sobre a estrutura dos dados caso o número de agrupamentos não seja o ideal. O problema quando se escolhe erroneamente o número de agrupamentos é que o método irá impor uma estrutura aos dados, no lugar de buscar a estrutura inerente a estes (FUNG, 2001; KAINULAINEN, 2002).

Entre os diferentes métodos não-hierárquicos, o k -médias (HARTINGAN; WONG, 1979) é o mais popular (MINGOTI, 2007; FERREIRA, 2011). O método é composto por quatro etapas: i) primeiramente escolhe-se k centroides para se inicializar o processo de partição; ii) cada elemento do conjunto de dados é, então, comparado com cada centroide inicial, por meio de uma medida de distância (em geral, a distância Euclidiana). O elemento é alocado ao grupo cuja distância é menor; iii) Para cada um dos n elementos amostrais recalculam-se os valores dos centroides para cada novo grupo formado, e repete-se o passo ii, considerando os centroides destes novos grupos; iv) Os passos ii e iii devem ser repetidos até que todos os elementos amostrais estejam “bem alocados” em seus grupos (MINGOTI, 2007).

Vidigal (2013) avaliou várias formas de agrupamento (hierárquicos - ligação simples, ligação completa, UPGMA e Ward; e não hierárquicos – K-médias e K-protótipos) com diferentes medidas de distância (euclidiana, euclidiana ao quadrado, euclidiana média, Mahalanobis, Manhattan, medidas combinadas e a de Gower) em dados simulados, e verificou que os algoritmos não-hierárquicos foram superiores aos hierárquicos.

2.3.3 Análise Conjunta (Variáveis Qualitativas e Quantitativas)

Na caracterização e avaliação do germoplasma os dados envolvem frequentemente medidas numéricas e distribuição em classes, e em muitos casos, combinações de tipos de variáveis diferentes (quantitativo e qualitativo) (MOHAMMADI; PRASANNA, 2003). Embora uma análise conjunta das variáveis quantitativas e qualitativas seja potencialmente um indicador mais completo da variabilidade existente nos bancos poucos trabalhos têm se utilizado desta estratégia devido principalmente à falta de conhecimento das técnicas estatísticas que permitem essa abordagem, a carência de softwares livres, juntamente com a tendência dos pesquisadores de darem mais importância àquelas variáveis que estejam diretamente relacionadas com os caracteres a ser melhorados em um programa de melhoramento (GONÇALVES et al., 2008; MOURA et al., 2010).

Uma técnica que permite a análise simultânea de dados quantitativos e qualitativos foi proposta por Gower (1971), no qual a similaridade entre os indivíduos X_l e X_k , $l \neq k$, é expresso na fórmula a seguir:

$$S_{lk} = \frac{\sum_{i=1}^p S_{lki} \delta_{lki}}{\sum_{i=1}^p \delta_{lki}}$$

em que S_{lki} é a similaridade entre os indivíduos X_l e X_k na variável i e δ_{lki} é uma função indicadora que indica se é possível, comparar tais indivíduos para determinada variável i .

No caso em que se têm variáveis contínuas, a similaridade δ_{lki} é expressa como:

$$S_{lki} = 1 - \frac{|x_l - x_k|}{R_i}$$

em que x_l e x_k são os valores assumidos pelos indivíduos X_l e X_k , $l \neq k$ para a variável i . R_i é a amplitude da variável i , fazendo com que o coeficiente S_{lki} esteja

entre 0 e 1. Para o caso das variáveis categóricas, a similaridade é dada por uma função que recebe 1 caso dois indivíduos sejam similares e 0, caso contrário.

Moura et al. (2010) ao avaliarem a coleção de germoplasma de *C.chinense*, com base em 44 descritores morfoagronômicos (37 qualitativos e sete quantitativos), verificaram que o agrupamento com base na distância de Gower resultou em maior eficiência na determinação da divergência genética entre os acessos avaliados.

Lawrence e Krzanowski (1966), que se basearam na fundamentação de Olkin e Tate (1961), propuseram o Location Model (LM), no qual, classifica n indivíduos quando p variáveis quantitativas e q variáveis qualitativas são obtidas em um ambiente. O LM combina níveis de todas as variáveis qualitativas em uma única variável multinomial W , com m níveis.

Franco et al. (1998) modificaram o LM e propuseram o *Modified Location Model* (MLM), assumindo que m níveis de variável W e variáveis p -multinormais para cada subpopulação são independentes. A estratégia MLM inclui duas etapas, na primeira, os grupos são definidos usando o método do Ward a partir da matriz de distância de Gower. Na segunda etapa, aplica-se o procedimento MLM, que estima a média do vetor das variáveis quantitativas para cada subpopulação independente dos valores W .

O PCAMIX é baseado na relação entre duas variáveis qualitativas ou entre uma variável qualitativa e uma quantitativa, por meio do coeficiente de associação. Para avaliar a associação, cada variável é representada por uma matriz de “quantificação”. Se a variável j_{th} é qualitativa, G_j irá denotar a matriz indicadora de ordem $n \times m_j$, sendo m_j é o número de categorias da variável j_{th} e n é o número de observações. D_j ($m_j \times m_j$) é definida com uma matriz diagonal de frequência das categorias da variável. J é a matriz de ordem $n \times n$, conhecida como operador de centralização e é dada por: $J = I - \frac{11'}{n}$. Aqui, a matriz de quantificação para a variável qualitativa j_{th} pode ser definida como:

$$S_j = J G_j D_j^{-1} G_j' . \quad (1)$$

Se a variável j_{th} é quantitativa e Z_j vetor coluna contém escores padronizados das n observações sobre a variável j , em seguida, a matriz de quantificação pode ser definida como:

$$S_j = n^{-1} Z_j Z_j' \quad (2)$$

Em termos das expressões (1) e (2), PCAMIX pode ser descrito

como o método que minimiza $f(x) = \sum \text{traces}(x' S_j X)$ sobre x , sujeito a $x'x = n/r$, onde x contém os escores dos componentes padronizados. A solução de x é dado pelo primeiro autovetor de $\sum S_j$, onde r é o número de autovalores que explica a variação máxima presente nos dados.

2.4 CAPSICUM SPP.

2.4.1 Origem e Botânica

De acordo com Eshbaugh (1993), a Bolívia é considerada o país de origem do gênero *Capsicum*. As espécies *C. annum* e *C. frutescens* foram domesticadas na Mesoamérica, enquanto *C. chinense*, *C. baccatum* e *C. pubescens* foram domesticadas na América do Sul, sendo o continente responsável pela alta diversidade genética de *C. chinense*, *C. frutescens*, *C. baccatum* e *C. pubescens* (CROSBY, 2008).

Os registros mais antigos do consumo de *Capsicum* datam de 9.000 A.C., encontrados em explorações arqueológicas realizadas em Tehuacán, México (RUFINO; PENTEADO, 2006). Na chegada dos primeiros europeus às Américas, as pimentas eram cultivadas pelos índios, ocupando lugar de grande importância na dieta alimentar, perdendo apenas para milho e mandioca. (HEISER, 1995). Logo após Colombo, ao retornar ao Novo Mundo em 1493, levando pimentas, o mesmo rapidamente se diversificou da Espanha para os outros países (BOSWELL, 1949). Em 1585 os portugueses as levaram para a Índia, posteriormente para Malaca, Indonésia e se espalhou para o Extremo Oriente (ANDREWS, 1993).

O cultivo das pimentas teve início em várias áreas utilizando diferentes espécies silvestres e essa domesticação acarretou em mudanças, principalmente nos frutos, que foram modificados, passando de pequenos, eretos, decíduos e vermelhos a grandes, frequentemente pendentes, não decíduos e com uma grande variedade de cores (HEISER, 1995).

O pimentão e as pimentas são classificados como Angiosperma, Dicotyledonea, divisão Spermatophyta, ordem Solanales, família Solanaceae,

subfamília Solanoideae e tribo Solaneae (VINALS et al., 1996). Segundo Bosland (1996), pertencendo também a essa família tem-se tomate (*Solanum lycopersicum*), a batata (*Solanum tuberosum*) e o tabaco (*Nicotiana tabacum*).

O gênero *Capsicum* é composto por 33 espécies, 27 silvestres e cinco domesticadas e usadas como vegetais frescos e temperos. As domesticadas são: *Capsicum annuum* L.; *Capsicum chinense* Jacq.; *Capsicum frutescens* L.; *Capsicum baccatum* L. e *Capsicum pubescens* Ruiz & Pav. (IBIZA et al., 2011).

Segundo Carvalho et al. (2003), as pimenteiras são plantas tipicamente herbáceas, seu tamanho varia de acordo com a espécie e as condições de cultivo, principalmente temperatura e irrigação. Suas folhas podem variar em formato, coloração, tamanho e pilosidade. As flores são tipicamente hermafroditas e preferencialmente autógamas.

2.4.2 Importância Econômica e Nutricional

De acordo com as estimativas da Fao para 2013 aproximadamente 65% da área cultivada com pimentas e pimentões se concentra no Continente Asiático, com produção de 18.021.978 toneladas. A segunda região mais importante é o continente Africano com 21% do total plantado e produção de 2.684.451 toneladas. (FAOSTAT, 2013).

No Brasil, a produção de pimenta vem crescendo muito nos últimos anos, com cultivos em regiões de clima subtropical como no Sul, ou de clima tropical como no Norte e Nordeste. No que diz respeito a agricultura familiar, tem sido um exemplo na integração pequeno agricultor-agroindústria, resultando na fixação de pequenos produtores rurais e suas famílias no campo, ajudando a evitar o êxodo rural. A contratação sazonal de mão-de-obra durante o período de colheita e o estabelecimento de novas indústrias processadoras é fundamental para geração de emprego para o cultivo dessa hortaliça (CARVALHO; BIANCHETTI, 2007).

As principais regiões brasileiras produtoras de pimenta são Sudeste e Centro Oeste. Os principais Estados produtores são Minas Gerais, Goiás, São Paulo, Ceará e Rio Grande do Sul. O mercado para as pimentas no Brasil está sofrendo grandes modificações pela exploração de novos tipos de pimentas e pelo desenvolvimento de produtos com grande valor agregado, a exemplo das conservas ornamentais (colorido de pimentas), geleias exóticas e outras formas processadas

(RUFINO; PENTEADO, 2006).

O gênero *Capsicum* exibe uma grande diversidade em relação a seus aspectos morfológicos, como forma, tamanho, cor, aroma, pungência e textura dos frutos, possuindo, tanto na forma fresca ou madura, pós ou extratos, entre outras utilidades (GOVINDARAJAN, 1987; GOVINDARAJAN; SATHYNARAYANA, 1991). Os representantes deste gênero são amplamente cultivados no mundo, sendo economicamente importantes para a indústria alimentícia, como temperos, condimentos e vegetais frescos e também para a indústria farmacêutica, por apresentar propriedades farmacológicas, além de apresentarem as vitaminas A, B1, B2, e E, principalmente, vitamina C. Possuem, ainda, proteínas, glicídios, lipídios, minerais, água e fibras (GELETA; LABUSCHAGNE, 2006).

O setor da saúde tem estimulado o aumento de pesquisas com a cultura, especialmente pela presença da substância química capsaicina, que confere o caráter ardido das pimentas, que tem sido bastante estudada. Já é conhecido que ela pode atuar como antioxidante, na cicatrização de feridas, na dissolução de coágulos sanguíneos, na prevenção da arteriosclerose e no controle do colesterol e também influencia na liberação de endorfinas, responsáveis pelo bem-estar e elevação do humor (OLIVEIRA et al., 2000).

2.4.3 Bancos de Germoplasma De *CAPSICUM* e sua Utilização

Somente a partir do final do século passado, esforços de países caribenhos e latino-americanos foram iniciados no sentido de preservar a diversidade genética em *Capsicum* spp., ainda encontrada com populações tradicionais, como ribeirinhos, indígenas e pequenos agricultores (HERNANDEZ, 2002; CARVALHO et al., 2003). No Brasil atividades conservacionistas realizadas com o germoplasma de pimentas da Amazônia se resumiam a coleta de morfotipos e a conservação em Bancos de Germoplasma (BAG), além de pequenas coleções, aliadas à conservação *in situ* realizada informalmente por populações tradicionais da região (CASALI; COUTO, 1984). Autores apontam que o Brasil deveria priorizar a coleta e a caracterização de germoplasma de *Capsicum*, por possuir uma grande variabilidade de tipos, inclusive de espécies silvestres e semidomesticadas, e considerando que o processo de erosão genética é rápido. (CARVALHO et al., 2003; BIANCHETTI; CARVALHO, 2005).

Existem várias coleções e alguns poucos Bancos Ativos de Germoplasma de *Capsicum* no Brasil. Porém, são formados, quase que exclusivamente por espécies domesticadas, não se verificando germoplasma de espécies silvestres que, em última análise, são fontes potenciais de genes de resistência, os quais poderão ser utilizados em programas de melhoramento (BIANCHETTI; CARVALHO, 2005). Entre as instituições brasileiras detentoras dos maiores bancos de germoplasma de *Capsicum* no Brasil, estão a Embrapa Hortaliças, com cerca de 650 acessos já caracterizados morfoagronomicamente (Embrapa, 2008), e a Universidade Federal de Viçosa, com 1.090 acessos, que foram caracterizados e avaliados até o ano de 2001 (SILVA et al., 2001).

Também há de se registrar o banco de germoplasma da UENF, que possui aproximadamente 250 acessos de *Capsicum* spp., caracterizados e avaliados para caracteres morfoagronômicos (SUDRÉ et al., 2010). Parte desses acessos também já foi caracterizada molecularmente (COSTA et al., 2006).

3 MATERIAL E MÉTODOS

3.1 MATERIAL VEGETAL

Foram avaliados 56 acessos de *Capsicum* spp. do banco de germoplasma da Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), usando o delineamento de blocos ao acaso, com três repetições e dez plantas por parcela, sendo 18 pertencente a *C. chinense* (acessos. 5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e acesso 58), 17 a *C. annuum* (acessos 2, 7, 8, 9, 10, 27, 28, 38, 41, 42, 43, 44, 45, 46, 52, 53 e acesso 55), 15 a *C. baccatum* (acessos 4, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 29, 35, 50 e acesso 56) e seis a *C. frutescens* (acessos 12, 16, 36, 39, 59 e acesso 60).

3.2 CARACTERÍSTICAS AGRONÔMICAS E MORFOLÓGICAS AVALIADAS

Os acessos foram caracterizados e avaliados com base em descritores morfológicos e agronômicos proposto pelo *BioversityInternational* (<http://www.bioversityinternational.org>) para *Capsicum* spp.. Para a caracterização morfoagronômica o experimento foi conduzido no município de Campos dos Goytacazes, Rio de Janeiro (21°45' S, 41°18' W).

Foram avaliados 25 descritores, sendo 14 morfológicos (qualitativos):

(MONTI et al., 2003)

Table 2. Descriptors studied and discrimination of the classes observed for 56 *Capsicum* accessions according to the IPGRI (1995).

Descriptor	Classes observed according to descriptors for <i>Capsicum</i> (IPGRI, 1995)
Morphological traits	
Stem color	1: green; 2: green with purple stripes; 3: purple
Anther color	2: yellow; 3: pale blue; 4: blue; 5: purple
Corolla color	1: white; 8: purple; 9: white with yellow-green spots; 10: white-green; 11: yellow with purple base; 12: purple with yellow base
Number of flowers per axil	1: one; 2: two; 3: three or more
Flower position	3: pendant; 5: intermediate; 7: erect
Plant growth habit	5: intermediate (compact); 7: erect
Immature fruit color	2: yellow; 3: green; 4: orange; 5: purple; 7: other
Fruit color at mature stage	1: white; 3: pale orange-yellow; 4: orange-yellow; 5: pale orange; 6: orange; 7: light red; 8: red; 9: dark red; 10: purple
Fruit shape	1: elongate; 2: almost round; 3: triangular; 4: campanulate; 5: blocky; 6: tomato-pepper; 7: ellipse; 8: Scotch bonnet
Fruit surface	1: smooth; 2: semi-wrinkled; 3: wrinkled
Number of locules	Observed cross-section of 10 cut fruits
Cotyledon color	1: light green; 5: purple; 7: variegated
Calyx annular constriction	0: absent; 1: present
Neck at base of fruit	0: absent; 1: present
Agronomic traits	
Plant height (cm)	Recorded when in 50% of plants the first fruit has begun to ripen
Plant canopy width (cm)	Measured immediately after first harvest, at the widest point
Days to flowering	Number of days from transplanting until 50% of plants have at least one open flower
Days to fruiting	Number of days from transplanting until 50% of plants bear mature fruits
Fruit length (cm)	Measured at the largest point in 10 ripe fruits
Fruit width (cm)	Measured at the widest point, as an average of 10 ripe fruits
Mean fruit weight (g)	Ratio of total fruit weight/total fruit number
1000-seed weight (g)	Estimated counting 250 seeds
Number of seeds per fruit	Counting the total number of seeds per fruit. Average of at least 10 fruits from 10 random plants
Fruit number per plant	Total number of fruits harvested in each plant
Fruit yield per plant (g)	Total weight of all harvested fruits in each plant

(SUDRE et al., 2010)

Para os caracteres agronômicos (quantitativos) foram avaliados os seguintes descritores:

- i) Comprimento do fruto (CF) - mensurado no maior comprimento em cm, quando o fruto estava maduro;
- ii) Diâmetro do fruto (DF) – mensurado na maior largura em cm, quando o fruto estava maduro;
- iii) Número de sementes por fruto (NSF) – obtido pela contagem do número total de sementes por fruto;
- iv) Altura de Planta (AP) – mensurado a maior altura da planta cm, quando 50% das plantas apresentaram frutos maduros;
- v) Diâmetro da copa (DC) – mensurado o maior diâmetro da copa em cm, quando 50% das plantas apresentaram frutos maduros;

- vi) Massa de mil sementes (MMS) – Foram pesadas 250 sementes, calculando-se o peso equivalente para 1000 sementes em gramas;
- vii) Dias para florescimento (DPF) – Número de dias do transplântio até 50% das plantas apresentarem pelo menos uma flor aberta;
- viii) Dias para frutificação (DPFR) – mensurado quando 50% das plantas da parcela apresentaram frutos maduros;
- ix) Números de frutos por planta (NFP) – somatório do número de frutos obtidos em todas as colheitas realizadas;
- x) Massa de frutos por planta (MFP) - somatório do peso de todos os frutos colhidos; e
- xi) Massa média por fruto (MMF) – Peso médio em gramas.

3.3 ANÁLISE ESTATÍSTICA

Para análise combinada das distâncias (quantitativas e qualitativas) foram consideradas seis medidas de distância para os dados quantitativos, a saber:

- i) Distância baseada na média das diferenças absoluta dos rank-padronizados (Gower)

$$A_1 = \frac{1}{p} \sum_{k=1}^p \frac{|X_{ik} - X_{jk}|}{r_k}$$

onde x_{ik} e x_{jk} são os i_{th} e j_{th} acessos da k_{th} variáveis quantitativas; r_k o ranqueamento das k_{th} variáveis; e p é o número total de variáveis quantitativas (GOWER, 1971).

- ii) Distância baseada na correlação de Pearson:

$$A_2 = (1 - r_{ij}^2)$$

onde r_{ij} é o produto da correlação (similaridade) entre i_{th} e j_{th} acessos, assim dissimilaridade = 1-similaridade.

- iii) Distância Kulczynski:

$$A_3 = 1 - 0.5 \left(\frac{\sum \min(x_{ij}, x_{ik})}{\sum x_{ij}} + \frac{\sum \min(x_{ij}, x_{ik})}{\sum x_{ik}} \right)$$

onde x_{ij} e x_{ik} são os i_{th} e j_{th} acessos;

- iv) Distância Canberra:

$$A_4 = \frac{1}{p} \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$$

onde x_{ik} e x_{jk} são os i_{th} e j_{th} acessos da k_{th} variáveis quantitativas; e p é o número total de variáveis quantitativas.

v) Distância Bray-Curtis:

$$A_5 = \frac{\sum_{k=1}^p |x_{ik} - x_{jk}|}{\sum_{k=1}^p (x_{ik} + x_{jk})}$$

onde x_{ik} e x_{jk} são os i_{th} e j_{th} acessos da k_{th} variáveis quantitativas; e p é o número total de variáveis quantitativas.

vi) Distância Morisita:

$$A_6 = 1 - \frac{2 \sum_{k=1}^p x_{ik} x_{jk}}{(\lambda_i + \lambda_j) \sum_{k=1}^p x_{ik} \sum_{k=1}^p x_{jk}}$$

onde x_{ik} e x_{jk} são os i_{th} e j_{th} acessos da k_{th} variáveis quantitativas; p é o

número total de variáveis quantitativas, $\lambda_i = \frac{\sum_{k=1}^p x_{ik} (x_{ik} - 1)}{\sum_{k=1}^p x_{ik} \sum_{k=1}^p (x_{ik} - 1)}$ e

$$\lambda_j = \frac{\sum_{k=1}^p x_{jk} (x_{jk} - 1)}{\sum_{k=1}^p x_{jk} \sum_{k=1}^p (x_{jk} - 1)}$$

Para os dados qualitativos foi utilizada a distância baseada na coincidência simples:

$$B_1 = \frac{1}{m} \sum_{k=1}^m d_k$$

Onde $d_k=0$ se $y_{ik}=y_{jk}$, do contrário $d_k=1$ (GOWER, 1971).

A amplitude dos elementos das seis matrizes de distância quantitativa (A_1 - A_6) e a distância qualitativa (B_1) situa-se entre 0 e 1. Assim, a combinação das várias matrizes distância foi computada somando as matrizes de distância correspondente para os dados qualitativos e quantitativos, a saber:

$$A_1 B_1 = ((a_{1:ij}) + (b_{1:ij}))$$

$$A_2 B_1 = ((a_{2:ij}) + (b_{1:ij}))$$

$$A_3 B_1 = ((a_{3:ij}) + (b_{1:ij}))$$

$$A_4 B_1 = ((a_{4:ij}) + (b_{1:ij}))$$

$$A_5 B_1 = ((a_{5:ij}) + (b_{1:ij}))$$

$$A_6 B_1 = ((a_{6:ij}) + (b_{1:ij}))$$

Onde (a_{1ij}) , (a_{2ij}) , (a_{3ij}) , (a_{4ij}) , (a_{5ij}) , (a_{6ij}) e (b_{1ij}) representam o ij th elementos das matrizes A_1 , A_2 , A_3 , A_4 , A_5 , A_6 e B_1 , respectivamente. Essas matrizes combinadas foram correlacionadas utilizando o teste de Mantel (1000 permutações).

Para agrupamento dos acessos de *Capsicum* spp., foram utilizados os agrupamentos hierárquicos (UPGMA e Ward) e K-médias. Posteriormente, realizou-se a correlação cofenética entre as matrizes de distância combinada com a matriz de agrupamento e análise do coeficiente aglomerativo (KAUFMAN; ROUSSEEUW, 1990).

O número ótimo de grupos (k) foi determinado como o valor de K no qual ocorre a mudança na área sob a função de densidade cumulativa (CDF) (ΔK) calculada por meio de diferentes valores de k . Suponha-se que M indica um resultado de agrupamento de fusão-consenso de ordem $N \times N$. Em seguida, um CDF empírico, definido no intervalo $[0, 1]$, é dado por:

$$CDF_{(c)} = \frac{\sum_{i < j} 1_{\{M(i,j) \leq c\}}}{\frac{N(N-1)}{2}}$$

Onde $1_{\{\dots\}}$ denota uma função indicadora, $M(i,j)$, com (i,j) sendo a entrada da matriz resultante da função de consenso M . A área sob a CDF corresponde a M computado usando a fórmula:

$$AUC = \sum_{i=2}^m [x_i - x_{i-1}] CDF(x_i)$$

Onde $\{x_1, x_2, \dots, x_m\}$ é o conjunto ordenado de entrada da matriz consenso M , com $m = N(N-1)/2$ (MONTI et al., 2003)

Após determinar o número ótimo de grupos, o melhor padrão de agrupamento do germoplasma foi determinado baseado na robustez do agrupamento. A robustez do agrupamento, foi definida pelo procedimento Simpson et al. (2010):

$$m(k) = \frac{1}{\frac{N_k(N_k-1)}{2}} \sum_{i < j} M(i,j)$$

O valor médio da robustez do agrupamento é calculado por meio dos k cluster usando o algoritmo de agrupamento. Esse algoritmo será realizado pelo pacote clusterCons do programa R.

Outro procedimento de agrupamento dos dados combinados foi realizado utilizando a análise do PCAmix utilizando o pacote PCAmixdata do programa R (<http://www.r-project.org>).

4 RESULTADOS E DISCUSSÃO

Houve correlação altamente significativa entre as matrizes de distância combinadas a 1% de probabilidade pelo teste de Mantel (Tabela 5.1). Os maiores valores de correlação foram verificados entre A1B1 x A4B1, A2B1 x A6B1 e A3B1 x A5B1, com 0,98, enquanto o menor valor foi verificado para A1B1 x A2B1, com 0,79.

Tabela 4.1 - Correlação entre matrizes de distâncias conjuntas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morista; e qualitativa B1: coincidência simples).

Matrizes	A2B1	A3B1	A4B1	A5B1	A6B1
A1B1	0,79**	0,94**	0,98**	0,90**	0,85**
A2B1		0,88**	0,84**	0,85**	0,98**
A3B1			0,96**	0,98**	0,94**
A4B1				0,92**	0,90**
A5B1					0,92**

** * Significativo a 1 e 5% de probabilidade, respectivamente, pelo teste de Mantel baseado em 1000 simulações

Essa alta correlação entre as distâncias combinadas deve-se, a diferentes fatores, tal como a semelhança entre algumas distâncias dos dados quantitativos, como por exemplo, Canberra, Bray-Curtis e Gower. A diferença entre Bray-Curtis e Canberra é a soma das distâncias ij , no qual, em Bray-Curtis a soma ocorre no interior da fração, enquanto no Canberra ocorre fora da fração. Em relação à distância de Gower, a diferença ocorre no seu denominador, sendo determinado pela amplitude dos acessos estudados em uma determinada variável k , enquanto para Bray-Curtis e Canberra esse denominador é a soma de i e j para variável k . Em relação às distâncias de Pearson, Morisita e Kulczynski há uma maior dessemelhança entre as mesmas e em comparação com Canberra, Bray-Curtis e Gower, porém, somente a distância combinada de Pearson/Coincidência Simples (A2B1) obteve correlação inferior a 0,9 quando associada com as demais distâncias combinadas (A1B1 x A2B1 = 0,79, A2B1 x A3B1 = 0,88, A2B1 x A4B1 = 0,84 e A2B1 x A5B1 = 0,85).

Para a maioria dos estudos de caracterização de germoplasma vegetal, utilizando análise conjunta dos dados quantitativos e qualitativos, a distância A1B1 é amplamente utilizada (GONÇALVES et al., 2008; ADEWALE et al., 2012;

SARTIE et al., 2012; SILVA et al., 2015). Entretanto, outras combinações podem ser empregadas visando definir com maior fidedignidade a dissimilaridade/similaridade entre os acessos.

Pelo coeficiente de correlação cofenética (CCC) entre as matrizes de agrupamentos hierárquicos e de distância combinada, o agrupamento UPGMA obteve os maiores valores (Tabela 5.2). Os valores de CCC para UPGMA variaram de 0,77 (A6B1) a 0,84 (A4B1), enquanto para Ward variaram de 0,60 (A2B1) a 0,75 (A1B1). De acordo com Rohlf (1992), valores $0,9 \geq \text{CCC}$ possuem um ajuste muito bom, $0,8 \leq \text{CCC} < 0,9$ bom ajuste, $0,7 \leq \text{CCC} < 0,8$ um ajuste ruim e $< 0,7$ muito ruim. Aplicando essa classificação de Rolf (1992) nos resultados obtidos, verifica-se que a maioria dos valores de CCC variou de bom ajuste a muito ruim.

Tabela 4.2 - Coeficiente de correlação cofenética entre as matrizes de agrupamento hierárquico (UPGMA e WARD) e de distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples).

Distância Genética	Agrupamento Hierárquico			
	UPGMA	Classificação Rolf	WARD	Classificação Rolf
Gower	0,83	Bom Ajuste	0,75	Ajuste Ruim
Pearson	0,78	Ajuste Ruim	0,60	Muito Ruim
Kulczynski	0,81	Bom Ajuste	0,68	Muito Ruim
Canberra	0,84	Bom Ajuste	0,75	Ajuste Ruim
Bray-Curtis	0,78	Ajuste Ruim	0,65	Muito Ruim
Morisita	0,77	Ajuste Ruim	0,64	Muito Ruim

Aplicando o coeficiente aglomerativo (CA), o agrupamento Ward obteve os maiores valores em relação ao UPGMA (Tabela 5.3). Diversos estudos na literatura também verificaram elevados valores de CA utilizando o Ward e tem usados esses resultados para justificar o uso desse agrupamento (CUSHMAN et al., 2010; NEGRO et al., 2010). Odonget al. (2011) avaliando coleções de coco, batata e feijão por meio de marcadores moleculares, verificaram também maiores valores de CCC para UPGMA (0,82, 0,89 e 0,92, respectivamente) quando comparado com Ward (0,74, 0,62 e 0,85, respectivamente). Entretanto, para CA houve o inverso, ou seja, elevados valores de Ward (0,97, 0,94 e 0,90, respectivamente) quando comparados com UPGMA (0,58, 0,77 e 0,67, respectivamente). Contudo, os autores sugerem que modificações devem ser realizadas antes que CA possa ser utilizado com Ward, em virtude, que esse coeficiente foi inicialmente proposto para descrever a robustez de estrutura hierárquica como obtido por UPGMA (KAUFMAN e

ROUSSEEUW, 1990). Para Ward, a distância no qual todos os grupos se unem é muitas vezes maior do que a distância no qual são unidos em um agrupamento pela primeira vez. Contudo, Mérigot et al. (2010) e Carteron et al. (2012) enfatizam que o CCC nem sempre é uma medida adequada para indicar qual o agrupamento representa fielmente a matriz de dissimilaridade inicial.

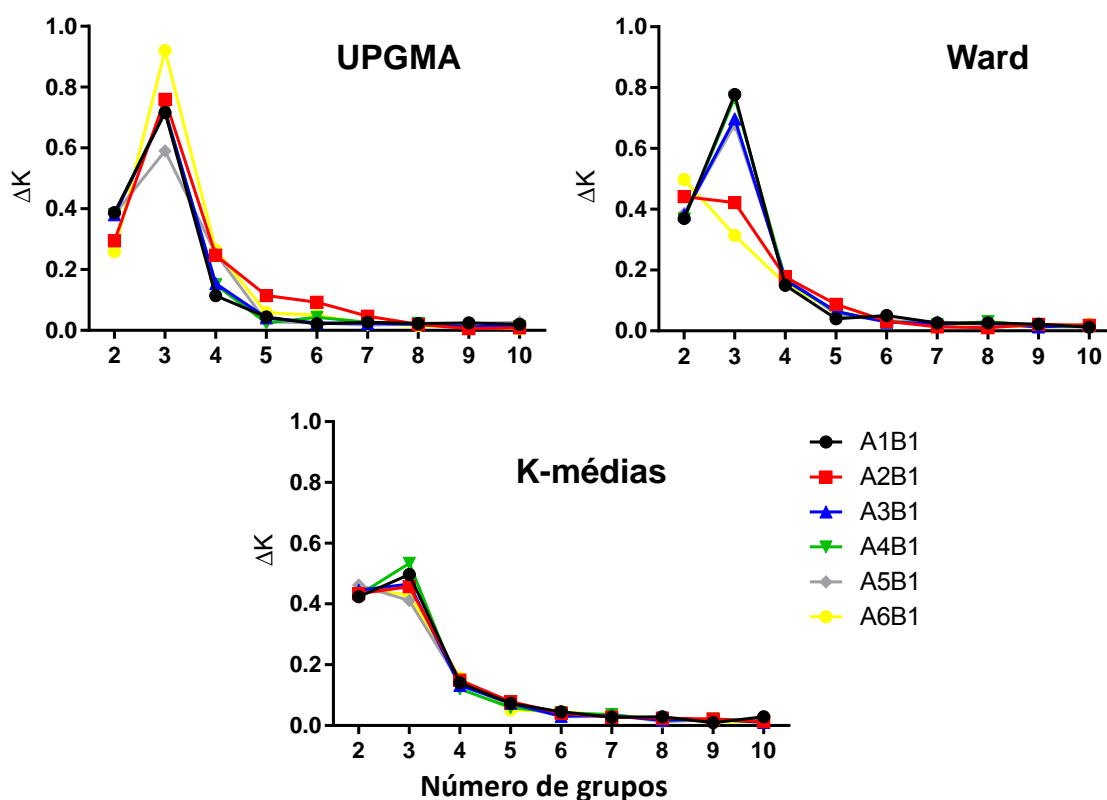
Tabela 4.3 - Coeficiente aglomerativo entre as matrizes de agrupamento hierárquico (UPGMA e WARD) e de distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples).

Distância Genética	Agrupamento Hierárquico	
	UPGMA	WARD
Gower	0,75	0,93
Pearson	0,81	0,93
Kulczynski	0,71	0,92
Canberra	0,75	0,93
Bray-Curtis	0,70	0,91
Morisita	0,76	0,93

Utilizando dados simulados, Odonget et al. (2011) verificaram que o agrupamento Ward foi eficiente na recuperação de sub-grupos originais quando os mesmos estavam presentes e distintamente separados. No entanto, na presença de muitos acessos únicos ou intermediários, os grupos formados pelo Ward não serão homogêneos. Neste caso, as diferenças na CCC entre UPGMA e Ward podem ser bastante úteis na decisão de qual selecionar. Em situações em que ambos UPGMA e Ward têm altos valores CCC ($\geq 0,8$), Ward terá muitas vantagens sobre UPGMA. Entretanto, em uma situação em que apenas UPGMA tem CCC $\geq 0,8$ e há uma grande diferença ($>0,1$) nos valores entre UPGMA e Ward, será preferível usar os grupos formados por UPGMA.

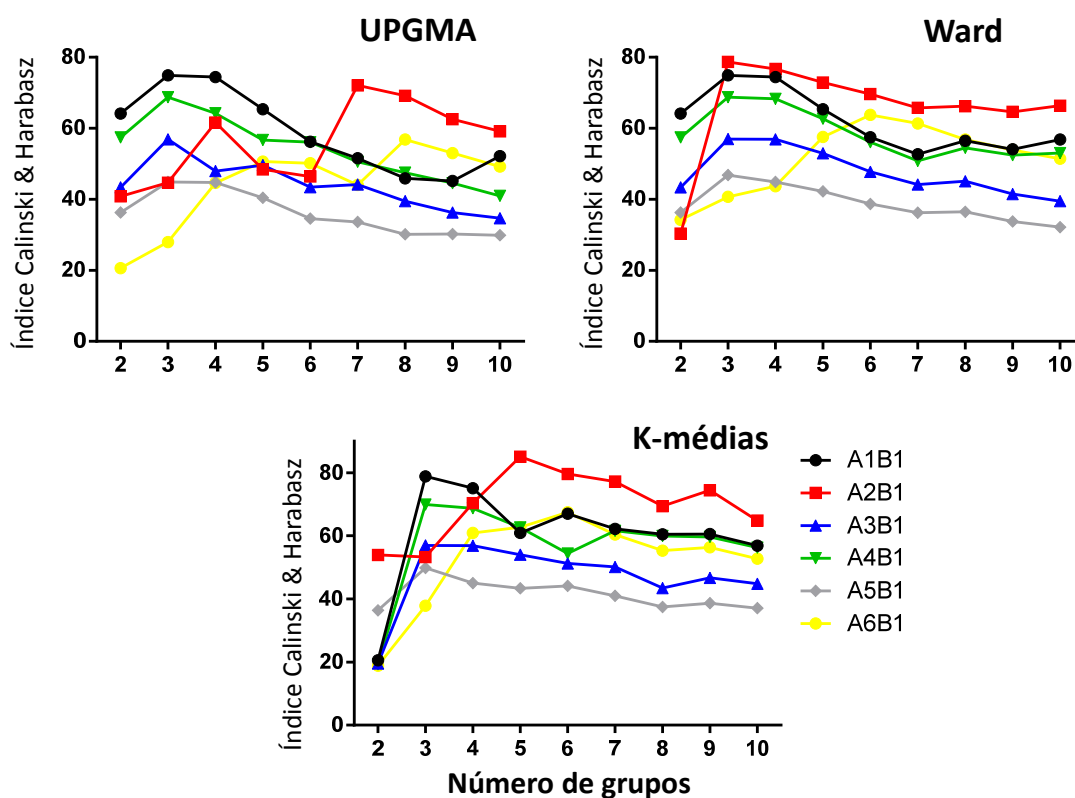
Na determinação do número ótimo de grupos utilizando as diferentes distâncias combinadas, três grupos foram considerados o ideal para maioria dos agrupamentos utilizando a representação gráfica da área sob a função da densidade acumulativa (ΔK) (Figura 5.1). Apenas para as distâncias A5B1 e A6B1 utilizando o Ward, e A5B1 utilizando K-médias não foi possível determinar o número ótimo de grupos. Sarkaret et al. (2014) utilizando essa metodologia em uma coleção de 219 acessos de arroz em seis distâncias combinadas (3 x 2 - quantitativa e qualitativa, respectivamente), verificaram que as distâncias o valor do pico de ΔK para o número de grupos foi determinado em três grupos.

Figura 4.1 - Representação gráfica da área sob a função da densidade cumulativa (FDC) (ΔK) em relação ao número de grupos (k) para os agrupamentos hierárquicos (UPGMA e Ward) e K-médias em relação às distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples).



Diversas metodologias têm sido propostas visando identificar o número ótimo de grupos. Milligan e Cooper (1985) apresentaram um estudo comparativo de 30 índices para determinação do número ótimo de grupos, no qual, verificaram que determinadas metodologias podem conduzir a resultados bastante discrepantes, sendo que os índices Calinski e Harabasz (1974), Duda e Hart's (1973) e índice-C apresentaram bons indicadores para determinação do número ótimo de grupos. Aplicando o índice Calinski e Harabasz (1974) para as diferentes distâncias combinadas nos agrupamentos UPGMA, Ward e k-médias, verifica-se também uma concordância na formação três grupos para maiorias das distâncias, exceto para distâncias A2B1 e A6B1 no agrupamento UPGMA, no qual, foi determinado em sete e oito grupos, respectivamente, A6B1 no agrupamento Ward (seis grupos foram identificados) e, A2B1 e A6B1 no agrupamento K-médias (cinco e seis grupos foram identificados, respectivamente) (Figura 5.2).

Figura 4.2 - Representação gráfica do índice Calinski&Harabasz (1974) em relação ao número de grupos (k) para os agrupamentos hierárquicos (UPGMA e Ward) e K-médias em relação às distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples).



Pela análise da robustez médias dos grupos, as distâncias A1B1 e A4B1 obtiveram os maiores valores para o agrupamento UPGMA, com 0,9723 e 0,9562, respectivamente, enquanto para as distâncias A2B1, A3B1, A5B1 e A6B1 os maiores valores foram obtidos pelo agrupamento Ward, com 0,9042, 0,9866, 0,9812 e 0,9631, respectivamente (Tabela 5.4). Sarkaret al. (2014) comparando a robustez dos grupos em sete agrupamentos em seis medidas de distâncias combinadas, verificaram que para três distâncias o agrupamento K-médias apresentou os maiores valores, enquanto para as demais distâncias o agrupamento ligação completa foi o melhor. Contudo, os agrupamentos UPGMA e Ward apresentaram valores próximos aos obtidos pelo K-médias e Ligação Completa.

Tabela 4.4 - Valor da robustez das médias dos grupos pelos agrupamentos UPGMA, Ward e K-médias utilizando as distâncias combinadas (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples).

Distâncias	Agrupamento		
	UPGMA	Ward	K-médias
A1B1	0,9723	0,8953	0,8916
A2B1	0,8564	0,9042	0,8370
A3B1	0,9138	0,9866	0,7204
A4B1	0,9562	0,9100	0,9054
A5B1	0,9252	0,9812	0,7812
A6B1	0,9304	0,9631	0,8431

Pelo agrupamento Ward, as distâncias A3B1, A5B1 e A6B1, que obtiveram os maiores valores médios de robustez, foram concordantes na classificação dos acessos nos grupos (Tabela 5.5). O 1º grupo (G2, G1, G1 e G2, respectivamente) foi constituído por 13 acessos de *C. annum*, sendo todos os pimentões (ac. 2, ac. 7, ac. 8, ac. 9, ac. 41, ac. 42, ac. 43, ac. 44, ac. 45 e ac. 46) reunidos nesse grupo, além dos acessos ac. 27, ac. 28 e ac. 55. O 2º grupo (G3, G2, G2 e G3, respectivamente) foi formado por 15 acessos, sendo todos pertencentes à *C. baccatum*, enquanto o 3º grupo alocou 28 acessos, sendo 18, seis e quatro pertencentes às espécies *C. chinense*, *C. frutescens* e *C. annum*, respectivamente.

As espécies que compõem o gênero *Capsicum* estão distribuídas em três complexos gênicos distintos com base na cruzabilidade. O complexo *C. annum* que reúne as espécies *C. annum*, *C. chinense* e *C. frutescens*; o complexo *C. baccatum* que reúne *C. baccatum* var. *baccatum* (forma silvestre) e *C. baccatum* var. *pendulum* (forma cultivada); e o complexo *pubescens* que reúne algumas espécies silvestres e apenas uma cultivada, *C. pubescens* (PICKERSGILL, 1991). Nesse contexto, as distâncias A3B1, A5B1 e A6B1 utilizando o agrupamento Ward foram eficientes na separação dos complexos *C. annum* e *C. baccatum* por meio dos descritores morfoagonômicos propostos pelo *BioversityInternational*.

Tabela 4.5 - Agrupamento hierárquico Ward usando distância conjunta (quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples) para 56 acessos de *Capsicum* spp.

Grupos	Acessos
A1B1	
G1	2, 7, 8, 9, 13, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G2	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58
G3	4, 10, 12, 14, 15, 16, 17, 19, 20, 21, 24, 25, 26, 29, 35, 36, 38, 39, 50, 52, 53, 56, 59 e 60
A2B1	
G1	5, 6, 10, 11, 12, 16, 18, 22, 23, 30, 31, 32, 33, 34, 36, 37, 38, 39, 47, 48, 49, 52, 53, 54, 57, 58, 59 e 60
G2	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G3	4, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 29, 35, 50 e 56
A3B1	
G1	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G2	4, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 29, 35, 50 e 56
G3	5, 6, 10, 11, 12, 16, 18, 22, 23, 30, 31, 32, 33, 34, 36, 37, 38, 39, 47, 48, 49, 52, 53, 54, 57, 58, 59 e 60
A4B1	
G1	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G2	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58
G3	4, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21, 24, 25, 26, 29, 35, 36, 38, 39, 50, 52, 53, 56, 59 e 60
A5B1	
G1	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G2	4, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 29, 35, 50 e 56
G3	5, 6, 10, 11, 12, 16, 18, 22, 23, 30, 31, 32, 33, 34, 36, 37, 38, 39, 47, 48, 49, 52, 53, 54, 57, 58, 59 e 60
A6B1	
G1	5, 6, 10, 11, 12, 16, 18, 22, 23, 30, 31, 32, 33, 34, 36, 37, 38, 39, 47, 48, 49, 52, 53, 54, 57, 58, 59 e 60
G2	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G3	4, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 29, 35, 50 e 56

À distância A2B1 utilizando o agrupamento Ward também foi concordante com as distâncias A3B1, A5B1 e A6B1 (Tabela 5.5). Contudo, as distâncias A1B1 e A4B1 foram discordantes e não eficientes na separação dos complexos de *Capsicum* spp..

Em relação ao agrupamento UPGMA, os acessos de *C. chinense* foram alocados isolados no mesmo grupo para a maioria das distâncias, exceto para distância A2B1 (Tabela 5.6). À distância A2B1 revelou a menor robustez entre as distâncias agrupadas pelo UPGMA, no qual, o grupo I alocou apenas o acesso 60

(*C. frutescens*), enquanto os grupos 2 e 3 foram discordantes em relação ao agrupamento das demais distâncias.

Os acessos de *C. baccatum* foram agrupados no mesmo grupo para todas as distâncias, contudo, não foram alocados isoladamente como verificados pelo agrupamento Ward. Pela distância A1B1 os acessos de *C. baccatum* foram agrupados com os seis acessos de *C. frutescens*, enquanto nas distâncias A3B1 e A4B1 reuniram além dos acessos *C. frutescens*, quatro acessos de *C. annuum* (ac. 10, ac. 38, ac. 52 e ac. 53). Para as distâncias A5B1 e A6B1 reuniram os acessos *C. baccatum* os acessos de *C. annuum*, sendo todos os pimentões e três acessos de pimenta (ac. 27, ac. 28 e ac. 55).

Comparando os agrupamentos obtidos pelo método Ward e UPGMA, verifica-se que o método Ward obteve maior eficiência em maximizar as dessemelhanças entre os complexos *C. annuum* e *C. baccatum* nas distâncias A2B1, A3B1, A5B1 e A6B1. Esse resultado é em função das características dos agrupamentos, em que, o método de Ward ou de variância mínima consiste em formar grupos a partir de pares que proporcionem a menor soma de quadrados, enquanto, no UPGMA a matriz de distância é atualizada calculando-se a média das distâncias entre os indivíduos de dois grupos (MOHAMMADI; PRASANNA, 2003; CROSSA; FRANCO, 2004). Odonget al. (2011) relatam que o UPGMA apresenta um bom desempenho na preservação das distâncias entre os acessos, enquanto o Ward foi excelente na recuperação dos grupos.

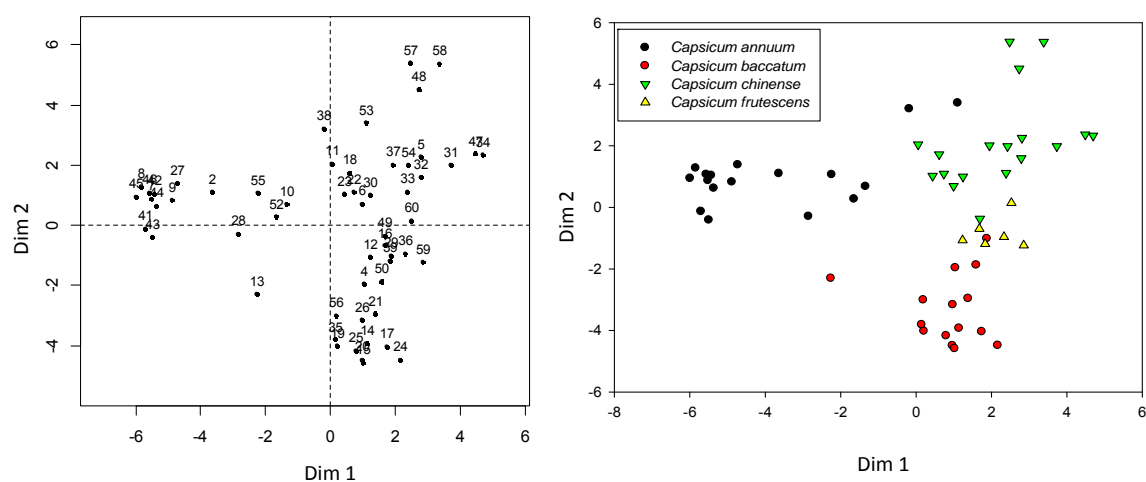
Tabela 4.6 - Agrupamento hierárquico UPGMA usando distância conjunta quantitativa: A1: Distância Gower, A2: Pearson; A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita; e qualitativa B1: coincidência simples) para 56 acessos de *Capsicum* spp.

Grupos	Acessos
A1B1	
G1	2, 7, 8, 9, 10, 27, 28, 38, 41, 42, 43, 44, 45, 46, 52, 53 e 55
G2	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58
G3	4, 12, 13, 14, 15, 16, 17, 19, 20, 21, 24, 25, 26, 29, 35, 36, 39, 50, 56, 59 e 60
A2B1	
G1	60
G2	2, 4, 6, 7, 8, 9, 11, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 37, 41, 42, 43, 44, 45, 46, 49, 50, 55 e 56
G3	5, 10, 12, 16, 31, 32, 33, 34, 36, 38, 39, 47, 48, 52, 53, 54, 57, 58 e 60
A3B1	
G1	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58
G2	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G3	4, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21, 24, 25, 26, 29, 35, 36, 38, 39, 50, 52, 53, 56, 59 e 60
A4B1 UPGMA	
G1	2, 7, 8, 9, 27, 28, 41, 42, 43, 44, 45, 46 e 55
G2	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58
G3	4, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21, 24, 25, 26, 29, 35, 36, 38, 39, 50, 52, 53, 56, 59 e 60
A5B1	
G1	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58
G2	10, 12, 16, 36, 38, 39, 52, 53, 59 e 60
G3	2, 4, 7, 8, 9, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 27, 28, 29, 35, 41, 42, 43, 44, 45, 46, 50, 55 e 56
A6B1	
G1	10, 12, 16, 36, 38, 39, 52, 53, 59 e 60
G2	2, 4, 7, 8, 9, 13, 14, 15, 17, 19, 20, 21, 24, 25, 26, 27, 28, 29, 35, 41, 42, 43, 44, 45, 46, 50, 55 e 56
G3	5, 6, 11, 18, 22, 23, 30, 31, 32, 33, 34, 37, 47, 48, 49, 54, 57 e 58

A análise do PCAMIX permitiu a separação dos acessos em relação a espécies, utilizando simultaneamente dados quantitativos e qualitativos (Figura 5.3). Para *C. annum*, os acessos ac. 2, ac. 7, ac. 8, ac. 9, ac. 41, ac. 42, ac. 43, ac. 44, ac. 45 e ac. 46, referente ao grupo pimentão, foram alocados próximos, enquanto para os demais acessos foi verificada uma maior dispersão, principalmente para ac. 38 e ac. 53 que ficaram alocados próximos aos acessos de *C. chinense*. Esse resultado foi concordante com os obtidos pelo agrupamento Ward utilizando as distâncias A2B1, A3B1, A5B1 e A6B1, no qual, os acessos do grupo pimentão ficaram alocados no mesmo grupo, enquanto os acessos ac. 38 e ac. 53 foram agrupados juntos com *C. chinense* e *C. frutescens*.

Para *C. baccatum* também houve uma aglutinação dos acessos, exceto para ac. 13 que ficou disperso. Esses resultados obtidos pelo PCAmix, no qual, tem como princípio a utilização das análises de componentes principais (ACP) e de correspondência múltipla (ACP), revela uma maior concordância com obtidos pelo agrupamento Ward, no qual, maximizou as diferenças dos complexos *C. annum* e *C. baccatum*. A análise PCAmix pode ser uma alternativa para análise simultânea dos dados quantitativos e qualitativos, visando uma comparação entre diferentes agrupamentos.

Figura 4.3 - Análise PCAMIX para 56 acessos de *Capsicum*



CONCLUSÕES

Houve elevada correlação entre as matrizes de distâncias combinadas.

Comparando os agrupamentos obtidos pelo método Ward e UPGMA, verifica-se que o método Ward obteve maior eficiência em maximizar as dessemelhanças entre os complexos *C. annuum* e *C. baccatum*.

A análise do PCAmix permitiu a separação dos acessos em relação a espécies, utilizando simultaneamente dados quantitativos e qualitativos, demonstrando ser uma alternativa para análise simultânea dos dados conjuntos, visando uma comparação entre diferentes agrupamentos.

O agrupamento Ward com as distâncias conjuntas Kulczynski, Bray-Curtis e Morisita(quantitativa), com coincidência simples(qualitativa) foram as que demonstraram ser as melhores combinações para realizar estudos de diversidade genética em *Capsicum* spp., quando se têm mistura de variáveis.

REFERÊNCIAS

- ADEWALE, B. D.; DUMET, D. J.; VROH-BI, I.; KEHINDE, O. B.; OJO, D. K.; ADEGBITE, A. E.; FRANCO, J. **Morphological diversity analysis of African yambean and prospects for utilization in germplasm conservation and breeding**. *Genetic Resources and Crop Evolution*. v.59, p. 927-936, 2012.
- ALVES, D. P.; ALVES, D. P.; SILVA, D.J.H.; MATTEDI, A.P. Conservação *ex situ* dos recursos filogenéticos. In: Telma Nair Santana Pereira. (Org.). **GERMOPLASMA: Conservação, Manejo e Uso no Melhoramento de Plantas**. 1ed. VIÇOSA: ARKA, v. 2, p. 60- 67, 2010.
- AMARAL JÚNIOR, A. T.; FREITAS JÚNIOR, S.P.; RANGEL, R.M.; PENA, G.F.; RIBEIRO, R.M.; MORAIS R.C.; SCHUELTER, A.R. Improvement of a popcorn population using selection indexes from a fourth cycle of recurrent selection program carried out in two different environments. ***Genetics and Molecular Research***, v. 9, n. 1, p. 340-347, 2010.
- ANDREWS, J. Diffusion of Mesoamerican food complex to southeastern Europe. ***The Geographical Reviews***, v. 83, p. 194–204, 1993.
- BIANCHETTI, L.B.; CARVALHO, S.I.C. Subsídio à coleta de germoplasma de espécies de pimenta e pimentões do gênero *Capsicum* (Solanaceae). In: WALTER, B.M.T.; CAVALCANTE, T.B. (Ed.). **Fundamentos para a coleta de germoplasma vegetal**. Brasília: Embrapa Recursos Genéticos e Biotecnologia, 2005. p.335-385
- BOSLAND, P. W. **Capsicums: Innovative Uses of an Ancient Crop**. Arlington, VA: ASHS Press, 1996. p. 479-487.
- BOSWELL, V.R. Garden pepper, both a vegetable and a condiment. Our vegetable travelers. ***The National Geographer***, v.96, p. 145–217, 1949.
- CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. ***Communications in statistics***, v. 3, p. 1-27, 1994.
- CALLOW, B. V.; FORD-LLOYD, B.; NEWBURY, H. J. Overview. In: CALLOW, B. V.; FORD-LLOYD, B.; NEWBURY, H. J. (Ed.) **Biotechnology and plant genetic resources: conservation and use**. Londres: CAB international, cap 01, p.05-10, 1997.
- CARTERON, A.; JEANMOUGIN, M.; LEPRIEUR, F.; SPATHARIS, S. Assessing the efficiency of clustering algorithms and goodness-of-fit measures using phytoplankton field data. ***Ecological Informatics***, v. 9. p. 64-68, 2012.
- CARVALHO, S.I. C.; BIANCHETTI, L. B.; BUSTAMANTE, P.G.; SILVA, D.B. S. Catálogo de germoplasma de pimentas e pimentões (*Capsicum* spp.) da Embrapa Hortaliças. Brasília: **Embrapa Hortaliças**, 2003, 49p. (Documentos 49).
- CARVALHO, S.I.C.; BIANCHETTI, L.B.; HENZ, G.P. Germplasm collection of *Capsicum* spp. maintained by Embrapa Hortaliças (CNPH). **Capsicum and Eggplant Newsletter**, v.22, p.17-20, 2005.

CARVALHO, S. I. C.; BIANCHETTI, L. B. - Sistema de produção de pimentas (*Capsicum* spp). Consumo: **Embrapa Hortaliças**, Sistemas de produção, 2 ISSN 1678- 880x Versão Eletrônica Novembro/2007. Disponível em: <
http://sistemasdeproducao.cnptia.embrapa.br/FontesHTML/Pimenta/Pimenta_ca_sicum_spp/importanciaeconomica.html

CASALI, V.W.; COUTO, F.A.A. Origem e botânica de *Capsicum*. **Informe Agropecuário**, v.10, n.11, p.8-10, 1984.

CHIORATO, A. F. **Divergência genética em acessos defeijoeiro (*Phaseolus vulgaris L.*) do Banco de Germoplasma do Instituto Agrônômico-IAC**. p. 85, 2004.

COSTA, F. R.; PEREIRA, T.N. S.; VITORIA, A, P.; CAMPOS, K.P.; RODRIGUES, R.; SILVA, D. H.; PEREIRA, M.G. Genetic diversity among *Capsicum* accessions using RAPD markers. **Crop Breeding and Applied Biotechnology**, v. 6, p. 18-23, 2006.

COSTA, A. M.; SPEHAR, C. R.; SERENO, J. R. B. (Ed.). **Conservação de recursos genéticos no Brasil**. Brasília, DF : Embrapa, 2012. p. 483-513.

CROSBY, K. M.; PROHENS, J.; NUEZ, F.; CARENA, M.J. (eds.). *Vegetables II*. **New York: Springer**, 2008. p. 221–248.

CROSSA, J.; FRANCO, J. **Statistical methods for classifying genotypes**. *Euphytica*, v. 137, p. 19-37, 2004

CUSHMAN, S. A.; MCKELVEY, K. S.; NOON, B, R.; MCGARIGAL, K. Use of abundance of one species as a surrogate for abundance of others. **Conserv Biol**, v. 24, p. 830–840, 2010.

DUDA, R. O.; HART, P. E. **Pattern Classification and Scene Analysis**. Wiley-Interscience Publication, New York, 1973.

EMBRAPA HORTALIÇAS: bases de dados (consultas “on-line”) <http://www.cnph.embrapa.br> em 17/01/2015.

ESHBAUGH, W. H. Peppers: history and exploitation of a serendipitous new crop discovery. In: Janick, J. e J.E. Simon (Eds). *New Crops*, 1993. p. 132-139.

FAO (2013) **Food and agriculture organization of the united nations**, versão eletrônica. Disponível em <http://faostat.fao.org/default.aspx>. Acesso em: fevereiro de 2015.

FARIA, P.N.; CECON, P.R.; SILVA, A.R.; FINGER, F.L.; SILVA, F.F.; CRUZ, C. D.; SÁVIO, F.L. Métodos de agrupamento em estudo de divergência genética de pimentas. **Horticultura Brasileira**, v.30, p.428-432, 2012.

FERREIRA, D. F. **Estatística Multivariada**, Lavras, 2011.

FRANCO, J.; CROSSA, J.; VILLASEÑOR, J.; TABAS, S.; EBERHART, S.A. Classifying genetic resources by categorical and continuous variables. **Crop Science**, v. 38, p. 1688-1696, 1998.

- FUNG, GLENN. A Comprehensive Overview of Basic Clustering Algorithms, 2001. Disponível em <<http://www.cs.wisc.edu/~gfung/clustering.pdf>>. Acesso em fev. de 2015. FUNG, G. A. **Comprehensive overview of basic clustering algorithms**. June 22, 2001.
- GELETA, L. F.; LABUSCHANGNE, M.T. Comparative performance and heterosis in single, three-way and double cross pepper hybrids. **The Journal of Agricultural Science**, v. 142, p. 659-663, 2006.
- GOEDERT, M. C. Conservação ex situ de recursos genéticos de plantas: caso Embrapa. **Magistra**. v. 18, p. 15-18, 2006.
- GONÇALVES, L. S. A.; RODRIGUES, R.; AMARAL JÚNIOR, A. T.; KARASAWA, M.; SUDRÉ, C. P. Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. **Genetics and Molecular Research**, v. 7, p. 1289-1297, 2008.
- GOVINDARAJAN, V. S. Capsicum production, technology, chemistry and quality. Part IV: evaluation of quality. **Crit Rev Food Science Nutrition**, v. 25, p. 158–282, 1987.
- GOVINDARAJAN, V. S.; SATHYANARAYANA, M. N. Capsicum production, technology, chemistry, and quality. Part V. Impact on physiology, pharmacology, nutrition, and metabolism; structure, pungency, pain, and desensitization sequences. **CRC Critical Rev. Food Sci. Nutr.** v. 29, p. 435–473, 1991.
- GOWER, J.C. A general coefficient of similarity and some of its properties. **Biometrics**, v.27, p. 857-875, 1971.
- HAIR, J.; ANDERSON, R. E.; TATHAN, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. 5ª ed.; Porto Alegre: Bookman; 2005.
- HARTIGAN, P.; WONG, M. A. A k-means clustering algorithm: algorithm AS 1366. **Applied Statistics**, n.28, p. 126-130, 1979.
- HAUSSMANN, B.I. G.; PARZIES, H. K.; PRESTERL, T.; SUSIC, Z.; IEDANER, T. Plant genetic resources in crop improvement (Review Article). **Plant Genetic Resources – Characterization and Utilization**, v. 2, n. 1, p. 3-21, 2004.
- HEISER, J. R. Peppers – *Capsicum* (Solanaceae). In: SMARTT, J.; SIMMONDS, N. W. **Evolution of crop plants**. London: Longman, p. 49- 51, 1995.
- HERNANDEZ, S.M. Genetic resources of chile (*Capsicum* spp.) in Mexico. IN: **Proceedings of the 16th international pepper conference**. Tampico, Tamaulipas, Mexico, 2002.
- IBIZA, V.P.; BLANCA, J.; CAÑIZARES, J.; NUEZ, F. Taxonomy and genetic diversity of domesticated *Capsicum* species in the andean region. **Genetic Resource and Crop Evolution**, v. 58, p. 12, 2011.

INTERNATIONAL BOARD FOR PLANT GENETIC RESOURCES. Consultative Group on International Agricultural Research. **Report of the third external review of the International Board for Plant Genetic Resources**. Rome, 1991, 85 p.

KAINULAINEN, JUKKA. **Clustering Algorithms: Basics and Visualization**, 2002. Disponível em <<http://www.niksula.cs.hut.fi/~jkainula/pdfs/clustering.pdf>>. Acesso em fev. de 2015.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data. **An introduction to cluster analysis**. Wiley, New York, 1990.

KNUDSEN, H. (Ed.). **Directorio de colecciones de germoplasma en América Latina y el Caribe**. Roma: IPGRI, 2000. 369 p.

LAWRENCE, C. J.; KRZANOWSKI, W. J. Mixture separation for mixed-mode data. **Stat. Comput.**, v. 6, p. 85-92, 1996.

LLERAS, E. Coleta de recursos genéticos vegetais. In: ARAUJO, S. M. C.; OSUNA, J. A (Eds.). **Encontro sobre recursos genéticos**, I, Jaboticabal, ANAIS..., Jaboticabal, 1988, p. 23-42.

MÉRIGOT, B.; DURBEC, J.P.; GAERTNER, J.C. On goodness-of-fit measure for dendrogram-based analyses. **Ecology**, v. 91, p. 1850–1859, 2010.

MILLIGAN, G.W.; COOPER, M. An examination of procedures for determining the number of cluster in a data set. **Psychometrika**, v.50, p. 159-179, 1985.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada**. Belo Horizonte, 2007.

MOHAMMADI, S. A.; PRASANNA, B. M. Analysis of genetic diversity in crop plants-salient statistical tools and considerations. **Crop Science**, v. 43, p. 1235-1248, 2003.

MOURA, M. C. C. L.; GONÇALVES, L. S. A.; SUDRÉ, C. P.; RODRIGUES, R.; AMARAL JÚNIOR, A. T.; PEREIRA, T. N. S. Algoritmo de Gower na estimativa da divergência genética em germoplasma de pimenta. **Horticultura Brasileira**. v. 28, p. 155-161, 2010.

MONTI, S.; TAMAYO, P.; MESIROV, J.; GOLUB, T. Consensus clustering: a resampling-based method for class Discovery and visualization of gene expression microarray data. **Machine Learning**, v. 52, p. 91-118, 2003.

NASS, L. L. Utilização de recursos genéticos vegetais no melhoramento. In NASS, L. L.; MELO, I. S.; VALADARES-INGLES, M. C.; VALOIS, A. C. C. (Eds.). **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação MT, 2001, p.29-56.

NEGRO, S. S.; CAUDRON, A. K.; DUBOIS, M.; DELAHAUT, P. & GEMMELL, N. **J. Correlation between male social status, testosterone levels, and parasitism in a dimorphic polygynous mammal**. PLoS ONE 5: e12507, p. (1-8), 2010.

NETO, L.G.P. **Germinação de sementes de soja armazenadas em banco de germoplasma**. Lavras: UFLA, 2004. 76. Dissertação de mestrado em Agronomia – Universidade Federal de Lavras, 2004.

ODONG, T.L.; VAN HEERWAARDEN, J.; JANSEN, J.; VAN HINTUM, T.J.L.; VAN EEUWIJK, F.A. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? **Theoretical and Applied Genetics**, v. 123, p. 195-205, 2011.

OLIVEIRA, A.B.; SILVA, A.M.; LOPES, C.A.; RIBEIRO, C.S.C.; LOPES, D.; CRUZ, D.M.R.; MARQUES, D.M.C.; FRANÇA, F.H.; REIFSCHNEIDER, F.J.B.; BUSO, G.S.C.; BIANCHETTI, L.B.; FERREIRA, M.E.; POZZOBON, M.T.; RESENDE, R. O.; CARVALHO, S.I.C.; PINHEIRO, V.L.; CASALI, V.W.D. **Capsicum: pimentase pimentões no Brasil**. EMBRAPA: CNPH, Brasília-DF, 2000.

OLKIN, I.; TATE, R. F. Multivariate correlation models with mixed discrete and continuous variables. **Annals of Mathematical Statistics**, v. 32, p. 448–465, 1961

PICKERSGILL, B. Cytogenetics and evolution of Capsicum L. In: TSUCHIYA, T.; GUPTA, P.K. **Chromosome engineering in plants: genetics, breeding, evolution**. Amsterdam: Elsevier, p.139-160, 1991.

RAMANATHA RAO. **Principles and concepts in plant genetic resources conservation and use**. In SAID SAAD M; RAMANATHA RAO V, eds. Establishment and management of field genebanks. p. 1-16, 2001.

ROHLF, F. J. NTSYS-pc **Numerical Taxonomy and Multivariate Analysis System**. Version 1.70. Exeter, Setauket, NY, 1992.

RUFINO, J. L. S.; PENTEADO, D. C. S. Importância econômica, perspectivas e potencialidades do mercado para a pimenta. **Informe Agropecuário**. v. 27, n. 235, p. 7-15, 2006.

SARKAR, R.K.; RAO, A.R.; WAHI, S.D.; BHAT, K.V. A comparative performance of clustering for mixture of qualitative and quantitative data – an application to black gram. **Plant Genetic Resource: Characterization and Utilization**. v. 9, n.4, p.523-527, 2011.

SARKAR, R.K.; MEHER, P.K.; WAHI, S.D.; MOHAPATRA, T.; RAO, A.R. An approach to the development of a core set of germplasm using a mixture of qualitative and quantitative data. **Plant Genetic Resource: Characterization and Utilization**. v. 12, p.1-8, 2014.

SARTIE, A.; ASIEDU, R.; FRANCO, J. Genetic and phenotypic diversity in a germplasm working collection of cultivated tropical yams (*Dioscorea* spp.). **Genetic Resources and Crop Evolution**, v. 59, n. 8, 2012.

SNEATH, P. H. A.; SOKAL, R. R. The comparison of dendrograms by objective methods. **Taxon**, vol. 11, p. 33-40, 1973.

SOKAL, R. R.; MICHENOR, C. D. A statistical method for evaluating systematic relationships. **Bulletin of the Society University of Kansas**, n. 38, p.1409-1438, 1958.

SNEATH, P. H. A. The application of computer to taxonomy. **Journal of General Microbiology**, v. 17, p. 201-226, 1957.

SUDRÉ, C. P.; GONÇALVES, L. S. A.; RODRIGUES, R.; DO AMARAL JÚNIOR, A. T.; RIVA-SOUZA, E. M.; BENTO, C.S. Genetic variability in domesticated *Capsicum spp.* as assessed by morphological and agronomic data in mixed statistical analysis. **Genetics and Molecular Research**, v. 9, p. 283-294, 2010.

VALLS, J. F. M. Caracterização de Recursos Genéticos Vegetais. In: NASS, L.L. (Org.). Recursos Genéticos Vegetais. Brasília: **Embrapa Recursos Genéticos e Biotecnologia**, 2007, p. 281-305.

VEIGA, R. F. A.; BARBOSA, W.; TOMBOLATO, A. F. C.; VALLS, J. F. M. Bancos de germoplasma: importância e organização. Em: ANA MARIA COSTA. (Org.). **Conservação de Recursos Genéticos no Brasil**. 1ed. PLANALTINA. EMBRAPA CERRADOS, v. 1, p. 105-125, 2012.

VICENTE, M.C.; GUZMÁN, F.A.; ENGELS, J.; RAMANATHA RAO, V. **Genetic Characterization and its use in decision making for the conservation of cropgermplasm**. In: THE ROLE OF BIOTECHNOLOGY, 2005, Turin. Proceedings... Turin: 2005. p.121-128.

VIDIGAL, B.C. **Avaliação de agrupamento em mistura de variáveis**. Dissertação (Estatística Aplicada e Biometria), Universidade Federal de Viçosa, 2013.

VIÑALS, F. N.; ORTEGA, R. G.; GARCIA, J.C. El cultivo de pimientos, chiles y ajies. Madrid: **Mundi-Prensa**. 1996, 607p.

WALTER, B. M. T.; CAVALCANTI, T. B. C.; BIANCHETTI, L. B. Princípios sobre coleta de germoplasma vegetal. In: NASS, L. L. (Ed.). Recursos genéticos vegetais. Brasília: EMBRAPA Recursos Genéticos e Biotecnologia, 2007. p. 193-229.

WARD JUNIOR, J. H. **Hierarchical grouping to optimize an objective function**. Journal of the American Statistical Association, v. 58, p. 236-244, 1963.

WITHERS, L. A. In vitro conservation and germplasm utilization. In: BROW, A. H. D.; FRANKEL, O. H.; MARSHALL, D. R.; WILLIAMS, J. T. **The use of plant genetic resources**. Cambridge: Cambridge University Press, 1989. p. 309-334.

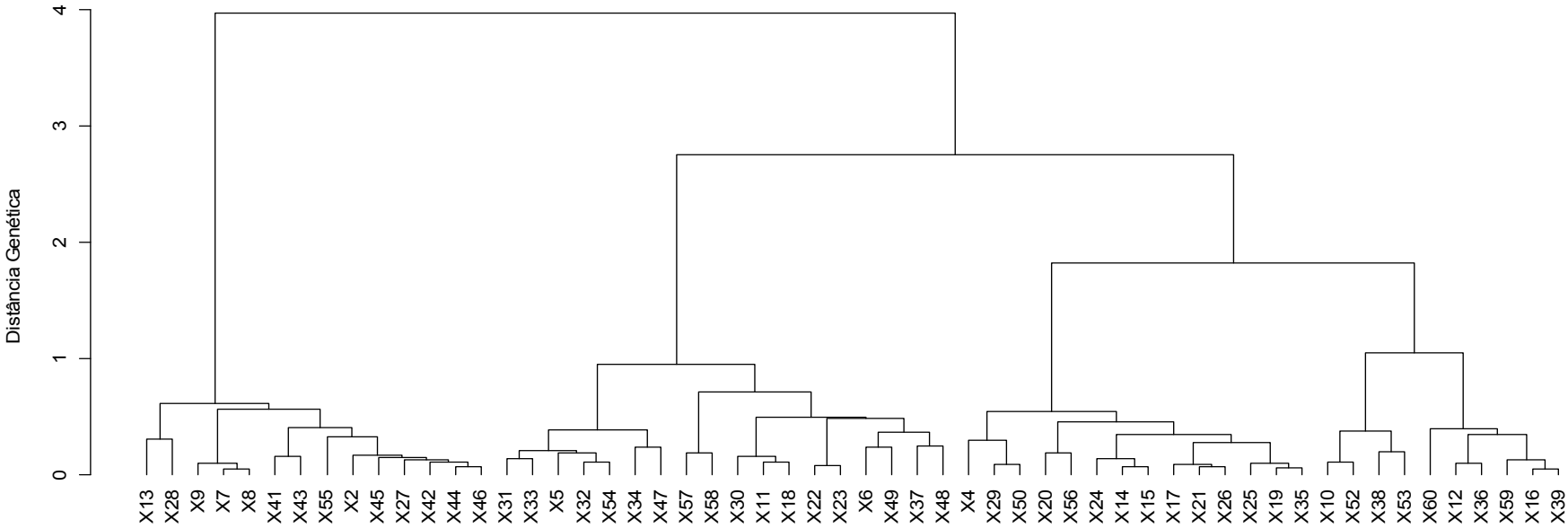
YOON, J. B.; YANG, D. C.; WAHNG, J. D.; PARK, H. G. Overcoming two postfertilization genetic barriers in interspecific hybridization between *C. annuum* L and *C. baccatum* L for introgression of Anthracnose resistance. **Breeding Science**, v. 56, p. 31-38, 2006.

ZIMMERMANN, M. J. O.; TEIXEIRA, M. G. Origem e evolução. In: ARAÚJO, R. S.; RAVA, C. A.; STONE, L. F.; ZIMMERMANN, M. J. O. eds. Cultura do feijoeiro comum no Brasil. Piracicaba: **Associação Brasileira para Pesquisa da Potassa e do Fosfato (POTAFOS)**. 1996, 786p.

APÊNDICES

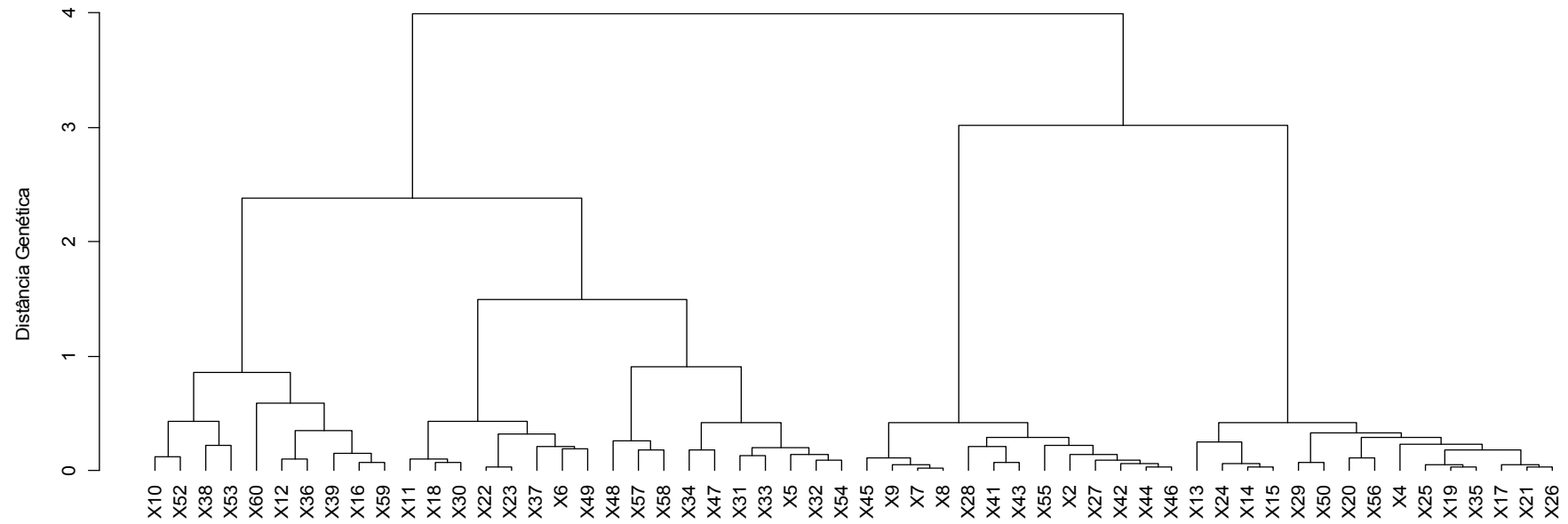
APÊNDICE A

Figura 1 – Agrupamento hierárquico Ward utilizando a distância combinada de Gower para 56 acessos de *Capsicum* spp



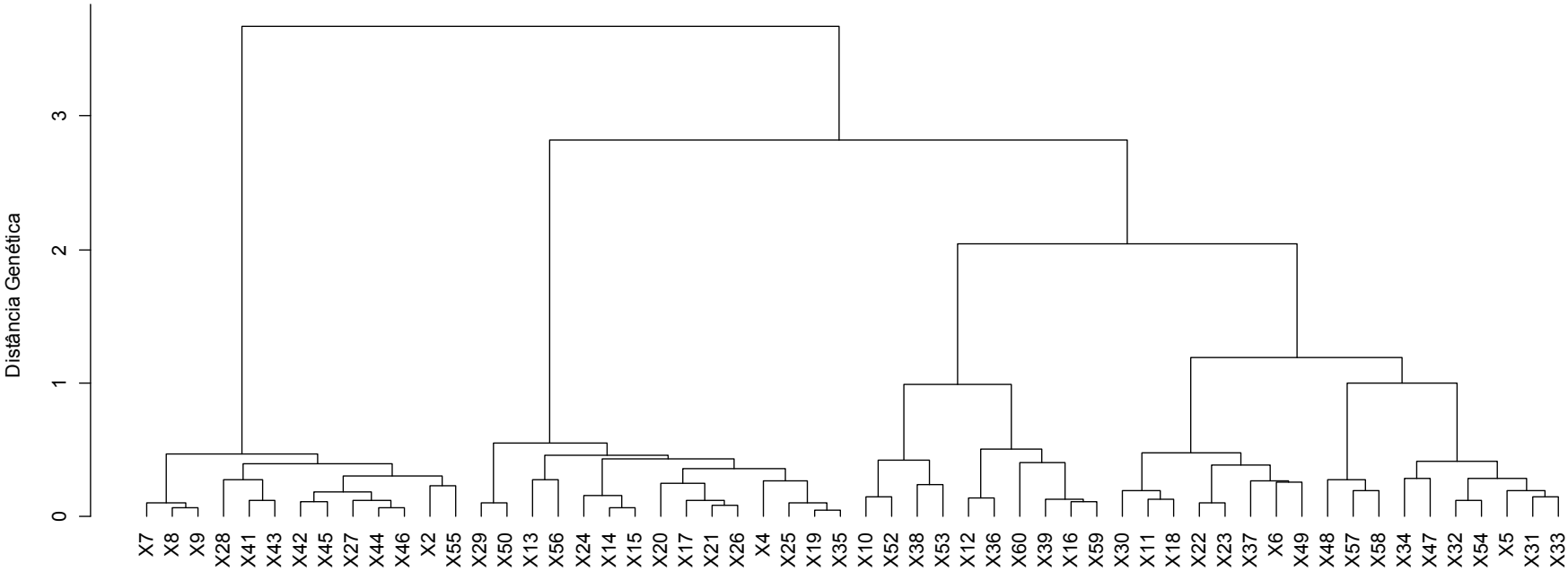
APÊNDICE B

Figura 2 – Agrupamento hierárquico Ward utilizando a distância combinada (Pearson e coincidência simples) para 56 acessos de *Capsicum* spp



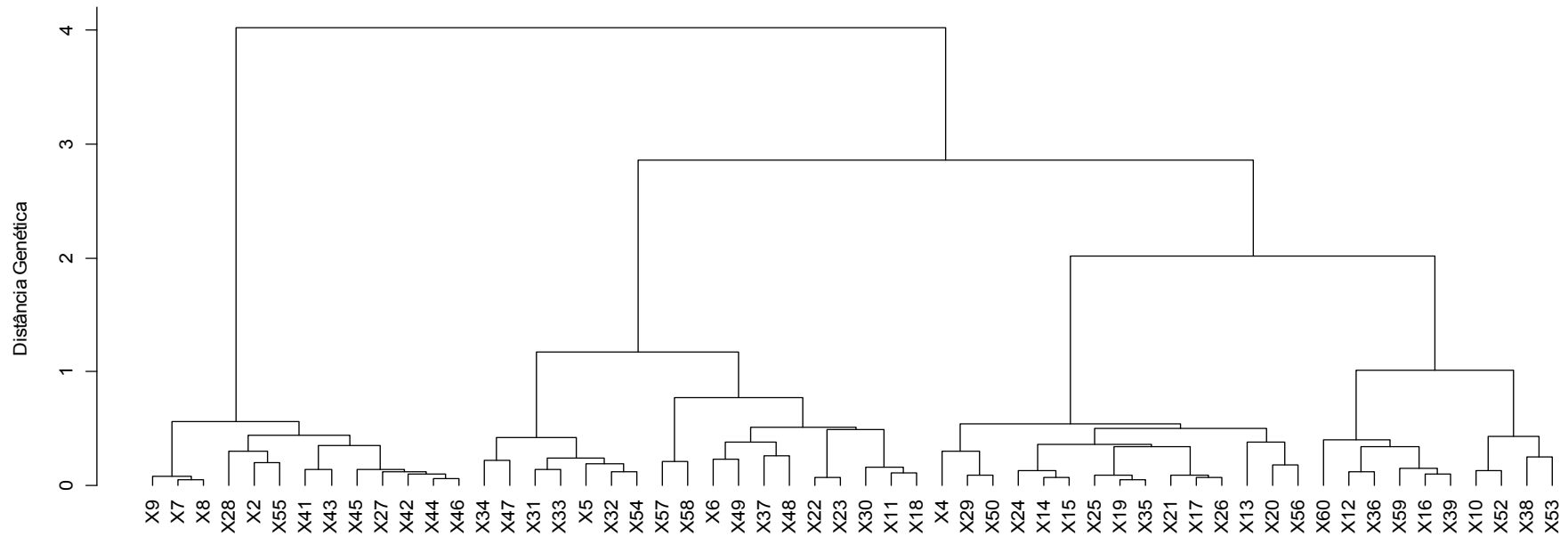
APÊNDICE C

Figura 3 – Agrupamento hierárquico Ward utilizando a distância combinada (Kulczynski e coincidência simples) para 56 acessos de *Capsicum* spp



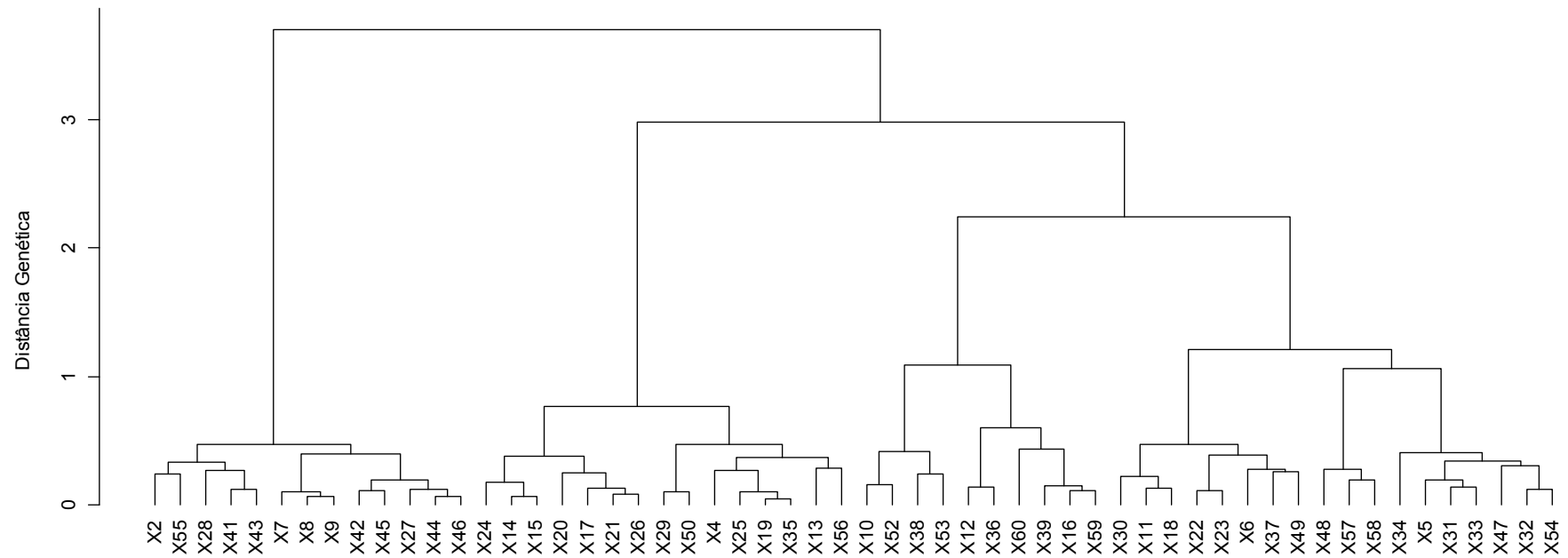
APÊNDICE D

Figura 4 – Agrupamento hierárquico Ward utilizando a distância combinada (Canberra e coincidência simples) para 56 acessos de *Capsicum* spp



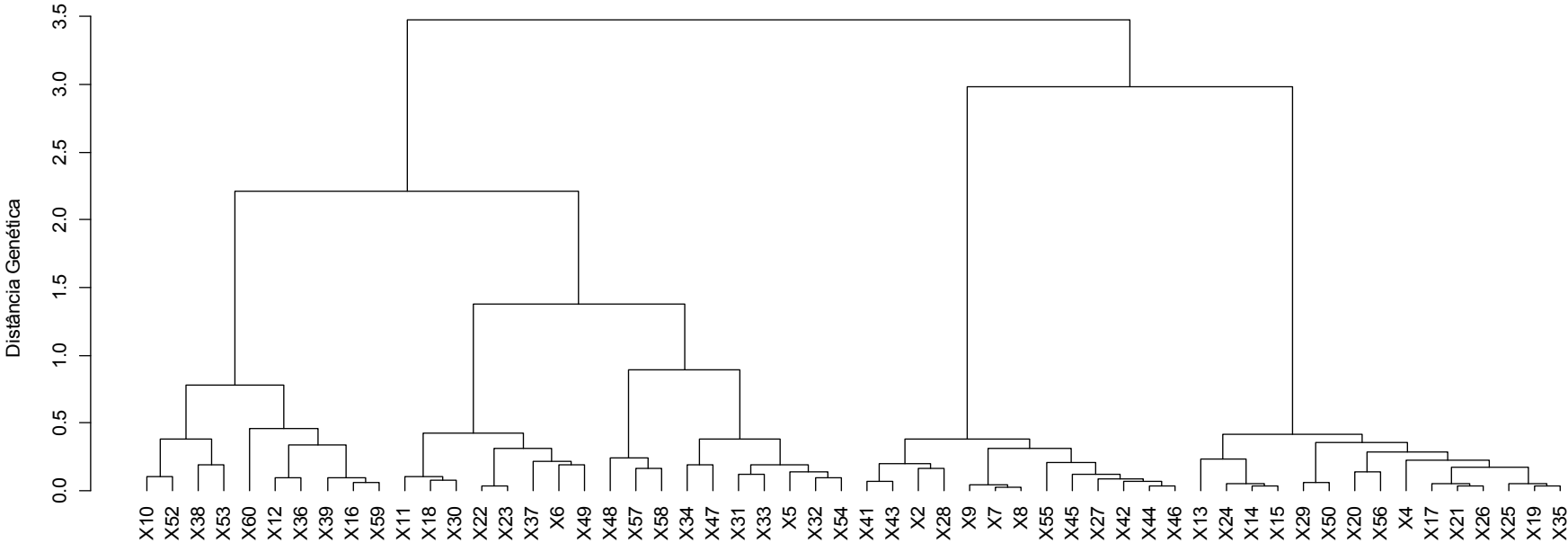
APÊNDICE E

Figura 5 – Agrupamento hierárquico Ward utilizando a distância combinada (Bray-Curtis e coincidência simples) para 56 acessos de *Capsicum* spp

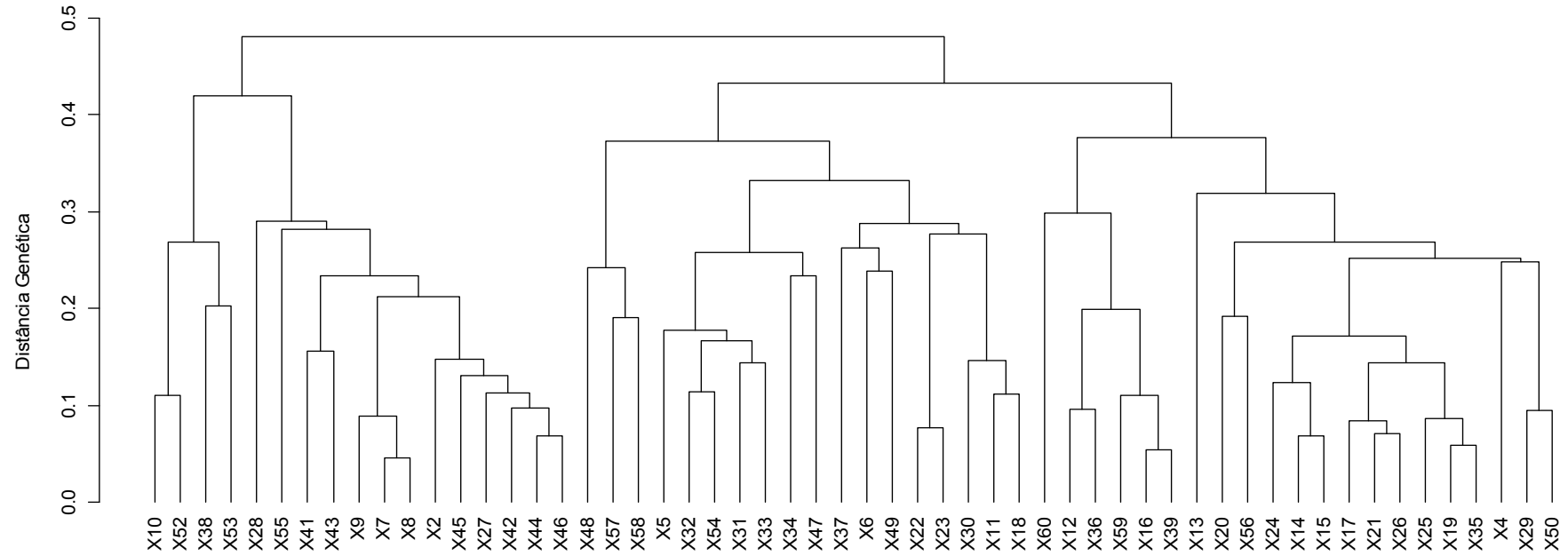


APÊNDICE F

Figura 6 – Agrupamento hierárquico Ward utilizando a distância combinada (Morisita e coincidência simples) para 56 acessos de *Capsicum* spp

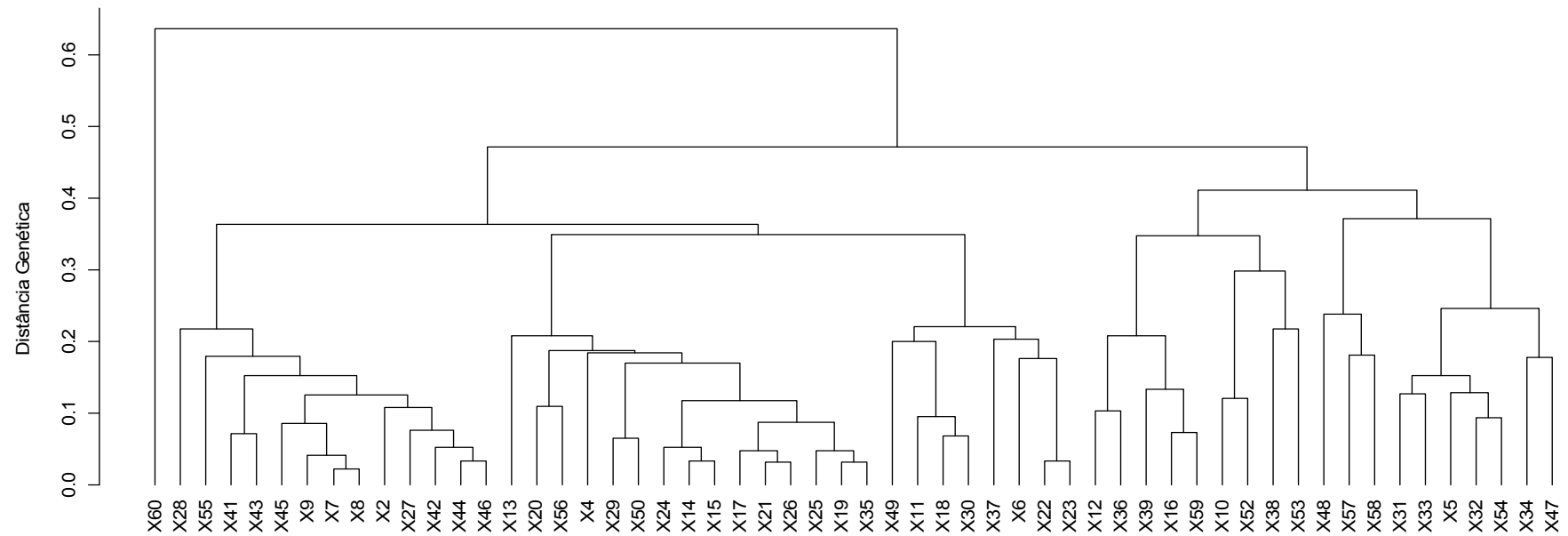


APÊNDICE G

Figura 7 – Agrupamento hierárquico UPGMA utilizando a distância combinada de Gower para 56 acessos de *Capsicum* spp

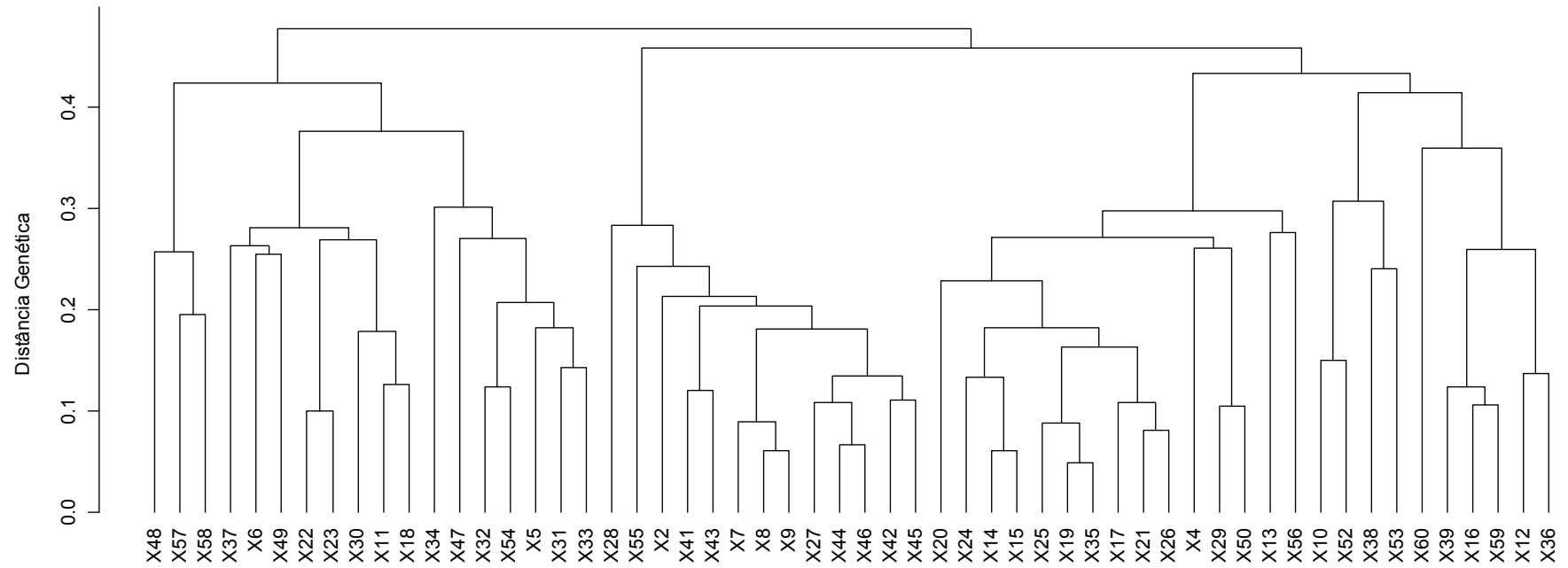
APÊNDICE H

Figura 8 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Pearson e coincidência simples) para 56 acessos de *Capsicum* spp



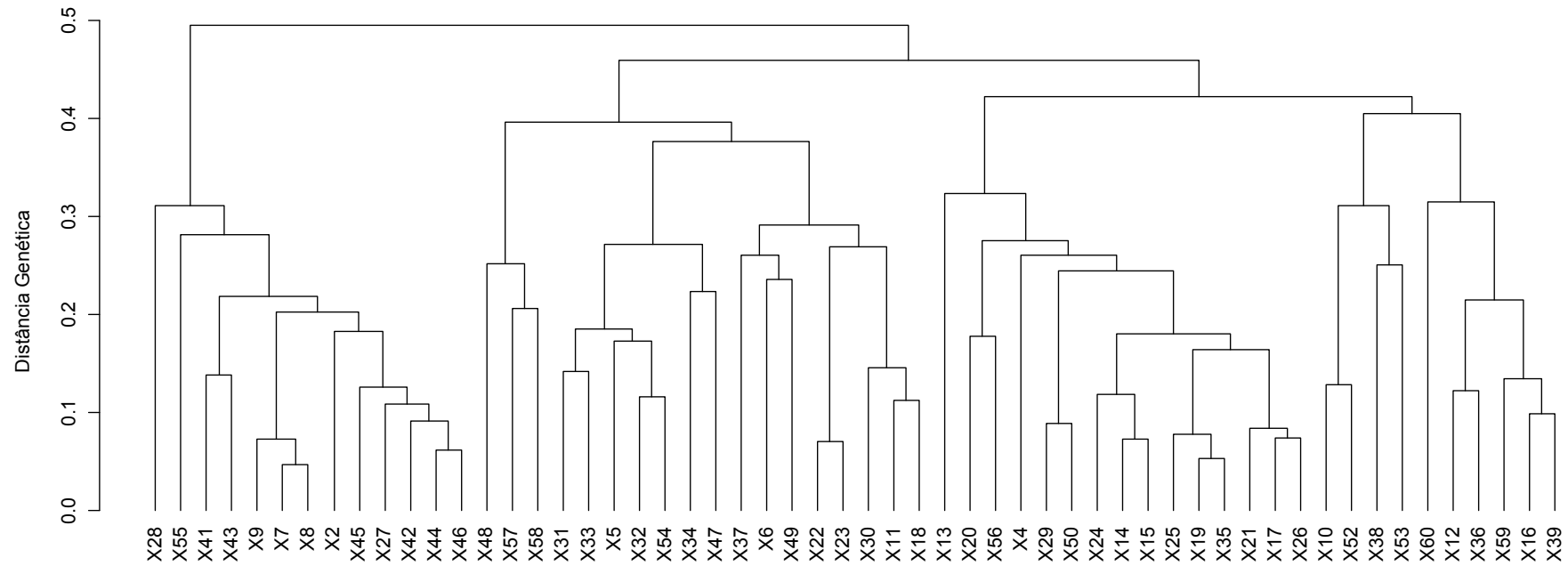
APÊNDICE I

Figura 9 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Kulczynski e coincidência simples) para 56 acessos de *Capsicum* spp



APÊNDICE J

Figura 10 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Canberra e coincidência simples) para 56 acessos de *Capsicum* spp



APÊNDICE K

Figura 11 – Agrupamento hierárquico UPGMA utilizando a distância combinada (Bray-Curtis e coincidência simples) para 56 acessos de *Capsicum* spp

