UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Leandro Henrique Santos Corrêa

FUNN-MG: Um pipeline para análise funcional e visual de metagenomas

Belém

2016

Leandro Henrique Santos Corrêa

FUNN-MG: Um pipeline para análise funcional e visual de metagenomas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.
Área de Concentração: Sitemas inteligentes; Bioinformática.

Universidade Federal do Pará – UFPA

Faculdade de Computação

Programa de Pós-Graduação

Orientador: Ronnie Cley de Oliveira Alves

Belém

2016

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Santos Corrêa, Leandro Henrique, 1988-
Funn-mg: um pipeline para análise funcional e visual de metagenomas / Leandro Henrique Santos Corrêa. - 2016.

Orientador: Ronnie Cley de Oliveira Alves.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação, Belém, 2016.

1. Bioinformática. 2. Redes. I. Título.
CDD 23. ed. 572.8633

Leandro Henrique Santos Corrêa

FUNN-MG: Um pipeline para análise funcional e visual de metagenomas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.
Área de Concentração: Sitemas inteligentes; Bioinformática.

Belém, 26 de Fevereiro de 2016:

Ronnie Cley de Oliveira Alves
Orientador

Professor
Claudomiro de Souza de Sales Junior

Professor
Guilherme Corrêa de Oliveira

Belém
2016

Este trabalho é dedicado às crianças adultas que, quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

- À minha família, a qual eu amo muito, em especial minha mãe Rosângela Santos.
- Ao meu orientador Ronnie Alves que não mediu esforços para o desenvolvimento desta pesquisa.
- Aos meus amigos que fizeram parte deste momento, sempre ajudando e incentivando.
- Àqueles amigos que já se foram, mas que se fazem presentes em lembranças e grandes momentos, em especial à Thamirys Oliveira.
- Aos colegas do grupo de pesquisa em metagenômica.
- À Fabiana Góes, amiga e companheira de estudos que me ajudou ao decorrer destes dois anos de mestrado.
- A José Flávio que me deu dicas valiosíssimas de programação, essenciais à essa pesquisa.
- Ao meu amigo Iraquitan Cordeiro Filho por ter me ajudado muitas vezes com problemas técnicos que surgiam no decorrer das análises.
- Ao Programa de Pós-Graduação em Ciência da Computação da UFPA, que faz um ótimo trabalho pelo desenvolvimento da ciência na região.
- Aos professores que por fim tornaram-se amigos.
- Aos meus amigos do Instituto Tecnológico Vale, ao qual me deram grande apoio no primeiro ano de pós-graduação.
- A todos os colegas da pós-graduação em Ciências da Computação da UFPA pelo convívio e aprendizado.

"Por vezes sentimos que aquilo que fazemos
não é senão uma gota de água no mar.
Mas o mar seria menor se lhe faltasse uma gota."
(Madre Teresa de Calcuta)

Resumo

Microrganismos habitam em diversos lugares. Embora saibamos que desempenham papéis importantes em vários ecossistemas, muito pouco se sabe sobre como essas comunidades complexas trabalham. Assim, para entender efetivamente redes metabólicas microbianas é preciso explorá-las por meio de uma abordagem de biologia de sistemas. Nós introduzimos uma estratégia baseada em análise estatísticas e métricas de teoria dos grafos para a análise funcional das redes metabólicas microbianas. Esta estratégia tem base em um grafo bipartido elaborado de forma adequada para i) avaliar a cobertura das vias metabólicas em conjunto de genes microbianos, bem como ii) a estabilidade da comunidade sobre perturbações topológicas nos elementos da rede associada. A partir da análise destas estratégias construímos um índice de representatividade das vias metabólicas encontradas na amostra, e submetemos duas amostras para análise da ferramenta. Os resultados preliminares identificaram vias metabólicas importantes relacionadas as comunidades, além de descartar vias metabólicas que não faziam sentido no contexto biológico das amostras.

Palavras chave: Metagenômica, Vias metabólcas, Redes, Análise computacional.

Abstract

Microorganisms abound everywhere. Though we know they play key roles in several ecosystems, too little is known about how these complex communities work. To act as a community they must interact with each other in order to achieve such community stability in which proper functions allows the microbial community to adapt in complex environment conditions. Thus, to effectively understand microbial metabolic networks one needs to explore them by means of a systems biology approach. We introduce a novel statistics and graph theory metric strategy for functional analysis of microbial metabolic networks. This strategy has it basis a bipartite graph properly devised to i) evaluate the cover of metabolic pathways in set of microbial genes, as well as ii) community stability over topological network perturbations in the associated microbial consortia. Based on the analysis of these strategies built a representative index of metabolic pathways found in the sample, and submit two samples for analysis tool. Preliminary results have identified important metabolic pathways related communities, and discard metabolic pathways that made no sense in the biological context of the samples.

Keywords: Metagenomics, Metabolic pathways, Networks, Computational analysis.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AMD	*Acid Mine Drainage*
BLAST	*Basic Local Alignment Search Tool.*
DCM	*Deep Chlorophyll Maximum.*
DNA	*Deoxyribonucleic acid.*
HMM	*Hidden Markov Models.*
FDR	*False Discovery Rate.*
FUNN-MG	*FUNctional Network analysis of MetaGenomics.*
HTS	*High-Throughput Sequencing.*
IMG/M	*Integrated Microbial Genomes System.*
KAAS	*KEGG Automatic Annotation Server.*
KEGG	*Kyoto Encyclopedia of Genes and Genomes.*
MGP	*Microbial Gene Pathway.*
ORF	*Open Read Frame.*
rRNA	*Ribosomal Ribonucleic Acid.*
SUR	*Surface.*
WGS	*Whole Genome Shotgun.*

Sumário

1 Apresentação da pesquisa

1 Introdução

Os microrganismos habitam qualquer parte da biosfera, incluindo solo, águas termais, fundo de oceanos, no alto da atmosfera, dentro de rochas profundas na crosta terrestre e em diversos organismos eucariontes pluricelulares (ex. humanos, animais, etc). Eles são extremamente adaptáveis às condições em que nenhum outro organismos seria capaz de sobreviver. Sua adaptabilidade é principalmente ligada devido ao fato de que e boa parte das vezes coabitam em comunidades complexas. As interações dentro das redes microbianas desempenham funções essenciais para a manutenção e sobrevivência da comunidade. Infelizmente, muito pouco se sabe sobre as interações microbianas.

Com o recente advento de tecnologias de sequenciamento de alto rendimento (HTS), abordagens metagenômicas de sequenciamento têm sido aplicadas para investigar caracterizações de diversas comunidades microbianas, incluindo marcadores de codificação de gene 16S rRNA e codificação de sequenciamento *Whole Genome Shotgun* (WGS) (HUGENHOLTZ; TYSON, 2008, p. 481). Além disso, o rápido desenvolvimento de numerosas ferramentas computacionais e metodologias têm sido explorados para a interpretação eficaz e visualização de perfis taxonômicos e metabólicos das comunidades microbianas complexas, colocadas em aplicações de perspectiva em vários domínios como a agricultura (FIERER et al., 2012), medicina (BäCKHED et al., 2005) e biomineralização (JOHNSTON et al., 2013).

O avanço tecnológico ocorrido nas últimas décadas propiciou a pesquisa metagenômica buscar o desenvolvimento de análises destas comunidades em determinados ambientes através de técnicas independentes de cultivo (amostras coletadas diretamente no meio ambiente, por exemplo), já que tradicionalmente a microbiologia baseava-se em culturas puras de crescimento em laboratório, ignorando a existência de muitos microrganismos que não se desenvolvem desta forma (HUGENHOLTZ; TYSON, 2008, p. 481).

Apesar do grande avanço nas tecnologias computacionais para a análise metagenômica, existe um espaço pouco explorado no desenvolvimento de métodos especificamente projetados para identificar as interações fundamentais em comunidades microbianas e, consequentemente, os genes associados a funções metabólicas essenciais (WOOLEY; GODZIK; FRIEDBERG, 2010) (PRAKASH; TAYLOR, 2012).

Neste sentido, nesta pesquisa é proposto um fluxo de análises computacionais em dados metagenômicos (*pipeline*) com base no material genético extraído e no conjunto de vias metabólicas anotadas. Considerando estes dados, o objetivo central é determinar

o repertório funcional destas comunidades com base na topologia da rede de interação metabólica.

A partir de bancos de dados de anotação funcional como KEGG (KANEHISA et al., 2012), foram construídas redes de interação a nível metabólico com a finalidade de identificar elementos centrais que executam importantes funções dentro da rede biológica. Medidas de redes utilizadas no estudo da teoria dos grafos e simulações de pertubações no sistema foram executadas para que os perfis de cada elemento da rede pudessem ser extraídos e posteriormente analisados. De antemão, a luz desta análise pôde-se identificar possíveis vias metabólicas que não possuem representatividade e significância estatística para a amostra em avaliação, portanto permitindo o refinamento da seleção de vias mais resilientes e importantes na comunidade microbiana presente na amostra.

2 Contextualização e terminologias do trabalho

2.1 A pesquisa metagenômica

Métodos de sequenciamento metagenômicos surgiram como uma abordagem revolucionária para investigar comunidades microbianas como sistemas altamente integrados. A análise de dados metagenômicos envolve a extração e sequenciamento de DNA em massa, além do suporte computacional de alto desempenho para associação de funções a cada sequência (DINSDALE et al., 2013, p. 1).

O fluxo de atividades depende do tipo de pesquisa. Geralmente, as pesquisas se dividem por etapas (Figura 1), destacam-se: amostragem, sequenciamento, montagem, classificação, predição de genes, anotação funcional e estatística.

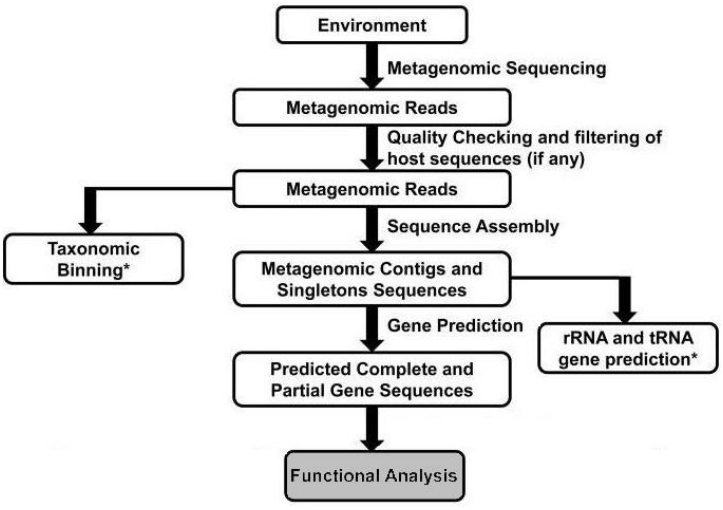

Figura 1: Fluxo de atividades da pesquisa metagenômica baseado em (PRAKASH; TAYLOR, 2012, p.2). Em destaque, a analise funcional, onde se situa o pipeline proposto nesta dissertação.

Na fase inicial as amostras são coletadas diretamente no ambiente e após um processo de filtragem e replicação do material genético são submetidas ao processo de sequenciamento dos dados. Existe uma restrição física dos sequenciadores que limita o sequenciamento de cadeias de nucleotídeos com um número muito grande de bases. Para contornar o problema foi desenvolvido uma nova geração de sequenciamento conhecido como *whole genome shotgun sequencing* (WGS). A nova geração de sequenciamento consiste em submeter a molécula de DNA a altas taxas de vibração fazendo com que ocorra fragmentação do material genético em vários pedaços de tamanhos aleatórios compatíveis com as limitações dos sequenciadores.

Os pedaços de material genético são sequenciados e representados por sequências contínuas de caracteres: A - adenina, C - citosina, T - timina e G - guanina, que representam as quatro bases que compõe o DNA. Os *reads*, como são conhecido estes fragmentos, são então combinados em trechos contínuos chamados *contigs*, de maneira que haja um consenso na sobreposição das cadeias de nucleotídeos, como mostra a Figura 2. Após esta etapa de montagem do conjunto de *reads*, análises para identificação de sequências codificantes, chamadas ORFs, e classificação taxinômica geralmente são realizadas.

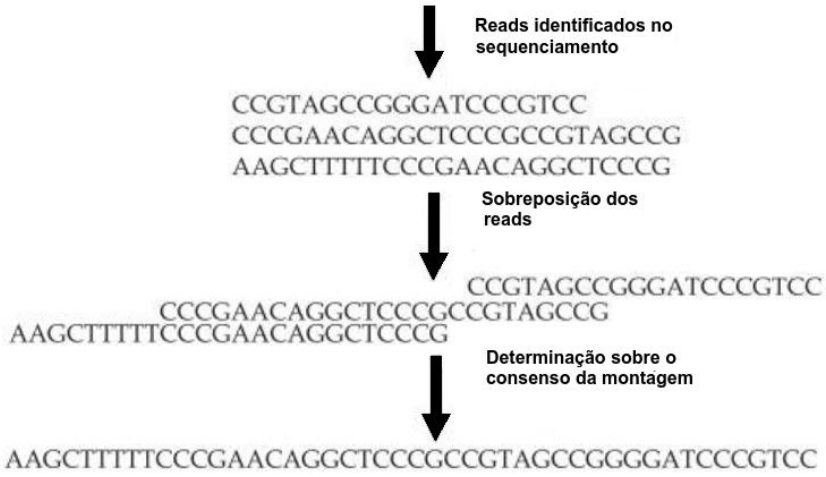

Figura 2: Determinação da melhor sobreposição do conjunto de *reads* no processo de montagem das sequências.

Geralmente, o estágio final do *pipeline* metagenômico é análise funcional das sequências obtidas. O primeiro nível de anotação funcional é atribuir funções biológicas aos ORFs. ORFs são estruturas delimitadas por duas trincas de nucleotídeos chamadas códon de iniciação e códon de parada, que quando possuem função de codificação de proteínas são obtidas após a fase de predição de genes. A presença de ORFs indica a existência de uma estrutura do material genético bem conservada que pode ou não ser transcrita ou traduzida.

O segundo nível de anotação, seria descobrir genes que constituem redes biológicas, tais como as vias metabólicas nos dados (WOOLEY; GODZIK; FRIEDBERG, 2010, p.8).

O desevolvimento da análise funcional metagenômica geralmente inclui pesquisas por homologia BLAST em banco de dados de genes ou proteínas como NCBI (WHEELER et al., 2008) ou KEGG, consultas HMMer em banco de dados como Pfam (FINN et al., 2008) e TIGRfam (SELENGUT et al., 2007), consultas em sistemas de anotação dedicados a metagenômica, como IMG/M (MARKOWITZ et al., 2008) e CAMERA (SUN et al., 2011), e outras estratégias que constantemente vêm acrescentando novas características na diversidade das análises, como mostra a Figura 3.

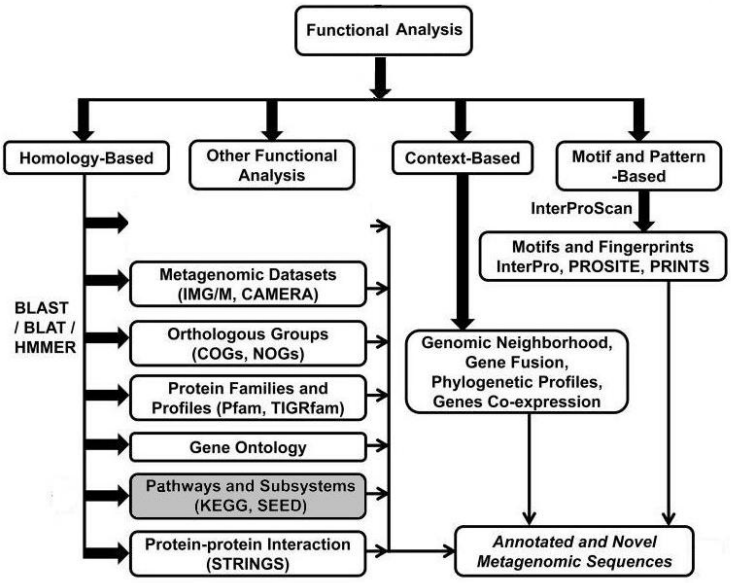

Figura 3: Principais abordagens em ferramentas de análise funcional em dados metagenômicos (PRAKASH; TAYLOR, 2012, p.2). Em destaque a abordagem utilizada nesta pesquisa.

Estas análises só podem ser realizadas graças ao desenvolvimento de tecnologias que permitem a análise, armazenamento e fácil disponibilização de grandes quantidades de dados, haja vista a complexa natureza dos dados metagenômicos. Pathguide[1] faz um levantamento dos repositórios disponíveis para análise de enriquecimento funcional e dá uma ideia do número de ferramentas que vêm sendo desenvolvidas na área.

2.2 Análise funcional a nível metabólico

Um ponto importante na análise de dados metagenômicos é a identificação do conjunto de vias metabólicas que compõe a comunidade. O cálculo da abundância relativa das vias dentro de uma única amostra fornece uma visão global do ambiente e é usado em muitos estudos e plataformas (ITAI; SHANRON, 2010, p.23).

Vias metabólicas são conjuntos de reações químicas onde o produto final de uma reação serve de substrato (ou reagente) à reação que lhe sucede, estando as reações inter-

[1] http://www.pathguide.org/

dependentes umas das outras. É importante salientar que ao menos uma destas reações deve ser irreversível, caso contrário consideramos estas reações como um ciclo fútil onde só há dissipação de energia. O complexo sistema de reações são importantes para estabilidade e sobrevivência dos organismos no meio ambiente. O ciclos do ácido cítrico e da degradação da glicólise em piruvato, por exemplo, são importantes no processo de geração de energia a nível celular.

Ao conjunto de todas as reações ocorridas no interior de um organismo vivo damos o nome de metabolismo, e se tratando de metagenômica em que o foco do estudo é a interação entre os organismos de uma comunidade, é importante a análise de como essas reações metabólicas entre diferentes organismos podem ser importantes para a sobrevivência destes e para a manutenção do equilíbrio com o meio onde vivem. Este é o principal objetivo da análise funcional a nível metabólico, busca-se por padrões de comportamento com base no perfil metabólico dos organismos, que possam servir de assinatura para registro de padrões de comportamento de uma comunidade. Por exemplo, em (NEISH, 2009), pesquisadores buscaram compreender padrões de comportamento em comunidades de microrganismos que vivem no intestino humano com intuito de encontrar características marcantes que possam servir para fins terapêuticos.

Um desafio importante neste modelo de análise é a identificação das vias mais relevantes que compõe o sistema. Para a análise de relevância, também chamada de enriquecimento funcional, é comum o uso de técnicas estatísticas com base na cobertura destas vias considerando um conjunto de dados de referência (banco de dados, amostras). Itai e Shanron (2010), por exemplo, apontam dois principais métodos que consideram um conjunto de genes e vias metabólicas (Figura 4): a) método independente, b) método de inserção. O modelo independente admite o cálculo da abundancia relativa por

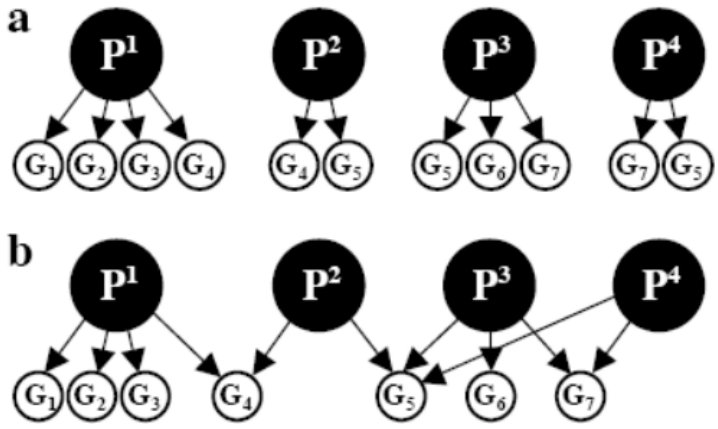

Figura 4: Modelos de enriquecimento funcional estatístico em dados metagenômicos a nível das vias metabólicas (ITAI; SHANRON, 2010, p. 24). a) Representação do modelo independente. b) Representação do modelo por inserção. Os vértices $\mathbf{P}^n$ representam as vias metabólicas, e os vértices $\mathbf{G}_n$ representam os genes.

organismo identificado na amostra, assumindo que a abundância entre os organismos são mutualmente independentes. Já o modelo por inserção calcula a abundância levando em consideração que o conjunto de genes é único, e as vias compartilham genes independentemente da espécie.

Os modelos estatísticos de anotação funcional tipicamente ignoram a topologia e tratam as vias metabólicas como conjunto de genes, **embora o perfil das comunidades geralmente seja destacado com base na interação entre as funções metabólicas dos organismos**. Um exemplo é o estudo da substância rizoxina. A substância é produzida pela bactéria *Burkholderia rhizoxinica*, encontrada em associação simbiótica com o fungo *Rhizopus microsporus*. Nessa relação ambos beneficiam-se da ação fitopatogênica da rizoxina contra mudas de arroz, que durante estágio de morte servem como alimento para a bactéria e para o fungo (LACKNER et al., 2011).

2.3 Modelo de redes na biologia de sistemas

Modelos de simulação de rede permitem um melhor entendimento das relações bioquímicas no processo metabólico seja de um organismo ou de uma comunidade. A reconstrução de vias metabólicas em suas respectivas reações e enzimas, torna mais simples o entendimento dos processos celulares o que podem permitir a identificação de características chave do metabolismo, tais como o aumento no crescimento, distribuição de recursos, a robustez da rede, e essencialidade gênica.

Em termos simplificados, a reconstrução de vias metabólicas recolhe toda a informação metabólica relevante conhecida de um organismo e a compila em um modelo matemático, usualmente em forma de grafo.

Um grafo consiste em dois conjuntos V e E, tal que os elementos V são chamados de vértices ou nodos; e os elementos E são chamados de arestas. Cada aresta conecta dois vértices, e possui um conjunto de um ou dois vértices associados (GROSS; YELLEN, 2004, p. 31).

Toda a rede pode ser escrita em forma de grafos, e a extração de algumas métricas peculiares à área pode ajudar a comparar e caracterizar diferente redes complexas (HUBER et al., 2007). A complexidade das interações torna muito difícil propor um modelo que explique o comportamento padrão de todas as redes biológicas, no entanto alguns marcadores podem oferecer *insights* sobre a arquitetura básica dos sistemas.

O modelo de redes aleatórias, por exemplo assume que um número fixo de vértices são conectados entre si, a mais notável propriedade deste modelo é a sua "democracia" ou característica uniforme, pois as ligações, neste modelo, são colocadas randomicamente entre os nodos.

Considerando as redes metabólicas que são o foco central desta pesquisa, existem

evidências de que estas redes são regidas por um modelo de livre escala (JEONG et al., 2000). Redes de livre escala são caracterizadas por terem seu grau de distribuição regido por uma lei de potência. A probabilidade de que um nodo esteja altamente ligado é estatisticamente mais significativa do que em um grafo de uma rede aleatória. As propriedades da rede frequentemente são determinadas por um número relativamente pequeno de nós altamente ligados que são conhecidos como *hubs* (BARABASI; OLTVAI, 2004, p. 102).

3 Trabalhos relacionados

A análise funcional a nível de comunidades microbianas representa um desafio computacional complexo. As análises geralmente são executadas em *pipelines*, e na maioria das vezes levam em conta uma lista de genes microbianos que é avaliada utilizando abordagens de análise a nível de vias metabólicas.

Ferramentas como MEGAN (HUSON et al., 2007) e MG-RAST (GLASS et al., 2010), providenciam resultados com base na cobertura destas vias em bancos de dados como KEGG ou SEED (OVERBEEK et al., 2005), contudo, eles se diferem a medida que MEGAN utiliza o banco de dados de proteína NCBI-NR para a anotação, enquanto MG-RAST utiliza anotação construída a partir de organismos do próprio banco de dados SEED (MITRA et al., 2011, p. 4). Projetos promissores como o pipeline HumanN (ABUBUCKER et al., 2012), vêm sendo desenvolvidos e aplicados com sucesso no Projeto Microbioma Humano (HMP). HumanN utiliza uma metodologia similar a destas ferramentas, no entanto, também oferece a exploração de métricas de diversidade ecológica antes de enumerar os perfis funcionais da comunidade.

Outro destaque é a ferramenta KAAS (MORIYA et al., 2007), que faz parte do conjunto de ferramentas disponibilizadas pelo banco de dados KEGG para análise funcional. Esta ferramenta tem papel importante na anotação de funções metabólicas a partir de consultas de sequencias de DNA e Proteínas. Contudo, os resultados da abordagem de mapeamento direto estão sujeitos a redundância e incompletude dos dados, já que algumas proteínas podem ser percursoras de múltiplas reações bioquímicas independentes, como no caso do ciclo Krebs encontrado no genoma humano, e no ciclo do Carboxilato redutivo encontrado em alguns microrganismos autotróficos.

> O ciclo carboxilato redutivo é uma via alternativa de fixação do carbono, atualmente encontrado apenas em certos microrganismos autotróficos. Na verdade, o ciclo carboxilato redutivo é essencialmente o inverso do ciclo de Krebs (ácido cítrico ou o ciclo de ácido tricarboxilico), a via final comum no metabolismo aeróbico para a oxidação de hidratos de carbono, ácidos graxos e aminoácidos, de modo que eles compartilham reações e papéis funcionais. Por esta razão, as proteínas responsáveis pela função normal do ciclo de Krebs podem ser erroneamente consideradas como evidências para a existência de um ciclo de Carboxilato redutivo no genoma humano. (YE; DOAK, 2009, p. 2).

A super estimativa do conjunto de vias pode sobrestimar a diversidade funcional das amostras, logo a representatividade das vias deve ser identificada de alguma forma, seja a partir de conjuntos de dados de referência, ou análise de biodiversidade das amostras.

Estratégias computacionais vêm ganhando espaço como forma de amenizar o problema. MinPath (YE; DOAK, 2009) propõe uma análise que busca deduzir o conjunto mínimo de vias necessárias para suportar um conjunto observado de funções. O método

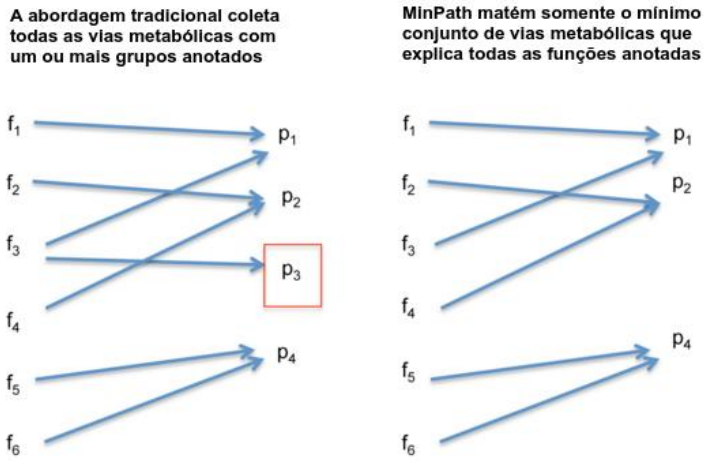

Figura 5: **Ilustração da metodologia da ferramenta MinPath**. Suponha que 6 famílias (grupos ortólogos, F1, ..., F6) são identificadas a partir de uma dada amostra de genes (por exemplo, os genes podem ser a partir de um genoma, ou amostrado a partir de uma metagenoma). A abordagem de mapeamento naïve (mostrada à esquerda) leva a uma reconstrução com 4 vias anotadas (P1, P2, P3, e P4). A abordagem utilizada pela ferramenta MinPath afirma que apenas três vias, P1, P2 e P4 são suficientes para explicar a existência das 6 famílias anotadas no conjunto de dados, e uma reconstrução conservadora das vias neste caso, deve ter apenas 3 vias (mostradas à direita). Como mostrado no trabalho, uma estimativa conservadora de tais vias fornece uma estimativa mais fiável da diversidade funcional de uma amostra (YE; DOAK, 2009, p. 2).

utiliza Programação Inteira[2] para decidir se uma via está presente, com base nas funções observadas. Note-se que, neste caso, a abundância relativa das diferentes funções não é tida em conta, e não existe qualquer estimativa da abundância relativa das diferentes vias encontradas (ITAI; SHANRON, 2010, p. 24).

MinPath (conjunto mínimo de *Pathways*), pode ser aproximadamente descrito da seguinte forma: dado um conjunto de vias de referência e um conjunto de proteínas ou grupos ortólgos (e suas funções previstas) que podem ser mapeados para uma ou mais vias, tenta-se encontrar o número mínimo de vias que podem explicar todas as funções (ver Figura 5).

[2] Programação Inteira pode ser entendido como um caso específico da Programação Linear, onde as variáveis devem assumir valores inteiros.

O *pipeline* FUNN-MG, proposto nesta pesquisa, tenta contornar o problema utilizando uma abordagem híbrida que considera tanto a abundância relativa das vias, como a natureza topológica da rede de interação.

Medidas topólogicas de redes de interação biológicas já são utilizadas em biologia de sistemas para análise a nível genético, como o trabalho de Goh et al. (2007), em que pesquisadores constroem uma rede de interação que explora a relação entre genes e desordens genéticas identificadas no DNA humano. Outro exemplo em que a topologia da rede é usada para analise funcional a nível genético é o trabalho de Vey e Moreno-Hagelsieb (2012). Neste trabalho os autores exploram a predição de *operons*[3] para derivar interações funcionais que são traduzidas e categorizadas de acordo com a anotação funcional do banco de dados IMG/M (MARKOWITZ et al., 2008). O resultado é uma coleção de redes discretas de ligações ponderadas com base na anotação, que são posteriormente analisadas visando identificar módulos de anotação que retratem a organização funcional e hierárquica das comunidades (VEY; MORENO-HAGELSIEB, 2012, p. 1).

Todos estes esforços são parte de uma pesquisa que abrange vários níveis e diversas metodologias, neste sentido, o uso de *pipelines* vêm crescendo. A modularização das etapas e o fácil gerenciamento das atividades (remoção e adição de novos modelos e estratégias de pesquisa) tornam promissora a criação de pipelines como padrão de desenvolvimento de ferramentas da área, mesmo na fase de análise funcional a nível metabólico, e não faltam exemplos na literatura de ferramentas que utilizam o modelo de análise e processamento sequencial (METAREP (GOLL et al., 2010), MetAMOS (TREANGEN et al., 2013), SmashCommunity (ARUMUGAM et al., 2010), etc).

4 Justificativa e contribuição à área

A reconstrução metabólica proporciona perspectivas sobre a funcionalidade de uma comunidade microbiana e pode ser utilizada para comparação entre várias comunidades. No entanto, a análise depende da qualidade dos dados, e do modo como as vias biológicas podem ser reconstruídas a partir de funções preditas.

As comunidades microbianas adaptam-se as condições ambientais modificando sua rede de expressão metabólica. Logo análises enviesadas podem gerar compreensões equivocadas a respeito da diversidade das amostras. Mesmo para genomas únicos completos, a reconstrução das vias nem sempre oferece uma imagem clara das funções biológicas do organismo, e a curadoria humana e verificação experimental é muitas vezes necessária (FRANCKE; SIEZEN; TEUSINK, 2005) (OBERHARDT et al., 2008). Assim, estratégias de enriquecimento das vias são aplicadas com intuito de diminuir o ruído nos resultados,

[3] *Operon* (em inglês) ou operão é um conjunto de genes nos procariontes e em alguns eucariontes que se encontram funcionalmente relacionados, contíguos e controlados coordenadamente.

o que pode levar melhores resultados e menor perda de tempo nas análises laboratoriais.

Boa parte das principais ferramentas de análise funcional metabólica em amostras metagenômicas, utiliza análise de cobertura das vias com relação a algum conjunto de referência. Os resultados esperados dessas ferramentas são descritivos. O foco na abordagem de mapeamento quase sempre é voltado à lista das vias catalogadas relacionadas com maior percentual de homologia às sequências consultadas. Os elementos mais significativos ficam por conta de gráficos de dispersão e representatividade percentual como mostra a Figura 6.

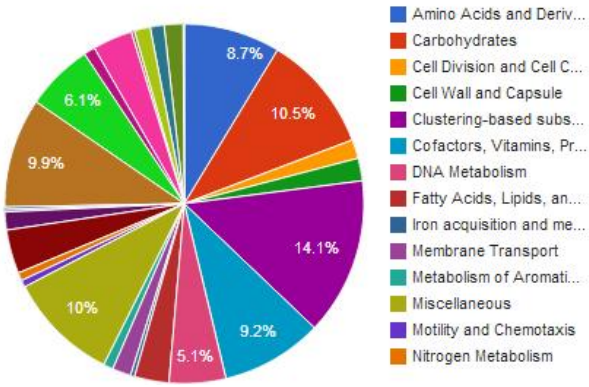

Figura 6: Porcentagem dos subsistemas da amostra do trabalho de Tyson et al. (2004) identificados pela ferramenta MG-RAST e obtidos com base no banco de dados SEED. Subsistemas são um conjunto de papéis funcionais encontrados em qualquer processo biológico comum, incluindo vias metabólicas, fenótipos, ou estruturas complexas de múltiplas subunidades.

Nesta presente proposta introduzimos uma abordagem com foco em redes de interação a partir de conceitos de teoria de grafos, simulação, técnicas de visualização da informação a nível computacional e indicadores estatísticos. A hipótese principal considerada no trabalho **é que a topologia dos elementos da rede, aliada a cobertura das vias em um banco de dados de referência, pode ajudar a construir um melhor indicador do grau de importância destas funções com relação a comunidade**.

Outro ponto importante é a representação visual dos elementos. A dispersão dos elementos no modelo de rede permite os pesquisadores avaliarem a interação entre genes e vias de maneira mais precisa, avaliando grupos de genes responsáveis por determinada função específica, através da interação direta com a ferramenta.

O *pipeline* FUNN-MG, ainda não dispõe de outros níveis de análises metagenômicas, como montagem, predição, análise taxonômica, e pesquisas de homologia. Portanto, etapas de "pré-processamento", como indicadas anteriormente devem adotar um padrão como (PYLRO et al., 2014) antes de ser submetido a análise da ferramenta FUNN-MG. Todas as análises desenvolvidas levam em conta conjunto de genes ou grupos ortólogos anotados no banco de dados KEGG.

Trabalhos recentes foram publicados, os quais propuseram a metodologia adotada nesta pesquisa. Em destaque o trabalho de Corrêa et al. (2014) que introduziu os conceitos de montagem da rede MGP (base da construção das analises desta pesquisa), e das avaliações estatísticas empregadas para avaliação de cobertura das vias metabólicas identificadas. A iniciativa da pesquisa foi parcialmente financiada Conselho Nacional de Desenvolvimento Científico e Tecnológico no âmbito do projecto BIOFLOWS [475620/2012-7].

5 Objetivos

5.1 Objetivo Geral

O objetivo deste presente trabalho é construir um *pipeline* para análise funcional e visual de metagenomas. Para tal, pretende-se criar uma rede de interação entre grupos de genes e vias metabólicas, e a partir de então, selecionar as vias mais representativas na amostra, com base na cobertura e na topologia da rede.

5.2 Objetivos específicos

- Analisar estudos metagenômicos com foco na diversidade funcional a partir da representatividade das vias metabólicas;
- Identificar e selecionar o conjunto de dados que serão manipulados;
- Avaliar estratégias computacionais que permitam uma exploração mais estruturada, tanto na construção como na exploração da rede de interação;
- Definir e executar um fluxo de análise de dados que permitam a identificação das vias com maior representatividade;
- Comparar os resultados obtidos com a execução do *pipeline* à outras abordagens;
- Disponibilizar um software livre e documentado.

6 Metodologia da Pesquisa

A pesquisa proposta neste trabalho pode ser classificada como explicativa. No decorrer do desenvolvimento das análises houve a necessidade da utilização de métodos experimentais de modelagem e simulação de redes para que as relações metabólicas identificadas na amostra fossem mapeadas, analisadas e posteriormente explicadas.

Quanto a metodologia, a partir de amostras de cadeias de nucleotídeos já sequenciadas, utilizou-se pesquisa de homologia em banco de dados de anotação para identificação

de genes (ou grupos ortólogos). A intenção inicial é verificar como as relações à nível metabólico se comportam na rede, e em um segundo momento formular indicadores que pudessem de certa forma fornecer um *ranking* de importância dos elementos desta rede. O método indutivo aplicado, permitiu análise e avaliação dos resultados após algumas estratégias computacionais adotadas, como: extração de medidas de rede, análises estatísticas, simulações, entre outros.

Os dados submetidos a avaliação são provenientes de amostras metagenômicas de oceanos (SUNAGAWA et al., 2015) e drenagem ácida de mina (AMD) (TYSON et al., 2004). As amostras foram escolhidas com base na complexidade de distribuição das espécies, baixa (AMD) e alta (oceano). A complexidade está diretamente ligada com a presença de populações dominantes na amostra (MAVROMATIS et al., 2007). A avaliação dos resultados foi feita em duas partes: (1) a comparação dos resultados com outra ferramenta de análise funcional a nível de vias metabólicas que se proponha a identificar o conjunto de vias mais representativas, como a ferramenta MinPath (YE; DOAK, 2009); (2) Comparação dos resultados obtidos tendo em conta o número de vias mais representativas e os respectivos metadados ambientais de onde as amostras foram coletadas.

O banco de dados KEGG foi utilizados para identificação das vias relacionadas às amostras. O KEEG é o banco de dados referência para analise de vias metabólicas, e vem sendo utilizado com sucesso pela comunidade de biologia de sistemas (KANEHISA et al., 2012). Os serviços de mapeamento destas vias foam feitos diretamente nos respectivos servidores ou através de interfaces de programação.

A ferramenta apresentada neste trabalho está disponibilizada em uma versão livre de acesso web <https://bitbucket.org/leandro_correa/funn-mg/>, a documentação que inclui os artefatos gerados no decorrer da fase de desenvolvimento será disponibilizada, e a plataforma de desenvolvimento está aberta a outros possíveis colaboradores que se interessarem pela proposta.

2 Materiais e métodos

A análise de dados metagenômicos é uma tarefa analítica complexa em ambos os sentidos, biológico e computacional. Em sequências baseadas em amostras metagnômicas, pesquisadores se concentram em encontrar o maior número das sequências genéticas padrão contendo as quatro bases de nucleótidos diferentes (A, C, G e T) que existem nas cadeias de DNA. As sequências podem ser analisadas de várias maneiras diferentes. Por exemplo, pesquisadores podem usar as sequências para analisar o genoma da comunidade como um todo, o que pode oferecer *insights* a respeito da ecologia, evolução e funcionalidade. Nesta pesquisa propõe-se o *pipeline* FUNN-MG que permite uma analise visual e funcional das vias mais representativas em comunidades microbianas.

O *pipeline* possui quatro principais tarefas (identificadas nos retângulos verdes de borda aredondada na Figura 7) que devem ser executadas sequencialmente: i) identificação de vias metabólicas, ii) avaliação estatística do conjunto de vias metabólicas identificadas, iii) construção da rede de interação gene/grupos-vias metabólicas, iv) extração de medidas para construção de um índice que represente o enriquecimento funcional destas vias.

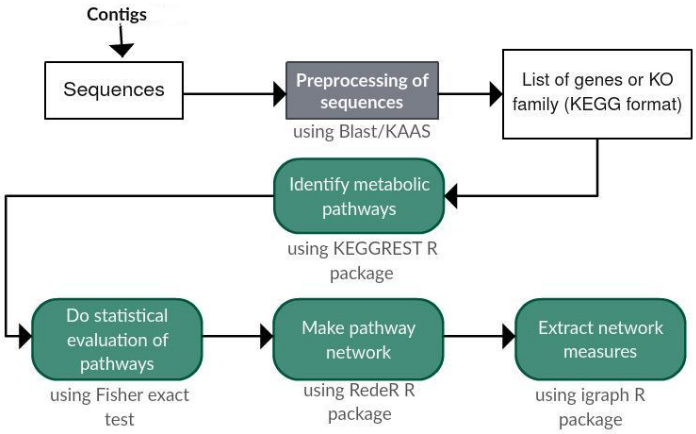

Figura 7: *Pipeline* FUNN-MG de análise fucional e visual metagenômica baseado na rede metabólica de comunidades de microrganismos.

1 Entrada de dados e pré-processamento das sequências

A análise pode ser feita de duas formas diferentes: i) através de consultas de genes, ii) através da consulta de grupos ortólogos. Grupos ortólogos são conjuntos de genes que são originários de um ancestral comum, logo, apesar de serem encontrados em espécies distintas possuem a mesma função biológica[1].

[1] Alguns genes conhecidos como prólogos também possuem (por vezes) a mesma função biológica.

No banco de dados KEGG, as funções à nível molecular específicas à um gene são armazenadas no banco de dados KO (KEGG *Orthology*) e associadas com seus respectivos grupos ortólogos, a fim de permitir a extensão de evidência experimental num organismo específico a outros organismos.

Até então, FUNN-MG não possui uma etapa de identificação dos genes ou grupos ortólogos a partir das sequências de DNA. As análises devem ser realizadas considerando o banco de dados KEGG, para isso sugere-se o uso da ferramenta BLAST (KENT, 2002) para identificação dos genes, ou a ferramenta KAAS (MORIYA et al., 2007) para a identificação dos grupos ortólogos.

Os dados de entrada devem seguir o padrão de arquivos CSV (*Comma-Separated Values*, do inglês "valores separados por vírgula") de acordo com a Tabela 1.

Tabela 1: Exemplo de um arquivo de entrada para a ferramenta FUNN-MG.

ID	KO	GENE	SCORE
376	K02986	hch:HCH_06194	313.153
377	K02948	tvi:Thivi_4224	197.593
378	K02952	win:WGP_0527	163.31
434	K02954	win:WPG_0527	167.548
572	K06168	syw:SYNW1640	711.835
603	K02970	sil:SPO0229	63.1586
754	NULL	phd:102343398	194.512

A tabela possui quatro colunas distintas:

- **Id**: O número de identificação da cadeia de nucleotídeos submetida a análise de identificação dos genes ou grupos ortólogos;
- **KO**: O identificador do respectivo grupo ortólogo da sequência;
- **Gene**: O gene identificado na consulta de homologia;
- **Score**: O numero de *hits* que indica o quão homólogo duas sequências são estruturalmente.

As informações de ID e SCORE são opcionais, pois não influenciam nos resultados da análise do pipeline FUNN-MG. No entanto, podem esclarecer os resultados obtidos com as análises, identificando as sequencias relacionadas às vias com um menor índice de enriquecimento e esclarecendo possíveis erro de anotação, por exemplo. As colunas GENE e KO são os respectivos *inputs* das análises propostas pelo pipeline.

2 Identificação das vias metabólicas

O primeiro passo a partir do conjunto de dados de entrada é a identificação e deleção dos elementos duplicados na amostra. A partir de então, consultas são realizadas com o intuito de identificar vias metabólicas que tenham anotação funcional comprovada para cada um dos genes ou grupos ortólogos. Mais uma vez a ferramenta utiliza o banco de dado KEGG, e os elementos que não obtém anotação funcional mapeados no banco de dados são automaticamente descartados[2].

Nesta fase, para a identificação da relação entre os dados de entrada e as vias metabólicas foi utilizado o auxílio do pacote KEGGREST (TENENBAUM, 2013) da linguagem de programação R (R Core Team, 2013).

3 Avaliação estatística das vias metabólicas

De posse do conjunto de vias metabólicas identificadas na amostra, FUNN-MG permite dois tipos de análises estatísticas dependendo do tipo de consulta: (1) a partir do conjunto de genes; (2) a partir dos grupos ortólogos.

A análise em função dos genes é mais específica, pois considera no cálculo a probabilidade de se encontrar espécies em cada uma das vias metabólicas encontradas na amostra, dado que, eventualmente, apenas um grupo selecionado de espécies terá uma via associada. Já para análise dos grupos ortólogos, por se tratar de genes que podem pertencer a diversas espécies diferentes, considera-se somente a variação dos elementos encontrados na amostra e na população, logo, não existe a possibilidade de considerarmos este viés na análise.

Para cada uma das consultas, três testes estatísticos foram utilizados para atribuir valor de representatividade às vias metabólicas encontradas:

- **Teste hipergeométrico**: indica a probabilidade da via em questão ser encontrada na amostra levando em consideração as relações anotadas no banco de dados KEGG, ou seja, enquanto menor a probabilidade, menores as chances da via ter sido encontrada ao acaso.
- **Teste Fisher**: procura garantir que a hipótese nula, nesse caso se há relação entre a via metabólica e o elemento em questão (gene ou grupo ortólogo), seja exata, garantindo um grau de significância. O cálculo é baseado no trabalho de Sreenivasaiah et al. (2012) com algumas adaptações.

[2] Vale ressaltar que os dados que não forem mapeados são guardados num arquivo de nome "not-found.tsv" para analises futuras.

- **FDR**: O cálculo do FDR (taxa de descoberta de falsos positivos) é realizado para avaliação dos resultados falsos positivos, que são geralmente encontrados quando existe um número alto de vias identificadas inicialmente.

Os modelos de testes estatísticos propostos são baseados no modelo independente (Figura 4) proposto por Itai e Shanron (2010). Para todas as vias metabólicas identificadas, uma a uma, foi calculado um *pvalue* que oferece uma estimativa inicial de significância da via. As peculiaridades de cada caso serão discutidas nas seções a seguir.

3.1 Avaliação estatística a partir dos grupos ortólogos

Os cálculos estatísticos são baseados em informações sobre a distribuição da relação entre grupos ortólogos e vias metabólicas no banco de dados KEGG.

Tabela 2: Tabela de anotação da quantidade de grupos ortólogos relacionados a cada via, tanto na amostra quanto no banco de dados KEGG. Em destaque a via path:map00910 *Nitrogen metabolism* que servirá de exemplo para demonstração dos cálculos estatísticos realizados. Os dados utilizados como exemplo são da amostra 4 DCM do projeto Tara ocean (SUNAGAWA et al., 2015).

Vias metabólicas	Amostra	KEGG
path:map00791	1	14
path:map00830	1	35
path:map00900	2	35
path:map00908	1	8
path:map00910	3	99
path:map03010	50	144

A Tabela 2, é um exemplo de tabela de anotação a nível de grupos ortólogos, que informa quantos grupos estão relacionado com a via metabólica, tanto na amostra, quanto no total de anotação do banco de dados KEGG. Neste exemplo ilustrativo, vamos imaginar que 83 grupos ortólogos foram encontrados na amostra, e 11950 tenham anotação no banco de dados KEGG. Logo, segundo a fórmula da distribuição hipergeométrica temos:

$$p = (X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \tag{2.1}$$

Assim, o resultado do teste da distribuição hipergeométrica da via path:map00910 (em destaque na Tabela 2) neste exemplo é:

$$p = (X = 3) = \frac{\binom{99}{3}\binom{11950-99}{83-3}}{\binom{11950}{83}} = 0.02652284$$

Para computar a função de distribuição hipergeométrica utilizou-se a função *dhyper*() nativa da linguagem de programação R.

O próximo teste executado na análise estatística visa garantir que existe relação entre a via metabólica identificada e o conjunto de grupos ortólogos obtidos na amostra. Para tanto, com auxílio da função *fisher.test*(), também nativa da linguagem de programação R, executa-se nessa análise o teste Fisher exato. O teste Fisher é feito com base em uma tabela de contingência. Tabelas de contingência são usadas em estatística para fornecer um resumo tabular de dados categóricos, as células representam o número de vezes que uma determinada combinação de variáveis ocorrem juntas em um conjunto de dados.

Nesse exemplo, a Tabela 3 detalha a tabela de contingência da via path:map00910 em destaque na Tabela 2.

Tabela 3: Tabela de contigência da via path:map00910 *Nitrogen metabolism*, considerando análise de grupos ortólogos. Entre parênteses os resultados obtidos a partir da Tabela 2. As letras, logo acima dos resultados, correspondem a fórmula matemática do cálculo do teste Fisher.

	Grupos relacionados à via	**Grupos não relacionados à via**	**Total de grupos**
Amostra	**a** **(3)**	**b** **(80)**	a + b (83)
População	**c** **(96)**	**d** **(11771)**	c + d (11867)
Total no KEGG	a + c (99)	b + d (11851)	n (11950)

A partir dos resultados, a tabela de contingência é submetida ao teste Fisher, segundo a fórmula:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{2.2}$$

A primeira e terceira linha da tabela, correspondem ao conjunto de grupos ortólogos identificados na amostra e no banco de dados KEGG, respectivamente. A população (linha 2) corresponde ao conjunto de grupos ortólogos pertencentes ao banco de dados KEGG, porém não encontrados na amostra. Matematicamente podemos escrever o exemplo em questão da seguinte forma:

$$p = \frac{\binom{3+80}{3}\binom{96+11771}{96}}{\binom{11950}{3+96}} = \frac{(3+80)!(96+11771)!(3+96)!(80+11771)!}{2!80!96!11771!11950!} = 0.03146282$$

Os valores obtidos são utilizados para análise de significância estatística ao final do processo, considera-se um grau de confiança de 95% ($pvalue < 0.05$).

Considerando conjunto maiores que cem vias metabólicas identificadas inicialemente, após os resultados dos testes hipergeométrico e Fisher, os resultados também são submetidos ao teste de identificação de possíveis falsos positivos (FDR) com auxílio da função *qvalue*() do pacote qvalue (STOREY, 2015) da linguagem R.

3.2 Avaliação estatística a partir do conjunto de genes

Para identificação do enriquecimento funcional a partir do conjunto de genes, considera-se apenas as espécies que possuem anotação funcional para cada via identificada na amostra.

Existem espécies na comunidade que não participam de todas as vias metabólicas associadas à amostra, portanto, as análises estatísticas em relação ao conjunto de genes e vias metabólicas também consideram a distribuição populacional das especies. A Tabela 4, mostra o exemplo do cálculo detalhado da via *Cyanoamino acid metabolism*, identificada na amostra AMD (TYSON et al., 2004) que possui 477 genes e 6 espécies distintas.

Tabela 4: Quantidade de genes relacionados a via *Cyanoamino acid metabolism* encontrados na amostra e no banco de dados KEEG. As espécies encontradas na amostra estão descritas na parte superior da tabela pelos seus respectivos indicadores no banco de dados KEGG.

	fac	**lfc**	**lfi**	**tac**	**tar**	**tvo**	**Total**
Amostra	2	0	0	1	0	0	**3**
KEGG	4	0	0	3	0	4	**11**

Neste exemplo, a via *Cyanoamino acid metabolism* tem anotação funcional no banco de dados KEGG para apenas três das seis espécies encontradas na amostra (Tabela 4 destaque em cinza), além de três genes encontrados do total de onze anotados (Tabela 4 destaque em laranjado). Com base nesta informação, o total de genes tanto na amostra quanto no banco de dados KEGG é computado, como mostra a Tabela 5.

Nota-se que o total computado tanto na amostra quanto na população (em cinza na Tabela 5) leva em consideração anotação da via para cada espécie no banco de dados KEGG. A partir dessas informações a tabela de contingência é construída (Tabela 6).

Assim como nos testes feitos a partir do conjunto de grupos ortólogos, os resultados são submetidos as análises hipergeométrica (função *dhiper*()), teste Fisher (função *fisher.test*()), e FDR (função *qvalue*()).

Tabela 5: Quantidade total de genes anotados no banco de dados KEGG e obtidos na amostra, para cada espécie. O total (em destaque laranjado) é calculado somente para as espécies que possuem anotação funcional no banco de dados KEGG com relação a via Cyanoamino acid metabolism (em cinza).

Id	Organismo	KEGG	Amostra
fac	*Ferroplasma acidarmanus*	596	120
lfc	*Leptospirillum ferrooxidans*	787	216
lfi	*Leptospirillum ferriphilum ML-04*	759	23
tac	*Thermoplasma acidophilum*	528	9
tar	*Thermoplasmatales archaeon BRNA1*	556	109
tvo	*Thermoplasma volcanium*	516	2
	Total	**1640**	**131**

Tabela 6: Tabela de contigência via *Cyanoamino acid metabolism*. Em destaque as variáveis usadas como parâmetro para o cálculo de enriquecimento segundo a função *fisher.test*() da linguagem R.

	Genes associados a via	**Genes não associados a via**	**Total de genes**
Amostra	**a (3)**	**b (128)**	a + b (131)
População	**c (8)**	**d (1501)**	c + d (1509)
Total no KEGG	a + c (11)	b + d (1629)	n (1640)

4 Construção da rede metabólica

Redes biológicas oferecem uma descrição quantificável das redes que caracterizam vários sistemas biológicos (BARABASI; OLTVAI, 2004, p. 102). A análise da natureza metabólica dos indivíduos de uma comunidade sugere que o modelo de representação de redes seja um ótimo indicador do comportamento dos microrganismos que compõe esta comunidade.

Nesta pesquisa, após a análise de enriquecimento funcional ao nível de cobertura das vias, que é segmentado em duas partes (genes e grupos ortólogos), FUNN-MG utiliza os mesmos critérios de construção e extração de medidas de rede para ambos os conjuntos. Como se sabe, grupos ortólogos são conjuntos de genes, logo os modelos descritos como gene-vias metabólicas, também também podem ser compreendidos como grupos ortólogos-vias metabólicas.

Assim, o *pipeline* FUNN-MG aborda a análise de redes funcionais com base no comportamento das redes de interação entre genes e vias metabólicas em amostras metagenômicas, e utiliza conceitos e propriedades da teoria dos grafos (HUBER et al., 2007)

para representar a comunidade microbiana. Dessa forma, uma estrutura de grafo bipartido não direcionado $MGP = (G, P, E)$ aqui nomeado *Microbial Gene Pathway* (MGP), - Figura 8 - divide os vértices em dois conjuntos disjuntos: (G)enes e (P)athways (vias metabólicas). Nessa representação as arestas (E) conectam um vértice (G) a um vértice (P), e os valores obtidos na fase de enriquecimento estatístico através do teste hipergeométrico são anotados em cada vértice (P).

A representação computacional é dada por uma matriz de índices $MI[i, j]$ (Figura 8 b), onde o índice i corresponde ao conjunto de genes, e j ao conjunto de vias metabólicas, logo:

$$MI_{(i,j)} = \begin{cases} 1 & \text{se existe relação anotada entre o gene } i \text{ e a via } j \text{ ;} \\ 0 & \text{caso contrário.} \end{cases}$$

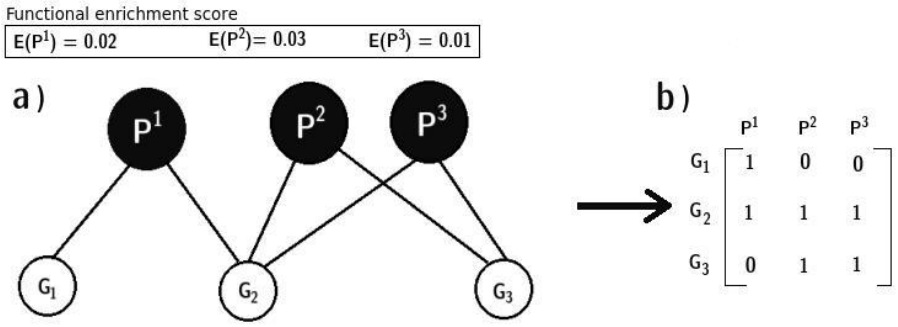

Figura 8: **a)** Grafo bipartido MGP com a representação de (G)enes e (P)athways (vias metabólicas). Acima, os valores do cálculo do enriquecimento funcional identificados para cada via. **b)** A matriz associada com as interações gene/grupo-vias metabólicas: (1) se existe relação anotada para o par gene/grupo-via; (0) se não existe relação.

A construção da rede além de ajudar na categorização dos elementos com base em medidas topológicas, permite também uma visualização das associações contidas na amostra em avaliação. A pesquisa da relação gene-vias, por exemplo, que considera as espécies identificadas, utiliza cores diferentes para destacar espécies distintas realçando as particularidades da amostra, como mostra a Figura 9.

O pacote *RedeR* (CASTRO et al., 2012) utilizado para visualização permite um sistema de visualização interativo, habilitando funções como: *zoom*, *pan*, destaque de vértices vizinhos, buscas, adição e deleção de componentes do grafo, etc.

5 Extração de medidas da rede

Após a construção da rede MGP as atenções são voltadas para a inspeção de pontos particulares, como a identificação de elementos *hub* ou vias importantes dentro da

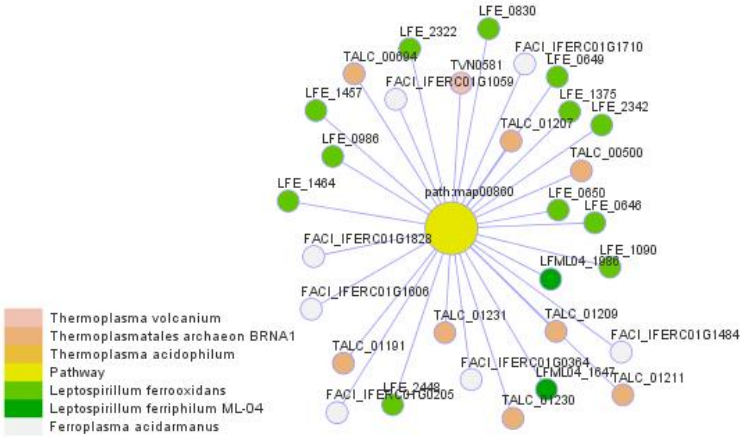

Figura 9: Visualização dos genes relacionados a via *Porphyrin and chlorophyll metabolism* (path:map00860) da amostra amostra AMD do trabaho de Tyson et al. (2004). A esquerda as espécies identificadas na amostras destacadas por cores distintas.

rede (BARABASI; OLTVAI, 2004). Estes pontos são explorados com auxílio de algumas medidas topológicas utilizadas no estudo da teoria dos grafos.

FUNN-MG calcula estas medidas topológicas com auxílio da biblioteca *igraph* (CSARDI; NEPUSZ, 2006) da seguinte forma: i) grau de conectividade (número de arestas ligadas a um vértice); ii) a centralidade de intermediação (proporção de caminhos mais curtos que passam através de um vértice); e iii) a conectividade da vizinhança (o número médio de vizinhos dos vizinhos de um vértice).

Os resultados são destacados e servem como medidas básicas que nos permitem comparar e caracterizar diferentes tipos de redes complexas como: redes aleatórias, livre escalas, hierárquicas etc (BARABASI; OLTVAI, 2004, p. 105).

Outra importante medida extraída da rede se refere ao grau de resiliência de cada via metabólica, ou seja, a capacidade de cada via metabólica continuar conectada a rede mesmo com a perda ou desligamento de algum gene, ou grupo ortólogo.

O cálculo da medida de resiliência é feito com base em simulações de pertubação na rede metabólica (NAVLAKHA et al., 2014). Se tratando do grafo MGP, bipartido, a rede é submetida a deleções aleatórias no conjunto de genes, um a um. A cada rodada o grau de conectividade de cada via é calculado até que a via tenha obtido o grau nulo. O resultado do número de deleções necessárias para que a via esteja totalmente desconexa à rede é anotado. A Figura 10 exemplifica o cálculo da medida de resiliência adotado neste trabalho.

A Figura 10 a) mostra uma rede contendo cinco vias metabólicas (em verde) e três genes (em branco). As figuras 10 b), 10 c) e 10 d), mostram a deleção aleatória sequencial dos três genes da rede. A Tabela 7 mostra a resiliência de cada via e o grau de conectividade a cada deleção.

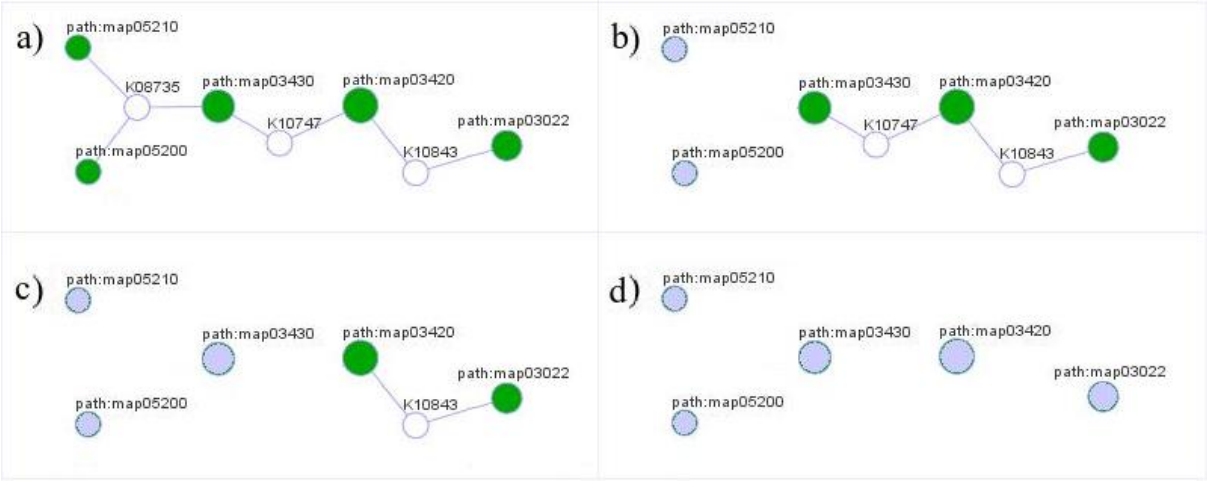

Figura 10: Teste de resiliência da rede executado pela ferramenta FUNN-MG para identificação da resiliência do conjunto de vias metabólicas. Os vértices em verde representam as vias metabólicas que possuem grau de conectividade maior que 0; os vértices em cinza representam as vias metabólicas com grau de conectividade igual a 0; os vértices em branco representam os genes identificados na amostra. 4 estágios (a, b, c, d) indicam a deleção simulada de um gene na rede monitorando o grau de conectividade das vias metabólicas.

O teste do modelo de deleção de genes é executado 100 vezes e a média do índice de resiliência é calculada.

A partir da média de deleção dos 100 testes do índice de resiliência, FUNN-MG calcula o índice que será utilizado para avaliação e seleção das vias. Existem duas hipóteses principais que são levadas em consideração para a construção do índice aqui chamado *funn-index*:

i) **O desligamento de conexões importantes da rede metabólica pode gerar um desequilíbrio das relações entre os grupos que compõe a comunidade**. Grupos de bactérias muitas vezes são responsáveis por funções metabólicas que podem gerar substratos importantes para a sobrevivência de outros grupos.

ii) Devido a grande diversidade de microrganismos que habitam a superfície da terra, existem funções distintas da maior parte de uma comunidade porém essenciais para

Tabela 7: Grau de conectividade de cada uma das vias da Figura 10 após as três deleções executadas. Em destaque a medida de resiliência ao final da simulação.

Via metabólica	Grau inicial	Deleção I	Deleção II	Deleção III	Resiliência obtida
path:map05210	1	0	0	0	**1**
path:map05200	1	0	0	0	**1**
path:map03430	2	1	0	0	**2**
path:map03420	2	2	1	0	**3**
path:map03022	1	1	1	0	**3**

explicar o comportamento fenotípico da relação comunidade-meio ambiente. **Poucas espécies, ou as vezes uma só, podem ter papel fundamental operando transformações a nível metabólico que podem ser periféricas a maior parte da rede de interação**.

Logo, a criação do índice foi feita a partir de dois indicadores: **(1) representatividade da via**, **(2) média dos testes do índice de resiliência**. O objetivo é encontrar um índice que possar ser representativo tanto ao nível da rede quanto ao nível da cobertura das vias. Então a fórmula para o *funn-index* é descrita da seguinte maneira:

Seja x o valor cálculo do teste hipergeométrico, e y o valor do cálculo da média do índice de resiliência, tem-se:

$$\textit{funn-index} = \log_{10} x \cdot y \tag{2.3}$$

A função logarítmica ajuda a normalizar os resultados do teste hipergeométrico, a fim de obter valores com poucas casas decimais e que indiquem uma relação direta de proporcionalidade no que diz respeito a importância das vias com base na cobertura.

Os resultados são normalizados em uma escala de 0 a 1, o score obtido é o fator de seleção das vias mais representativas na amostra. **Para garantir que todas as vias que obtiveram o pvalue $<$ 0.05 no teste Fisher estejam na seleção final, o *threshold* de corte é escolhido a partir do menor *funn-index* entre as vias classificadas pelo teste Fisher.**

3 Resultados

Para avaliação da ferramenta FUNN-MG foram utilizados dois conjuntos de amostras: Oceano (SUNAGAWA et al., 2015), e drenagem ácida de mina (TYSON et al., 2004). As escolhas das amostras foram feitas com base nas respectivas complexidades: baixa (drenagem ácida de mina), e alta (oceano). A complexidade das amostras é definida de acordo com o número de espécies dominantes, ou seja, a abundância relativa de cada espécie com relação ao conjunto de *reads* identificados na fase de sequenciamento (MAVROMATIS et al., 2007).

As amostras foram submetidas às duas abordagens disponibilizadas pela ferramenta FUNN-MG: i) a nível da interação entre genes-vias metabólicas; ii) a nível da interação entre grupos ortólogos-vias metabólicas. Os resultados obtidos foram comparados com análises executadas pela ferramenta MinPath (YE; DOAK, 2009), a partir do banco de dados KEGG. Os critérios para validação foram o número de vias enriquecidas pelas ferramentas, o modelo topológico da rede de interação (subseção 2.3), e a qualidade destas vias considerando o contexto biológico das amostras.

Todos os testes realizados nesta pesquisa foram executados em um terminal Linux Ubuntu 14.04.3 LTS, 8gb RAM, processador Intel(R) Core(TM) i5-3317U CPU @ 1.70GHz. A portabilidade da ferramenta para outros sistemas operacionais, e outras configurações de memória ainda estão em fase de teste.

1 Análise das amostras de Oceano do projeto Tara *ocean*

O projeto Tara *ocean* contou com uma expedição que explorou os oceanos do ano de 2009 à 2013 coletando amostras de vários pontos do mundo (PESANT et al., 2015). Para cada uma das 139 amostras recolhidas, foram extraídos *reads* a partir de sequenciadores de tecnologia Illumina, que contem assinatura para genes 16 sRNA (LOGARES et al., 2013).

Após a etapa de sequenciamento, os *reads* foram montados e mapeados para estimativa das abundâncias funcionais e taxonômicas de cada amostra utilizando a ferramenta MOCAT (KULTIMA et al., 2012). O conjunto não redundante de genes funcionais foi anotado a partir do banco de dados KEGG (v62) utilizando a ferramenta SMASH v1.6 (ARUMUGAM et al., 2010).

Os testes executados com a ferramenta FUNN-MG nesta fase utilizaram duas amostras da estação 4 (DCM e SUR) localizadas ao norte do oceano atlântico, sul de Portugal.

A Figura 11 mostra a quantidade de vias identificadas na estação 4 (amostras SUR e DCM), considerando cada abordagem, bem como a frequência dos graus dos vértices identificados na rede MGP.

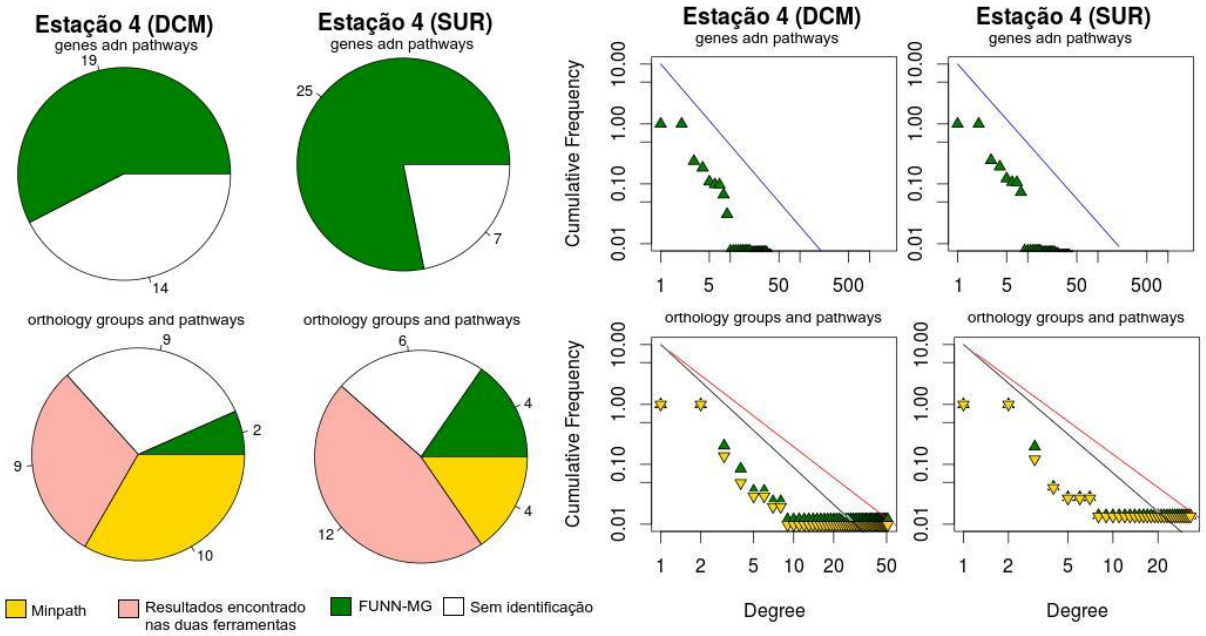

Figura 11: Análise da comparação entre os resultados das amostras 4 DCM e 4 SUR, obtidos a partir das ferramentas FUNN-MG (considerando genes e grupos ortólogos), e MinPath. Na parte de cima da imagem, os gráficos representam a análise a nível da interação entre genes e vias metabólicas, na parte de baixo a interação entre grupos ortólogos e vias metabólicas. Ao lado esquerdo os *piecharts* mostram a quantidade de vias metabólicas identificadas após o enriquecimento funcional em cada ferramenta: MinPath (amarelo), FUNN-MG (verde), nas duas abordagens (rosa), e as vias não identificadas em nenhuma das duas ferramentas (branco). Ao lado direito é mostrado a frequência acumulativa dos graus identificados nas redes a partir das vias resultantes das duas análises, MinPath em amarelo, e FUNN-MG em verde. As retas que cortam os gráficos representam o comportamento *power law* esperado nas redes mapeadas para as duas ferramentas (em vermelho e azul para ferramenta FUNN-MG, e em preto para ferramenta MinPath).

No primeiro momento da análise é possível visualizar que o número de vias encontradas, quando considerado a análise a nível do conjunto de genes, é maior que a análise via grupos ortólogos. Isso se deve ao fato de que os grupos ortólogos são grupos de genes que podem pertencer a uma lista de espécies, logo é menos comum que participem de vias metabólicas mais específicas.

Em um segundo momento, observando os gráficos à direita da Figura 11, nota-se que a frequência relativa dos graus de conectividade dos elementos da rede, indicam que existem muitos elementos com baixo grau de conectividade em detrimento de poucos elementos com alto grau de conectividade. Este comportamento não uniforme é comum em redes que são caracterizadas por um grau de distribuição em lei de potência (*power law*).

As retas que cortam os gráficos são indicadores do comportamento padrão de distribuições em lei de potência para cada conjunto de vértices identificados em cada análise.

Muito embora esta afirmação ainda possa ser debatida na comunidade científica, acredita-se que a maioria das redes biológicas possuem caraterísticas topológicas de rede de livre escala, que são caracterizadas por terem seu grau de distribuição regido por uma lei de potência (BARABASI; OLTVAI, 2004, p. 104).

1.1 Amostra 4 DCM

No próximo exemplo (Figura 12) destacamos uma sub-rede obtida a partir da amostra 4 DCM. Comparou-se os resultados das vias metabólicas obtidas com as duas ferramentas: MinPath (à esquerda), e FUNN-MG (à direita).

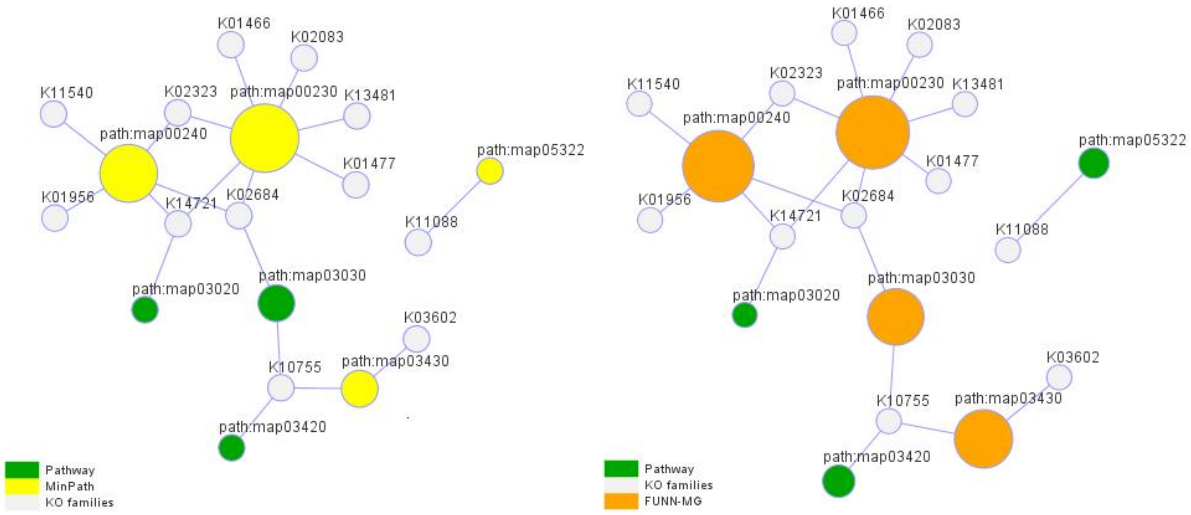

Figura 12: Sub-rede da amostra 4 DCM considerando a análise de grupos ortólogos da ferramenta MinPath (à esquerda) e da ferramenta FUNN-MG (à direita). Em verde as vias metabólicas não enriquecidas, em amarelo as vias metabólicas enriquecidas pela ferramenta MinPath, em laranjado as vias metabólicas enriquecidas pela ferramenta FUNN-MG, em branco os grupos ortólogos anotados no banco de dados KEGG identificados na amostra. Destaque para os tamanhos dos vértices que destacam a importância da via segundo análise da de cada ferramenta (no caso da ferramenta MinPath o tamanho do vértice é relacionado com o grau de conectividade).

A rede *MGP* da amostra 4 DCM conta com 83 grupos ortólogos ligados a 30 vias metabólicas distintas (Apêndice A, Tabela 13). A sub-rede em destaque na Figura 12 é um subconjunto da amostra 4 DMC, e contém 7 vias metabólicas ligadas à 12 grupos ortólogos. Neste exemplo as duas ferramentas identificaram 4 vias enriquecidas na análise, no entanto não houve consenso na escolha, como mostra a Tabela 8.

Logo abaixo a descrição básica das vias identificadas na sub-rede.

Tabela 8: Identificação das funções das vias metabólicas da Figura 12. Em rosa as vias enriquecidas pelas duas ferramentas, em laranjado a via enriquecida somente pela ferramenta FUNN-MG, e em amarelo a via enriquecida somente pela ferramenta MinPath.

	Id	Função	Categoria
1	path:map00230	Purine metabolism	Nucleotide metabolism
2	path:map00240	Pyrimidine metabolism	Nucleotide metabolism
3	path:map03020	RNA polymerase	Genetic Information Processing
4	**path:map03030**	**DNA replication**	**Genetic Information Processing**
5	path:map03420	Nucleotide excision repair	Genetic Information Processing
6	path:map03430	Mismatch repair	Genetic Information Processing
7	**path:map05322**	**Systemic lupus erythematosus**	**Human Diseases**

- ***Purine and Pyrimidine metabolism***: pirimidinas e purinas são bases nitrogenadas de estruturas heterocíclicas que compõe o nucleotídeo. Adenina (A) e Guanina (G) são purinas que ligam-se às pirimidinas Citosina (C), Timina (T), ou uracila (U).

- ***RNA polymerase***: RNAs polimerases são enzimas responsáveis pela síntese de RNA a partir de sequências de DNA ou RNA.

- ***DNA replication***: corresponde a fase de replicação da molécula de DNA e RNA, que ocorre quando pirimidinas se unem às purinas correspondentes através de ligações de hidrogênio, denominadas emparelhamento de base.

- ***Mismatch repair*** e ***Nucleotide excision repair***: tratam sobre a reparação de emparelhamentos incorretos do DNA. Basicamente, um sistema de reconhecimento e reparação de inserção errada, deleção, e má incorporação de bases que podem surgir durante a fase de recombinação.

- ***Systemic lupus erythematosus***: Lúpus Eritematoso Sistêmico (LES) é uma doença auto-imune em que o sistema imunitário do corpo ataca erroneamente o tecido saudável. Ela pode afetar a pele, articulações, rins, cérebro e outros órgãos.

As categorias apresentadas na Tabela 8 foram obtidas segundo a referência do banco de dados KEGG. Na subamostra da Figura 12, a via *Systemic lupus erythematosus* se distingue das demais vias pois é a única que corresponde a uma via específica de doenças humanas, o que não faz sentido se tratando de amostras de microrganismos.

Duas hipóteses podem explicar o erro no mapeamento da via: erro de anotação à nível de homologia, ou erro de anotação à nível de identificação de relações metabólicas.

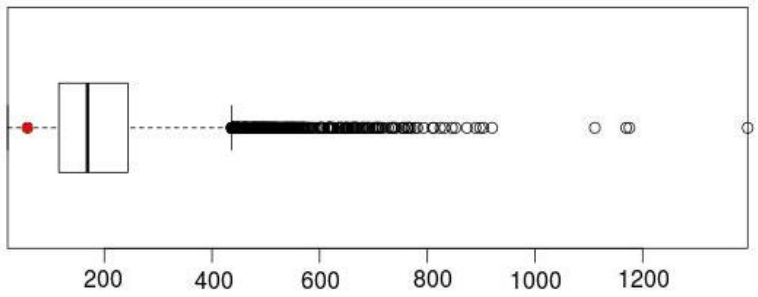

Figura 13: Boxplot da distribuição dos scores de anotação da amostra 4 DCM, em destaque em vermelho o score de anotação pro grupo K11088, único relacionado a via *Systemic lupus erythematosus*.

Na Figura 13, o boxplot mostra a distribuição dos *scores* de anotação do conjunto de grupos ortólogos da amostra 4 DCM, e em vermelho o grupo ortólogo K11088 (*score* = 46.595) único relacionado a via *Systemic lupus erythematosus* na amostra. Os *scores* de anotação são percentuais de acertos relacionados a fase de consulta por homologia das sequências no banco de dados KEGG. Analisando o boxplot é possível observar que o *score* de anotação do grupo K11088 é muito inferior a média da amostra, o que corrobora a hipótese de que nesse caso tenha havido erro de anotação da sequência.

1.2 Amostra 4 SUR

No segundo exemplo para amostras da estação 4, destacamos uma sub-rede obtida a partir da amostra 4 SUR (Figura 14). Mais uma vez comparou-se os resultados das

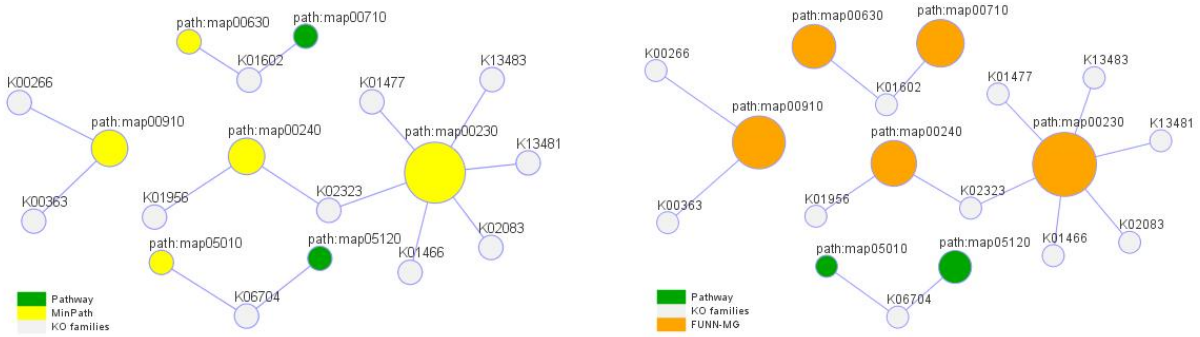

Figura 14: Sub-rede da amostra 4 SUR considerando a análise de grupos ortólogos da ferramenta MinPath (à esquerda) e da ferramenta FUNN-MG (à direita). Em verde as vias metabólicas não enriquecidas, em amarelo as vias metabólicas enriquecidas pela ferramenta MinPath, em laranjado as vias metabólicas enriquecidas pela ferramenta FUNN-MG, em branco os grupos ortólogos anotados no banco de dados KEGG identificados na amostra. Destaque para os tamanhos dos vértices que destacam a importância da via segundo análise da de cada ferramenta.

vias metabólicas obtidas com as duas ferramentas: MinPath (à esquerda), e FUNN-MG (à direita). Foram identificados 58 grupos ortólogos relacionados a 26 vias metabólicas distintas (Apêndice A Tabela 12). A Figura 12 é uma sub-rede extraída que contém 7 vias metabólicas e 11 grupos ortólogos. Os resultados são descritos na Tabela 9, e a definição básica das vias identificadas é descrita logo abaixo.

Tabela 9: Identificação das funções das vias metabólicas da Figura 14. Em rosa as vias enriquecidas pelas duas ferramentas, em laranjado a via enriquecida somente pela ferramenta FUNN-MG, e em amarelo a via enriquecida somente pela ferramenta MinPath.

	Id	Função	Categoria
1	path:map00230	Purine metabolism	Nucleotide metabolism
2	path:map00240	Pyrimidine metabolism	Nucleotide metabolism
3	path:map00630	Glyoxylate and dicarboxylate metabolism	Carbohydrate metabolism
4	**path:map00710**	**Carbon fixation in photosynthetic organisms**	**Energy metabolism**
5	path:map00910	Nitrogen metabolism	Energy metabolism
6	**path:map05010**	**Alzheimer's disease**	**Human Diseases**
7	path:map05120	Epithelial cell signaling in Helicobacter pylori infection	Infectious diseases

- ***Glyoxylate and dicarboxylate metabolism***: descreve uma variedade de reações envolvendo glioxilato e dicarboxilato. Glioxilato é a base conjugada do ácido glioxílico. Da mesma forma descarboxila são as bases conjugadas de ácidos dicarboxílicos, uma classe geral de compostos orgânicos que contêm dois grupos de ácido carboxílico, tais como ácido oxálico ou ácido succínico.

- ***Carbon fixation in photosynthetic organisms***: é a denominação para o processo que remete ao fato do organismo absorver dióxido de carbono. Este processo ocorre graças ao Ciclo de Calvin (ou como é mais comumente conhecido: fase escura da fotossíntese), onde o dióxido de carbono é ligado à um receptor, originando como produto de uma reação de redução uma molécula de CH_2O (um carboidrato). Porém, sinteticamente podemos dizer que o ciclo de Calvin consiste numa cadeia de reações químicas que utilizam as moléculas de CO_2 para formar moléculas mais complexas, tais como aminoácidos, ácidos graxos e carboidratos.

- ***Nitrogen metabolism***: o processo biológico do ciclo do nitrogênio é uma interação complexa entre muitos microrganismos que catalisam reações diferentes, onde o nitrogênio é encontrado em vários estados de oxidação que variam de +5 em nitrato a -3 em amônia. O processo de redução do nitrogênio molecular atmosférico em amônia, é uma forma biologicamente útil incorporada à aminoácidos e outros compostos vitais.

- ***Alzheimer's disease***: A doença de Alzheimer (DA) é uma doença crônica que destrói lentamente neurônios e provoca grave deficiência cognitiva.
- ***Epithelial cell signaling in Helicobacter pylori infection***: H. pylori é uma bactéria em forma de espiral que cresce no trato digestivo e tem uma tendência para atacar o revestimento do estômago.

Na subamostra em destaque na Figura 14, temos duas vias muito importantes: *Carbon fixation in photosynthetic organisms* e *Nitrogen metabolism*, que possivelmente tem relações com comunidades de plânctons que habitam os oceanos (VARGAS et al., 2015). De outro modo, a ferramenta MinPath identificou a via *Alzheimer's disease*, que por se tratar de doença degenerativa humana não faz sentido de ser mapeada na amostra.

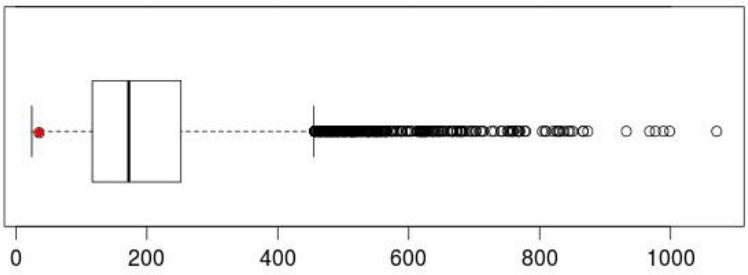

Figura 15: Boxplot da distribuição dos scores de anotação da amostra 4 SUR, em destaque em vermelho o score de anotação pro grupo K06704, único relacionado a via *Alzheimer's disease*.

A Figura 15 mostra a distribuição dos *scores* de anotação da amostra 4 SUR, em vermelho o *score* de anotação do grupo K06704 (*score* = 28.490), o único que está ligado a via *Alzheimer's disease*.

Assim como na amostra 4 DCM, acredita-se que um erro de anotação pode ser responsável pelo mapeamento da via. Erros de anotação são muito comuns em amostras de alta complexidade. As amostras 4 DCM e 4 SUR possuem respectivamente 584 e 548 espécies de procariotos identificadas, nenhuma das espécies é dominante na amostra. Sequências obtidas a partir de um sistema complexo sem espécie dominante não contem grandes fragmentos genômicos de qualquer população componente utilizando tecnologias atuais (KUNIN et al., 2008, p. 7), a identificação de genes com fragmentos menores potencializa o erro na fase de anotação.

2 Amostra de drenagem ácida de mina

O segundo conjunto de dados metagenômicos selecionados para estudo experimental é o biofilme de Drenagem Ácida de Mina (AMD) (NCBI, 2006). Este projeto de sequenciamento de biofilme foi projetado para explorar a distribuição e a diversidade

de vias metabólicas em biofilmes acidófilos. O metagenoma AMD foi montado em 2425 *contigs* distribuídos ao longo de cinco espécies principais como mostra a Tabela 10.

Tabela 10: Número de *contigs* por espécie da amostra AMD identificados no trabalho de Tyson et al. (2004).

Nome das espécies	Número de contigs
Ferroplasma acidarmanus Type I	412
Ferroplasma sp. Type II	118
Leptospirillum sp. Group II 5-way CG	79
Leptospirillum sp. Group III	959
Thermoplasmatales archaeon Gpl	857

A primeira etapa de anotação envolveu consultas de homologia realizadas a partir da ferramenta KAAS, que identificou 383 grupos ortólogos não redundantes na amostra. Na segunda parte, os genes relacionados aos grupos ortólogos que pertenciam às espécies análogas ao banco de dados KEGG foram filtrados, totalizando 477 genes para seis espécies distintas, como mostra a Figura 16.

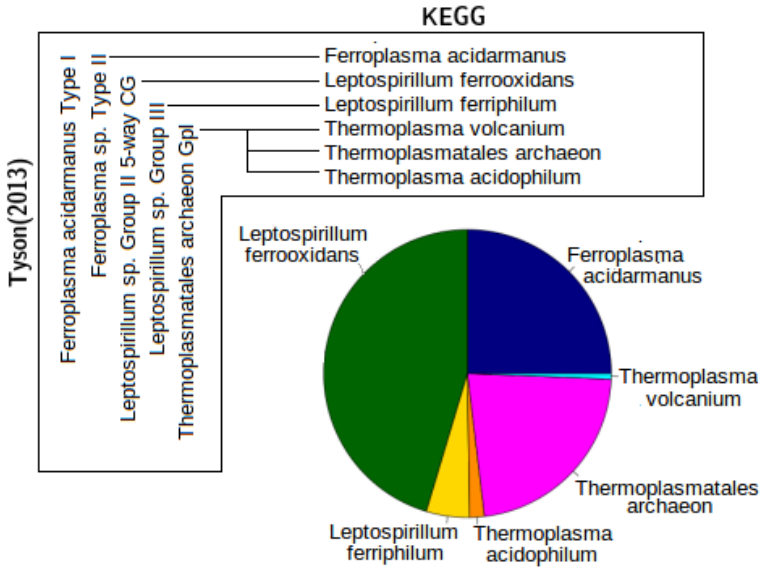

Figura 16: O dendrograma no topo destaca a associação entre espécies no metagenoma AMD e suas espécies-alvo no banco de dados KEGG. Ao fundo, um *piechart* da distribuição dos 477 genes identificados na amostra.

Inicialmente 117 vias metabólicas foram mapeadas (Apêndice B, Tabela 13, 14, 15 e 16) a partir dos 383 grupos ortólogos destacados. Destas, 55 vias foram identificadas pelas duas ferramentas, 10 vias foram identificadas somente pela ferramenta MinPath, 31 vias somente pela ferramenta FUNN-MG, e 21 vias não foram identificadas por nenhuma das duas ferramentas, como mostra a Figura 17.

Desta vez, a identificação de vias pela metodologia utilizada nesta pesquisa foi maior do que a análise feita pela ferramenta MinPath. A explicação se deve ao fato da

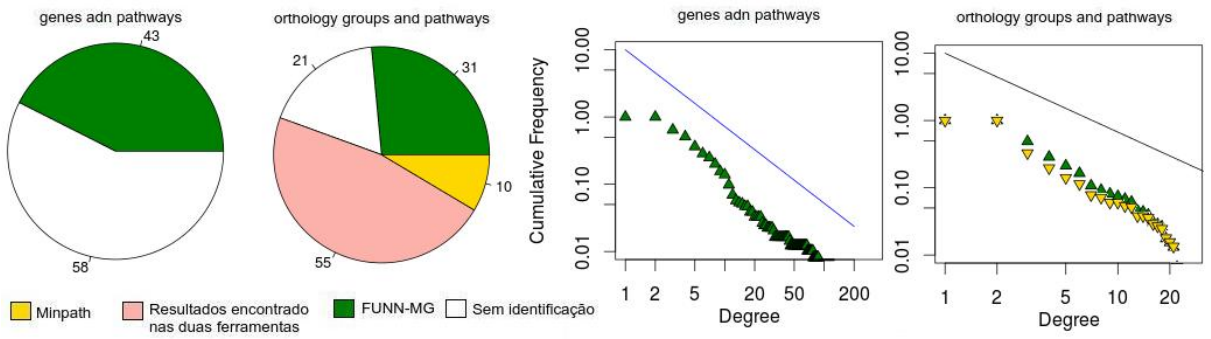

Figura 17: Análise dos resultados das amostras AMD (NCBI, 2006) obtidos a partir das ferramentas FUNN-MG (considerando genes e grupos ortólogos), e MinPath. Ao lado esquerdo os *piecharts* mostram a quantidade de vias metabólicas identificadas após o enriquecimento funcional em cada ferramenta: mais à esquerda considerando a interação genes-vias metabólicas, e ao lado a interação grupos ortólogos-vias metabólicas. Ao lado direito é mostrado a frequência acumulativa dos graus identificados nas redes a partir das vias resultantes das duas análises, MinPath em amarelo, e FUNN-MG em verde. As retas que cortam os gráficos representam o comportamento *power law* esperado nas redes mapeadas para as duas ferramentas.

maior interação entre os elementos da rede, o que de certa forma aumenta os respectivos índices de resiliência que contribui diretamente para aumento do *funn-index*, que é o indicador de enriquecimento para as vias.

Os resultados obtidos nas duas análises também indicam o comportamento *power law* dos elementos da rede MGP da amostra AMD. Nota-se na Figura 17 (à direita), que a distribuição da frequência acumulada dos vértices que compõe a rede, tem o comportamento semelhante às retas que indicam o comportamento padrão de distribuições *power law*.

Com a diminuição da complexidade da amostra é perceptível um comportamento mais linear na distribuição dos vértices na rede. No exemplo das amostras 4 DCM e 4 SUR da Figura 11, é visível a divisão dos elementos da rede em dois grupos bem distintos: vértices altamente conectados, e vértices com grau de distribuição bem próximo ao eixo x. Neste exemplo, a amostra AMD apresenta grupos de vértices com grau de conexão intermediário, localizados entre os dois extremos. O comportamento também é característico de redes de livre escala, e é explicado devido o aumento do número de vias metabólicas identificadas.

No próximo exemplo, a Figura 18 mostra a rede MGP da amostra AMD. Em destaque os genes envolvidos na via metabólica enriquecida *Reductive carboxylate cycle (CO2 fixation)*.

A via em questão foi identificada somente pela ferramenta FUNN-MG, nas duas análises (genes e grupos ortólogos). Na análise do conjunto de genes, 9 genes da espécie

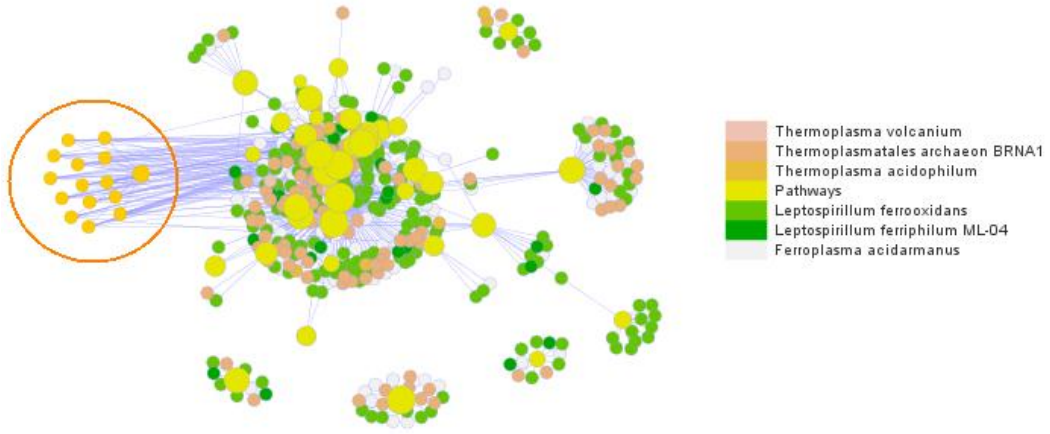

Figura 18: *Microbial Genes Pathways network* da amostra ADM (NCBI, 2006). Do lado esquerdo destaca-se os genes identificados relacionados à via enriquecida *Reductive carboxylate cycle (CO2 fixation)*. À diretia a legenda dos vértices segundo as espécies identificadas na amostra.

Ferroplasma acidarmanus e 5 genes da espécie *Thermoplasma acidophilum* estão correlacionados à via.

Os biofilmes acidófilos são comunidades auto-sustentáveis que crescem no subsolo profundo e não recebem entradas significativas de fixação de carbono ou de nitrogênio provenientes de fontes externas. A via metabólica *Reductive carboxylate cycle (CO2 fixation)* encontrada em procariotos, metaboliza o CO_2 em componentes orgânicos atavés do ciclo de Calvin-Benson-Bassham. Mais informações sobre a via podem ser obtidas diretamente no catálogo de vias metabólicas do banco de dados KEGG, no endereço: <http://www.genome.jp/dbget-bin/www_bget?map00720>.

Algumas vias metabólicas sem aparente ligação com o ambiente AMD também foram encontradas pela ferramenta FUNN-MG, como a via *Pathways in cancer* categorizada como *doença humana* pelo banco de dados KEGG. Quando há um grande número

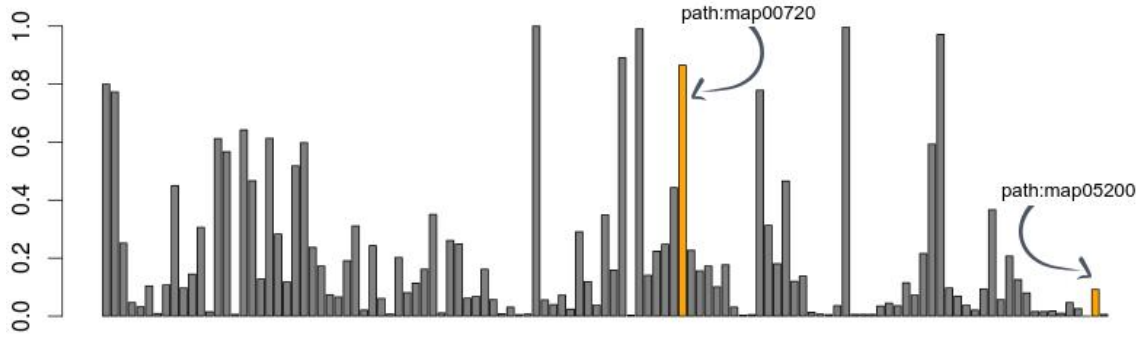

Figura 19: Distribuição dos índices de eriquecimento funcional obtidos com a ferramenta FUNN-MG para o conjunto de 117 vias da amostra AMD. Em destaque as vias *Reductive carboxylate cycle (CO2 fixation)* (path:map00720) e *Pathways in cancer* (path:map05200).

de conexões na rede, o *threshold* com base nos valores de confiança obtidos com o teste

Fisher pode superestimar o conjunto de vias. Neste exemplo a via *Reductive carboxylate cycle (CO2 fixation)* possui um funn-index = 0.865, ao passo que a via *Pathways in cancer* possui funn-index = 0.092. Seus respectivos funn-idex são discrepantes, como mostra a Figura 19, no entanto as duas se enquadram como possíveis vias pertencentes à comunidade.

Análises biológicas que consideram as espécies e os metadados locais podem ajudar a definir um *threshold* mais próximo ao perfil metabólico das amostras.

4 Conclusões

1 Discussão dos resultados

A proposta deste estudo foi avaliar o perfil das comunidades microbianas, destacando as vias metabólicas mais relevantes com base na topologia da rede de interação gene(ou grupo ortólogo)-via metabólica. Foi submetido três amostras de complexidades diferentes para análise da estratégia proposta, alguns pontos merecem destaque:

i) A estratégia adotada que considera tanto a abundância relativa das vias, como a natureza topológica da rede de interação obteve resultados satisfatórios, identificando vias metabólicas importantes no contexto das amostras utilizadas e descartando vias sem sentido no contexto biológico. Em alguns casos, conseguiu-se identificar possíveis erros de anotação a partir da análise de representatividade das vias metabólicas.

ii) Uma limitação encontrada foi a definição de um *threshold* que tornasse viável a identificação do conjunto de vias em amostras com baixo e alto grau de complexidade. Para amostras de baixas complexidade o aumento do conjunto de vias metabólicas influenciou os resultados da análise o que impactou no alto número de vias enriquecidas (75,5%). No entanto o score apresentado na pesquisa apresentou discrepâncias entre as vias que apresentavam e as que não apresentavam ligações com o contexto biológico das amostras.

iii) A visualização da relação entre genes, grupos ortólogos e vias metabólicas no modelo de grafos, ajudou a destacar alguns resultados importantes apresentados no trabalho, principalmente na identificação visual das vias menos enriquecidas através do tamanho dos vértices da rede.

2 Trabalhos futuros

Como perspectiva de próximas abordagens, pretendemos observar, utilizando o índice de enriquecimento proposto nesta pesquisa, sub-redes metabólicas amostradas aleatoriamente a partir de uma rede em escala metagenômica.

Nós identificamos que no decorrer da pesquisa em amostras de baixa complexidade, por vezes um gene é relacionado a várias vias metabólicas distintas. É sabido que existem genes responsáveis por várias funções (GIANOULIS et al., 2009), no entanto, é pouco provável que todas essas funções sejam de fato exercidas pela comunidade.

A análise destas sub-redes levarão em conta que cada gene (ou grupo ortólogo) deve ser responsável por apenas uma função metabólica na rede. As vias mais relevantes serão comparadas com a abordagem atual proposta nesta pesquisa.

Referências

ABUBUCKER, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology*, Public Library of Science, v. 8, n. 6, p. e1002358+, jun. 2012. ISSN 1553-7358. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.1002358>. Citado na página 22.

ARUMUGAM, M. et al. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, Oxford University Press, v. 26, n. 23, p. 2977–2978, dez. 2010. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btq536>. Citado 2 vezes nas páginas 24 e 39.

BäCKHED, F. et al. Host-Bacterial Mutualism in the Human Intestine. *Science*, American Association for the Advancement of Science, Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108, USA., v. 307, n. 5717, p. 1915–1920, mar. 2005. ISSN 1095-9203. Disponível em: <http://dx.doi.org/10.1126/science.1104816>. Citado na página 16.

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, Nature Publishing Group, v. 5, n. 2, p. 101–113, fev. 2004. ISSN 1471-0056. Disponível em: <http://dx.doi.org/10.1038/nrg1272>. Citado 4 vezes nas páginas 22, 34, 36 e 41.

CASTRO, M. A. et al. RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome biology*, v. 13, n. 4, p. R29+, abr. 2012. ISSN 1465-6914. Disponível em: <http://dx.doi.org/10.1186/gb-2012-13-4-r29>. Citado na página 35.

CORRÊA, L. et al. Funn-mg: A metagenomic systems biology computational framework. In: CAMPOS, S. (Ed.). *Advances in Bioinformatics and Computational Biology*. Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8826). p. 25–32. ISBN 978-3-319-12417-9. Disponível em: <http://dx.doi.org/10.1007/978-3-319-12418-6_4>. Citado na página 26.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006. Disponível em: <http://igraph.org>. Citado na página 36.

DINSDALE, E. A. et al. Multivariate analysis of functional metagenomes. *Frontiers in genetics*, v. 4, 2013. ISSN 1664-8021. Disponível em: <http://dx.doi.org/10.3389/fgene.2013.00041>. Citado na página 17.

FIERER, N. et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 109, n. 52, p. 21390–21395, dez. 2012. ISSN 1091-6490. Disponível em: <http://dx.doi.org/10.1073/pnas.1215210110>. Citado na página 16.

FINN, R. D. et al. The Pfam protein families database. *Nucleic acids research*, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton Hall, Hinxton,

Cambridgeshire, CB10 1SA, UK., v. 36, n. Database issue, p. D281–D288, jan. 2008. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkm960>. Citado na página 19.

FRANCKE, C.; SIEZEN, R. J.; TEUSINK, B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, v. 13, n. 11, p. 550–558, nov. 2005. ISSN 0966842X. Disponível em: <http://dx.doi.org/10.1016/j.tim.2005.09.001>. Citado na página 24.

GIANOULIS, T. A. et al. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of Sciences, v. 106, n. 5, p. 1374–1379, fev. 2009. ISSN 1091-6490. Disponível em: <http://dx.doi.org/10.1073/pnas.0808022106>. Citado na página 50.

GLASS, E. M. et al. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor protocols*, v. 2010, n. 1, p. pdb.prot5368+, jan. 2010. ISSN 1559-6095. Disponível em: <http://dx.doi.org/10.1101/pdb.prot5368>. Citado na página 22.

GOH, K.-I. et al. The human disease network. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 104, n. 21, p. 8685–8690, maio 2007. ISSN 0027-8424. Disponível em: <http://dx.doi.org/10.1073/pnas.0701361104>. Citado na página 24.

GOLL, J. et al. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics*, Oxford University Press, v. 26, n. 20, p. 2631–2632, out. 2010. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btq455>. Citado na página 24.

GROSS, J. J.; YELLEN. *Handbook of graph theory*. 5. ed. [S.l.: s.n.], 2004. v. 1. Citado na página 21.

HUBER, W. et al. Graphs in molecular biology. *BMC Bioinformatics*, European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge CB10 1SD, UK. huber@ebi.ac.uk., v. 8, n. Suppl 6, p. S8+, 2007. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-8-s6-s8>. Citado 2 vezes nas páginas 21 e 34.

HUGENHOLTZ, P.; TYSON, G. W. Microbiology: Metagenomics. *Nature*, Nature Publishing Group, v. 455, n. 7212, p. 481–483, set. 2008. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/455481a>. Citado na página 16.

HUSON, D. H. et al. Megan analysis of metagenomic data. *Genome research*, Cold Spring Harbor Lab, v. 17, n. 3, p. 377–386, 2007. Citado na página 22.

ITAI; SHANRON. *Computational Methods forMetagenomic Analysis*. Tese (Doutorado) — Israel Institute of Tecnology, 2010. Citado 4 vezes nas páginas 19, 20, 23 e 31.

JEONG, H. et al. The large-scale organization of metabolic networks. *Nature*, Nature Publishing Group, Department of Physics, University of Notre Dame, Indiana 46556, USA., v. 407, n. 6804, p. 651–654, out. 2000. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/35036627>. Citado na página 22.

JOHNSTON, C. W. et al. Gold biomineralization by a metallophore from a gold-associated microbe. *Nat Chem Biol*, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., advance online publication, fev. 2013. Disponível em: <http://dx.doi.org/10.1038/nchembio.1179>. Citado na página 16.

KANEHISA, M. et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, Oxford University Press, v. 40, n. Database issue, p. D109–D114, jan. 2012. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkr988>. Citado 2 vezes nas páginas 17 e 27.

KENT, J. J. BLAT - the BLAST-like alignment tool. *Genome research*, Cold Spring Harbor Laboratory Press, Department of Biology and Center for Molecular Biology of RNA, University of California-Santa Cruz, Santa Cruz, CA 95064, USA. kent@biology.ucsc.edu, v. 12, n. 4, p. 656–664, abr. 2002. ISSN 1088-9051. Disponível em: <http://dx.doi.org/10.1101/gr.229202>. Citado na página 29.

KULTIMA, J. R. et al. MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS ONE*, Public Library of Science, v. 7, n. 10, p. e47656+, out. 2012. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0047656>. Citado na página 39.

KUNIN, V. et al. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, American Society for Microbiology, v. 72, n. 4, p. 557–578, dez. 2008. ISSN 1098-5557. Disponível em: <http://dx.doi.org/10.1128/mmbr.00009-08>. Citado na página 45.

LACKNER, G. et al. Complete Genome Sequence of Burkholderia rhizoxinica, an Endosymbiont of Rhizopus microsporus. *Journal of bacteriology*, v. 193, n. 3, p. 783–784, fev. 2011. ISSN 1098-5530. Disponível em: <http://dx.doi.org/10.1128/jb.01318-10>. Citado na página 21.

LOGARES, R. et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol*, p. n/a, out. 2013. Disponível em: <http://dx.doi.org/10.1111/1462-2920.12250>. Citado na página 39.

MARKOWITZ, V. M. et al. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research*, v. 36, n. suppl 1, p. D534–D538, jan. 2008. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkm869>. Citado 2 vezes nas páginas 19 e 24.

MAVROMATIS, K. et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Meth*, Nature Publishing Group, Department of Energy Joint Genome Institute (DOE-JGI), 2800 Mitchell Drive, Walnut Creek, California 94598, USA., v. 4, n. 6, p. 495–500, jun. 2007. ISSN 1548-7091. Disponível em: <http://dx.doi.org/10.1038/nmeth1043>. Citado 2 vezes nas páginas 27 e 39.

MITRA, S. et al. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, v. 12, n. Suppl 1, p. S21+, 2011. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-12-s1-s21>. Citado na página 22.

MORIYA, Y. et al. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan., v. 35, n. Web Server issue, p. W182–W185, jul. 2007. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkm321>. Citado 2 vezes nas páginas 22 e 29.

NAVLAKHA, S. et al. Topological properties of robust biological and computational networks. *Journal of The Royal Society Interface*, The Royal Society, v. 11, n. 96, p. 20140283+, jul. 2014. ISSN 1742-5662. Disponível em: <http://dx.doi.org/10.1098/rsif.2014.0283>. Citado na página 36.

NCBI, B. M. N. C. for B. I. *Metagenomics: Sequences from the Environment [Internet]*. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK6860/>. Citado 5 vezes nas páginas 10, 45, 47, 48 e 61.

NEISH, A. S. Microbes in gastrointestinal health and disease. *Gastroenterology*, v. 136, n. 1, p. 65–80, jan. 2009. ISSN 1528-0012. Disponível em: <http://dx.doi.org/10.1053/j.gastro.2008.10.080>. Citado na página 20.

OBERHARDT, M. A. et al. Genome-Scale Metabolic Network Analysis of the Opportunistic Pathogen Pseudomonas aeruginosa PAO1. *Journal of Bacteriology*, American Society for Microbiology, v. 190, n. 8, p. 2790–2803, abr. 2008. ISSN 1098-5530. Disponível em: <http://dx.doi.org/10.1128/jb.01583-07>. Citado na página 24.

OVERBEEK, R. et al. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research*, Oxford University Press, Fellowship for Interpretation of Genomes, 15W155 81st Street, Burr Ridge, IL 60527, USA., v. 33, n. 17, p. 5691–5702, jan. 2005. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gki866>. Citado na página 22.

PESANT, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, v. 2, p. 150023+, maio 2015. ISSN 2052-4463. Disponível em: <http://dx.doi.org/10.1038/sdata.2015.23>. Citado na página 39.

PRAKASH, T.; TAYLOR, T. D. Functional assignment of metagenomic data: challenges and applications. *Briefings in Bioinformatics*, Oxford University Press, v. 13, n. 6, p. 711–727, jul. 2012. ISSN 1477-4054. Disponível em: <http://dx.doi.org/10.1093/bib/bbs033>. Citado 3 vezes nas páginas 16, 17 e 19.

PYLRO, V. S. et al. Data analysis for 16s microbial profiling from different benchtop sequencing platforms. *Journal of Microbiological Methods*, v. 107, p. 30 – 37, 2014. ISSN 0167-7012. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167701214002528>. Citado na página 25.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Citado na página 30.

SELENGUT, J. D. et al. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Research*, Oxford University Press, v. 35, n. suppl 1, p. D260–D264, jan. 2007. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkl1043>. Citado na página 19.

SREENIVASAIAH, P. K. K. et al. IPAVS: Integrated Pathway Resources, Analysis and Visualization System. *Nucleic acids research*, Oxford University Press, v. 40, n. Database issue, p. D803–D808, jan. 2012. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkr1208>. Citado na página 30.

STOREY, J. *qvalue: Q-value estimation for false discovery rate control*. [S.l.], 2015. R package version 2.0.0. Disponível em: <http://qvalue.princeton.edu/,http://github.com/jdstorey/qvalue>. Citado na página 33.

SUN, S. et al. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Research*, Oxford University Press, v. 39, n. suppl 1, p. D546–D551, jan. 2011. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkq1102>. Citado na página 19.

SUNAGAWA, S. et al. Structure and function of the global ocean microbiome. *Science*, American Association for the Advancement of Science, v. 348, n. 6237, p. 1261359+, maio 2015. ISSN 1095-9203. Disponível em: <http://dx.doi.org/10.1126/science.1261359>. Citado 3 vezes nas páginas 27, 31 e 39.

TENENBAUM, D. *KEGGREST: Client-side REST access to KEGG*. [S.l.], 2013. R package version 1.8.0. Disponível em: <http://www.bioconductor.org/packages/release/bioc/html/KEGGREST.html/>. Citado na página 30.

TREANGEN, T. et al. Metamos: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, v. 14, n. 1, p. R2, 2013. ISSN 1465-6906. Disponível em: <http://genomebiology.com/2013/14/1/R2>. Citado na página 24.

TYSON, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, v. 428, n. 6978, p. 37–43, mar. 2004. ISSN 1476-4687. Disponível em: <http://dx.doi.org/10.1038/nature02340>. Citado 9 vezes nas páginas 8, 9, 11, 25, 27, 33, 36, 39 e 46.

VARGAS, C. D. et al. Eukaryotic plankton diversity in the sunlit ocean. *Science*, American Association for the Advancement of Science, v. 348, n. 6237, p. 1261605, 2015. Citado na página 45.

VEY, G.; MORENO-HAGELSIEB, G. Metagenomic annotation networks: construction and applications. *PloS one*, Public Library of Science, v. 7, n. 8, p. e41283, 2012. Citado na página 24.

WHEELER, D. L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA., v. 36, n. Database issue, jan. 2008. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkm1000>. Citado na página 19.

WOOLEY, J. C.; GODZIK, A.; FRIEDBERG, I. A Primer on Metagenomics. *PLoS Comput Biol*, Public Library of Science, v. 6, n. 2, p. e1000667+, fev. 2010. ISSN 1553-7358. Disponível em: <http://dx.doi.org/10.1371/journal.pcbi.1000667>. Citado 2 vezes nas páginas 16 e 18.

YE, Y.; DOAK, T. G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS computational biology*, Public Library of Science, v. 5, n. 8, p. e1000465, 2009. Citado 4 vezes nas páginas 22, 23, 27 e 39.

Anexos

ANEXO A – Resultados (1)

Comparação entre as ferramenta minPath e FUNN-MG, nas amostras do projeto Tara Ocean: <http://ocean-microbiome.embl.de/companion.html>.

Estação 4: A amostra da estação estação 4 DCM foi submetida às ferramentas MinPath e FUNN-MG, considerando a interação grupos ortólogos-vias metabólicas. O threshold de corte obtido foi igual a 0.205

Tabela 11: Anotação funcional para as vias do projeto Tara ocean, amostra 4 DCM, de 1-30. Em rosa as vias identificadas nas duas ferramentas, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, e em amarelo as vias identificadas somente pela ferramenta minPath.

	pathways	functions	index-funn
1	path:map00196	Photosynthesis - antenna proteins	**0.487**
2	path:map00230	Purine metabolism	**0.766**
3	path:map00240	Pyrimidine metabolism	**0.575**
4	path:map00290	Valine, leucine and isoleucine biosynthesis	**0.366**
5	path:map00330	Arginine and proline metabolism	**0.087**
6	path:map00362	Benzoate degradation via hydroxylation	**0.015**
7	path:map00561	Glycerolipid metabolism	**0.012**
8	path:map00627	1,4-Dichlorobenzene degradation	**0.024**
9	path:map00630	Glyoxylate and dicarboxylate metabolism	**0.037**
10	path:map00633	Trinitrotoluene degradation	**0.075**
11	path:map00650	Butanoate metabolism	**0.022**
12	path:map00660	C5-Branched dibasic acid metabolism	**0.477**
13	path:map00680	Methane metabolism	**0.005**
14	path:map00710	Carbon fixation in photosynthetic organisms	**0.071**
15	path:map00770	Pantothenate and CoA biosynthesis	**0.272**
16	path:map00791	Atrazine degradation	**0.101**
17	path:map00830	Retinol metabolism	**0.066**
18	path:map00900	Terpenoid biosynthesis	**0.298**
19	path:map00908	Zeatin biosynthesis	**0.133**
20	path:map00910	Nitrogen metabolism	**0.289**
21	path:map03010	Ribosome	**1**
22	path:map03020	RNA polymerase	**0**
23	path:map03022	Basal transcription factors	**0.072**
24	path:map03030	DNA replication	**0.205**
25	path:map03420	Nucleotide excision repair	**0.022**
26	path:map03430	Mismatch repair	**0.258**
27	path:map03440	Homologous recombination	**0.035**
28	path:map04150	mTOR signaling pathway	**0.069**
29	path:map04910	Insulin signaling pathway	**0.023**
30	path:map05322	Systemic lupus erythematosus	**0.015**

Estação 4: A amostra da estação estação 4 SUR foi submetida às ferramentas MinPath e FUNN-MG, considerando a interação grupos ortólogos-vias metabólicas. O threshold de corte obtido foi igual a 0.055

Tabela 12: Anotação funcional para as vias do projeto Tara ocean, amostra 4 SUR, de 1-26. Em rosa as vias identificadas nas duas ferramentas, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, e em amarelo as vias identificadas somente pela ferramenta minPath.

	pathways	functions	index-funn
1	path:map00196	Photosynthesis - antenna proteins	**0.591**
2	path:map00230	Purine metabolism	**0.821**
3	path:map00240	Pyrimidine metabolism	**0.095**
4	path:map00290	Valine, leucine and isoleucine biosynthesis	**0.394**
5	path:map00330	Arginine and proline metabolism	**0.005**
6	path:map00362	Benzoate degradation via hydroxylation	**0.036**
7	path:map00627	1,4-Dichlorobenzene degradation	**0.05**
8	path:map00630	Glyoxylate and dicarboxylate metabolism	**0.069**
9	path:map00633	Trinitrotoluene degradation	**0.457**
10	path:map00650	Butanoate metabolism	**0.001**
11	path:map00660	C5-Branched dibasic acid metabolism	**0.5**
12	path:map00680	Methane metabolism	**0.007**
13	path:map00710	Carbon fixation in photosynthetic organisms	**0.113**
14	path:map00770	Pantothenate and CoA biosynthesis	**0.031**
15	path:map00790	Folate biosynthesis	**0.014**
16	path:map00791	Atrazine degradation	**0.122**
17	path:map00830	Retinol metabolism	**0.062**
18	path:map00900	Terpenoid biosynthesis	**0.35**
19	path:map00908	Zeatin biosynthesis	**0.055**
20	path:map00910	Nitrogen metabolism	**0.159**
21	path:map03010	Ribosome	**1**
22	path:map03430	Mismatch repair	**0.055**
23	path:map04150	mTOR signaling pathway	**0.114**
24	path:map04910	Insulin signaling pathway	**0.051**
25	path:map05010	Alzheimer's disease	**0**
26	path:map05120	Epithelial cell signaling in Helicobacter pylori infection	**0.035**

ANEXO B – Resultados (2)

Comparação entre as ferramenta minPath e FUNN-MG, nas amostras AMD (NCBI, 2006), considerando a análise a partir dos grupos ortólogos. Em destaque o índice obtido para cada via de acordo com a ferramenta FUNN-MG. O *threshold* de corte nesta análise ficou em 0.036. Em amarelo as vias identificadas somente pela ferramenta minPath, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, em vermlho as vias identificadas pelas duas ferramentas.

Tabela 13: Anotação funcional para as vias AMD, de 1-27. Em rosa as vias identificadas nas duas ferramentas, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, e em amarelo as vias identificadas somente pela ferramenta minPath.

	pathways	functions	index-funn
1	path:map00010	Glycolysis / Gluconeogenesis	**0.8**
2	path:map00020	Citrate cycle (TCA cycle)	**0.773**
3	path:map00030	Pentose phosphate pathway	**0.252**
4	path:map00040	Pentose and glucuronate interconversions	**0.047**
5	path:map00051	Fructose and mannose metabolism	**0.032**
6	path:map00052	Galactose metabolism	**0.104**
7	path:map00053	Ascorbate and aldarate metabolism	**0.009**
8	path:map00061	Fatty acid biosynthesis	**0.108**
9	path:map00071	Fatty acid metabolism	**0.45**
10	path:map00072	Synthesis and degradation of ketone bodies	**0.097**
11	path:map00130	Ubiquinone biosynthesis	**0.145**
12	path:map00190	Oxidative phosphorylation	**0.306**
13	path:map00195	Photosynthesis	**0.015**
14	path:map00230	Purine metabolism	**0.612**
15	path:map00240	Pyrimidine metabolism	**0.567**
16	path:map00253	Tetracycline biosynthesis	**0.006**
17	path:map00260	Glycine, serine and threonine metabolism	**0.642**
18	path:map00280	Valine, leucine and isoleucine degradation	**0.467**
19	path:map00281	Geraniol degradation	**0.128**
20	path:map00290	Valine, leucine and isoleucine biosynthesis	**0.613**
21	path:map00300	Lysine biosynthesis	**0.283**
22	path:map00310	Lysine degradation	**0.118**
23	path:map00330	Arginine and proline metabolism	**0.518**
24	path:map00340	Histidine metabolism	**0.598**
25	path:map00350	Tyrosine metabolism	**0.237**
26	path:map00360	Phenylalanine metabolism	**0.173**
27	path:map00362	Benzoate degradation via hydroxylation	**0.074**

Tabela 14: Anotação funcional para as vias AMD, de 28-68. Em rosa as vias identificadas nas duas ferramentas, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, e em amarelo as vias identificadas somente pela ferramenta minPath.

	pathways	functions	index-funn
28	path:map00363	Bisphenol A degradation	**0.066**
29	path:map00380	Tryptophan metabolism	**0.191**
30	path:map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	**0.311**
31	path:map00401	Novobiocin biosynthesis	**0.021**
32	path:map00410	beta-Alanine metabolism	**0.243**
33	path:map00430	Taurine and hypotaurine metabolism	**0.061**
34	path:map00440	Aminophosphonate metabolism	**0.007**
35	path:map00450	Selenoamino acid metabolism	**0.202**
36	path:map00460	Cyanoamino acid metabolism	**0.081**
37	path:map00471	D-Glutamine and D-glutamate metabolism	**0.114**
38	path:map00480	Glutathione metabolism	**0.162**
39	path:map00500	Starch and sucrose metabolism	**0.351**
40	path:map00510	N-Glycan biosynthesis	**0.011**
41	path:map00520	Nucleotide sugars metabolism	**0.261**
42	path:map00521	Streptomycin biosynthesis	**0.249**
43	path:map00523	Polyketide sugar unit biosynthesis	**0.062**
44	path:map00540	Lipopolysaccharide biosynthesis	**0.068**
45	path:map00550	Peptidoglycan biosynthesis	**0.162**
46	path:map00561	Glycerolipid metabolism	**0.057**
47	path:map00562	Inositol phosphate metabolism	**0.008**
48	path:map00564	Glycerophospholipid metabolism	**0.031**
49	path:map00590	Arachidonic acid metabolism	**0.005**
50	path:map00591	Linoleic acid metabolism	**0.007**
51	path:map00620	Pyruvate metabolism	**1**
52	path:map00621	Biphenyl degradation	**0.056**
53	path:map00622	Toluene and xylene degradation	**0.04**
54	path:map00623	2,4-Dichlorobenzoate degradation	**0.073**
55	path:map00624	1- and 2-Methylnaphthalene degradation	**0.024**
56	path:map00625	Tetrachloroethene degradation	**0.291**
57	path:map00626	Naphthalene and anthracene degradation	**0.119**
58	path:map00627	1,4-Dichlorobenzene degradation	**0.038**
59	path:map00630	Glyoxylate and dicarboxylate metabolism	**0.349**
60	path:map00633	Trinitrotoluene degradation	**0.159**
61	path:map00640	Propanoate metabolism	**0.891**
62	path:map00643	Styrene degradation	**0.003**
63	path:map00650	Butanoate metabolism	**0.991**
64	path:map00660	C5-Branched dibasic acid metabolism	**0.141**
65	path:map00670	One carbon pool by folate	**0.224**
66	path:map00680	Methane metabolism	**0.248**
67	path:map00710	Carbon fixation in photosynthetic organisms	**0.443**
68	path:map00720	Reductive carboxylate cycle (CO2 fixation)	**0.865**

Tabela 15: Anotação funcional para as vias AMD, de 69-108. Em rosa as vias identificadas nas duas ferramentas, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, e em amarelo as vias identificadas somente pela ferramenta minPath.

	pathways	functions	index-funn
68	path:map00720	Reductive carboxylate cycle (CO2 fixation)	**0.865**
69	path:map00730	Thiamine metabolism	**0.227**
70	path:map00740	Riboflavin metabolism	**0.156**
71	path:map00750	Vitamin B6 metabolism	**0.173**
72	path:map00760	Nicotinate and nicotinamide metabolism	**0.101**
73	path:map00770	Pantothenate and CoA biosynthesis	**0.177**
74	path:map00785	Lipoic acid metabolism	**0.031**
75	path:map00790	Folate biosynthesis	**0.003**
76	path:map00830	Retinol metabolism	**0.005**
77	path:map00860	Porphyrin and chlorophyll metabolism	**0.779**
78	path:map00900	Terpenoid biosynthesis	**0.313**
79	path:map00903	Limonene and pinene degradation	**0.181**
80	path:map00910	Nitrogen metabolism	**0.466**
81	path:map00920	Sulfur metabolism	**0.12**
82	path:map00930	Caprolactam degradation	**0.138**
83	path:map00940	Phenylpropanoid biosynthesis	**0.013**
84	path:map00941	Flavonoid biosynthesis	**0.007**
85	path:map00950	Alkaloid biosynthesis I	**0.005**
86	path:map00960	Alkaloid biosynthesis II	**0.036**
87	path:map00970	Aminoacyl-tRNA biosynthesis	**0.996**
88	path:map00980	Metabolism of xenobiotics by cytochrome P450	**0.006**
89	path:map00982	Drug metabolism - cytochrome P450	**0.006**
90	path:map00983	Drug metabolism - other enzymes	**0.006**
91	path:map01040	Biosynthesis of unsaturated fatty acids	**0.035**
92	path:map01051	Biosynthesis of ansamycins	**0.045**
93	path:map01055	Biosynthesis of vancomycin group antibiotics	**0.036**
94	path:map02010	ABC transporters - General	**0.115**
95	path:map02020	Two-component system - General	**0.073**
96	path:map02030	Bacterial chemotaxis - General	**0.216**
97	path:map02040	Flagellar assembly	**0.593**
98	path:map03010	Ribosome	**0.971**
99	path:map03020	RNA polymerase	**0.097**
100	path:map03022	Basal transcription factors	**0.069**
101	path:map03030	DNA replication	**0.038**
102	path:map03050	Proteasome	**0.021**
103	path:map03060	Protein export	**0.094**
104	path:map03070	Type III secretion system	**0.367**
105	path:map03320	PPAR signaling pathway	**0.057**
106	path:map03410	Base excision repair	**0.207**
107	path:map03420	Nucleotide excision repair	**0.126**
108	path:map03430	Mismatch repair	**0.08**

Tabela 16: Anotação funcional para as vias AMD, de 109-117. Em rosa as vias identificadas nas duas ferramentas, em laranjado as vias identificadas somente pela ferramenta FUNN-MG, e em amarelo as vias identificadas somente pela ferramenta minPath.

	pathways	functions	index-funn
109	path:map03440	Homologous recombination	**0.016**
110	path:map04910	Insulin signaling pathway	**0.016**
111	path:map04920	Adipocytokine signaling pathway	**0.018**
112	path:map04930	Type II diabetes mellitus	**0.01**
113	path:map05010	Alzheimer's disease	**0.007**
114	path:map05012	Parkinson's disease	**0.026**
115	path:map05111	Vibrio cholerae pathogenic cycle	**0**
116	path:map05200	Pathways in cancer	**0.092**
117	path:map05210	Colorectal cancer	**0.006**