JULIANA SARAGIOTTO SILVA

**Métricas de análise de redes sociais e sua aplicação em redes de interação
biológicas: metodologia de aplicação e estudos de caso**

São Paulo
2014

JULIANA SARAGIOTTO SILVA

**Métricas de análise de redes sociais e sua aplicação em redes de interação
biológicas: metodologia de aplicação e estudos de caso**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do título de Doutor em Ciências

Orientador: Prof. Dr. Antonio Mauro
Saraiva

São Paulo

2014

JULIANA SARAGIOTTO SILVA

**Métricas de análise de redes sociais e sua aplicação em redes de interação
biológicas: metodologia de aplicação e estudos de caso**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção
do título de Doutor em Ciências

Área de Concentração: Engenharia de
Computação

Orientador: Prof. Dr. Antonio Mauro
Saraiva

São Paulo

2014

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 05 de novembro de 2014.

Assinatura do autor _____

Assinatura do orientador _____

Catologação-na-publicação

Silva, Juliana Saragiotto

**Métricas de análise de redes sociais e sua aplicação em redes de interação biológicas: metodologia de aplicação e estudos de caso / J.S. Silva. – versão corrigida -- São Paulo, 2014.
171 p.**

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Redes sociais (Análise) 2.Redes de interação 3.Redes complexas 4.Métricas 5.BPMN 6.Informática para a biodiversidade I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico este trabalho aos meus pais e a um anjo amigo que me ensinaram que a
essência da vida está nas coisas simples da vida:
um abraço que acolhe, um sorriso que incentiva e em mãos que sustentam!

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus pelo dom da vida e por estar sempre presente em minha vida, a cada amanhecer, dando-me força e iluminando meus passos.

Ao Prof. Dr. Antonio Mauro Saraiva, meu orientador de doutorado (e um bom amigo) que, com sabedoria e paciência, compartilhou seus conhecimentos e se tornou um facilitador neste processo de pesquisa. Obrigada por me ajudar a descobrir o melhor caminho a trilhar e, também, por me fazer acreditar, a cada reunião de trabalho, que no final tudo daria certo!

À Profa. Dra. Tatiana Teixeira Torres, do Departamento de Genética do Instituto de Biociências da Universidade de São Paulo (IB/USP), pela acolhida em seu laboratório, troca de experiências, apresentação das particularidades das Redes de Interação de micro-organismos e disponibilização de um ambiente agradável de pesquisa, em conjunto com seus orientados de graduação (Marina, Fabrício e Raquel), mestrado (Sophie) e doutorado (Nancy – co-orientação). Dentro desta equipe, um agradecimento especial à minha parceira e amiga de pesquisa, Nancy de Castro Stoppe, que muito me ensinou sobre Análises Microbiológicas e que me possibilitou abarcar neste universo de pesquisa por meio da aplicação dos recursos de Análise de Redes Sociais em Redes de Interação de micro-organismos. Gostaria de registrar, também, um agradecimento à sua orientadora Profa. Dra. Laura Maria Mariscal Ottoboni, do Centro de Biologia Molecular e Engenharia Genética da Universidade Estadual de Campinas (CBMEG/UNICAMP), por participar deste processo de colaboração.

Às Profas. Dras. Astrid de Matos Peixoto Kleinert e Tereza Cristina Giannini, do Laboratório de Abelhas do Instituto de Biociências da Universidade de São Paulo (BeeLab/IB/USP), pelo aprendizado, troca de experiências e a oportunidade de realização de um trabalho colaborativo sobre Redes de Interação abelha-planta no Brasil.

Ao Núcleo de Pesquisa em Biodiversidade e Computação da Universidade de São Paulo (BioComp-USP), por proporcionar um ambiente favorável para o desenvolvimento desta pesquisa, a partir de um processo colaborativo transdisciplinar (Biologia, Computação, Engenharia e Estatística) que foi a base de sustentação para o desenvolvimento da metodologia aqui proposta.

Ao Prof. Dr. Marco A. R. Mello, do Laboratório de Ecologia do Mutualismo do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais (LEM/ICB/UFMG), pelas valiosas contribuições no entendimento das métricas de Análise de Redes e de suas possibilidades de aplicação no domínio biológico.

Ao Dr. Raúl Jiménez Rosenberg, da *Comisión Nacional para el Conocimiento y Uso de la Biodiversidad* (CONABIO, México-DF), que trouxe valiosas contribuições da área de Análise Estatística para apoiar as atividades de Análise dos Dados desta pesquisa científica.

À Profa. Maria Cristina V. Borba, responsável pelos cursos de Redação Acadêmica e Técnicas de Apresentação na Escola Politécnica da USP, que com maestria, atenção e carinho revisou todas as minhas produções científicas em inglês.

Aos membros da banca do exame de qualificação, Profas. Dras. Anna Helena Reali Costa e Tatiana Teixeira Torres, pelas valiosas contribuições que foram importantes para o delineamento final desta pesquisa de doutorado. Aos membros da banca examinadora da defesa, pela disponibilidade na leitura desta tese e por toda a contribuição que certamente deixarão para a continuidade da pesquisa.

Aos coordenadores do Programa de Doutorado Interinstitucional da Escola Politécnica da USP com a Universidade Federal de Mato Grosso (Dinter EPUSP-UFMT), Prof. Dr. André Riyuiti Hirakawa (EPUSP) e Profa. Dra. Patricia Cristiane de Souza (UFMT), pelo empenho e dedicação na condução das atividades deste programa. Aos professores da EPUSP que se deslocaram até Cuiabá para ministrar as disciplinas do Programa, proporcionando um ambiente rico de troca de experiências e aprendizado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT) pelo aporte financeiro a este Programa Dinter EPUSP-UFMT.

Aos amigos do Dinter EPUSP-UFMT, em especial aos que fizeram esta caminhada ser mais suave, por meio de conversas, incentivos, companheirismo, enfim, acreditando que chegaríamos “vivos” até o final: Clovis Santos, Eunice Nunes, Evandro Freiburger, Ivairton Santos, João Paulo Pretti, Luciana Borges, Mara Dota, Raul Santos, Soraia Prietch e Wannessa Fonseca. Não poderia deixar

de mencionar os amigos do Interlab/USP, que sempre se fizeram presentes e estiveram na torcida: Ana Cláudia, Cleber, Makoto e Mariza.

Aos professores do Laboratório de Automação Agrícola da Escola Politécnica da USP (LAA/EPUSP) – Profs. Drs. André Riyuiti Hirakawa, Carlos Eduardo Cugnasca e Pedro Luiz Pizzigatti Corrêa – pela acolhida, convivência e troca de experiências. Também gostaria de agradecer a todos os colegas do LAA/EPUSP, alunos de mestrado e/ou doutorado, que durante minha estada na USP, compartilharam comigo seu dia-a-dia, desafios e conquistas, em particular: Allan Koch, Daniel Lins, Leonardo Campos, Luiz Lamardo, Marcelo Barros, Michel Bieleveld e Wilian Costa.

À equipe técnica e administrativa do LAA/EPUSP e Biocomp-USP, Edson de Souza, Lourdes Takigami e Suzano Bitencourt que, com dedicação, sempre me ajudaram a solucionar diversas situações, ao longo desta caminhada do doutorado. Um agradecimento especial à Lourdes, pelas partilhas, companheirismo e cuidados para comigo – uma grande amiga. E, também, ao Bruno Nunes Silva, pelo auxílio nas atividades desenvolvidas no BioComp-USP.

À equipe da secretaria do Departamento de Engenharia da Computação e Sistemas Digitais da Escola Politécnica da USP (PCS/EPUSP), que estiveram sempre disponíveis a ajudar e orientar acerca das questões do dia-a-dia do curso: Aleksandro Silva, Leia Sicília, Mayra Onishi, Rubens Toledo e Tania Silva. Um agradecimento especial à Leia, pela alegria com que sempre me acolheu, pela solicitude e amizade.

Aos colegas de trabalho da Coordenadoria de Tecnologia da Informação da Secretaria de Estado de Saúde de Mato Grosso (SES/MT), pela torcida, incentivo e compreensão, especialmente Alane, Dejair, Fatinha, Kelly, Luiz Viana, Manoel Neto, Ricardo e Sizi.

Aos colegas de trabalho e aos alunos do Departamento de Área de Informática do Instituto Federal de Mato Grosso (DAI/IFMT), que sempre estiveram na torcida e souberam entender minhas ausências, durante o meu processo de afastamento para as atividades de pesquisa.

À família Segue-me, pelo apoio, partilhas e orações; um agradecimento especial aos tios Beto e Dora, Luiz e Lucilene, e aos primos Becc, Dani Maia, Fabíola e Newton.

Aos familiares, que ao longo desses cinco anos de caminhada, estiveram sempre me apoiando e compreenderam minhas ausências, em função do doutorado. Um agradecimento especial aos meus irmãos Maria Cristina, Luiz Eduardo e João Henrique, bem como aos meus cunhados Plínio e Stefânia (futura mamãe do Bento), por entenderem este processo que estava vivenciando e pela ajuda nos momentos que precisei. Aos meus pais, Domicio e Inês, pela paciência, dedicação e amor demonstrados a cada gesto e cuidado para comigo e, também, para que o ambiente aqui em casa fosse o mais propício a este processo de pesquisa. Aos meus afilhados Gabriel (7 anos) e Luiza (5 anos) que, mesmo sem entender porque a madrinha passava horas na frente do computador, traziam alegria ao meu dia-a-dia e a certeza de que tudo daria certo. Deixo aqui uma frase da Luiza, quando eu lhe contei que o meu curso começou quando ela nasceu: “Madrinha, o seu curso é do meu tamanho...muito grande...vai demorar para terminar?”

Por fim, a todos os professores que, direta ou indiretamente, contribuíram para que a realização deste sonho se tornasse uma realidade: desde a minha formação inicial (Educação Infantil, Ensino Fundamental e Médio – escolas estaduais em Icaraíma/PR), passando pela Graduação (Ciência da Computação – UFMT), Especialização (Gestão em Projetos Inovadores – UFMT) e Mestrado (Política Científica e Tecnológica – UNICAMP).

A todos vocês, meus sinceros agradecimentos!

RESUMO

Diversos pesquisadores têm se utilizado do recurso de Redes de Interação na área de Biodiversidade para analisar o papel das espécies na estrutura da uma rede – cujos fundamentos conceituais são os mesmos das Redes Sociais (como *Facebook*, *LinkedIn*, entre outras). Nesse sentido, algoritmos, métricas e recursos computacionais e estatísticos provenientes da área de Análise de Redes Sociais (*Social Network Analysis* – SNA) são ferramentas importantes para endereçar/apoiar estudos com interações. Assim sendo, o objetivo desta tese é propor uma metodologia para aplicação das métricas de SNA em estudos com Redes de Interação biológicas no domínio da Informática para a Biodiversidade. A metodologia está formalizada por meio da Notação para Modelagem de Processos de Negócio (BPMN - *Business Process Model and Notation*) e estruturada em quatro etapas: (i) mapeamento dos tipos de dados e de interação disponíveis; (ii) definição das perguntas-chave a serem respondidas e das variáveis de análise; (iii) escolha das métricas de SNA adequadas ao contexto da pesquisa; e (iv) realização de análises biológicas com o apoio de SNA. Como recursos materiais foram utilizadas as métricas de SNA, bem como um conjunto de ferramentas computacionais (como os pacotes do *R* e os programas *Dieta*, *Pajek* e *Ucinet*) e de Análise Estatística (como a Análise Exploratória de Dados e a Análise Multivariada de Dados). Esta proposta nasceu de um processo de colaboração com pesquisadores de diversas áreas do conhecimento, a partir de projetos desenvolvidos no Núcleo de Pesquisa em Biodiversidade e Computação da USP (BioComp-USP), o que trouxe uma base de sustentação a esta metodologia. Para avaliar a adequação desta proposta a necessidades reais de pesquisa a metodologia foi aplicada a três estudos de caso com Redes de Interação microbiológicas. Os resultados mostram os benefícios que a disponibilização de um método sistematizado para guiar os passos de uma pesquisa pode trazer a um pesquisador – seja em função do aporte de recursos recomendados, seja pelo processo de organização das atividades de pesquisa. Além disso, verifica-se a possibilidade de transposição desta proposta a outros domínios do conhecimento ainda não explorados, como em Agrobiodiversidade.

Palavras-Chave: Análise de Redes Sociais. Redes de Interação. Redes Complexas. Métricas. BPMN. Informática para a Biodiversidade.

ABSTRACT

Several researchers have used Interaction Networks resources in the Biodiversity area for analyzing the role of species in network structure – their conceptual foundations are the same as those in Social Networks (such as Facebook, LinkedIn, among others). Thus, algorithms, metrics, and statistical and computational resources from the Social Network Analysis (SNA) area are important tools for addressing this issue. Therefore, the aim of this thesis is to propose a methodology for applying SNA metrics to biological interaction network studies in the Biodiversity Informatics domain. The methodology is formalized by means of Business Process Model and Notation (BPMN) and structured in four steps: (i) mapping the data types and the interaction available; (ii) defining the key-questions to be answered and the analysis variable; (iii) choosing the SNA metrics appropriate to the context of the research; and (iv) performing the biological analysis with the support of SNA. As material resources, the SNA metrics were used, as well as a set of computational (such as *R* packages, *Dieta*, *Pajek* and *Ucinet* software) and Statistical Analysis (Exploratory Data Analysis and Multivariate Data Analysis) tools. This proposal generated a process collaboration with researchers from different knowledge areas, by means of projects developed at the Research Center on Biodiversity and Computing at USP (BioComp-USP), which provided a support base to this methodology. To assess the suitability of this proposal to the real research needs, it was applied to three case studies with microbiological Interaction Networks. The results show the benefits that providing a systematic method to guide the steps of one research can bring to a researcher – be it due to the support of the resources recommended, be it by the organization of the research activities. In addition, there is the possibility of applying this methodology to unexplored knowledge fields, such as Agrobiodiversity.

Keywords: Social Network Analysis. Interaction Networks. Complex Networks. Metrics. BPMN. Biodiversity Informatics.

LISTA DE ILUSTRAÇÕES

Figura 2.1	– Exemplo de uma interação polinizador-planta.	26
Figura 2.2	– Exemplo da presença de <i>E.coli</i> em uma amostra.	27
Figura 2.3	– Exemplo de um grafo para uma Rede de Interação polinizador-planta.	33
Figura 2.4	– Exemplo da representação de um gráfico <i>boxplot</i>	43
Figura 2.5	– Exemplo do uso do gráfico <i>boxplot</i> para representação da frequência de distribuição de grupos filogenéticos (A, B1, B2 e D) de <i>E.coli</i>	44
Figura 2.6	– Exemplo do uso do gráfico <i>scatterplot</i> para representação da correlação entre grupos filogenéticos (A, B1, B2 e D) de <i>E.coli</i>	45
Figura 2.7	– Distribuição de probabilidade <i>Gamma</i> para o grupo filogenético A de <i>E.coli</i>	47
Figura 2.8	– Exemplo da aplicação do método <i>Clustering Visualizations of Multidimensional Data</i> (reordenando as variáveis pelo índice <i>kmeans</i>).	50
Figura 3.1	– <i>Pajek</i> : interface do programa e representação do grafo de uma Rede de Interação.	58
Figura 4.1	– Metodologia para aplicação de métricas de SNA em biodiversidade (modelada em BPMN).	65
Figura 4.2	– Etapa1 da metodologia: Mapeamento dos tipos de dados e de interação (modelada em BPMN).	65
Figura 4.3	– Exemplo parcial de uma matriz de interação ponderada de visitantes-florais x plantas.	66
Figura 4.4	– Etapa2 da metodologia: Definição das perguntas-chave e variáveis (modelada em BPMN).	68
Figura 4.5	– Etapa2-subprocesso: Escolher variáveis de análise (modelada em BPMN).	68
Figura 4.6	– Etapa3 da metodologia: Escolha das métricas de SNA adequadas (modelada em BPMN).	80

LISTA DE ILUSTRAÇÕES *(continuação)*

Figura 4.7	– Etapa4 da metodologia: Realização de análises biológicas com SNA (modelada em BPMN).	80
Figura 5.1	– Etapa1 da metodologia: trilha do Estudo de Caso 1.	83
Figura 5.2	– Etapa2 da metodologia: trilha do Estudo de Caso 1.	83
Figura 5.3	– Etapa2 da metodologia (subprocesso – escolher variáveis de análise): trilha do Estudo de Caso 1.	84
Figura 5.4	– Etapa3 da metodologia: trilha do Estudo de Caso 1.	85
Figura 5.5	– Etapa4 da metodologia: trilha do Estudo de Caso 1.	85
Figura 5.6	– Metodologia para aplicação de métricas de SNA em biodiversidade: trilha do Estudo de Caso 1.	86
Figura 5.7	– Etapa1 da metodologia: trilha do Estudo de Caso 2.	87
Figura 5.8	– Etapa2 da metodologia (subprocesso – escolher variáveis de análise): trilha do Estudo de Caso 2.	87
Figura 5.9	– Etapa3 da metodologia: trilha do Estudo de Caso 2.	88
Figura 5.10	– Etapa1 da metodologia: trilha do Estudo de Caso 3.	90
Figura 5.11	– Etapa2 da metodologia (subprocesso – escolher variáveis de análise): trilha do Estudo de Caso 3.	91
Figura 5.12	– Etapa3 da metodologia: trilha do Estudo de Caso 3.	93

LISTA DE TABELAS

Tabela 2.1	– Tipos de dados presentes nos estudos de Redes de Interação mapeados.	31
Tabela 2.2	– Mapeamento dos programas, técnicas e algoritmos utilizados nos estudos.	32
Tabela 2.3	– Formas de representação de uma Rede de Interação.	35
Tabela 2.4	– Alguns elementos básicos da BPMN.	51
Tabela 4.1	– Resumo de recursos utilizados para aplicar EDA.	70
Tabela 4.2	– Resumo de recursos utilizados para aplicar Análise Multivariada de Dados.	73
Tabela 4.3	– Resumo de métricas para analisar as interações no nível da rede.	75
Tabela 4.4	– Resumo de métricas para analisar as interações no nível da espécie.	78
Tabela 5.1	– Lista dos Estudos de Caso desenvolvidos.	81

LISTA DE ABREVIATURAS E SIGLAS

BeeLab/IB/USP	Laboratório de Abelhas do IB/USP
BioComp-USP	Núcleo de Pesquisa em Biodiversidade e Computação da Universidade de São Paulo
BPMN	<i>Business Process Model and Notation</i> (Notação para Modelagem de Processos de Negócio, pois inicialmente era descrita como <i>Business Process Modeling Notation</i> – versão 1.0)
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CBMEG/ UNICAMP	Centro de Biologia Molecular e Engenharia Genética da UNICAMP
CDF	<i>Cumulative Distribution Function</i> (Função de Distribuição Cumulativa)
CETESB	Companhia Ambiental do Estado de São Paulo
CONABIO	<i>Comisión Nacional para el Conocimiento y Uso de la Biodiversidad</i> (México, Distrito Federal)
DAI/IFMT	Departamento de Área de Informática do IFMT
Dinter	Doutorado Interinstitucional
ECDF	<i>Empirical CDF</i>
EDA	<i>Exploratory Data Analysis</i> (Análise Exploratória de Dados)
EPUSP	Escola Politécnica da USP
FAO	<i>Food and Agriculture Organization of the United Nations</i> (Organização das Nações Unidas para Alimentação e Agricultura)
FAPEMAT	Fundação de Amparo à Pesquisa do Estado de Mato Grosso
IB/USP	Instituto de Biociências da USP
IFMT	Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso
IQR	<i>Interquartile range</i> (amplitude inter-quartis)
K-S test	Teste estatístico <i>Kolmogorov-Smirnov</i>
LAA/EPUSP	Laboratório de Automação Agrícola da Escola Politécnica da USP
LEM/ICB/UFMG	Laboratório de Ecologia do Mutualismo do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais

LISTA DE ABREVIATURAS E SIGLAS (*continuação*)

MLE	Maximum-Likelihood Estimation (Estimativa por Máxima Verossimilhança – método estatístico)
MMA	Ministério do Meio Ambiente
NODF	<i>Nestedness metric based on Overlap and Decreasing Fill</i> (uma métrica para cálculo do aninhamento)
NTC	<i>Nestedness Temperature Calculator</i> (um programa de Análise de Redes)
OMG	<i>Object Management Group</i>
PCS/EPUSP	Departamento de Engenharia da Computação e Sistemas Digitais da Escola Politécnica da USP
SBIAgro	Congresso Brasileiro de Agroinformática
SES/MT	Secretaria de Estado de Saúde de Mato Grosso
SNA	<i>Social Network Analysis</i> (Análise de Redes Sociais)
TSI	<i>Trophic State Index</i> (Índice de Estado Trófico – permite avaliar o grau de eutrofização de corpos d'água)
UFMT	Universidade Federal de Mato Grosso
UNICAMP	Universidade Estadual de Campinas
UPGMA	<i>Unweight Pair Group Method with Arithmetic mean</i> (um método de clusterização hierárquica)
USP	Universidade de São Paulo
WebBee	Uma rede de informações sobre abelhas, baseada na <i>web</i>
WQI	<i>Water Quality Index</i> (Índice de Qualidade de Água)

SUMÁRIO

1 INTRODUÇÃO	19
1.1 Motivação da Pesquisa.....	19
1.2 Objetivo.....	20
1.3 Justificativa	20
1.4 Estrutura da tese.....	21
2 REVISÃO DA LITERATURA.....	23
2.1 Iniciativas para estudos e a conservação da biodiversidade	23
2.1.1 A importância dos estudos sobre polinizadores.....	23
2.1.2 Redes de Interação entre polinizadores e plantas	25
2.1.3 Redes de Interação de microorganismos	26
2.2 Técnicas de Análise de Redes de Interação biológicas.....	28
2.3 Redes Sociais.....	34
2.3.1 Fundamentos e conceitos de Redes Sociais	34
2.3.2 Métricas de Análise de Redes Sociais.....	36
2.3.2.1 Métricas para análise no nível da rede (<i>network level</i>).....	36
2.3.2.2 Métricas para análise no nível da espécie ou do nó (<i>species/node level</i>).....	38
2.4 Análise Estatística.....	41
2.4.1 Análise Exploratória de Dados.....	41
2.4.1.1 Visualização da distribuição dos dados por meio de <i>bloxplot</i>	42
2.4.1.2 Análise de correlação entre variáveis por meio de <i>scatterplot</i> ...	44
2.4.1.3 Modelo de Distribuição de Probabilidade	45
2.4.2 Análise Multivariada de Dados.....	47
2.4.2.1 <i>Mantel Test</i>	48
2.4.2.2 <i>Clustering Visualizations of Multidimensional Data</i>	49
2.5 Notação para Modelagem de Processos de Negócio	50
3 MATERIAIS E MÉTODOS.....	53
3.1 Revisão da literatura	53
3.2 Experimentos realizados.....	54
3.2.1 Descrição dos dados	55
3.2.2 Ferramentas de apoio (recursos materiais)	55
3.2.2.1 Métricas de SNA.....	56
3.2.2.2 Ferramentas Computacionais	56
3.2.2.3 Ferramentas de Análise Estatística	58

3.2.3 Interação com especialistas em biodiversidade (recursos humanos)	59
3.2.4 Manuscritos desenvolvidos	60
3.3 Formulação da proposta de uma metodologia.....	61
3.4 Validação da metodologia proposta por meio de estudos de caso	62
4 UMA METODOLOGIA PARA APLICAÇÃO DE MÉTRICAS DE ANÁLISE DE REDES SOCIAIS EM BIODIVERSIDADE.....	63
4.1 Etapa 1: Mapeamento dos tipos de dados e de interação disponíveis	64
4.2 Etapa 2: Definição das perguntas-chave a serem respondidas e das variáveis de análise	66
4.3 Etapa 3: Escolha das métricas de SNA adequadas ao contexto da pesquisa .	74
4.4 Etapa 4: Realização de análises biológicas com o apoio de SNA	79
5 DISCUSSÃO DA METODOLOGIA PROPOSTA A PARTIR DE ESTUDOS DE CASO COM REDES DE INTERAÇÃO MICROBIOLÓGICAS	81
5.1 Estudo de Caso 1: Stoppe <i>et al.</i> (2014a)	82
5.2 Estudo de Caso 2: Silva <i>et al.</i> (2014).....	86
5.3 Estudo de Caso 3: Stoppe <i>et al.</i> (2014b)	89
6 CONSIDERAÇÕES FINAIS E CONCLUSÕES	95
6.1 Considerações Finais	95
6.2 Conclusões	96
REFERÊNCIAS	99
APÊNDICE A – <i>Paper Genetics and Molecular Biology</i> (co-autoria).....	108
APÊNDICE B – <i>Paper Complex Networks V</i> (CompleNet2014) – (1ª autora).126	
APÊNDICE C – <i>Paper SpringerPlus</i> (2ª autora – <i>equal contribution</i>).....	137

1 INTRODUÇÃO

Este capítulo tem por objetivo apresentar uma visão geral do trabalho, tendo como pano de fundo a motivação, o objetivo e a justificativa da pesquisa. Descreve, também, a estrutura da tese.

1.1 Motivação da Pesquisa

Diversos pesquisadores (BASCOMPTE; JORDANO, 2007; BASCOMPTE; JORDANO; OLESEN, 2006; KRUG; ALVES-DOS-SANTOS, 2010; MEMMOTT *et al.*, 2007; OLESEN *et al.*, 2007; VÁZQUEZ *et al.*, 2009), nesta última década, têm se utilizado do recurso de Redes de Interação para analisar os fatores que contribuem ou influenciam na manutenção da biodiversidade – uma rede heterogênea composta por *links* densos e esparsos (OLESEN *et al.*, 2007).

Os conceitos, ferramentas e métricas utilizadas nos estudos de Redes de Interação têm sua origem no domínio das Ciências Sociais, particularmente na Análise de Redes Sociais (*Social Network Analysis* – SNA), que se utiliza dos conceitos da teoria de grafos e das técnicas e recursos computacionais, para a análise da estrutura das conexões de atores em redes (SCOTT, 2000). A Análise de Redes Sociais é considerada uma das aplicações de maior amplitude no domínio da ciência de redes para o estudo de conexões e relacionamentos de pessoas (HANSEN; SHNEIDERMAN; SMITH, 2011).

Os seguintes desafios são apresentados para esta área de SNA: (i) compreender como as Redes Sociais prosperam, enfraquecem ou se transformam; (ii) fornecer um conjunto de conceitos e métricas para um estudo sistematizado deste processo dinâmico de mudanças na estrutura das Redes Sociais; e (iii) trabalhar com Redes Sociais complexas e esparsas na descoberta de padrões, tendências, agrupamentos (*clusters*), lacunas e discrepâncias na estrutura da rede (HANSEN; SHNEIDERMAN; SMITH, 2011). Estes mesmos desafios emergem nos estudos sobre biodiversidade (como em MEMMOTT; WASER; PRICE, 2004; OLESEN; ESKILSEN; VENKATASAMY, 2002; OLESEN *et al.*, 2007).

Em face desta realidade, surge a oportunidade de se apropriar dos conhecimentos, métricas e recursos computacionais (disponíveis na área de SNA) para aplicação em outros domínios, como nas Redes de Interação de biodiversidade, tornando-se um campo factível para a experimentação científica.

Outra questão que emerge neste contexto da Informática para a Biodiversidade (*Biodiversity Informatics*) é uma tendência à disponibilização de bancos de dados com informações mais detalhadas acerca do processo de interação entre os organismos (tais como a frequência/abundância da interação) – o que potencializa a realização de estudos mais detalhados acerca da estrutura da rede e, também, a busca por conhecimentos não explorados.

1.2 Objetivo

Propor uma metodologia para apoiar os pesquisadores em biodiversidade na aplicação das métricas de Análise de Redes Sociais (*Social Network Analysis* – SNA) em Redes de Interação biológicas, servindo de guia para pesquisas dessa natureza. Como objetivo específico tem-se a aplicação da metodologia em três estudos de caso de Redes de Interação microbiológicas para avaliar a adequação da proposta.

1.3 Justificativa

Embora diversos pesquisadores em Análise de Redes Ecológicas já estejam usando os recursos de SNA para apoiar o desenvolvimento de seus trabalhos (ARAÚJO *et al.*, 2008; BASCOMPTE; JORDANO, 2007; BEZERRA; MACHADO; MELLO, 2009; OLESEN *et al.*, 2007), ainda no domínio da Biologia, observa-se que em diversas áreas isto ainda não é uma realidade, como em análises microbiológicas sobre qualidade da água. Outra questão que emerge neste domínio do conhecimento é a ausência de um instrumental metodológico, que subsidie o especialista em biodiversidade no processo de aplicação das métricas de SNA, justificando, assim o desenvolvimento desta pesquisa.

Além disso, é importante registrar que este projeto de pesquisa está sendo desenvolvido no contexto do Núcleo de Pesquisa em Biodiversidade e Computação da Universidade de São Paulo (BioComp-USP), o que lhe propicia um ambiente favorável à interação com pesquisadores de diversas áreas do conhecimento (na própria universidade e no exterior) e, assim, traz novos olhares e demandas que delineiam e sustentam o processo de construção desta proposta.

Por fim, pesquisas dessa natureza vêm ao encontro das necessidades locais e regionais do estado de origem da pesquisadora (Mato Grosso), que necessita qualificar capital humano com competências em Informática para a

Biodiversidade, com vistas à criação de grupos de pesquisa que desenvolvam trabalhos nesta área.

1.4 Estrutura da tese

Esta tese encontra-se estruturada em seis Capítulos, incluindo os Capítulos de Introdução (1) e de Considerações Finais e Conclusões (6).

O Capítulo 2 apresenta uma revisão da literatura nas áreas de conhecimento que fundamentam esta pesquisa. Desta forma, num primeiro momento, descreve algumas iniciativas para estudos e conservação da biodiversidade (Informática para a Biodiversidade); na sequência, identifica os tipos de dados disponíveis nos estudos de Redes de Interação, bem como as formas de apresentação, técnicas, algoritmos e programas aplicados nesses estudos; depois, apresenta os elementos conceituais das Redes Sociais, com um olhar particular para os recursos disponíveis no domínio de SNA (conceitos e métricas); por fim, propõe duas abordagens de apoio ao processo de construção desta pesquisa (a Análise Estatística e a Notação para Modelagem de Processos de Negócio).

O Capítulo 3 tem por objetivo descrever os materiais e métodos utilizados em todas as fases de desenvolvimento desta pesquisa: (i) revisão da literatura; (ii) experimentação; (iii) formulação da proposta da metodologia; e (iv) validação da metodologia proposta.

Na sequência, o Capítulo 4 tem como finalidade apresentar a proposta de uma metodologia para aplicação de métricas de SNA, em estudos de Redes de Interação no domínio da biodiversidade; esta metodologia está estruturada em 04 etapas: (i) mapeamento dos tipos de dados e de interação disponíveis; (ii) definição das perguntas-chave a serem respondidas e das variáveis de análise; (iii) escolha das métricas de SNA adequadas ao contexto da pesquisa; e (iv) realização de análises biológicas com o apoio de SNA. É importante registrar que esta metodologia consiste na contribuição original desta tese, para a área de Informática para a Biodiversidade, e foi desenvolvida a partir de iniciativas realizadas no âmbito do BioComp-USP.

O Capítulo 5 se dedica a discutir a metodologia proposta, a partir de estudos de caso com Redes de Interação microbiológicas, analisando os passos sugeridos em cada etapa da metodologia, durante o processo de aplicação das métricas de SNA.

Por fim, têm-se as Referências que subsidiaram o embasamento teórico desta pesquisa, bem como os Apêndices que fazem parte do trabalho.

2 REVISÃO DA LITERATURA

O objetivo deste capítulo é descrever os marcos teórico-conceituais que fundamentam esta pesquisa.

Desta forma, a primeira seção (2.1) se dedica a apresentar as iniciativas que estão sendo realizadas com o apoio das Tecnologias da Informação, com vistas à conservação e manutenção da biodiversidade (Informática para a Biodiversidade). A segunda seção (2.2) tem por finalidade identificar os estudos que trabalham com Redes de Interação biológicas, analisando sistematicamente os tipos de dados e de interações disponíveis, bem como as formas de apresentação, técnicas, algoritmos e programas aplicados nos estudos. A seção seguinte (2.3) apresenta conceitos e métricas de Redes Sociais, com potencial para aplicação no contexto de Redes de Interação biológicas. Já a quarta seção (2.4), traz os recursos utilizados no domínio da Estatística para apoiar as atividades de Análise Exploratória de Dados e de Análise Multivariada de Dados. Por fim, a quinta seção (2.5) apresenta uma notação utilizada na área da Computação para Modelagem de Processos de Negócio (*Business Process Model and Notation* – BPMN).

2.1 Iniciativas para estudos e a conservação da biodiversidade

Esta seção apresenta uma visão geral sobre a importância dos estudos com polinizadores para o contexto da conservação biológica e, na sequência, traz o conceito de Redes de Interações. É importante registrar que este processo de revisão da literatura teve início a partir desta temática (em especial pelas Redes de Interação polinizador-planta) em função de ser uma área, no domínio da Biologia, que dispõe de diversos estudos que se utilizam do recurso de Redes de Interação para apoiar as atividades de Análise de Dados.

2.1.1 A importância dos estudos sobre polinizadores

Os serviços de polinização, por insetos ou outros animais, são responsáveis por 75% da manutenção das plantas agrícolas necessárias para a alimentação da população mundial, bem como de muitas plantas utilizadas na indústria farmacêutica (MMA, 2010). Serviços, estes, que ao longo dos anos têm declinado em decorrência de fatores como mudanças no uso da terra, aumento de concentração de dióxido de

carbono na atmosfera e introdução de plantas e polinizadores exóticos dentro dos ecossistemas naturais (MEMMOTT; WASER, 2002).

De acordo com Imperatriz-Fonseca, Saraiva e Gonçalves (2007) existem inúmeras ações em torno do tema polinizadores, envolvendo a comunidade científica e os tomadores de decisão de diversas nações, como a Iniciativa Internacional dos Polinizadores, a qual tem promovido trabalhos visando a conservação e o uso sustentado da biodiversidade. Um exemplo de uma iniciativa regional, decorrente dessas ações mundiais, é a Iniciativa Brasileira de Polinizadores, que tem organizado a comunidade científica em torno de um tema de grande relevância para a sobrevivência das espécies, do qual a Universidade de São Paulo tem participado desde a sua concepção.

Saraiva (2003) e Canhos *et al.* (2004) descrevem quais iniciativas locais, regionais e globais estão sendo desenvolvidas em prol da integração de dados de biodiversidade, bem como o fato de que um dos principais desafios reside na organização e estruturação de uma abordagem global, por meio de esforços colaborativos. De acordo com Saraiva *et al.* (2003) estes são alguns dos elementos que estão conduzindo à construção de um novo campo de pesquisa, denominado “Informática para a Biodiversidade” (*Biodiversity Informatics*), que busca a gestão e o uso eficiente de informações de biodiversidade a partir de um *background* computacional de sustentação – tais como protocolos e padrões para integração e troca de dados¹.

Neste contexto, as Tecnologias da Informação têm tido um papel importante, uma vez que têm possibilitado avanços relevantes no acesso aos dados dispersos em todo o planeta (IMPERATRIZ-FONSECA; SARAIVA; GONÇALVES, 2007). Como exemplo de uma contribuição do uso das Tecnologias da Informação como suporte aos estudos de polinização, no Brasil, pode-se mencionar a *WebBee* – uma rede de informações sobre abelhas, baseada na *web*, que foi desenvolvida com o objetivo de organizar conhecimentos e facilitar o compartilhamento de informações sobre esses polinizadores (SARAIVA *et al.*, 2003); este trabalho é fruto da cooperação dos pesquisadores do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica (Laboratório de Automação Agrícola) e do

¹ Exemplos de protocolos e padrões para integração e troca de dados podem ser consultados em Silva *et al.* (2010).

Departamento de Ecologia do Instituto de Biociências (Laboratório de Abelhas), ambos da Universidade de São Paulo.

É por meio de iniciativas como esta que, atualmente, a comunidade científica brasileira dispõe de bases de dados de ocorrências de certas espécies de polinizadores; além disso, esforços adicionais têm sido feitos para o registro de ocorrências de interação polinizador-planta – tanto no processo de polinização agrícola, quanto ambiental. No entanto, não se pode deixar de mencionar que, embora existam esforços colaborativos globais e regionais, as bases de conhecimentos existentes no mundo, frequentemente, estão dispersas e inacessíveis para as pessoas que necessitam de determinada informação para introduzir apropriadamente práticas amigáveis de polinização dentro da gestão agrícola (FAO, 2008) – e este é um dos desafios emergentes para este campo de pesquisa.

2.1.2 Redes de Interação entre polinizadores e plantas

De acordo com Giannini (2011), desde meados do século XIX, os estudos de Darwin sobre a origem das espécies por seleção natural sugerem que a diversidade decorre das interações entre organismos, por exemplo, em um processo de evolução mútua entre as espécies relacionadas (mutualismo) – como nas interações entre polinizadores e plantas.

Para Memmott (1999), em uma interação mutualista polinizador-planta os polinizadores visitam as flores das plantas em busca de alimento, normalmente pólen ou néctar (como pode ser observado na Figura 2.1). Deste modo, configura-se uma relação mutualista benéfica, quando os visitantes florais se alimentam e, as plantas, obtêm como retorno deste serviço a polinização – processo fundamental para a sobrevivência das espécies.

Figura 2.1 - Exemplo de uma interação polinizador-planta

Fonte: Webbee (2011)

Sobre este processo de interação polinizador-planta, Saraiva *et al.* (2003) chamam a atenção para o fato das abelhas serem reconhecidas como os principais polinizadores de muitas espécies de plantas e, em função disso, influenciarem diretamente na produtividade agrícola e no meio ambiente – por isso a crescente preocupação em se conservar a biodiversidade.

No entanto, a realidade de muitos países, como o Brasil, é preocupante. De acordo com Freitas *et al.* (2009), em 2009, a Floresta Amazônica sofria as mais altas taxas de desmatamento das Américas e o Brasil foi considerado o país onde este fator acontecia com maior intensidade. Os autores também identificaram que tanto os biomas da Amazônia, quanto do Cerrado, estavam sendo rapidamente substituídos por pastagens e plantações de soja, gerando um risco iminente de extinção de espécies de abelhas – algumas até sem serem conhecidas.

Face ao exposto, nota-se a importância da conservação dos ecossistemas para a manutenção do processo de interação entre polinizadores e plantas, comumente descritas na literatura como Redes de Interação polinizador-planta.

2.1.3 Redes de Interação de micro-organismos

As interações ecológicas dos micro-organismos com hospedeiros ocorrem entre os diferentes seres vivos, micro-organismos e plantas, micro-organismos e animais. Com relação às associações de animais e bactérias, pode-se citar os micro-

organismos comensais que habitam o intestino dos animais de sangue quente e têm papel importante na homeostase dos seus hospedeiros. Adicionalmente, eles são utilizados como indicadores de contaminação fecal, sendo *Escherichia coli* (*E.coli*) o mais utilizado na avaliação da qualidade microbiológica da água (CARLOS *et al.*, 2010).

Uma forma de separar as *E.coli* de acordo com a sua filogenia é por meio dos grupos filogenéticos, que são divididos em quatro agrupamentos principais: A, B1, B2 e D. Tem sido observado que essa separação está associada com a patogenicidade ou não das *E.coli*, havendo os grupos associados com linhagens patogênicas extraintestinais (B2), linhagens comensais (A e B1) e linhagens patogênicas intestinais (B1 e D). (STOPPE *et al.*, 2014a)

Nesse sentido, a partir do mapeamento da distribuição dos sub(grupos) filogenéticos em amostras de *E.coli*, alguns pesquisadores tem se utilizado deste recurso para tentar identificar fontes de contaminação de água (como WHITE *et al.*, 2011). Por meio de estudos dessa natureza também se torna possível identificar, por exemplo, como a proliferação de micro-organismos no ambiente podem gerar desdobramentos para a saúde humana (RADHOUANI *et al.*, 2014).

Como este processo de interação entre micro-organismos é de ordem microscópica, seu processo de identificação ocorre, em geral, com algum mecanismo auxiliar de apoio – por meio de amostras de corpos d'água, fezes de hospedeiros, entre outros. Na Figura 2.2 consta um exemplo de identificação da presença da bactéria de *E.coli* em uma placa de coleta de amostra.

Figura 2.2 - Exemplo da presença de *E.coli* em uma amostra

Fonte: Adiba *et al.* (2010)

Estudos com Redes de Interação de micro-organismos são trazidos para discussão, neste trabalho, em função das possibilidades de aplicação dos recursos de SNA para este domínio da Biologia – em particular, para apoiar estudos com análise de qualidade de água, por meio da identificação de padrões de frequência de distribuição de (sub)grupos filogenéticos em função de outras variáveis de análise (como clima, localização geográfica, entre outras).

Até o momento foram trazidos para o centro da análise estudos sobre conservação da biodiversidade, com um olhar particular para as Redes de Interação biológicas. Na sequência, são apresentados os tipos de dados e de interações presentes em diversos estudos da área de Informática para a Biodiversidade, bem como as formas de apresentação, técnicas, algoritmos e programas comumente utilizados.

2.2 Técnicas de Análise de Redes de Interação biológicas

Ao se analisar os estudos disponíveis, na literatura contemporânea, nos últimos anos (1999-2011), que trazem como elemento central a temática de Redes de Interação biológicas, nota-se a presença de contribuições de diversos grupos de autores, como: (i) Jane Memmott, Nickolas M. Waser e colaboradores; (ii) Diego P. Vázquez e Marcelo A. Aizen; e (iii) Jens M. Olesen, Jordi Bascompte, Pedro Jordano, Thomas M. Lewinsohn e colaboradores.

Os estudos aqui mapeados foram trazidos para análise na perspectiva de se identificar, sistematicamente, os algoritmos, programas, técnicas, tipos de dados e formas de apresentação aplicados, como descrito a seguir.

Memmott (1999), em um de seus estudos iniciais, constatou que diversos autores sugeriram o uso de Redes de Interação para o estudo de polinizadores, no domínio da Biologia, para analisar: (i) distribuição de interações entre conjunto de plantas e animais mutualistas; e (ii) níveis de generalização em interações polinizador-planta (n plantas são visitadas por *um* polinizador e n polinizadores visitam *uma* espécie de planta – é a regra, ao invés de ser a exceção).

Complementarmente, Memmott e Waser (2002) analisaram como a introdução intencional ou não-intencional de plantas e polinizadores não-nativos, dentro dos ecossistemas naturais, poderiam impactar os serviços de polinização;

isto foi possível a partir da preparação de um banco de dados contendo uma lista de flores (plantas) e seus visitantes florais (insetos), por meio do qual a Rede de Interação foi construída, descrevendo quais visitantes e flores foram observadas – neste estudo faltaram informações sobre a abundância dos insetos e flores, bem como a frequência relativa de suas interações.

Nesta mesma direção, Memmott, Waser e Price (2004) investigaram como as redes de polinização respondiam à perda de espécies, uma vez que o declínio de recursos florais é uma ameaça comumente enfrentada pelos insetos polinizadores; desta forma, os registros de flores e seus visitantes foram armazenados em um banco de dados e, para a remoção dos polinizadores, foram utilizados três diferentes algoritmos: (i) *at random*; (ii) *from least-linked* (mais especialistas no processo de visitação das plantas) *to most-linked* (mais generalistas); e (iii) *from most-linked to least-linked*. Dentre os resultados, os autores conseguiram identificar um padrão de interação em que os especialistas tendem a se associar com os generalistas, ao contrário do comportamento previsto pela visão clássica².

Ademais, alguns anos depois, Memmott *et al.* (2007) exploraram o potencial de ruptura de uma interação mutualista (entre plantas e seus polinizadores), em função de uma previsão de mudança climática; neste estudo, os autores constataram, por meio de simulações, que mudanças climáticas têm o potencial de perturbar a sobreposição temporal entre os polinizadores e os seus recursos alimentares florais, assumindo que não há mecanismos de compensação; além disso, conseguiram prever que uma fração substancial de polinizadores, especialmente os mais especializados, com pequenas amplitudes de dieta, são vulneráveis a perturbações desse tipo.

Já as pesquisas de Vázquez e Aizen (2003, 2004) com Redes de Interação polinizador-planta, trouxeram para o centro da análise um elemento adicional: os modelos nulos (*null model*) – que são descritos como uma ferramenta de apoio à análise, que pode ser usada na ausência de um mecanismo de hipótese (VÁZQUEZ; AIZEN, 2003).

Com relação às contribuições de Olesen (e seus colaboradores), é importante mencionar, num primeiro momento, o estudo publicado em 2002

² De que as plantas são agrupadas de acordo com o padrão, isto é, que os especialistas se associam com os especialistas, e os generalistas com os generalistas.

(OLESEN; ESKILDSSEN; VENKATASAMY, 2002), em que os autores examinaram as correlações ecológicas de padrões de generalização em redes de polinizadores – especialmente como padrões de interação variavam com a latitude, altitude e insularidade; no experimento, a ausência/presença da relação mutualista foi registrada por meio de dados binários (0/1) e, a ocorrência de interação na rede, representada na forma de um grafo bipartite; neste estudo os autores consideraram, também, o número de visitas encontradas, bem como o número total de interações na rede e a conectância (a porcentagem de todas as possíveis interações, dentro de uma rede, que são realmente estabelecidas).

Em outro trabalho, Bascompte, Jordano e Olesen (2006) procuraram entender como as redes mutualistas (espécie vegetal e animal) estavam organizadas e, como tal, uma organização afetava a coexistência das espécies; para tanto compilaram dados quantitativos de 19 Redes de Interação polinizador-planta, por meio do uso de matrizes ponderadas (contendo a frequência de visitação dos polinizadores às espécies de plantas) e, a partir de então, calcularam o índice de assimetria para cada par da interação. Embora os autores não tenham descrito os recursos tecnológicos utilizados, o estudo foi selecionado por ter trazido um elemento novo para a análise: o fato das interações assimétricas em redes mutualistas “parecerem ser a chave” para a diversidade e a coexistência das espécies; a partir deste trabalho, verifica-se a necessidade de se levar em consideração as técnicas computacionais que permitam analisar a presença de interações assimétricas.

Igualmente, Lewinsohn *et al.* (2006) propuseram um *framework* global para avaliar conjuntos de interações animais-plantas, de modo a detectar padrões de interação associados a processos de co-evolução e de ecologia (como exemplo, pode-se mencionar os padrões de interação de nidificação); deste modo, se utilizaram dos recursos de matrizes, redes bipartidas ou conjuntos multivariados, para realizarem as análises de correspondência de padrões.

Em um estudo publicado no ano seguinte, Olesen *et al.* (2007) analisaram 51 redes de polinização, incluindo quase 10.000 espécies e 20.000 *links*, a partir do uso da técnica de detecção de modularidade; utilizaram, então, o algoritmo *simulated annealing* (que trabalha com uma matriz binária de ausência/presença) e, deste modo, conseguiram identificar o papel dos atores em

uma Rede de Interação polinizador-planta – as espécies que desempenhavam a função de *hub*, *conector* ou ambos.

Por fim, a pesquisa de Krug e Alves-dos-Santos (2010) utilizou o recurso de Redes de Interação para analisar como a extinção de uma espécie polinizadora pode afetar a estrutura da rede; para tanto, criaram um banco de dados, contendo: (i) número total de visitas feitas por uma abelha a uma planta; (ii) matriz de ausência/presença, para a análise do grau de interação (número de interações entre espécies de abelhas e plantas); e (iii) grau médio de interações, por meio da média aritmética dos valores das espécies envolvidas (abelhas e plantas). A construção da Rede de Interação foi realizada a partir da utilização do pacote “*bipartite*”³ do programa *R* (DORMANN *et al.*, 2009).

Com base nessa revisão da literatura, foi possível identificar que boa parte dos estudos mapeados compreendiam a análise de tolerância da Rede de Interação frente à extinção/inclusão de uma espécie – analogamente ao estudo de Kumar, Novak e Tomkins (2006), que mostraram que o potencial de um ator em uma Rede Social pode ser a causa de sucesso ou declínio de uma organização. Já em relação aos tipos de dados presentes nos estudos, os mais recorrentes foram descritos na Tabela 2.1.

Tabela 2.1 - Tipos de dados presentes nos estudos de Redes de Interação mapeados

Tipo de dado	Descrição
1. Espécie polinizadora e espécie visitada	1. Identificação do polinizador e da planta polinizada.
2. Frequência de interação entre espécies	2. Número total de visitas do polinizador à planta – para analisar a intensidade/força de relações mutualistas.
3. Conectância	3. Indica a porcentagem de todas as possíveis interações, que são realmente estabelecidas, dentro de uma rede.
4. Índice de assimetria	4. Se uma planta depende fortemente de um polinizador, o polinizador depende fracamente da planta.

Fonte: autor

Embora nem todos os estudos discorram sobre os programas, técnicas e/ou algoritmos utilizados para a Análise de Redes, aqueles que mencionaram encontram-se mapeados na Tabela 2.2. Com relação às ferramentas computacionais, foi possível observar uma tendência crescente na disponibilização

³ De acordo com Dormann e Gruber (2011), o pacote “*bipartite*” do programa *R* dispõe de funções para visualizar redes e calcular uma série de índices comumente usados para descrever padrões em redes ecológicas.

de soluções abertas, com extensa documentação, interface gráfica e compatibilidade para importação/exportação de dados com outros programas. Além disso, o que se pode notar é o crescente uso de modelos nulos e de algoritmos randômicos, nas atividades de previsão de interação polinizador-planta.

Tabela 2.2 - Mapeamento dos programas, técnicas e algoritmos utilizados nos estudos

Programas utilizados (software)	Técnicas e algoritmos utilizados
Análise de Redes de Interação: <ul style="list-style-type: none"> – Pacote “<i>bipartite</i>” do programa <i>R</i>: conjunto de funções, de código livre, utilizadas para análise de redes ecológicas; – <i>Nestedness Calculator</i>: software livre, utilizado para calcular o índice de aninhamento de <i>habitats</i>, a partir de uma matriz de interação. Análise de Redes Sociais: <ul style="list-style-type: none"> – <i>Ucinet</i>: software proprietário, com interface <i>Windows</i>, para análise de dados de Redes Sociais; – <i>Pajek</i>: software livre, para uso não-comercial, com interface <i>Windows</i>, para análise de redes com grandes conjuntos de dados. 	Para simular a retirada de polinizadores da rede: <ul style="list-style-type: none"> – Algoritmo <i>at random</i>; – Algoritmo <i>from least-linked to most-linked</i>; – Algoritmo <i>from most-linked to least-linked</i>. Para randomizar interações polinizador-planta: <ul style="list-style-type: none"> – Algoritmos escritos em <i>MatLab</i>. Para a detecção de modularidade: <ul style="list-style-type: none"> – Algoritmo <i>simulated annealing</i>, que trabalha com uma matriz binária de ausência/presença. Para a geração de conjuntos de dados ao acaso: <ul style="list-style-type: none"> – Modelos nulos – ferramenta de apoio à análise, que pode ser usada na ausência de um mecanismo de hipótese (VÁZQUEZ; AIZEN, 2003).

Fonte: autor

No que se refere à forma de apresentação e visualização das relações presentes nas redes estudadas, observa-se que boa parte dos autores se utiliza dos seguintes recursos: (i) matriz de interação binária (também conhecida como matriz de ausência/presença) – na qual as linhas representam, por exemplo, os polinizadores e as colunas, por exemplo, as plantas e, quando existe interação entre espécies, a célula contém o valor “1”, caso contrário, o valor é “0” (VÁZQUEZ; AIZEN, 2003); e (ii) grafo – um diagrama em que os vértices representam os atores da rede (no caso de Redes de Interação de polinizadores – os vértices representam as plantas ou os insetos) e as linhas conectadas entre estes vértices, representam a relação entre eles (SCOTT, 2000). Em geral, um grafo é construído com o auxílio de uma ferramenta computacional (*software*) a partir de fontes de dados provenientes de matrizes de interação. Um exemplo de um grafo, para uma Rede de Interação polinizador-planta pode ser visualizado na Figura 2.3.

É importante mencionar, também, que alguns estudos (BASCOMPTE; JORDANO; OLESEN, 2006; VÁZQUEZ; AIZEN, 2004) trazem como ferramenta de apoio o uso de matrizes de frequência de interação (matrizes ponderadas) – em que

o número “1” da matriz binária é substituído por um valor que represente o número de vezes em que duas espécies foram observadas interagindo (frequência de interação).

Figura 2.3 - Exemplo de um grafo para uma Rede de Interação polinizador-planta

Fonte: adaptado de Memmott e Waser (2002, p. 2398)

Todos esses instrumentos de apoio ao processo de construção da análise de uma Rede de Interação – desde o uso de algoritmos, matrizes, grafos, entre outros – foram fundamentais para estudos de diversas naturezas. Para ilustrar, pode-se mencionar, por exemplo, a definição de um novo padrão de comportamento na relação mutualista, de que as espécies especialistas tendem a se associar com espécies generalistas (MEMMOTT; WASER; PRICE, 2004); ou ainda, que a possibilidade de se quantificar a transferência de pólen, mostrando qual espécie de inseto poliniza qual espécie de planta, potencializa o uso de rede de polinização em culturas agrícolas comerciais (MEMMOTT, 1999).

Por fim, ao se investigar o potencial de aplicação das métricas de SNA no domínio biológico, descobriu-se que diversos conceitos e algoritmos utilizados nas pesquisas sobre biodiversidade decorrem da área de Redes Sociais. Como exemplo, pode-se mencionar o uso de grafos para a representação das Redes de Interação biológicas, com uma tendência à utilização de algoritmos para detecção de *clusters* (agrupamentos). Além disso, foi possível constatar que as métricas de SNA configuram-se como um instrumento importante para mensurar, de forma sistemática, as interdependências entre polinizadores e plantas e, ao mesmo tempo, as espécies vitais à rede.

Como foram apresentados, até o momento, os estudos que gravitam em torno do tema de Análise de Redes de Interação biológicas, o foco da análise será deslocado, a partir de agora, para o domínio das Redes Sociais – pois os fundamentos conceituais presentes nas Redes de Interação são os mesmos das Redes Sociais.

2.3 Redes Sociais

Esta seção se dedica a apresentar, num primeiro momento, as origens e os conceitos básicos que permeiam a área de Redes Sociais e, na sequência, as métricas comumente usadas em Análise de Redes Sociais (SNA), com o objetivo de se identificar aquelas com potencial para aplicação em biodiversidade.

2.3.1 Fundamentos e conceitos de Redes Sociais

Segundo Hansen, Shneiderman e Smith (2011), a ciência das Redes Sociais é considerada relativamente nova, com raízes no início do século XX, as quais estão fundamentadas em dois séculos de trabalho de matemáticos acerca de topologia e teoria de redes. Em especial, no trabalho seminal de Euler (1736), que buscou identificar se um único caminho pode ser percorrido, ao longo das sete pontes de Königsberg (que conectavam ilhas no rio Pregel), sem cruzar nenhuma ponte mais de uma vez; a solução a este problema levou ao desenvolvimento da matemática presente na teoria de grafos, que conhecemos atualmente. Outras pesquisas, anos mais tarde, também foram de grande importância para os avanços nesta área, como o trabalho de Erdős e Rényi (1959) sobre grafos randômicos.

No início do século XXI, a ciência das Redes Sociais tem florescido ao lado de uma nova cultura global de ampla conectividade de rede nos processos de comunicação, advinda do uso das Tecnologias da Informação e Comunicação (HANSEN; SHNEIDERMAN; SMITH, 2011). Scott (2000) complementa que, o método de Análise de Redes emergiu no domínio das Ciências Sociais, para estudos qualitativos das conexões de atores em Redes Sociais, a partir dos recursos disponíveis na teoria de grafos atrelados às técnicas e aos recursos computacionais.

Uma vez que se tenha uma breve noção da origem desta área de pesquisa (pois não será objeto deste trabalho um aprofundamento neste tema), torna-se importante, a partir de agora, conhecer os conceitos em que se fundamentam as Redes Sociais.

Uma Rede Social pode ser descrita como um conjunto de atores (pessoas ou grupos) que se conectam por meio de relações sociais – sejam elas de amizade, trabalho em grupo ou a simples troca de informações (BRAGA; GOMES; RUEDIGER, 2008). De maneira mais formal, os principais componentes de uma rede são os vértices (nós, elementos ou atores da rede) e as conexões entre eles

(interações, relações, *edges*, *links*) – um exemplo de uma Rede de Interação entre polinizadores e plantas foi demonstrado na Figura 2.3.

Em geral, tanto nos estudos de SNA como de Redes de Interação biológicas, uma rede é representada na forma de grafo e/ou de matriz. Em alguns casos, também se faz necessário incluir alguma propriedade à linha utilizada na construção do grafo (para representar a direção ou a frequência da interação). Todos esses elementos estão descritos na Tabela 2.3.

Tabela 2.3 - Formas de representação de uma Rede de Interação

Representação	Descrição
1. Grafo	1. Compreende um conjunto de elementos (nós/vértices) e a interação entre eles (linhas).
2. Tipos de linhas utilizadas na construção de um grafo	2.1. <i>Undirected Lines</i> – apenas a representação da presença/ausência da interação é importante, independente da direção; 2.2. <i>Directed Lines</i> – a direção da interação torna-se relevante; 2.3. <i>Valued Lines</i> – registra a intensidade da interação (frequência), por meio de valores numéricos (assim, não se utiliza de entradas binárias na matriz).
3. Matriz de interação binária ou ponderada	3.1. Matriz de interação binária – uma forma de representação de um grafo por meio de uma matriz, em que os valores armazenados nas células representam a ausência/presença (0/1) da interação; 3.2. Matriz de interação ponderada – uma forma de representação de um grafo valorado, por meio de uma matriz cujas células contêm valores numéricos (frequência de interação) ao invés de simples entradas binárias.

Fonte: Scott (2000)

A exemplo do que se observa nas pesquisas com Redes de Interação biológicas, verificou-se a prevalência da utilização tanto do recurso de grafos, quanto de matrizes de interação binárias e/ou ponderadas, dependendo da necessidade de pesquisa.

Como já foram apresentados os fundamentos e conceitos de Redes Sociais, torna-se necessário descrever, a partir de agora, o conjunto de métricas (índices) disponíveis no domínio de SNA e que podem vir a apoiar as atividades com Redes de Interação biológicas. A seleção pelo conjunto de métricas, apresentadas a seguir, decorre de um processo de interação com especialistas da área de Biodiversidade, bem como de uma revisão sistemática na literatura disponível (1999-2011) acerca de estudos com Redes de Interação biológicas.

2.3.2 Métricas de Análise de Redes Sociais

De acordo com Dormann *et al.* (2009), nos últimos anos, muitas análises de redes ecológicas têm introduzido novos índices (métricas)⁴ para descrever a estrutura da rede; muitas dessas métricas provêm do domínio das Redes Sociais. Além disso, no domínio das Redes Sociais, existe um conjunto adicional de índices dedicados a analisar as características de cada vértice da rede, como descrito por Butts (2008) – alguns dos quais também estão sendo aplicados no domínio da biodiversidade. Este conjunto de métricas disponíveis para se analisar a estrutura da rede têm seus fundamentos conceituais alicerçados no domínio de Redes Sociais, por isso são denominadas, nesta pesquisa, como métricas de Análise de Redes Sociais – ou, simplesmente, métricas de SNA.

Dentre o conjunto de métricas de SNA disponíveis, observa-se a presença de dois grupos distintos, em função do aspecto da rede que se deseja analisar: (i) nível da rede (*network level*) – congrega as métricas que permitem avaliar a estrutura da rede como um todo; e (ii) nível da espécie/nó (*species/node level*) – corresponde às métricas dedicadas a analisar o papel de cada espécie/nó (vértice) da rede. Cada um desses níveis de análise é descrito a seguir.

2.3.2.1 Métricas para análise no nível da rede (*network level*)

Uma métrica no nível da rede permite analisar aspectos mais gerais da estrutura da Rede de Interação, com o objetivo de se conhecer a rede como um todo.

No trabalho de Dormann *et al.* (2009) pode-se encontrar, de forma detalhada, diversas métricas dessa natureza, tais como: conectância (*connectance*), aninhamento (*nestedness*) e sobreposição de nicho (*niche overlap*) – que comumente são utilizadas nos estudos de Redes de Interação biológicas; cada uma dessas métricas está descrita a seguir.

A métrica conectância (*connectance*) estima a proporção das interações possíveis, que de fato são realizadas; é calculada a partir da fórmula⁵ $C = L/(i*j)$, que compreende a soma das interações realizadas (L) dividida pelo produto do número de espécies presentes na matriz de interação: número de linhas

⁴ Os termos “métrica” e “índice” são usados como sinônimos, na literatura de Análise de Redes (como em DORMANN *et al.*, 2009), para expressar um cálculo matemático ou estatístico sobre algum aspecto da rede; nesta pesquisa, por padronização, optou-se pelo uso do termo “métrica”.

⁵ Mello (2012) chama a atenção para o fato de que existem variações da fórmula para calcular a métrica conectância, em função do tipo de rede que se quer analisar (unipartida ou bipartida) e da natureza de suas conexões (direcionadas, não direcionadas, entre outras).

(*i*) multiplicado pelo número de colunas (*j*) (DORMANN *et al.*, 2009). A conectância é frequentemente usada para análises de co-ocorrência de espécies, em Redes de Interação ecológicas; mais detalhes sobre o significado ecológico da conectância pode ser visto em Jordano (1987).

A métrica aninhamento (*nestedness*) compreende um padrão de interação no qual os elementos especialistas da rede bipartida interagem com espécies, que formam subconjuntos bem-definidos de espécies, com as quais interagem os elementos generalistas da rede (BASCOMPTE; JORDANO, 2007); desta forma, as conexões mais especialistas representam um subconjunto das conexões mais generalistas. De acordo com Dormman *et al.* (2009) e Bascompte e Jordano (2007) esta métrica, originalmente, foi proposta para o estudo da estrutura da rede com o objetivo de identificar o padrão de interação de espécies que viviam entre ilhas. É importante registrar que, uma estrutura de Rede de Interação aninhada favorece a redução da competição e o aumento no número de espécies coexistindo (BASTOLLA *et al.*, 2009), tornando a comunidade mais robusta à extinção (MEMMOTT; WASER; PRICE, 2004).

O cálculo do aninhamento, em geral, é realizado a partir das métricas: (*i*) T (*Temperature*)⁶ – que representa o complemento do aninhamento, então $N = 1 - T$ (ATMAR; PATTERSON, 1993); ou (*ii*) *NODF* (*Nestedness metric based on Overlap and Decreasing Fill*) – uma métrica mais recentemente desenvolvida, que incorporou novos recursos, que a métrica anterior não contemplava, como calcular o aninhamento de forma independente, entre linhas e colunas, o que permite avaliar, por exemplo, o aninhamento entre locais (composição de espécies) ou apenas entre as espécies (ocupação de espécies) (ALMEIDA-NETO *et al.*, 2008). Um estudo comparativo sobre estas métricas pode ser visto em Ulrich, Almeida-Neto e Gotelli (2009).

Por fim, a métrica sobreposição de nicho (*niche overlap*) indica similaridade no padrão de interação entre espécies de mesmo nível trófico; compreende valores entre “0” (quando não há uso comum de nicho) e “1” (quando ocorre uma perfeita sobreposição de nicho) (DORMANN *et al.*, 2009). A

⁶ É importante chamar a atenção para o fato de que diversos trabalhos (ALMEIDA-NETO; GUIMARAES; LEWINSOHN, 2007; ALMEIDA-NETO *et al.*, 2008; RODRÍGUEZ-GIRONÉZ; SANTAMARÍA, 2006; ULRICH; ALMEIDA-NETO; GOTELLI, 2009) apontam que a métrica T possui lacunas em suas bases conceituais; assim sendo, o que se tem observado, atualmente, é o uso mais comum da métrica *NODF*, no lugar da métrica T . Além disso, a métrica T só pode ser calculada para matrizes binárias, enquanto a métrica *NODF* também pode ser usada para matrizes ponderadas.

sobreposição de nicho pode ser calculada por meio das métricas E (Especialização) e C_{ws} (Coeficiente de clusterização/agrupamento ponderado), disponíveis em Araújo *et al.* (2008).

Uma vez que foram descritas algumas características e métricas no nível da rede, torna-se necessário, a partir de agora, apresentar os recursos disponíveis para analisar especificidades de cada elemento de uma Rede de Interação – conforme disposto na seção a seguir.

2.3.2.2 Métricas para análise no nível da espécie ou do nó (*species/node level*)

Uma métrica no nível da espécie ou do nó permite analisar algum aspecto particular da rede, a partir da identificação do papel de cada vértice da rede (espécie/nó).

Dentre o conjunto de métricas de SNA disponíveis, para analisar o papel de uma espécie/nó da rede, aquelas que comumente são identificadas nos estudos com Redes de Interação ou com potencial para aplicação neste domínio, são as seguintes: dependência (*dependence*), força de interação (*interaction strength*), métricas de centralidade (*degree*, *betweenness centrality*, *closeness centrality* e *eigenvector centrality*) e *w-clique* (*cohesive subgroups*). Todas estas métricas encontram-se descritas a seguir.

A métrica que permite identificar o quanto cada espécie da rede é dependente da outra é denominada dependência (*dependence*). Jordano (1987) descreve que a dependência de uma espécie i sobre uma espécie j (d_{ij}) é considerada a proporção de todas as interações da espécie i em relação à espécie j . Assim sendo, a fórmula usada para calcular esta métrica é $d_{ij} = N_{ij}/N_i$, onde N_{ij} representa o número de interações observadas entre a espécie i e a espécie j , e N_i o número total de interações identificadas para a espécie i (JORDANO; VÁZQUEZ; BASCOMPTE, 2009).

Em conjunto com a métrica dependência, muitos autores se utilizam da métrica força de interação (*interaction strength*) para mapear as espécies mais fortes na estrutura da rede. Deste modo, a força de interação de uma espécie j compreende a soma de todas as dependências da espécie j com cada espécie i ($strength_{(j)} = \sum d_{ij}$) e representa o efeito de uma espécie j na rede como um todo (VÁZQUEZ; MORRIS; JORDANO, 2005; WOOTTON; EMMERSON, 2005). Por exemplo, em uma Rede de Interação mutualista polinizador-planta, a força de

interação é usualmente medida como a contribuição dos polinizadores para a manutenção das espécies de plantas (VÁZQUEZ; MORRIS; JORDANO, 2005). Já em uma Rede de Interação de micro-organismos, a força de interação de um determinando subgrupo filogenético de *E.coli* corresponde ao seu efeito na estrutura da rede como um todo (SILVA *et al.*, 2014).

Ainda neste contexto, diversos autores (FREEMAN, 1978/79; HANSEN; SHNEIDERMAN; SMITH, 2011; SCOTT, 2000) descrevem um grupo adicional de métricas como medidas de centralidade⁷, dentre as quais estão: grau (*degree*), centralidade de intermediação (*betweenness centrality*), centralidade por proximidade (*closeness centrality*) e centralidade de autovetor (*eigenvector centrality*). As três primeiras métricas (*degree*, *betweenness centrality* e *closeness centrality*) estão fundamentadas no trabalho de Freeman (1978/79) e, de acordo com Butts (2008), refletem um “enxugamento” padronizado de uma série de medidas similares utilizadas em trabalhos anteriores. Já a última métrica de centralidade (*eigenvector centrality*) foi trazida para a discussão por ser a base para a métrica *PageRank*, que é considerada o coração da ferramenta de busca do Google. Cada uma dessas métricas é descrita, em detalhes, a seguir.

A métrica grau (k) representa o número de conexões que incidem sobre um vértice; assim sendo, o grau de vértice i é descrito como k_i . Para grafos direcionados, esta métrica pode ser desdobrada em: (i) *in-degree* – número de linhas que apontam para o vértice; e (ii) *out-degree* – número de linhas que partem de um determinado vértice até os adjacentes (HANSEN; SHNEIDERMAN; SMITH, 2011). É importante registrar que o grau de interação é uma das métricas utilizadas de forma recorrente nas análises de Redes de Interação biológicas, para identificar espécies que se relacionam com o maior número de vértices na rede (por exemplo: polinizadores/micro-organismos mais abundantes na rede).

Já a métrica centralidade de intermediação (*betweenness centrality*) permite analisar quão vital é um vértice na rede, isto é, identificar os elementos da rede que possuem a função de *bridge*; seus valores variam entre “0” (caso o vértice seja retirado, nada se altera na rede) e “1” (alto nível de centralidade entre pares de vértices na rede) (HANSEN; SHNEIDERMAN; SMITH, 2011). O cálculo desta

⁷ Detalhes adicionais acerca das fórmulas para o cálculo das métricas de centralidade estão descritos em Butts (2008). Já a aplicação prática, em uma ferramenta computacional, está disponível em Hansen, Shneiderman e Smith (2011).

métrica está baseado nas distâncias geodésicas (o menor caminho conectando um par de vértices na rede); assim sendo, é uma métrica muito útil para se analisar pontos de controle de comunicação em uma rede (FREEMAN, 1978/79).

Outra métrica disponível é a centralidade por proximidade (*closeness centrality*) que tem por objetivo calcular quão próximo cada vértice está de outro na estrutura da rede; o menor valor que esta métrica pode assumir é “1” (distância geodésica, isto é, menor distância entre dois vértices) (HANSEN; SHNEIDERMAN; SMITH, 2011). No entanto, Butts (2008) chama a atenção para o fato de esta métrica (*closeness centrality*) possuir uma fragilidade na definição do valor da proximidade, para grafos que não estão fortemente conectados (possuem vértices desconectados); neste caso, para qualquer vértice para o qual falte um caminho, o valor da proximidade será igual à zero, gerando uma exceção – em função disso esta métrica é menos frequentemente usada que as duas outras métricas de centralidade (*degree* e *betweenness centrality*).

Por fim, a centralidade de autovetor (*eigenvector centrality*) permite medir a importância de um vértice na estrutura da rede; seus valores variam entre “0” (menor relevância para a rede) e “1” (maior relevância) (HANSEN; SHNEIDERMAN; SMITH, 2011). De acordo com Butts (2008) um vértice é fundamental (importante – *coreness*) para a rede na medida em que ele tem muitas conexões com outros vértices centrais, ou ainda, quando ele consegue se conectar a outros vértices da rede por meio de uma multiplicidade de pequenos caminhos. Hansen, Shneiderman e Smith (2011) chamam a atenção para o fato de o Google utilizar uma variação da métrica *eigenvector centrality*, descrita como *PageRank*TM. Esta métrica é utilizada pelo Google para examinar a estrutura da rede (*links da web*) e, a partir de então, identificar a importância das páginas *web* para uma determinada pesquisa que se esteja realizando (GOOGLE, 2011). Detalhes adicionais sobre o cálculo da métrica *PageRank* estão disponíveis em Brin e Page (1998).

Uma vez que as medidas de centralidade foram apresentadas, resta descrever duas métricas utilizadas para identificação de padrões de agrupamento na estrutura da rede: *clique* e *w-clique*.

O *clique* é uma métrica de SNA que permite identificar os subgrupos coesos (*clusters*) na estrutura da rede. Um *clique* é composto por um conjunto de três ou mais vértices da rede totalmente conectados entre si (NOOY; MRVAR;

BATAGELJ, 2005). Entretanto, como esta métrica não leva em conta a abundância da interação para identificar os *clusters* na rede, Araújo *et al.* (2008) desenvolveram uma abordagem alternativa chamada *w-clique* para ser usada com matrizes ponderadas (abundância/frequência de interação). Com relação a esta nova abordagem, os *clusters* representam os grupos de vértices nos quais todos os vértices são conectados entre si por “conexões fortes” (conexão cujo valor/peso é maior ou igual à média de valores da rede). No trabalho de Mello (2012) está descrito o processo para a identificação dos subgrupos coesos (*w-cliques*) em uma Rede de Interação ponderada. Além disso, um estudo mais detalhado sobre a identificação de *cliques* pode ser visto em Everett e Borgatti (1998).

Foram aqui apresentados, até o momento, os conceitos e ferramentas provenientes do domínio de SNA que são a base de sustentação desta pesquisa de doutorado. Ademais, outros temas necessitam ainda ser trazidos para este contexto (como os recursos disponíveis em Análise Estatística) por se entender que são elementos que podem enriquecer o processo de Análise de Dados, nesta pesquisa.

2.4 Análise Estatística

Dentre os diversos recursos disponíveis no domínio de Análise Estatística, em função do escopo desta pesquisa, esta revisão da literatura esteve centrada em duas principais abordagens: (i) a Análise Exploratória de Dados; e (ii) a Análise Multivariada de Dados – para apoiar as atividades de conhecimento dos dados e, também, de escolha das variáveis de análise.

2.4.1 Análise Exploratória de Dados

A Análise Exploratória de Dados (*Exploratory Data Analysis* – EDA) é uma abordagem estatística comumente usada para analisar conjuntos de dados (bancos de dados) por meio da sumarização de suas principais características, a partir do uso de diversas técnicas gráficas (como *boxplot*, *scatterplot*, entre outras); embora um modelo estatístico possa ser usado, o objetivo principal desta abordagem é identificar o que os dados podem nos dizer (CHATFIELD, 1995).

A EDA nasceu dos estudos de Tukey (1977), com o objetivo de incentivar os pesquisadores para as atividades de exploração dos dados, a partir de uma perspectiva focada nos dados (Problema => Coleta de Dados => Análise =>

Modelo => Conclusões), ou seja, a coleta de dados é seguida pela atividade de análise, com o objetivo de inferir que modelo seria mais apropriado, em função da natureza dos dados. Nesse sentido, a EDA surge como uma abordagem alternativa à Análise Clássica de Dados, em que a atividade de coleta de dados era seguida pela imposição de um modelo probabilístico/determinístico, análise e conclusões – focados nos parâmetros desse modelo (Problema => Coleta de Dados => Modelo => Análise => Conclusões) (NIST/SEMATECH, 2003).

Desta forma, a abordagem de EDA pode ser usada pelo pesquisador como recurso adicional de apoio ao processo de conhecimento da natureza dos dados que se esteja trabalhando. Nesse sentido, deve englobar as seguintes tarefas: (i) visualização da distribuição dos dados – usando, por exemplo, o recurso gráfico *boxplot*; (ii) análise de correlação entre variáveis – a partir do uso de matriz de dispersão (*scatterplot*); e (iii) identificação do modelo de distribuição de probabilidade (como *Poisson*, *Gamma distribution*, entre outros). Cada um desses recursos encontra-se descrito a seguir.

2.4.1.1 Visualização da distribuição dos dados por meio de *boxplot*

Uma das primeiras atividades no processo de EDA consiste na exploração da natureza dos dados, isto é, visualizar graficamente como os dados estão distribuídos; a técnica gráfica comumente utilizada para esta finalidade é o *boxplot* (diagrama de caixa).

O *boxplot* é um recurso visual, proveniente da estatística descritiva, usado para descrever graficamente grupos de dados numéricos, por meio de três quartis, como disposto na Figura 2.4 (KRZYWINSKI; ALTMAN, 2014): o primeiro quartil (Q1) corresponde ao quartil inferior (25% dos menores valores dos dados); o segundo quartil (Q2 ou *m*) compreende à mediana dos dados (*median* – representada por uma linha dentro da caixa do gráfico); e o terceiro quartil (Q3) corresponde ao quartil superior (75% dos valores observados). Neste gráfico também está representada a amplitude inter-quartis (IQR – *interquartile range*), que cobre os 50 % dos valores centrais da distribuição (representando a caixa do gráfico, $IQR=Q3-Q1$).

Figura 2.4 - Exemplo da representação de um gráfico *boxplot*

Fonte: adaptado de Krzywinski e Altman (2014)

Nota: Q1 representa o primeiro quartil (25% dos menores valores dos dados); m ou Q2 corresponde à mediana dos dados; Q3 corresponde ao terceiro quartil (75% dos valores observados); *IQR* representa a amplitude inter-quartil, ou seja, a região que cobre 50% dos valores centrais da distribuição dos dados. Nesta figura também estão representados os *outliers* (desvios extremos da média) presentes na distribuição dos dados. Consta ainda um exemplo de um gráfico *boxplot*, diagramado com um corte/entalhe (*notch*) – recurso comumente usado quando se necessita comparar dois ou mais gráficos, pois facilita o processo de análise das medianas (se os entalhes não se sobrepõem há uma forte evidência de que as medianas são diferentes).

Além disso, o *boxplot* mostra os *outliers* (desvios extremos da média) presentes na distribuição dos dados, o que ajuda a identificar o grau de dispersão e assimetria no banco de dados. A Figura 2.5 corresponde a um exemplo de um gráfico *boxplot* apresentando a frequência de distribuição dos grupos filogenéticos de *E.coli* (STOPPE *et al.*, 2014b); neste exemplo é possível identificar a presença de *outliers* na distribuição dos grupos filogenéticos B1 e B2 – os quais são representados por meio de um asterisco (*).

Existe uma variação do gráfico *boxplot* que consiste na inclusão de um corte/entalhe (*notch*) em cada uma das laterais da caixa, como pode ser observado nas Figuras 2.4 e 2.5. Desta forma, ao se comparar dois gráficos com o uso deste recurso, se os entalhes não se sobrepõem há uma forte evidência de que as medianas são diferentes (CHAMBERS *et al.*, 1983), facilitando, assim, o processo de análise.

Figura 2.5 - Exemplo do uso do gráfico *boxplot* para representação da frequência de distribuição de grupos filogenéticos (A, B1, B2 e D) de *E.coli*

Fonte: Stoppe *et al.* (2014b)

2.4.1.2 Análise de correlação entre variáveis por meio de *scatterplot*

O *scatterplot* é uma técnica gráfica de EDA usada para gerar matrizes/gráficos de dispersão, a partir dos quais se torna possível analisar a correlação entre um conjunto de variáveis.

Aqui vale um parêntese – um coeficiente de correlação consiste em uma medida estatística para identificar o grau de correlação entre duas variáveis ou dois conjuntos de dados. Dentre os diversos coeficientes disponíveis na literatura o mais comum, especialmente em Análise de Dados biológicos (como em DUTILLEUL *et al.*, 2000), é o coeficiente de correlação de *Pearson* (também conhecido como *Pearson product-moment correlation coefficient*). De acordo com Kader e Franklin (2008), este coeficiente pode assumir valores entre “-1” e “+1”; enquanto um valor extremo (-1 ou +1) indica forte correlação linear entre os dados (correlação perfeita = variáveis dependentes), um valor igual à zero (“0”) demonstra que nenhuma correlação foi encontrada; por conseguinte, a correlação positiva (+) indica uma tendência linear crescente e, a negativa (-), uma tendência linear decrescente. Detalhes adicionais sobre coeficientes de correlação podem ser visto em Dowdy, Weardon e Chilko (2004).

A Figura 2.6 compreende um exemplo de um gráfico *scatterplot* apresentando a correlação entre os grupos filogenéticos de *E.coli*, descritos em

Stoppe *et al.* (2014b). A partir deste exemplo é possível identificar uma forte correlação linear entre os grupos filogenéticos A e B2 (-0.74) e, desta forma, pode-se inferir que estas variáveis são dependentes.

Figura 2.6 - Exemplo do uso do gráfico *scatterplot* para representação da correlação entre grupos filogenéticos (A, B1, B2 e D) de *E.coli*

Fonte: Stoppe *et al.* (2014b)

Nota: Este *scatterplot* está dividido em três partes, representando: (i) no painel superior à diagonal principal – o valor resultante da análise de correlação entre as variáveis; (ii) na diagonal principal da matriz – o histograma da frequência de distribuição dos grupos filogenéticos A, B1, B2 e D; e (iii) no painel inferior à diagonal principal – a função *smooth* (suave) que mostra se os dados apresentam uma distribuição contínua.

2.4.1.3 Modelo de Distribuição de Probabilidade

No domínio da Estatística, constam diversos modelos de distribuição de probabilidade, divididos em dois grupos principais: distribuições discretas (como *Poisson distribution*) e distribuições contínuas (como *Gamma distribution*). Enquanto o primeiro grupo lida com eventos que ocorrem em espaços amostrais contáveis, o segundo, lida com eventos que ocorrem em um espaço amostral contínuo. A seguir são apresentados dois exemplos de modelos de distribuição de probabilidade, comumente descritos na literatura, e que podem ser usados para Análise de Dados.

Poisson é um modelo de distribuição discreta, que expressa a probabilidade de um determinado número de eventos ocorrerem num intervalo fixo de tempo e/ou de espaço, se estes eventos ocorrerem com uma taxa média conhecida e independentemente do tempo decorrido desde o último evento (HAIGHT, 1967). Um teste simplificado para identificar se o modelo de distribuição dos dados é do tipo *Poisson*, é por meio da comparação dos valores da média e da variância dos dados, que devem ser iguais (WEISSTEIN, 2014b).

Por outro lado, *Gamma* é um modelo de distribuição de probabilidade contínua, mais genérico, muito utilizado quando não se conhece o padrão de distribuição dos dados. Este modelo está baseado nos parâmetros “*shape*” (forma) e “*scale*” (escala), que se desdobram em outros dois parâmetros: “*rate*” ($1/\text{scale}$) e “*mean*” (shape/rate). Basicamente o processo de identificação de um modelo de distribuição de probabilidade *Gamma* parte da distribuição do conjunto de dados da amostra (*Empirical CDF* ou *Empirical Cumulative Distribution Function*) e o compara com uma distribuição teórica *Gamma CDF* (*Cumulative Distribution Function* ou Função de Distribuição Cumulativa) – que é gerada por meio dos parâmetros “*shape*” e “*scale*” (ou suas derivações) a partir de algum algoritmo de otimização (como o *Maximum-Likelihood Estimation* – MLE). Este processo de comparação entre as duas amostras de distribuição é realizado por meio do teste *Kolmogorov-Smirnov* (*K-S test*), gerando como resultado um atributo denominado *p-value* (quanto menor o seu valor mais similares são as distribuições). Deste modo, se for constatada similaridade entre as distribuições, o modelo de distribuição dos dados é do tipo *Gamma*.

No exemplo demonstrado na Figura 2.7, como os valores de distribuição dos dados do grupo filogenético A (*Empirical CDF*) descrevem uma curva de distribuição muito próxima à *Gamma CDF*, pode-se afirmar que o modelo de distribuição de probabilidade deste conjunto de dados é do tipo *Gamma*.

Detalhes adicionais sobre estes recursos do domínio da Estatística podem ser vistos nos trabalhos de: Aksoy (2000), Jambunathan (1954), NIST/Sematech (2003) e Weisstein (2014a) – modelo de distribuição *Gamma*; Lilliefors (1967) e NIST/Sematech (2003) – teste *Kolmogorov-Smirnov*.

Figura 2.7 - Distribuição de probabilidade *Gamma* para o grupo filogenético A de *E.coli*

Fonte: Stoppe et al. (2014b)

Nota: A distribuição teórica *Gamma CDF* (*Cumulative Distribution Function* ou Função de Distribuição Cumulativa) está representada por meio de uma linha contínua vermelha e a *Empirical CDF* (frequência de distribuição do grupo filogenético A) pelos pontos azuis. Esta distribuição teórica *Gamma CDF* foi gerada por meio dos parâmetros “*shape*” e “*rate*” (cujos valores estão indicados no *label* do eixo y), a partir do algoritmo de otimização MLE (*Maximum-Likelihood Estimation*). Por fim, para identificar a similaridade entre as duas amostras de distribuição (*Gamma CDF* e *Empirical CDF*) foi realizado o teste *Kolmogorov-Smirnov*, obtendo como resultado o atributo *p-value* (indicado no *label* do eixo x).

Uma vez que a Análise Exploratória de Dados foi apresentada, outros recursos provenientes da área de Estatística podem ser incorporados ao processo de escolha das variáveis de análise, com o objetivo de dar uma maior robustez à pesquisa, como a Análise Multivariada de Dados – a qual é apresentada a seguir.

2.4.2 Análise Multivariada de Dados

Dentre os diversos métodos de Análise Multivariada de Dados, disponíveis no domínio da Análise Estatística, pode-se mencionar o *Mantel test* (MANTEL, 1967) e o *Clustering Visualizations of Multidimensional Data* (HURLEY, 2004) – que são objeto de estudo, nesta pesquisa, em função dos inúmeros recursos que

disponibilizam para apoiar as atividades de Análise de Dados. Assim sendo, cada um desses métodos é aqui descrito em detalhes.

2.4.2.1 *Mantel Test*

O teste de *Mantel* (MANTEL, 1967) permite identificar a significância estatística entre duas matrizes de dissimilaridade, a partir da aplicação do recurso de análise de correlação. Desta forma, torna-se necessário, num primeiro momento, compreender como se dá este processo de geração de uma matriz de dissimilaridade sobre uma Rede de Interação – como é descrito a seguir.

Atualmente existem diversas medidas de associação entre conjuntos de dados, disponíveis na literatura, dentre as quais encontram-se os coeficientes de similaridade e dissimilaridade (distância). Segundo Legendre e Legendre (2012) o coeficiente de similaridade (S) varia entre “1” (quando dois objetos são idênticos) e “0” (quando dois objetos são completamente diferentes), correspondendo ao complemento do coeficiente de dissimilaridade (D). Nesse sentido, o coeficiente de similaridade é representado pela seguinte equação: $S=1-D_{(normalizada)}$; assim sendo, um valor de dissimilaridade igual à “1” indica que os objetos são diferentes e um valor igual à “0” denota completa similaridade (semelhança) entre os objetos.

Para o cálculo de uma matriz de (dis)similaridade há a necessidade de aplicação de um método matemático de apoio (por exemplo, *Bray-Curtis*, *Jaccard*, *Euclidiano*, entre outros), cuja escolha deve ser realizada em função do tipo de dado que se esteja trabalhando. Desta forma, algumas recomendações devem ser seguidas (LEGENDRE; LEGENDRE, 2012): o método *Bray-Curtis* deve ser utilizado quando se trabalhar com dados de abundância (frequência de interação); o método *Jaccard*, quando se trabalhar com dados de ausência/presença (0/1); e o método *Euclidiano* para dados espaciais, como distância entre locais (sítios).

Uma vez que se tenha disponível as matrizes de dissimilaridade, sobre cada um dos conjuntos de dados, procede-se à realização do teste de *Mantel*, utilizando-se de algum método de correlação linear – como, por exemplo, o coeficiente de *Pearson*⁸ (descrito na seção 2.4.1.2 Análise de correlação entre variáveis por meio de *scatterplot*). Caso haja significância no teste de *Mantel* (“ p ”<0.05.), deve-se observar o valor do coeficiente de correlação gerado – que,

⁸ No trabalho de Dutilleul *et al.* (2000) podem ser vistos detalhes adicionais sobre a aplicação do teste de *Mantel* e do coeficiente de *Pearson* em estudos biológicos.

como já descrito anteriormente, varia entre “-1” e “1” (valores extremos indicam forte correlação positiva ou negativa); um valor igual à zero indica que não há correlação entre as matrizes.

2.4.2.2 Clustering Visualizations of Multidimensional Data

O outro método de Análise Multivariada, denominado *Clustering Visualizations of Multidimensional Data*, consiste num recurso gráfico aplicado para a construção de uma análise de agrupamento hierárquico a partir de matrizes de similaridade, permutando as variáveis em um arranjo mais efetivo, ao invés de simplesmente usar as variáveis na ordem disponível no banco de dados (HURLEY, 2004).

A inovação desse método de análise de visualização de *cluster* reside no processo de reordenação das variáveis, considerando algum índice na construção dos *clusters* (como “*single link cluster*”, “correlação de *Pearson*”, “distância Euclidiana”, entre outros) – o que não ocorre com os métodos que usam a ordem padrão das variáveis (disponível no banco de dados). No entanto, Hurley (2004) chama a atenção para o fato de que a aplicação deste método com variáveis categóricas, com poucos níveis de classificação (por exemplo, área de habitação = rural/urbana), pode gerar um viés nos resultados, mostrando agrupamentos (*clusters*) sem uma boa acurácia.

Dentre as técnicas gráficas disponíveis para este propósito, têm-se os *scatterplots* (matrizes/gráficos de dispersão), os quais são comumente usados para a visualização de dados em Análise Multivariada. Um exemplo da aplicação deste método *Clustering Visualizations of Multidimensional Data*, por meio do uso do recurso gráfico *scatterplot*, está disponível na Figura 2.8. É importante registrar que, neste método de agrupamento de clusterização hierárquica, os *clusters* mais próximos da diagonal principal da matriz de dispersão (*scatterplot*) compreendem as variáveis fortemente correlacionadas; além disso, as cores dos *clusters* representam maior similaridade no valor do índice calculado para os componentes deste grupo.

Uma vez que foram descritos os recursos provenientes do domínio da Estatística, que são utilizados como aporte a esta pesquisa, resta apresentar uma última ferramenta de apoio: a Notação para Modelagem de Processos de Negócio (BPMN).

Figura 2.8 – Exemplo da aplicação do método *Clustering Visualizations of Multidimensional Data* (reordenando as variáveis pelo índice *kmeans*)

Fonte: autor, a partir dos dados de Stoppe *et al.* (2014b)

Nota: Nesta matriz de dispersão (*scatterplot*), as mesmas cores nos *clusters* representam maior similaridade no valor do índice calculado para os componentes deste agrupamento; neste caso, os componentes próximos da diagonal principal da matriz consistem nas variáveis altamente relacionadas. Deste modo, nota-se, por exemplo, a presença de uma forte correlação entre os componentes do *micro-cluster* compreendido entre as cidades “Oslo_and_Telemark” (identificador 27) até Barcelona (identificador 32) – todas são cidades européias; isto ocorreu pelo fato da ordenação da matriz ter sido baseada na distância geográfica entre estes locais. Neste exemplo, a matriz de similaridade utilizada para a aplicação do método de Hurley (2004) foi gerada a partir do método *Bray-Curtis*.

2.5 Notação para Modelagem de Processos de Negócio

O padrão de Notação para Modelagem de Processos de Negócio (*Business Process Model and Notation* – BPMN, inicialmente descrito como *Business Process Modeling Notation*) foi desenvolvido pela organização internacional *Object Management Group*

(OMG) com o objetivo de prover uma notação de fácil entendimento por todos os usuários participantes do negócio – desde os técnicos até os usuários finais.

A especificação atualmente disponível é a BPMN 2.0 (OMG, 2011), que estende o escopo e as capacidades da versão BPMN 1.2. e provê cinco categorias básicas (objetos de fluxo, dados, objetos de conexão, raias e artefatos), as quais se subdividem em diversos elementos. Na Tabela 2.4 estão dispostos apenas os elementos básicos da BPMN que são utilizados nesta pesquisa – Capítulo 4.

Tabela 2.4 - Alguns elementos básicos da BPMN

Categoria	Elemento	Descrição	Notação
	Evento (<i>event</i>)	Representa alguma coisa que “acontece” durante um fluxo de processo; dentre os tipos disponíveis, estão os eventos de início (<i>start</i>) e fim (<i>end</i>).	Início Fim
Objetos de Fluxo (<i>flow objects</i>)	Atividade (<i>activity</i>)	Consiste em um termo genérico para uma atividade desenvolvida em um processo; caso possa ser subdivida em tarefas, recebe a denominação de SubProcesso. Enquanto a Atividade Humana deve ser executada por um especialista do domínio, a Atividade de Serviço pode ser automatizada.
	Gatilho (<i>gateway</i>)	Representa controle de fluxos de sequência em processos (desvios); Exemplos de dois tipos de desvios existentes, dentre o conjunto disponível: (i) <u>inclusivo</u> – cada caminho é independente, mas ao menos um deve ser seguido; e (ii) <u>exclusivo</u> (decisão) – apenas um caminho pode ser seguido.	Desvio Inclusivo Desvio Exclusivo
Objetos de Conexão (<i>connecting objects</i>)	Fluxo de sequência (<i>sequence flow</i>)	Mostra a ordem em que as atividades são realizadas em um processo.
Raias (<i>swimlanes</i>)	Piscina (<i>pool</i>)	Atua como um <i>container</i> gráfico para um conjunto de atividades de outros <i>pools</i> .
Artefatos (<i>artifacts</i>)	Anotação de texto (<i>text annotation</i>)	Provê informação adicional para o leitor do diagrama BPMN.

Fonte: OMG (2011) e BonitaSoft (2013)

É importante registrar que a BPMN consta nesta pesquisa de doutorado por ter sido o recurso de apoio escolhido para a diagramação da metodologia proposta, com o detalhamento de cada uma das etapas (SubProcessos) que a compõem, com suas respectivas atividades, eventos, fluxos de sequência e desvios aplicados em sua construção.

Este Capítulo de Revisão da Literatura teve por objetivo apresentar o arcabouço conceitual em que se fundamenta esta pesquisa. O Capítulo seguinte se dedica a apresentar os materiais e métodos aplicados no desenvolvimento desta pesquisa.

3 MATERIAIS E MÉTODOS

Este capítulo descreve os recursos materiais e metodológicos utilizados ao longo de todo o processo de desenvolvimento desta pesquisa de doutorado.

Para facilitar a condução dos trabalhos, as atividades foram divididas em quatro etapas: (i) revisão da literatura – para estudo dos conceitos básicos, recursos e ferramentas disponíveis; (ii) experimentação – para aprender por meio da experiência (aprender-fazendo); (iii) formulação da proposta de uma metodologia para aplicação das métricas de SNA em biodiversidade – contribuição da pesquisa; e (iv) validação da metodologia proposta – por meio de estudos de casos. Cada uma dessas etapas metodológicas é descrita a seguir.

3.1 Revisão da literatura

Esta primeira etapa do trabalho teve por objetivo o estudo dos elementos basilares que fundamentam esta pesquisa, passando por diversas temáticas: Redes de Interação biológicas, Análise de Redes Sociais, Análise Estatística e Modelagem de Processos de Negócio. Este processo de revisão da literatura se fez necessário, no início desta pesquisa, em função da necessidade de se conhecer em profundidade cada um dos temas em questão e, por conseguinte, mapear os elementos teórico-conceituais do domínio desta pesquisa.

Deste modo, num primeiro momento foi realizada uma revisão sistemática da bibliografia disponível em periódicos internacionais, anais de eventos nacionais e bancos de teses de universidades brasileiras (1999-2011), que gravitassem em torno da temática de Redes de Interação biológicas. A partir de então, foi realizado o mapeamento dos conceitos, tipos de dados e de interações, algoritmos, programas e técnicas utilizadas nesses estudos, bem como das formas de apresentação e visualização de uma Rede de Interação.

Na sequência, para adentrar no domínio de Redes Sociais, alguns passos se fizeram necessários: (i) revisão dos estudos disponíveis na literatura contemporânea que se utilizavam de algum recurso de SNA (tais como: métricas, algoritmos e ferramentas); (ii) identificação dos fundamentos conceituais da área de Redes Sociais e SNA (variáveis de análise); (iii) mapeamento das métricas de SNA, comumente empregadas para se observar aspectos da estrutura da rede (escolha e teste de algumas métricas); e (iv) avaliação das possibilidades de uso dos conceitos

e métricas de SNA em Redes de Interação biológicas, buscando identificar o potencial de sua aplicação neste contexto.

É importante registrar que a escolha pelas ferramentas computacionais, de apoio ao cálculo das métricas de SNA, levou em consideração o contexto do pesquisador biólogo (ferramentas de seu domínio ou que fossem de fácil usabilidade) e a possibilidade de interoperar os dados (importação e exportação para diversos padrões de uso comum). Além disso, optou-se pela escolha de ferramentas que fossem *software* livres ou de domínio público, e que tivessem uma extensa documentação.

Outro tema que também foi objeto de estudo, nesta etapa de revisão da literatura, foi a Análise Exploratória de Dados (*Exploratory Data Analysis* – EDA) – uma abordagem estatística alternativa à Análise Clássica de Dados, geralmente utilizada para se conhecer a natureza dos dados e apoiar o processo de seleção de ferramentas e técnicas estatísticas apropriadas; ainda no contexto da Estatística, foi realizado um estudo acerca de alguns métodos disponíveis em Análise Multivariada de Dados, em função de se dispor nesta pesquisa de doutorado de um conjunto diversificado de variáveis de análise e, em função disso, necessitar identificar correlação entre elas. Por fim, com o objetivo de se conhecer uma ferramenta utilizada para a formalização de modelagem de fluxos de processos, foi realizado um estudo sobre a Notação para Modelagem de Processos de Negócio (BPMN – *Business Process Model and Notation*).

Cada um desses conceitos que fundamentam esta pesquisa de doutorado encontra-se descrito, em detalhes, no Capítulo 2 (Revisão da Literatura). Além disso, nesta etapa de investigação do campo de pesquisa, foram desenvolvidos dois trabalhos: (i) Silva e Saraiva (2011) – um artigo completo publicado e apresentado VIII Congresso Brasileiro de Agroinformática (SBIAgro 2011); e (ii) Silva e Saraiva (2012) – um resumo publicado e apresentado no X Encontro sobre Abelhas.

3.2 Experimentos realizados

Esta segunda etapa da pesquisa teve por finalidade criar um ambiente favorável ao aprendizado, por meio da experiência, no domínio da Informática para a Biodiversidade. Isto só foi possível a partir de um processo de forte interação com especialistas em biodiversidade, que se desdobrou em atividades de estudo e

aplicação de métricas de SNA em bancos de dados de Redes de Interação de micro-organismos e de polinizadores.

A seguir são apresentados os tipos de dados e de recursos (materiais e humanos) utilizados nesta etapa da pesquisa, bem como os estudos de caso (manuscritos) desenvolvidos a partir desses experimentos.

3.2.1 Descrição dos dados

Os bancos de dados utilizados nos ensaios desta pesquisa correspondem a matrizes (binárias e ponderadas) bipartidas, em formato 'xls', contendo dados de Redes de Interação de polinizadores ou de micro-organismos – disponibilizados por pesquisadores que desenvolvem seus projetos no contexto do Núcleo de Pesquisa em Biodiversidade e Computação da Universidade de São Paulo (BioComp-USP).

Quando se trabalhou com Redes de Interação de polinizadores, tinha-se em um dos lados da matriz a representação dos polinizadores/visitantes-florais (por exemplo, nas linhas) e, do outro (nas colunas), a representação das plantas; igualmente, na intersecção de cada linha e coluna (célula) constava o valor da interação (0/1 – para matriz binária; ou um número representativo da frequência/abundância da interação – para matriz ponderada). Um exemplo de um banco de dados de polinizadores, utilizados nesta etapa da pesquisa, está disponível em Silva e Saraiva (2011).

Por outro lado, os bancos de dados das Redes de Interação de micro-organismos, sempre correspondiam a matrizes ponderadas (abundância de interação), cujas colunas representavam os (sub)grupos filogenéticos de *E.coli* e, as linhas, a cada análise, correspondiam a uma variável diferente (tais como: hospedeiros, corpos d'água, sítios/locais, hábitos alimentares, clima, entre outras). Exemplos das Redes de Interação de micro-organismos, utilizadas nesta pesquisa, estão descritos em Silva *et al.* (2014) e Stoppe *et al.* (2014a, 2014b).

3.2.2 Ferramentas de apoio (recursos materiais)

Como recursos materiais de apoio foram utilizados as métricas de SNA, bem como um conjunto de ferramentas computacionais e de Análise Estatística, as quais são descritas a seguir.

3.2.2.1 Métricas de SNA

Para realizar um estudo comparativo da estrutura de diversas Redes de Interação de polinizadores (análises no nível da rede – *network level*) foram utilizadas as métricas *connectance*, *nestedness*, *niche overlap*. Já para as diversas outras análises realizadas, com Redes de Interação de micro-organismos, que compreendiam estudos sobre o papel de determinados elementos da rede (*species/node level*), foram aplicadas as métricas *dependence*, *strength*, *betweenness centrality* e *w-clique*.

Esse conjunto de métricas encontra-se descrito, em detalhes, no Capítulo de Revisão da Literatura (seção 2.3.2 Métricas de Análise de Redes Sociais); já a forma de cálculo de cada uma delas consta no Capítulo 4 (seção 4.3 Etapa 3: Escolha das métricas de SNA adequadas ao contexto da pesquisa).

3.2.2.2 Ferramentas Computacionais

Para o cálculo das métricas de SNA foram usadas as seguintes ferramentas computacionais de apoio:

- (i) os pacotes “*bipartite*” (DORMANN *et al.*, 2009) e “*sna*” (BUTTS, 2008) do programa *R* (*The R Foundation for Statistical Computing*) – para analisar aspectos gerais (*network level*) e específicos (*species level*) da estrutura das Redes de Interação;
- (ii) os programas *Dieta* (ARAÚJO *et al.*, 2008), *Pajek* (BATAGELJ; MRVAR, 1998) e *Ucinet* (BORGATTI; EVERETT; FREEMAN, 2002) – para identificar padrões de agrupamento em redes ponderadas (*w-cliques*) e visualizar os resultados na forma de grafo.

Os passos que devem ser seguidos para o cálculo das métricas de SNA, por meio de cada uma dessas ferramentas computacionais, encontram-se descritos nas Tabelas 4.3 e 4.4 do Capítulo 4 desta tese (seção 4.3 Etapa 3: Escolha das métricas de SNA adequadas ao contexto da pesquisa).

Para a geração dos *scripts* no *R* foi utilizado como apoio a ferramenta *RStudio*⁹ (versão 0.97.312 2009-2012), que é um programa *open source* para o ambiente de estatística computacional *R*.

⁹ O programa *RStudio*, nas modalidades *desktop* ou *server*, encontra-se disponível para *download* em: <<http://www.rstudio.com/ide/download/>>. Acesso em: 06 maio 2014.

Com relação ao conjunto de programas usados para a geração dos *w-cliques*, primeiramente foi aplicado o *Dieta*¹⁰ (*Dieta1.exe*, versão de 30 out. 2008), na sequência o *Pajek*¹¹ (versão 2.05, set. 2011) e, por fim, o *Ucinet*¹² (versão 6.199, 1999-2007).

O programa *Pajek* (BATAGELJ; MRVAR, 1998), disposto na Figura 3.1, também é muito usado para a geração de grafos em 2D e 3D, em função de dispor de algoritmos para a geração automática de *layouts* da rede, como Kamada e Kawai (1989) e Fruchterman e Reingold (1991); além disso, disponibiliza recursos para o cálculo e a plotagem de diversas métricas de SNA, como as métricas de centralidade *betweenness* e *closeness centrality* – durante esta pesquisa a ferramenta *Pajek* foi utilizada para a geração do grafo *betweenness centrality*, disposto em Silva *et al.* (2014). Informações adicionais, sobre os recursos disponíveis neste programa, podem ser consultadas em Batagelj e Mrvar (2014).

A escolha por estas ferramentas, aqui descritas, ocorreu em função de disporem de uma interface gráfica para interação do usuário (o que não ocorre com pacotes baseados em linguagens de *script* e/ou programação) e, assim, facilitarem a aprendizagem e o uso – em especial, pelos pesquisadores em biodiversidade.

¹⁰ O código fonte, manual e arquivo executável do programa *Dieta* estão disponíveis para *download* na seção *Suplement* de Araújo *et al.* (2008), em: <<http://esapubs.org/archive/ecol/E089/115/suppl-1.htm>>. Acesso em: 06 maio 2014.

¹¹ *Pajek*: palavra eslovena para aranha (*spider*) – um programa para análise e visualização de redes complexas, disponibilizado livremente para usuários não comerciais. *Download* disponível em: <<http://pajek.imfm.si/doku.php?id=download>>. Acesso em: 06 maio 2014.

¹² O *Ucinet* é um programa para Análise de Redes Sociais (SNA) e encontra-se disponível para *download* em: <<https://sites.google.com/site/ucinetsoftware/downloads>>. Acesso em: 06 maio 2014.

Figura 3.1 - Pajek: interface do programa e representação do grafo de uma Rede de Interação

Fonte: autor, a partir dos dados de Saraiva e Silva (2010)

3.2.2.3 Ferramentas de Análise Estatística

Como ferramenta de apoio à atividade de escolha das variáveis de análise, foram utilizados os recursos provenientes da área de Estatística, em particular a Análise Exploratória de Dados (*Exploratory Data Analysis* – EDA) e a Análise Multivariada de Dados (*Mantel Test* e *Clustering Visualizations of Muldimensional Data*).

Para a aplicação da EDA, foram seguidos os passos metodológicos sugeridos por esta abordagem: Identificação do Problema => Coleta de Dados => Análise => Modelo => Conclusões. Como a identificação do problema e a coleta de dados já haviam sido realizadas (no escopo dos projetos do BioComp-USP), a proposta aqui, nesta pesquisa de doutorado, foi se dedicar às atividades de análise dos dados, identificação do modelo de distribuição de probabilidade, para então, em conjunto com os especialistas em biodiversidade, gerar as conclusões derivadas do problema de pesquisa. Nesse sentido, foram usados os seguintes recursos de apoio: (i) *boxplot* – para a visualização da distribuição dos dados; (ii) *scatterplot* – para analisar a correlação entre as variáveis; e (iii) *Poisson* e *Gamma distribution* – para identificar o modelo de distribuição de probabilidade dos dados. As orientações sobre a forma de aplicação de cada um desses recursos de EDA (pacotes e funções do programa *R*, bem como os passos a serem seguidos) estão dispostas na Tabela

4.1 do Capítulo 4 desta tese (seção 4.2 Etapa 2: Definição das perguntas-chave a serem respondidas e das variáveis de análise).

O outro recurso presente no domínio da Estatística e que também foi aplicado nesta pesquisa é a Análise Multivariada de Dados, por meio dos métodos *Mantel Test* (MANTEL, 1967) e *Clustering Visualizations of Muldimensional Data* (HURLEY, 2004).

Para o teste de *Mantel* foram geradas, inicialmente, as matrizes de dissimilaridade dos dados e, posteriormente, realizado o cálculo de *Mantel*. Para a aplicação do método *Clustering Visualizations of Muldimensional Data* foi gerada a matriz de similaridade dos dados (complemento da matriz de dissimilaridade) e, na sequência, reordenadas as variáveis da matriz (a partir de algum índice de clusterização hierárquica) para serem apresentadas em um gráfico de dispersão (*scatterplot*).

Nos dois casos foram usadas como ferramentas de apoio as funções disponíveis nos pacotes do programa *R*, as quais estão detalhadas na Tabela 4.2 do Capítulo 4 (seção 4.2 Etapa 2: Definição das perguntas-chave a serem respondidas e das variáveis de análise), em conjunto com os passos que devem ser seguidos para aplicação de cada uma delas em um conjunto de dados.

3.2.3 Interação com especialistas em biodiversidade (recursos humanos)

Os experimentos desenvolvidos nesta pesquisa de doutorado foram realizados a partir da interação com especialistas de diversas áreas do conhecimento: Ecologia, Genética, Análise de Redes, Análise Estatística e Ciência da Computação – tornando-se este um diferencial na pesquisa, pois foi estabelecido um ambiente favorável de cooperação multidisciplinar.

Os estudos com Redes de Interação microbiológicas foram realizados a partir da parceria com as pesquisadoras Profa. Dra. Tatiana T. Torres (Departamento de Genética do Instituto de Biociências da USP) e Doutoranda Nancy C. Stoppe (Centro de Biologia Molecular e Engenharia Genética da UNICAMP), com o objetivo de entender os tipos de dados utilizados e identificar os problemas de pesquisa que pudessem ser analisados a partir do uso de métricas de SNA. Ainda nesta etapa da pesquisa, houve também uma forte interação com os seguintes pesquisadores: (i) Prof. Dr. Marco A. R. Mello, especialista em Análise de Redes Ecológicas (Laboratório de Ecologia do Mutualismo do Instituto de Ciências

Biológicas da UFMG), que muito contribuiu para o entendimento das métricas de SNA e de suas possibilidades de aplicação neste domínio microbiológico; e (ii) Dr. Raúl Jiménez Rosenberg, especialista em Análise Estatística (*Comisión Nacional para el Conocimiento y Uso de la Biodiversidad* – CONABIO, México, Distrito Federal), que trouxe grandes contribuições acerca do processo de escolha das variáveis de análise a partir do uso dos recursos de Análise Exploratória de Dados (EDA) e de Análise Multivariada (em particular, *Clustering Visualizations of Multidimensional Data*).

Ademais, os estudos com Redes de Interação de polinizadores foram realizados a partir de um processo de colaboração com as pesquisadoras Profas. Dras. Astrid M. P. Kleinert e Tereza C. Giannini (Laboratório de Abelhas do Instituto de Biociências da USP), com o objetivo de aplicar os recursos disponíveis no domínio de SNA como uma das ferramentas de análise para entender o papel de uma espécie não nativa (*A. mellifera*) e, outra, nativa (*T. spinipes*) nas Redes de Interação abelha-planta brasileiras. Este trabalho, que ainda está em andamento, conta com a participação do pesquisador Prof. Dr. Lucas A. Garibaldi (*Universidad Nacional de Río Negro*, Argentina), especialista em Análise Estatística aplicada à Conservação Ecológica, que trouxe um ferramental de apoio desta área de conhecimento como contribuição para este trabalho, bem como valorosas orientações durante o processo de análise dos dados e resultados.

3.2.4 Manuscritos desenvolvidos

A concretização dos estudos realizados nesta etapa da pesquisa, bem como o processo de colaboração com especialistas de diversas áreas do conhecimento, culminou com o desenvolvimento de cinco manuscritos:

- (i) “*Clustering of water bodies in unpolluted and polluted environments based on Escherichia coli phylogroup abundance using a simple interaction database*” (STOPPE et al., 2014a) – Apêndice A;
- (ii) “*Social Network Analysis Metrics and Their Application in Microbiological Network Studies*” (SILVA et al., 2014) – Apêndice B;
- (iii) “*Complex pattern of Escherichia coli phylogenetic group distribution in worldwide samples of human feces and wastewater treatment plants*” (STOPPE et al., 2014b) – Apêndice C;

- (iv) “*Native and non-native supergeneralist bee species have different impacts on plant-bee networks*” (GIANNINI *et al.*, 2014) – *in preparation*; e
- (v) “*A Methodology for Applying Social Network Analysis Metrics on Biodiversity*” (SILVA; SARAIVA, 2015) – *submitted*.

Aqui vale um parêntese – os estudos de caso gravitaram, em sua maioria, no domínio das Redes de Interação de micro-organismos, por serem um campo pesquisa ainda pouco explorado pelos recursos de SNA e haver uma necessidade crescente para identificação de padrões de agrupamento e co-ocorrência de micro-organismos. Já o trabalho com Redes de Interação de polinizadores (GIANNINI *et al.*, 2014) não será aqui apresentado, pois ainda encontra-se em fase de revisão.

Concluída esta etapa de experimentação, cujo objetivo foi a realização de diversos ensaios com dados de biodiversidade, para que fosse possível aprender por meio da experiência, o passo seguinte consistiu na formulação da proposta de uma metodologia para aplicação de métricas de SNA em Redes de Interação biológicas, como descrito a seguir.

3.3 Formulação da proposta de uma metodologia

A terceira etapa desta pesquisa de doutorado teve por finalidade elaborar a proposta de uma metodologia para apoiar os pesquisadores em biodiversidade, a partir de todas as lições apreendidas no processo de experimentação realizado na etapa anterior.

Nesse sentido, a metodologia proposta compreende uma sequência de ações e passos a serem realizados, para aplicação das métricas de SNA em estudos de Redes de Interação biológicas, tendo como aporte as ferramentas disponíveis no domínio da Ciência da Computação e da Estatística. A formalização de cada uma das etapas desta metodologia foi realizada por meio da Notação para Modelagem de Processos de Negócio (BPMN – *Business Process Model and Notation*), por dispor de recursos visuais que facilitam o entendimento dos fluxos a serem seguidos/escolhidos, por qualquer pesquisador que quera dispor deste instrumental de apoio. A ferramenta computacional utilizada para a construção de

cada um dos diagramas em BPMN foi o programa *open source Bonita BPM 6.0* (BONITASOFT, 2013).

É importante registrar que a construção desta metodologia, desde seu início, teve a colaboração de especialistas em biodiversidade, que ajudaram na formalização e validação deste instrumento; esta forte interação com pesquisadores contribuiu para a identificação de lacunas, bem como, para a inclusão de recursos necessários para o contexto estudado – isto ocorreu, por exemplo, quando se identificou a necessidade da inclusão de uma abordagem estatística (como a Análise Exploratória de Dados), para potencializar o entendimento sobre os dados e as estruturas subjacentes a eles.

Ademais, esta metodologia configura-se como a contribuição original desta pesquisa de doutorado e foi concebida a partir das necessidades reais emanadas da área de Biodiversidade. Uma versão preliminar desta proposta foi apresentada no IX Congresso Brasileiro de Agroinformática (SBIAgro 2013), como artigo completo (SILVA; SARAIVA, 2013); a sua versão final está descrita no Capítulo 4 desta tese e, também, em Silva e Saraiva (2015).

3.4 Validação da metodologia proposta por meio de estudos de caso

A última etapa desta pesquisa teve por finalidade avaliar a adequação da metodologia proposta, a partir de um processo de revisitação dos estudos de caso desenvolvidos (provas de conceito).

Como está descrito no Capítulo 5, os estudos de caso foram submetidos à metodologia, seguindo os fluxos de cada processo de negócio estabelecido; assim, conseguiu-se identificar as contribuições que a metodologia proposta traz para os pesquisadores e tomadores de decisão na área de Informática para a Biodiversidade, guiando outros esforços de aplicação.

Ao término deste Capítulo 3, espera-se que o leitor tenha tido uma visão geral de todos os passos metodológicos seguidos e os recursos materiais aplicados, ao longo do desenvolvimento desta pesquisa de doutorado. O próximo Capítulo se dedica a apresentar a contribuição desta pesquisa: a proposta de uma metodologia orientativa para apoiar/guiar os pesquisadores em biodiversidade na aplicação das métricas de SNA em Redes de Interação biológicas.

4 UMA METODOLOGIA PARA APLICAÇÃO DE MÉTRICAS DE ANÁLISE DE REDES SOCIAIS EM BIODIVERSIDADE

Tendo em vista a multiplicidade de recursos e ferramentas disponíveis na área de Análise de Redes Sociais (*Social Network Analysis* – SNA), este capítulo propõe uma metodologia para aplicação de métricas de SNA em pesquisas com Redes de Interação biológicas.

Esta proposta de metodologia nasceu de um processo de colaboração com pesquisadores das áreas de Computação e Biologia, a partir de projetos desenvolvidos no Núcleo de Pesquisa em Biodiversidade e Computação (BioComp-USP)¹³ – um ambiente favorável à interação transdisciplinar, que tornou possível a transposição de recursos disponíveis na área de Redes Sociais (como as métricas de SNA) para o domínio da biodiversidade. Dentre os estudos emanados deste processo de interação, que serviram de base para o delineamento desta metodologia, pode-se mencionar: na área de Ecologia – Silva e Saraiva (2012), Giannini *et al.* (2014); e na área de Microbiologia – Silva *et al.* (2012, 2014), Stoppe *et al.* (2014a, 2014b).

A metodologia foi formalizada por meio da Notação para Modelagem de Processos de Negócio (BPMN – *Business Process Model and Notation*) e está estruturada em quatro etapas: (i) Etapa 1 – mapeamento dos tipos de dados e de interação disponíveis; (ii) Etapa 2 – definição das perguntas-chave e das variáveis de análise; (iii) Etapa 3 – escolha das métricas de SNA adequadas; e (iv) Etapa 4 – realização de análises biológicas com SNA.

É importante mencionar que esta metodologia foi concebida tendo como público alvo o especialista em biodiversidade (biólogo), que trabalha com Redes de Interação biológicas, para que possa servir como guia para o uso das métricas de SNA em suas pesquisas e apoiá-lo no processo de tomada de decisão. Desta forma, todas as etapas dessa metodologia devem ser conduzidas por este especialista – pois é ele quem conhece em profundidade a natureza dos dados e o domínio da pesquisa.

Cada uma das etapas, componentes da metodologia, é apresentada na Figura 4.1 e está descrita, em detalhes, nas seções a seguir.

¹³ <http://www.biocomp.org.br/>

4.1 Etapa 1: Mapeamento dos tipos de dados e de interação disponíveis

Esta primeira etapa da metodologia compreende as atividades de mapeamento dos tipos de dados e de interação disponíveis para a realização da pesquisa (conforme disposto na Figura 4.2).

A primeira atividade, em uma pesquisa com dados de interação, compreende a tarefa exploratória de conhecimento dos dados que se têm disponíveis. Em geral, os biólogos dispõem de uma Rede de Interação bipartida (por exemplo: visitante-floral x planta, reservatório de água x grupos filogenéticos) organizada por meio de uma matriz de interação binária (0/1) ou de uma matriz de interação ponderada (frequência/abundância da interação – como demonstrado na Figura 4.3); neste exemplo, as linhas da matriz representam as plantas (flores produtoras de óleo) e, as colunas, os visitantes florais (abelhas); os dados compreendem a frequência de interação entre o visitante floral e a planta, constituindo uma matriz de interação ponderada.

É importante lembrar que a matriz de interação deve ser preparada no formato adequado à ferramenta computacional, que será utilizada para calcular as métricas de SNA. Em geral, corresponde a um arquivo texto (txt), contendo os dados da matriz bipartida (com ou sem os *labels* de linhas/colunas, dependendo do programa) e os separadores específicos para os dados (tabulações/espços em branco) – detalhes adicionais acerca do formato adequado para preparar uma matriz de interação, aos diversos programas de Análise de Redes, podem ser vistos em Mello (2012).

Outra questão que deve ser levada em conta, ainda neste processo de mapeamento e preparação dos dados, quando se trabalhar com matriz ponderada (frequência/abundância de interação), é se os dados estão na sua forma bruta (*raw data*) ou normalizados (*normalized data* – frequência relativa). Legendre e Legendre (2012) sugerem que os dados de frequência de distribuição sejam normalizados, pois refletem melhor o papel de cada espécie da rede, que os dados brutos; evitando, assim, algum viés no processo de análise. Desta forma, dentro dos fluxos desta etapa da metodologia, há uma verificação sobre a necessidade de se normalizar os dados – ficando a decisão a cargo do especialista em biodiversidade.

Figura 4.1 - Metodologia para aplicação de métricas de SNA em biodiversidade (modelada em BPMN)

Figura 4.2 - Etapa1 da metodologia: Mapeamento dos tipos de dados e de interação (modelada em BPMN)

Uma vez que se tenha preparado a(s) matriz(es) no formato adequado, para um determinado conjunto de dados, caso surja a necessidade de estruturar outras matrizes (para novos conjuntos de dados), o fluxo retorna para a etapa de identificação do tipo da matriz de interação, recomeçando a sua dinâmica de atividades.

Figura 4.3 - Exemplo parcial de uma matriz de interação ponderada de visitantes-florais x plantas

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		<i>Centris aenea</i>	<i>Centris fuscata</i>	<i>Centris caxienseis</i>	<i>Centris tarsata</i>	<i>Centris flavifrons</i>	<i>Centris trigonoides</i>	<i>Centris obsoleta</i>	<i>Epicharis sp2</i>	<i>Apis mellifera</i>	<i>Centris sp3</i>	<i>Centris sp1</i>	<i>Xylocopa sp</i>	<i>Xylocopa grisescens</i>
2	<i>Diplopterys pubipetala</i>	1368	1364	740	460	416	256	0	328	364	368	0	0	0
3	<i>Byrsonima gardnerana</i>	924	320	2108	464	284	0	0	0	28	0	0	0	0
4	<i>Banisteriopsis muricata</i>	396	468	108	140	272	652	912	364	44	0	368	164	84
5	<i>Heteropterys sp1</i>	764	680	528	308	404	300	0	28	0	0	0	76	0
6	<i>Heteropterys sp2</i>	740	656	528	332	0	324	0	116	116	0	0	0	0
7	<i>Dicella bracteosa</i>	556	512	356	132	524	0	0	0	0	0	0	0	0
8	<i>Carolus chasei</i>	604	452	432	200	0	0	0	0	0	0	0	0	0
9	<i>Stigmaphyllon paralias</i>	504	816	0	0	0	0	0	0	0	0	0	0	0
10	<i>Banisteriopsis stellaris</i>	292	300	244	116	0	0	0	0	0	0	0	0	0
11	<i>Banisteriopsis schizopter</i>	228	224	124	120	0	0	0	0	0	0	0	0	0

Fonte: adaptado de Bezerra, Machado e Mello (2009)

Aqui vale um parêntese – nesta metodologia não foram tratadas questões acerca da qualidade dos dados presentes na matriz de interação, pois entende-se que o pesquisador, por conhecer em profundidade a natureza dos dados e o seu problema de pesquisa, já disponha de um banco de dados que tenha sido validado anteriormente.

Como produto final desta primeira etapa da metodologia, dispõe-se de uma ou mais matrizes de interação, preparadas no formato adequado aos programas que se precisar trabalhar, nas etapas subsequentes. O próximo passo (Etapa 2) consiste na definição das perguntas-chave e das variáveis de análise, como descrito na seção a seguir.

4.2 Etapa 2: Definição das perguntas-chave a serem respondidas e das variáveis de análise

Este é um momento de suma importância para a realização da pesquisa, pois compreende a etapa em que se delineiam as perguntas-chave que se deseja responder com o apoio das métricas de SNA, diante de uma necessidade de

pesquisa. Aqui são eleitas as variáveis, dentre as disponíveis no banco de dados, que farão parte da análise. A Figura 4.4 apresenta o diagrama com as duas atividades componentes desta etapa da metodologia: (i) definir as perguntas-chave a responder; e (ii) escolher as variáveis participantes da análise.

A primeira atividade, nesta etapa da metodologia, consiste na definição das perguntas-chave que necessitam ser respondidas, no desenvolvimento da pesquisa; esta é uma atribuição eminentemente do especialista em biodiversidade, pois é ele quem tem conhecimento da natureza do problema e das necessidades da pesquisa. Exemplos de perguntas-chave, em estudos com Redes de Interação de polinizadores, estão presentes em: (i) Kleinert e Giannini (2012) – “Como identificar as espécies de abelhas mais generalistas em pesquisas conduzidas em diferentes localidades no Brasil?”; e (ii) Krug e Alves-dos-Santos (2010) – “Como a extinção de uma espécie polinizadora pode afetar a estrutura da Rede de Interação?”.

O próximo passo, ainda nesta etapa da metodologia, compreende a escolha das variáveis participantes da análise – o detalhamento das atividades componentes deste subprocesso da metodologia está disposto na Figura 4.5. Em geral, observa-se um número considerável de variáveis disponíveis nos bancos de dados de Redes de Interação, tais como: (i) frequência/abundância de interação; (ii) local de coleta – sítios; (iii) clima; (iv) temperatura; (v) precipitação; entre outras.

Este processo de escolha das variáveis não é algo trivial, por isso merece atenção e o uso de recursos adicionais de apoio, como aqueles provenientes do domínio da Análise Estatística – que permitem, por exemplo, realizar a análise de correlação entre as variáveis, facilitando a identificação das variáveis dependentes e independentes no banco de dados e, por conseguinte, o processo de escolha daquelas que farão parte da análise.

A metodologia sugere que seja utilizada a abordagem de Análise Estatística, como ferramenta de apoio a este processo de eleição das variáveis. Dentre os diversos recursos disponíveis nesta área do conhecimento, temos a Análise Exploratória de Dados (EDA) e a Análise Multivariada de Dados. Os passos que devem ser seguidos na aplicação de cada uma dessas ferramentas de apoio é descrito, em detalhes, a seguir.

Figura 4.4 - Etapa2 da metodologia: Definição das perguntas-chave e variáveis (modelada em BPMN)

Fonte: autor

Figura 4.5 - Etapa2-subprocesso: Escolher variáveis de análise (modelada em BPMN)

Fonte: autor

A abordagem de EDA (TUKEY, 1977) deve ser usada para ajudar o pesquisador a conhecer a natureza dos dados, englobando as seguintes tarefas: visualização da distribuição dos dados (usando *boxplot*), análise de correlação entre variáveis (usando *scatterplot*) e, por fim, identificação do modelo de distribuição de probabilidade (*Poisson*, *Gamma distribution*, entre outros). As orientações acerca da aplicação de EDA em um conjunto de dados estão descritos a seguir e, também, sumarizados na Tabela 4.1.

Esta primeira etapa de EDA, que compreende a tarefa de visualização da distribuição dos dados por meio da geração de um gráfico *boxplot*, pode ser realizada a partir do uso da função “*boxplot*”, disponível no pacote “*graphics*” (R DEVELOPMENT CORE TEAM, 2012b) do programa *R*; caso seja necessário criar um gráfico *boxplot* com a inclusão de um corte/entalhe (*notch*), basta que o parâmetro “*notch*” seja indicado como “*true*”.

Na sequência, a tarefa de identificação de análise de correlação entre variáveis por meio da geração de um gráfico *scatterplot* pode ser realizada a partir aplicação da função “*pairs*”, disponível no pacote “*graphics*” (R DEVELOPMENT CORE TEAM, 2012b) do programa *R*; esta função dispõe de alguns parâmetros auxiliares que permitem a inclusão de outras informações no *scatterplot* em cada uma das partes da matriz: (i) painel superior à diagonal principal – parâmetro “*upper.panel*”; (ii) diagonal principal da matriz – parâmetro “*diag.panel*”; e (iii) painel inferior à diagonal principal – parâmetro “*lower.panel*”. Desta forma, além de plotar a correlação entre as variáveis, pode-se incluir, por exemplo, o histograma da frequência de distribuição dos dados e outra função de distribuição que necessitar. Um exemplo da aplicação desses recursos encontra-se disposto na Figura 2.6 (Capítulo Revisão da Literatura).

A última tarefa no processo de EDA consiste na identificação do modelo de distribuição de probabilidade dos dados; para tanto, recomenda-se que esta atividade tenha início a partir de um modelo clássico (como *Poisson distribution*) e, se necessário, seja realizado um teste com um modelo mais geral (como *Gamma distribution*).

Para identificar se o modelo de distribuição dos dados é do tipo *Poisson*, basta realizar um teste simplificado por meio da comparação dos valores dos parâmetros média (*mean*) e variância (*variance*). Se os valores

desses parâmetros forem iguais o modelo de distribuição dos dados é *Poisson*. Para cálculo dos valores dos parâmetros “*mean*” e “*variance*” basta usar a função “*ddply*” (WICKHAM, 2011), disponível no pacote “*plyr*” do programa *R*.

Tabela 4.1 - Resumo de recursos utilizados para aplicar EDA

Recurso	Como aplicar o recurso?
Boxplot	Função “ <i>boxplot</i> ” do pacote “ <i>graphics</i> ” (R DEVELOPMENT CORE TEAM, 2012b) do programa <i>R</i> .
Scatterplot	Função “ <i>pairs</i> ” do pacote “ <i>graphics</i> ” (R DEVELOPMENT CORE TEAM, 2012b) do programa <i>R</i> .
Poisson distribution (modelo de distribuição de probabilidade discreta)	<p>1º passo: Usar a função “<i>ddply</i>” (WICKHAM, 2011) do pacote “<i>plyr</i>” do programa <i>R</i>, para calcular os valores dos parâmetros média (“<i>mean</i>”) e variância (<i>variance</i> – “<i>var</i>”), para o conjunto de dados em estudo.</p> <p>2º passo: Se os valores dos parâmetros “<i>mean</i>” e “<i>variance</i>” forem iguais, o modelo de distribuição de probabilidade é <i>Poisson</i>.</p>
Gamma distribution (modelo de distribuição de probabilidade contínua)	<p>1º passo: Usar a função “<i>fitdistr</i>” do pacote “<i>mass</i>” (RIPLEY <i>et al.</i>, 2013) do programa <i>R</i>, que implementa o algoritmo de otimização <i>Maximum-Likelihood Estimation</i> (MLE), para o cálculo dos valores dos parâmetros “<i>shape</i>” e “<i>rate</i>”;</p> <p>2º passo: Aplicar a função “<i>pgamma</i>”, disponível no pacote “<i>stats</i>” (R DEVELOPMENT CORE TEAM, 2012a) do programa <i>R</i>, para gerar uma amostra de distribuição teórica <i>Gamma CDF</i> (<i>Cumulative Distribution Function</i>);</p> <p>3º passo: Usar a função “<i>ks.test</i>” (teste <i>Kolmogorov-Smirnov</i>) do pacote “<i>stats</i>” (R DEVELOPMENT CORE TEAM, 2012a) do programa <i>R</i> para comparar o conjunto de dados em estudo (<i>Empirical CDF</i>) com a distribuição teórica <i>Gamma</i> (<i>Gamma CDF</i>);</p> <p>4º passo: Analisar o resultado do parâmetro de saída <i>p-value</i> da função “<i>ks.test</i>” – quanto menor o seu valor, mais similares são as distribuições dos dados, confirmando-se que o modelo de distribuição de probabilidade é do tipo <i>Gamma</i>.</p>

Fonte: autor

Já para testar se um modelo de distribuição de probabilidade é do tipo *Gamma*, uma série de passos se fazem necessários, a começar pelo cálculo dos parâmetros “*shape*” (forma) e “*scale*” (escala), que se desdobram em outros dois parâmetros: “*rate*” (1/*scale*) e “*mean*” (*shape/rate*) – que são elementos basilares deste modelo de distribuição. Sugere-se que sejam escolhidos os parâmetros “*shape*” e “*rate*” para a análise, pois eles podem ser calculados para o modelo de distribuição *Gamma* a partir do algoritmo de otimização *Maximum-Likelihood Estimation* (MLE); este algoritmo está implementado na função “*fitdistr*” do pacote “*mass*” (RIPLEY *et al.*, 2013) do programa *R*. Assim sendo, basta que sejam informados, como parâmetro de entrada para a função “*fitdistr*”, a distribuição do conjunto de dados e o nome

do modelo de distribuição que se está testando (“*gamma*”). Na sequência, deve-se gerar uma amostra de distribuição teórica *Gamma CDF* (*Cumulative Distribution Function* – Função de Distribuição Cumulativa), baseada nos valores de distribuição dos dados e dos parâmetros “*shape*” e “*rate*”, utilizando como base a função “*pgamma*”, disponível no pacote “*stats*” (R DEVELOPMENT CORE TEAM, 2012a) do programa *R*. Por fim, o teste *Kolmogorov-Smirnov* (*K-S test*) deve ser usado para comparar a distribuição do conjunto de dados em estudo (*Empirical CDF* ou *ECDF*) com a distribuição teórica *Gamma* (*Gamma CDF*) e, assim, identificar quão similares são as distribuições a partir do valor do atributo *p-value*; quanto menor o valor do *p-value*, mais similares são as distribuições. O teste *K-S* pode ser realizado a partir da função “*ks.test*”, que também encontra-se disponível no pacote “*stats*” (R DEVELOPMENT CORE TEAM, 2012a) do programa *R*.

Ainda no domínio da Análise Estatística, a abordagem de Análise Multivariada dispõe de um conjunto de métodos de apoio, dentre os quais esta metodologia sugere a aplicação dos seguintes: (i) *Mantel test* (MANTEL, 1967); e (ii) *Clustering Visualizations of Muldimensional Data* (HURLEY, 2004).

O teste de *Mantel* pode ser usado, a partir da comparação de duas matrizes de dissimilaridade (tais como: sítios e dados de abundância de espécies; ou sítios e clima), para auxiliar no processo de identificação das variáveis que possuem forte correlação – ou seja, aquelas variáveis que influenciam no comportamento da outra (variáveis dependentes). Por exemplo, ao se constatar uma forte correlação entre dados de distribuição de espécies/organismos com um determinado clima, pode-se inferir que existe um padrão de distribuição, em função deste clima. Detalhes acerca dos recursos e passos a serem seguidos para aplicar o teste de *Mantel* em uma Rede de Interação estão descritos a seguir e, também, sumarizados na Tabela 4.2.

Para a geração das matrizes de dissimilaridade, sobre determinados conjuntos de dados, recomenda-se utilizar a função “*vegdist*”, disponível no pacote “*vegan*” de ecologia de comunidades (OKSANEN, 2013)¹⁴ do programa *R*. Esta função espera receber, ao menos, dois parâmetros de

¹⁴ O pacote *vegan* disponibiliza ferramentas para estudos em ecologia de comunidades, como análise de (dis)similaridade; tais ferramentas podem ser aplicadas para outros tipos de dados, em diversas áreas do conhecimento.

entrada: (i) a matriz de interação – contendo, por exemplo, a frequência de distribuição de grupos filogenéticos de *E.coli* em reservatórios de água; e (ii) o método matemático usado para o cálculo da matriz de dissimilaridade¹⁵ – como *Bray-Curtis* (“*bray*”), *Jaccard* (“*jaccard*”), *Euclidiano* (“*euclidean*”), entre outros. Uma vez que as matrizes de dissimilaridade estejam criadas, procede-se à execução do teste de *Mantel*, por meio da função “*mantel*”, também disponível no pacote “*vegan*” (OKSANEN, 2013) do programa *R*. Esta função espera receber quatro parâmetros de entrada: (i) primeira matriz de dissimilaridade – M1; (ii) segunda matriz de dissimilaridade – M2; (iii) método escolhido para realizar a correlação entre as duas matrizes – por exemplo, “*pearson*”; e (iv) número de permutações (randomizações) realizadas pela função, para validar a significância do teste – recomenda-se utilizar 999.

Como resultado do teste de *Mantel* são gerados valores para as variáveis “*r*” (valor do coeficiente de correlação) e “*p*” (significância do teste de *Mantel*); neste caso, como o coeficiente de correlação (“*r*”) varia de -1 à 1, um valor igual à zero indica que nenhuma correlação foi encontrada e, valores extremos (-1 ou 1) indicam forte correlação entre as matrizes; já com relação à variável “*p*”, um valor significativo para o teste de *Mantel* consiste em “*p*” < 0.05. É importante registrar que a análise de correlação por meio do teste de *Mantel* é usualmente recomendada para análise de dados de abundância e densidade.

Já a proposta de Hurley (2004) – *Clustering Visualizations of Multidimensional Data* – parte de uma matriz de similaridade dos dados e realiza a aplicação de algum índice de clusterização hierárquica para reordenar as variáveis da matriz, com o objetivo de identificar *clusters* comuns no conjunto de dados; por conseguinte, se utiliza do recurso gráfico *scatterplot* para a apresentação dos resultados (os *clusters* são identificados em função do valor de similaridade entre os elementos componentes da análise). Cada um dos passos necessários para a aplicação deste método está descrito a seguir e, também, sumarizado na Tabela 4.2.

¹⁵ Embora o pacote “*vegan*” do programa *R* trabalhe apenas com matrizes de dissimilaridade (que identifica as diferenças entre conjuntos de dados dispostos em uma matriz bipartida), existem outros programas que podem apresentar testes de similaridade (que identificam as semelhanças entre conjuntos de dados, ao invés das diferenças). Desta forma, é importante esclarecer que uma matriz de similaridade representa o complemento de uma matriz de dissimilaridade (1- matriz de dissimilaridade). Detalhes adicionais sobre os métodos matemáticos disponíveis para o cálculo de uma matriz de dis(similaridade) podem ser consultados em Legendre e Legendre (2012).

No processo de geração da matriz de similaridade (que é o complemento da matriz de dissimilaridade), recomenda-se o uso da seguinte fórmula de apoio: $(1 - \text{vegdist}(\text{matrix}, "Bray"))$; neste exemplo, o método matemático escolhido é o “Bray-Curtis” mas, caso seja outro, basta realizar a substituição do mesmo, como já descrito anteriormente (no teste de *Mantel*). Na sequência, para a reordenação das variáveis, a partir de algum índice de clusterização hierárquica, pode-se utilizar o pacote “gclus” (HURLEY, 2012) do programa *R* e, então, proceder à escolha da função que mais se adeque à análise (tais como, “order.clusters”, “reorder.hclust” ou qualquer outra função que se encontra disponível neste pacote). Por fim, para a apresentação do resultado desta análise, recomenda-se o uso da função “plotcolors”, também disponível no pacote “gclus” (HURLEY, 2012) do programa *R*, para construir a matriz de dispersão (*scatterplot*), cujas cores dos *clusters* representam uma maior similaridade no valor do índice calculado para os elementos daquele agrupamento; além disso, os *clusters* mais próximos da diagonal principal da matriz compreendem as variáveis altamente correlacionadas.

Tabela 4.2 - Resumo de recursos utilizados para aplicar Análise Multivariada de Dados

Recurso	Como aplicar o recurso?
Mantel test (MANTEL, 1967)	<p><u>1º passo:</u> Função “vegdist” do pacote “vegan” (OKSANEN, 2013) do programa <i>R</i>: para a geração das matrizes de dissimilaridade; a escolha do método matemático (parâmetro “m”) para cálculo da matriz de dissimilaridade deve levar em consideração a natureza dos dados – como exemplo, <i>Bray-Curtis</i> (“bray”), <i>Jaccard</i> (“jaccard”), <i>Euclidiano</i> (“euclidean”), entre outros (LEGENDRE; LEGENDRE, 2012).</p> <p><u>2º passo:</u> Função “mantel” do pacote “vegan” (OKSANEN, 2013) do programa <i>R</i>: para a execução do teste de <i>Mantel</i>, tendo como resultados os valores de “r” (-1 a 1) e “p” (significância “p” < 0.05).</p>
Clustering Visualizations of Multidimensional Data (HURLEY, 2004)	<p><u>1º passo:</u> Função “vegdist” do pacote “vegan” (OKSANEN, 2013) do programa <i>R</i>: para a geração da matriz de similaridade – que é o complemento da matriz de dissimilaridade $(1 - \text{vegdist}(\text{matrix}, "Bray"))$;</p> <p><u>2º passo:</u> Pacote “gclus” (HURLEY, 2012) do programa <i>R</i>: escolher uma função auxiliar (como “order.clusters” ou “reorder.hclust”) para realizar a reordenação das variáveis da matriz, a partir de algum índice de clusterização hierárquica;</p> <p><u>3º passo:</u> Função “plotcolors”, também disponível no pacote “gclus” (HURLEY, 2012) do programa <i>R</i>: para construir a matriz de dispersão (<i>scatterplot</i>) com os componentes da análise, agrupados em <i>clusters</i>, de acordo com a similaridade entre eles; os <i>clusters</i> mais próximos da diagonal principal da matriz compreendem as variáveis altamente correlacionadas.</p>

Fonte: autor

É importante registrar que outras ferramentas também podem ser usadas para apoiar esta atividade de escolha das variáveis de análise, tais como: (i) técnicas de mineração de dados (*Data Mining*) – para identificação de padrões sobre os dados disponíveis; (ii) análise de correspondência (*Correspondence Analysis*) – para identificar correlação entre variáveis; e (iii) outras que se tiver disponível, no momento da análise.

Assim que forem realizados todos os procedimentos necessários à aplicação das ferramentas estatísticas de apoio (ou ainda outro recurso que se tenha disponível), cabe ao pesquisador observar os resultados gerados e eleger as variáveis a serem utilizadas na análise. Aqui vale um parêntese – em alguns casos, o especialista em biodiversidade opta pela escolha intencional das variáveis (independente do uso de algum recurso computacional e/ou estatístico de apoio), por conhecer em profundidade a natureza do problema e já ter plena certeza das variáveis a serem utilizadas na análise.

Ao término desta etapa, o passo metodológico seguinte compreende a escolha das métricas de SNA que mais se adéquam aos tipos de dados disponíveis e à(s) pergunta(s) de pesquisa (Etapa 3 da metodologia).

4.3 Etapa 3: Escolha das métricas de SNA adequadas ao contexto da pesquisa

Esta terceira etapa da metodologia (detalhada na Figura 4.6) tem por objetivo auxiliar o pesquisador a responder a seguinte questão: “Dentre o conjunto de recursos disponíveis no domínio de SNA, qual métrica escolher para apoiar as análises em biodiversidade?”.

Esta atividade de escolha das métricas mais adequadas ao contexto emerge como um desafio ao pesquisador, por ser primordial para o resultado da pesquisa; por isso, a esta etapa da metodologia, deve ser dedicada uma atenção especial. Assim sendo, a primeira decisão a ser tomada é sobre qual aspecto da Rede de Interação se deseja analisar: a estrutura da rede como um todo (*network level*) ou o papel de cada espécie/nó na rede (*species/node level*)?

Para analisar as interações no nível da rede, isto é, aspectos gerais da estrutura da rede, utiliza-se como recurso de apoio o pacote

“*bipartite*” (DORMANN *et al.*, 2009) do programa *R* (*The R Foundation for Statistical Computing*); dentre as métricas disponíveis, pode-se mencionar: conectância (*connectance*), aninhamento (*nestedness*), sobreposição de nicho (*niche overlap*), entre outras¹⁶. Cada um dessas métricas encontra-se descrita no Capítulo de Revisão da Literatura (seção 2.3.2.1 Métricas para análise no nível da rede – *network level*) e, além disso, um resumo de suas principais características está disposto na Tabela 4.3. É importante registrar que a opção pelo uso dessas métricas no nível da rede é sugerida quando se deseja comparar diversas Redes de Interação, avaliando como cada rede está estruturada.

Tabela 4.3 - Resumo de métricas para analisar as interações no nível da rede

Métrica	Fórmula	Como calcular?
Connectance (C)	$C = L / (i*j)$ – disponível em Dormann <i>et al.</i> (2009).	Função “ <i>networklevel</i> ” do pacote “ <i>bipartite</i> ” do programa <i>R</i> (DORMANN <i>et al.</i> , 2009), com o parâmetro <i>index</i> ="ALL" – para calcular todos os índices disponíveis no nível da rede, inclusive conectância.
Nestedness (N)	Calculada por meio das métricas <i>T</i> ($N = 1 - T$) (ATMAR; PATTERSON, 1993) ou <i>NODF</i> (disponível em ALMEIDA-NETO <i>et al.</i> , 2008).	Programa ANINHADO (GUIMARAES; GUIMARAES, 2006): permite calcular as métricas <i>T</i> e <i>NODF</i> ; ou Programa NTC - <i>Nestedness Temperature Calculator</i> (ATMAR; PATTERSON, 1995): permite calcular a métrica <i>T</i> ; ou Função “ <i>nestednodf</i> ” do pacote “ <i>vegan</i> ” do programa <i>R</i> (OKSANEN, 2013): permite calcular a métrica <i>NODF</i> para matrizes binárias e ponderadas.
Niche Overlap (NO)	Calculada por meio das métricas <i>E</i> e C_{ws} (ARAÚJO <i>et al.</i> , 2008).	Programa <i>Dieta</i> (ARAÚJO <i>et al.</i> , 2008): geradas automaticamente a partir da execução dos passos do programa <i>Dieta</i> e disponibilizadas em um de seus arquivos de saída – ‘Indices[filename].txt’.

Fonte: autor

Já as métricas no nível da espécie permitem um olhar mais detalhado sobre algum elemento específico da rede. Dentre as métricas comumente utilizadas, estão: dependência (*dependence*), força de interação

¹⁶ Todas as métricas (índices) disponíveis no pacote “*bipartite*” do programa *R* estão descritas em Dormann *et al.* (2009).

(*strength*), *w-clique (cohesive subgroups)* e as métricas de centralidade (*degree*, *betweenness centrality*, *closeness centrality* e *eigenvector centrality*).

Cada uma dessas métricas encontra-se descrita no Capítulo de Revisão da Literatura (seção 2.3.2.2 Métricas para análise no nível da espécie ou do nó – *species/node level*) e, além disso, as fórmulas e ferramentas utilizadas para o cálculo de cada uma delas estão dispostas na Tabela 4.4. Mesmo assim, é importante registrar comentários adicionais acerca da especificidade de algumas métricas em particular, como disposto a seguir.

A métrica *dependence* permite analisar o quanto cada espécie depende da outra com que se relaciona (exemplo: dentre as interações que um visitante floral realiza na rede, qual planta ele tem maior dependência – maior frequência de interação). Além disso, a métrica *strength* representa o efeito de um nó na rede como um todo (WOOTTON; EMMERSON, 2005) e, por isso, é muito usada para analisar as espécies com maior força de interação na estrutura da rede.

Com relação às métricas de centralidade usadas de forma recorrente, tem-se *degree* e *betweenness centrality*. Enquanto a métrica *degree* permite identificar as espécies que interagem com o maior número de elementos da rede (exemplo: as plantas mais visitadas), a *betweenness centrality* permite avaliar quão vital é um vértice para a rede, ou seja, aqueles vértices que são centrais na estrutura da rede, desempenhando a função de *bridge* (HANSEN; SHNEIDERMAN; SMITH, 2011). Assim sendo, tanto a métrica *strength*, quando a *betweenness centrality* são relevantes para o estudo de espécies invasoras ou em processo de extinção.

Por fim, vale à pena chamar a atenção para a métrica *w-clique* (ARAÚJO *et al.*, 2008) – um recurso importante para a descoberta de padrões de agrupamento em redes ponderadas (dados de frequência/abundância de interação). Apesar do processo de identificação dos *w-cliques* (forma de cálculo) estar sumarizado na Tabela 4.4, em função da particularidade deste processo, vale à pena realizar um detalhamento dos passos a serem seguidos, a partir das recomendações de Mello (2012):

- (i) submeter a rede bipartida ponderada ao programa *Dieta* (ARAÚJO *et al.*, 2008) para a geração da matriz binária de conexões fortes da rede (cujas células contêm valor 0/1

indicando ausência/presença de conexões fortes) – disponível no arquivo de saída 'B[filename].mat';

- (ii) submeter a matriz de conexões fortes ao programa *Pajek* (BATAGELJ; MRVAR, 1998) para transformá-la em uma rede bipartida não-direcionada – aplicando a transformação de arcos em arestas e salvando a rede com a extensão '.net'; e
- (iii) submeter a rede gerada no *Pajek* ao programa *Ucinet* (BORGATTI; EVERETT; FREEMAN, 2002) para a identificação dos *w-cliques* e a geração do diagrama de *clusters* (dendograma) com os *cliques*.

Todos esses programas estão baseados na teoria de Redes Complexas para identificação de padrões de agrupamento (*clusters*). Mello (2012) complementa que apesar de o programa *Ucinet* implementar o conceito clássico de *clique* (no momento da identificação dos subgrupos coesos), o resultado deste processamento corresponde aos *w-cliques* pois a matriz em questão já foi submetida ao programa Dieta (que deixou no conjunto de dados apenas as conexões fortes).

Por fim, é importante registrar que, dependendo da natureza da pesquisa, pode-se optar num primeiro momento pela utilização de algumas métricas no nível da rede e, num segundo momento, outras métricas no nível da espécie (para analisar algum componente da rede, que desempenha um papel importante para o contexto da análise).

Desta forma, estando com as métricas de SNA calculadas (no nível da rede e/ou da espécie) encaminha-se à última etapa desta metodologia (Etapa 4), como descrito a seguir.

Tabela 4.4 - Resumo de métricas para analisar as interações no nível da espécie

Métrica	Fórmula	Como calcular?
Dependence (d_{ij})	$d_{ij} = N_{ij}/N_i$ – disponível em Jordano, Vázquez e Bascompte (2009).	Função “ <i>specieslevel</i> ” do pacote “ <i>bipartite</i> ” do programa R (DORMANN <i>et al.</i> , 2009), com o parâmetro <i>index</i> ="ALL" – para calcular todos os índices disponíveis no nível da espécie.
Strength (s_i)	$s_i = \sum d_{ij}$ – disponível em Wootton e Emmerson (2005) e em Vázquez, Morris e Jordano (2005).	
Degree (k_i)	<p>k_i = número total de conexões estabelecidas com o vértice i (FREEMAN, 1978/79).</p> <p><u>Obs.:</u> Para grafos direcionados, pode ser desdobrado em:</p> <p>(i) <i>in-degree</i> – número de conexões que chegam até o vértice; e</p> <p>(ii) <i>out-degree</i> – número de conexões que partem de um vértice até os adjacentes (HANSEN; SHNEIDERMAN; SMITH, 2011).</p>	<p>Função “<i>degree</i>” do pacote “<i>sna</i>” do programa R (BUTTS, 2008).</p> <p><u>Obs.:</u> O parâmetro “<i>cmode</i>” desta função indica o tipo de grau de centralidade que está sendo calculado:</p> <p>(i) “<i>indegree</i>” – para grau de entrada do vértice;</p> <p>(ii) “<i>outdegree</i>” – para grau de saída do vértice; e</p> <p>(iii) “<i>freeman</i>” – para o grau total do vértice (grau de entrada + grau de saída).</p>
Betweenness Centrality	Cálculo baseado nas distâncias geodésicas (o menor caminho conectando um par de vértices) e está descrito em Freeman (1978/79).	Função “ <i>betweenness</i> ” do pacote “ <i>sna</i> ” do programa R (BUTTS, 2008); ou programa <i>Pajek</i> (BATAGELJ; MRVAR, 1998) – permite o cálculo da métrica e sua representação na forma de grafo.
Closeness Centrality	Cálculo baseado na proximidade entre os vértices na estrutura da rede e está descrito em Freeman (1978/79).	Função “ <i>closeness</i> ” do pacote “ <i>sna</i> ” do programa R (BUTTS, 2008); ou programa <i>Pajek</i> (BATAGELJ; MRVAR, 1998) – permite o cálculo da métrica e sua representação na forma de grafo.
Eigenvector Centrality	O cálculo para identificar os vértices que são considerados os núcleos centrais da rede (<i>coreness</i>) está descrito em Butts (2008).	Função “ <i>evcent</i> ” do pacote “ <i>sna</i> ” do programa R (BUTTS, 2008).
W-Clique	O cálculo para identificar os subgrupos coesos na rede (<i>cliques</i>) leva em consideração a abundância da interação e está disponível em Araújo <i>et al.</i> (2008).	<p><u>1º passo:</u> Programa <i>Dieta</i> (ARAÚJO <i>et al.</i>, 2008): arquivo de saída ‘B[filename].mat’, que representa a matriz binária de conexões fortes;</p> <p><u>2º passo:</u> Programa <i>Pajek</i> (BATAGELJ; MRVAR, 1998): para transformar a matriz de conexões fortes em uma rede não-direcionada; e</p> <p><u>3º passo:</u> Programa <i>Ucinet</i> (BORGATTI; EVERETT; FREEMAN, 2002): para identificar os <i>w-cliques</i> e gerar o dendograma.</p>

Fonte: autor

4.4 Etapa 4: Realização de análises biológicas com o apoio de SNA

A última etapa desta metodologia tem por objetivo realizar as análises biológicas, com base nos resultados gerados pelas etapas anteriores – cada uma das atividades componentes desta etapa está representada na Figura 4.7.

Desta forma, o primeiro passo nesta etapa consiste na atividade de obter os resultados das métricas de SNA calculadas e, na sequência, usar o recurso de representação das interações na forma de grafo (gráficos das Redes de Interação) – o que facilita a interpretação de aspectos particulares da estrutura da rede e o diálogo entre os especialistas participantes da pesquisa. Como ferramenta computacional de apoio para a plotagem dos grafos da rede recomenda-se o uso do programa *Pajek*, desenvolvido por Batagelj e Mrvar (1998).

A partir de então, torna-se necessário realizar as análises biológicas para identificar os novos conhecimentos gerados, bem como os resultados traduzidos nas métricas e nos grafos. Por exemplo, um elemento com maior força na rede pode representar uma espécie mais generalista e predominante no domínio da análise; portanto, caso seja uma espécie invasora ao ecossistema, um olhar especial deve ser dado a este elemento da rede.

Percorridas todas as etapas da metodologia, caso seja necessário, o pesquisador pode voltar a alguma etapa anterior (como indicado na Figura 4.1) e, assim, realizar um maior refinamento da análise.

Face ao exposto neste Capítulo 4, dispõe-se de um instrumento metodológico para apoiar um pesquisador em biodiversidade que tenha interesse em realizar análise de dados de Redes de Interação biológicas, com o aporte dos recursos de SNA e de Análise Estatística. A partir de agora, torna-se necessário avaliar a adequação desta proposta, com estudos de caso emanados de problemas reais de pesquisa – como descrito no Capítulo a seguir.

Figura 4.6 - Etapa3 da metodologia: Escolha das métricas de SNA adequadas (modelada em BPMN)

Fonte: autor

Figura 4.7 - Etapa4 da metodologia: Realização de análises biológicas com SNA (modelada em BPMN)

Fonte: autor

5 DISCUSSÃO DA METODOLOGIA PROPOSTA A PARTIR DE ESTUDOS DE CASO COM REDES DE INTERAÇÃO MICROBIOLÓGICAS

Este capítulo se dedica a apresentar três estudos de caso (Tabela 5.1), à luz da metodologia descrita no Capítulo 4, com o objetivo de discutir e avaliar a adequação deste instrumento a estes estudos, de modo a identificar lacunas ou pontos de melhoria.

Nesse sentido, é importante esclarecer que o foco, aqui, não será detalhar os estudos de caso pois eles estão apresentados nos Apêndices (A, B e C) desta tese, mas sim, os caminhos percorridos na metodologia em cada trabalho.

Tabela 5.1 - Lista dos Estudos de Caso desenvolvidos

Referência	Título do artigo	Autores	Instituições envolvidas
Stoppe et al. (2014a)	<i>Clustering of water bodies in unpolluted and polluted environments based on Escherichia coli phylogroup abundance using a simple interaction database</i>	STOPPE, N.C.; SILVA, J.S. ; TORRES, T.T.; CARLOS, C.; HACHICH, E.M.; SATO, M.I.Z.; SARAIVA, A.M.; OTTOBONI, L.M.M	Biocomp-USP CETESB IFMT UNICAMP
Silva et al. (2014)	<i>Social Network Analysis Metrics and Their Application in Microbiological Network Studies</i>	SILVA, J.S. ; STOPPE, N.C.; TORRES, T.T.; OTTOBONI, L.M.M.; SARAIVA, A.M.	Biocomp-USP IFMT UNICAMP
Stoppe et al. (2014b) – in preparation	<i>Complex pattern of Escherichia coli phylogenetic group distribution in worldwide samples of human feces and wastewater treatment plants</i>	STOPPE, N.C.*; SILVA, J.S.* ; ROSENBERG, R.J.; CARLOS, C.; HACHICH, E.M.; SATO, M.I.Z.; SARAIVA, A.M.; OTTOBONI, L.M.M.; TORRES, T.T. [<i>*equal contribution</i>]	Biocomp-USP CETESB CONABIO(México) IFMT UNICAMP

Fonte: autor

Estes estudos de caso emanaram de um processo colaborativo de pesquisadores do BioComp-USP com pesquisadores de outras instituições (CETESB, CONABIO-México, IFMT e UNICAMP), por meio de uma atuação transdisciplinar (Biologia, Ciência da Computação, Engenharia e Estatística) o que trouxe uma base de sustentação sólida para avaliar a adequação desta metodologia a este contexto.

A partir de agora cada um destes trabalhos será revisitado tendo como base os passos orientativos da metodologia proposta no Capítulo 4, com a indicação

das trilhas seguidas – com destaque para os passos percorridos por cada estudo de caso.

5.1 Estudo de Caso 1: Stoppe *et al.* (2014a)

Este trabalho teve por objetivo estabelecer uma classificação de *E.coli* (provenientes de amostras de corpos d'água) baseada em subgrupos filogenéticos e tentar associá-los com fontes de poluição.

Conforme proposto na metodologia (Capítulo 4), a primeira etapa deste trabalho compreendeu as atividades de mapeamento dos tipos de dados e de interação. Nesse sentido, ao se conhecer o banco de dados disponível (planilha do Excel), observou-se que ele estava organizado em uma matriz de interações microbiológicas, cujas linhas correspondiam a corpos d'água (*water bodies* – amostras de água) e, colunas, a subgrupos filogenéticos de *E.coli* (*phylogenetic subgroups*). Como cada célula da matriz continha um número inteiro positivo (representando a frequência de ocorrência do subgrupo filogenético no corpo d'água correspondente), a matriz de interação era do tipo ponderada (abundância de interação).

Na sequência, como disposto na Figura 5.1 (indicações em vermelho), os dados foram normalizados, para evitar um viés na amostra (conforme orientado por LEGENDRE; LEGENDRE, 2012); e, então, a matriz de interação foi preparada para ser utilizada no programa *Dieta*, copiando os dados em um simples arquivo de texto (txt) e levando em consideração as seguintes recomendações de Mello (2012): (i) os *labels* das linhas e colunas devem ser suprimidos; (ii) o separador usado, entre os dados, deve ser o espaço em branco; e (iii) o separador decimal deve ser o ponto (.) e, não, a vírgula (,).

Como também se tinha disponível um conjunto de outras variáveis correlacionadas a estes corpos d'água – como *average WQI* (*Water Quality Index*), *average TSI* (*Trophic State Index*), *main land use* (*urban/rural*) – algumas delas foram usadas para preparar matrizes de apoio; neste caso (como indicado pelas setas azuis na Figura 5.1), os dados não foram normalizados, pois correspondiam a valores categóricos (exemplos de valores de *WQI*: 1 – *very bad*; 2 – *bad*; 3 – *acceptable*; 4 – *good*; e 5 – *very good*). Além disso, como estas matrizes seriam usadas no programa *R*, o separador usado entre os dados foi a tabulação e os *labels* das linhas e colunas foram mantidos no arquivo texto (txt).

O produto final desta etapa compreendeu três matrizes de interação: (i) corpos d'água x subgrupos filogenéticos – para uso no programa *Dieta*; (ii) corpos d'água x *WQI* – para uso no programa *R*; e (iii) corpos d'água x *TSI* – para uso no programa *R*.

Figura 5.1 - Etapa1 da metodologia: trilha do Estudo de Caso 1

Fonte: autor

O passo seguinte, na condução deste trabalho, consistiu nas atividades de definição das perguntas-chave a serem respondidas e, também, de escolha das variáveis de análise (Figura 5.2). Como pergunta-chave de pesquisa, dispunha-se da seguinte: “É possível estabelecer uma classificação de amostras de *E.coli* (provenientes de corpos d'água) baseada em subgrupos filogenéticos e tentar associá-los com fontes de poluição?”.

Figura 5.2 – Etapa2 da metodologia: trilha do Estudo de Caso 1

Fonte: autor

Tendo como base esta pergunta de pesquisa, teve início o processo de escolha das variáveis participantes da análise, trilhando os passos indicados na Figura 5.3 (sinalizações em vermelho). Para apoiar esta atividade optou-se pelo uso de um recurso proveniente do domínio da Estatística denominado Análise Multivariada de Dados; nesse sentido, foi aplicado o método *Mantel test*, com o objetivo de analisar a correlação entre as seguintes variáveis: corpos d'água e os índices de qualidade de água (*WQI* e *TSI*). Uma vez que foi confirmada a correlação

entre eles, procedeu-se ao uso da distribuição dos subgrupos filogenéticos, em substituto aos índices *WQI* e *TSI*, na perspectiva de se utilizar deles como uma ferramenta complementar para análise da qualidade da água. Por conseguinte, as variáveis de análise eleitas foram os corpos d'água (*water bodies*) e a distribuição dos subgrupos filogenéticos (*phylogenetic subgroups distribution*), finalizando, assim, esta etapa da pesquisa.

Figura 5.3 - Etapa2 da metodologia (subprocesso – escolher variáveis de análise): trilha do Estudo de Caso 1

Fonte: autor

O passo seguinte na condução deste trabalho (Etapa 3 da metodologia), consistiu na atividade de escolha das métricas de SNA adequadas ao contexto da pesquisa. A primeira atividade consistiu em decidir o aspecto da rede que se desejava analisar; como demonstrado na Figura 5.4 (sinalizações em vermelho), a trilha seguida foi “analisar interações no nível da espécie (*species level*)”, pois havia a necessidade de se identificar os seguintes aspectos da rede: (i) padrões de agrupamento por meio do cálculo da métrica *w-clique* – programas *Dieta* (ARAÚJO *et al.*, 2008), *Pajek* (BATAGELJ; MRVAR, 1998) e *Ucinet* (BORGATTI; EVERETT; FREEMAN, 2002); e (ii) padrão de clusterização hierárquica por meio do método clássico de análise denominado UPGMA (*Unweight Pair Group Method with Arithmetic mean* – SOKAL; MICHENER, 1958).

Aqui vale um parêntese – a opção pela escolha da métrica *w-clique* decorreu da necessidade que se tinha, neste trabalho, de identificar algum padrão de agrupamento sobre os dados, a partir de uma simples Rede de Interação de micro-organismos (contendo a frequência de distribuição dos subgrupos filogenéticos); nesse sentido, os conhecimentos gerados por meio da aplicação desta métrica de SNA são descritos a seguir.

Figura 5.4 – Etapa3 da metodologia: trilha do Estudo de Caso 1

Fonte: autor

Como última etapa da pesquisa, conforme sugerido pela metodologia, procedeu-se à realização de análises biológicas com o apoio de SNA. De tal modo, foram percorridas todas as atividades propostas (Figura 5.5) e chegou-se aos seguintes resultados:

- (i) resposta biológica – a métrica *w-clique* pode ser usada como um potencial classificador de poluição de corpos d’água e, assim, ser utilizada como uma ferramenta complementar para avaliação de grau de poluição da água; e
- (ii) novo conhecimento – a partir de um simples banco de dados de interação foi possível descobrir novos padrões de agrupamento (*clusters*) para os corpos d’água, o que não ocorreu com os métodos tradicionais (como UPGMA).

Figura 5.5 – Etapa4 da metodologia: trilha do Estudo de Caso 1

Fonte: autor

Com base nas análises biológicas realizadas – as quais foram enriquecedoras ao processo de análise – não houve a necessidade de voltar a alguma etapa anterior (como disposto na Figura 5.6 – sinalizações em vermelho);

assim sendo, chegou-se ao término da aplicação de todos os passos da metodologia proposta a este trabalho.

Figura 5.6 - Metodologia para aplicação de métricas de SNA em biodiversidade: trilha do Estudo de Caso 1

Fonte: autor

Face ao exposto, nota-se que houve adequação da metodologia às necessidades de pesquisa do Estudo de Caso 1 (STOPPE *et al.*, 2014a). A seguir, outro trabalho é apresentado, com a finalidade dar continuidade ao processo de avaliação deste instrumento metodológico.

5.2 Estudo de Caso 2: Silva *et al.* (2014)

Este trabalho se propôs a aplicar os conceitos e métricas de SNA em Redes de Interação microbiológicas para identificar padrões de agrupamento, além de descobrir novos conhecimentos sobre a estrutura dos subgrupos filogenéticos.

A primeira etapa deste trabalho, assim como no Estudo de Caso anterior, consistiu nas atividades de mapeamento dos tipos de dados e de interação (Etapa 1 da metodologia proposta). Apesar do banco de dados deste estudo corresponder ao mesmo do anterior (Estudo de Caso 1), a trilha seguida nesta etapa foi uma versão mais simplificada (conforme disposto na Figura 5.7), pois apenas se fez necessário preparar as matrizes ponderadas para um mesmo conjunto de dados – os quais foram normalizados e, na sequência, organizados para serem usados nos programas *Dieta* e *R*.

Figura 5.7 - Etapa1 da metodologia: trilha do Estudo de Caso 2

Fonte: autor

A segunda etapa, na condução desta pesquisa, consistiu nas atividades de definição das perguntas-chave a serem respondidas e, também, de escolha das variáveis de análise (Figura 5.2, mesma trilha do Estudo de Caso 1). Em princípio foi definida a seguinte pergunta de pesquisa: “É possível aplicar os conceitos e métricas de SNA em Redes de Interação microbiológicas para identificar padrões de agrupamento, além de descobrir novos conhecimentos sobre a estrutura dos subgrupos filogenéticos?”.

Como este foi um dos primeiros trabalhos em que foram aplicadas as métricas de SNA no domínio microbiológico, optou-se pela escolha intencional das variáveis, a partir do conhecimento do especialista em biodiversidade (como indicado na Figura 5.8 – trilha em vermelho); neste caso, as variáveis de análise eleitas foram os corpos d’água (*water bodies*) e a distribuição dos subgrupos filogenéticos (*phylogenetic subgroups distribution*) nestes respectivos locais.

Figura 5.8 - Etapa2 da metodologia (subprocesso – escolher variáveis de análise): trilha do Estudo de Caso 2

Fonte: autor

O passo seguinte na condução deste trabalho (Etapa 3 da metodologia) consistiu na atividade de escolha das métricas de SNA adequadas ao contexto da pesquisa. Como primeira atividade procedeu-se à escolha do aspecto da Rede de Interação que se desejava analisar; como demonstrado na Figura 5.9, a opção seguida foi realizar a análise no nível da espécie (*species level*), em função da necessidade de se usar um conjunto de métricas que permitissem identificar características de cada espécie/nó da rede.

Dentre as ferramentas de apoio disponíveis para o cálculo das métricas de SNA optou-se pelo uso daquelas que permitiam analisar os seguintes aspectos: (i) papel de um nó na rede – pacotes “*sna*” (BUTTS, 2008) e “*bipartite*” (DORMANN *et al.*, 2009) do programa *R*; e (ii) padrões de agrupamento – programas *Dieta* (ARAÚJO *et al.*, 2008), *Pajek* (BATAGELJ; MRVAR, 1998) e *Ucinet* (BORGATTI; EVERETT; FREEMAN, 2002).

Figura 5.9 – Etapa3 da metodologia: trilha do Estudo de Caso 2

Fonte: autor

Como última etapa da pesquisa, conforme sugerido pela metodologia (Etapa 4), procedeu-se à realização de análises biológicas com o apoio de SNA. De tal modo, foram percorridas todas as atividades propostas (Figura 5.6 – mesma trilha do Estudo de Caso 1) e os resultados encontrados foram os seguintes:

- (i) respostas biológicas – a métrica *betweenness centrality* permitiu confirmar que subgrupos filogenéticos têm um papel central na Rede de Interação; a partir das métricas *dependence* e *interaction strength* foi possível identificar os elementos mais generalistas na rede (subgrupos filogenéticos ocorrendo na maioria dos corpos d’água) e os predominantes (quais subgrupos filogenéticos são

importantes para certos corpos d'água), permitindo realizar análise da tolerância da Rede de Interação em relação à extinção/inclusão de um elemento na rede; e

- (ii) novos conhecimentos gerados – as métricas de SNA podem ser usadas como uma nova ferramenta para visualização de associação entre subgrupos filogenéticos e corpos d'água e, ao mesmo tempo, para ajudar a identificação de elementos com função de ponte (*bridge*) na Rede de Interação.

Tendo concluída a última etapa da pesquisa, chega-se à finalização da execução de todos os processos propostos pela metodologia (conforme disposto na Figura 5.6) e obtém-se como resultado diversos elementos que podem apoiar um tomador de decisão na área de biodiversidade – confirmando, mais uma vez, a adequabilidade da metodologia proposta a este contexto. Na sequência é apresentado um terceiro Estudo de Caso, também com a finalidade de continuar o processo de avaliação deste instrumento metodológico, a partir de problemas reais de pesquisa.

5.3 Estudo de Caso 3: Stoppe *et al.* (2014b)

O objetivo deste trabalho foi avaliar a distribuição filogenética de amostras de *E.coli* de humanos e de esgotos (*wastewater*); além disso, também tinha por finalidade realizar meta-análises na distribuição de grupos filogenéticos de humanos (provenientes de amostras de diversos países), procurando por padrões de associação entre os grupos filogenéticos e a localização geográfica (ou clima/hábitos alimentares/local de habitação) que pudessem indicar adaptação local.

Como primeiro passo nesta pesquisa (como indicado na Figura 5.10), foi realizada a atividade de conhecer os dados da Rede de Interação, que correspondiam à frequência de distribuição de grupos filogenéticos de *E.coli* (A, B1, B2 e D) isoladas em fezes humanas e de esgotos (provenientes do mapeamento de estudos feitos em diversas partes do mundo). Na sequência, para os dados de grupos filogenéticos (*phylogenetic group*) e clima (*Koppen climate classification*) foi preparada uma matriz binária, indicando a presença/ausência do grupo filogenético, em cada tipo de clima constante no banco de dados. Além disso, foram preparadas três matrizes ponderadas, a partir da normalização dos dados, contendo: (i) grupos

filogenéticos x hábitos alimentares (*feeding habits – western or other*); (ii) grupos filogenéticos x local de habitação (*living area – urban or rural*); e (iii) grupos filogenéticos x localização geográfica (*geographic coordinates*).

Este conjunto de matrizes (binárias e ponderadas) foi preparado no formato adequado tanto para o uso no programa *R*, quanto no programa *Dieta*, seguindo as mesmas recomendações já descritas nos Estudos de Caso 1 e 2.

Figura 5.10 - Etapa1 da metodologia: trilha do Estudo de Caso 3

Fonte: autor

A segunda etapa desta pesquisa, conforme proposto na metodologia, compreendeu as atividades de definição das perguntas-chave e de escolha das variáveis de análise (Figura 5.2 – mesma trilha do Estudo de Caso 1). Assim sendo, num primeiro momento, o estudo esteve centrado na definição das seguintes perguntas-chave: (i) “É possível identificar a distribuição de grupos filogenéticos entre amostras de *E.coli* levando em consideração a localização geográfica, o clima e os hábitos alimentares?”; e (ii) “A distribuição de grupos filogenéticos de *E.coli* (provenientes de amostras de esgotos) pode ser utilizada em substituto às amostras de fezes humanas?”.

Num segundo momento, conforme disposto na Figura 5.11, a partir das variáveis disponíveis no banco de dados, foi seguida uma trilha particular do subprocesso de escolha das variáveis de análise, que até então não havia sido percorrida pelos demais estudos de caso – a qual será detalhada a seguir.

Figura 5.11 - Etapa2 da metodologia (subprocesso – escolher variáveis de análise): trilha do Estudo de Caso 3

Fonte: autor

Como ferramentas de apoio, a esta atividade de seleção das variáveis, optou-se fortemente pelo uso da abordagem de Análise Estatística (por meio dos recursos de Análise Exploratória de Dados e de Análise Multivariada de Dados), bem como de uma ferramenta complementar (técnica de mineração de dados – *Data Mining*). Isto ocorreu pelo fato de este Estudo de Caso ter a colaboração de um especialista em Análise Estatística, que trouxe diversas contribuições para esta etapa da pesquisa.

Os seguintes recursos provenientes do domínio de EDA foram aplicados neste estudo de caso: (i) *boxplot* – para visualizar a distribuição dos dados; (ii) *scatterplot* – para analisar a correlação entre as variáveis; e (iii) *Poisson* e *Gamma distribution* – para identificar o modelo de distribuição de probabilidade dos dados. Ademais, foram aplicados dois métodos disponíveis no contexto de Análise Multivariada de Dados: (i) teste de *Mantel* (MANTEL, 1967) – para tentar identificar a correlação entre grupos filogenéticos x hábitos alimentares (ou clima, local de habitação e localização geográfica); (ii) *Clustering Visualizations of Multidimensional Data* (HURLEY, 2004) – para tentar identificar a correlação entre as variáveis por meio da aplicação de algum índice de clusterização hierárquica para reordenação destas variáveis (como grupos filogenéticos x clima – ou hábitos alimentares, local de habitação e localização geográfica).

A outra ferramenta de apoio utilizada decorre do domínio da Ciência da Computação e corresponde à técnica de mineração de dados denominada *data mining*, que também pode ser usada quando se necessita identificar padrões sobre

os dados disponíveis, pois permite a aplicação de algoritmos de classificação¹⁷. Neste estudo de caso foi usado o algoritmo de árvore de classificação (*classification tree*), a partir do pacote “*rpart*” (THERNEAU; ATKINSON; RIPLEY, 2014) do programa *R*. No entanto, é importante mencionar que a decisão pela utilização da técnica de mineração de dados deve levar em conta o valor do atributo *misclassification error* sobre os dados de teste (percentual de dados no qual o algoritmo não realiza uma boa classificação); assim, após a execução do algoritmo, este atributo deve retornar o menor valor possível, para que a árvore de classificação resultante tenha um resultado satisfatório para a análise – o percentual adequado depende muito da área de estudo e da natureza do problema.

É importante esclarecer que, neste Estudo de Caso, optou-se pela utilização do recurso de *Data Mining* por uma sugestão de um dos especialistas participantes da análise, como uma tentativa adicional neste processo de identificação de um padrão de classificação sobre os dados disponíveis.

Por fim, ao término desta etapa de experimentação, tendo como aporte um arcabouço de ferramentas de apoio, em especial da área de Análise Estatística, os resultados foram analisados e a seleção das variáveis realizadas, tais como: distribuição dos grupos filogenéticos (*phylogenetic group distribution*), clima (*Koppen climate classification*), hábitos alimentares (*feeding habits*), local de habitação (*living area*) e coordenadas geográficas (*geographic coordinates*).

A próxima etapa, neste Estudo de Caso, compreendeu a atividade de escolha das métricas de SNA adequadas ao contexto da pesquisa (Etapa 3 da metodologia proposta). Nesse sentido, como apresentado na Figura 5.12, a opção seguida foi “analisar interações no nível da espécie (*species level*)”, pois havia a necessidade de se tentar identificar padrões de agrupamento na Rede de Interação por meio do cálculo da métrica *w-clique* – programas *Dieta* (ARAÚJO *et al.*, 2008), *Pajek* (BATAGELJ; MRVAR, 1998) e *Ucinet* (BORGATTI; EVERETT; FREEMAN, 2002).

¹⁷ No programa *R* existem disponíveis diversos pacotes que implementam os algoritmos de mineração de dados, tais como: *randomForest* (*classification with Random Forest*), *klaR* (*Naive Bayes with Kernel*), *e1071* (*naiveBayes* e *SVM – Support Vector Machines*).

Figura 5.12 – Etapa3 da metodologia: trilha do Estudo de Caso 3

Fonte: autor

Na última etapa da pesquisa, como proposto pela metodologia (Etapa 4), foram realizadas as análises biológicas com o apoio de SNA, sendo percorridas todas as trilhas sugeridas (Figura 5.6). Deste modo, ao contrário da hipótese prevista, não foi identificado um padrão de distribuição dos grupos filogenéticos em função da localização geográfica ou de qualquer outra variável de análise. No entanto, a partir do conhecimento dos especialistas em biodiversidade conseguiu-se inferir os seguintes resultados de análise: (i) a não identificação de um padrão de distribuição pode ser por causa da falta de padrão, ou ainda, pelo fato de não haver diferença dos grupos filogenéticos, em relação ao clima, hábitos alimentares, local de habitação e localização geográfica; (ii) os grupos filogenéticos podem não ter uma resolução suficiente para mostrar um padrão de distribuição, em função da localização geográfica; e, por fim, (iii) o aspecto regional reforça o uso de amostras de esgotos como substituto por contaminação fecal humana.

Face ao exposto o que se aprendeu, por meio deste processo de experimentação e interação com pesquisadores, é que muitas vezes, pode-se chegar ao final do estudo e as métricas de SNA não conseguirem explicar algum aspecto dos dados, ou até, não conseguirem identificar um padrão sobre os mesmos. No entanto, um resultado negativo, pode também representar uma resposta biológica relevante para a pesquisa e, por isso, não deve ser descartado no processo de análise.

Por fim, pode-se concluir que os elementos emanados deste processo de avaliação da metodologia proposta corroboram a necessidade de se dispor de

um instrumento metodológico de apoio, que sirva como guia aos pesquisadores que trabalham (ou venham a trabalhar) com Redes de Interação biológicas. Aqui vale um parêntese – embora não foi apresentado aqui a trilha do Estudo de Caso com polinizadores (GIANNINI *et al.*, 2014, em função do trabalho estar em processo de revisão), a metodologia também se mostrou aplicável a este experimento, pois foi possível a utilização das métricas de SNA para apoiar as atividades de Análise de Dados com as Redes de Interação de polinizadores disponíveis.

Chega-se, então, ao término da apresentação de todos os Capítulos que compõem a base de sustentação desta pesquisa (Revisão da Literatura, Materiais e Métodos, Resultados e Discussão). Assim sendo, resta apresentar as Considerações Finais e Conclusões, advindas deste processo de pesquisa, que são descritas a seguir.

6 CONSIDERAÇÕES FINAIS E CONCLUSÕES

Este último Capítulo tem por objetivo apresentar as considerações finais e conclusões desta pesquisa de doutorado, tendo como base os elementos derivados de cada etapa deste trabalho.

6.1 Considerações Finais

Este trabalho teve por objetivo apresentar a proposta de uma metodologia para apoiar os pesquisadores em biodiversidade na aplicação das métricas de Análise de Redes Sociais (*Social Network Analysis* – SNA) em Redes de Interação biológicas, servindo de guia para pesquisas dessa natureza.

Para tanto, a pesquisa foi desenvolvida a partir dos seguintes passos metodológicos: (i) revisão da literatura – para o entendimento do arcabouço conceitual em que se fundamentava esta pesquisa; (ii) experimentos realizados com métricas de SNA – para aprender com a experiência; (iii) formulação da proposta de uma metodologia – contribuição do trabalho; e (iv) validação da metodologia por meio de estudos de caso.

Nesse sentido, por meio de um levantamento bibliográfico foi possível identificar que existem diversas iniciativas para estudos e conservação da biodiversidade, tendo como suporte as Tecnologias da Informação (Informática para a Biodiversidade). Ademais, o que se constatou é que os recursos utilizados no domínio de Análise de Redes de Interação emanam da área de SNA, em especial as métricas.

No entanto, apesar de a comunidade científica utilizar fortemente os conceitos, algoritmos e métricas de SNA como ferramentas de apoio em Análise de Redes de Interação ecológicas (em especial, nos estudos com polinizadores), não foi encontrado um método sistematizado para aplicação de tais recursos no domínio da biodiversidade, nem mesmo a contribuição que outras áreas do conhecimento (como a abordagem de Análise Estatística) podem trazer para pesquisas dessa natureza.

Considerando esta lacuna, foram realizados diversos experimentos (aqui descritos como estudos de caso), emanados de problemas reais de pesquisa, que possibilitaram aprender por meio da experiência os caminhos que podem ser seguidos durante o processo de aplicação das métricas de SNA no domínio da

biodiversidade. Este processo de experimentação levou em consideração a natureza dos dados e as necessidades da pesquisa, por meio dos quais se conseguiu identificar as possibilidades de aplicação desses recursos.

Assim sendo, foi elaborada a proposta de uma metodologia para aplicação das métricas de SNA em Redes de Interação biológicas, a partir de um processo colaborativo com especialistas em diversas áreas do conhecimento (Biologia, Ciência da Computação, Engenharia e Estatística – do BioComp-USP e de outras instituições), que participaram de todas as etapas de sua formulação, trazendo um arcabouço de sustentação a esta proposta. É importante registrar que este processo de colaboração trouxe contribuições para o alcance da missão do BioComp-USP – que é o desenvolvimento de pesquisas abrangentes que promovam o desenvolvimento de propostas para um conjunto amplo de demandas, que possam apoiar a formação de recursos humanos, disseminação do conhecimento e a tomada de decisão em biodiversidade, com vistas à sua preservação e o uso sustentável (BIOCOMP, 2014).

Esta metodologia foi estruturada em quatro etapas e formalizada em BPMN – o que possibilitou a representação gráfica de todos os seus processos (subprocessos), atividades, sequência de fluxos e eventos, facilitando a compreensão dos passos a serem seguidos, em cada uma das etapas propostas. Foram incluídas, também, justificativas e recomendações sobre as possibilidades de trilhas a serem percorridas, em função da realidade de pesquisa.

6.2 Conclusões

A partir dos elementos aqui apresentados, pode-se concluir que a realização de uma pesquisa com Redes de Interação biológicas tendo um método sistematizado de apoio (para guiar os passos a serem seguidos) traz uma base de sustentação ao pesquisador, seja em função da facilidade de organização das etapas da pesquisa (que pode envolver diversos especialistas), ou ainda, pela disponibilização de um conjunto de ferramentas de apoio ao processo de Análise de Dados.

Por meio do processo de construção de cada Estudo de Caso foi possível confirmar, por exemplo, que o uso de diversos recursos de apoio (como aqueles provenientes da Estatística) fortalece o processo de escolha das variáveis de análise. Deste modo, a escolha intencional das variáveis de análise pode não trazer resultados satisfatórios, justificando o uso de ferramentas de apoio a esta

etapa da pesquisa. Além disso, as discussões dos resultados à luz da metodologia permitiram revisitar os estudos de caso para avaliar a adequação da proposta para apoiar as pesquisas com Redes de Interação biológicas.

É importante registrar, também, que o uso da Notação para Modelagem de Processos de Negócio (BPMN) facilitou a comunicação entre os especialistas participantes da pesquisa, uma vez que este recurso permitiu o detalhamento gráfico de todas as atividades componentes de cada etapa da metodologia. Nesse sentido, a sistematização que a metodologia oferece permite que o pesquisador na escolha de uma trilha (caminho a ser percorrido) visualize melhor as ferramentas, recursos e métodos aplicáveis em cada etapa, ao longo do desenvolvimento da pesquisa. Isto é relevante, pois viabiliza a possibilidade de outros pesquisadores replicarem o estudo, se seguirem os mesmos passos e disporem do mesmo conjunto de dados – trazendo uma base de sustentação científica ao trabalho.

Por fim, esta metodologia deixa claro que quando o pesquisador dispõe de dados de interação no formato de uma matriz binária ou ponderada, torna-se possível aplicar as métricas de SNA em seu estudo, levando-se em consideração as etapas metodológicas aqui propostas – em especial quando há a necessidade de se identificar padrões de agrupamento na rede e, também, a correlação entre as variáveis.

Face ao exposto, esta metodologia consiste na contribuição original desta pesquisa de doutorado para a área de Informática para a Biodiversidade e, assim, espera-se que esta proposta sirva de base para novos estudos e o desenvolvimento de soluções computacionais para apoiar as ações em torno da conservação da biodiversidade.

Como trabalhos futuros pretende-se aplicar esta metodologia em domínios ainda não explorados, como na Agrobiodiversidade – para apoiar análises qualitativas de elementos da interação (como localização geográfica, aspectos temporais, produtividade, entre outros), ao invés de simplesmente analisar a ocorrência/abundância da interação. Uma possibilidade seria a utilização desta metodologia em contextos agrícolas para analisar, por exemplo, o papel dos polinizadores em determinadas culturas, a partir da disponibilização de bancos de dados de frequência de interação abelha-planta contendo dados de produtividade, bem como outras variáveis de análise.

Ademais, vislumbra-se como oportunidade de pesquisa a realização de estudos com Redes de Interação biológicas com outras dimensões e com um maior volume de dados (como aqueles provenientes de sensores), que permitam avaliar a escalabilidade das métricas de SNA. Neste sentido, pode-se propor uma evolução da metodologia levando em consideração, além dos elementos aqui propostos, outros processos de Análise de Dados disponíveis na área de Ciência da Computação (como no domínio da Mineração de Dados), pois há uma tendência crescente em se trabalhar com grandes volumes de dados (*Big Data*).

REFERÊNCIAS

- ADIBA, S.; NIZAK, C.; VAN-BAALEN, M.; DENAMUR, E.; DEPAULIS, F. From grazing resistance to pathogenesis: the coincidental evolution of virulence factors. **PLoS ONE**, v. 5, n. 8, article ID e11882, 2010.
- AKSOY, H. Use of gamma distribution in hydrological analysis. **Turk J Engin Environ Sci**, v. 24, 419-428, 2000.
- ALMEIDA-NETO, M.; GUIMARAES, P.R.; LEWINSOHN, T.M. On nestedness analyses: rethinking matrix temperature and anti-nestedness. **Oikos**, v. 116, n. 4, p. 716-722, 2007.
- ALMEIDA-NETO, M.; GUIMARAES, P.; GUIMARAES, P.R.; LOYOLA, R.D.; ULRICH, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. **Oikos**, v. 117, n. 8, p. 1227-1239, 2008.
- ARAÚJO, M.S.; GUIMARÃES, P.R.; SVANBÄCK, R.; PINHEIRO, A.; GUIMARÃES, P.; REIS, S.F.; BOLNICK, D.I. Network analysis reveals contrasting effects of intraspecific competition on individual vs. population diets. **Ecology**, v. 89, n. 7, p. 1981-1993, 2008.
- ATMAR, W.; PATTERSON, B.D. The measure of order & disorder in the distribution of species in fragmented habitat. **Oecologia**, v. 96, n. 3, p. 373-382, 1993.
- ATMAR, W.; PATTERSON, B.D. **The nestedness temperature calculator**: a visual basic program, including 294 presence-absence matrices. [Software]. AICS Research, Inc., University Park, NM and The Field Museum, Chicago, IL: 1995. Disponível em: <<http://aics-research.com/nestedness/tempcalc.html>>. Acesso em: 27 abr. 2014.
- BASCOMPTE, J.; JORDANO, P.; OLESEN, J. M. Asymmetric coevolutionary networks facilitate biodiversity maintenance. **Science**, v. 312, n. 5772, p. 431-433, 2006.
- BASCOMPTE, J.; JORDANO, P. Plant-animal mutualistic networks: the architecture of biodiversity. **Annual Review of Ecology, Evolution, and Systematics**, v. 38, p. 567-593, 2007.
- BASTOLLA, U.; FORTUNA, M.A.; PASCUAL-GARCIA, A.; FERRERA, A.; LUQUE, B.; BASCOMPTE, J. The architecture of mutualistic networks minimizes competition and increases biodiversity. **Nature**, v. 458, p. 1018-1020, 2009.
- BATAGELJ, V.; MRVAR, A. Pajek: program for large network analysis. **Connections**, v. 21, n. 2, p. 47-57, 1998.
- BATAGELJ, V.; MRVAR, A. **Pajek and Pajek – XXL**: programs for analysis and visualization of very large networks – Reference manual. Version 3.15. [Software]. 04 Mar. 2014. Disponível em <<http://mrvar.fdv.uni-lj.si/pajek/pajekman.pdf>>. Acesso em: 04 maio 2014.

BEZERRA, E.L.S.; MACHADO, I.C.S.; MELLO, M.A.R. Pollination networks of oil-flowers: a tiny world within the smallest of all worlds. **Journal of Animal Ecology**, v. 78, n. 5, p. 1096-1101, 2009.

BIOCOMP. Núcleo de Pesquisa em Biodiversidade e Computação. São Paulo. Disponível em: <<http://www.biocomp.org.br/>>. Acesso em: 10 maio 2014.

BONITASOFT. **Bonita BPM 6.0**: documentation. Version 6.0.4. [Software]. Bonita Soft S.A., 2013. Disponível em: <<http://documentation.bonitasoft.com/>>. Acesso em: 07 abr. 2014.

BORGATTI, S.P.; EVERETT, M.G.; FREEMAN, L.C. **Ucinet 6 for Windows**: software for social network analysis. Version 6.199. [Software]. Harvard, MA: Analytic Technologies, 2002.

BRAGA, M.J.C.; GOMES, L.F.A.M.; RUEDIGER, M.A. Mundos pequenos, produção acadêmica e grafos de colaboração: um estudo de caso dos Enanpads. **Revista de Administração Pública (RAP)**, FGV, Rio de Janeiro, v. 42, n.1, p. 133-154, jan./fev. 2008.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer Networks and ISDN Systems**, v. 30, n. 1-7, p. 107-117, 1998.

BUTTS, C.T. Social network analysis with sna. **Journal of Statistical Software**, v. 24, n. 6, p. 1-51, 2008.

CANHOS, V.P.; SOUZA, S.; GIOVANNI, R.; CANHOS, D.A.L. Global biodiversity informatics: setting the scene for a “new world” of ecological modeling. **Biodiversity Informatics**, v.1, p. 1-13, 2004.

CARLOS, C.; AIRES, M.M.; STOPPE, N.C.; HACHICH, E.M.; SATO, M.I.Z.; GOMES, T.A.T.; AMARAL, L.A.; OTTOBONI, L.M.M. *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. **BMC Microbiology**, v. 10, article ID 161, 2010.

CHAMBERS, J.M.; CLEVELAND, W.S.; KLEINER, B.; TUKEY, P.A. **Graphical methods for data analysis**. Belmont, CA: Wadsworth & Brooks/Cole, 1983. (Statistics/Probability Series)

CHATFIELD, C. **Problem Solving**: A statistician's guide. 2nd ed. London: Chapman and Hall/CRC, 1995. (Series Chapman & Hall/CRC Texts in Statistical Science)

DORMANN, C.F.; FRÜND, J.; BLÜTHGEN, N.; GRUBER, B. Indices, graphs and null models: analyzing bipartite ecological networks. **The Open Ecology Journal**, v.2, p. 7-24, 2009.

DORMANN, C.F.; GRUBER, B. **Package bipartite**. Version 1.17. [Pacote do programa R]. 10 Aug. 2011. Disponível em: <<http://cran.r-project.org/web/packages/bipartite/bipartite.pdf>>. Acesso em: 27 out. 2011.

DOWDY, S.; WEARDON, S.; CHILKO, D. Distributions of two variables. In: _____. **Statistics for Research**. 3rd ed. Hoboken, NJ.: John Wiley & Sons, 2004. cap. 9, p. 211-264. (Wiley Series in Probability and Statistics)

DUTILLEUL, P.; STOCKWELL, J.D.; FRIGON, D.; LEGENDRE, P. The mantel test versus pearson's correlation analysis: assessment of the differences for biological and environmental studies. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 5, n. 2, p. 131-150, 2000.

ERDÖS, P.; RÉNYI, A. On random graphs I. **Publicationes Mathematicae**, University of Debrecen, v. 6, p. 290-297, 1959.

EULER, L. Solutio problematis ad geometriam situs pertinentis. **Commentarii Academiae Scientiarum Imperialis Petropolitanae**, v. 8, p. 128-140, 1736.

EVERETT, M.G.; BORGATTI, S.P. Analyzing clique overlap. **Connections**, v. 21, n. 1, p. 49-61, 1998.

FAO. Food and Agriculture Organization of the United Nations. Knowledge management of pollination services. In: **Pollination services for sustainable agriculture**. [Folder]. Rome, Italy, 2008. Disponível em: <http://www.fao.org/fileadmin/templates/agphome/documents/Biodiversity-pollination/Pollination-FolderFlyer_web.pdf>. Acesso em: 20 maio 2014.

FREEMAN, L.C. Centrality in social networks conceptual clarification. **Social Networks**, v.1, n. 3, p. 215-239, 1978/79.

FREITAS, B.M.; IMPERATRIZ-FONSECA, V.L.; MEDINA, L.M.; KLEINERT, A.M.P.; GALETTO, L.; NATES-PARRA, G.; QUEZADA-EUÁN, J.J.G. Diversity, threats and conservation of native bees in the Neotropics. **Apidologie**, v. 40, n. 3, p. 332-346, 2009.

FRUCHTERMAN, T.M.J.; REINGOLD, E.M. Graph drawing by force-directed placement. **Software: Practice and Experience**, v. 21, n. 11, p. 1129-1164, 1991.

GUIMARAES, P.R.; GUIMARAES, P. Improving the analyses of nestedness for large sets of matrices. **Environmental Modelling and Software**, v. 21, n. 10, p. 1512-1513, 2006.

GIANNINI, T.C. **Distribuição geográfica de abelhas e plantas associadas através de modelagem computacional**. 2011. 140 p. Tese (Doutorado) – Departamento de Ecologia, Instituto de Biociências, Universidade de São Paulo, São Paulo, 2011.

GIANNINI, T.C.; GARIBALDI, L.A.; ACOSTA, A.L.; SILVA, J.S.; MAIA, K.P.; SARAIVA, A.M.; GUIMARAES, P.R.; KLEINERT, A.M.P. Native and non-native supergeneralist bee species have different impacts on plant-bee networks. **Ecography**, 2014. (*unpublished*).

GOOGLE. **Por que usar o Google**. Disponível em: <http://www.google.com/intl/pt-BR/why_use.html>. Acesso em: 04 maio 2011.

HAIGHT, F.A. **Handbook of the poisson distribution**. New York: John Wiley & Sons, 1967.

HANSEN, D.L.; SHNEIDERMAN, B.; SMITH, M.A. **Analysing social media networks with NodeXL: insights from a connected world**. Amsterdam: Morgan Kaufmann, 2011.

HURLEY, C.B. Clustering visualizations of multidimensional data. **Journal of Computational and Graphical Statistics**, v. 13, n. 4, p. 788-806, 2004.

HURLEY, C. **gclus: clustering graphics**. Version 1.3.1. [Pacote do programa R]. 25 June 2012. Disponível em: <<http://cran.r-project.org/web/packages/gclus/index.html>>. Acesso em: 04 maio 2014.

IMPERATRIZ-FONSECA, V.L.; SARAIVA, A.M.; GONÇALVES, L. A iniciativa brasileira de polinizadores e os avanços para a compreensão do papel dos polinizadores como produtores de serviços ambientais. **Biosci. J.**, Uberlândia, v. 23, sup. 1, p. 100-106, 2007.

JAMBUNATHAN, M.V. Some properties of beta and gamma distributions. **The Annals of Mathematical Statistics**, University of Mysore, v. 25, n. 2, p. 401-405, 1954.

JORDANO, P. Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution. **The American Naturalist**, v. 129, n. 5, p. 657-677, 1987.

JORDANO, P.; VÁZQUEZ, D.; BASCOMPTE, J. Redes complejas de interacciones mutualistas planta-animal. In: MEDEL, R.; AIZEN, M.A.; ZAMORA, R. (Eds.). **Ecología y evolución de interacciones planta-animal**. Santiago de Chile: Editorial Universitaria, 2009. cap. 1, p. 17-41.

KADER, G.D.; FRANKLIN, C.A. The evolution of pearson's correlation coefficient. **Mathematics Teacher**, v. 102, n. 4, Data Analysis and Probability, p. 292-299, 2008.

KAMADA, T.; KAWAI, S. An algorithm for drawing general undirected graphs. **Information Processing Letters**, v. 31, n. 1, p. 7-15, 1989.

KLEINERT, A.M.P.; GIANNINI, T.C. Generalist bee species on Brazilian bee-plant interaction networks. **Psyche**, v. 2012, article ID 291519, p. 1-7, 2012.

KRUG, C.; ALVES-DOS-SANTOS, I. Como a extinção de *Apis mellifera* L. pode afetar a estrutura de uma rede de interação abelha-planta em mata com araucária? In: ENCONTRO SOBRE ABELHAS: GENÉTICA E BIOLOGIA EVOLUTIVA DE ABELHAS, 9., 2010, Ribeirão Preto. **Anais...** Ribeirão Preto: FUNPEC, 2010. p. 304-306.

KRZYWINSKI, M.; ALTMAN, N. Points of significance: visualizing samples with box plots. **Nature Methods**, v. 11, n. 2, p. 119-120, 2014.

KUMAR, K.; NOVAK, J.; TOMKINS, A. Structure and evolution of online social networks. In: ACM SIGKDD INTERNATIONAL CONFERENCE OF KNOWLEDGE DISCOVERY AND DATA MINING (KDD 2006), 12., 2006, Philadelphia, PA. **Proceedings...** Philadelphia, PA., 2006. p. 611-617.

LEGENDRE, P.; LEGENDRE, L. Ecological resemblance. In: _____. **Numerical Ecology**: developments in environmental modelling. 3rd english ed. Amsterdam: Elsevier, 2012. cap. 7, p. 265-335.

LEWINSOHN, T.M.; PRADO, P.I.; JORDANO, P.; BASCOMPTE, J.; OLESEN, J.M. Structure in plant-animal interaction assemblages. **Oikos**, v. 113, n. 1, p. 174-184, 2006.

LILLIEFORS, H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. **Journal of the American Statistical Association**, v. 62, n. 318, p. 399-402, 1967.

MANTEL, N. The detection of disease clustering and a generalized regression approach. **Cancer Research**, v. 27, n. 2, p. 209-220, 1967.

MELLO, M.A.R. **Guia para análise de redes ecológicas**. [S.l.]: 30 Aug. 2012. Disponível em: <<http://marcomello.casadosmorcegos.org/Teaching.html>>. Acesso em: 11 set. 2012.

MEMMOTT, J. The structure of a plant-pollinator food web. **Ecology Letters**, v. 2, n. 5, p. 276-280, 1999.

MEMMOTT, J.; CRAZE, P.G.; WASER, N.M.; PRICE, M.V. Global warming and the disruption of plant-pollinator interactions. **Ecology Letters**, v. 10, n. 8, p. 710-717, 2007.

MEMMOTT, J.; WASER, N.M. Integration of alien plants into a native flower-pollinator visitation web. **Proc. R. Soc. Lond. B**, v. 269, n. 1508, p. 2395-2399, 2002.

MEMMOTT, J.; WASER, N.M.; PRICE, M.V. Tolerance of pollination networks to species extinctions. **Proc. R. Soc. Lond. B**, v. 271, n. 1557, p. 2605-2611, 2004.

MMA. Ministério do Meio Ambiente. **Projeto polinizadores do Brasil**: sobre o projeto. Disponível em: <<http://www.polinizadoresdobrasil.org.br>>. Acesso em: 11 nov. 2010.

NIST/SEMATECH. Exploratory data analysis. In: _____. **e-Handbook of statistical methods**. [S.l.]: [s.n.], 2003. cap. 1. Disponível em: <<http://www.itl.nist.gov/div898/handbook/>>. Acesso em: 30 mar. 2013.

NOOY, W.; MRVAR, A.; BATAGEL, J.V. **Exploratory network analysis with Pajek**. Cambridge: Cambridge University Press, 2005.

OKSANEN, J. **Multivariate analysis of ecological communities in R: vegan tutorial**. [Version] [Pacote do programa R]. 08 Feb. 2013. Disponível em: <<http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf>>. Acesso em: 29 jun. 2013.

OLESEN, J.M.; BASCOMPTE, J.; DUPONT, Y.L.; JORDANO, P. The modularity of pollination networks. **PNAS**, v. 104, n. 50, p. 19891-19896, 2007.

OLESEN, J.M.; ESKILSEN, L.I.; VENKATASAMY, S. Invasion of pollination networks on oceanic islands: importance of invader complexes and endemic super generalists. **Diversity and Distributions**, v. 8, n. 3, p. 181-192, 2002.

OMG. Object Management Group. **Business Process Model and Notation (BPMN)**. Version 2.0. 03 Jan. 2011. Disponível em: <<http://www.omg.org/spec/BPMN/2.0/>>. Acesso em: 03 abr. 2014.

R DEVELOPMENT CORE TEAM and contributors worldwide. **The R stats package**: R statistical functions. Version 2.15.0. [Pacote do programa R]. 30 Mar. 2012, 2012a. Disponível em: <<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html>>. Acesso em: 20 maio 2014.

R DEVELOPMENT CORE TEAM and contributors worldwide. **The R graphics package**: R functions for base graphics. Version 2.15.0. [Pacote do programa R]. 30 Mar. 2012, 2012b. Disponível em: <<http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/00Index.html>>. Acesso em: 20 maio 2014.

RADHOUANI, H.; SILVA, N.; POETA, P.; TORRES, C.; CORREIA, S.; IGREJAS, G. Potential impact of antimicrobial resistance in wildlife, environment and human health. **Frontiers in Microbiology**, v. 5, article ID 23, 2014.

RIPLEY, B.; VENABLES, B.; HORNIK, K.; GEBHARDT, A., FIRTH, D. **Support functions and datasets for venables and ripley's MASS**. Version 7.3-23. [Pacote do programa R]. 08 Jan. 2013. Disponível em: <<http://www.stats.ox.ac.uk/pub/MASS4/>>. Acesso em: 05 abr. 2014.

RODRÍGUEZ-GIRONÉZ, M.A.; SANTAMARÍA, L. A new algorithm to calculate the nestedness temperature of presence-absence matrices. **J. Biogeogr.**, v. 33, n. 5, p. 924-935, 2006.

SARAIVA, A.M. **Tecnologia da informação na agricultura de precisão e biodiversidade**: estudos e proposta de utilização de Web services para desenvolvimento e integração de sistemas. 2003. 187 p. Tese (Livre Docência) – Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica, Universidade de São Paulo, São Paulo, 2003.

SARAIVA, A.M.; IMPERATRIZ-FONSECA, V.L.; CUNHA, R.S.; CARTOLANO-JÚNIOR, E.A. WebBee – a Web-based information network on Bees. **Revista de Engenharia de Computação e Sistemas Digitais**, PCS, EPUSP, São Paulo, v. 1, p. 77-86, 2003.

SARAIVA, A.M.; SILVA, J.S. **Protocol for handling pollinator interaction data:** user needs assessment and inventory of the ways data is currently handled. Version 1.0. [Report]. Developed by the Universidade de São Paulo with support from FAO, São Paulo, 2010.

SCOTT, J. **Social network analysis:** a handbook. 2nd ed. London: Sage, 2000.

SILVA, J.S.; SANTOS, R.T.; SARAIVA, A.M.; CORRÊA, P.L.P. Biodiversity Informatics: global initiatives for data integration. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT (CONTECSI), 7., 2010, São Paulo. **Proceedings...** São Paulo, 2010. p. 1180-1193.

SILVA, J.S.; SARAIVA, A.M. Métricas de análise de redes sociais aplicadas no estudo de redes de interação polinizador-planta. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA (SBIAgro 2011), 8., 2011, Bento Gonçalves, RS. **Anais...** Bento Gonçalves, RS, 2011.

SILVA, J.S.; SARAIVA, A.M. Social networks analysis metrics and its application in pollinator-plant interaction networks studies. In: ENCONTRO SOBRE ABELHAS: UM SÉCULO DE PESQUISAS NO BRASIL, 10., 2012, Ribeirão Preto. **Anais...** Ribeirão Preto: FUNPEC, 2012. p. 441.

SILVA, J.S.; SARAIVA, A.M. Metodologia para aplicação de métricas de análise de redes sociais em biodiversidade. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA (SBIAgro 2013), 9., 2013, Cuiabá. **Anais...** Cuiabá, 2013.

SILVA, J.S.; SARAIVA, A.M. A Methodology for Applying Social Network Analysis Metrics on Biodiversity. **IEEE Latin America Transactions**, 2015. (*unpublished*).

SILVA, J.S.; SARAIVA, A.M.; STOPPE, N.C.; TORRES, T.T.; OTTOBONI, L.M.M. Social network analysis metrics and its application in microbiological interaction networks studies. In: INTERNATIONAL CONFERENCE ON ECOLOGICAL INFORMATICS: INFORMING DECISIONS ON BIODIVERSITY AND NATURAL RESOURCES CONSERVATION (ISEI 2012), 8., 2012, Brasília. **Proceedings...** Brasília, 2012. p. 167-169.

SILVA, J.S.; STOPPE, N.C.; TORRES, T.T.; OTTOBONI, L.M.M.; SARAIVA, A.M. Social network analysis metrics and their application in microbiological network studies. In: INTERNATIONAL WORKSHOP ON COMPLEX NETWORKS (CompleNet 2014), 5., 2014, Bologna, Italy. **Proceedings...** Bologna: Springer International Publishing, Series Studies in Computational Intelligence, v. 549, Bologna, 2014. p. 251-260.

SOKAL, R.R.; MICHENER, C.D. A statistical method for evaluating systematic relationships. **Science Bulletin**, Univ. Kansas, v. 38, n. 22, p. 1409-1438, 1958.

STOPPE, N.C.; SILVA, J.S.; TORRES, T.T.; CARLOS, C.; HACHICH, E.M.; SATO, M.I.Z.; SARAIVA, A.M.; OTTOBONI, L.M.M. Clustering of water bodies in unpolluted and polluted environments based on *Escherichia coli* phylogroup abundance using a

simple interaction database. **Genetics and Molecular Biology**, v. 37, n. 4, 2014a. No prelo.

STOPPE, N.C.; SILVA, J.S.; ROSENBERG, R.J.; CARLOS, C.; HACHICH, E.M.; SATO, M.I.Z.; SARAIVA, A.M.; OTTOBONI, L.M.M.; TORRES, T.T. Complex pattern of *Escherichia coli* phylogenetic group distribution in worldwide samples of human feces and wastewater treatment plants. **Springer Plus**, 2014b. (*unpublished*)

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. **Package 'rpart'**: recursive partitioning and regression trees. Version 4.1-8. [Pacote do programa R]. 28 Mar. 2014. Disponível em: <<http://cran.r-project.org/web/packages/rpart/rpart.pdf>>. Acesso em: 27 abr. 2014.

TUKEY, J.W. **Exploratory data analysis**. Reading, MA: Addison-Wesley, 1977.

ULRICH, W.; ALMEIDA-NETO, M.; GOTELLI, N.J. A consumer's guide to nestedness analysis. **Oikos**, v. 118, n. 1, p. 3-17, 2009.

VÁZQUEZ, D.P.; AIZEN, M.A. Null model analyses of specialization in plant–pollinator interactions. **Ecology**, v. 84, n. 9, p. 2493-2501, 2003.

VÁZQUEZ, D.P.; AIZEN, M.A. Asymmetric specialization: a pervasive feature of plant–pollinator interactions. **Ecology**, v. 85, n. 5, p. 1251-1257, 2004.

VÁZQUEZ, D.P.; BLÜTHGEN, N.; CAGNOLO, L.; CHACOFF, N.P. Uniting pattern and process in plant–animal mutualistic networks: a review. **Annals of Botany**, v. 103, n. 9, p. 1445-1457, 2009.

VÁZQUEZ, D.P.; MORRIS, W.F.; JORDANO, P. Interaction frequency as a surrogate for the total effect of animal mutualists on plants. **Ecology Letters**, v. 8, n. 10, p. 1088-1094, 2005.

WEBBEE. Rede de informações sobre abelhas, baseada na web. **Foto**: Marilda Cortopassi Laurino. Disponível em: <http://www.usp.br/webbee/fichas/ab16/l_principal_paratrigona_subnuda.jpg>. Acesso em: 19 set. 2011.

WEISSTEIN, E.W. **Gamma distribution**. [S.l.]: MathWorld, A Wolfram Web Resource, 2014a. Disponível em: <<http://mathworld.wolfram.com/GammaDistribution.html>>. Acesso em: 23 maio 2014.

WEISSTEIN, E.W. **Poisson distribution**. [S.l.]: MathWorld, A Wolfram Web Resource, 2014b. Disponível em: <<http://mathworld.wolfram.com/PoissonDistribution.html>>. Acesso em: 23 maio 2014.

WHITE, A.P.; SIBLEY, K.A.; SIBLEY, C.D.; WASMUTH, J.D.; SCHAEFER, R.; SURETTE, M.G.; EDGE, T.A.; NEUMANN, N.F. Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. **Applied and Environmental Microbiology**, v. 77, n. 21, p. 7620-7632, 2011.

WICKHAM, H. The split-apply-combine strategy for data analysis. **Journal of Statistical Software**, v. 40, n. 1, p. 1-29, 2011.

WOOTTON, J.T.; EMMERSON, M. Measurement of interaction strength in nature. **Annual Review of Ecology, Evolution, and Systematics**, v. 36, p. 419-444, 2005.

APÊNDICE A – *Paper Genetics and Molecular Biology* (co-autoria)

[No prelo]

Clustering of water bodies in unpolluted and polluted environments based on Escherichia coli phylogroup abundance using a simple interaction database

Genetics and Molecular Biology Volume 37 (4), 2014. No prelo.

Available in: <http://www.gmb.org.br/onlineversion.html>

Clustering of water bodies in unpolluted and polluted environments based on *Escherichia coli* phylogroup abundance using a simple interaction database

Nancy de Castro Stoppe • Juliana Saragiotto Silva • Tatiana Teixeira Torres • Camila Carlos • Elayse Maria Hachich • Maria Inês Zanoli Sato • Antonio Mauro Saraiva • Laura Maria Mariscal Ottoboni

N.C. Stoppe • C.Carlos • L.M.M.Ottoboni

Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas - UNICAMP, C.P. 6010, 13083-875 Campinas, SP, Brasil

E-mail: ottoboni@unicamp.br

Telephone: +55 19 35211036

Fax: +55 19 35211089

WWW.cbmeg.unicamp.br

N.C. Stoppe • J.S.Silva • T.T. Torres • A.M.Saraiva • L.M.M. Ottoboni

Núcleo de Pesquisa em Biodiversidade e Computação (BioComp-USP) - Universidade de São Paulo – USP, São Paulo, SP, Brasil

E.M.Hachich • M.I.Z.Sato

Departamento de Análises Ambientais, Companhia Ambiental do Estado de São Paulo - CETESB, São Paulo, SP, Brasil

T.T.Torres

Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo – USP, São Paulo, S. P., Brasil

J.S.Silva

Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT), Cuiabá, MT, Brasil

A.M.Saraiva

Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da USP, São Paulo, SP, Brasil

Abstract

Different types of water bodies including lakes, streams, and coastal marine waters are often susceptible to fecal contamination from a range of point and nonpoint sources and have been evaluated using a fecal indicator microorganism. *Escherichia coli* is the most common fecal indicator used, but it does not allow the discrimination of the source pollution by traditional cultivation methods. The use of triplex PCR seems to be a cheap and rapid approach, and it enabled us to separate water bodies according to geographical location. Besides, the phylogenetic distribution of *E. coli* subgroups in isolated from water samples shows higher frequencies of subgroups A₁ and B2₃ in rivers associated to human pollution sources, while subgroups D₁ and D₂ are associated to pristine sites and B1 to domesticated animal sources suggesting their use as a first screening for pollution source identification. We also proposed a simple classification based on phylogenetic subgrouping distribution using w-clique that allows differentiate polluted from nonpolluted sites.

KEY WORDS: *E.coli* • phylogenetic groups • pollution sources • interaction networks • Social Network Analysis

Introduction

The microbiological quality of water is usually evaluated by means of fecal indicator microorganisms and *E. coli* has often been used because it is a normal inhabitant of most warm-blooded animals intestinal tract. However, the traditional methods used hitherto have not allowed the differentiation among host sources. Reliable and accurate fecal source identification methods are extremely important to control fecal contamination from relevant animal sources, to protect recreational water users from waterborne pathogens, and to preserve the integrity of drinking water supplies (Rosley and Bukh 2011; USEPA 2005).

Clermont *et al.* (2000) developed a method for the assignment of *E. coli* isolates in four major phylogenetic groups A, B1, D, or B2. Because of its simplicity and rapidity, it has been widely used with different purposes, including ecological niche differentiation, propensity to cause diseases and fecal source tracking (Carlos *et al.* 2010; Escobar-Paramo *et al.* 2006; Figueira *et al.* 2011; Gordon *et al.* 2008; Johnson *et al.* 2004; Orsi *et al.* 2007, 2008; Ratajczak *et al.* 2010; Walk *et al.* 2007). It is based on triplex PCR, and uses the combination of two genes (*chuA*, the outer-membrane hemin receptor gene, and *yjaA*, which encodes an uncharacterized protein) and a DNA fragment that has been recently identified as part of a putative lipase esterase gene (Tenallion *et al.*

2010). To improve the discriminative power of analyses when several isolates per sample were considered, Escobar-Páramo et al. (2004) proposed the use of all the combinations of genetic markers and thus, seven subgroups (A₀, A₁, B₁, B₂, B₃, D₁ and D₂) were defined.

Furthermore, algorithms, metrics and computational resources for analyzing interaction networks appear as an important tool to systematically measure interdependencies among molecular markers and water bodies. Similarly, researchers in the Ecology field (Memmott et al. 2007; Bascompte et al. 2006) have already exploited the interaction networks resources for analyzing the visitor-plant interactions, focusing on the factors that have contributed and influenced biodiversity maintenance. The conceptual foundations of these tools are the same as in Social Network Analysis (SNA) – that provides algorithms and metrics to characterize the network structure, as well as to identify cohesive subgroups.

The aim of this work was to establish a classification of *E.coli* strains isolated from water bodies based on phylogenetic subgroups and try to associate with the pollution source, by means of metric *w-clique*.

Material and methods

Sample collection

Water from twelve rivers and reservoirs with different pollution levels in the State of São Paulo (Figure 1 and Table 1) were collected in sterilized bottles according to Standard Methods (APHA 2010). They are part of sampling sites of a Surface Waters Monitoring Program of State of São Paulo that has been conducted by CETESB (State of São Paulo Environmental Agency). The monitoring includes physical, chemical and biological analysis in the twenty-two Watershed Management Units (WMU) located in the State of São Paulo. Two indexes are currently used for the evaluation of domestic effluents dilution and the trophic state of the water bodies. The water quality index (WQI) is obtained from the combination of a set of variables including pH, dissolved oxygen, biological oxygen demand, thermotolerant coliforms or *E.coli*, water temperature, total nitrogen, total phosphorus, total solids and turbidity. The trophic state index (TSI), on the other hand, is based on the concentrations of chlorophyll and phosphorus. WQI ranges from 0 to 100 and five intervals indicate the water quality. Values in the range 0-19 indicate a very bad quality; 20-36, bad; 37-51, acceptable; 52-79, good and 80-100 indicate a very good quality. TSI ranges

from 47 to 67 with values below 47 indicating an ultraoligotrophic environment; 48-52, oligotrophic; 53-59, mesotrophic; 60-63, eutrophic and values above 67 indicating a hypereutrophic environment. The sampling sites located on reservoirs (GUAR0502, GUAR0601, BILL2251 and BILL2801) are mainly used as recreational water areas and were evaluated by *E.coli* levels. The identification of the main pollution source in each site was based on historical data and sampling observations. The samples were collected between July 2009 and April 2010 (CETESB 2010, 2011).

Strains isolation

Samples were analyzed using the membrane filter technique according to U.S. Environmental Protection Agency method 1603 (USEPA 2002). Briefly, volumes from 0.01 to 100 ml were filtered onto 0.45 μ m membrane and incubated on modified mTEC agar at 35 ± 0.5 °C for 2h and at 44.5 ± 0.2 °C for 22-24h. Approximately ten typical colonies (red to magenta) from each sample were streaked on LES ENDO agar (Difco), incubated for 24 h at 35 °C and tested for citrate utilization, lactose fermentation, oxidase, L-lysine decarboxylase, motility, glucose and sucrose fermentation, tryptophan deamination, indole production, urea hydrolysis and sulfide production. Typical *E. coli* profile was reisolated on nutrient agar, incubated for 24 h at 35 °C and kept at -70 °C in tryptic soy broth (Difco) with 10% glycerol (v/v) for further analysis (ATCC 2010).

Phylogenetic grouping

Genomic DNA from strains was isolated with the Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer's instructions and the phylogenetic grouping of *E.coli* strains were determined as previously described by Clermont *et al.* (2000). The strains were assigned to the phylogenetic subgroups according to the combination of PCR products of the genes *chuA*, *yjaA* and DNA fragment TspE4.C2 as the following: A₀ (-/-/-); A₁ (-/+/-); B₁ (-/-/+); B₂ (+/+/-); B₂₃ (+/+/+); D₁ (+/-/-) and D₂ (+/-/+) (Escobar-Páramo *et al.* 2004).

Statistical analysis

The chi-square test was used to determine if the differences in the distribution of phylogenetic subgroups among rivers and reservoirs were significant. Correlation analysis was performed by the Mantel test through comparison of two dissimilarity matrices, calculated from

vegdist function (with bray-curtis index as parameter), using the community ecology package vegan for R program (Oksanen, 2011). Following, we prepared the similarity matrices as a complement of dissimilarity matrix (1-vegdist(matrix, “bray”)) because we are using an index that ranges between 0 and 1. The similarity matrix of the phylogenetic subgroups was clustered using UPGMA (unweight pair group method with arithmetic mean) algorithm and the dendrogram was built using DendroUPGMA (Garcia-Vallve et al, 1999).

Social Network Analysis metric

The SNA metric *clique* has been used to identify the cohesive subgroups (clusters) in the network structures. According to Nooy et al. (2005), a clique is composed by a set of three or more vertices totally connected to each other. However, this metric often relies on binary interaction matrix instead of frequency interaction matrix, which does not take into account interaction abundance to identify the clusters (Araújo et al. 2008). As an alternative, these researchers developed the *w-clique* approach, i.e., vertex groups in which all vertices are connected to each other by “strong” interactions (the weights of which are higher than the average network weight).

Data analysis

The data used in the present study come from a bipartite microbiological interaction database, composed by a weighted matrix (isolates abundance), in which rows were represented by water bodies and columns by phylo-groups (Table 1); this matrix was prepared to be used in the computer program *Dieta1.exe*¹⁸, in order to identify niche overlap (cohesive subgroups) in this interaction network. In this analysis, the following parameter values¹⁹ were used: data type – integers; running a Monte Carlo bootstrap – 1000 replications; the diet proportion method calculation – numerical sum; weight factor – 5; printing the file bootstrap.txt. Among the results, an output file ‘B[file_name].mat’ was generated that corresponds to a binary matrix (0/1), in which cells (containing the number one) represent interactions the weights of which are higher than the average network weight (*w-cliques*).

¹⁸ The program “Dieta1” was developed by Araújo et al. (2008) and allows analyzing individual specialization and cohesive subgroups in weighted networks – it was based on the complex network theory. All the files of this toll (computer program, source code and manual) are available in the supplement section of Araújo et al. (2008).

¹⁹ More details about these parameters can be seen in the Dieta1.exe manual, available in the supplement section of Araújo et al. (2008).

Following the identification of niche overlap, we used the *Pajek* program²⁰ (Batagelj and Mrvar 1998) to transform the network from arcs to edges and then save the matrix in a '.net' file extension. Finally, the matrix was submitted to the *Ucinet* program²¹ for the identification of *w-cliques* (Borgatti et al, 2002). This tool generated two output files: the first, which shows the cliques found, with the identification of the individual memberships of each clique and other clique overlap analyses²²; and the second, which represents the cluster diagram (dendrogram).

Results and discussion

Five hundred and forty three strains were isolated from twelve rivers and reservoirs. They were classified according to the phylogenetic subgroups and their distribution can be seen in Figure 2. The observed distribution of the phylogenetic subgroups among rivers was different than the expected frequencies and was statistically significant ($\chi^2 = 217.22$, $df = 66$, $P < 0.005$). The majority of the environmental strains belongs to subgroup A₀, even though river and reservoir sites have different sources of pollution. The distribution of the phylogenetic subgroups in environmental samples has not shown similar patterns in previous studies. Group B1 is the most common (over 70%) in beaches in California and surrounding environmental waters of sewage treatment plants in Australia (Hamilton et al, 2010; Anastasi et al, 2012). In surface water samples (lakes and rivers) half of the isolates belong to phylo-group B1 suggesting its as the most frequent group in environmental samples (Hamelin et al, 2007; Power et al, 2005; Walk et al, 2007). However, other factors can influence the structure of an *E. coli* population in water, such as the hydrological conditions on the watershed and geographical location (Ratajczak et al, 2010; Tenaillon et al, 2010). Indeed, some authors showed that subgroup A₀ is more environmentally adapted (Figueira et al. 2011; Higgins et al. 2007; Walk et al. 2009), while B2₂ is the least common subgroup found in rivers and reservoirs (Figueira et al. 2011), in line with the results observed in the present study.

²⁰ Available online for free download in <<http://pajek.imfm.si/doku.php>>.

²¹ Available online for free download in <<http://www.analytictech.com/downloaduc6.htm>>.

²² More details about clique overlap analyses can be seen in Everett and Borgatti (1998).

Only seven strains isolated from different sites (TIET2050, TIET3120, TGDE0900, JAMI02100, PTEI02900 and BILL2801) presented a *chuA*-, *yjaA*+, TspE4.C2+ profile (data not shown) and according to Rodrigues-Siek et al. (2005) are assigned as B1 group. Gordon et al. (2008) characterized 662 *E.coli* strains from different hosts and environmental strains and did not observe any matching this profile. The lower frequency observed here indicates the rare occurrence of this profile which was also observed by Higgins et al. (2007) that found only one out of 68 strains isolated from surface water samples.

Although the majority of human strains belong to group A, Orsi et al. (2007) and Carlos et al. (2010) suggested the use of group B2 as an indicator of human pollution source, due to the recurrence in this host. Indeed, the present data also show high prevalence of this group in the sites strongly impacted by human sources (BILL2801, BILL2251, GUAR0502 and GUAR0601).

Ten years ago, the Billings and Guarapiranga Reservoirs, presented group A as the most frequent, followed by B1, D and the group B2 was the least (Orsi et al. 2007). Our results also showed the predominance of group A, but a decrease of B1 and a significant increase of the frequency of group B2. This group seems to be associated with human pollution sources, which have grown in the last years. The population living in surrounding dam areas increased 24% and 30% in ten years, for Billings and Guarapiranga Reservoirs, respectively. Part of this population has access to neither sewage collection nor wastewater treatment and this might explain the obtained data. Furthermore, WQI changed from good to bad in these sites in the last ten years, although even in 2000 the sites did not meet the recreational waters legislation, which evaluates fecal contamination by fecal bacteria levels (Brazil, 2000).

The sites with expected domesticated animal pollution sources (TIET3120, TGDE0900, AGUA2800 and JAMI2100) did not present a similar pattern of phylo-group distribution. Notwithstanding, TIET3120 and TGDE0900 are located downstream to cities that discharge untreated wastewater, which makes phylo-groups distribution analogous to the sites with human pollution sources. Meanwhile, AGUA2800 and JAMI2100 are located in areas with agricultural activities and have seasonal presence of cattle and the most frequent subgroups were A₀ and B1, as observed by others (Carlos et al. 2010; Higgins et al. 2007; Ishii et al. 2007).

Some of the pristine sites (IPIR0018 and PBAL0014) showed a high frequency of group D, which was also observed by Higgins et al. (2007) in a non-polluted site suggesting association with wildlife. Surprisingly, the other fresh sites (TIET2050 and PTEI2900) showed higher

frequencies of groups A and B2, which are similar to the human source sites. The former water bodies are located inside an environmental protection area (from source to mouth) and are truly pristine, while the latter can have some contribution of other pollution sources, despite good quality indexes that these sites presented during the studied period (Table 1) (CETESB 2010, 2011). As these sites are not located inside protected areas, they could be affected by nonpoint pollution sources or even illegal discharges.

The high frequency of subgroup A_0 in all sites can bias the phylogenetic group distribution in the water samples with different pollution sources. If this subgroup is removed from all sites, subgroups A_1 and $B2_3$ are more closely related to expected human sources, $B1$ to domesticated animal sources and D_1 and D_2 to pristine environments suggesting this classification as a first screening for pollution source identification.

The contamination of the sites was not detected by the two water quality indexes, suggesting that other tools should be used for pollution evaluation. Mantel correlation test between quality indexes (WQI and TSI) and phylogenetic subgroup distribution showed, respectively, either a weak ($r=0.36$, $P=0.014$) or no correlation ($r=0.2537$, $P=0.074$), indicating that both indexes miss important information for pollution evaluation. As phylogenetic subgroup distribution seems to be a suitable tool to identify sources of pollution, it should be included for pollution classification of water bodies.

To try and cluster the rivers according to pollution source and degree, we evaluated our data using correspondence analysis, but a spread distribution was observed (data not shown). Similarity matrix among the isolates was calculated and clustered by UPGMA and showed two groups (Figure 3). Two pristine sites (IPIR0018 and PBAL0014) and one animal site (AGUA2800) belong to the first cluster, which was unexpected since these sites do not share similar characteristics. This is also observed in the other group where sites with different degrees and sources of pollution were clustered, suggesting that this tool cannot explain the observed data.

The impact of geographical location seems to be an important factor in the distribution of phylogenetic subgroups. The *w-clique* metric clustered the water bodies in two groups (Figure 4). The first group contained the water bodies belonging to the São Paulo Metropolitan area (BILL2801, BILL2251, GUAR0502, GUAR0601, TIET2050, TIET3120 and TGDE0900) and the other clustered the water bodies located far from the metropolitan region (PTEI2900, JAMI2100, IPIR0018 and PBAL0014). In our case, geographical location reflects the degree of pollution,

since São Paulo Metropolitan area has long suffered from air and water pollution as most industrialized areas. Interesting to notice is in the sites where nonpoint animal pollution sources were expected (TGDE0900, JAMI2100 and AGUA2800), no specific behaviour was observed indicating that animal sources in these sites seems to be less important. Similar cluster results were observed if the strains belonged to subgroup A₀ was removed (data not shown).

Thus, by means of cliques identification, it was possible to discover new patterns in a simple interaction database, such as clustering of water bodies (in nonpolluted and polluted environments), based on the abundance of phylo-groups. This clustering was not recovered using traditional methods (Mantel correlation test and correlation analysis), which shows the innovative contribution of this research.

Besides, our results showed that the quality indexes could not reflect all aspects of the evaluation of domestic effluents dilution and the trophic state of the water bodies, since TIET2050 and TGDE00900 presented good average scores (Table 1) and had phylogenetic group distribution more related to polluted sites according to w-clique classification. Our results suggest that w-clique classification can be used as a potential pollution classifier in sampling sites as a complementary tool for pollution degree evaluation.

Acknowledgements

This work was supported by FAPESP (Project 2007/55312-6). NCS has received a fellowship from FAPESP (Project 2010/15222-0). LMMO has received a research fellowship from CNPq. The Research Center on Biodiversity and Computing (BioComp) was supported by the USP Provost's Office for Research. We thank Dr. Marco Mello and his Guide for Ecological Networks Analysis for helping our data analysis.

References

- Anastasi, EM, Matthews, B, Stratton, HM, Katouli, M (2012) Pathogenic *Escherichia coli* found in sewage treatment plants and environmental waters. *Appl Environ Microbiol* 78: 5536-5541.
- APHA (American Public Health Association) (2010) Samples. In: _____ Standard methods for examination of water and wastewater. Section 9060. Available at: <http://standardmethods.org>. Accessed 18 Feb 2010.
- Araújo MS, Guimarães PR Jr, Svanbäck R, Pinheiro A, Guimarães P, Reis SFD, Bolnick DI (2008) Network analysis reveals contrasting effects of intraspecific competition on individual versus population diets. *Ecology* 89: 1981-1993.

ATCC (American Type Culture Collection) (2010) Microbial media and supplements. Available at: <http://www.atcc.org/CulturesandProducts/Microbiology/MicrobialMediaandSupplements/tabid/186/Default.aspx>. Accessed 07 Mar 2010.

Batagelj V, Mrvar A (1998) Pajek – program for large network analysis. *Connections* 21: 47–57.

Bascompte J, Jordano P, Olesen JM (2006) Asymmetric Coevolutionary Networks Facilitate Biodiversity Maintenance. *Science*, 312: 431-433.

Borgatti, SP, Everett, MG, Freeman, LC (2002) Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Brazil (2000). Ministério do Meio Ambiente. Conselho Nacional do Meio Ambiente. Resolução CONAMA 274, de 29 de novembro de 2000. Define os critérios de balneabilidade em águas brasileiras. Available at: <http://www.mma.gov.br/port/conama/legiano1.cfm?codlegitipo=3&ano=2000>. Accessed 27 Sep 2012.

Carlos C, Pires MM, Stoppe NC, Hachich EM, Sato M.Z, Amaral TATG, Amaral, LA, Ottoboni LMM (2010) *Escherichia coli* phylogenetic group determination and application in the identification of the major animal source of fecal contamination. *BMC Microbiol* 10: 161.

CETESB (Companhia Ambiental do Estado de São Paulo) (2010) Relatório de qualidade das águas superficiais no estado de São Paulo 2009. São Paulo: CETESB, 310p. (<http://www.cetesb.sp.gov.br/publicacoes/publicacoes.asp>).

CETESB (Companhia Ambiental do Estado de São Paulo) (2011) Relatório de qualidade das águas superficiais no estado de São Paulo 2010. São Paulo: CETESB, 298p. (<http://www.cetesb.sp.gov.br/publicacoes/publicacoes.asp>).

Clermont O, Bonacorsi S, Bingen E. (2000) Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* 66: 4555-4558.

Escobar-Páramo P, Grenet K, LeMenac'h A, Rode L, Salgado E, Amorin C, Gouriou S, Picard B, Rahimy C, Andreumont A, Denamur E, Ruimy R (2004) Large-scale population structure of human commensal *Escherichia coli* isolates. *Appl Environ Microbiol* 70: 5698-5700.

Escobar-Páramo P, LeMenac'h A, LeGall T, Amorin C, Gouriou S, Picard B, Skurnik D, Denamur E (2006) Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environ Microbiol* 8:1975-1984.

Everett MG, Borgatti SP (1998) Analyzing Clique Overlap. *Connections* 21: 49-61

Figueira V, Sera E, Manaia CM (2011) Differential patterns of antimicrobial resistance in population subsets isolated from waste- and surface waters. *Sci Total Environ* 409: 1017-1023.

Garcia-Vallve S, Palau J, Romeu A (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol Biol Evol* 9: 1125-1134 (Available at: <http://genomes.urv.cat/UPGMA>).

Gordon DM, Clermont O, Tolley H, Denamur E. (2008) Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* 10: 2484-2496.

Hamelin, K, Bruant, G, El-Shaarawi, A, Hill, S, Edge, T.A, Fairbrother, J, Harel, J, Maynard, C, Masson, L, Brouseau, R (2007) Occurrence of virulence and antimicrobial resistance genes in

Escherichia coli isolates from different aquatic ecosystems within the St. Clair River and Detroit River areas. *Appl Environ Microbiol* 73: 477-484.

Hamilton MJ, Hadi AZ, Griffith JF, Ishii, S, Sadowsky MJ (2010) Large scale analysis of virulence genes in *Escherichia coli* strains isolated from Avalon Bay, CA. *Water Res* 44: 5463-5473.

Higgins J, Hohn C, Hornor S, Frana M, Denver M, Joerger R (2007) Genotyping of *Escherichia coli* from environmental and animal samples. *J Microbiol Meth* 70: 227-235.

Ishii S, Meyer KP, Sadowsky, MJ (2007) Relationship between phylogenetic groups, genotypic clusters, and virulence factors of *Escherichia coli* strains from diverse human and animal sources. *Appl Environ Microbiol* 73: 5703-5710.

Johnson LAK, Brown MB, Carruthers EA, Ferguson JA, Dobek PE, Sadowsky MJ (2004) Sample size, library composition, and genotypic diversity among natural populations of *Escherichia coli* from different animals influence accuracy of determining sources of fecal pollution. *Appl Environ Microbiol* 70: 4478-4485.

Li B, Sun J-Y, Han, L-Z, Huang XH, Fu Q, Ni Y-X (2010) Phylogenetic groups and pathogenicity islands markers in fecal *Escherichia coli* isolates from asymptomatic humans in China. *Appl Environ Microbiol* 76: 6698-6670.

Memmott J, Craze PG, Waser NM, Price, MV (2007) Global warming and the disruption of plant–pollinator interactions. *Ecology Letters* 10: 710–717.

Nooy W, Mrvar A, Batagelj V (2005) *Exploratory Network Analysis with Pajek*. Cambridge University Press, 334p.

Oksanen J (2011) Multivariate analysis of ecological communities in R: vegan tutorial. Version of Oct 30, 2011. Available at: <http://cc.oulu.fi/~jarioksa/opetus/metodi/vegan tutor.pdf>. Accessed 08 Oct 2012.

Orsi RH, Stoppe NC, Sato MIZ, Gomes, TAT, Prado PI, Manfio GP, Ottoboni LMM (2007) Genetic variability and pathogenicity potential of *Escherichia coli* isolated from recreational water reservoirs. *Res Microbiol* 158: 420-427.

Orsi RH, Stoppe NC, Sato MIZ, Prado PI, Ottoboni LMM (2008) Phylogenetic group distribution among *Escherichia coli* isolated from rivers in Sao Paulo State, Brazil. *World J Microbiol* 24: 1573-1577.

Power ML, Littlefield-Wyer J, Gordon DM, Veal DA, Slade MB (2005) Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environ Microbiol* 7: 631-640.

Ratajczak M, Laroche E, Berthe T, Clermont O, Pawlak B, Denamur E, Petit F (2010) Influence of hydrological conditions on the *Escherichia coli* population structure in the water of a creek on a rural watershed. *BMC Microbiol* 10: 222.

Rodrigues-Siek K, Giddings CW, Doetkott C, Johnson, TJ, Fakhr MK, Nolan LK (2005) Comparison of *Escherichia coli* isolates implicated in human urinary tract infection and avian colibacillosis. *Microbiol* 151: 2097-2110.

Roslev P, Bukh AS (2011) State of the art molecular markers for fecal pollution source tracking in water. *Appl Microbiol Biotechnol* 89: 1341-1355.

Tenallion O, Skurnik D, Picard B, Denamur E. (2010) The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 8: 207-217.

USEPA (US Environmental Protection Agency) (2002) Method 1603 *Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant *Escherichia coli* Agar (Modified m-TEC), EPA 821-R-02-023. Cincinnati: U.S Environmental Protection Agency, Office of Research and Development.

USEPA (US Environmental Protection Agency) (2005) Microbial source tracking guide document. Office of Research and Development, Washington, DC EPA-600/R-05/064, 131p.

Walk ST, Alm EW, Calhoun LM, Mladonicky JM, Whittam TS (2007) Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ Microbiol* 9: 2274-2288.

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS (2009) Cryptic lineages of the genus *Escherichia*. *Appl Environ Microbiol* 75: 6534-6344.

Figure 1 – Sampling sites location in the Watershed Management Units

Figure 2 – Distribution of *E. coli* phylogenetic subgroups in rivers and reservoirs. The circle size is proportional to phylogenetic subgroup frequency.

Figure 3 – Dendrogram by UPGMA cluster analysis method

Figure 4 – Dendrogram showing water body clusters according to *w-cliques*

Table 1 – Sampling sites from rivers and reservoirs and phylogenetic subgroups distribution

Short name	River or reservoir	Source of pollution	average WQI	average TSI	Main land use	geographical coordinates	Number of isolates	Phylogenetic subgroup						
								A ₀	A ₁	B1	B ₂	B ₂ ₃	D ₁	D ₂
BILL2801	Billings Reservoir – beach in front of a Wastewater Treatment Plant	Human	bad	hypereutrophic	Urban area	23°46'37''S/ 46°32'01''W	51	13	22	5	1	7	1	2
BILL2251	Billings Reservoir – pier at Engineer Institute camp area	Human	acceptable	eutrophic	Urban area	23°44'46''S/ 46°38'25''W	53	13	14	5	2	18	0	1
GUAR0502	Guarapiranga Reservoir – Castelo country club	Human	bad	eutrophic	Urban area	23°42'53''S/ 46°42'58''W	52	14	14	4	0	15	3	2
GUAR0601	Guarapiranga Reservoir – Odair restaurant	Human	bad	eutrophic	Urban area	23°41'57''S/ 46°44'41''W	49	11	17	5	2	9	3	2
TIET3120	Tiete River – downstream to a WTP	Animal	bad	hypereutrophic	Urban and industrial area	23°30'11''S/ 46°20'13''W	52	17	11	6	2	9	4	3
TGDE0900	Tanque Grande Reservoir – near to Guarulhos	Animal	good	oligotrophic	Water source protection area	23°22'38''S/ 46°27'35''W	51	19	20	2	3	4	3	0
AGUA2800	Aguapei River – city of Junqueiropolis	Animal	good	mesotrophic	Agricultural and livestock area	21°13'15''S/ 51°29'52''W	54	25	0	12	0	4	6	7
JAMI2100	Jaguari-Mirim River – close to a farm	Animal	good	mesotrophic	Agricultural area	22°04'56''S/ 46°43'13''W	22	17	3	2	0	0	0	0
TIET2050	Tiete River – near to the river's spring	Pristine	good	mesotrophic	Green belt area	23°33'54''S/ 46°00'57''W	46	18	8	4	1	11	3	1
PTEI0900	Paratei River – in an environmental protection area	Pristine	good	mesotrophic	Environmental protection area	23°12'14''S/ 46°00'50''W	44	27	10	3	1	2	0	1
IPIR0018	Ipiranga River – inside an environmental protection area	Pristine	very good	ultraoligotrophic	Environmental protection area	23°20'9.4''S/ 45°08'1.4''W	39	26	0	2	0	2	3	6
PBAL0014	Pau de Bala Stream – inside an environmental protection area	Pristine	very good	ultraoligotrophic	Environmental protection area	23°19'57''S/ 45°07'9.6''W	23	9	0	0	0	1	9	4

APÊNDICE B – *Paper Complex Networks V* (CompleNet2014) – (1ª autora)

Artigo aceito para apresentação no CompleNet2014 e
publicado no *Journal Studies in Computational Intelligence* #

***Social Network Analysis Metrics and Their Application in Microbiological
Network Studies***

Studies in Computational Intelligence Volume 549, 2014, pp 251-260

Available in: http://link.springer.com/chapter/10.1007/978-3-319-05401-8_24

Social Network Analysis metrics and their application in microbiological network studies

Juliana Saragiotto Silva, Nancy de Castro Stoppe, Tatiana Teixeira Torres, Laura Maria Mariscal Ottoboni and Antonio Mauro Saraiva

Abstract In the last decade, several researchers have been using interaction networks resources to investigate of the role of biologic interactions in biodiversity maintenance. The conceptual foundations are the same as in Social Networks (such as Facebook), that have brought a set of metrics to study the network structure and the function of each node in the network. Thus, the aim of this work was to assess the application of Social Network Analysis (SNA) concepts and metrics in microbiological interaction networks, to identify patterns of cohesive subgroups, besides discovering new knowledge regarding the underlying structure of subgroups. We built a bipartite microbiological interaction database containing frequency of phylogenetic subgroups in water bodies and applied the following SNA metrics: dependence distribution, strength, betweenness centrality and clique. The *sna* package for the *R* program, *Pajek*, *Dieta* and *Ucinet* programs were the tools used. Among the results, we found that SNA concepts and metrics are extremely useful in microbiological studies to understand the correlation between each node in the network (the generalist and the predominant nodes), as well as to analyze the co-occurrence pattern of microorganisms in the network (cohesive subgroups).

Juliana S. Silva · Nancy C. Stoppe · Tatiana T. Torres · Antonio M. Saraiva
Research Center on Biodiversity and Computing (BioComp-USP), Universidade de São Paulo (USP), São Paulo, SP, Brazil

Juliana S. Silva
Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT), Cuiabá, MT, Brazil, email: juliana.silva@cba.ifmt.edu.br

Nancy C. Stoppe · Laura M. M. Ottoboni
Centro de Biologia Molecular e Engenharia Genética (CBMEG), Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil, email: ncstoppe@gmail.com, ottoboni@unicamp.br

Tatiana T. Torres
Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo (USP), São Paulo, SP, Brazil, email: tttores@ib.usp.br

Antonio M. Saraiva
Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica, Universidade de São Paulo (USP), São Paulo, SP, Brazil, email: saraiva@usp.br

1 Introduction

The concepts, metrics and tools commonly used in interaction network studies have their foundations in Social Network Analysis (SNA), which uses the graph theory concepts, computing techniques and resources, to analyze the network structure and its interactions [26]. The SNA is considered one of the broadest applications in the network science field to research human relationships and connections [16]. Among the challenges that posed in this area are: (i) how networks improve, weaken or transform; and (ii) how to tackle the dynamic process of changes in the network structure. These same challenges arise in several biological contexts, such as in pollination studies [3, 29], that use interaction network analysis as a resource for studying the factors that have contributed and influenced biodiversity maintenance.

In the ecological field, interaction network is defined as a set of species (pollinators and plants) which are connected by means of links, which represent interactions [3]. Thus, researchers apply the algorithms, metrics and software commonly available in the SNA area to identify the role of each species in the network, as well as the properties of the network structure (e.g. connectance, number of species at each trophic level, among others). Furthermore, in the microbiological context, network analysis has become increasingly common in the last decade, and has been used in protein interactions with other macromolecules, such as carbohydrates, nucleic acids, lipids and other chemical molecules (metabolites and drugs) [31]. Besides, network analysis can also be used with microorganisms in more complex data, as phylogenetic relationships [22], microbial source tracking [21], soil bacterial diversity [25], health and disease variants [10, 19], and microbiota diversity [2, 22].

Concerning microbial indicators, they have been used to survey water quality as surrogates to detect the presence of pathogens of fecal origin. However, these indicators do not provide any information about the origin of fecal pollution, i.e., whether the host source is human, cattle or birds, or even a combination of these [14]. Phylogenetic groups were proposed as a screening tool in source tracking due to rapidity and simplicity [9]. Nevertheless, new approaches should be used to evaluate pollution sources relationships. Considering this limitation, we deem that network analysis is an important tool for addressing this issue, particularly as it has not been used yet in interactions between phylogenetic subgroups and environmental sites. This proof-of-principle study will also provide a framework to evaluate the potential use of network metrics in more complex biological problems.

Therefore, our aim is to assess the application of SNA concepts and metrics in microbiological interaction networks, to identify patterns of cohesive subgroups, besides discovering new knowledge regarding the underlying structure of subgroups.

2 Material and Methods

2.1 Sample description

Phylogenetic subgroups (A_0 , A_1 , B_1 , B_{2_2} , B_{2_3} , D_1 and D_2) described by [11] and [12] were identified from *Escherichia coli* strains isolated from twelve water bodies with different pollution levels in the State of São Paulo as previously described by [27]. The phylogenetic subgroups were used as biomarkers in the rivers and reservoirs studied, which are Tietê River (TIET2050 and TIET3120), Paratei River (PTEI2900), Ipiranga River (IPIR0018), Pau de Bala Stream (PBAL0014), Aguapei River (AGUA2800), Jaguari Mirim River (JAMI2100), Tanque Grande Reservoir (TGDE0900), Billings Reservoir (BILL2801 and BILL2251) and Guarapiranga Reservoir (GUAR0502 and GUAR0601).

These data were organized in a bipartite microbiological interaction database (water bodies x phylogenetic subgroups) – available as part of a project of the Research Center on Biodiversity and Computing (BioComp-USP), under development by the authors. This dataset is composed of a weighted matrix (phylogenetic group abundance), instead of a binary matrix, in which rows were represented by water bodies, and columns by phylogenetic subgroups; therefore, each cell contains a positive integer representing the frequency of occurrence of a phylogenetic subgroup in the corresponding water body.

2.2 SNA metrics

The SNA metrics was the method used to analyze the role of each node on the microbial interaction network (species level); it was chosen because it allows identifying the following characteristics of the network nodes: *dependence*, *strength*, *betweenness centrality* and *w-clique* (cohesive subgroups) – which are described below.

In a weighted network, the *dependence* of i node on j node (d_{ij}) is considered the proportion of all interactions of the i node with j node [17]. Therefore, the formula used is $d_{ij} = N_{ij}/N_i$, where N_{ij} represents the number of interactions observed between node i and j , and N_i the total number of interactions identified to node i [18].

The interaction *strength* of the j node comprises the sum of all dependencies of that j node with each i node ($strength_{(j)} = \sum d_{ij}$) and represents the effect of one j node with the total population of the network [30, 28]. For instance, in plant-pollinator mutualism networks, it is usually measured as the contribution of a pollinator to the maintenance of the plant species [28]. In our network, is the effect of one j phylogenetic subgroup in the network structure.

The *betweenness centrality* metric allows analyzing how vital a node is in the network; values near 1 indicate the node that has the bridge function in the network (if it is removed, any node would be disconnected from the other nodes in the network), and 0 indicate nodes with low betweenness centrality (if the node is removed, the network structure would not be altered) [16]. The calculation of this metric is based on the geodesic distances (the shortest paths between a given pair of nodes in the network) and it is described in [13].

The last metric applied was the *w-clique* that allows identifying the cohesive subgroups (clusters) in the network structure; it is composed of a set of vertices (nodes) that are connected to each other by strong interactions, i.e., the weights of which are higher than the average network weight [1]. This metric is an alternative approach to the *clique*²³ metric, because it takes into account the interaction abundance (weighted matrices) to identify the network clusters.

2.3 Tools applied to calculate the metrics

To calculate the SNA metrics in the network, we used the following tools: *sna* package for R program (The R Foundation for Statistical Computing), *Pajek*, *Dieta* and *Ucinet* programs.

The *sna* package [8] has a set of tools for Social Network Analysis that allows calculating the metrics related to the nodes level in the network, such as *dependence* and *strength*; from this metrics we can analyze the role of each node in the interaction networks.

The *Pajek*²⁴ program [4] – Slovene word for spider – is a computational tool used for performing large network analysis and drawing the network structure in graph form [24]. In this context, it was used to calculate the *betweenness centrality* metric and its representation in graph form.

The *Dieta*²⁵ program [1] allows analyzing individual specialization and identifying cohesive subgroups (*w-cliques*) in weighted networks; it is based on the complex network theory. It is used in combination with *Ucinet*²⁶ [7] – a software that has incorporated a range of SNA metrics, such as cohesive subgroups and measure of centrality [6]. We firstly used the *Dieta* program to verify the nodes that have been connected to each other by strong interactions and, in sequence, we applied the resulting matrix in the *Ucinet* program to identify the *w-cliques*.

²³ A clique is composed by a set of three or more vertices (nodes) totally connected to each other on the network [24].

²⁴ Available online for free download at <<http://pajek.imfm.si/doku.php>>.

²⁵ Available online for free download at <<http://esapubs.org/archive/ecol/E089/115/Dieta1.exe>>.

²⁶ Available online for free download in <<http://www.analytictech.com/downloaduc6.htm>>. A Ucinet tutorial is available at [15].

2.4 Data analysis

From the bipartite microbiological database (phylogenetic subgroups frequency in water bodies), available in an excel spreadsheet, we organized the data, sorted the rows and columns in descending order of degree (number of connections of each node), as recommended by [23], in order to draw the standard graphs. Following these spreadsheets, we prepared two matrices formats and saved in text files: (i) the first, in which labels of columns and rows are represented, besides the values of frequencies – to be used in the *R* program (*matrix_r.txt*); and (ii) the second, in which only the row labels and values of frequencies remain to be used in the *Dieta* program (*matrix_dieta.txt*).

Firstly, we use the matrix (prepared to the *R* program) and applied the *sna* package to calculate the node level metrics, specially the *dependence* distribution and the *strength* of each node in the network. Furthermore, this matrix was also used in the *Pajek* program to draw the network graphs, and to calculate and plot the *betweenness centrality* metric values in graph form.

Another application performed in these networks was the use of the *Dieta* program, in order to identify niche overlap (cohesive subgroups). More details about the parameters used in the *Dieta* program can be seen in its manual, available in the supplementary section of [1]. The output file generated by *Dieta* corresponds to a binary matrix representing the interactions with strong connections, i.e., interactions whose values were greater than or equal to the network average (*w-clique*). This matrix was opened with the *Pajek* program to transform it into an undirected network, because it was a niche overlap network. Finally, we used the *Ucinet* program for the identification of cliques and for drawing the clusters by means of a dendrogram.

3 Results and Discussion

Among the results, by analyzing *dependence* distribution and *strength* metrics of the network nodes (Table 1), it was possible to identify the generalist nodes (e.g., phylogenetic subgroups occurring in most of the water body nodes) and the predominant nodes (e.g., which phylogenetic subgroups were important for certain water body groups). In this case, Table 1 represents the *dependence* distribution of each water body by phylogenetic subgroups, and the phylogenetic subgroups strength metric (sum of the dependencies of all *i* water bodies by each *j* phylogenetic subgroup). Thus, it can be observed that although the BILL2801 reservoir has a positive value of dependence for all phylogenetic subgroups, it has the highest *dependence* of A_1 (0.4314), because this phylogenetic subgroup is the most frequent in this reservoir.

Table 1. Dependence distribution of water bodies by phylogenetic subgroups, and phylogenetic subgroups strength metric

Dependence of water bodies by phylogenetic subgroups ^(a)	A ₀	B ₁	B _{2₃}	D ₂	A ₁	D ₁	B _{2₂}
BILL2801	0.2549	0.0980	0.1373	0.0392	0.4314	0.0196	0.0196
GUAR0601	0.2245	0.1020	0.1837	0.0408	0.3469	0.0612	0.0408
TIET2050	0.3913	0.0870	0.2391	0.0217	0.1739	0.0652	0.0217
TIET3120	0.3269	0.1154	0.1731	0.0577	0.2115	0.0769	0.0385
BILL2251	0.2453	0.0943	0.3396	0.0189	0.2642	0.0000	0.0377
GUAR0502	0.2692	0.0769	0.2885	0.0385	0.2692	0.0577	0.0000
TGDE0900	0.3725	0.0392	0.0784	0.0000	0.3922	0.0588	0.0588
PTEI2900	0.6136	0.0682	0.0455	0.0227	0.2273	0.0000	0.0227
AGUA2800	0.4630	0.2222	0.0741	0.1296	0.0000	0.1111	0.0000
IPIR0018	0.6667	0.0513	0.0513	0.1538	0.0000	0.0769	0.0000
PBAL0014	0.3913	0.0000	0.0435	0.1739	0.0000	0.3913	0.0000
JAMI2100	0.7727	0.0909	0.0000	0.0000	0.1364	0.0000	0.0000
Phylogenetic subgroup strength ^(b)	4.9920	1.0455	1.6539	0.6969	2.4529	0.9188	0.2399

Notes:

^(a) The sum total of the dependence values of each water body should be equal to 1. High value of water body dependence (closer to 1) comprises higher association with the phylogenetic subgroup; whilst low value of water body dependence (closer to 0) consists of lower association with the phylogenetic subgroup.

^(b) The strength metric represents the abundance of phylogenetic subgroups (based on their frequency values) in each water bodies; thus, high strength values can demonstrate the nodes that are more generalists or more predominant in the microbiological network. $\text{Strength}(j) = \sum d(ij)$.

In our network, A₀ was the phylogenetic subgroup with the highest *strength* (4.9920). This subgroup has the highest abundance in the water bodies sampled and showed the highest interaction frequency values. However, not necessarily, the most frequent and ubiquitous node displays the highest strength. For instance, the phylogenetic subgroup A₁ which is not present in some reservoirs (AGUA2800, IPIR0018 and PBAL0014) has the second highest *strength* in the network (2.4529). This higher *strength* resulted from the highest *dependence* values in those water bodies in which it is present (as in BILL2801, TGDE0900 and GUAR0601). From this analysis, we can conclude that the *strength* metric is directly related to the presence of association (interaction) and with the *dependence* value. Moreover, by means of the *dependence* and *strength* metrics, it is possible to note the role of each network node and then analyze the network tolerance regarding to the extinction/inclusion of a node – similarly to analyzing the potential of an actor in Social Networks studies, as performed by [20].

Another result found in this study refers to the *betweenness centrality* metric. As it can be seen in Figure 1, this metric allows verifying that the phylogenetic subgroups have a central role in the network – specially A₀, B₁, B_{2₃} and D₂ –, the most centralized in the graph and with the largest circles). An analogy to Social

Networks would be to know the most popular people in the network (those that have many friends). In our context, by means of this metric we can verify that the water body interactions have been established through the phylogenetic subgroups, as expected. An additional result to support this kind of analysis is the graphical representation of the metric values, as can be done with the *Pajek* program.

Fig. 1. *Betweenness centrality* graph – phylogenetic subgroups (A₀, A₁, B₁, B₂, B₃, D₁, D₂) and water bodies (AGUA2800, BILL2251, BILL2801, GUAR0502, GUAR0601, IPIR0018, JAMI2100, PBAL0014, PTEI2900, TGDE0900, TIET2050, TIET3120)

Once the relationship between water bodies and phylogenetic subgroups was confirmed (by means of *dependence*, *strength* and *betweenness centrality* metric), we found a metric that can support the identification of phylogenetic subgroup co-occurrence patterns in water bodies, from a clustering method. The first alternative was the use of statistical methods – commonly applied for cluster identification – such as Correspondence Analysis (CA) and Unweight Par Group Method with Arithmetic mean (UPGMA). As described by [27], the CA did not recover any consistent grouping. The UPGMA clustering of the dissimilarities split the nodes into two groups, but these groups did not reflect features that are of interest for water studies, such as geographic location, source of pollution or abiotic factors.

Finally, by means of *cliques* identification, it was possible to discover new knowledge regarding a simple interactions database, such as grouping patterns of water bodies (in protected environments and contaminated environments), based on the abundance of phylogenetic subgroups. From the identification of network *cliques* (nodes connected by stronger interactions – greater than or equal to the network average, as showed in Table 2), we could find the cohesive subgroups in the network and it was an innovative contribution for this research area.

It is noteworthy that there is a set of metrics (measures) for network analysis. However, the choice should take into account the specific problem and the dataset. Besides, it is possible to analyze at the network level (network structure) or at the

node level (the role of one node in the network) – this also depends on the context analysis. More details about a set of metrics can be seen in [5, 29].

Table 2. The output file ‘B[file_name].mat’ generated by Dieta program that corresponds to a binary matrix representing the presence (1) or absence (0) of the interactions with strong connections

*Vertices	12	*Matrix	1	2	3	4	5	6	7	8	9	10	11	12
1	BILL2801	1	0	1	1	1	1	1	1	1	0	0	0	0
2	GUAR0601	2	1	0	1	1	1	1	1	0	0	0	0	0
3	TIET2050	3	1	1	0	1	1	1	1	1	1	0	0	0
4	TIET3120	4	1	1	1	0	1	1	1	1	1	0	0	0
5	BILL2251	5	1	1	1	1	0	1	1	1	0	0	0	0
6	GUAR0502	6	1	1	1	1	1	0	1	1	0	0	0	0
7	TGDE0900	7	1	1	1	1	1	1	0	1	0	0	0	0
8	PTEI2900	8	1	0	1	1	1	1	1	0	0	1	0	1
9	AGUA2800	9	0	0	1	1	0	0	0	0	0	1	1	0
10	IPIR0018	10	0	0	0	0	0	0	0	1	1	0	1	1
11	PBAL0014	11	0	0	0	0	0	0	0	0	1	1	0	0
12	JAMI2100	12	0	0	0	0	0	0	0	1	0	1	0	0

Note: The left side of this table (a) shows the vertices (nodes) of the network and the right side of this table (b) demonstrate the binary matrix. This matrix gets the result of processing of Dieta program from the bipartite microbiological interaction database (containing the distribution of phylogenetic subgroups in these water bodies). In this case, the nodes between 1 and 7 represent the “contaminated environments” and, the others, the “protected environments”.

Another finding in this research concerning computational tools is that there is a trend towards having graphical and user-friendly interfaces, based on open source codes and having extensive documentation, supporting data import/export to other software, metrics calculation and interactions representation in graph form. As an example, we can mention the *Pajek* program that allows a set of tools to be accessed only through menu options (such as *betweenness centrality*) and also provides a set of graphs for visualizing the results (networks).

4 Conclusion

From this research, we found that SNA concepts and metrics can be used as a tool in microbiological studies, allowing the visualization of phylogenetic subgroups and water bodies associations. For instance, they can help to identify pollution sources (*w-clique*), higher/lower association between water bodies and phylogenetic subgroups (*dependence*), abundance of phylogenetic subgroups in each water bodies (*strength*) and, moreover, the node that has the bridge function

in the network. However, is important to emphasize the support of experts in Biodiversity Informatics and microbiological fields, to identify which kind of new knowledge would be generated with biological significance.

In conclusion, a considerable part of the microbiological Interaction Networks studies requirements, especially microbial genetics, concerns the need to identify subgroups (clusters) and co-occurring microorganisms – that have already been exploited in the Social Networks and Ecological Networks field. Thus, these findings will guide further developments of the application of SNA tools and metrics to microbiological studies.

Acknowledgments This research was developed in the Research Center on Biodiversity and Computing, of the University of São Paulo (BioComp-USP). JSS is a PhD student supported by CAPES, FAPEMAT and IFMT. NCS is a fellowship FAPESP, TTT and LMMO are researcher fellowship CNPq. The authors thank Dr. Marco A. R. Mello (UFMG) and his Guide for Ecological Networks Analysis (<http://marcomello.casadosmorcegos.org/Redes.html>) for helping with the data analysis.

References

1. Araújo, M.S., Guimarães, P.R., Svanbäck, R., Pinheiro, A., Guimarães, P., Reis, S.F., Bolnick, D.I.: Network analysis reveals contrasting effects of intraspecific competition on individual vs. population diets. *Ecology* 89: 1981-1993 (2008)
2. Bapteste, E., Bicep, C., Lopez, P.: Evolution of genetic diversity using networks: the human gut microbiome as a case study. *Clin. Microbiol. Infect.*, 18: 40-43 (2012)
3. Bascompte, J., Jordano, P.: Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.*, 38, 567-93 (2007)
4. Batagelj, V., Mrvar, A.: Pajek – program for large network analysis. *Connections* 21: 47-57 (1998)
5. Blüthgen, N., Fründ, J., Vázquez, D.P., Menzel, F.: What do interaction network metrics tell us about specialization and biological traits. *Ecology* 89:3387-3399 (2008) <http://dx.doi.org/10.1890/07-2121.1>.
6. Borgatti, S.P., Everett, M.G., Freeman, L.C.: UCINET 5 for Windows: Software for Social Network Analysis (USER'S GUIDE) (1999). Available at: <http://www.analytictech.com/ucinet6/Ucinet_Guide.doc>. Accessed in: 15 Nov 2012.
7. Borgatti, S.P., Everett, M.G., Freeman, L.C.: Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies (2002)
8. Butts, C.T.: Social Network Analysis with sna. *Journal of Statistical Software*, 24, 1-51 (2008)
9. Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I.Z., Gomes, T.A.T., Amaral, L.A., Ottoboni, L.M.M.: *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiol* 10, 161 (2010)
10. Chen, M., Cho, J., Zhao, H.: Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.*, 7: e1001353 (2011)
11. Clermont, O., Bonacorsi, S., Bingen, E.: Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* 66: 4555-4558 (2000)
12. Escobar-Páramo, P., Grenet, K., LeMenac'h, A., Rode, L., Salgado, E., Amorin, C., Gouriou, S., Picard, B., Rahimy, C., Andreumont, A., Denamur, E., Ruimy, R.: Large-scale

- population structure of human commensal *Escherichia coli* isolates. *Appl Environ Microbiol* 70: 5698-5700 (2004)
13. Freeman, L.C.: Centrality in social networks: conceptual clarification. *Social Networks*, 1(3): 215-239 (1978/1979)
 14. Hagedorn, C., Blanch, A.R., Harwood, V.J.: *Microbial source tracking: methods, application, and case studies*. New York: Springer, 642p. (2011)
 15. Hanneman, R.A., Riddle, M.: *Introduction to social network methods*. Riverside, CA: University of California, Riverside, (published in digital form at <<http://faculty.ucr.edu/~hanneman/>>) (2005)
 16. Hansen, D.L., Shneiderman, B., Smith, M.A.: *Analysing social media networks with NodeXL: insights from a connected world*. Morgan Kaufmann, Amsterdam. 284 p. (2011)
 17. Jordano, P.: Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution. *American Naturalist* 129: 657-677 (1987)
 18. Jordano, P., Vázquez, D., Bascompte, J.: Redes complejas de interacciones mutualistas planta-animal. *In*: R. Mendel, M.A. Aizen and R. Zamora (eds). *Ecología y evolución de interacciones planta-animal*. 1ª ed. – Santiago de Chile: Universitaria. p. 17-41 (2009)
 19. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y. et al.: Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491: 119-124 (2012)
 20. Kumar, K., Novak, J., Tomkins, A.: Structure and evolution of online social networks. *In*: *Proc. of ACM SIGKDD Intl. Conf. of Knowledge Discovery and Data Mining*, p.611-617, New York (2006)
 21. Lee, J.E., Lee, S., Sung, J., Ko, G.P.: Analysis of human and animal fecal microbiota for microbial source tracking. *The ISME J.*, 5: 362-365 (2011)
 22. Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, S., Schlegel, M., Tucker, T.A., Schrenzel, M.D., Knight, R., Gordon, J.I.: Evolution of mammals and their gut microbes. *Science*, 320: 1647-1651 (2008)
 23. Mello, M.A.R.: *Guia para Análise de Redes Ecológicas*. Version March 3 (2013). Available at: <<http://marcomello.casadosmorcegos.org/Redes.html>> Accessed in: 01 Apr 2013.
 24. Nooy, W., Mrvar, A., Batagelj, V.: *Exploratory Network Analysis with Pajek*. Cambridge University Press, 334p. (2005)
 25. Roesch, L.F.W., Fulthorpe, R.R., Pereira, A.B., Pereira, C.K., Lemos, J.N., Barbosa, A.D., Suleiman, A.K.A., Gerber, A.L., Pereira, M.G., Loss, A., Costa, E.M.: Soil bacteria community abundance and diversity in ice-free areas of Keller Peninsula, Antarctica. *Appl. Soil Ecol.*, 61: 7-15 (2012)
 26. Scott, J.: *Social Network Analysis: A handbook*. 2nd edition. London: Sage (2000)
 27. Stoppe, N.C., Silva, J.S., Torres, T.T., Carlos, C., Hachich, E.M., Sato, M.I.Z., Saraiva, A.M., Ottoboni, L.M.M.: Clustering of water bodies in unpolluted and polluted environments based on *Escherichia coli* phylogroup abundance using a simple interaction database, *BMC Microbiology*. (unpublished).
 28. Vázquez, D.P., Morris, W.F., Jordano, P.: Interaction frequency as a surrogate for the total effect of animal mutualists on plants. *Ecology Letters* 8: 1088-1094 (2005)
 29. Vázquez, D.P., Blüthgen, N., Cagnolo, L., Chacoff, N.P.: Uniting pattern and process in plant-animal mutualistic networks: a review. *Annals of Botany* 103: 1445-1457 (2009)
 30. Wootton, J.T., Emmerson, M.: Measurement of interaction strength in nature. *Annual Review of Ecology, Evolution and Systematics* 36:419-444 (2005)
 31. Yamada, T., Bork, P.: Evolution of biomolecular networks – lessons from metabolic and protein interactions. *Nature Rev. Mol. Cell Biol.*, 10: 791-803 (2009)

APÊNDICE C – *Paper SpringerPlus* (2ª autora – equal contribution)

[*Draft version – unpublished*]

***Complex pattern of Escherichia coli phylogenetic group distribution in
worldwide samples of human feces and wastewater treatment plants***

Complex pattern of *Escherichia coli* phylogenetic group distribution in worldwide samples of human feces and wastewater treatment plants

Nancy de Castro Stoppe^{1,2} • Juliana Saragiotto Silva^{2,3} • Raúl Jiménez Rosenberg⁴ • Camila Carlos¹ • Elayse Maria Hachich⁵ • Maria Inês Zanoli Sato⁵ • Antonio Mauro Saraiva⁶ • Laura Maria Mariscal Ottoboni¹ • Tatiana Teixeira Torres^{2,7}

¹: Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brasil

²: Núcleo de Pesquisa em Biodiversidade e Computação (BioComp-USP) - Universidade de São Paulo (USP), São Paulo, SP, Brasil

³: Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT), Cuiabá, MT, Brasil

⁴: Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), México, D.F., México

⁵: Departamento de Análises Ambientais, Companhia Ambiental do Estado de São Paulo - CETESB, São Paulo, SP, Brasil

⁶: Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da USP, São Paulo, SP, Brasil

⁷: Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo (USP), São Paulo, SP, Brasil

Corresponding author: T.T.Torres

Departamento de Genética e Biologia Evolutiva - IBUSP, São Paulo – SP, Brasil

E-mail: tttorres@ib.usp.br

Running title: Phylogroups in feces and sewage

31	
32	ABSTRACT
33	
34	

INTRODUCTION

Escherichia coli (*E.coli*) is the predominant aerobic microorganism in the gastrointestinal tract of warm-blooded animals and has been extensively used as a microbial indicator of fecal contamination. Among the tools for studying *E.coli* population genetics, phylogrouping triplex PCR has been widely applied due to its simplicity and rapidity (Clermont et al. 2000; Tenaillon 2010). This method can quickly assign *E. coli* isolates to one of the four major phylogenetic groups: A, B1, D, or B2 allowing its use for different purposes including population genetics, classification in extra-intestinal pathogenic and commensal strains and host source relationships (Carlos et al. 2010, Duriez et al., 2001; Escobar-Páramo et al. 2004; Gordon et al. 2008; Orsi et al. 2007; Tenaillon et al., 2010; Unno et al. 2009).

The method uses the combination of two genes (*chuA* and *yjaA*) and a DNA fragment that has been identified as part of a putative lipase esterase gene⁷¹ (Clermont et al., 2000; Tenaillon et al., 2010).

Studies with PG and host source have been conducted and showed human commensal strains belonging to group A (Escobar-Paramo et al. 2006; Li et al. 2010) while strains isolated from animals are in group B1 (Carlos et al. 2010; Higgins et al. 2007; Ishii et al. 2007) suggesting an association between phylogenetic groups and host sources.

Phylo-groups D and B2 were predominant in *E.coli* isolated from wastewater samples even from different climate zones, where about 50% of isolates were classified in these groups in a temperate area and more than 95% in a subtropical region (Anastasi et al. 2010; Mokracka et al. 2011).

The presence of *E.coli* in aquatic environments indicates that the water has been contaminated with fecal material (human and other animals). Reference libraries are built using isolates taken from fecal fresh samples. Nevertheless, some researchers have been used wastewater isolates as surrogates for human contamination because they assume that this type of isolates are more representative of strains found in the environment and due to no necessity of ethical committee approval (Ahmed et al. 2009; Duran et al. 2006; Fremaux et al. 2009; Jiang et al. 2007; Kaneene et al. 2007; Kelty et al. 2012; Silkie and Nelson, 2009; USEPA, 2005).

Analysis of human commensal *E.coli* population structure in different climate regions showed that subjects in tropical areas exhibit more-diverse *E.coli* microbiota than those in temperate areas. Besides the geography and climate, other factors contribute to this difference, such as diet and hygiene level. Isolates from the phylogenetic groups A and B1 were prevalent in tropical areas (Escobar-Páramo et al. 2004).

Other factors might be correlated to phylogenetic distribution in commensal *E.coli* isolates, such as geographic and climate factors. Bailey et al. (2010a) studied worldwide phylogroups distribution, but they did not evaluated this relationship with geographic location, neither distance nor location in the same continent. According to Escobar-Páramo et al. (2004) analysis, the geographical and climatic location of human populations can have an important role in structuring the *E.coli* PG distribution. These authors observed that French population are distinct from American and African ones, but no further information trying to correlate distance were studied. They also observed the 10 different locations was separated in two clusters, tropical and temperate zones. However, no specific climate classification was used not allowing separation in stricter climate group. Li et al. (2010) compared human commensal *E.coli* isolated from China with those studied by Escobar-Páramo et al. (2004) and observed despite China belongs to temperate belt, it has PG distribution similar to tropical one. Again, no stricter climate classification was used.

There is not a consensus if microorganisms exhibit any biogeographic patterns and these studies of microbial biogeography have been done mainly with free-living microorganisms (Martiny et al, 2006; Fierer, 2008). Papke and Ward (2004) states that in a permanent symbiotic association between a microorganism and a host, it is possible the symbiont evolves in parallel with its host. However, no information could be found for commensal relationships between microorganisms and their hosts.

Although dispersal limitation and selection are the main processes influencing the distribution of free-living microbes, the host microenvironment also play a major role in structuring commensal isolates. Hence, we tested for the presence of discontinuity against the null hypothesis that the distribution of phylogenetic groups is limited by dispersal across space.

The aims of this work were to evaluate the phylogenetic distribution in *E.coli* strains isolated from humans and wastewater samples. We also compared phylogenetic distribution between isolates from human and wastewater samples. Finally, we performed meta-analysis in PG distribution from human *E.coli* worldwide isolates, looking for patterns of association between the commensal isolates and host geographical location or climate or host feeding habits or host living area that might indicate local adaptation.

MATERIAL AND METHODS

***E.coli* isolation from human feces**

Feces from 116 adult humans were collected using sterile swabs and Cary-Blair transport medium. They were streaking on Endo agar LES (Difco) and incubated for 24 h at 35 °C. Three lactose-positive colonies (pink to dark-red with a metallic surface sheen) were picked from each sample and tested for citrate utilization, lactose fermentation, oxidase, L-lysine decarboxylase, motility, glucose and sucrose fermentation, tryptophan deamination, indole production, urea hydrolysis and sulfide production. One typical *E. coli* profile strain from each individual host was reisolated on nutrient agar, incubated for 24 h at 35 °C and kept at -70 °C in tryptic soy broth (Difco) with 10% glycerol (v/v) for further analysis (ATCC 2010).

The Research Ethics Committee of the State University of Campinas School of Medical Sciences has approved the present study (Permission 049/09) and all participants gave their informed written consent. Human samples were collected from healthy human distributed almost equally in males and females living in São Paulo Metropolitan area. The age interval of the subjects ranged from 19 to 79 years old (average = 43), with BMI (body mass index) ranging from 17.5 to 40.9 (average = 26). The majority of them (>89%) have omnivore food habitats.

***E.coli* isolation from raw wastewater**

Two to three collections were done from five wastewater treatment plants (WTP) São Paulo Metropolitan area (Parque Novo Mundo, Jesus Netto, Barueri, São Lourenço da Serra and Vinhedo). All the wastewater treatment plants received domestic discharges, and Parque Novo Mundo and Barueri plants also received industrial discharges. Raw wastewater samples were collected in sterilized bottles according to Standard Methods (APHA, 2010) and samples were analyzed using the membrane filter technique according to U.S. Environmental Protection Agency method 1603 (USEPA 2002) as previously described by Stoppe et al. (2014).

DNA isolation and Phylogenetic grouping (PG)

Genomic DNA from the *E.coli* strains was isolated with the Wizard Genomic DNA Purification Kit (Promega) according the manufacturer's instructions. The PG of *E.coli* strains was determined by triplex PCR as previously described by Clermont et al. (2000) and Stoppe et al. (2014).

Meta-analysis of human isolates

Meta-analysis was performed using human data (Table 1). Data from commensal *E.coli* strains isolated from healthy humans was analysed using Koppen climate classification (Peel *et al.*, 2007), feeding habits (western or other), living area (urban or rural), and geographical distance (km).

a) Analysis of Variance (ANOVA)

The ANOVA model was performed to analyzing the variance of PG data distribution and the interaction among other factors (city, country, Koppen climate classification, feeding habits and living area). It was used the function “*aov*” available in the “*stats*” package from R program (R Development Core Team, 2012).

b) Exploratory data analysis (EDA)

In order to perform EDA, firstly we used boxplot to see the data distribution, in sequence apply correlation analysis using scatterplot, and finally tested a probability distribution model.

Data distribution visualization through the bloxplot

Visualization of PG data distribution was done using the graphical technique notched *boxplot* by means of “*boxplot*” function available in the “*graphics*” package (R Development Core Team, 2012) from R program; in this work we used the parameter “*notch*” as “*true*” in order to draw a notch in each side of boxes. As described by Chambers *et al.* (1983), when you compare two plots, this resource facilitates the visual identification of ‘strong evidence’ that the two medians differ only seeing if the two plots do not overlap.

Scatterplot for correlation analysis

We applied the “*pairs*” function, available in the “*graphics*” package (R Development Core Team, 2012) from R program, with the following arguments: (1) “*upper.panel*” – the correlation analysis between variables; (2) “*diag panel*” – the histogram function, representing the frequency of distribution of the data; and (3) “*lower.panel*” – the smooth function. In this case, the variables were displayed in the scatterplot matrix with the ordering which was in the dataset.

Probability Distribution Model: Poisson and Gamma Distribution

First of all, it was performed the Poisson test. To calculate the values of mean and variance we use the *ddply* function (Wickham, 2011) available in the “*plyr*” package from R program.

As the probability distribution model of the PG data was not known, we search a more general probability model distribution such as *Gamma*. It was used the “*shape*” and “*rate*” parameters, due to the optimization algorithm applied (maximum-likelihood estimation - MLE) is covered by “*fitdistr*” function available in the “*mass*” package from R (RIPLEY *et al.*, 2013). In sequence, it was generated a sample of theoretical *Gamma* distribution (Cumulative Distribution Function – CDF) for the parameters “*shape*” and “*rate*” using the “*pgamma*” function available in the package “*stats*” (R Development Core Team, 2012) from R program. Finally, the Kolmogorov-Smirnov test (K-S test) was performed, using the “*ks.test*” function also available in the package “*stats*”, to compare each PG data distribution (empirical CDF) with a theoretical *Gamma* distribution (*Gamma* CDF) to indicate their similarity, though the *p-value* attribute; how smaller the *p-value*, more similar the distributions are.

c) Identification of patterns (Cliques and data mining)

Social Network Analysis metric

W-clique metric was used to identify the cohesive subgroups (clusters) in the network (Araujo *et al.*, 2008). As described by Stoppe *et al.*, (2014) and Silva *et al.* (2014), the *Dieta* program was firstly used to verify the nodes that have been connected to each other by strong interactions – represented by one binary matrix (0/1), in which cells containing the number 1 represented interactions whose weights were higher than the average network weight (*w-cliques*) (Araujo *et al.*, 2008). In sequence, the *Pajek* program was used to transform the network from arcs to edges (Batagelj and Mrvar 1998). Finally, it was applied the resulting matrix in the *Ucinet* program (Borgatti *et al.*, 2002) to identify the *w-cliques* though the cluster diagram (dendrogram).

Data Mining Classification

The algorithm of Data Mining Classification used was the “*Decision Tree*” It was used the package “*rpart*” from R program, to constructed the regression tree; we uses different values for training and

testing data to try to achieve a lower value of the “*misclassification error*” – the percentage of data that the tree does not classify properly (Therneau et al., 2013).

d) Multivariate analysis

The statistical analysis was performed based in two methods of multivariate analysis: Mantel test (Oksanen, 2013) and Clustering Visualizations of Multidimensional data (Hurley, 2004).

Mantel test

The Mantel test was applied to analyze the correlation analysis between two dissimilarity matrices, based on Pearson’s product-moment correlation. The dissimilarity matrices were calculated with the “*vegdist*” function, using the community ecology package “*vegan*” in the R program (Oksanen, 2013), with the following coefficients as parameters: Bray-Curtis (abundance data) and Jaccard (presence-absence data) (Legendre e Legendre, 2012).

Clustering Visualizations of Multidimensional Data

In order to perform the Clustering Visualizations of Multidimensional Data method, the dissimilarity matrices of phylogenetic group distribution (raw data) were built using the Bray-Curtis coefficient (Legendre and Legendre, 2012). The similarity matrices were prepared as a complement to the dissimilarity matrices (1-vegdist (matrix, “Bray”)).

The permutation of the variables (hierarchical clustering order) was based on the classification of the data – i.e. Koppen climate classification, feeding habits, living area, and geographic distance, which was displayed by the graphical method *scatterplot* matrix, which is commonly used for displaying multivariate data.

RESULTS

PG distribution in human isolates

One hundred and sixteen strains isolated from different individuals. The PG distribution observed was 48.3% belongs to group A, 5.2% of B1, 16.4% of B2 and 30.2% of D (Table 1). No significant differences were observed among PG distribution and gender, age or BMI.

219

220 *PG distribution in wastewater isolates*

221 One hundred and fifty strains were isolated from raw wastewater samples. Of these strains, 44% belonged
 222 to group A, 39% to group D and 9% and 8% to groups B2 and B1, respectively.

223 The PG distribution in wastewater was similar to humans, in which groups A and D were predominant.

224

225 *Comparison of distribution of phylogroups between human and wastewater isolates*

226 There was a strong positive correlation between the PG distribution in wastewater and human hosts
 227 (Mantel statistic $r = 0.607$, $P = 0.046$).

228

229 *Comparison of worldwide PG distribution in humans and wastewater*

230 We also have selected additional published data on PG distribution from WTP and human commensal
 231 isolates from Australia, Portugal, Spain and United States (Anastasi et al. 2010; Bailey et al. 2010a,b;
 232 Boczek et al. 2006; Figueira et al. 2011; Hannah et al. 2009; Johnson et al. 2005; Logue et al. 2012;
 233 Machado et al. 2005; Moreno et al. 2009; Sabate et al. 2008; Sannes et al. 2004; Silva et al. 2010;
 234 Valverde et al. 2009; Vollmerhausen et al. 2011; Zhang et al. 2002). However, no significant correlation
 235 was observed in Australia (Mantel statistic $r = 0.003$, $P = 0.454$), Portugal, (Mantel statistic $r = -0.396$, P
 236 $= 0.600$), Spain (Mantel statistic $r = -0.023$, $P = 0.575$) or United States (Mantel statistic $r = 0.070$, $P =$
 237 0.467).

238

239 *Meta-analysis of human isolates*

240 We compared our distribution of *E.coli* phylogenetic groups with commensal isolates from other world-
 241 wide studies. Data were selected according to the following criteria: (1) *E.coli* were isolated from feces
 242 from healthy individuals (both sexes), (2) one isolate per individual or single representative of strains was
 243 obtained, (3) the classification in PG was done according to Clermont et al. (2000), (4) approximated
 244 geographical localization (city or region) where the strains were obtained and recorded. Some data were
 245 reanalyzed in two or more studies but we were careful not to repeat these datasets. We combined data of
 246 isolates from the same city or region (Table 1 and Figure 1). We also used the geographical coordinates
 247 and local climate as parameters to discover patterns in the PG distribution.

248 Exploratory data analysis was performed for analyzing data, and among the graphical techniques, *boxplot*
 249 and *scatterplot* were selected. Figure 2 shows notched *boxplot* of PG distribution. A was the most

common type with a median of 36.1%. It was followed by PG D (median 21.5%), B2 (median of 20%) and B1 (median of 16.4%). The widest frequency range was observed to the PG A which was from 8.7 to 79.3%, while PG B2 ranged from 0 to 53.9%, D from 0 to 45.7% and B1 from 0 to 34%. PG A and B1 apparently had the most different frequencies while B2 and D had the most common frequencies and PG B1 and B2 presented outliers results.

Scatterplot with the correlation analysis among PG and other factor can be seen in Figure 3. A strong negative correlation was observed between groups A and B2, but no other strong significant correlations were observed among phylogroups (Figure 3a). When new variables are added, for instance, feeding habits, living area, or geographical location, the correlation among phylogroups did not change, suggesting the other variables are independent (Figure 3b).

PG data distribution did not have a discrete probability distribution, due to the values of mean and variance were not equal, which means it did not present a Poisson distribution. The probabilistic data distribution model for all phylogenetic groups in human data, using the optimization algorithm maximum-likelihood estimation (MLE), was *Gamma* distribution (Figure 4).

Trying to cluster PG distribution according to geographical location and climate, we used *w-clique* metric. However, clusters identified in the dendrogram did not show any cluster (cohesive subgroup) that showed geographical or climate classification meaning (Figure 5a and b). We also analyzed more than one factor simultaneously: geographical location and feeding habits; geographical location and living area; geographical location, feeding habits and living area; climate and feeding habits; climate and living area, climate feeding habits and living area. Also, none of them showed any cluster that showed some meaning (data not shown).

Data mining classification using decision tree algorithm was applied trying to determine if the frequency of specific PG could determine the pattern of host living area (urban or rural). It was used different values for training and testing data in order to reach the lowest value of for the misclassification error. Although it found a value of 11% of misclassification error, in practice, when we cross-validating with our data, less than 70% of data fitted correctly. We believed that this occurred due to the algorithm used automatically selects only part of the data that will participate in the construction of the tree (some PGs) and so it can choose any PG that is not present everywhere and therefore does not represent the entire sample space, generating an misclassification error higher than expected – depending on the nature of data. Examples of decision trees generated from some tests can be seen on Figure 6.

We observed a weak correlation between geographical location and PG distribution (Mantel statistic $r = 0.2059$, $P = 0.007$) and moderate correlation between climate and PG distribution (Mantel statistic $r = 0.399$, $P = 0.001$).

The multivariate analysis using the clustering visualizations of multidimensional data method also did not show any relationship between PG and feeding habits, living area, Koppen climate groups, and geographical distance (Figures 7, 8, 9 and 10).

Hurley (2004) proposed that the use of categorical variables with few levels could bias the results showing inaccurate clusters. As feeding habits and living areas have only two classes each, western and other diets and urban and rural areas respectively, it might explain the obtained results (Figures 7 and 8). However, Koppen climate and geographical distance have several categories and despite only micro-clusters were observed, in human commensal *E.coli* was not related to either climate or geographical distance (Figures 9 and 10).

DISCUSSION

Commensal populations of *E.coli* have stable genetic isolates with very low rates of recombination leading to clonal population structure and allowing the characterization of major phylogenetic groups (Selander and Levin 1980; Tenaillon 2010), which means as a rule, each host has only one PG. As we isolated the *E.coli* strains from 116 different individuals we have a representative PG distribution in human adults from São Paulo city.

The phylogenetic group frequency in human commensal *Escherichia coli* strains was first presented by Duriez et al. (2001) where the phylogenetic groups A and B1 were the most common and the rarest was B2. However, Bailey et al. (2010a) compared distribution of human commensal *E.coli* phylogenetic groups and, as observed in our study, the majority of them showed group A predominance and underrepresentation of B1 type.

We observed a higher frequency of the groups A and D were predominant in wastewater samples, despite whereas previous studies found that either B2 and D groups were predominant in wastewater (Anastasi et al. 2010; Mokracka et al. 2011) or A and B1 groups (Figueira et al. 2011). This observed difference might suggest that geographical location and climate play a role in determining PG distribution, since environmental strains seems to be more susceptible to these factors. Some authors have suggested group A as the most environmentally adapted (Skurnik et al, 2008; Anastasi et al, 2012). The PG distribution we observed in wastewater samples from Brazil are in line with some of the previously published data;

similar distributions were observed in Portugal (Figueira et al, 2011) and Spain (Sabate et al, 2008), but not in Australia (Anastasi et al, 2010) and United States (Boczek et al, 2006).

Sewage samples provide bacterial isolates that have been used to surrogate human feces samples (USEPA, 2005). *E.coli* commensal isolates collected in Australia, Portugal, Spain and United States presented PG distribution did not show significant correlation with PG distribution in isolates from WTPs samples. Conversely, our data, both human and wastewater samples presented a strong positive correlation reinforcing the use of isolates from urban wastewater treatment plants as surrogates for human fecal contamination. Our samples, feces and wastewater, were collected from the same area (São Paulo Metropolitan region) and at the same period, which was not observed for other studies. If it is true, the use of sewage samples as surrogates to humans may be limited.

Duriez et al. (2001) proposed that studies with *E. coli* commensal strains should include geographic, socioeconomic and medical information since these enteric isolates are believed to be the reservoirs of pathogenic strains. We compiled data from the literature that used Clermont phylogenetic groups (Clermont et al, 2000), trying to find out if any PG pattern might be related to climate, feeding habits, living area and/or geographical location. Some data are available in the literature relating Clermont phylogroups with climate (Escobar-Páramo et al, 2004; Gordon et al, 2005; Li et al, 2010), living area (Nowrouzian et al, 2005, 2009; Karami, 2007; Pallechi et al, 2007; Vinue et al, 2008; Bartolini et al, 2009; Hannah et al, 2009; Bailey et al, 2010b; Nicolas-Chanoine et al, 2013), and feeding habits (Gordon et al, 2005). However, the comparisons took into account only qualitative differences and none of them have made statistical analysis.

Boxplot results presented PG A and B1 without overlapping frequencies and can indicate their differential distribution worldwide, in opposite what was observed for PG B2 and D,

The negative correlation observed for PG A and B2 might be related to their phylogenetic evolution history. PG B2 was the most ancestral PG while A and B1 is the freshest ones (Gordon et al, 2008). However, PG B1 had neither a positive correlation with PG A nor a negative correlation with PG B2.

Gamma distribution was the probabilistic distribution model best fitted in our data, confirmed by performing the calculation of “*shape*” and “*rate*” parameters (using the optimization algorithm MLE) and running the K-S test (to identify the *p-value*). As Gamma is a continuous distribution model more generalist it is usually identified in contexts similar to that in which the data pattern is not known. The problem is that a generalist model not giving much information to describe the PG data.

We also try to cluster PG distribution using SNA using w-clique metrics, It was successfully used by Stoppe et al (2014) in rivers samples, which was enabling in differentiation of polluted and unpolluted sites. Unfortunately, any meaning cluster neither geographical nor climate was observed in present data.

A promising tool for simple classification is a decision tree (algorithm for performing data Mining Classification). Unfortunately, our data did not present a good level of confidence making data mining difficult to use, in this context.

Mantel test is widely used in Ecological studies trying to correlate genetic markers and geographical distance. Notwithstanding, meta-analysis data showed a weak correlation between geographical location and PG distribution using Mantel test. It suggests the distance between the sites might not influence the PG distribution. Some authors proposed the influence of climate where the human hosts were living and the *E.coli* PG distribution (Escobar-Páramo et al, 2004; Gordon et al, 2005; Li et al, 2010). Our analysis showed a moderate correlation between climate and PG distribution suggesting this factor can have more influence on PG distribution than geographical distance.

We also did a meta-analysis comparison using the clustering multivariate analysis and PG did not show sub structuring with none of parameters evaluated, even with geographic distance and climate, which have more levels of classification. Clustering multivariate analysis has a high potential for comparison data, but as far as we know, none reports are available in the literature being the present study a seminal work using this innovative approach. The Clustering Visualizations of Multidimensional Data is a graphical resource applied for build a hierarchical cluster analysis on the similarity matrix, permuting the variables in a more effective arrangement instead of simply using variables in their ordering in the dataset (Hurley, 2004). The innovation of this cluster analysis method is reordering the variables available regarding some merit index (like single link cluster, Pearson correlation, Euclidean distance, among others) to build the clusters, which does not occur with the methods that using the standard variable order (in the dataset).

The absence of patterns observed to geographic distance or climate might be related to the used tool. Clermont phylogroups might not have the enough resolution to reveal the patterns, since they have only four groups. Recently, Clermont et al (2013) proposed the use of eight phylogroups instead of the regular four types improving their specificity and detection. When databases using the eight phylogroups are available, they can also be tested using multivariate analysis.

Microbial biogeography studies should record data from different distances, habitats and environmental conditions, to better understand its field area. Biogeographic theories used for macroorganisms should not

be applied to microorganisms, since the biodiversity at the microbial level is fundamentally different from that of macroorganisms (Martiny et al, 2006; Fierer, 2008)

Our spatial autocorrelation analysis did not reveal any specific patterns. There is substructuring of the PG distribution but geographical distance does not limit *E. coli* dispersion. For free-living microorganisms, dispersion over large distances is extremely rare, but for commensal microbes it can make much easier because they can travel with their hosts. However, no studies on microbial biogeography related to commensal microorganisms and their hosts are available in literature. The absence of patterns observed with the use of different tools (including the multivariate analysis) for PG distribution leads us to infer that no pattern exists. To confirm this, other studies using other marker tools and with other commensal relationships should be conducted.

ACKNOWLEDGMENTS

São Paulo Research Foundation (FAPESP) (Project 2007/55312-6) supported this work. NCS received a fellowship from FAPESP (Project 2010/15222-0). LMMO and TTT received research fellowships from National Counsel of Technological and Scientific Development (CNPq). JSS is a PhD student supported by CAPES, FAPEMAT, and IFMT. The Research Center on Biodiversity and Computing (BioComp) was supported by the USP Provost's Office for Research.

REFERENCES

- Abdul-Razzaq, M.S. and Abdul-Lateef, L.A. (2011) Molecular phylogeny of *Escherichia coli* isolated from clinical samples in Hilla, Iraq. *African J Biotechnol* **10**, 15783-15787.
- Ahmed, W., Goonetilleke, A., Powell, D., Chauhan, K. and Gardner, T. (2009) Comparison of molecular markers to detect fresh sewage in environmental waters. *Water Res* **43**, 4908-4917.
- Anastasi, E.M., Matthews, B., Gundogdu, A., Vollmerhausen, T.L., Ramos, N.L., Stratton, H., Ahmed, W. and Katouli, M. (2010) Prevalence and persistence of *Escherichia coli* strains with uropathogenic virulence characteristics in sewage treatment plants. *Appl Environ Microbiol* **76** 5788-5786.
- Anastasi, E.M., Matthews, B., Stratton, H. and Katouli, M. (2012) Pathogenic *E.coli* found in sewage treatment plants and environmental waters *Appl Environ Microbiol* **78** 5536-5541.

- 403 APHA (American Public Health Association) (2010) Samples. In: _____ *Standard methods for*
 404 *examination of water and wastewater*. Section 9060. Available at: <<http://standardmethods.org>>.
 405 Accessed 18 Feb 2010.
- 406 Araújo, M.S., Guimarães, P.R., Svanbäck, R., Pinheiro, A., Guimarães, P., Reis, S.F., Bolnick, D.I.
 407 (2008) Network analysis reveals contrasting effects of intraspecific competition on individual vs.
 408 population diets. *Ecology* **89** 1981-1993.
- 409 ATCC (American Type Culture Collection) (2010) *Microbial media and supplements*. Available at:
 410 <[http://www.atcc.org/CulturesandProducts/Microbiology/MicrobialMediaandSupplements/tabid/186/Def](http://www.atcc.org/CulturesandProducts/Microbiology/MicrobialMediaandSupplements/tabid/186/Default.aspx)
 411 [ault.aspx](http://www.atcc.org/CulturesandProducts/Microbiology/MicrobialMediaandSupplements/tabid/186/Default.aspx)>. Accessed 07 Mar 2010.
- 412 Bailey, J.K., Pinyon, J.L., Annantham, S. and Hall, R.M. (2010a) Distribution of human commensal
 413 *Escherichia coli* phylogenetic groups. *J Clin Microbiol* **48**, 3455-3456.
- 414 Bailey, J.K., Pinyon, J.L., Annantham, S. and Hall, R.M. (2010b) Commensal *Escherichia coli* of healthy
 415 humans: a reservoir for antibiotic-resistance determinants. *J Med Microbiol* **59** 1331-1339.
- 416 Bartoloni, A., Palecchi, L., Rodriguez, L., Fernandez, C., Mantella, A., Bartalesi, F., Strohmeyer, M.,
 417 Kristiansson, C., Gotuzzo, E., Paradisi, F. and Rossolini, G.M. (2009) Antibiotic resistance in a remote
 418 Amazonas community. *Int J Antimicrob Agents* **33** 125-129.
- 419 Batagelj, V., Mrvar, A. (1998) Pajek – program for large network analysis. *Connections* **21** 47–57.
- 420 Boczek. L. Johnson, C.H., Rice, E.W. and Kinkle, B.K (2006) The widespread occurrence of the
 421 enterohemolysin gene *ehlyA* among environmental strains of *Escherichia coli*. *FEMS Microbiol Lett* **254**
 422 281-284.
- 423 Borgatti, S.P., Everett, M.G., Freeman, L.C. (2002) Ucinet for Windows: Software for Social Network
 424 Analysis. Harvard, MA: Analytic Technologies.
- 425 Carlos, C., Aires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I.Z., Gomes, T.A.T., Amaral, L.A. and
 426 Ottoboni, L.M.M. (2010) *Escherichia coli* phylogenetic group determination and its application in the
 427 identification of the major animal source of fecal contamination. *BMC Microbiol* **10** 161.
- 428 Chatfield, C. (1995) Problem Solving: A Statistician's Guide. Chapman and Hall, 2nd edition.
 429 ISBN 0412606305.

- 430 Clermont, O., Bonacorsi, S. and Bingen, E. (2000) Rapid and simple determination of the *Escherichia*
 431 *coli* phylogenetic group. *Appl Environ Microbiol* **66** 4555-4558.
- 432 Clermont, O., Christensonm J.K., Denamur, E., Gordon, D.M. (2013) The Clermont *Escherichia coli*
 433 phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ*
 434 *Microbiol Rep* **5** 58-65.
- 435 Damborg, P., Nielsen, S.S. and Guardabassi, L. (2009) *Escherichia coli* shedding patterns in humans and
 436 dogs: insights into within-household transmission of phylotypes associated with urinary tract infections.
 437 *Epidemiol Infect* **137** 1457-1464.
- 438 Duran, M., Haznedaroglu, B.Z. and Zitomer, D.H. (2006) Microbial source tracking using host specific
 439 FAME profiles of fecal coliforms. *Water Res* **40** 67-74.
- 440 Duriez, P. Clermont, O., Bonacorsi, S., Bingen, E., Chaventre, A., Elion, J., Picard, B. and Denamur, E.
 441 (2001) Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically
 442 distinct human populations. *Microbiology* **147** 1671-1676.
- 443 Escobar-Páramo, P., Grenet, K., Menac'h, A.L., Rode, L., Salgado, E., Amorin, C., Gouriou, S., Picard,
 444 B., Rahimy, M.S., Andremont, A., Denamur, E. and Ruimy, R. (2004) Large-scale population structure of
 445 human commensal *Escherichia coli* isolates. *Appl Environ Microbiol* **70** 5698-5700.
- 446 Fierer, N. (2008) Microbial biogeography: patterns in microbial diversity across space and time. In:
 447 *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back*. K.
 448 Zengler (editor). ASM Press, Washington DC,pgs. 95-115.
- 449 Figueira, V., Serra, E. and Manaia, C.M. (2011) Differential patterns of antimicrobial reistance in
 450 population subsets of *Escherichia coli* isolated from waste- and surface waters. *Sci Total Environ* **409**
 451 1017-1023.
- 452 Fremaux, B., Gritzfeld, J. and Yost, C.K. (2009) Evaluation of host-specific *Bacteroidales* 16S rRNA
 453 gene markers as a complementary tool for detecting fecal pollution in a prairie watershed. *Water Res* **43**
 454 4838-4849.
- 455 Gordon, D.M., Stern, S.E. and Collignon, P.J. (2005) Influence on the age and sex of human hosts on the
 456 distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology* **151** 15-23.

- 457 Gordon, D.M., Clermont, O., Tolley, H. and Denamur, E. (2008) Assigning *Escherichia coli* strains to
 458 phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* **10**
 459 2484-96.
- 460 Grasselli, E., François, P., Gutacker, M., Gettler, E., Benagli, C., Convert, M., Boerlin, P., Schrenzel, J.
 461 and Piffaretti, J.C. (2009) Evidence of horizontal gene transfer between human and animal isolates
 462 *Escherichia coli* strains identified by microarray. *FEMS Immunol Med Microbiol* **53** 351-358.
- 463 Grude, N., Potaturkina-Nesterova, N.I., Jenkins, A., Strand, L., Nowrouzian, F.L., Nyhus, J. and
 464 Kristiansen, B.E. (2007) A comparison of phylogenetic group, virulence factors and antibiotic resistance
 465 in Russian and Norwegian isolates of *Escherichia coli* from urinary tract infection. *Clin Microbiol Inf* **13**
 466 208-211.
- 467 Hannah, E.L., Johnson, J.R., Angul, F., Haddadin, B., Williamson, J. and Samore, M.H. (2009) Molecular
 468 analysis of antimicrobial-susceptible and -resistant *Escherichia coli* from retail meats and human stool
 469 and clinical specimens in a rural community setting. *Foodborne Pathog Dis* **6** 285-295.
- 470 Harada, K., Okada, E., Shimizu, T., Takaoka, Y., Sawada, T. and Takahashi, T. (2012) Antimicrobial
 471 resistance, virulence profiles, and phylogenetic groups of *Escherichia coli* isolates: A comparative
 472 analysis between dogs and their owners in Japan. *Comp Immunol Microbiol Inf Dis* **35** 139-144.
- 473 Hurley, C.B. (2004) Clustering Visualizations of Multidimensional Data. *Journal of Computational and*
 474 *Graphical Statistics*, **13**:4 788-806.
- 475 Jakobsen, I., Kurbasic, A., Skjot-Rasmussen, L., Ejrnaes, K., Porsbo, L.J., Pedersen, K., Jensen, L.B.,
 476 Emborg, H.D., Agerso, Y., Olsen, K.E.P., Aarestrup, F.M., Frimodt-Moller, N. and Hammerum, A.M.
 477 (2010) *Escherichia coli* isolates from broiler chicken meat, broiler chickens, pork, and pigs share
 478 phylogroups and antimicrobial resistance with community-dwelling humans and patients with urinary
 479 tract infection. *Foodborne Pathog Dis* **7** 537-547.
- 480 Jiang, S.C., Chu, W., Olson, B.H., He, J.-W., Choi, S., Zhang, J., Le, J.Y. and Gedalanga, P.B. (2007)
 481 Microbial source tracking in a small southern California urban watershed indicates wild animals and
 482 growth as the source of fecal bacteria. *Appl Microbiol Biotechnol* **76** 927-934.
- 483 Johnson, J.R., Owens, K., Gajewski, A. and Kuskowski, M.A. (2005) Bacterial characteristics in relation
 484 to clinical sources of *Escherichia coli* isolates from women with acute cystitis or pyelonephritis and
 485 uninfected women. *J Clin Microbiol* **43** 6064-6072.

- 486 Kanamuru, S., Kurazono, H., Nakano, M., Teral, A., Ogawa, O. and Yamamoto, S. (2006) Subtyping of
 487 uropathogenic *Escherichia coli* according to the pathogenicity island encoding uropathogenic-specific
 488 protein: Comparison with phylogenetic groups. *Int J Urol* **13** 754-760.
- 489 Kaneene, J., Miller, R.A., sayah, R., Johnson, Y.J., Gilliland, D. and Gardiner, J.C. (2007) Considerations
 490 when using discriminant function analysis of antimicrobial resistance profiles to identify sources of fecal
 491 contamination of surface water in Michigan. *Appl Environ Microbiol* **73** 2878-2890.
- 492 Karami, N. (2007) Antibiotic resistance and fitness of *Escherichia coli* in the infantile commensal
 493 microbiota. PhD thesis, University of Gothenburg, 99p. Available at: <http://hdl.handle.net/2077/4418>.
- 494 Kelty, C.A., Varma, M., Sivaganesan, M., Haugland, R.A. and Shanks, O.C. (2012) Distribution of
 495 genetic marker concentrations for fecal indicator bacteria in sewage and animal feces. *Appl Environ*
 496 *Microbiol* **78** 4225-4232.
- 497 Lee, S., Yu, J.K., Park, K., Oh, E.J., Kim, S.Y. and Park, Y.J. (2012) Phylogenetic groups and virulence
 498 factors in pathogenic and commensal strains of *Escherichia coli* and their association with *bla*_{CTX-M}. *Ann*
 499 *Clin Lab Sci* **40** 361-367.
- 500 Legendre, P., Legendre, L. (2012) Numerical Ecology. Oxford: Elsevier, 3rd edition, pages 265-335.
- 501 Li, B., Sun, J.Y., Han, L.Z., Huang, X.H., Fu, Q. and Ni, Y.X. (2010) Phylogenetic groups and
 502 pathogenicity islands markers in fecal *Escherichia coli* isolates from asymptomatic humans in China.
 503 *Appl Environ Microbiol* **76** 6698-6700.
- 504 Logue, C.M., Doetkott, C., Mangiamale, P., Wannemuehler, Y.M., Johnson, T.J., Tivendale, K.A., Li, G.,
 505 Sherwood, J.S. and Nolan, L.K. (2012) Genotypic and phenotypic traits that distinguish neonatal-
 506 meningitis-associated *Escherichia coli* from fecal *E.coli* isolates from healthy human hosts. *Appl Environ*
 507 *Microbiol* **78** 5824-5830.
- 508 Luo, Y., Cui, S., Li, J., Yang, J., Lin. L., Hu, C., Lin, S., Ye, L., Zhao, Q. and Ma, Y. (2011)
 509 Characterization of *Escherichia coli* isolates from healthy food handlers in hospital. *Microb Drug Resist*
 510 **17** 443-448.
- 511 Machado, E., Canton, R., Baquero, F., Galan, J.C., Rollan, A., Peixe, L. and Coque, T.M. (2005) Integron
 512 content of extended-spectrum- β -lactamase-producing *Escherichia coli* strains over 12 years in a single
 513 hospital in Madrid, Spain. *Antimicrob Agents Ch* **49** 1823-1829.

- 514 Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-
 515 Devine, M.C, Kane, M., Krumins, J.A., Kuske, C.R., Morin, P.J., Naeem, S., Ovreas, L., Reysenbach,
 516 A.L., Smith, V.L and Staley, J.T. (2006) Microbial biogeography: putting microorganisms on the map.
 517 *Nat. Rev. Microbiol.*, **4** 102-112.
- 518 Mereghetti, L., Tayoro, J., Watt, S., Lanotte, P., Loulergue, J., Perrotin, D. and Quentin, R. (2002)
 519 Genetic relationship between *Escherichia coli* strains isolated from the intestinal flora and those
 520 responsible for infections diseases among patients hospitalized in intensive care units. *J Hosp Inf* **52** 43-
 521 51.
- 522 Mokracka, J., Kokzura, R., Jablonska, L. and Kaznowski, A. (2011) Phylogenetic groups, virulence genes
 523 and quinolone resistance of integron-bearing *Escherichia coli* strains isolated from a wastewater
 524 treatment plant. *Antonie van Leeuwenhoek* **99** 817-824.
- 525 Moreno, E., Johnson, J.R., Perez, T., Pratts, G., Kuskowski, M.A. and Andreu, A. (2009) Structure and
 526 urovirulence of the fecal *Escherichia coli* population among healthy women. *Microb Inf* **11** 274-280.
- 527 Nicolas-Chanoine, M.-H., Gruson, C., Bialek-Davenet, S., Bertrand, X., Thomas-Jean, F.T., Bert, F.,
 528 Mayat, M., Meiller, E., Marcon, E., Danchin, N., Noussair, L., Moreau, R., Leflon-Guibout, V. (2013) 10-
 529 fold increase (2006-11) in the rate of healthy subjects with extended-spectrum β -lactamase-producing
 530 *Escherichia coli* faecal carriage in a Parisian check-up centre. *J Antimicrob Chemother* **68** 562-568.
- 531 Nooy, W., Mrvar, A., Batagelj, V. (2005) Exploratory Network Analysis with Pajek. Cambridge
 532 University Press, 334p.
- 533 Nowrouzian, F.L., Wold, W.E. and Adlerberth, I. (2005) *Escherichia coli* strains belonging to
 534 phylogenetic group B2 have superior capacity to persist in the intestinal microflora to infants. *J. Infec. Dis*
 535 **191**, 1078-1083.
- 536 Nowrouzian, F.L., Adlerberth, I. and Wold, W.E. (2006) Enhanced persistence in the colonic microbiota
 537 of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to
 538 colonic cells. *Microbes Infect* **8** 834-840.
- 539 Nowrouzian, F.L., Ostblom, A.E., Wold, W.E. and Adlerberth, I. (2009) Phylogenetic group B2
 540 *Escherichia coli* strains from the bowel microbiota of Pakistani infants carry few virulence genes and lack
 541 of capacity for long-term persistence. *Clin Microbiol Infect* **15**, 466-472.

- Obata-Yasuoka, M., Ba-Thein, W., Tsukamoto, T., Yoshikawa, H. and Hayashi, H. (2002) Vaginal *Escherichia coli* share common virulence factor profiles, serotypes and phylogeny with other extraintestinal *E.coli*. *Microbiology* **148** 2745-2752.
- Oksanen, J. (2013) Multivariate Analysis of Ecological Communities in R: vegan tutorial. Version February 08, 2013. Available at: <<http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf>> Accessed in: July 29, 2013.
- Orsi, R.H., Stoppe, N.C., Sato, M.I.Z., Gomes, T.A.T., Prado, P.I., Manfio, G.P. and Ottoboni, L.M.M (2007) Genetic variability and pathogenicity potential of *Escherichia coli* isolated from recreational water reservoirs. *Res Microbiol* **158** 420-427.
- Pallechi, L., Lucchetti, C., Bartolini, A., Bartalesi, F., Mantella, A., Gamboa, H., Carattoli, A., Paradisi, F. and Rossolini, G.M. (2007) Population structure and resistance genes in antibiotic-resistant bacteria from a remote community with a minimal antibiotic exposure. *Antimicrob Agents Ch* **51** 1179-1184.
- Papke, R.T. and Ward, D.M. (2004) The importance of physical isolation to microbial diversification *FEMS Microbiol Ecol* **48** 293-303.
- Peel, M.C., Finlayson, B.L. and McMahon, T.A. (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci* **11** 1633-1644.
- Petersen, A.M., Nielsen, E.M., Litrup, E., Brynskov, L., Mirsepasi, H. and Krogfelt, K.A. (2009) A phylogenetic group of *Escherichia coli* associated with active left-sided inflammatory bowel disease. *BMC Microbiol* **9** 171.
- R Development Core Team and contributors worldwide (2012) *The R Graphics Package: R functions for base graphics*. Version: 2.15.0. Publication on: 2012-03-30. Available at: <R-core@r-project.org>. Accessed 07 Apr 2014.
- Riccobono, E., Pallechi, L., Mantella, A., Bartalesi, F., Zeballos, I.C., Trigoso, C., Villagran, A.L., Bartolini, A. and Rossolini, G.M. (2012) Carriage of antibiotic-resistant *Escherichia coli* among healthy children and home-raised chickens: a household study in a resource-limited setting. *Microb Drug Resist* **18** 83-87.

- 568 Ripley, B., Venables, B., Hornik, K., Gebhardt, A., Firth, D. (2013) *Support Functions and Datasets for*
 569 *Venables and Ripley's MASS*. Version 7.3-23. Publication on: 2013-01-08. Available at:
 570 <<http://www.stats.ox.ac.uk/pub/MASS4/>>. Accessed 05 Apr 2014.
- 571 Sabate, M., Prats, G., Moreno, E., Balleste, E., Blanch, A.R. and Andreu, A. (2008) Virulence and
 572 antimicrobial resistance profiles among *Escherichia coli* strains isolated from human and animal
 573 wastewater. *Res Microbiol* **159** 288-293.
- 574 Sannes, M.R., Kuskowski, M.A., Owens, K., Gajewski, A. and Johnson, J. (2004) Virulence factors profile
 575 and phylogenetic background of *Escherichia coli* isolates from veterans with bacteremia and uninfected
 576 control subjects. *J Inf Dis* **190** 2121-2128.
- 577 Selander, R.K. and Levin, B.R. (1980) Genetic diversity and structure in *Escherichia coli* populations.
 578 *Science* **210** 545-547.
- 579 Silkie, S. and Nelson, K.L. (2009) Concentrations of host-specific and generic fecal markers measured by
 580 quantitative PCR in raw sewage and fresh animal feces. *Water Res* **43**, 4860-4871.
- 581 Silva, J.S., Stoppe, N.C., Torres, T.T., Ottoboni, L.M.M., Saraiva, A.M. (2014) Social Network Analysis
 582 Metrics and Their Application in Microbiological Network Studies. In: 5th International Workshop on
 583 Complex Networks (CompleNet2014), 2014, Bologna, Italy. Proceedings..., *Studies in Computational*
 584 *Intelligence* **549**, 251-260.
- 585 Skurnik, D., Bonet, D., Bernède-Baudin, C.; Michel, R., Baleire, C., Chau, F., Mohler, J., Jarlier, V.,
 586 Boutin, J.-P., Moreau, B., Guilemot, D., Denamur, E., Andreumont, A. and Ruimy, R. (2008)
 587 Characteristics of human intestinal *Escherichia coli* with changing environments. *Environ. Microbiol* **10**
 588 2132-2137.
- 589 Sorsa, L.J., Feldmann, F., Hildinger, K., Dufke, S. and Schubert, S. (2007) Characterization of four novel
 590 genomic regions of uropathogenic *Escherichia coli* highly associated with the extraintestinal virulent
 591 phenotype: A jigsaw puzzle of genetic modules. *J Med Microbiol* **297**, 83-95.
- 592 Stoppe, N.C., Silva, J.S., Torres, T.T., Carlos, C., Hachich, E.M., Sato, M.I.Z., Saraiva, A.M. and
 593 Ottoboni, L.M.M. (2014) Clustering of water bodies in unpolluted and polluted environments based on
 594 *Escherichia coli* phylogroup abundance using a simple interaction database. *Genetics and*
 595 *Molecular Biology* (in press).

- 596 Tenaillon, O., Skurnik, D., Picard, B and Denamur, E. (2010) The population genetics of commensal
597 *Escherichia coli*. *Nat Rev Microbiol* **8**, 207-217.
- 598 Therneau, T., Atkinson, B., Ripley, B. (2013) *Recursive partitioning and regression trees*. Version 4.1-1.
599 Publication on: 2013-03-20. Available at: <<http://cran.r-project.org/web/packages/rpart/>>. Accessed 20
600 Apr 2014.
- 601 Unno, T., Han, D., Jang, J., Lee, S.N, Ko, G.P., Choi, H.Y., Kim, J.H., Sadowsky, M.J. and Hur, H.G.
602 (2010) Absence of *Escherichia coli* phylogenetic group B2 strains in human and domesticated animals
603 from Jeonnam Province, Republic of Korea. *Appl Environ Microbiol* **75**, 5659-5666.
- 604 USEPA (US Environmental Protection Agency) (2002) *Method 1603 Escherichia coli (E. coli) in water*
605 *by membrane filtration using modified membrane-thermotolerant Escherichia coli Agar (Modified m-*
606 *TEC)*, EPA 821-R-02-023. Cincinnati: U.S Environmental Protection Agency, Office of Research and
607 Development.
- 608 USEPA (US Environmental Protection Agency) (2005) *Microbial source tracking guide document*.
609 Office of Research and Development, Washington, DC EPA-600/R-05/064, 131p.
- 610 Valverde, A. Canton, R., Garcillan-Barcia, M., Novais, A., Gallan, J.C., Alvarado, A., de la Cruz, F.,
611 Baquero, F. and Coque, T.M. (2009) Spread of *bla*_{CTX-M-14} is driven mainly by IncK plasmids
612 disseminated among *Escherichia coli* phylogroups A, B1 and D in Spain. *Antimicrob Agents Ch* **53**,
613 5204-5212.
- 614 Vinue, L., Sáenz, Y., Somalo, S., Escudero, E., Moreno, M.A., Ruiz-Larrea, F. and Torres, C. (2008)
615 Prevalence and diversity of integrons and associated resistance genes in faecal *Escherichia coli* isolates of
616 healthy humans in Spain. *J Antimicrob Chemother* **62** 934-937.
- 617 Vollmerhausen, T.L., Ramos, N.L., Gundogdu, A., Robinson, W., Brauner, A. and Katouli, M. (2011)
618 Population structure and uropathogenic virulence-associated genes of faecal *Escherichia coli* from
619 healthy young and elderly adults. *J Med Microbiol* **60**, 574-581.
- 620 White, A.P., Sibley, K.A., Sibley, C.D., Wasmuth, J.D., Schaefer, R., Surette, M.G., Edge, T.A. and
621 Neumann, N.F. (2011) Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle
622 adaptations but not host specificity. *Appl Environ Microbiol* **77**, 7620-7632.

623 Wickham, H. (2011) The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical*
624 *Software*, **40**(1), 1-29. Available at: <<http://www.jstatsoft.org/v40/i01/>>. Accessed 05 Apr 2014.

625 Zhang, L., Foxman, B. and Marrs, C. (2002) Both urinary and rectal *Escherichia coli* isolates are
626 dominated by strains of phylogenetic group B2. *J Clin Microbiol* **40**, 3951-3955.

627

628

Table 1 – Distribution of commensal *E.coli* phylogenetic groups isolated from human feces

#	city or country	region,	Koppen climate classification ^a	feeding habits	living area	approximated geographic coordinates	phylogenetic group (%)				Reference
							A	B1	B2	D	
1	São Paulo, Brazil		Aw	Western	Urban	23° 32' S 46° 38' W	48.3	5.2	16.4	30.2	this study
2	Sydney, Australia		Cfa	Western	Urban	33° 52' S 151° 12' E	37.1	9.5	29.5	23.8	Bailey et al. 2010a,b
3	Maroochydore, Australia		Cfa	Western	Urban	26° 39' S 153° 05' E	15.2	10.6	28.5	45.7	Vollmerhausen et al. 2011
4	Cotonou, Benin		Aw	Other	Urban	06° 22' N 02° 26' E	50.0	32.6	17.4	0.0	Escobar-Páramo et al. 2004
5	Alto de los Zarcos, Bolivia		Aw	Other	Rural	21° 28' S 63° 54' W	77.0	10.0	5.0	8.0	Pallechi et al. 2007
6	Villamontes, Bolivia		Aw	Other	Urban	21° 15' S 63° 32' W	79.3	0.0	0.0	20.7	Riccobono et al. 2012
7	São Paulo, Brazil		Aw	Western	Urban	23° 32' S 46° 38' W	40.4	8.5	12.8	38.3	Carlos et al. 2010
8	Calgary, Canada		Dfc	Western	NA	51° 02' N 114° 03' W	13.9	12.0	53.7	20.4	White et al. 2011
9	Fuzhou, China		Cfa	Other	Urban	26° 04' N 119° 17' E	43.7	23.4	16.0	16.9	Li et al. 2010
10	Beijing, China		Dwa	Other	Urban	39° 54' N 116° 24' E	12.0	0.0	47.8	40.2	Luo et al. 2011
11	Bogota, Colombia		Cfb	Western	Urban	04° 35' N 74° 04' W	57.1	3.6	25.0	14.3	Escobar-Páramo et al. 2004
12	Olib and Silba, Croatia		Csa	Western	NA	44° 22' N 14° 43' E	35.1	31.6	19.3	14.0	Duriez et al. 2001
13	Copenhagen, Denmark		Cfb	Western	NA	55° 40' N 12° 34' E	20.5	22.2	29.2	28.1	Damborg et al. 2009; Jakobsen et al. 2010; Petersen et al. 2009; Duriez et al. 2001; Escobar-Páramo et al. 2004; Leflon-Guibout et al. 2008
14	Paris, France		Cfb	Western	Urban	48° 51' N 02° 15' E	48.3	9.0	18.4	24.4	Escobar-Páramo et al. 2004
15	Brittany, France		Cfb	Western	Rural	48° 12' N 02° 55' W	28.0	26.0	24.0	22.0	Escobar-Páramo et al. 2004
16	Brest, France		Cfb	Western	Rural	48° 23' N 04° 29' W	14.3	23.8	33.3	28.6	Escobar-Páramo et al. 2004
17	Tours, France		Cfb	Western	Urban	47° 23' N 00° 37' E	25.0	21.0	29.0	25.0	Escobar-Páramo et al. 2004
18	Western France		Cfb	Western	Urban	48° 06' N 01° 41' W	48.0	12.0	20.0	20.0	Mereghetti et al. 2002
19	National Park, French Guiana		Af	Other	Rural	02° 35' N 53° 33' W	63.4	20.4	3.2	12.9	Escobar-Páramo et al. 2004
20	Munchengladbach, Germany		Cfb	Western	NA ^b	48° 08' N 11° 34' E	43.3	27.0	16.2	13.5	Sorsa et al. 2007
21	Babylon, Iraq		BWh	Other	Urban	32° 28' N 44° 33' E	60.0	30.0	0.0	10.0	Abdul-Razzaq and Abdul-Lateef, 2011; Harada et al. 2012; Obata-Yasuoka et al. 2002
22	Tokyo, Japan		Cfa	Other	Urban	35° 41' N 139° 41' E	16.6	9.4	50.3	23.8	Kanamaru et al. 2006
23	Kyoto, Japan		Cfa	Other	NA	35° 00' N 135° 46' E	42.0	32.0	20.0	6.0	Lee et al. 2010
24	Seoul, Korea		Dwa	Other	Urban	37° 34' N 126° 58' E	38.0	18.0	22.9	21.0	Unno et al. 2010
25	Jeonnam Province, Korea		Dwa	Other	NA	34° 52' N 126° 59' E	29.8	34.0	0.0	36.2	Duriez et al. 2001
26	Bay Reserve, Mali		As	Other	NA	13° 00' N 05° 00' E	23.6	58.2	1.8	16.4	Grude et al. 2007
27	Oslo and Telemark, Norway		Dfb	Western	NA	59° 54' N 10° 45' E	30.0	25.0	15.0	30.0	Nowrouzian et al. 2009
28	Lahore, Pakistan		BSh	Other	Urban	31° 32' N 74° 20' E	47.0	18.0	12.0	23.0	Silva et al. 2010
29	Villa Real, Portugal		Csa	Western	NA	41° 18' N 07° 31' W	8.6	10.3	65.5	15.5	Bartoloni et al. 2009
30	Angaiza, Peru		Af	Other	Rural	03° 35' S 71° 36' W	72.0	17.0	3.0	8.0	Machado et al. 2005; Valverde et al., 2009
31	Madrid, Spain		Csa	Western	Urban	40° 25' N 03° 42' W	39.5	18.4	15.8	26.3	Moreno et al. 2009
32	Barcelona, Spain		Csa	Western	Urban	41° 23' N 02° 10' E	33.0	19.0	17.0	31.0	Karami 2007; Nowrouzian et al. 2005, 2006
33	Göteborg, Sweden		Cfb	Western	NA	57° 42' N 11° 58' E	30.5	12.7	42.5	14.3	Grasselli et al. 2009
34	Geneva and Ticino, Switzerland		Cfb	Western	Urban	46° 05' N 07° 30' E	40.0	10.0	20.0	30.0	Hannah et al. 2009
35	Boise, United States		BSk	Western	Rural	43° 33' N 116° 22' W	26.2	13.9	28.7	31.1	Johnson et al. 2005; Logue et al. 2012; Sannes et al. 2004; Zhang et al. 2002
36	Minneapolis, United States		Dfb	Western	Urban	44° 59' N 93° 22' W	13.6	15.0	53.1	18.4	Zhang et al. 2002
37	Ann Arbor, United States		Dfa	Western	Urban	42° 14' N 83° 44' W	20.5	12.5	47.7	19.3	Logue et al. 2012
38	Fargo, United States		Dfb	Western	Urban	46° 52' N 96° 47' W	16.2	15.7	53.9	14.2

Figure 2 – Notched *boxplot* of meta-analysis PG distribution

Figure 3 – *Scatterplot* showing (a) correlation among phylogroups and (b) correlation among phylogroups, feeding habits, living area and geographical location. In both *scatterplots* the figures are divided in three parts, representing: (1) upper panel – the correlation analysis between the variables; (2) diagonal panel – the histogram of the PG distribution data; and (3) lower panel – the smooth function of the data.

Figure 4 – *Gamma* distribution of meta-analysis phylogenetic distribution. *Gamma* CDF is represented by continuous red line and PG empirical CDF by blue dots. The values of “*shape*” and “*rate*” parameters, discovered from *gamma* distribution to a specific sample space, are showed on y-axis label; and the *p-value* (generated by K-S test) is indicated on x-axis label.

Figure 5 – Dendrogram showing clusters according to w -clique metric, considering PG distribution data and countries (a) or Koppen climate classification (b)

Train 69

Train 90

Train 95

Figure 6 – Application of Data Mining to living area classification using decision tree algorithm. The PG A, B1 and D was selected by algorithm to build the decision tree, when performing 69% rows for training and 31% for testing (a), and 90% for training and 10% for testing (b). Another example was the selection of PG A, B1 and B2 using 95% of rows for training and 5% for testing (c).

Figure 7 – *Scatterplot* matrix showing multivariate analysis between phylogenetic grouping and feeding habits using the Clustering Visualizations of Multidimensional data (Hurley, 2004). Same color in the *scatterplot* clusters shows more similar index value; thus, by means of Bray-Curtis coefficient, the data have the same pattern (“*shape*” and “*rate*”) of phylogroup distribution values. In this case, the components display close to the *scatterplot* matrix diagonal consist the highly related variables.

Figure 8 – *Scatterplot* matrix showing multivariate analysis between phylogenetic grouping and living area using the Clustering Visualizations of Multidimensional data (Hurley, 2004). Same color in the *scatterplot* clusters shows more similar index value; thus, by means of Bray-Curtis coefficient, the data have the same pattern (“*shape*” and “*rate*”) of phylogroup distribution values. In this case, the components display close to the *scatterplot* matrix diagonal consist the highly related variables. Thus, we note the presence of a micro-cluster (cohesive subgroup) among countries with identifier 33 and 38; however, as encompassing sites from different parts of the world, we cannot notice a significant pattern.

Figure 9 – *Scatterplot* matrix showing multivariate analysis between phylogenetic grouping and Koppen climate classification using the Clustering Visualizations of Multidimensional data (Hurley, 2004). Same color in the *scatterplot* clusters shows more similar index value; thus, by means of Bray-Curtis coefficient, the data have the same pattern (“*shape*” and “*rate*”) of phylogroup distribution values. In this case, the components display close to the *scatterplot* matrix diagonal consist the highly related variables.

Figure 10 – *Scatterplot* matrix showing multivariate analysis between phylogenetic grouping and geographical distance using the Clustering Visualizations of Multidimensional data (Hurley, 2004). Same color in the *scatterplot* clusters shows more similar index value; thus, by means of Bray-Curtis coefficient, the data have the same pattern (“*shape*” and “*rate*”) of phylogroup distribution values. In this case, the components display close to the *scatterplot* matrix diagonal consist the highly related variables.