UNIVERSIDADE DE CAXIAS DO SUL

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

COORDENADORIA DE PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

NÍVEL DOUTORADO

Desenvolvimento de workflows científicos para a geração e análise de diferentes redes de interatomas

DANIEL LUIS NOTARI

CAXIAS DO SUL

2012

DANIEL LUIS NOTARI

Desenvolvimento de workflows científicos para a geração e análise de diferentes redes de interatomas

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando a obtenção do grau de Doutor em Biotecnologia.

Orientador: Prof. Dr. Diego Bonatto

Caxias do Sul

2012

DANIEL LUIS NOTARI

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
UCS - BICE - Processamento Técnico

N761d Notari, Daniel Luis
Desenvolvimento de workflows científicos para a geração e ánalise de diferentes redes de interatomas / Daniel Luis Notari. -- 2012.
xviii, 161 f. : il. ; 30 cm

Dissertação (Doutorado) – Universidade de Caxias do Sul, Programa de Pós-Graduação em Biotecnologia, 2012.
"Orientação: Prof. Dr. Diego Bonatto".

1. Bioinformática. 2. Biotecnologia. 3. Biologia. 4. Proteínas. I. Título

CDU 2.ed. : 57:004

Índice para o catálogo sistemático:

1. Bioinformática	57:004
2. Biotecnologia	57.08
3. Biologia	57
4. Proteínas	577.112

Catalogação na fonte elaborada pela bibliotecária
Márcia Servi Gonçalves – CRB 10/1500

Desenvolvimento de workflows científicos para a geração e análise de diferentes redes de interatomas

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando a obtenção do grau de Doutor em Biotecnologia.

Orientador: Prof. Dr. Diego Bonatto

TESE APROVADA EM 20 DE DEZEMBRO DE 2012

Comissão examinadora:

Prof. Dr. Hugo Verli

Prof. Dr. Sergio Echeverrigaray Laguna

Prof. Dra. Helena Grazziotin Ribeiro

Dedico este trabalho a todas as pessoas que acreditam na realização de seus sonhos.

AGRADECIMENTOS

Em primeiro lugar quero agradecer a toda a minha família, em especial a minha esposa Rafaela.

Aos meus colegas professores do Centro de Computação e Tecnologia da Informação, em especial a Helena, o Alexandre Khron, a Carine, Maria de Fátima, Cadinho, pelo apoio técnico e emocional dispendido ao longo dos últimos anos. Aos meus colegas professores do Instituto de Biotecnologia, em especial ao Sérgio, a Ana, a Scheila e ao Ghünther pela oportunidade de conhecer uma nova e fantástica área de conhecimento.

Aos alunos de iniciação científica e de trabalho de conclusão dos cursos de Ciência da Computação, Sistemas de Informação e Biologia que toparam o desafio de realizar o seu trabalho final em uma área totalmente diferente do seu conhecimento. Um muito obrigado aos alunos Samuel, Maurício, Renata, Jaqueline, Vanessa, Bruno, Joice, Virgínia, Cristian e Douglas.

Quero fazer um agradecimento muito especial ao meu orientador Prof. Dr. Diego Bonatto por toda a paciência e ensinamentos passados ao longo destes quatro anos. Por que não é fácil fazer um trabalho destes em uma área diferente da sua formação de graduação e mestrado. Muito obrigado por tudo.

E, por fim, que agradecer também a todos que me acompanharam e que de alguma forma contribuíram para a concretização deste trabalho. Aos professores e funcionários do Programa de Pós-Graduação em Biotecnologia (PPGBio), aos demais professores, colegas e amigos.

ESTRUTURA DA TESE

Esta tese de doutorado está caracterizada da seguinte maneira: uma introdução, os objetivos, seguida de uma revisão bibliográfica geral e de dois capítulos principais. Na parte final, encontra-se uma discussão geral, as conclusões obtidas, as perspectivas e três adendos. Os dois capítulos aparecem na forma de artigos científicos e seguem as normas utilizadas pelos periódicos para os quais foram e/ou serão submetidos.

Na introdução são levantados aspectos importantes sobre a bioinformática e biologia de sistemas para a análise de redes de interação de proteínas.

A seguir, aparecem os dois capítulos principais: O primeiro capítulo, intitulado **"DIS2PPI: A WORKFLOW DESIGNED TO INTEGRATE PROTEOMIC AND GENETIC DISEASE DATA"** foi aceito para publicação com modificações no periódico *International Journal of Knowledge Discovery in Bioinformatics*. Este artigo apresenta uma forma de gerar uma rede de interação entre doenças monogênicas e proteínas específicas a partir de múltiplos conjuntos de dados. Para tanto, um programa organizado na forma de um *workflow* científico que consulta diferentes bancos de dados biológicos, tais como o OMIM (*Online Mendelian Inheritance in Man*) e STRING (*Known and Predicted Protein-Protein Interactions*) foi gerado.

O segundo capítulo, intitulado **"NET2HOMOLOGY: A WORKFLOW DESIGNED TO ANALYZE CONSERVED BIOLOGICAL PROCESS BETWEEN TWO DIFFERENT SPECIES USING INTERACTOMIC NETWORKS"** será submetido a um periódico internacional de bioinformática (ainda não definido). Este artigo permite comparar duas redes de interação de proteínas geradas para dois organismos diferentes a fim de verificar processos biológicos conservados. Este *workflow* recebe uma rede de interação de proteínas de um organismo

como dados de entrada. Um alinhamento de sequências usando o programa Psi-Blast é feito com estas proteínas usando um outro organismo de interesse. O resultado do Psi-Blast apresenta proteínas distantemente relacionadas entre os organismos envolvidos. O *workflow* extrai as informações necessárias deste resultado para pesquisar redes de interação de proteínas para cada organismo no banco de dados STRING. Estas redes são comparadas usando os programas NetworkBlast ou PathBlast.

Seguindo a ordem estabelecida, a tese apresenta a discussão geral inter-relacionando os conhecimentos dos dois capítulos e o relato das conclusões e das perspectivas pretendidas. A seguir são apresentados três anexos. O primeiro anexo apresenta o trabalho intitulado "Toxicological effects of the main carcinogenic substances in tobacco smoke on human embryonic development", o qual participo como co-autor e que se encontra em fase de revisão no periódico PLoS ONE. Neste artigo é descrito o mecanismo de ação de várias pequenas moléculas presentes no tabaco no desenvolvimento embrionário utilizando, para tanto, ferramentas de biologia de sistemas. O segundo anexo apresenta o Portal IntergenicDB, que prove acesso ao banco de dados IntergenicDB. Este banco de dados foi criado para auxiliar no estudo de sequências promotoras de DNA de organismos procariontes. O portal fornece uma interface de consulta a genes, a organismos, a sequências de DNA e a métodos de predição. O terceiro anexo apresenta ferramentas de consultas a dados biológicos desenvolvidos para auxiliar no desenvolvimento das ferramentas DIS2PPI e NET2HOMOLOGY. Consultas podem ser feitas aos bancos de dados STRING e PDB, bem como a execução de uma consulta ao BLAST.

ÍNDICE

LISTA DE TABELAS

CAPÍTULO I

LISTA DE FIGURAS

REVISÃO BIBLIOGRÁFICA GERAL

CAPÍTULO I

Figure 3. (A) Descriptions of the major proteins gathered from xeroderma pigmentosum data by Dis2PPI choosing *Homo sapiens* as the model species. The major PPI network generated from STRING 9.0 prospection by Dis2PPI. The gathered PPI network was visualized in Cytoscape (B).

Figure 4. The Java source code used to access the ESearch tool web service to obtain a disease report (A) and the Java source code to access the String database to obtain a protein description (B).

Figure 5. Example of protein data extraction and processing using Dis2PPi software and expert advises. In (A) the OMIM's report for xeroderma pigmentosum was used to extract all potential protein names (B). After selecting the correct protein names (C), the list was subjected to STRING database in order to generate a PPI network (D).

Figure 6. In (A), xeroderma pigmentosum PPI network was acquired for XPA, XPC and XPD groups. Then, a merged PPI network obtained using Merge Cytoscape plugin (B).

Figure 7. Cluster and centrality analyses of XP network. The three major clusters and the bottlenecks nodes are shown in the network by different node shapes and by a gray color, respectively (inset box).

Figure 8. Cluster and centrality analyses of XP-CS network. The four major clusters and the bottlenecks nodes are shown in the network by different node shapes and by a gray color, respectively (inset box).

CAPÍTULO II

Figure 1: Fluxogram of Net2Homology scientific workflow**.**

Figure 2: In the first experiment, xeroderma pigmentosum *Homo sapiens* PPI network (A) used as input to execute Psi-Blast, and xeroderma pigmentosum *Mus musculus* PPI network (B) was obtained after extract information from Psi-Blast and query STRING

database. NetworkBlast had runned with (A) and (B) networks, and generates the PPI network resultant (C) that shows conserved protein complex across these two networks. The second experiment uses cicle-oxigenase-1 enzyme *Homo sapiens* PPI network (D) as input to execute Psi-Blast, and cicle-oxigenase-1 enzyme *Mus musculus* PPI network (E) was obtained after extract information from Psi-Blast and query STRING database. NetworkBlast had runned with (D) and (E) networks, and generates the PPI network resultant (F) that shows conserved protein complex across these two networks.

LISTA DE ABREVIATURAS

API	*Application Program Interface*
BiNGO	*Biological Networks Gene Ontology*
BioSystems	*Biological Systems*
BLAST	*Basic Local Alignment Search Tool*
CentiScape	*Centralities for Cytoscape*
COG	*Clusters of Orthologous Groups*
CMR	*Comprehensive Microbial Resourse*
DDBJ	DNA *Data Bank of Japan*
DER	Diagrama Entidade-Relacionamento
e-PCR	*Electronic PCR*
EMBL-EBI	*European Bioinformatics Institute*
ENA	*European Nucleotide Archive*
E-Utilities	*Entrez Programming Utilities*
GEO	*Gene Expression Omnibus*
GO	*Gene Ontology (Ontologia Gênica)*
HMM	*Hidden Markov Model*
HUPO	*The Human Proteome Organization*
INSDC	*International Nucleotide Sequence Database Collaboration*
IP	*Internet Protocol*
LINES	*Long Interspersed Sequences*
MCODE	*Molecular Complex Detection*
NCBI	*National Center for Biotechnology Information*
OMIM	*Online Mendelian Inheritance in Man*

PCR	*Polymerase Chain Reaction* (Reação da cadeia de polimerase)
PDB	*Protein Data Bank*
PPI	*Protein-Protein Interaction* (Interação proteína-proteína)
PRIDE	*PRoteomics IDEntifications*
ProtClustDB	*Entrez Protein Clusters DataBase*
ProtMap	*Protein Map*
Psi-Blast	*Position-Specific Iterated BLAST*
Redes BA	Redes Barabási-Albert
SGBDR	Sistema Gerenciador de Banco de Dados Relacional
SINES	*Short Interspersed Sequences*
STRING	*Search Tool for the Retrieval of Interacting Genes/Proteins*
STS	*Sequence Tagged Sites* (Sequências em Sítios Marcados)
SVM	*Support Vector Machine*
UniProtKB	UniProt Knowledgebase
VOG	*Viral COG*
W3C	*World Wide Web Consortium*
WGS	*Whole Genome Shotgun*

RESUMO

Workflows científicos são projetados para realizar experimentos *in silico* com o intuito de processar e analisar uma grande quantidade de dados usando simulação computacional. Os experimentos são organizados como uma sequência de etapas que caracterizam um fluxo de execução, onde em cada etapa utilizam-se diferentes *softwares*. Estes *softwares* não possuem um modelo de representação de dados comum. Por isto, quando um *workflow* científico é construído com o objetivo de integrar e processar dados, cada etapa precisa analisar a estrutura dos dados, processá-los e preparar os mesmos de acordo com a estrutura necessária para a execução da próxima etapa do *workflow*. Desta maneira, os *workflows* científicos usam diferentes algoritmos provenientes das áreas da matemática e da ciência da computação, os quais são capazes de processar, armazenar e transformar os dados utilizados em informação útil para diferentes análises de pesquisadores de biologia. Adicionalmente, a biologia de sistemas é uma subárea da bioinformática que ajuda a compreender as interações observadas entre populações de moléculas, células e organismos resultantes do aumento da complexidade biológica. Assim, nesta tese de doutorado, buscou-se desenvolver *workflows* científicos utilizando-se ferramentas de bioinformática e biologia de sistemas para comparar processos biológicos através da geração de redes de interação proteína-proteína. Para tanto, o programa Dis2PPI foi gerado por meio de um *workflow* científico que possibilitou integrar os bancos de dados de doenças monogênicas OMIM (*Online Mendelian Inheritance in Man*) com o programa de metabuscas STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*). O objetivo desta integração foi gerar uma rede de interação proteína-proteína associada com doenças de natureza monogênica. Como prova de conceito, o programa Dis2PPI foi utilizado para a obtenção de redes de interação para as síndromes xeroderma pigmentosa e de

Cockayne, duas doenças monogênicas raras e cujos fenótipos estão associados com acúmulos de danos de DNA, câncer, progeria e neurodegeneração. Uma vez geradas, estas redes foram avaliadas com o uso ferramentas de biologia de sistemas para análise de ontologia gênica e de centralidade no programa Cytoscape. Da mesma forma, o programa Net2Homology foi gerado através de um desenho de um *workflow* científico que possibilitasse integrar o resultado de um alinhamento de sequências de proteínas, usando o programa NCBI PSI-BLAST, com o programa de metabuscas STRING. Neste sentido, o programa Net2Homology possibilitou recuperar duas redes de interação proteína-proteína associada a dois organismos diferentes. Este mesmo programa foi utilizado para a obtenção de redes de interação para a doença xeroderma pigmentosa e para a enzima ciclo-oxigenase-1 cruzando as informações dos organismos *Homo sapiens* e *Mus musculus*. As redes de interação obtidas foram comparadas visando obter os processos biológicos evolutivamente conservados.

Palavras-chaves: redes de interação proteína-proteína, *workflow* científico, biologia de sistemas.

ABSTRACT

Scientific workflows can be used to design *in silico* experiments to process and analyse a great amount of data using computational simulation. In this sense, all *in silico* experiments are organized as a sequence of steps that characterize an execution flow, where each steps employ different softwares. It is important to note that these softwares do not employ a common structure or model to represent data. Thus, taking into account the diversity of softwares and data representation, a scientific workflow should be built to integrate and process data, where each step must be able to analyse the data structure, to process these data and finally generate a new data structure that provide the requirements needed to execute the next step of the workflow. It is important to note that scientific workflows use mathematics and computer sciences methods capable to process, storage, and transform these data into useful information. In addition, a subarea of Bioinformatics termed Systems biology helps to understand the interactions observed between populations of molecules, cells and organisms as the result of the increase of biological complexity. Thus, the purpose of this work was to develop scientific workflows using bioinformatics and system biology softwares with the purpose to compare different biological processes from the generation of protein-protein interaction (PPI) networks. In this sense, the Dis2PPI software was design as a scientific workflow to integrate OMIM (*Online Mendelian Inheritance in Man*) database with metasearch program STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) in order to retrieve protein-protein interaction networks associated to monogenic diseases. The Dis2PPI was used to generate PPI networks for xeroderma pigmentosum and cockayne syndromes, two rare monogenic diseases associated with DNA damage, cancer, progeria, and neurodegeneration. The PPI networks generated were analyzed with Cytoscape program and specific plugins that

evaluate gene ontologies and centrality analysis. By its turn, Net2Homology program was designed as a scientific workflow to integrate NCBI PSI-BLAST sequence alignments with STRING metasearch program to recovery PPI networks associate with two different species. In this sense, Net2Homology was used to obtain PPI network for xeroderma pigmentosum disease and cyclooxigenase-1 enzyme by crossing *Homo sapiens* and *Mus musculus* information. A comparison between these networks was realized to verify the major evolutively conserved biological process.

Key words: protein-protein interaction network, scientific workflow and systems biology.

1. INTRODUÇÃO

A bioinformática fornece um conjunto variado de ferramentas para armazenar, processar, analisar e avaliar uma grande quantidade de dados através de experimentos *in silico* executados por meio de uma simulação computacional. Além disto, programas computacionais são devsenvolvidos para fazer inferência de arquivos de dados, gerar conexões entre estes e derivar predições úteis e interessantes para as ciências biológicas, uma vez que a maioria dos bancos de dados possui uma estrutura própria, dificultando a integração entre eles. Neste sentido, a Biologia de Sistemas auxilia a compreender como funcionam as interações entre populações de moléculas, células e organismos devido ao aumento da complexidade dos processos biológicos.

Assim, por definição, um banco de dados é um conjunto de informações integradas e que tem por objetivo atender a uma comunidade de usuários. O conceito de banco de dados é usado na área de computação para representar um conjunto de dados relacionados. Por sua vez, na área das ciências biológicas, um banco de dados compreende os dados de ácidos nucleicos e sequências de proteínas, estruturas moleculares e funções, bem como dados de interações entre estas moléculas em diferentes contextos biológicos, tais como vistas em inúmeras patologias. Desta maneira, um banco de dados biológico permite o armazenamento de diferentes tipos de dados onde o mais importante é fornecer um mecanismo de consulta que seja eficiente para a coleta de informações específicas a partir destes bancos.

A grande variedade de ferramentas de bioinformática disponíveis para pesquisar e analisar bancos de dados (DNA, RNA e proteínas) é uma tarefa relativamente simples para um pesquisador quando usadas isoladamente. Quando pesquisadores desejam usar mais de uma ferramenta para realizar experimentos *in silico*, surgem problemas tais como os diferentes formatos de entrada e de saída de dados empregados pelos bancos de

dados, o que dificulta o uso destas informações em outras ferramentas de bioinformática e, por fim, a dificuldade de integração dos dados entre as diferentes ferramentas de bioinformática.

Assim, uma forma de construir experimentos que envolvam o uso de diversas ferramentas e formatos heterogêneos de dados é através de um *workflow*. Um *workflow* visa automatizar procedimentos, onde documentos, dados, informações ou tarefas são distribuídos entres os participantes, de acordo com um conjunto pré-definido de regras, para se alcançar o objetivo proposto.

Os *workflows* científicos são usados para descrever experimentos *in silico* construídos através de uma sequência de passos que caracterizam um fluxo de execução. Neste caso, cada um destes passos pode usar diferentes ferramentas e bancos de dados de bioinformática, onde sempre há a manutenção de um registro dos passos executados, o que permite repetir o *workflow* com pequenas variações nos parâmetros empregados.

Assim, esta tese mostra a geração de *workflows* científicos para estudo de doenças monogênicas, os quais possuem um forte impacto em pesquisas básicas e aplicadas nas áreas das Ciências Biológicas. Para o desenvolvimento destes *workflows* utilizou-se ferramentas de bioinformática e de biologia de sistemas para comparar processos biológicos através da geração de redes de interação proteína-proteína. Foram obtidas redes para as síndromes xeroderma pigmentosa e de Cockayne, duas doenças monogênicas raras e cujos fenótipos estão associados com acúmulos de danos de DNA, câncer, progeria e neurodegeneração.

2. OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver *workflows* científicos para a prospecção de dados proteômicos com o propósito de gerar redes de interação entre proteínas para estudos comparativos de processos biológicos.

2.1.1 OBJETIVOS ESPECÍFICOS

- Desenhar um *workflow* utilizando os programas NCBI Eutils WSDL, Entrez Utilities Web Service, Apache Axis for Java e URL Internet Access que possibilitem integrar os bancos de dados OMIM com o programa de metabuscas STRING a fim de recuperar uma rede de interação proteína-proteína associada com doenças de natureza monogênica;
- Utilizar o *workflow* gerado para a construção do programa Dis2PPI na linguagem de programação Java, que permita a recuperação e o uso de redes interatômicas obtidas em associação com o programa Cytoscape;
- Utilizar o programa Dis2PPI para obtenção de redes de interação para as doenças Xeroderma Pigmentosa e síndrome de Cockayne.
- Analisar as redes de interação obtidas para as doenças Xeroderma Pigmentosa e síndrome de Cockayne usando ferramentas de biologia de sistemas presentes no programa Cytoscape. Neste caso, especificamente, fazer análises de ontologia gênica e de centralidade.
- Desenhar um *workflow* utilizando os programas Entrez Utilities Web Service, PDB Web Service, Apache Axis for Java e URL Internet Access que possibilitem integrar um alinhamento de sequências de proteínas usando o

programa NCBI PSI-BLAST com o programa de metabuscas STRING a fim de recuperar duas redes de interação proteína-proteína associada a dois organismos diferentes;

- Utilizar o *workflow* gerado para a construção do programa Net2Homology na linguagem de programação Java que permita o uso das redes obtidas em associação com os programas NetworkBlast e Cytoscape;
- Utilizar o programa Net2Homology para obtenção de redes de interação para a doença Xeroderma pigmentosa e para a enzima Ciclo-oxigenase-1 cruzando as informações dos organismos *Homo sapiens* e *Mus musculus*.
- Analisar as redes de interação obtidas para as doenças Xeroderma pigmentosa e para a enzima Ciclo-oxigenase-1 cruzando as informações dos organismos *Homo sapiens* e *Mus musculus* visando obter processos biológicos conservados através do uso do programa NetworkBlast.

3. REVISÃO BIBLIOGRÁFICA GERAL

Este capítulo apresentará os portais de bioinformática, os bancos de dados e as ferramentas de bioinformática, bem como descreverá a biologia de sistemas e suas ferramentas, as características e topologias de redes biológicas e, por fim, introduzirá a possibilidade de integração das ferramentas de bioinformática e biologia de sistemas através do uso de *workflows* científicos.

3.1 PORTAIS DE BIOINFORMÁTICA

O INSDC[1] *(International Nucleotide Sequence Database Collaboration*) é uma associação entre o NCBI[2] (*National Center for Biotechnology Information*), o ENA[3] (*European Nucleotide Archive*) e o DDBJ[4] (*DNA Data Bank of Japan*). Os três institutos realizam uma colaboração internacional para manter um conjunto de dados biológicos onde estes dados estão em constante sincronia (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008). Além destes portais, o portal do ExPasy será apresentado nas próximas seções.

3.1.1 NCBI

O NCBI tem por objetivo fornecer e desenvolver novas tecnologias para o armazenamento e análise de dados de biologia molecular, bioquímica e genética, facilitando o uso dessas bases dados e programas para os pesquisadores (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008). Alguns dos serviços disponibilizados pelo NCBI são o PubMed[5] e o Entrez[6]. O PubMed é uma das ferramentas de busca

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

[2] http://www.ncbi.nlm.nih.gov/

[3] http://www.ebi.ac.uk/ena/

[4] http://www.ddbj.nig.ac.jp/

[5] http://www.ncbi.nlm.nih.gov/sites/pubmed

[6] http://www.ncbi.nlm.nih.gov/sites/gquery

oferecidas pelo NCBI que fornece acesso a citações na literatura biomédica contidos no banco de dados MEDLINE, revistas científicas e livros *on-line* (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008). O Entrez é um serviço que fornece mecanismos de busca entre os diversos bancos de dados disponibilizados pelo NCBI. Ele integra a literatura científica, bancos de dados de DNA e proteínas, estruturas tridimensionais (3-D) de proteínas, genomas completos e informações taxonômicas dos organismos conhecidos no seu sistema de busca (Barnes e Gray, 2003; Jones e Pevzner, 2004; Wheeler et al., 2006; Lesk, 2008). O Entrez fornece um *site* para consultas na Internet assim como um suporte a chamada de *web services* ou a resolução de solicitações de consulta através da montagem de uma URL (*Uniform Resource Locator*) (Barnes e Gray, 2003; Jones e Pevzner, 2004; Wheeler et al., 2006; Lesk, 2008). As ferramentas do Entrez podem ser usadas para a construção de programas específicos para pesquisa e análise de dados biológicos.

3.1.2 ENA

O ENA é um portal que provê informações sobre sequências de nucleotídeos, desenvolvido e mantido pelo EMBL-EBI (*European Bioinformatics Institute*) (Barnes e Gray, 2003; Lesk, 2008). Os serviços do ENA são constantemente aprimorados para refletir o crescente volume de dados, sempre melhorando a tecnologia de anotação genômica e ampliando as aplicações onde são armazenadas (Barnes e Gray, 2003; Lesk, 2008). O ENA possui atualmente cerca de 255,1 milhões de sequências[7] armazenadas.

3.1.3 DDBJ

O DDBJ é um portal japonês que oferece ferramentas de pesquisa e análise para sequências de DNA e proteínas, bem como fornece acesso a diversas ferramentas de bioinformática. Além disto, o DDBJ fornece o único banco de dados de sequências de

[7] Dados retirados do site http://www.ebi.ac.uk/ena/about/statistics em 17/10/2012.

nucleotídeos da Ásia (Barnes e Gray, 2003). Os dados submetidos ao DDBJ são processados e disponibilizados utilizando o seu formato proprietário de distribuição através do uso de arquivos textos (*flat files*) (Barnes e Gray, 2003). O arquivo texto inclui a sequência de nucleotídeos, as informações do provedor, referências, os organismos-fonte e as características definidas pela tabela de características do DDBJ/EMBL/GenBank para descrever a natureza biológica das sequências de nucleotídeos (Barnes e Gray, 2003).

Os arquivos texto são usados com muita frequência para submeter ou adquirir informações de bancos de dados biológicos. O formato de arquivo texto chamado de FASTA é um tipo de arquivo comum utilizado para armazenar sequências de ácidos nucleicos, aminoácidos e RNA (Pearson e Lipman, 1988). A Figura 1 mostra algumas informações das proteínas ATM, BRCA1 e CDKN1A do organismo *Homo sapiens* no formato de arquivo FASTA. As linhas iniciadas com o caractere '>' indicam a descrição de um gene, proteína ou RNA armazenado e as demais linhas contêm a sequência de dados que descrevem um gene, proteína ou RNA (Lesk, 2008).

```
>ATM   (Homo sapiens)   (STRING-Protein-Id:9606.ENSP00000278616)
MSLVLNDLLICCRQLEHDRATERKKEVEKFKRLIRDPETIKHLDRHSDSKQG(....)
>BRCA1         (Homo sapiens)   (STRING-Protein-Id:9606.ENSP00000350283)
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLL(....)
>CDKN1A        (Homo sapiens)   (STRING-Protein-Id:9606.ENSP00000244741)
MSEPAGDVRQNPCGSKACRRLFGPVDSEQLSRDCDALMAGCIQEARERW(....)
```

Figura 1: Exemplo de um arquivo FASTA para as proteínas ATM, BRCA1 e CDKN1A de *Homo sapiens*.

3.1.4 EXPASY

O ExPASy[8] é um portal de recursos de bioinformática que fornece acesso a bancos de dados científicos e ferramentas de software para diferentes áreas das ciências

[8] http://www.expasy.org/

humanas (Artimo et al., 2012). As ferramentas incluem proteomas, genomas, filogenia, biologia de sistemas, transcriptomas, entre outros (Artimo et al., 2012).

3.2 BANCOS DE DADOS

Os portais de bioinformática organizam os dados de diferentes organismos na forma de bancos de dados de genomas, de transcriptomas, de proteomas, de sequências intergênicas, entre outras classificações existentes. Os bancos de dados biológicos não são exclusivos de portais, uma vez que existe uma grande variedade de trabalhos publicados sobre a criação, armazenamento e consulta de bancos de dados para diferentes dados biológicos (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008). O Anexo 2 apresenta um exemplo de um banco de dados criado para sequências intergênicas. Nesta seção serão conceituados os tipos de banco de dados de genomas, de transcriptomas, de proteômas e de sequências intergênicas. Apenas alguns bancos de dados existentes serão explicados uma vez que esses foram utilizados no desenvolvimento deste trabalho.

3.2.1 GENOMA

O genoma é o conjunto das informações genéticas de um organismo que está codificada no seu DNA (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008). Considerando a espécie humana, o projeto genoma humano[9] teve a participação de empresas e pesquisadores do mundo inteiro. O propósito deste sequênciamento foi mapear todos os genes de nossa espécie (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008).

As informações obtidas com o sequênciamento estão sendo organizadas em diversos bancos de dados biológicos. Atualmente, existem muitos bancos de dados

[9] http://www.genome.gov/10001772

biológicos disponibilizados *on-line*. Por exemplo, o GenBank[10] é um banco de dados de sequências de DNA e uma coleção de anotações de todas as sequências públicas de DNA de diferentes organismos. O GenBank tem sido expandido para incluir dados de expressão de sequências, dados de sequências de proteínas, estrutura tridimensional de proteína, taxonomia, informações sobre diferentes tipos de interações biológicas e literatura biomédica (Jones e Pevzner, 2004; Lesk, 2008). O GenBank possui atualmente mais de 380 mil organismos sequênciados (Benson et al., 2011). Além disto, há, aproximadamente, mais de 13 milhões de registros de sequências no banco de dados tradicional[11] e, quase 62 milhões de registros de sequências incompletas submetidas no WGS[12] (*Whole Genome Shotgun*).

3.2.2 TRANSCRIPTOMA

Por analogia com o termo genoma, criou-se o termo transcriptoma que representa o conjunto completo dos transcritos presentes em uma célula (Lesk, 2008). Estes bancos de dados têm como objetivo coletar e classificar dados referentes aos diferentes tipos de RNA encontrados em vários organismos, além de identificar e caracterizar potenciais peptídeos e/ou proteínas que estão codificadas em moléculas de RNA mensageiro (Barnes e Gray, 2003; Lesk, 2008). Outro objetivo é reconhecer quais são os genes expressos em diferentes tecidos e fases de desenvolvimento, gerando uma visão mais abrangente da forma como os organismos regulam a expressão de diferentes genes para uma condição ambiental e/ou fisiológica específica (Lehninger et al., 2005).

Um exemplo de banco de dados de transcriptoma é o GEO[13] (*Gene Expression Omnibus*) que é um repositório de dados público para armazenar informações sobre

---

[10] http://www.ncbi.nlm.nih.gov/genbank

[11] Dados extraídos do site http://www.ncbi.nlm.nih.gov/genbank em 24/09/2012.

[12] http://www.ncbi.nlm.nih.gov/genbank/wgs

[13] http://www.ncbi.nlm.nih.gov/geo/

expressões de gene e arranjos de hibridização (Edgar et al., 2002). O GEO possui cerca de mais 38 mil séries armazenadas e quase um milhão de modelos[14].

3.2.3 PROTEOMA

Um proteoma[15] tem o objetivo de identificar, localizar e analisar funções das proteínas envolvidas com as células. As análises de uma proteína envolvem a separação, identificação e quantificação de várias amostras de proteínas presentes em uma amostra (Lesk, 2008). Outras informações armazenadas visam registrar as relações entre as proteínas para uma determinada função celular (Barnes e Gray, 2003). As informações de um proteoma podem armazenar padrões de expressões de proteínas obtidos com o uso de técnicas de biologia molecular e, com o intuito de complementar as informações de projetos de genomas e transcriptomas (Lesk, 2008).

O proteoma humano é organizado pelo HUPO[16] (*The Human Proteome Organization*), uma organização científica internacional[17] com cooperação e colaboração de várias organizações de pesquisa com o objetivo de desenvolver novas tecnologias, técnicas e treinamento. Um proteoma pode possuir informações sobre análise de estrutura, de modificações pós-traducionais, predições e interações de proteínas conhecidas, entre outras (Barnes e Gray, 2003; Jones e Pevzner, 2004; Lesk, 2008). Os bancos de dados PDB, ProtClustDB, PRIDE e Swiss-Prot são alguns dos bancos de dados de proteínas existentes.

O PDB[18] (*Protein Data Bank*) é um banco de dados que armazena estruturas tridimensionais de proteínas, ácidos nucleicos, vírus e carboidratos (Berman, 2000). No

---

[14] Dados retirados do site http://www.ncbi.nlm.nih.gov/geo/summary/ em 26/09/2012.

[15] Definição extraída do site http://www.hupo.org/overview/background/proteomics.asp em 19/12/2012.

[16] http://www.hupo.org/research/hpp/

[17] Definição extraída do site http://www.hupo.org/overview/background/ em 19/12/2012.

[18] http://www.rcsb.org/pdb

PDB há cerca de 84 mil estruturas depositadas[19] sobre quase 55 mil organismos[20] diferentes.

O ProtClustDB[21] (*Entrez Protein Clusters DataBase*) é um banco de dados que fornece mecanismos de consulta para informações de anotação, domínio e estruturas de proteínas, além de publicações sobre estes temas e uma coleção de sites para ferramentas disponíveis na web (Klimke et al., 2009). Este banco de dados possui uma coleção de agrupamentos de sequências de proteínas com informações verificadas e outras não (Klimke et al., 2009). Possui mais de 4,6 milhões de proteínas armazenadas, além de informações a respeito de mais de 627 mil agrupamentos de proteínas[22].

O banco de dados PRIDE[23] (*PRoteomics IDEntifications*) é um repositório para dados proteômicos que incluem identificadores de proteínas, peptídeos e modificações pós-traducionais, sendo possível realizar consultas por espécie, tipo de célula, termos de ontologia gênica e doenças (Vizcaino et al., 2010). Atualmente este banco de dados possui informações a respeito de mais de 26 mil experimentos, mais de onze milhões de identificadores de proteínas e mais 63 milhões de identificadores de peptídeos[24].

O Swiss-Prot[25] é um banco de dados de anotações de sequências de proteínas manualmente anotadas e uma seção revisada do UniProtKB (UniProt[26] *Knowledgebase*) do portal ExPasy (Bairoch et al., 2004). As suas anotações descrevem funções de uma proteína, modificações pós-traducionais, domínios e *sites*, estrutura secundária de uma

---

[19] Dados retirados do site http://www.rcsb.org/pdb/home/home.do em 24/09/2012.

[20] Dados retirados do item "By Source Organism (Gene Source)" em 24/09/2012 do site http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html

[21] http://www.ncbi.nlm.nih.gov/proteinclusters

[22] Estatísticas retiradas do site http://www.ncbi.nlm.nih.gov/genomes/prkstats.html em 19/12/2012.

[23] http://www.ebi.ac.uk/pride/

[24] Estatísticas extraídas do site http://www.ebi.ac.uk/pride/ em 19/10/2012.

[25] http://web.expasy.org/docs/swiss-prot_guideline.html

[26] http://www.uniprot.org/

proteína, entre outras informações (Bairoch et al., 2004). O Swiss-Prot possui mais de 538 mil sequências armazenadas, compreendendo mais de 191 milhões de aminoácidos sumarizados em mais de 213 mil referências[27].

3.2.4 SEQUÊNCIAS INTERGÊNICAS

As regiões promotoras, encontradas em uma sequência de DNA, representam um mecanismo de regulação da expressão gênica. Estas regiões localizam-se à montante em uma determinada região codificadora e interagem com a enzima RNA polimerase (RNAp), desencadeando o processo de expressão gênica (Zaha, 1996, Alberts et al., 1999; Lehninger et al., 2005). Deste modo, os bancos de dados de promotores auxiliam nos estudos sobre a regulação gênica, sendo importante organizar dados sobre promotores pertencentes a organismos que ainda não possuem repositórios públicos (Avila e Silva, 2011). Além disto, as sequências de nucleotídeos de um organismo podem ser usadas para definir um modelo de sequências para iniciar a transcrição em outro organismo (Zaha, 1996; Alberts et al., 1999; Lehninger et al., 2005).

O RegulonDB[28] é um banco de dados que possui informações sobre as regiões promotoras da bactéria *Escherichia coli*. Outro tipo de banco de dados é o que armazena informações sobre sequências repetidas não-codificantes presentes no DNA que formam grandes frações dos genomas de eucariotos, sendo que a sequência repetida pode ser curta (*Short Interspersed Sequences* - SINES) ou longa (*Long Interspersed Sequences* – LINES) (Lesk, 2008). Este tipo de banco de dados pode ser acessado pelas ferramentas do NCBI.

[27] Estatísticas extraídas do site http://web.expasy.org/docs/relnotes/relstat.html em 19/10/2012.

[28] http://regulondb.ccg.unam.mx/

3.2.5 OUTROS TIPOS DE BANCO DE DADOS

O OMIM[29] (*Online Mendelian Inheritance in Man*) é um banco de dados que contêm um catálogo de doenças genéticas mantido pelo NCBI (Barnes e Gray, 2003; Lesk, 2008; Amberger et al., 2009). As pesquisas podem ser realizadas no *site* por palavras-chave ou através do uso das ferramentas de consulta do Entrez (Barnes e Gray, 2003; Lesk, 2008; Amberger et al., 2009).

O BioSystems[30] (*Biological Systems*) é um banco de dados que fornece um acesso integrado a sistemas biológicos cruzando informações de diversos bancos de dados e ferramentas usando os recursos do NCBI (Geer et al., 2010). É possível consultar informações dos sistemas biológicos e seus componentes (genes, proteínas e moléculas), bem como, acessar informações sobre doenças (genes, biomarcadores e drogas) e publicações (Geer et al., 2010).

O PubChem[31] é um banco de dados com dados experimentais que provê acesso a informações sobre atividades biológicas de pequenas moléculas (Bolton et al., 2008). Este banco de dados está dividido em três diferentes seções com informações para substâncias, compostos e ensaios de avaliação da atividade biológica (Bolton et al., 2008; Wang et al., 2012). Uma das ferramentas disponibilizadas é a pesquisa por estruturas químicas baseada em similaridade (Bolton et al., 2008; Wang et al., 2012). O PubChem possui atualmente mais de 100 milhões de substâncias[32], contêm mais de 500 mil descrições de protocolos de ensaios biológicos abrangendo mais de cinco mil proteínas e 30 mil genes, e cobre mais de 130 milhões de registros de atividades biológicas armazenadas (Wang et al., 2012).

---

[29] http://www.ncbi.nlm.nih.gov/omim

[30] http://www.ncbi.nlm.nih.gov/biosystems/

[31] http://pubchem.ncbi.nlm.nih.gov/

[32] Estatística extraídas do site http://pubchem.ncbi.nlm.nih.gov/ em 20/10/2012.

O banco de dados SWISS-MODEL[33] armazena informações sobre anotações tridimensionais de modelos de comparação de estrutura de proteínas gerados a partir do modelo SWISS-MODEL de homologia (Kiefer, 2009). Este banco de dados contêm informações sobre modelos teoricamente calculados, sendo que os mesmos podem conter falsos positivos (Kiefer, 2009). Atualmente, este banco de dados possui[34] mais de 3,2 milhões de entradas de modelos para mais de 2,3 milhões de sequências únicas armazenadas no banco de dados UniProt.

3.3 FERRAMENTAS DE BIOINFORMÁTICA

Nesta seção serão descritas as ferramentas de bioinformática dos portais NCBI e Expasy. Estes portais fornecem diversas ferramentas para manipulação de genomas, proteomas e transcriptomas, bem como ferramentas para acesso aos seus dados, para alinhamento de sequências e para análise de interações de proteínas. Uma pequena parte destas ferramentas será apresentada nesta seção, sendo que algumas delas foram utilizadas no desenvolvimento deste trabalho.

3.3.1 GENOMAS

O PCR[35] (*Polymerase Chain Reaction)* é uma técnica de amplificação do DNA na qual se obtêm várias cópias de uma fita de DNA (Lehninger et al., 2005). O PCR é uma técnica muito utilizada em laboratórios de biologia molecular. No entanto, é possível simular um PCR usando a ferramenta e-PCR (*Electronic* PCR) que permite identificar sequências em sítios marcados (STS) dentro de sequências de DNA (Rotmistrovsky et al., 2004).

---

[33] http://swissmodel.expasy.org/repository/

[34] Estatísticas retiradas do site http://swissmodel.expasy.org/repository/ em 20/10/2012.

[35] http://www.ncbi.nlm.nih.gov/sutils/e-pcr/

Uma forma de analisar um genoma é verificar os aminoácidos gerados após a transcrição e tradução de uma molécula codificante de DNA. Neste sentido, existem diversas ferramentas de bioinformática que permitem fazer a transcrição e a tradução *in silico*. Um exemplo é a ferramenta ProtMap[36] (*Protein Map*) que utiliza genomas completos de proteínas agrupadas através de um método baseado em grupos de similaridade de sequências (COG - *Clusters of Orthologous Groups*) ou de vírus (VOG - *Viral* COG) (Klimke et al., 2009). O ProtMap mostra o mapeamento de cada proteína baseado no COG/VOG com o seu genoma e apresenta todos os segmentos vizinhos do genoma para membros de um grupo específico de proteínas relacionadas (Klimke et al., 2009).

3.3.2 ACESSO AOS BANCOS DE DADOS DO NCBI

A ferramenta E-Utilities[37] (*Entrez Programming Utilities*) é um conjunto de programas que fornece acesso a consultas baseadas no Entrez e acesso a diversos bancos de dados do NCBI (Sayers e Wheeler, 2004). O E-Utilities caracteriza-se por ser uma ferramenta que oferece programas de acesso do lado de um servidor *web* (Sayers e Wheeler, 2004), ou seja, é necessário fazer um acesso remoto usando um programa cliente para acessar as informações desejadas. Normalmente, este acesso é feito usando uma sintaxe fixa de URL (como um endereço *web* qualquer que usa-se para fazer uma consulta na internet) com um conjunto de parâmetros pré-estabelecidos (Sayers e Wheeler, 2004). As ferramentas do E-Utilities são utilizadas neste trabalho para fazer consultas a algumas bases de dados (OMIM e GenBank) e serviços do NCBI (alinhamento de sequências e pesquisa de sequência de aminoácidos).

[36] http://www.ncbi.nlm.nih.gov/sutils/protmap.cgi

[37] http://www.ncbi.nlm.nih.gov/books/NBK25501/

3.3.3 ALINHAMENTO DE SEQUÊNCIAS

O BLAST[38] (*Basic Local Alignment Search Tool*) usa métodos de similaridade para comparação e alinhamento de sequências de nucleotídeos ou de proteínas (Altschul, 1990; Altschul, 1997). O resultado é baseado em cálculos estatísticos após a obtenção de casamentos (*matches*) significativos entre as sequências (Altschul, 1990; Altschul, 1997). O algoritmo processa os dados utilizando a heurística comparativa por pares de segmentos, realizando testes exaustivos com os dados encontrados através do acesso aos diferentes bancos de dados públicos e tendo como finalidade encontrar o menor número de *gaps* (falhas) nos fragmentos resultantes (Altschul, 1990; Altschul, 1997). O BLAST pode ser usado para inferir relacionamentos funcionais e evolucionários entre sequências, bem como identificar membros de uma família de genes (Altschul, 1990; Altschul, 1997).

O Psi-Blast (*Position-Specific Iterated* BLAST) é um dos programas disponíveis para encontrar proteínas distantemente relacionadas ou encontrar novos membros de famílias de proteínas (Gish e States, 1993). As pesquisas podem ser realizadas no *site* através do *upload* de dados de genes ou proteínas através do uso das ferramentas de consulta do Entrez (Barnes e Gray, 2003; Lesk, 2008; Amberger et al., 2009).

O alinhamento pode ser aplicado para encontrar um gene em um genoma, identificar éxons, identificar domínios proteicos ou identificar possíveis homólogos em um banco de dados (Barnes e Gray, 2003; Lesk, 2008). A busca de similaridade pode ser aplicada na predição de genes, predição de função, predição de estrutura e na inferência de árvores filogenéticas (Barnes e Gray, 2003; Lesk, 2008). Através do alinhamento de sequências podem-se identificar algumas características, tais como: a) duas sequências similares e, com a mesma função, podem não ter o mesmo ancestral; b)

[38] http://blast.ncbi.nlm.nih.gov/

duas sequências são homólogas se elas possuem uma sequência ancestral comum; c) duas sequências são parálogas quando se obtém homologia por duplicação; d) duas sequências são ortólogas quando se identifica a homologia por especiação; e, e) se duas (ou mais) sequências são parecidas, onde estas podem ser homólogas e/ou podem ter funções similares e/ou podem ter a mesma estrutura (Barnes e Gray, 2003; Lesk, 2008).

O *e-value* representa o número esperado de alinhamentos locais distintos em um casamento (*match*) realizado para a comparação de duas sequências aleatórias de quaisquer tamanhos (Altschul, 1990; Altschul, 1997). Em outras palavras, o valor do *e-value* descreve o nível de ruído (*hit*) do alinhamento podendo variar entre zero e o número de sequências procuradas no banco de dados (Barnes e Gray, 2003; Lesk, 2008).

Outras ferramentas para realizar alinhamento de sequências são o COBALT e o CLUSTAL. A ferramenta COBALT[39] permite realizar múltiplos alinhamentos de sequências de proteínas (Papadopoulos e Agarwala, 2007). Esta ferramenta procura uma coleção de pares derivados de pesquisas em bancos de dados, similaridade de sequências e entradas do usuário onde estes dados são combinados e incorporados a um alinhamento de sequências progressivo (Papadopoulos e Agarwala, 2007). A ferramenta Clustal[40] é um programa para alinhamento de múltiplas sequências desenvolvido para diferentes sistemas operacionais (Larkin, 2007).

3.3.4 PREDIÇÃO DE INTERAÇÃO DE PROTEÍNAS

O STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) é uma ferramenta para realizar metabuscas com o intuito de se obter redes de interação de proteínas (Jensen et al., 2009). O STRING possui um banco de dados sobre predições e

[39] http://www.ncbi.nlm.nih.gov/tools/cobalt

[40] http://www.clustal.org/clustal2/

interações de proteínas conhecidas, envolvendo associações físicas e funcionais (Jensen et al., 2009). Atualmente, o STRING possui informações sobre mais de 1,1 mil organismos e mais de cinco milhões de proteínas. Por outro lado, a ferramenta SBRT[41] (*Systems Biology Research Tool*), disponível no portal Expasy, permite executar diversos métodos para análise de redes usando teoria dos grafos, geometria e análise combinatória (Wright and Wagner, 2008).

3.4 BIOLOGIA DE SISTEMAS

Com o sucesso do Projeto Genoma Humano e o estabelecimento das técnicas de análise de larga escala de sistemas biológicos, tornou-se necessário entender como ocorrem às interações entre os diferentes tipos de moléculas de origem biológica, assim como para possibilitar visualizar, de forma mais ampla, a função exercida por genes, proteínas e outras moléculas em um determinado contexto biológico (Barabási e Oltvai, 2004). Um grande desafio da biologia é compreender como populações de organismos são estruturadas a partir das suas interações com o meio ambiente e como sistemas complexos tem evoluído (Karsenti, 2012). Sistemas complexos e organismos vivos, em particular, surgem a partir da ocorrência de processos dinâmicos simultâneos em diferentes escalas de tempo (Karsenti, 2012).

Uma das ferramentas usadas para trabalhar nestas pesquisas é a Biologia de Sistemas, uma área interdisciplinar com o objetivo de compreender como funcionam as interações entre populações de moléculas, células e organismos devido ao aumento da complexidade de processos biológicos (Karsenti, 2012). O estudo das interações entre os componentes de um sistema biológico envolvem a teoria dos grafos, matemática discreta e processamento computacional (Barabási e Oltvai, 2004; O'Malley e Dupré, 2005; Siegal et al., 2007). Assim, o uso destas ferramentas e técnicas aliadas à relação

[41] http://www.ieu.uzh.ch/wagner/software/SBRT/index.html

entre a topologia de uma rede biológica e suas propriedades funcionais ou evolucionárias foram utilizadas para descrever as interações de sistemas celulares através das redes de livre-escala (Rosvall e Sneppen, 2003; Barabási e Oltvai, 2004; O'Malley e Dupré, 2005; Siegal et al., 2007).

Esta seção apresenta inicialmente a teoria dos grafos, na sequência as redes biológicas, redes de livre-escala e finaliza abordando a topologia local e global de uma rede.

3.4.1 TEORIA DOS GRAFOS

Um grafo é uma forma de representar relacionamentos existentes entre pares de objetos (Goodrich e Tamassia, 2007). Um grafo é definido geometricamente como um conjunto de pontos no espaço que são interconectados por um conjunto de linhas (Gibbons, 1994). Formalmente, um grafo G (V, E) consiste de um conjunto de vértices[42] V = {v1, v2,...} e um conjunto de arestas[43] E = {e1, e2,...}, onde uma aresta representa um conjunto de pares não dirigidos de elementos de V (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). Os vértices representam os objetos, as arestas representam relações entre os objetos e cada aresta define uma relação simétrica (grafos não dirigidos; Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011).

As ligações entre os vértices podem ser dirigidos, e neste caso chamam-se arcos, ou não dirigidos, e neste caso chamam-se arestas (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011).

Um grafo pode ser representado visualmente usando pequenos círculos para vértices e retas ou curvas para arestas. A Figura 2 apresenta um exemplo de um grafo G

[42] Um vértice também pode ser chamado ponto, nodo ou nó. Para a explicação da teoria dos grafos adoraremos o termo vértice e a partir da explicação do seu uso na biologia de sistemas, utilizaremos o termo nó.

[43]Uma aresta também é chamada de elo, ligação ou conector. Para a explicação da teoria dos grafos adoraremos o termo aresta e a partir da explicação do seu uso na biologia de sistemas utilizaremos o termo conector ou ligação.

= (V, E) e o cálculo das medidas do grau de um vértice, do caminho entre dois vértices, do comprimento entre dois vértices e a menor distância entre dois vértices.

O grau (ou conectividade) de um vértice é calculado pela equação $\text{grau}(v) = \sum \text{arestas}(v)$ onde deve somar o número de vezes que *v* é usado como uma ligação para as arestas, ou seja, esta medida indica o número de ligações de um vértice tem com os outros vértices (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). A Figura 2 mostra os valores calculados dos graus do nós 1, 2, 3, 4, 5 e 6 na quarta linha à direita.

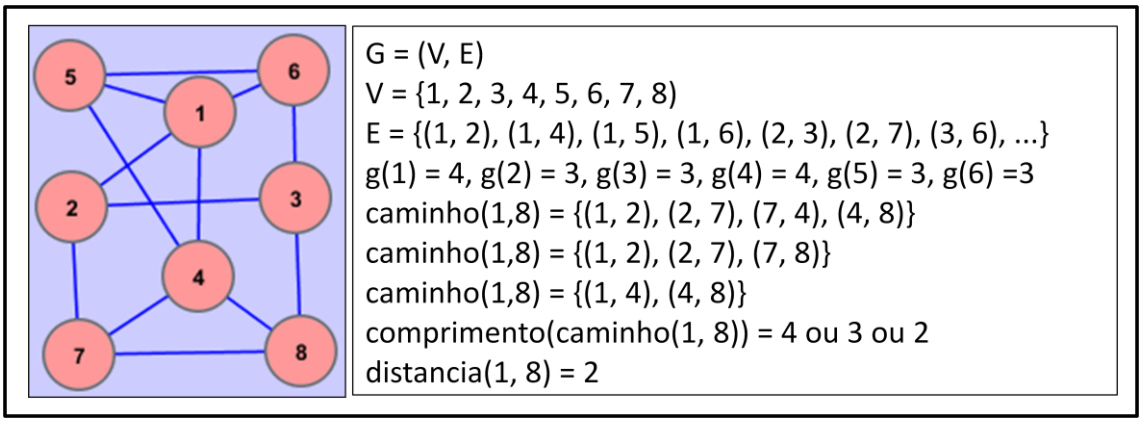


Figura 2: Exemplo de um grafo G = (V, E), onde V representa os vértices e E representa as arestas entre os vértices.

O grau de distribuição de um vértice é calculado pela equação $\text{grau}_{\text{dist(v)}} = \frac{\text{grau(v)}}{\sum \text{vertices}}$ onde se calcula o grau do vértice *v* e divide-se número total de vértices (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). Para o grafo da Figura 2, por exemplo, o vértice v(1) possui grau(1) = 4 e tem um grau de distribuição de 0,5 por causa da divisão deste valor por 8, que é o número total de vértices.

Um caminho de um grafo é uma sequência de vértices com ligações existentes partindo-se de um vértice A e chegando a um vértice B, sendo que a equação $\text{caminho}(v1, vn) = \{(v1, v2), (v2, v3), (v3, v4), \ldots\}$ permite determinar o caminho

como uma sequência de vértices v_1, v_2,.., v_k, tal que v_i, $v_i + 1 \in E$, sendo $1 \leq i < |k - 1|$, o caminho de v_1 a v_k onde k–1 é o comprimento do caminho (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). A Figura 2 mostra três possíveis caminhos para o mesmo par de vértices (1, 8) nas linhas cinco a sete à direita.

Quando se tem um caminho em que o vértice inicial e o final são os mesmos, denomina-se de ciclo (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). O caminho mais significativo representa a média sobre todos os caminhos mais curtos possíveis entre todos os pares de vértices e o cálculo é feito pela equação $caminho_{mais_{significativo}} = media(\sum_{i=1}^{n} < distancia(v1, v2))$ (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011).

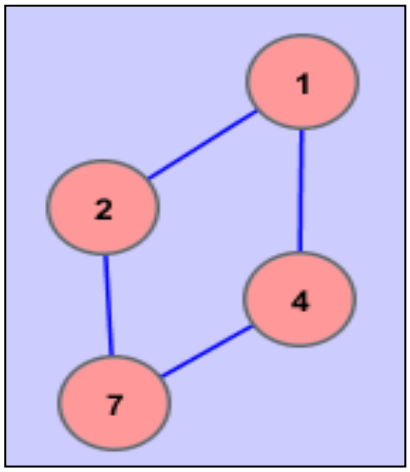
O comprimento de um caminho é calculado pela equação $\text{comprimento}(\text{vi}, \text{vn}) = \sum_{\text{i}=1}^{\text{n}} \text{arestas}(\text{caminho}(\text{vi}, \text{vn}))$ onde deve-se somar o número de ligações entre as arestas do vértice inicial até o vértice final (Gibbons, 1994). A Figura 2 mostra o valor calculado do comprimento de um caminho na penúltima linha à direita, sendo que três valores são mostrados, um para cada caminho exemplificado. O menor destes valores representa o caminho mais curto entre dois vértices, isto porque que pode existir mais de um caminho entre dois vértices em um mesmo grafo, conforme mostra a Figura 2 (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011).

A distância entre dois vértices é calculada pela equação $\text{distancia}(\text{v}, \text{w}) = menor(\text{comprimento}(\text{caminho}(\text{v}, \text{w})))$ onde se devem calcular todos os caminhos entre os vértices *v* e *w*, sendo que o comprimento do menor caminho entre eles é o resultado da equação (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). A Figura 2 mostra o valor calculado para a distância na última linha à direita.

O diâmetro de um grafo é representado pela equação $diametro = maior(\forall distancia(v, w))$ onde se deve primeiro encontrar o caminho

mínimo entre cada par de vértices de um grafo (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). A maior distância encontrada entre qualquer par de vértices é o diâmetro do grafo (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). O caminho representado pelas arestas E = {(1, 5), (5, 6), (6, 3), (3, 8), (8, 4), (4, 7), (7, 2), (2, 1)} representa a maior distância para o grafo da Figura 2. Com isto, o diâmetro do grafo é 8.

Um grafo é conexo (ou conectado) se, para quaisquer dois vértices, existir um caminho entre eles (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). O subgrafo mostrado na Figura 3 é um exemplo de um grafo conexo.



G’ = (V’, E’)
V’ = {1, 2, 4, 7)
E’ = {(1, 2), (1, 4), (2, 7), (4, 7)}

Figura 3: Um subgrafo G’ = (V’, E’) do grafo G = (V, E) com os vértices V’ e as arestas E’.

Um subgrafo é um grafo obtido a partir da remoção de vértices e arestas do grafo original (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). Formalmente, um subgrafo é um subconjunto do grafo original onde um subgrafo G’ = (V’, E’) de um grafo G = (V, E) é um grafo tal que V’ $\subseteq$ V e E’ $\subseteq$ E (Gibbons, 1994; Goodrich e Tamassia, 2007; Shaffer, 2011). A Figura 3 mostra um subgrafo do grafo da Figura 2 com os seus vértices e arestas.

Denomina-se clique de um grafo G = (V, E) um subgrafo de G que seja completo (Gibbons, 1994). Um grafo é completo quando cada par distinto de vértices define uma aresta de tal forma que um vértice possui conexão com todos os outros vértices (Gibbons, 1994). O subgrafo mostrado na Figura 3 é um exemplo de um clique.

3.4.2 REDES BIOLÓGICAS

Uma rede complexa descreve uma grande variedade de sistemas na natureza e na sociedade fortemente conectados (Barabási e Albert, 2002), tais como redes sociais, redes celulares, redes de computadores, redes de interação de proteínas e a Internet, entre outros exemplos. O comportamento de sistemas complexos, desde a célula até a Internet, caracteriza-se por atividades orquestradas onde muitos componentes interagem um com o outro através de interações de um par de componentes (Barabási e Albert, 2002; Barabási e Oltvai, 2004). Estes componentes representam um conjunto de nós que são conectados com outros componentes por uma ligação, onde cada ligação representa interações entre os dois componentes, sendo que, desta forma, o conjunto de todos os nós e todas as ligações entre os nós formam uma rede (Barabási e Albert, 2002; Barabási e Oltvai, 2004).

Desta forma, uma rede pode representar um sistema biológico completo, onde os nós e suas ligações podem ser diferentes entidades biológicas, tais como moléculas, células, entre outras (Weston e Hood, 2004; Siegal et al., 2007) No contexto da análise de dados proteômicos, os nós são representados pelas proteínas e os conectores pelas reações químicas (Barabási e Albert, 2002; Rosvall e Sneppen, 2003; Barabási e Oltvai, 2004; O'Malley e Dupré, 2005; Siegal et al., 2007).

Diversas pesquisas mostram que redes reais, independente de sua idade, função ou escopo, convergem para uma arquitetura homogênea, uma característica que permite a pesquisadores de diferentes áreas usarem a teoria de redes como um paradigma

comum (Barabási, 2009). Diversas pesquisas feitas nos últimos anos mostram que as redes de livre-escala ajudaram a proliferar o estudo sobre as redes em diversos campos da ciência, um novo campo de pesquisa com os seus desafios e habilidades (Barabási, 2009). As redes de livre-escala são o tópico da próxima seção.

3.4.3 REDES DE LIVRE-ESCALA

Paul Erdös e Alfréd Rényi (1960) inicialmente estabeleceram os conceitos matemáticos da topologia de redes de livre-escala que têm sua distribuição de conectividade regida pela lei de potência definida pela equação $P(k) = k^{y}$, onde k representa o número de nós e *y* representa as ligações (Barabási e Albert, 1999; Rosvall e Sneppen, 2003; Barabási e Oltvai, 2004; Barabási, 2009).

A observação inicial do grau de distribuição de nós, em redes baseados na lei de potência, foi feita por (Barabási e Albert, 1999). Os autores discutiram que as redes reais de livre-escala são baseadas em dois mecanismos genéricos de muitas redes reais (Barabási e Albert, 1999; Barabási e Albert, 2002): i) uma rede começa com um número fixo *n* de nós que, então, são aleatoriamente conectados sem modificar *n*. Muitas redes reais descrevem sistemas abertos que crescem com a inclusão contínua de novos nós. Assim, uma rede é criada inicialmente com um núcleo pequeno de nós, ocorrendo um aumento gradual de nós durante o ciclo de vida da rede; e, ii) os modelos de rede assumem a probabilidade de que dois nós são conectados independe do grau dos nós, ou seja, novas ligações são feitas aleatoriamente.

A análise das redes de livre-escala envolve o estudo do fenômeno de interação molecular através da integração multinível de dados e modelos (Barabási e Oltvai, 2004; O'Malley e Dupré, 2005; Siegal et al., 2007). A rede livre de escala mais conhecida é a rede Barabási-Albert (BA) que possui dois mecanismos básicos (Barabási e Albert, 1999; Barabási e Albert, 2002; Barabási e Oltvai, 2004): i) o crescimento da

rede significa que a rede emerge através da ligação de nós subsequentes; e, ii) novos nós ligam-se, preferencialmente, aos nós mais conectados, ou seja, os nós com maior grau de distribuição. Um modelo baseado nestes dois mecanismos indica que o desenvolvimento de grandes redes é governado por um fenômeno de auto-organização que está além de características específicas de sistemas individuais (Barabási e Albert, 1999). A Figura 4 apresenta um exemplo da rede BA onde os nós cinco e quinze são os mais conectados da rede e, por outro lado, os nós terminais[44] são os menos conectados. Os nós altamente conectados tem uma maior probabilidade de ter um efeito na funcionalidade da rede do que nós esparsos, como por exemplo, proteínas altamente conectadas tendem a ser mais letais quando forem eliminadas (*knocked out*) (Siegal et al., 2007).

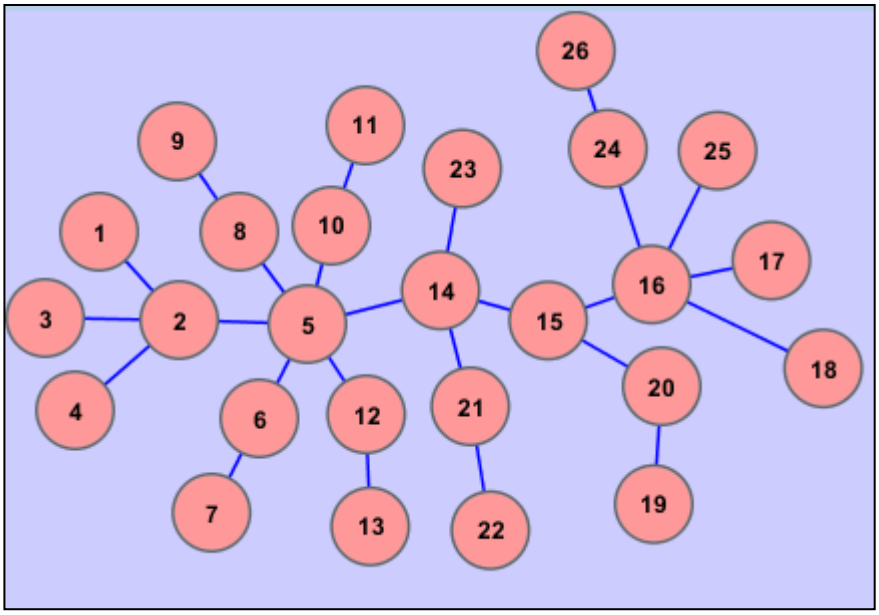


Figura 4: Exemplo de uma rede do tipo Barabási-Albert.

O grau de distribuição de um nó pode determinar diferentes classes de redes, ou seja, este valor pode nos ajudar a determinar a topologia de uma rede biológica (Barabási e Albert, 2002; Barabási e Oltvai, 2004). O valor desta medida representa que

[44] Os nós terminais são os nós nas extremidades do grafo. Neste grafo, os nós 9, 4, 22 e 18 são alguns dos nós terminais.

um nó é denominado de *hub*, que se caracteriza por ter um grande número de ligações e por manter os nós unidos (Barabási e Albert, 2002; Barabási e Oltvai, 2004). Esses nós representam pontos de controle da rede (Rosvall e Sneppen, 2003; Barabási e Oltvai, 2004). Desta maneira, analisando a equação da lei de potência, quanto menor for o valor de *y*, mais importante é o papel dos *hubs* na rede, por outro lado, se o valor de *y* for menor que três, os *hubs* são irrelevantes; se o valor de *y* for maior que dois e menor que três, há uma hierarquia de *hubs*, onde o *hub* mais conectado está em contato com uma pequena fração de todos os nós; e, por fim, se *y* for igual a dois, o *hub* está em contato com a maior parte de todos os nós (Barabási e Albert, 2002; Barabási e Oltvai, 2004). Em geral, uma rede de livre-escala deve possuir um valor de *y* menor que três e, tem como principal característica nós com significativas diferenças no grau de conectividade entre os nós (Barabási e Albert, 2002; Barabási e Oltvai, 2004).

As redes de livre-escala predizem que os nós que aparecem antes na história são os nós mais conectados da rede (Barabási e Oltvai, 2004). Assim, as proteínas evolutivamente mais antigas têm mais conexões quando comparadas com proteínas recentes (Barabási e Oltvai, 2004). Essa descoberta empírica demonstra uma preferência por formação de novas conexões com proteínas evolutivamente antigas (Barabási e Oltvai, 2004), enfatizando que estes nós, com um número maior de conexões, podem desempenhar funções bioquímicas importantes na célula (Siegal et al., 2007.). Em grafos dirigidos, as interações entre dois nós têm uma direção bem definida, como por exemplo, a direção do fluxo de material de um substrato para um produto em uma reação metabólica (Barabási e Oltvai, 2004). Em grafos não-dirigidos, as ligações não têm uma direção assinalada, por exemplo, em uma rede de interação de proteínas, uma ligação representa uma relação mútua de amarração bilateral, se a proteína A amarra-se

a proteína B, então a proteína B também se amarra a proteína A (Barabási e Oltvai, 2004).

As redes de livre-escala representam as redes de interação entre proteínas (PPI), visto que as redes PPI também seguem a lei de potência (Barabási e Oltvai, 2004; Siegal et al., 2007), sendo que a análise de uma rede pode levar em consideração as informações a respeito da sua topologia, que pode ser classificada em local ou global (Aittokallio e Schwikowski, 2006).

3.4.4 TOPOLOGIA LOCAL DE UMA REDE

A análise de topologia local visa descobrir padrões individuais de interação que carreguem alguma informação significativa sobre o seu papel no mecanismo celular através da interpretação de subgrafos obtidos através de *hubs*[45], motivos ou cliques (Aittokallio e Schwikowski, 2006).

Os motivos representam um grupo de moléculas (nós) ligadas com o intuito de trabalharem juntas para realizar uma função biológica, ou seja, representam processos biológicos comuns (Barabási e Oltvai, 2004). Para identificar os motivos que caracterizam uma rede é necessário determinar todos os subgrafos de uma rede. Em seguida, uma rede aleatória é gerada mantendo-se inalterados o número de nós, ligações e grau de distribuição. Subgrafos que aparecem significativamente mais frequentes na rede real comparando-se a rede aleatória são designados como candidatos a serem classificados como motivos (Barabási e Oltvai, 2004). Os motivos normalmente são associados com alguma função biológica otimizada (Barabási et al., 2011).

Em termos de centralidades, esta é uma medida quantitativa da posição relativa de um nó em relação aos outros nós, e pode ser usada para estimar a importância relativa ou o papel global de um nó na rede (Aittokallio e Schwikowski, 2006). As

[45] O conceito de *hubs* foi apresentado na seção 1.4.3.

medidas de centralidade estão baseadas na conectividade do nó (grau de centralidade), o menor caminho de um nó em relação a outros nós (medida chamada de *closeness*) ou o número de caminhos mínimos que atravessam o nó (medida chamada de *betweenness;* Aittokallio e Schwikowski, 2006). *Closeness* é definida pela soma das distâncias de um nó para todos os outros nós pela equação $closeness(v) = \sum_{i=0}^{n} \text{comprimento(v, vi)}$ (Everett and Borgatti, 2012). Por outro lado, a medida de *betweenness* é uma medida relacionada à centralidade dos nós na rede e é calculado a partir do número de caminhos mínimos que passam por um nó (Yu et al., 2007; Rahn et al., 2010). Esta medida é calculada pela equação da distância d(v, w) apresentada na seção 3.4.1.

Nós que possuem muitos caminhos mínimos que passam por eles possuem maior *betweenness*, indicando uma maior importância para a estrutura da rede (Yu et al., 2007). As proteínas que possuem um alto grau de nó determinam a existência de um *hub* enquanto que as proteínas que possuem um valor alto de *betweenness* controlam o fluxo de informação em uma rede e, assim, quanto maior for este valor, maior é a chance de uma proteína conectar diferentes agrupamentos ou processos biológicos (Rahn et al., 2010).

É possível cruzar diferentes tipos de medidas de centralidades para obter maiores informações sobre quais nós seriam importantes para a rede, como por exemplo, cruzar os dados de grau de nós e *betweenness* para obtenção dos chamados nós gargalos, também conhecidos como *bottlenecks* (Yu et al., 2007).

Os gargalos são conectores chave entre proteínas com propriedades funcionais e dinâmicas interessantes e, correspondem a componentes dinâmicos de interação na rede (Yu et al., 2007). Em redes dirigidas, o valor da medida de *betweenness* (nodo associado ao gargalo) é muito mais significativo do que o grau de um nó (tipicamente um *hub*, neste caso; Yu et al., 2007).

A Figura 5 apresenta um exemplo de uma rede biológica onde o nó 15 representa um *hub* na rede porque possui o maior grau de todos os nós da rede e, o círculo preto indica um ponto de gargalo na rede através da análise de centralidade.

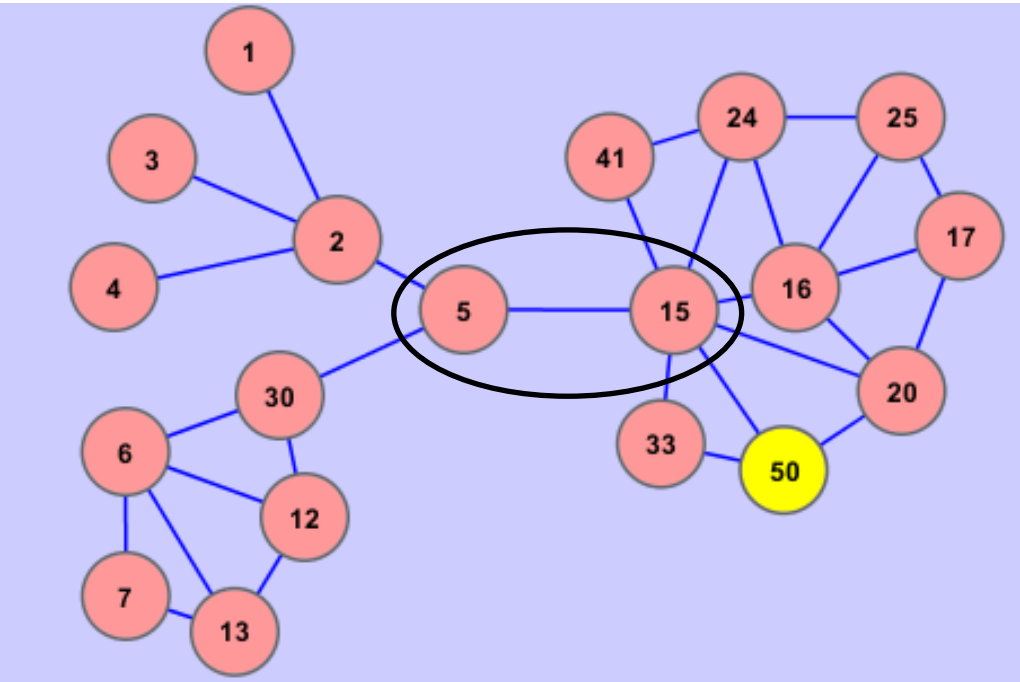


Figura 5: Exemplo de uma rede biológica com um ponto de gargalo na rede obtida através da análise de centralidade.

3.4.5 TOPOLOGIA GLOBAL DE UMA REDE

A topologia global tenta caracterizar o comportamento de uma célula como um todo, ou seja, tenta averiguar a organização hierárquica de uma célula procurando por motivos ou agrupamentos em uma rede através de grupos interconectados envolvidos em uma função celular comum (Aittokallio e Schwikowski, 2006). As medidas analisadas na topologia global envolvem o grau de distribuição[46] e coeficiente de agrupamento em redes celulares e proteomas (Aittokallio e Schwikowski, 2006).

O grau de distribuição é usado para descobrir padrões de interação baseado na análise de centralidade (Aittokallio e Schwikowski, 2006). O grau de distribuição

[46] O cálculo do grau de distribuição foi apresentado na seção 1.4.2.

também fornece uma medida de conectividade da rede, onde uma rede com pouca conectividade de nós possui poucos nós com um grau alto de conectividade na rede. Por outro lado, redes com uma média alta de conexões entre os nós possuem poucos nós com um grau baixo de conectividade, sendo esta uma característica das redes aleatórias, também conhecidas como redes randômicas (Siegal et al., 2007). A consequência da presença desta propriedade em uma rede é que poucos *hubs* altamente conectados mantêm o controle de toda a rede juntos (Barabási et al., 2011).

O coeficiente de agrupamento é calculado pela equação $coeficiente_{agrupamento(v)} = \frac{grau(v)}{\sum_{i=1}^{n} arestas(v)}$, onde o número de ligações dos vizinhos de um nó (calculado pelo grau do nó) dividido pelo número máximo de ligações destes nós (número de arestas de um nó; Aittokallio e Schwikowski, 2006). O coeficiente de agrupamento visa descobrir padrões de *pathways* através da decomposição da rede em grupos, sendo que para analisar um padrão de *pathways* é necessário lembrar que um caminho é um conjunto de nós únicos (sem repetição) conectados através de um grafo dirigido sem ramificações ou círculos (Aittokallio e Schwikowski, 2006). Desta forma, um *pathway* em uma rede celular representada por um grafo representa a transformação de um caminho de um nutriente em um produto final na rede metabólica ou uma série de modificações pós-traducional do sentido de um sinal para o destino pretendido de um sinal de tradução da rede (Albert, R., 2005).

3.5 FERRAMENTAS DE BIOLOGIA DE SISTEMAS

Existem várias ferramentas para a análise de biologia de sistemas, tais como Medusa, Pajek e Cytoscape. A ferramenta Medusa[47] usa um método de integração de dados baseado em modelos de aprendizado de motivos e, usa fatores de transcrição

[47] http://cbio.mskcc.org/leslielab/software/medusa

através da junção de sequências promotores e expressão gênica (Kundaje et al., 2004). A ferramenta Pajek[48] é uma ferramenta para visualização e análise de grandes redes, especialmente redes sociais (Batagelj and Mrvar, 2002). O software Cytoscape[49] permite visualizar as informações referentes a uma rede de interação das proteínas e fazer análises de biologia de sistemas usando os *plugins* desenvolvidos para aplicações em conjunto com este software (Shannon et al., 2003).

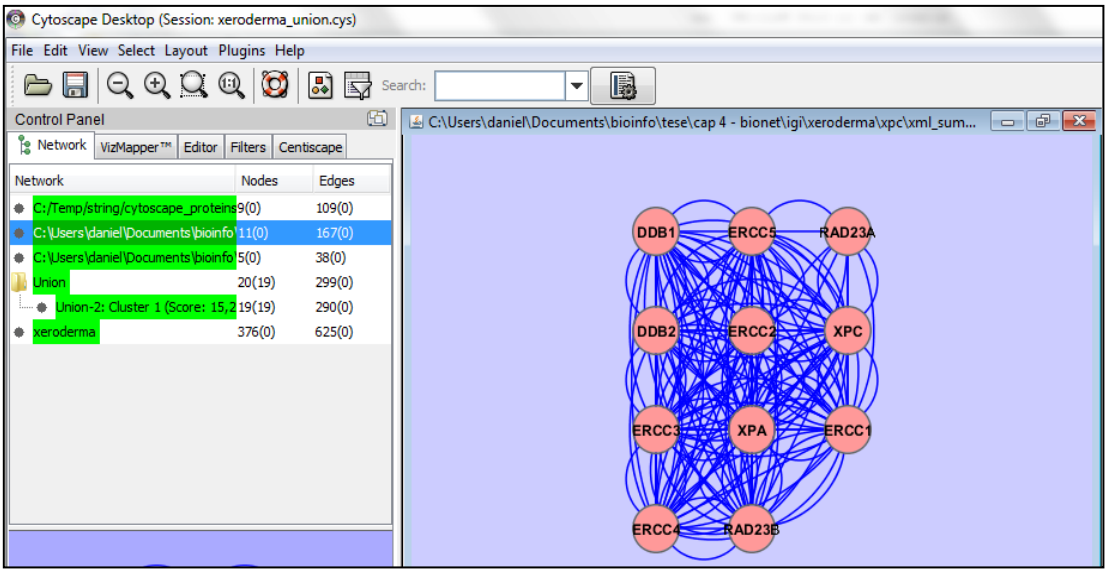


Figura 6: Rede PPI da doença xeroderma pigmentosa importada para o programa Cytoscape.

O Cytoscape foi utilizado neste trabalho como ferramenta de análise de biologia de sistemas porque permite buscar diretamente uma rede de interação de proteínas e possui um grande conjunto de *plugins* para analisar agrupamentos, ontologia gênica e centralidade em uma rede. Outro fator importante é que o Cytoscape pode ser acessado por outros programas desenvolvidos na forma de *workflows*. Esta abordagem foi utilizada neste trabalho, onde uma rede resultante de um algoritmo pode ser visualizada diretamente no Cytoscape, sem a necessidade de uma intervenção manual. A Figura 6 apresenta uma rede de interação de proteínas para a doença xeroderma pigmentosa,

[48] http://pajek.imfm.si/doku.php

[49] http://www.cytoscape.org

importada para o Cytoscape a partir de uma pesquisa realizada no banco de dados STRING.

Alguns *plugins* disponíveis no *site* do Cytoscape permitem realizar análise de agrupamentos em redes de interação de proteínas, bem como, inferir as funções biológicas dos seus componentes e análise de centralidade. O MCODE[50] (*Molecular Complex Detection*) é o *plugin* utilizado para detectar regiões densamente conectadas em grandes redes de interação de proteínas que representam complexos moleculares, ou seja, esta análise pode levar a detecção de agrupamentos (Bader e Hogue, 2003).

A Figura 7 apresenta o resultado do uso do plugin MCODE para análise da rede de interação de proteínas para a doença xeroderma pigmentosa. À esquerda são apresentados os agrupamentos da rede e, à direita são apresentados os agrupamentos juntos na rede. Os números de 1 a 4 indicam os pontos de conexão da rede denominados de *hubs*.

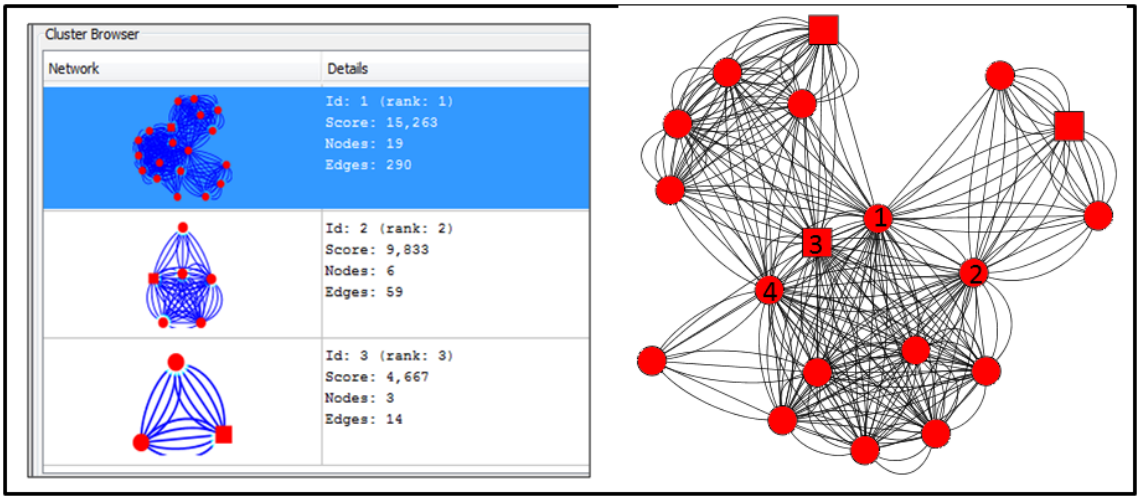
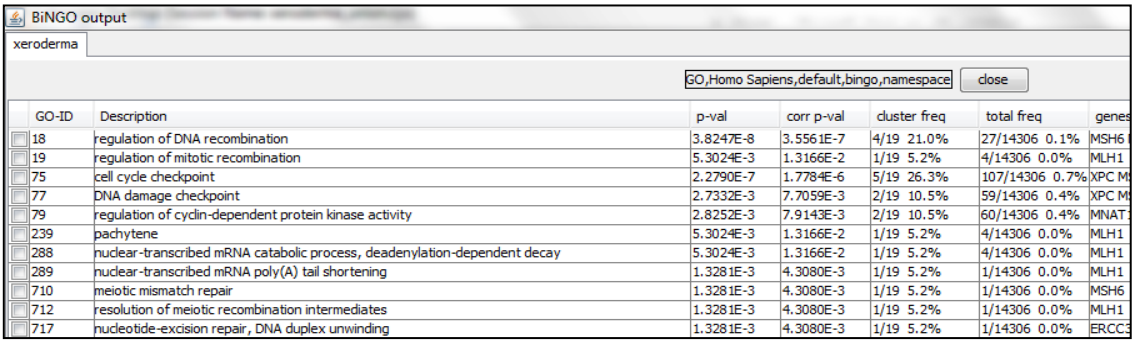


Figura 7: Análise de agrupamentos de uma rede PPI usando o plugin MCODE para a identificação de pontos de conexão da rede denominados de *hubs*.

[50] http://www.allegroviva.com/allegromcode

O *plugin* BiNGO[51] (*Biological Networks Gene Ontology*) permite determinar as categorias de ontologia gênica (GO) presentes em um conjunto de genes ou em um sub-grafo biológico de uma rede (Maere, 2005). Isto é feito através da identificação de *hubs* em uma rede com o intuito de permitir a análise de suas funcionalidades e conexões usando a classificação determinada pelo Consórcio de Ontologia Gênica[52] (Maere, 2005). O objetivo deste consórcio é produzir um vocabulário controlado para ser usado em eucariotos, mesmo sabendo-se que, o conhecimento sobre os genes e papéis das proteínas em células cresce e muda constantemente (The Gene Ontology Consortium et al., 2000). As ontologias servem para classificar as redes conforme suas funções biológicas, que podem ser componente celular, processo biológico ou função molecular (The Gene Ontology Consortium et al., 2000).

A Figura 8 apresenta a análise de ontologia gênica da doença xeroderma pigmentosa (Figura 6). O resultado mostra o código e a descrição da funcionalidade de cada proteína, bem como alguns valores estatísticos e o gene envolvido na função biológica como resultado do uso do *plugin* BINGO.


BiNGO output

xeroderma

GO,Homo Sapiens,default,bingo,namespace | close

GO-ID	Description	p-val	corr p-val	cluster freq	total freq	genes
18	regulation of DNA recombination	3.8247E-8	3.5561E-7	4/19 21.0%	27/14306 0.1%	MSH6
19	regulation of mitotic recombination	5.3024E-3	1.3166E-2	1/19 5.2%	4/14306 0.0%	MLH1
75	cell cycle checkpoint	2.2790E-7	1.7784E-6	5/19 26.3%	107/14306 0.7%	XPC M
77	DNA damage checkpoint	2.7332E-3	7.7059E-3	2/19 10.5%	59/14306 0.4%	XPC M
79	regulation of cyclin-dependent protein kinase activity	2.8252E-3	7.9143E-3	2/19 10.5%	60/14306 0.4%	MNAT
239	pachytene	5.3024E-3	1.3166E-2	1/19 5.2%	4/14306 0.0%	MLH1
288	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	5.3024E-3	1.3166E-2	1/19 5.2%	4/14306 0.0%	MLH1
289	nuclear-transcribed mRNA poly(A) tail shortening	1.3281E-3	4.3080E-3	1/19 5.2%	1/14306 0.0%	MLH1
710	meiotic mismatch repair	1.3281E-3	4.3080E-3	1/19 5.2%	1/14306 0.0%	MSH6
712	resolution of meiotic recombination intermediates	1.3281E-3	4.3080E-3	1/19 5.2%	1/14306 0.0%	MLH1
717	nucleotide-excision repair, DNA duplex unwinding	1.3281E-3	4.3080E-3	1/19 5.2%	1/14306 0.0%	ERCC3

Figura 8: Análise de ontologia gênica para da rede PPI da doença xeroderma pigmentosa.

[51] http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html

[52] http://www.geneontology.org/

A última medida a ser apresentada é a análise de centralidade (Figura 9) conforme foi descrita na seção 3.4.5. Esta análise pode ser feita usando o programa CentiScaPe[53] (*Centralities for Cytoscape*) que calcula a medida de diferentes parâmetros de centralidade para uma rede e, permite ao usuário analisar as relações existentes entre os dados experimentais fornecidos pelos usuários, além de identificar os nós da rede que são relevantes para analisar visões topológicas (Scardoni et al., 2009).

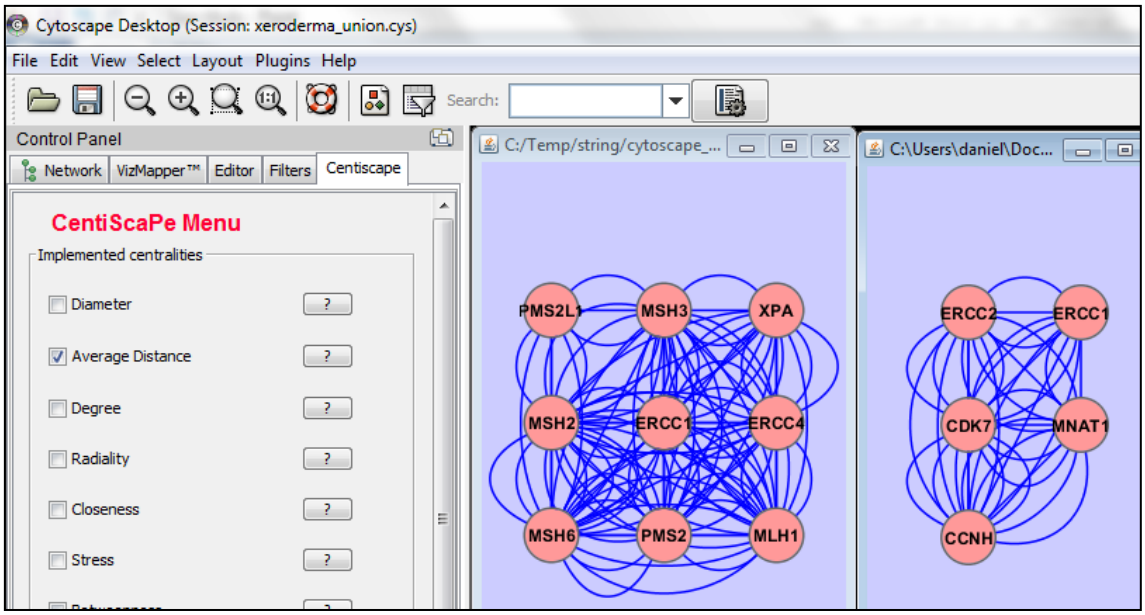


Figura 9: Análise de centralidade usando o *plugin* CentiScape no programa Cytoscape.

A Figura 9 apresenta o resultado do uso do *plugin* CentiScape para o cálculo da distância média (rede apresentada à direita) para a rede de interação de proteína da doença xeroderma pigmentosa (apresentada à esquerda).

3.6 WORKFLOWS CIENTÍFICOS

Um *workflow* diz respeito à automatização de procedimentos, onde documentos, informações ou tarefas são passadas entre os participantes de acordo com um conjunto pré-definido de regras, para se alcançar ou contribuir no objetivo global de um negócio

[53] http://www.cbmc.it/~scardonig/centiscape/centiscape.php

(Wfmc, 2004; Sommerville, 2007). Apesar de um *workflow* permitir a organização manual de procedimentos, na prática a maioria dos *workflows* são organizados dentro de um contexto do sistema de informação para prover um apoio automatizado aos procedimentos (Wfmc, 2004; Sommerville, 2007).

Workflows científicos são usados para descrever experimentos *in silico* em bioinformática (Silva, 2006; Freire et. al, 2006; Davidson et, al, 2007; Silva et. al, 2010; Chirigati and Freire, 2012) e, o seu uso tem crescido muito nos últimos anos (Freire et. al, 2006; Davidson et, al, 2007; Silva et. al, 2010; Linke et al., 2011; Chirigati and Freire, 2012). Estes *workflows* são construídos através de uma sequência de passos que caracterizam um fluxo de execução onde, cada um destes passos usam programas de terceiros (Silva, 2006). Cada execução do *workflow* é considerada um experimento e, de acordo com o resultado alcançado, os parâmetros podem ser revistos e o *workflow* é novamente executado ou, eventualmente, é preciso realizar algumas execuções parciais novamente (Silva, 2006; Freire et. al, 2006; Davidson et, al, 2007; Silva et. al, 2010; Chirigati and Freire, 2012). A execução de um determinado *workflow* pode não terminar ou não oferecer resultados conclusivos, porém é importante manter registrado os experimentos com execução bem sucedida, bem como aqueles com execuções defeituosas (Meyer et al., 2004; Freire et. al, 2006; Davidson et, al, 2007; Silva et. al, 2010; Chirigati and Freire, 2012).

Existem programas específicos para a criação e execução do *workflow*, tais como, Taverna (Oinn et al., 2004), Kepler (Altintas et al., 2004; Wang et al., 2012a), VisTrails (Callahan et al., 2006), Galaxy (Goecks et al., 2010) e Conveyor (Linke et al., 2011). Estas ferramentas permitem desenhar o *workflow* e fornecem mecanismos para acessar os programas de terceiros usando *web services*, chamadas remotas via URL ou

chamadas via linha de comando (Altintas et al., 2004; Oinn et al., 2004; Callahan et al., 2006; Linke et al., 2011; Wang et al., 2012a).

Para usar uma das ferramentas citadas anteriormente, é necessário conhecer e compreender o seu funcionamento interno, visando usar e integrar corretamente os softwares de terceiros, como é possível observar na explicação do uso do Kepler (Wang et al., 2012a). Por si só, compreender o funcionamento dos programas de terceiros é uma tarefa árdua e, quando juntamos ambos, a complexidade pode aumentar muito (Goecks et al., 2010). Uma alternativa a isto é desenvolver o *workflow* sem o uso destas ferramentas utilizando uma linguagem de programação para desenvolver o seu próprio workflow (Köster and Rahmann, 2012). Neste caso, a complexidade de usar programas de terceiros continua, mas o desenvolver já conhece a linguagem de programação. Entretanto se não a conhecer, a complexidade é a mesma da situação anterior.

Um *workflow* é construído a partir de vários passos ligados que consumem entradas, processam e convertem dados e produzem resultados (Linke et al., 2011; Wang et al., 2012a). Os mais simples *workflows* são cadeias lineares de passos para converter entradas de dados para saídas desejadas (Linke et al., 2011; Wang et al., 2012a). Procedimentos mais complexos podem incluir laços de repetição, ramificações, processamento paralelo e condicional, leitura de várias fontes de dados e, devem escrever diferentes saídas de dados usando diferentes formatos (Linke et al., 2011; Wang et al., 2012a).

Muitos experimentos *in silico* em bioinformática são feitos usando o conceito de *workflow* para gerar *workflows* genéricos ou específicos. Exemplos de workflows genéricos foram desenvolvidos para usar a linguagem de programação *phyton* para o processamento de agrupamento de dados (Köster and Rahmann, 2012), ou para integrar vários programas e ferramentas de bioinformática usando o Taverna. Desta forma, há a

possibilidade do usuário de bioinformática usar estas ferramentas para montar o seu próprio *workflow* (Bruin et al., 2012). Uma abordagem para detecção probabilística de motivos (Claeys et al., 2012) e uma análise de perda e ganho de genes em genomas distantemente relacionados (Ptitsyn and Moroz, 2012) são exemplos de *workflows* com propósito específico.

Capítulo I

DIS2PPI: A WORKFLOW DESIGNED TO INTEGRATE PROTEOMIC AND GENETIC DISEASE DATA

DANIEL LUÍS NOTARI[1], SAMUEL BRANDO OLDRA[1], CRISTIAN REOLON[1], MAURICIO ADAMI MARIANI[1], DIEGO BONATTO[2],*

[1]CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO E INSTITUTO DE BIOTECNOLOGIA, UNIVERSIDADE DE CAXIAS DO SUL, CAXIAS DO SUL, BRASIL.HTTP://WWW.UCS.BR

[2]CENTRO DE BIOTECNOLOGIA DO ESTADO DO RIO GRANDE DO SUL, SALA 219, DEPARTAMENTO DE BIOLOGIA MOLECULAR E BIOTECNOLOGIA, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, UFRGS, AVENIDA BENTO GONÇALVES 9500 - PRÉDIO 43421, CAIXA POSTAL 15005, CEP 91509-900, PORTO ALEGRE, RIO GRANDE DO SUL, BRASIL.HTTP://WWW.UFRGS.BR

Manuscrito aceito no periódico: ***International Journal of Knowledge Discovery in Bioinformatics (IJKDB)***

Dis2PPI: A workflow designed to integrate proteomic and genetic disease data

Summary

Experiments in bioinformatics are based on protocols that employ different steps for data mining and data integration, collectively known as computational workflows. Considering the use of databases in the biomedical sciences software that is able to query multiple databases is desirable. Systems biology, which encompasses the design of interactomic networks to understand complex biological processes, can benefit from computational workflows. Unfortunately, the use of computational workflows in systems biology is still very limited, especially for applications associated with the study of disease. To address this limitation, we designed Dis2PPI, a workflow that integrates information retrieved from genetic disease databases and interactomes. Dis2PPI extracts protein names from a disease report and uses this information to mine protein-protein interaction (PPI) networks. The data gathered from this mining can be used in systems biology analyses. To demonstrate the functionality of Dis2PPI for systems biology analyses, we mined information about xeroderma pigmentosum and Cockayne syndrome, two monogenic diseases that lead to skin cancer when the patients are exposed to sunlight and neurodegeneration**.**

Keywords: Data integration, system biology, interactomes networks.

1. INTRODUCTION

Less than two percent of DNA actually encodes proteins from human genome and more than fifth percent consists of repeated sequences (transposons, movable DNA sequences) (Lehninger et al., 2005). So, there are many researches to be done to discovery DNA and proteins undefined yet. Biological databases are important tool used for bioinformatics analyses like that mentioned above. All biological databases are built taking into consideration the amount of information generated, its logical structure, and the diversity of tools used to gain access to and analyze the information stored (Lesk, 2008). Thus, any bioinformatics application must be able to share and/or integrate data from more than one biological database. To perform this task, the biological data present in the initial database should be read and processed, the output

result should be usable as input for another database, and all data should be collected. This process is repeated when a third database is available, resulting in time-consuming data mining and analysis with results that are not always satisfactory.

Unfortunately, the diversity of available tools and the unique logical structure of each database make the simultaneous retrieval of information from multiple sources a difficult task for the user. Two important biological databases containing useful information are (i) Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim), which describes monogenic diseases in humans and their cellular and molecular components (Amberger et al., 2009), and (ii) STRING [http://string-db.org/], a database of known and predicted protein-protein interactions from a large number of organisms, including direct (physical) and indirect (functional) association (Jensen et al., 2009). Unfortunately, the crosstalk between these two databases is very limited, although both contain similar information. This heterogeneous task character a workflow process.

Workflows are used in many bioinformatics experiments to follow a protocol to establish the steps that must be taken and the order in which these steps must be undertaken by the user to execute that protocol. Thus, the organization of these steps in an automatic manner is desirable in many research areas within bioinformatics and the biological sciences. Peng et al. (2009) and Downing et al. (2009) have used computational workflows to define steps for biological experiments using databanks. The results of these experiments were then used to query databases. For example, Peng et al. (2009) developed workflow steps that verify post-transcriptional gene regulation at the miRNA level in complex human diseases, focusing on liver tissue biopsies contaminated or uncontaminated by the hepatitis C virus. Workflows can also be applied to health population analyses, as demonstrated by Downing et al. (2009). One example of this application is the creation of a database to query heterogeneous data sources and extract useful disease information following dynamic neuroscience (Cheung et al., 2009), genomic (Nuzzo and Riva, 2009), and computerized clinical guidelines (Damiani et al., 2010). Another interesting application of the use of workflows is to analyze microscopy images, leading to multiple classifications and the elucidation of complex phenotypes (Misselwitz et al., 2010).

Interestingly, many applications that use web services to extract and integrate data from biological databases (e.g., genomic or proteomic) could also use workflows to define the main steps used for computational processing. Silva et al. (2006) defined a

workflow management engine independent of the data source that is based on a set of scientific web services, a set of template data web services, and a database manager. Workflows designed for data management (downloading, importing, extracting, integrating, and normalizing), data and metadata annotation assignments, queries, and custom and meta-analyses of gene expression datasets have been developed (Bisognin et al., 2009).

One of the primary purposes of a scientific workflow is to analyze the results generated from multiple simultaneous database mining operations. Some areas of bioinformatics, including systems biology, can make use of workflows to better integrate mined biological data in useful models. Systems biology enables the study of interactions among the different components of biological entities on a large scale (Weston and Hood, 2004; Barabasi and Oltvai, 2004), and the use of analytical high-throughput methods generates a large amount of information about the behavior of an organism under particular physiological conditions (Barabasi, 2002).

To help researchers integrate different types of high-throughput data, software and algorithms have been developed to facilitate the analysis and design of biological models (Zöllner et al., 2005; Gehlenborg et al., 2010; Linke et al., 2011), including those that integrate data on monogenic diseases and protein-protein networks. Some prominent examples of these systems include the NCBI Entrez Utilities Web Service (Wheeler et al., 2006), which enables developers to access DNA, RNA, protein, and genetic disease data; the Agilent Literature Search Software, which can generate PPI networks from literature data mining [http://www.agilent.com/labs/research/litsearch.html]; and DisGeNET, software that can query and analyze a network representation of human gene-disease networks (Bauer-Mehren et al., 2010). Using these programs, it is possible to query a database with the name of a disease and obtain a protein-protein interaction (PPI) network, but each software program mentioned here must be used separately, and users must query, select, process, and prepare data from each program separately. None of the software programs mentioned is able to query multiple databases simultaneously to generate a diseasome network. A diseasome network is a graph of disorders and disease genes linked by known disorder–gene associations that offers a platform from which to explore, in a single-graph theoretic framework, all known phenotype and disease gene associations, indicating the common genetic origin of many diseases (Goh et al., 2007).

To better generate diseasome networks from multiple high-throughput data, Dis2PPI allows the user to generate a diseasome network using a single program that queries OMIM and STRING databases and/or other databanks and software.

2. MATERIAL AND METHODS

2.1 DATABASES

The OMIM database was queried using human genetic diseases as search terms. We chose xeroderma pigmentosum and Cockayne syndrome as diseases for which we mined information from OMIM. The STRING database was used to query protein-protein interaction (PPI) networks using the names of proteins extracted from our query of OMIM genetic diseases. In this case, proteins related to xeroderma pigmentosum and Cockayne syndrome were used to query the STRING database. The parameters used in STRING software were as follows: all prediction methods enabled, excluding text mining; 20 to 50 interactions; degree of confidence, high (0.700); and a network depth equal to 1. The OMIM database were accessed by NCBI Eutils WSDL[54] specification using Entrez Utilities Web Service[55] with Apache Axis2 for Java (Amberger et al., 2009).

2.2 DIS2PPI DESIGN AND VALIDATION

The scientific workflow integrated in the Dis2PPI[56] software [http://www.danlian.com.br/dis2ppi] is explained in Fig. 1. Initially, a genetic disease search term is input by the user (Fig. 1). We used xeroderma pigmentosum and Cockayne syndrome, two monogenic diseases, to mine for information about the major PPI networks associated with these diseases. Briefly, Dis2PPI submits a genetic disease as a query to the OMIM database using web services. The output from this first analysis is an HTML file with a report of all monogenic diseases that match the original input. The software then presents a report for the user's evaluation. This report is presented as a tree window, allowing the user to select the desired data. After the user selects the desired data, Dis2PPI proceeds by extracting protein names from the report and

[54] Details is available at http://www.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/eutils.wsdl

[55] Details is available at http://www.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/DOC/esoap_java_help.html

[56] Dis2PPi software is distributed under GPL license terms.

displaying the results for the user as a new tree window. The user can then add or remove duplicate or synonymous protein names from a chosen organism. The resulting list of selected proteins is used as input to the STRING database using a Uniform Resource Locator (URL) method, returning a PPI network file that matches the protein name input supplied by the user. The resulting network can be displayed in a browser or can be imported with the Cytoscape (Shannon et al., 2003; http://www.cytoscape.org) tool for further systems biology analyses (e.g., modularity or centralities analyses).

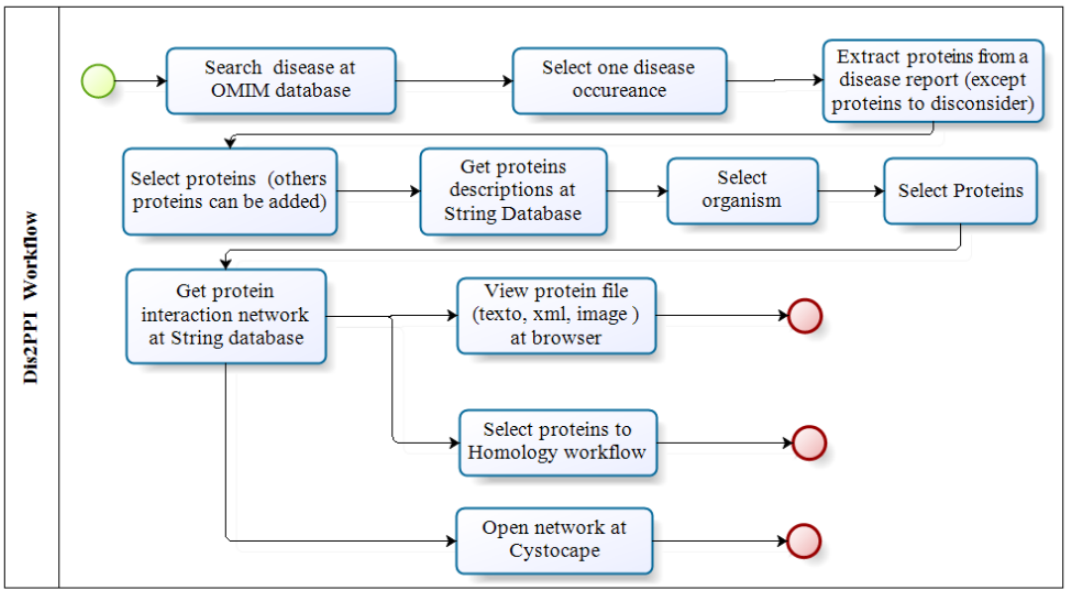


Figure 1. A workflow describing the major steps of Dis2PPI.

2.3. CLUSTER ANALYSIS

The binary PPI network obtained from the initial search was analyzed using the Cytoscape plugin Molecular Complex Detection (MCODE; Bader and Hogue, 2003), which is available at http://baderlab.org/Software/MCODE. MCODE is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions according to given parameters stipulated by the researcher (Bader and Hogue, 2003). The parameters for cluster finding were as follows: loops included; degree cutoff 2; expansion of a cluster by one neighbor shell allowed (fluff option enabled); deletion of a single connected node from clusters (haircut option enabled); node density cutoff, 0.1; node score cutoff, 0.2; kcore, 2; and maximum network depth, 100.

2.4. CENTRALITY ANALYSIS

Centrality analysis was performed using the program CentiScaPe (Compe and Egly, 2012; http://profs.sci.univr.it/~scardoni/centiscape/centiscapedownload.php). In this analysis, CentiScaPe scanned the network, and measured the significance levels of node degree and betweenness to establish the most important nodes (proteins) in the network. Thus, it is possible to detect the most relevant proteins for a determined pathway or connection. Betweenness indicate the number of shortest paths that go through each protein (Scardoni et al., 2009). Thus, proteins with a high betweenness score control the information flow through a network. The higher a protein's betweenness score, the higher the probability that the protein connects to different clusters or biological processes, which characterize so-called bottlenecks. Finally, the node degree is a simple measure that indicates the number of connections that involve a specific protein. Proteins with a high node degree are called hubs (Scardoni et al., 2009) and have key regulatory functions in the cell. Thus, the local topologies of the network (hubs and bottlenecks) were defined considering the threshold generated by node degree and betweenness centralities calculated by CentiScape 1.0. In this sense, bottlenecks were defined as a node with a value above the threshold calculated for node degree and betweenness, and hubs are nodes with high node degree and low betweenness.

3. RESULTS

3.1 DIS2PPI IMPLEMENTATION

Dis2PPI software was designed as a scientific workflow tool used to search and integrate genetic disease information with protein-protein interaction (PPI) networks. Dis2PPI allows the user to evaluate the major biological pathways and protein complexes that can be affected by a specific genetic disease or physiological condition. Dis2PPI employs five primary steps used to mine disease-protein interaction networks (Fig. 1): (i) searching for a specific disease, primarily human genetic diseases; (ii) extracting a list of proteins names associated with the disease; (iii) searching for protein-protein interactions (PPI); (iv) visualizing the resulting PPI network; and (v) proceeding with a systems biology analysis.

The first step receives a user input of a monogenic disease, which is used to access the OMIM database to obtain a description of the disease. After the desired

description is selected by the user (Fig. 2A), the software extracts protein names from the disease description (Fig. 2B) and uses this as input to access the STRING database remotely (Fig. 3A) and capture PPI networks. After the PPI networks are gathered (Fig. 3B), the user can obtain a text file containing information about the disease-genetic networks identified and upload the text file to various programs for network analyses (e.g., Cytoscape).

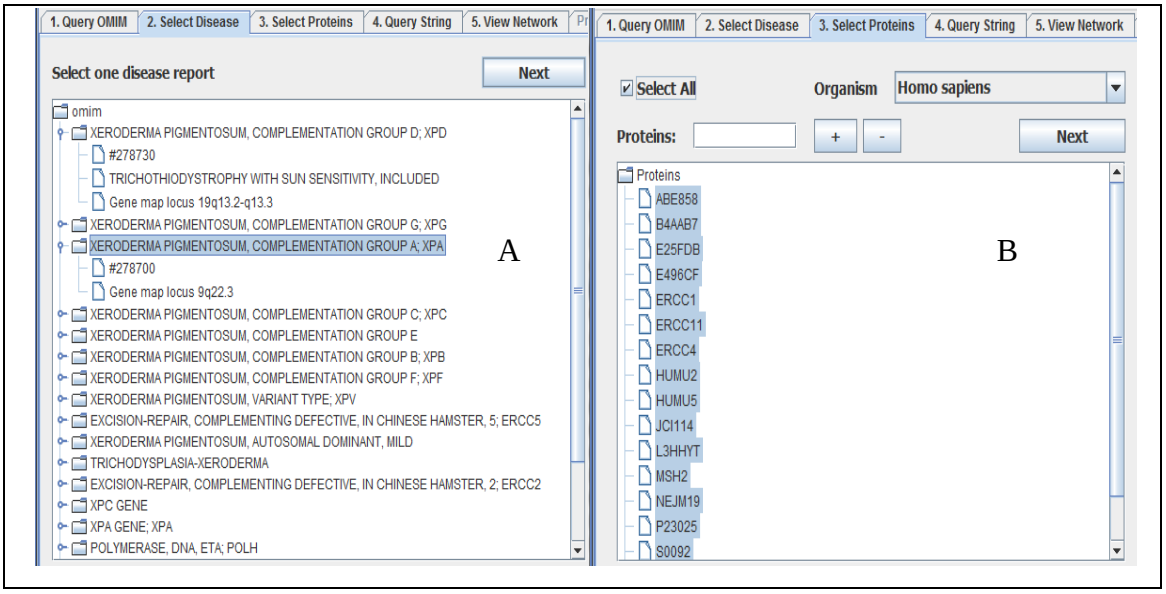


Figure 2. (A) The initial analysis of xeroderma pigmentosum type XPA performed by Dis2PPI software. Dis2PPI lists different diseases that are related to Xeroderma pigmentosum, allowing the user to select a specific genetic pathology. Once selected, the user also chooses the species whose proteins will be investigated (B).

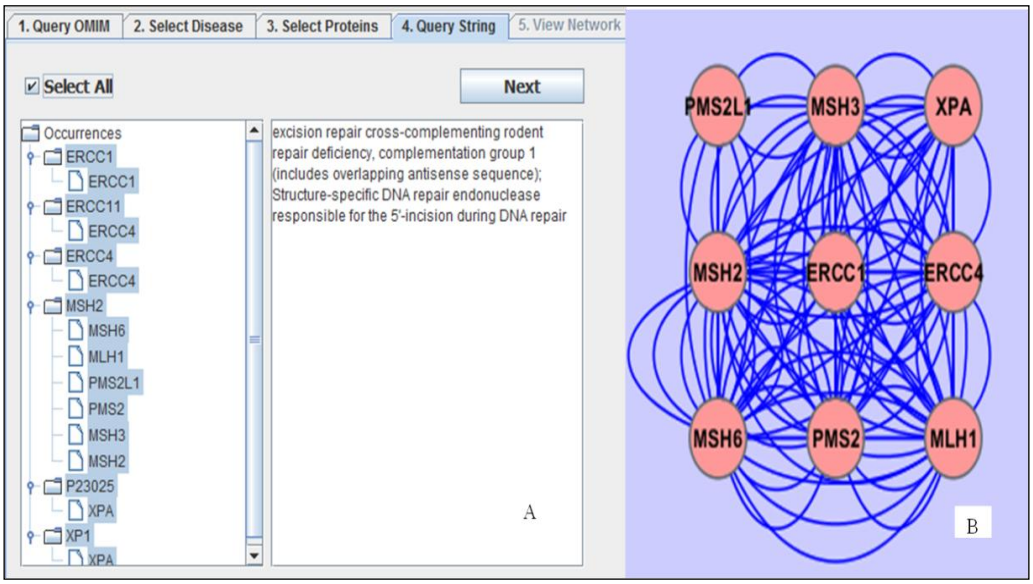


Figure 3. (A) Descriptions of the major proteins gathered from xeroderma pigmentosum data by Dis2PPI choosing *Homo sapiens* as the model species. The major PPI network generated from STRING 9.0 prospection by Dis2PPI. The gathered PPI network was visualized in Cytoscape (B).

3.2 DATA EXTRACTION AND AUTOMATIC DATA INTEGRATION

To effectively implement the steps described previously, Dis2PPI makes use of regular expressions to extract biological information for data mining and systems biology analyses. Regular expression is a computational method to look for patterns in a text file, i.e. this method is interesting to scan unknown text to find expressions that matches with a target search. In our case, the regular expression used was ([A-Z]{1}[A-Z0-9]{1,5}), which extracts a protein identifier that begins with a capital letter and whose next five characters can be capital letters or numbers. When a regular expression is executed, many terms are correct proteins names like CCNH and CDK7. But it is possible to obtain false-positive proteins names like HTML tags or proteins names that do not match with the chosen organism. More details are related on section 3.4. Regular expressions were also used in Dis2PPI implementation to access String databases to correct mount URL (e_([A-Za-z0-9]{2}[A-Za-z0-9_]+)\\.) to download the network image file from String database.

OMIM and String databases were accessed using remote calls. Web services were used to access the OMIM database using Entrez Program Utilities (Amberger et al., 2009) tools (Fig. 4A), and URL access method was used to access the STRING database (Fig. 4B).

A

```
EUtilsServiceStub service = new EUtilsServiceStub();
EUtilsServiceStub.ESearchRequest req = createSearchRequest(qr);
EUtilsServiceStub.ESearchResult res = service.run_eSearch(req);
listIdsDisease = processSearchResultRequest(res);
return listIdsDisease;
```

B

```
URL url = new URL("http://string-db.org/api/tsv-no-header/resolve?identifier=" +
                  vo.getTerm() + "&species=" + vo.getSpecie());
HttpURLConnection connection = (HttpURLConnection) url.openConnection();
connection.addRequestProperty("Request-Method", "GET");
connection.setDoInput(true);
connection.setDoOutput(false);
connection.connect();
BufferedReader reader = new BufferedReader(new InputStreamReader(url.openStream()));
```

Figure 4. The Java source code used to access the ESearch tool web service to obtain a disease report (A) and the Java source code to access the String database to obtain a protein description (B).

The extract proteins step (Fig. 1) was implemented as a regular expression (presented above) to extract all proteins cited in the HTML-format disease report generated by the OMIM database (Fig. 2B). When the regular expression was executed, proteins appeared in the results, including A188, APOA1, B15, CLEC16, and many others. However, the algorithm can also mine for non-protein names, which can be excluded by the researcher to avoid the inclusion of non-protein names and expressions in the results of the analysis. The Dis2PPI software access the String database to query all proteins names extracted. The validate proteins names and its descriptions were shown to user (Fig. 3A). Proteins names that did not have a query match on String database were discarded to avoid user work with false-positive proteins names.

The get proteins description step (Fig. 1) was implemented to separate the query results. Any line from the query results contains information that is then combined with the query parameters. The separation of these two forms of information is needed to enable the user to select all proteins (of the user's choice) with which to generate the protein-protein interaction network (Fig. 3A).

The get protein interaction network step (Fig. 1) was implemented as a regular expression to create the URL string needed to access the STRING database and obtain a PPI network XML flat file to use with Cytoscape software. Also, an image file was downloaded to display to the user (Fig. 3B) at web browser.

3.3 A CASE ANALYSIS WITH DIS2PPI – XERODERMA PIGMENTOSUM

To evaluate the functionality of Dis2PPI, the disease xeroderma pigmentosum type XPA was used as the initial query. The data returned from the OMIM included several entries, being the first selected (Fig. 2A). After selected all protein names and *Homo sapiens* as the model organism to be searched (Fig. 2B), we mined PPI network information using STRING database. The data gathered from STRING include many proteins and their descriptions (Fig. 3A). Proteins that were not found to belong to a major PPI network were discarded in this analysis. Once generated, the PPI network was uploaded in Cytoscape (Fig. 3B).

A PPI network was acquired for xeroderma pigmentosum type XPA using Dis2PPI software. The whole process of protein extraction from OMIM report's diseases followed by PPI network generation from STRING database is shown on Fig. 5A to D. An important point is that all data gathered is validated by the user. Dis2PPI extracted all candidate proteins (Fig. 5B) from OMIM's report (Fig. 5A). The expert

can remove or add protein names to avoid false-positive information. In this example, ABE858, B4AAB7, E25FDB, E496CF, JCI113, L3HHYT, S00928, SO161, S77E, TTD1, TTD2, XP4 and XP8 were queried to STRING database as part of *Homo sapiens* proteome, and the results indicated that these names are not valid proteins identifications.

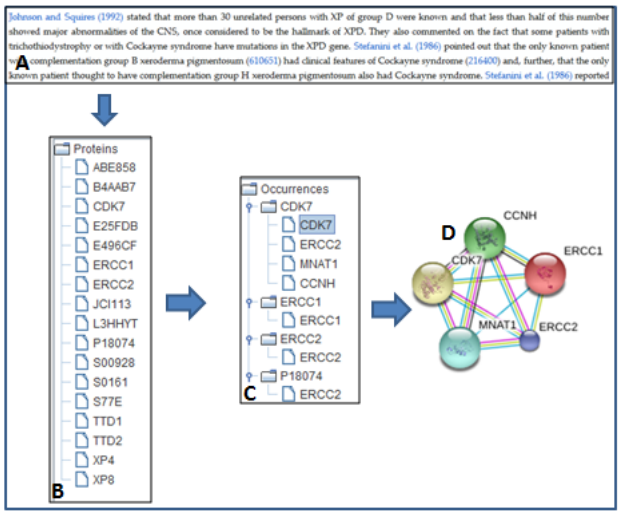


Figure 5. Example of protein data extraction and processing using Dis2PPi software and expert advises. In (A) the OMIM's report for xeroderma pigmentosum was used to extract all potential protein names (B). After selecting the correct protein names (C), the list was subjected to STRING database in order to generate a PPI network (D).

On the other hand, CDK7, MNAT1, ERCC1, ERCC2 and P18074 are valid proteins names, as indicated by STRING database (Fig. 5C). At this point, the user can select these proteins to obtain a xeroderma pigmentosum (XP) PPI network. Dis2PPI generates a new XP PPI network (Fig. 5D) by querying STRING database, allowing the user to apply different systems biology analyses.

3.4 XERODERMA PPI NETWORK SYSTEM BIOLOGY ANALYSIS

Dis2PPI was used to obtain three different XP networks for XPA, XPC and XPD groups (Fig. 6A). The Merge Cytoscape plugin (Shannon et al., 2003) was used to create a new PPI network using union method to merge all XP networks (Fig. 6B).

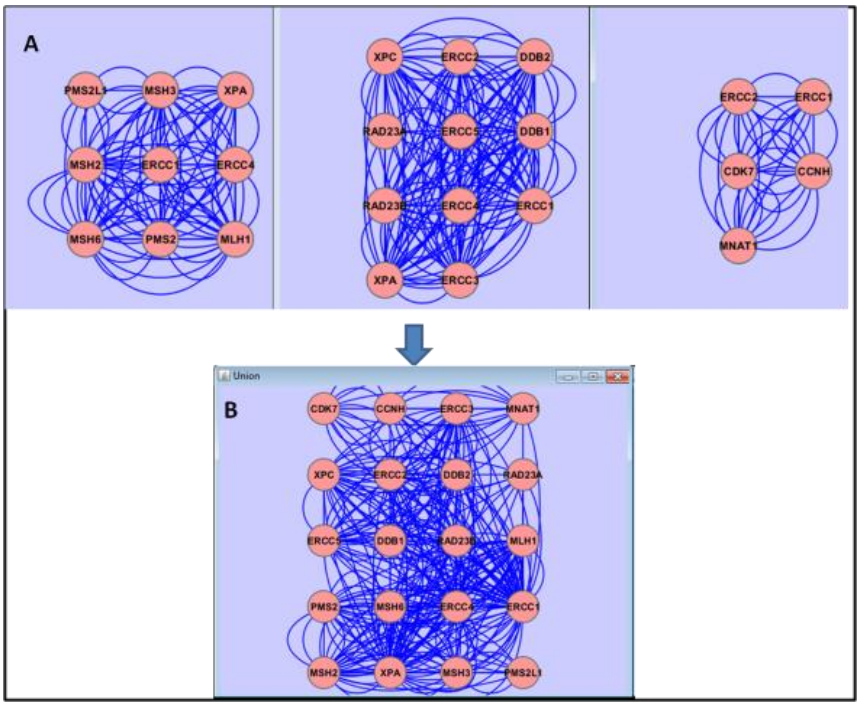


Figure 6. In (A), xeroderma pigmentosum PPI network was acquired for XPA, XPC and XPD groups. Then, a merged PPI network obtained using Merge Cytoscape plugin (B).

The next step was to analyze the merged XP network (Fig. 6B) in terms of cluster number and node centralities. Considering cluster analysis, it was possible identify three major clusters (Fig. 7), which are associated to: (i) DNA damage removal by nucleotide excision repair (NER), (ii) DNA mismatch repair (MMR), and (iii) RNA transcription and NER (Table 1). In addition, centrality analysis was performed in order to identify the major nodes within the XP network. Centrality data indicated the presence of four major bottlenecks that are associated to the cluster 1, necessary for DNA damage recognition and repair. These bottleneck proteins are ERCC1, 2 and 4, and XPA proteins (Fig. 7). In addition to NER, one interesting cluster analyzed from major XP network is responsible for MMR process (Cluster 2; Fig. 7). Moreover, unclustered proteins like RAD23A and PMS2L1 were found within major XP network (Fig. 7).

Table 1. Specific gene ontology (GO) classes derived from physical protein-protein interactions observed in different subgraphs.

Cluster number	GO biological process category	GO number	*p*-value	Corrected *p*-value[a]	k[b]	f[c]
Cluster 1	Nucleotide-excision repair, DNA damage removal	718	3.56×10^{-30}	1.03×10^{-27}	10	21
	Response to UV	9411	2.40×10^{-15}	8.47×10^{-14}	7	62
Cluster 2	Mismatch repair	6298	5.28×10^{-15}	1.28×10^{-12}	5	22
Cluster 3	Nucleotide-excision repair, DNA damage removal	718	2.73×10^{-9}	3.84×10^{-7}	3	21
	RNA elongation from RNA polymerase II promoter	6368	3.54×10^{-8}	1.69×10^{-6}	3	48

[a]Values calculated from p-value after FDR application.

[b]Total number of proteins found in the network that belong to a specific GO.

[c]Total number of proteins belonging to a specific GO.

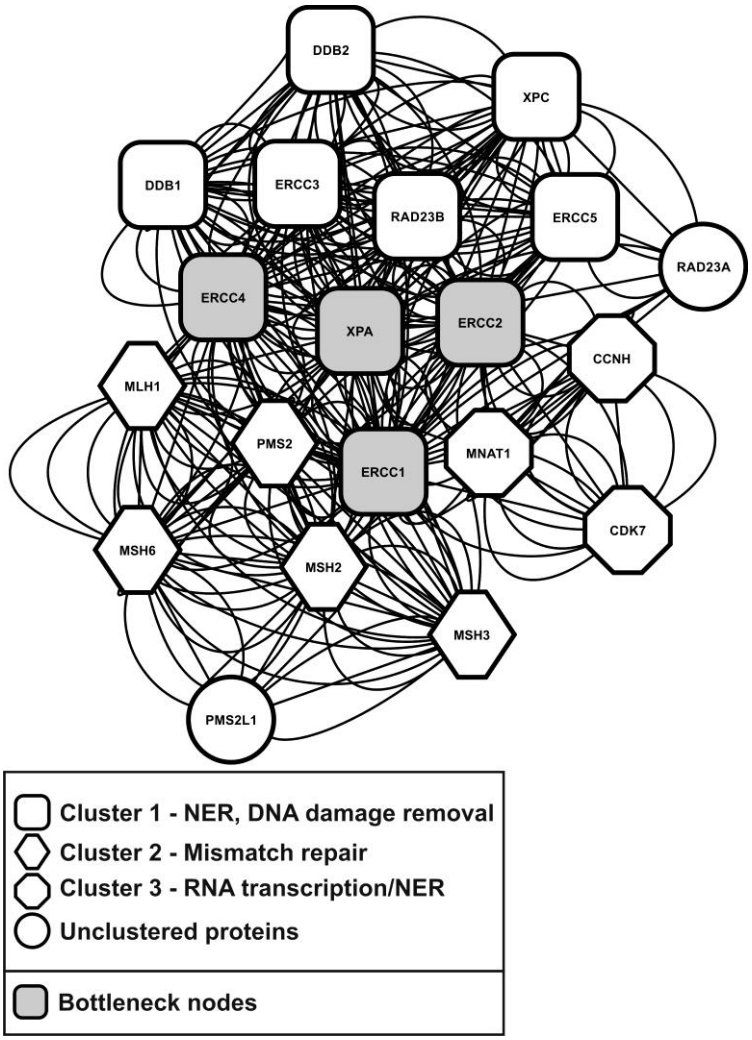


Figure 7. Cluster and centrality analyses of XP network. The three major clusters and the bottlenecks nodes are shown in the network by different node shapes and by a gray color, respectively (inset box).

3.5 EXPANDING THE XERODERMA PIGMENTOSUM NETWORK BY ADDING COCKAYNE SYNDROME-ASSOCIATED PROTEINS

In order to improve the systems biology analysis of XP network (Figure 7), we expanded the original XP network with proteins associated with Cockayne syndrome (CS), also a monogenic disease related to NER and MMR pathways.

The data gathered allow to the generation of a XP-CS network, which contains 81 nodes and 907 edges (Figure 8).

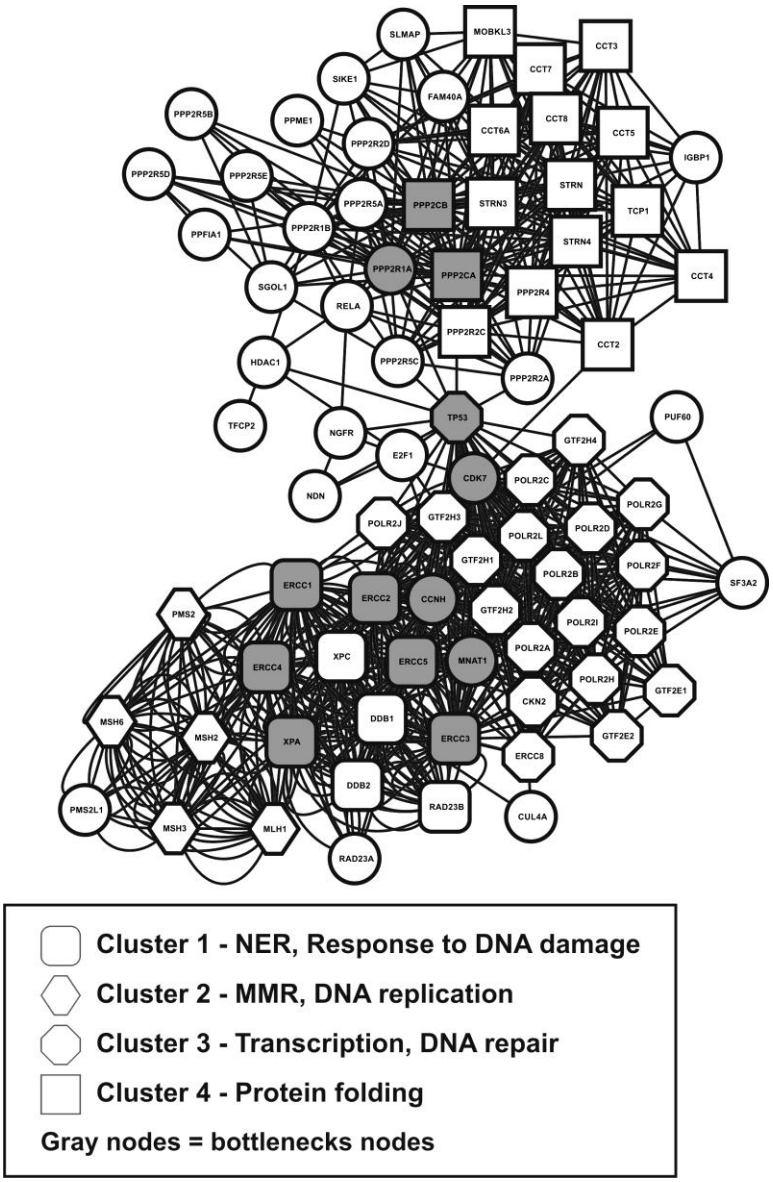


Figure 8. Cluster and centrality analyses of XP-CS network. The four major clusters and the bottlenecks nodes are shown in the network by different node shapes and by a gray color, respectively (inset box).

Cluster analysis of XP-CS network (Figure 8) indicated the presence of four major subgraphs, which are associated to: (i) NER pathway and response to DNA damage, (ii) MMR pathway and DNA replication, (iii) transcription and DNA repair processes, and (iv) protein folding mechanism (Table 2).

Table 2. Specific gene ontology (GO) classes derived from physical protein-protein interactions observed in different subgraphs of XP-CS network.

Cluster number	GO biological process category	GO number	*p*-value	Corrected *p*-value[a]	k[b]	f[c]
Cluster 1	Nucleotide-excision repair, DNA damage removal	718	3.06×10^{-30}	5.15×10^{-28}	10	21
	Response to UV	9411	4.41×10^{-14}	6.17×10^{-13}	6	Five
Cluster 2	Mismatch repair	6298	2.20×10^{-14}	5.61×10^{-13}	5	29
	DNA replication	6263	2.41×10^{-12}	4.10×10^{-11}	5	71
Cluster 3	RNA elongation from RNA polymerase II promoter	6368	1.95×10^{-41}	3.82×10^{-39}	17	47
	Nucleotide-excision repair, DNA damage removal	718	1.54×10^{-8}	1.04×10^{-7}	4	21
Cluster 4	Protein folding	6457	6.14×10^{-12}	9.76×10^{-10}	8	182

[a]Values calculated from p-value after FDR application.

[b]Total number of proteins found in the network that belong to a specific GO.

[c]Total number of proteins belonging to a specific GO.

Centralities analysis pointed to the major bottleneck nodes previously indentified only in the XP network (Figure 7) and also five new bottleneck nodes, which are TP53, CDK7, CCNH, MNAT1, PPP2R1A, PPP2CA, PPP2CB, ERCC4, and ERCC5. Some bottlenecks nodes, like those associated to the serine/threonine-protein phosphatase 2A catalytic subunit alpha and beta isoforms (PPP2CA and PPP2CB, respectively) and serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform (PPP2R1A) were not previously described for XP and CS diseases.

4. DISCUSSION

The NCBI Entrez Utilities Web Service enables developers to access Entrez Utilities via the Simple Object Access Protocol (SOAP) [http://www.w3.org/TR/soap/]. The current version of the NCBI Entrez Utilities Web Service provides access to DNA, RNA, and protein data from many species and allows users to query the OMIM database (Amberger et al., 2009). Dis2PPI uses the Entrez utilities web service to directly access OMIM disease reports. In this sense, other plug-ins commonly used to search for PPI networks associated with diseases include the Agilent Literature Search software (Agilent, 2012) for Cytoscape (Shannon et al., 2003) and DisGeNET (Bauer-Mehren et al., 2010).

Agilent Literature Search software generates an overview network of gene/protein associations. When a query is entered, it is submitted to multiple user-selected search engines, and the retrieved results (documents) are fetched from their respective sources. Each document is then parsed into sentences and analyzed for protein-protein associations. Agilent Literature Search software uses a set of "context" files (lexicons) for defining protein names (and aliases) and association terms (verbs) of interest (Agilent, 2012). Associations extracted from these documents are collected into a Cytoscape network. The sentences and source hyperlinks for each association are further stored as attributes of the corresponding Cytoscape edges (Agilent, 2012). The differences between Agilent and Dis2PPI are as follows: (i) Dis2PPI specifically queries the OMIM about diseases, whereas Agilent is an engine used to query many sources for information about genes and proteins; (ii) Dis2PPI extracts proteins from a disease report based on the use of regular expressions to identify protein names, and users can add or remove proteins associated with a specific disease; (iii) Dis2PPI does not use verbs as search terms.

By its turn, DisGeNET is another Cytoscape plug-in used to query and analyze a network representation of human gene-disease databases (Bauer-Mehren et al., 2010). DisGeNET allows the user to query specific diseases and their respective associated genes (Bauer-Mehren et al., 2010). This plug-in represents gene-disease associations in terms of bipartite graphs and provides gene-centric and disease-centric representations of the data (Bauer-Mehren et al., 2010). DisGeNET assists the user in the interpretation and exploration of complex human diseases with respect to their genetic origins using a variety of built-in functions (Bauer-Mehren et al., 2010). Dis2PPI allows the user to

mine for protein-disease associations in different species by generating a protein interaction network representation. While DisGeNET provides a gene- and disease-centric association, Dis2PPI have a interatomic approach by considering not only gene data, but also proteomic and literature data.

Thus, considering the algorithm implemented in Dis2PPI software, we prompt us to evaluate the functionality of the program by prospecting the major proteins related to xeroderma pigmentosum (XP) disease and then expanding the XP network by adding proteins that are related to Cockayne Syndrome (CS).

XP is a genetically heterogeneous autosomal recessive disorder characterized by increased sensitivity to sunlight with the development of carcinomas at an early age (Satokata et al., 1992) and characterized by increased sensitivity to ultraviolet (UV) irradiation and increased risk of skin cancer resulting from a defect in DNA repair pathways, especially nucleotide (NER) and base excision repair (BER) (Li et al., 1993). Patients of the group A, the most common form in Japan, show involvement of the central and peripheral nervous systems in addition to cutaneous lesions (Kanda et al., 1990). XPC is the most common form of XP in the caucasian population, accounting for over a third of all cases in this group, where cell lines from these patients were the least sensitive to UV-irradiation compared to 4 other cell lines, and exhibited a near-normal level of XPC mRNA (Li et al., 1993). Linkage of trichothiodystrophy (TTD) and xeroderma pigmentosum of complementation group D of XP (Nuzzo et al., 1986) and the ERCC2 DNA repair gene corrects the defect in XPD cells (Flejter et al., 1992).

Both NER and MMR are two DNA repair pathways evolutively conserved that are found in both prokaryotes and eukaryotes (Friedberg et al., 2005). NER and MMR are essential for the remotion of different types of base lesions generated on DNA, including ultraviolet light (UV)-induced cyclobutane pyrimidine dimer (CPD) and pyrimidine (6-4) pyrimidone photoproduct (6-4PP), as well as other bulky base adducts that can be induced by numerous chemical compounds (Friedberg, 2001; Hoeijmakers, 2001; Friedberg et al., 2005). It is known that defects in NER are associated with several human autosomal recessive hereditary disorders, such as XP. In this sense, the proteins ERCC1 and ERCC4 and XPA (Fig. 7) composes a ternary complex that recognizes the DNA damages induced by UV light (Park and Sancar, 1994), where ERCC1 and ERCC4 are two structure-specific DNA repair endonucleases responsible for the 5’-incision during DNA repair (Rahn et al., 2010), while XPA, a 32 kDa zinc metalloprotein, is the sole recognition factor required for NER activities and is believed

to play a role in verification of DNA damage (Friedberg et al., 2005). By its turn, ERCC2 is an ATP-dependent 5'-3' DNA helicase (White, 2009), component of the core-TFIIH basal transcription factor and is also involved in NER. In fact, these four proteins are essential for NER pathway, and their homologous are found in virtually all eukaryotes. Moreover, these proteins were found participating in different DNA repair pathways, like DNA recombination (Costa et al., 2003; Naegeli and Sugasawa, 2011) and mismatch repair (MMR). Interestingly, these four proteins where identified as bottleneck in our centrality analysis, reinforcing the major position that these proteins display on XP network (Fig. 7).

In addition to ERCC1, 2 and 4, and XPA, two proteins, MNAT1 and CCNH, which are found belonging to cluster 3 (Fig. 7), composes the catalytic subunit of the CDK-activating kinase (CAK) enzymatic complex (Drapkin et al., 1996) that regulates the cyclin-dependent kinase (CDK) 7. How MNAT1 and CCNH are connected to NER are unknown, opening a new perspective for the study of these two proteins in the context of XP disease.

Considering the CAK complex, the importance of this protein complex for NER have been described using an elegant DNA repair in vitro protocol with purified NER and CAK proteins (Araújo et al., 2000). The CAK proteins act during the initial phase of recruitment of NER proteins at the DNA lesion, where it is firstly recognized by the binding of XPC-RAD23B complex and/or the lesion sensor UV light-damaged DNA-binding protein (UV-DDB) complex, composed by DDB1 and XPE proteins. Interestingly, both complexes (XPC-RAD23B and UV-DDB) are found represented in our network (Fig. 7). It is important to note that XPC-RAD23B complex mediates the recruitment of the transcription initiation factor TFIIH to the DNA lesion, leading DNA to unwinding by the helicase activity of XPD and the ATPase activity of XPB (Compe and Egly, 2012). Once DNA opening was completed, XPA is recruited to the site of lesion, leading to the release of the CAK complex. The release of CAK enlarge DNA opening and promotes the recruitment of the XPF-excision repair cross complementing 1 (ERCC1) complex and XPG, releasing XPC-RAD23B and promoting DNA repair (Compe and Egly, 2012).

While RAD23A is found with XPC to repair DNA damages induced by UV light (Li et al., 1997), PMS2L1 belongs to the postmeiotic segregation increased 2-like proteins group that participate into MMR. Considering PMS2 group, it has been described its association with human DNA polymerase eta (Polη), which have an

essential role in translesion DNA synthesis, a process necessary for DNA damage tolerance (Kanao et al., 2009). Despite the fact that MMR is an important excision DNA damage pathway (Li, 2008), their roles on XP are less known. Experimental data using knockout transgenic murine models for XPA gene indicated that MMR cooperatively works with NER (Yoshino et al., 2002). Thus, our data prospection validated the importance of MMR for XP disease, serving as a model for new biochemical and genetic studies with human cells from XP patients.

Interestingly, CS is also an autosomal recessive progeroid disorder that affects the development and maintenance of different tissues (Weidenheim et al., 2009; Natale, 2011). All CS patients display neurological dysfunctions and a phenotype that resemble aging, gait defects and ocular and skeletal abnormalities, leading to patient death during childhood (Weidenheim et al., 2009; Natale, 2011). In this sense, mutations on ERCC6 and ERCC8, two proteins that participate in a specific NER pathway termed transcription-coupled repair (TC-NER; Lainé and Egly, 2006) leads to the development of CS. The TC-NER pathway is also associated with XP and with a less understood disease termed combined XP-CS disease, whose patients display clinical features of both XP and Cockayne syndromes (Lainé and Egly, 2006).

When considering TC-NER pathway, it is important to note that genes that are transcribed within a cell are repaired in a fast pace when compared to silenced regions of chromatin (Lainé and Egly, 2006). This fast response to specific DNA damages induced by UV light or chemical compounds ensures that RNA polymerase complex can resume the transcription of genes that are needed to maintain cell homeostasis. Thus, disturbances on TC-NER as observed in XP and CS diseases lead to a delay in the transcription of major genes, culminating in cell cycle arrest and/or cell death (Lainé and Egly, 2006).

Our systems biology analysis of XP-CS network (Figure 8) revealed the presence of biological processes associated to TC-NER as expected (e.g., NER and RNA elongation from RNA polymerase II promoter; Table 2), but also indicated the participation of protein folding as an important mechanism for XP and CS diseases (Table 2).

The protein folding process is represented by serine/threonine protein phosphatase 2 catalytic and regulatory isoforms (PPP2C and PPP2R, respectively; Figure 8) as well as by the calmodulin binding proteins named striatin (STRNs; Figure 8). Interestingly, both PPP2C/PPP2R and STRNs compose a protein complex named

striatin-interacting phosphatase and kinase (STRIPAK; Hyodo et al., 2012), which is necessary for the completion of cytokinesis. Defects on STRIPAK complex leads to formation of multinucleate cells (Hyodo et al., 2012), a condition that was already observed for cells defective in ERCC1 (Rageul et al., 2011). Noteworthy, murine models defective for STRNs and PPP2 complexes display problems in spine morphogenesis and synaptic signaling (Bartoli et al., 1999; Chen et al., 2012), similar to those neuropathological features observed in XP and CS diseases (Itoh et al., 1999). Finally, the data gathered for XP and CS diseases by using Dis2PPI software point to new biological processes and proteins previously not related for both monogenic syndromes, which can be used to design new experimental procedures in order to understand the pleiotropic effect of mutations associated to excision repair pathway.

5. CONCLUSION

Dis2PPI software was developed as a scientific workflow including both manual and automatic activities that can be used by the researcher. Automatic activities are used to query data from a public database (OMIM) to obtain a list of proteins and to analyze the relationships between these proteins and a disease. Automatic activities also perform data extraction using regular expressions and the creation of HTML files to obtain a protein interaction network from the STRING database. Manual activities are used to show the data to the user and allow the user to choose, add, or remove data and press a button to continue the workflow process.

Dis2PPI allows the user to mine protein-disease associations in different species and generate a protein-protein interaction network. The results gathered by the researcher can be analyzed by applying systems biology methods using dedicated software such as Cytoscape or Medusa. Finally, the results gathered by Dis2PPI can be used by the researcher to design validation experiments.

Acknowledgments

This work was supported by research grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Programa Institutos Nacionais de Ciência e Tecnologia (INCT de Processos Redox em Biomedicina-Redoxoma; grant number 573530/2008-4). We want to thank our colleagues, Helena

Ribeiro and Alexandre Nascimento, for their help with the development of the Dis2PPI software and with the writing of this article.

References

Agilent Literature Search Software, (2012). URL: <http://www.agilent.com/labs/research/litsearch.html>.

Amberger, J., Bocchini, C. A., Scott, A. F. and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Research,* 37(suppl 1): D793-D796.

Araújo, S. J., Tirode, F., Coin, F., Pospiech, H., Syväoja, J. E., Stucki, M., Hübscher, U., Egly, J. M., Wood, R. D. (2000). Nucleotide excision repair of DNA with recombinant human proteins: definition of the minimal set of factors, active forms of TFIIH, and modulation by CAK. *Genes & Development*, 14(3), 349-59.

Bader, G. D., Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics,* 4, 2.

Barabasi A. L. (2002). Statistical physics of complex networks. *Reviews of Modern Physics*, 74, 47-97.

Barabasi A. L., Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature ReviewsGenetics*, 5, 101-113.

Bartoli, M., Ternaux, J. P., Forni, C., Portalier, P., Salin, P., Amalric, M., Monneron, A. (1999). Down-regulation of striatin, a neuronal calmodulin-binding protein, impairs rat locomotor activity. *Journal of Neurobiology*, 40, 234-243.

Bauer-Mehren, A., Rautschka, M., Sanz, F. Furlong, L. (2010). DisGeNET - a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, 26(22), 2924-2926.

Bisognin, A., Coppe, A., Ferrari, F., Risso, D., Romualdi, C., Bicciato, S. and Bortoluzzi, S. (2009). A-MADMAN: Annotation-based microarray data meta-analysis tool. *BMC Bioinformatics*, 10, 201.

Chen, Y. K., Chen, C. Y., Hu, H. T., Hsueh, Y. P. (2012). CTTNBP2, but not CTTNBP2NL, regulates dendritic spinogenesis and synaptic distribution of the striatin-PP2A complex. *Molecular Biology of the Cell*, 23, 4383-4392.

Cheung, K., Frost, R., Marshall, M. S., Prud'hommeaux, E., Samwald, Matthias, Zhao, J. and Paschke, A. (2009). A journey to Semantic Web query federation in the life sciences. *BMC Bioinformatics*, 10 (Suppl 10), S10.

Compe, E. and Egly, J. M. (2012). TFIIH: when transcription met DNA repair. *Nature Reviews Molecular Cell Biology*, 13(6), 343-54.

Costa, R. M., Chigançcas, V., Galhardo, R. S., Carvalho, H. and Menck, C. F. (2003). The eukaryotic nucleotide excision repair pathway. *Biochimie*, 85(11), 1083-99.

Damiani, G., Pinnarelli, L., Colosimo, C., Almiento, R., Sicuro, L., Galasso, R., Sommella, L. and Ricciardi, W. (2010). The effectiveness of computerized clinical guidelines in the process of care: a systematic review. *BMC Health Services Research*, 10, 2.

Downing, G., Boyle, S. N., Brinner, K. M. and Osheroff, J. A. (2009). Information management to enable personalized medicine: stakeholder roles in building clinical decision support. *BMC Medical Informatics and Decision Making*, 9, 44.

Drapkin, R., Le Roy, G., Cho, H., Akoulitchev, S. and Reinberg, D. (1996). Human cyclin-dependent kinase-activating kinase exists in three distinct complexes. *Proceedings of the National Academy of Sciences of USA*, 93(13), 6488-93.

Flejter, W. L., McDaniel, L. D., Johns, D., Friedberg, E. C., Schultz, R. A. (1992). Correction of xeroderma pigmentosum complementation group D mutant cell phenotypes by chromosome and gene transfer: involvement of the human ERCC2 DNA repair gene. *Proceedings of the National Academy of Sciences of USA*, 89: 261-265.

Friedberg, E. C. (2001). How nucleotide excision repair protects against cancer. *Nature Reviews Cancer*, 1, 22–33.

Friedberg, E. C., Walker, G. C. and Siede, W. (2005). DNA Repair and Mutagenesis. *American Society of Microbiology Press*, 662–669.

Gehlenborg, N., O'Donoghue, S. I., Baliga, N.S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., Gavin, A. –C. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7, S56-S68.

Goh K. I., Cusick M. E., Valle D., Childs B., Vidal M., Barabási A. L. (2007). The Human Disease Network. *Proceedigns of the National Academy of Sciences of USA*, 104, 8685-8690.

Hoeijmakers, J. H. J. (2001). Genome maintenance mechanisms for preventing cancer. *Nature*, 411, 366–374.

Hyodo, T., Ito, S., Hasegawa, H., Asano, E., Maeda, M., Urano, T., Takahashi, M., Hamaguchi, M., Senga, T. (2012). Misshapen-like kinase 1 (MINK1) is a novel

component of striatin-interacting phosphatase and kinase (STRIPAK) and is required for the completion of cytokinesis. *Journal of Biological Chemistry*, 287, 25019-25029.

Itoh, M., Hayashi, M., Shioda, K., Minagawa, M., Isa, F., Tamagawa, K., Morimatsu, Y., Oda, M. (1999). Neurodegeneration in hereditary nucleotide repair disorders. *Brain & Development*, 21, 326-333.

Jensen L. J., Kuhn M., Stark M., Chaffron S., Creevey C., Muller J., Doerks T., Julien P., Roth A., Simonovic M., Bork P., von Mering C. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37 (suppl 1), D412-6.

Kanao, R., Hanaoka, F. and Masutani, C. (2009). A novel interaction between human DNA polymerase eta and MutLalpha. *Biochemical and Biophysical Research Communications*, 389(1), 40-5.

Kanda, T., Oda, M., Yonezawa, M., Tamagawa, K., Isa, F., Hanakago, R., Tsukagoshi, H. (1990). Peripheral neuropathy in xeroderma pigmentosum. *Brain*, 113: 1025-1044.

Lehninger, A. L.; Cox, M. M.; Nelson, D. L. (2005). *Principles of Biochemistry*. 4. ed. New York: Worth.

Lesk, A. (2008). *Introduction to Bioinformatics. University of Cambridge*. New York: Oxford University Press.

Li, G. M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Research*, 18(1), 85-98.

Li, L., Bales, E. S., Peterson, C. A., Legerski, R. J. (1993). Characterization of molecular defects in xeroderma pigmentosum group C. *Nature Genetics*, 5, 413-417.

Li, L., Lu, X., Peterson, C. and Legerski, R. (1997). XPC interacts with both HHR23B and HHR23A in vivo. *Mutation Research/DNA Repair*, 383(3), 197-203.

Linke, B., Giegerich, R., Goesmann, A. (2011). Conveyor: a workflow engine for bioinformatic analyses. *Bioinformatics*, 27, 903-911.

Misselwitz, B., Strittmatter, G., Periaswamy, B., Schlumberger, M. C., Rout, S., Horvath, P., Kozak, K. and Hardt, W. D. (2010). Enhanced CellClassifier: a multi-class classification tool for microscopy images. *BMC Bioinformatics*, 11, 30.

Naegeli, H. and Sugasawa, K. (2011). The xeroderma pigmentosum pathway: decision tree analysis of DNA quality. *DNA Repair*, 10(7), 673-83.

Natale, V. (2011). A comprehensive description of the severity groups in Cockayne syndrome. *American Journal of Medical Genetics Part A*, 155A, 1081-1095.

Nuzzo, A. and Riva, A. (2009). Genephony: a knowledge management tool for genome-wide research. *BMC Bioinformatics*, 10, 278.

Nuzzo, F., Stefanini, M., Colognola, R., Zei, G., Santachiara, A. S., Lagomarsini, P., Casati, S., Marinoni, S. (1986). Association of two rare hereditary disorders, xeroderma pigmentosum and trichothiodystrophy, in three families from north-east Italy. *International Congress on Human Genetics*, Berlin, 249.

Park, C. H. and Sancar, A. (1994). Formation of a ternary complex by human XPA, ERCC1, and ERCC4(XPF) excision repair proteins *Proceedings of the National Academy of Sciences of USA*, 91(11), 5017-21.

Peng, X., Li, Y., Walters, K. A., Rosenzweig, E. R., Lederer, S. L., Aicher, L. D., Proll, S. and Katze, M. G. 2009. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, 10, 373.

Rageul, J., Frëmin, C., Ezan, F., Baffet, G., Langouët, S. (2011) The knock-down of ERCC1 but not of XPF causes multinucleation. *DNA Repair,* 10, 978-990.

Rahn, J. J., Adair, G. M. and Nairn, R. S. (2010). Multiple roles of ERCC1-XPF in mammalian interstrand crosslink repair. *Environmental and Molecular Mutagenesis*, 51(6), 567-81.

Satokata, I., Tanaka, K., Okada, Y. (1992). Molecular basis of group A xeroderma xeroderma pigmentosum: a missense mutation and two deletions located in a zinc finger consensus sequence of the XPAC gene. *Human Genetics*, 88: 603-607.

Scardoni, G., Petrerlini, M., Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe. *Bioinformatics*, 25, 2857-2859.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research,*13(11):2498-504.

Silva, F. N., Cavalcanti, M. C., Dávila, A. M. R. (2006). In services: data management for in silico workflows. IEEE: Proceedings of the 17th International Conference on *Database and Expert Systems Applications* (DEXA'06), 72, 206 - 210.

Weidenheim, K. M., Dickson, D. W., Rapin, I. (2009) Neuropathology of Cockayne syndrome: Evidence for impaired development, premature aging, and neurodegeneration. *Mechanisms of Ageing and Development*, 130, 619-636.

Weston, A. D., Hood, L. (2004). Systems Biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of Proteome Research*, 3, 179-196.

Wheeler D. L., Barrett T., Benson D. A., Bryant S. H., Canese K., Chetvernin V., Church D. M., DiCuccio M., Edgar R., Federhen S., Geer L. Y., Kapustin Y., Khovayko O., Landsman D., Lipman D. J., Madden T., Maglott D. R., Ostell J., Miller V., Pruitt K. D., Schuler G. D., Sequeira E., Sherry S. T., Sirotkin K., Souvorov A., Starchenko G., Tatusov R. L., Tatusova T. A., Wagner L.and Yaschenko E. (2006). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35 (suppl 1), D5-D12.

White, M. F. (2009). Structure, function and evolution of the XPD family of iron-sulfur-containing 5'-->3' DNA helicases. *Biochemical Society Transactions*, 37, 547-51.

Yoshino, M., Nakatsu, Y., te Riele, H., Hirota, S., Kitamura, Y. and Tanaka, K. (2002). Additive roles of XPA and MSH2 genes in UVB-induced skin tumorigenesis in mice. *DNA Repair*, 1(11), 935-40.

Zöllner, F., Neumann, S., Kummert, F., Sagerer, G. (2005). Database driven test case generation for protein-protein docking. *Bioinformatics*, 21, 683-684.

Capítulo II

NET2HOMOLOGY: A WORKFLOW DESIGNED TO ANALYZE CONSERVED BIOLOGICAL PROCESS BETWEEN TWO DIFFERENT SPECIES USING INTERACTOMIC NETWORKS

DANIEL LUÍS NOTARI[1] and DIEGO BONATTO[2]

[1] Centro de Computação e Tecnologia da Informação e Instituto de Biotecnologia, Universidade de Caxias do Sul, Brasil. http://www.ucs.br

[2] Centro de Biotecnologia do Estado do Rio Grande do Sul, sala 219, Departamento de Biologia Molecular e Biotecnologia, Universidade Federal do Rio Grande do Sul, UFRGS, Avenida Bento Gonçalves, 9500, Prédio 43421, Caixa Postal 15005, Cep 91509-900, Porto Alegre, Rio Grande do Sul, Brasil. http://www.ufrgs.br

Manuscrito em fase de redação que será submetido no formato *short-communication* a um periódico que ainda não foi escolhido.

Net2Homology: A workflow designed to analyze conserved biological process between two different species using interactomic networks

Daniel Luís Notari[1], Diego Bonatto[2]

[1]Centro de Computação e Tecnologia da Informação e Instituto de Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul, Brasil.

[2]Centro de Biotecnologia do Estado do Rio Grande do Sul, Departamento de Biologia Molecular e Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

*** - Address to which correspondence should be sent:**

Diego Bonatto

Centro de Biotecnologia da UFRGS - Sala 219

Departamento de Biologia Molecular e Biotecnologia

Universidade Federal do Rio Grande do Sul - UFRGS

Avenida Bento Gonçalves 9500 - Prédio 43421

Caixa Postal 15005

Porto Alegre – Rio Grande do Sul

BRAZIL

91509-900

Phone: (+55 51) 3308-6080

Fax: (+55 51) 3308-7309

E-mail: diegobonatto@gmail.com

Contract/grant sponsor: CNPQ, Universidade de Caxias do Sul (UCS)

Abstract

Homology can be defined as the most important principle in comparative biology. In this sense, there are several methods to evaluate homology among different species, e.g. the comparison of DNA and amino acid sequences by using softwares like Basic Local Alignment Search Tool (BLAST), among others. An interesting approach is to discovery conserved proteins between two different species by using interactomic network comparison, allowing also to discover conserved pathways and biological processes for a specific network of proteins or DNA-proteins. In order to explore the potential of network comparison, a scientific workflow called Net2Homology was designed to generate a bioinformatic protocol to evaluate the similarity found between two PPI networks from different species, allowing the search of conserved motifs and subgraphs, which are related to conserved biological processes. Net2Homology program were used to obtain protein-protein interaction networks to xeroderma pigmentosum diseases and cicle-oxigenase-1 enzime by *Homo sapiens* and *Mus musculus* cross-species. All networks were compared using NetworkBlast program and the output shows conserved biological process.

Keywords: Cross-species network comparison, conserved proteins, homology, system biology.

1. INTRODUCTION

Homology can be defined as the most important principle in comparative biology (Bock, 1974), indicating an entire hierarchy of biological patterns and processes, from molecular level (e.g., protein structure and functions) to morphological elements (Hall, 1994). In addition, the analysis of sequence homology of a known protein generally offers important data that lead to the discovery of how new sequenced genes and genomes are organized and their potential functions within a cell (Altschul et al., 1990). Moreover, the sequence homology analysis allows the discover of evolutive processes related to a family of proteins, like the presence of orthologous and paralogous members of this family (Bandyopadhyay et al., 2012).

Searching homology within and between genomes is a common technique used to study several aspects of its evolution, such as genomic rearrangements or genome duplication (Simillion et al., 2008). For this purpose, several softwares were developed, like I-ADHoRe, which detects highly degenerated homology relations within and between genomes (Simillion et al., 2008). Another way to discovery homology is to compare sequences using similarity methods among proteins and evaluate measures like distance (Bhadra et al., 2006). BLAST is an algorithm that finds regions of local similarity among sequences by comparison of nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches (Altschul, 1990). Also, BLAST can infer functional and evolutionary relationships between sequences as well as help identify members of gene families (Altschul, 1990).

Many tools use similarity searches to evaluate homology between proteins like CDART (Conserved Domain Architecture Retrieval Tool) performs similarity searches of the NCBI Entrez Protein Database based on domain architecture (Geer et al., 2002), QPath searches the network for homologous pathways (Shlomi et al., 2006), LMNAST (Local Modular Network Alignment Similarity Tool) applies a BLAST-like heuristic to gene order and arrangement (Quan and Bentley, 2011) and AlignNemo uses the concept of semantic similarity applied to Gene Ontology vocabularies (Ciriello et al., 2012).

The search for protein function across species can be used as a strategy for identifying functionally related proteins that supplements sequence-based comparisons with information on conserved protein interactions (Bandyopadhyay et al., 2012). The database and online resource STRING (Szklarczyk et al., 2011) generalizes access to

protein interaction data by integrating known and predicted interactions from a variety of sources and information of structural information including homology models (Szklarczyk et al., 2011).

A third method possible to be used to find out homology is to construct networks (or interactome) from protein-protein interactions (PPI; Jonsson et al., 2006). PPI network can be built by integrating experimental protein interactions data from many species, where the protein interactions were given confidence scores based on their homology (Jonsson et al., 2006).Network comparison is possible using a set of different tools like: (i) APID2NET that provide interactive analysis of PPI networks (Hernandez-Toro et al., 2007), (ii) PathBlast that executes a network alignment and search tool for comparing protein interaction networks (Kelley et al., 2004), (iii) NetworkBlast, which is used to identifying protein complexes in several PPI networks from different species (Kalaev et al., 2008; Sharan et al., 2005), and (iv) AlignNemo, which uncovers subnetworks of proteins that relate in biological function and topology of interactions (Ciriello et al., 2012).

Our approach presents a different way to analyze cross-species homology by using a PPI network from a source specie to search similarity against another target specie by applying NCBI BLAST tool. We use e-value score to identify proteins that are evolutionary conserved, where these proteins were used to get to PPI networks from STRING database for each species. The network comparison is made using NetworkBlast algorithm to look for complexes of evolutionarily conserved across the two networks and apply systems biology tools to evaluate conserved motifs and biological processes (Kalaev et al., 2008; Sharan et al., 2005).

2. MATERIALS AND METHODS

2.1 WORKFLOW DESIGN

Net2Homology was implemented as a scientific workflow (Fig. 1). This workflow can be divided in five major steps: i) protein selection, ii) protein similarity search, iii) query PPI networks, iv) cross-species network comparison, and v) system biology analysis.

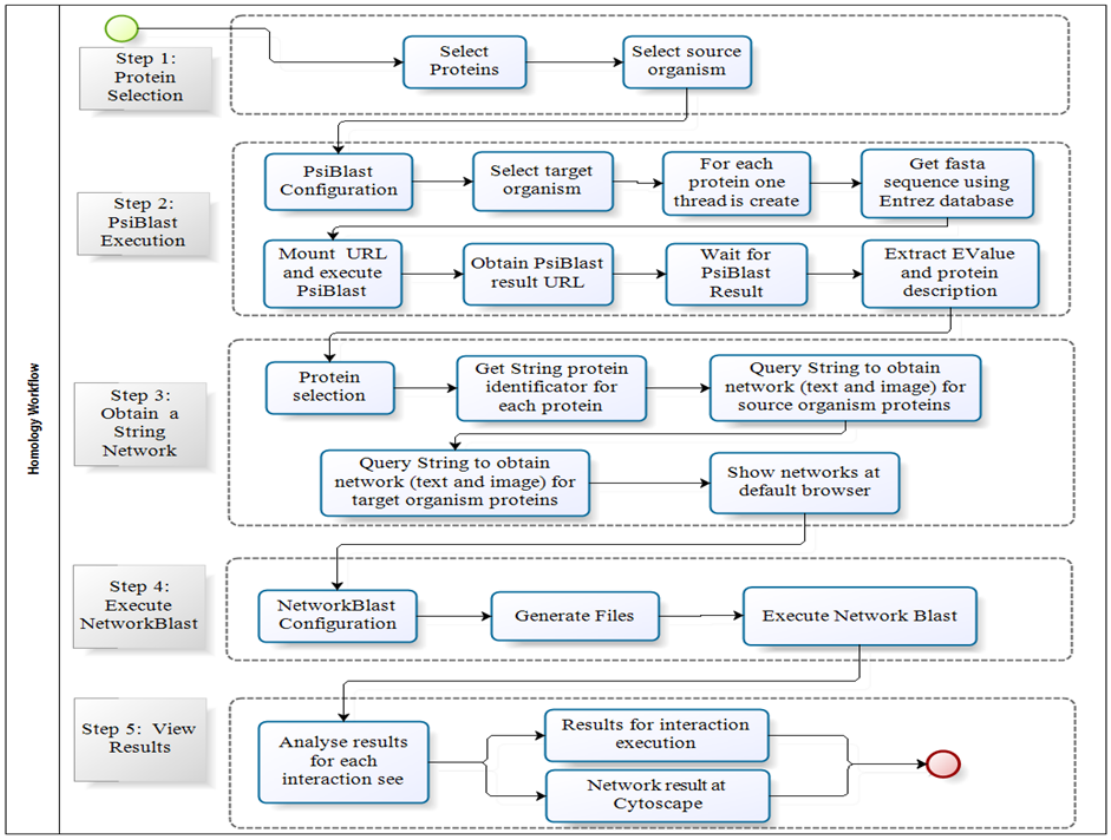


Figure 1: Fluxogram of Net2Homology scientific workflow.

2.2 PROTEIN SELECTION

Net2Homology user must obtain a flat file from a PPI network. This file must be formatted as STRING network file format (Szklarczyk et al., 2011), and its content should have, at least, two columns, separated by space or tab character, with proteins identifiers that represents one interaction to particular specie. After load proteins file, it is necessary to select proteins and define the source specie.

2.3 PROTEIN SIMILARITY SEARCH

Following the workflow, the protein alignment between two species will be executed using Psi-Blast tool. The goal on this step is to obtain proteins distantly related or to find out a new member of a protein family (Gish and States, 1993). Also, it is possible to identify proteins domains (Gish and States, 1993). There are settings necessary to execute this step and the first parameter to define target specie. Others parameters are database, algorithm, threshold and file result format.

For each protein a new computational thread is created to get sequence FASTA (Pearson and Lipman, 1988), configuration URL to execute the alignment using NCBI

ENTREZ tool (Wheeler et al., 2006) and obtain an RID (Request IDentifier) number (McGinnis and Madden, 2004). This number is used to get the alignment result and many attempts are made to get it. After the file result arrives, the last activity is to extract e-value. E-value it the score of expect value and percentage identity matched (McGinnis and Madden, 2004).

2.4 QUERY PPI NETWORKS

All proteins and its e-value are present to the user, and he must select proteins of his interesting. In the third step, proteins STRING identification (Szklarczyk et al., 2011) are, initially, queried from each protein and then, a PPI network was queried from STRING to source and target species using this identifiers.

2.5 CROSS-SPECIES NETWORK COMPARISON

The cross-species comparison is made using two PPI networks and NetworkBlast tool. To use this tool is necessary to configurate some parameters and generate three files (Kalaev et al., 2008). The main parameters are (Sharan et al., 2005): i) density of protein complexes, ii) the rate of false negatives in each network, iii) the sequence similarity threshold, v) the way experiments are categorized for interaction reliability computation, and v) number of interactions. After that, the tool will generate three files (Kalaev, 2008): i) first file contains all sourced proteins edges and the score combined value obtained from STRING PPI network description, ii) the second file contains the same information as first file but using target network, and iii) the last file contains the proteins inputted by the user plus e-value and protein description obtained with Psi-Blast execution.

The NetworkBlast result is two files for each interaction defined where the first file contains the execution report and the second file shows the resultant network with conserved proteins (Kalaev, 2008). This file is automatic opened at Cytoscape tool (Shannon et. al, 2003) for system biology analysis.

2.6 SYSTEM BIOLOGY ANALYSIS

Systems biology enables the study of interactions among the different components of biological entities on a large scale (Weston and Hood, 2004; Barabasi and Oltvai, 2004), and the use of analytical high-throughput methods generates a large amount of information about the behavior of an organism under particular physiological

conditions (Barabasi, 2002). The PPI network resultant can be analyzed to look for evolution motifs (Bailey and Gribskov, 1998; Kuang, 2005) and subgraphs (Kuang, 2005).

3. RESULTS

The PPI networks outputs from Net2Homology program execution for xeroderma pigmentosum type XPD disease and cicle-oxigenase-1 enzyme are shown in Fig. 2. The first experiment uses all proteins from xeroderma pigmentosum *Homo sapiens* PPI network (Fig. 2A) to execute Psi-Blast with a set of parameters (organism *Mus musculus*, database swissprot protein sequences, blastp algorithm and threshold 0.005). After Psi-Blast execution, e-value, protein descriptions and fasta sequence was extract from each protein report. Second and third data were queried on STRING database to obtain proteins names. These proteins queried at STRING database to get a xeroderma pigmentosum *Mus musculus* PPI network (Fig. 2B). After that, NetworkBlast was executed with these two set of proteins and e-values. The PPI network (Fig. 2C) execution result shows conserved protein complex across these two networks.

The second experiment uses whole proteins from cicle-oxigenase-1 enzyme *Homo sapiens* PPI network (Fig. 2D) to execute Psi-Blast with a set of parameters (organism *Mus musculus*, database non-redundant protein sequences, blastp algorithm and threshold 0.005). After Psi-Blast execution, e-value, protein descriptions and fasta sequence was extract from each protein report. Second and third data were queried on STRING database to obtain proteins names. These proteins queried at STRING database to get a cicle-oxigenase-1 enzyme *Mus musculus* PPI network (Fig. 2E). After that, NetworkBlast was executed with these two set of proteins and e-values. The PPI network (Fig. 2F) execution result shows conserved protein complex across these two networks.

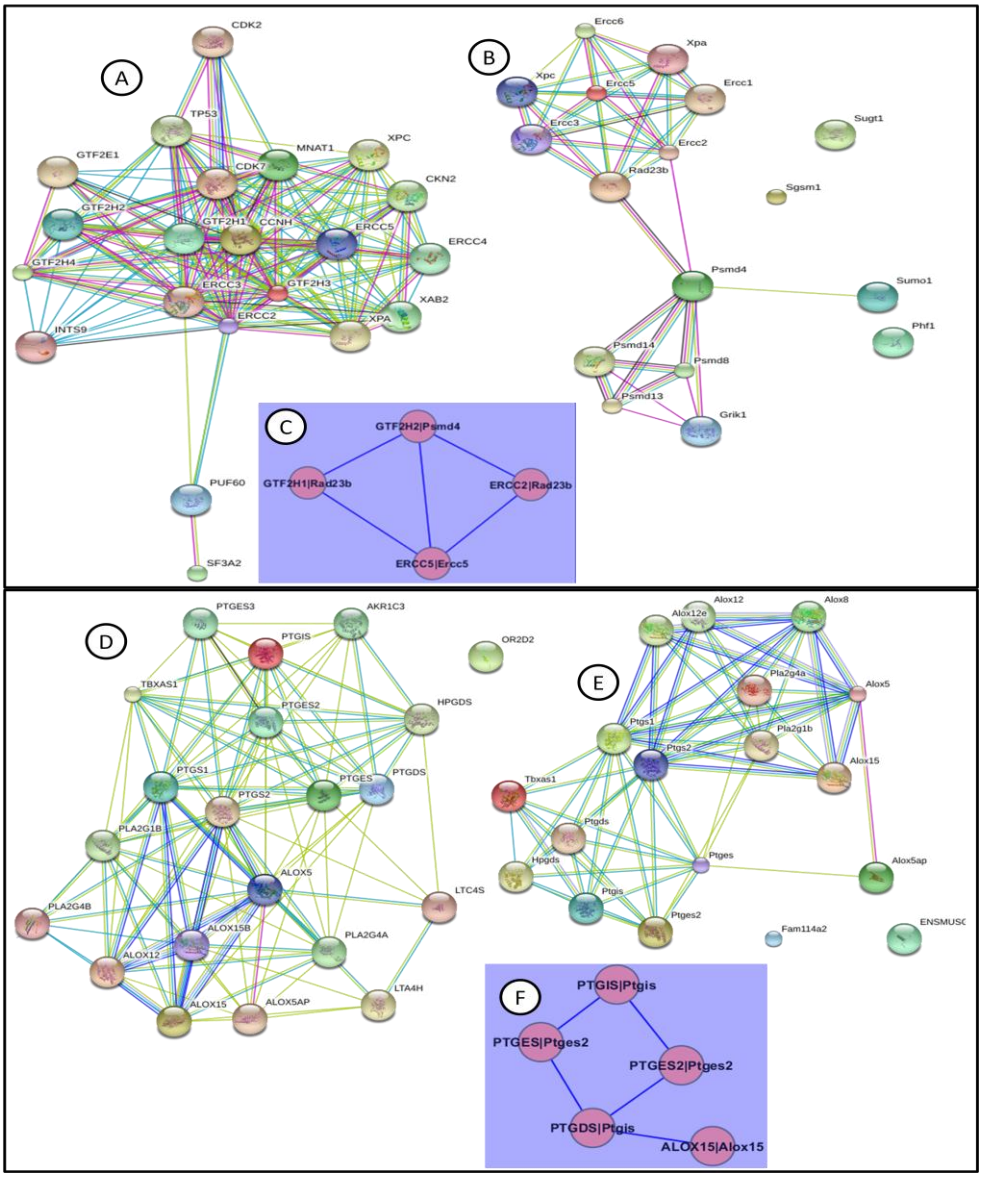


Figure 2: In the first experiment, xeroderma pigmentosum *Homo sapiens* PPI network (A) used as input to execute Psi-Blast, and xeroderma pigmentosum *Mus musculus* PPI network (B) was obtained after extract information from Psi-Blast and query STRING database. NetworkBlast had runned with (A) and (B) networks, and generates the PPI network resultant (C) that shows conserved protein complex across these two networks. The second experiment uses cicle-oxigenase-1 enzyme *Homo sapiens* PPI network (D) as input to execute Psi-Blast, and cicle-oxigenase-1 enzyme *Mus musculus* PPI network (E) was obtained after extract information from Psi-Blast and query STRING database. NetworkBlast had runned with (D) and (E) networks, and generates the PPI network resultant (F) that shows conserved protein complex across these two networks.

4. DISCUSSION

Net2Homology was implemented as a desktop tool using java programming language and a scientific workflow was used to integrate different tools. This workflow allows executing the steps using just one program. SWISS-MODEL (Arnold et al., 2006) uses a workflow to integrate some tools, programs and databases to aid in the process to homology protein modeling.

Net2Homology generates a random PPI networks from species defined because the user's selection made on step three and data inputted. Using the same idea, Jonsson, et al. (2006) constructed PPI networks using the rat proteome used for cancer studies, and also APID2NET allows to create a PPI network from data by querying APID server (Hernandez-Toro et al., 2007).

Net2Homology uses e-value score supplied by PSI-BLAST execution to select proteins because the percentage identity matched (McGinnis and Madden, 2004) and to obtain proteins distantly related or to find out new member to a protein family (Gish and States, 1993). On the other hand, BLAT (The BLAST-Like Alignment Tool) uses measures like mismatch schemes, number of required index matches (hits), and extends these into high-scoring pairs (HSPs) for analyzing vertebrate genomes and cross-species protein alignment (Kent, 2002). Another measure useful is p-values used in the context of sequence homology searches to present a simple and efficient algorithm for computing the distribution of this statistic (Bailey and Gribskov, 1997).

Net2Homology tool uses NCBI protein databases to align sequences to look for evoluted proteins complex crossing two species. i-ADHoRe tool first identifies pairs of homologous segments from a dataset containing one or more genomes, after that aligning these segment pairs such that homologous genes are placed in the same column (Simillion et al., 2008).

Net2Homology executes a PSI-BLAST once for each inputed proteins and use e-value scores to select proteins to get a PPI network. Sometimes, it is not always possible to identify distantly related protein domains by sequence search techniques, because intermediate sequences possess features of more than one protein and facilitate detection of remotely related proteins (Bhadra et al., 2006). A approach to test this scenarios was developed to execute PSI-BLAST for many generations, that establishes relationships by cascading an entirely sequence analysis based method where

intermediate sequences serve as links and bridge the sequence gap between proteins (Bhadra et al., 2006).

Most homology searches involve comparing a single sequence to a database of known sequences (Bailey and Gribskov, 1997). This would often bring some inconsistence score because to identify very distant homologs using sequence similarly is limited and confined to the most important portions of the molecule (Bailey and Gribskov, 1997). When two unrelated sequences are compared, numerous chances for apparent matches arise, causing a severe noise problem in homology searches of large databases (Bailey and Gribskov, 1997). Besides that, cross-species analysis allows coping with the high levels of noise that are typical to these data (Kalaev et al., 2008). This noise can be reduced by searching patterns named motifs. Motifs describe the key, defining portions of a family of molecules (Bailey and Gribskov, 1997). By comparing networks drawn from different species it is possible to reduce measurement noise and to reinforce the common signal present in the networks (Sharan et al., 2005).

Net2Homology result is a PPI network that can be shown at Cytoscape tool and this network can be analyzed using system biology Cytoscape plugins. For example, APID2NET is a Cytoscape tool to provide interactive analysis of PPI networks like genetic ontology terms, experimental methods that validate each interaction, find out concurrent functional and structural attributes along all protein pairs in a network and analyze hubs, annotations and homology (Hernandez-Toro et al., 2007).

CONCLUSIONS

Net2Homology program were used to obtain protein-protein interaction networks to xeroderma pigmentosum diseases and cicle-oxigenase-1 enzime by *Homo sapiens* and *Mus musculus* cross-species. These PPI networks were compared using NetworkBlast program and the output shows conserved biological process.

Acknowledgments

I want to thank my colleague Alexandre Nascimento to help on the java development of this software.

References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215 (3), 403–410.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling. Bioinformatics, 22 (2), 195-201.

Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics, 14(1), 48-54.

Bandyopadhyay, S., Sharan, R. and Ideker, T. (2012). Systematic identification of functional orthologs based on protein network comparison. Genome Research, 16, 428-435.

Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 5, 101-113.

Barabasi, A. L. (2002). Statistical physics of complex networks Reviews of Modern Physics, 74, 47.

Bhadra, R., Sandhya, S., Abhinandan, K. R., Chakrabarti, S., Sowdhamini, R. and Srinivasan, N. (2006). Cascade PSI-BLAST web server: a remote homology search tool for relating protein domains. Nucleic Acids Research, 34, W143–146.

Bock, W. J. (1974). Philosophical foundations of classical evolutionary classification. Systematic Zoology, 22, 375–392.

Ciriello, G., Mina, M., Guzzi, P. H., Cannataro, M. and Guerra, C. (2012). AlignNemo: A Local Network Alignment Method to Integrate Homology and Topology. PLoS ONE, 7 (6), e38107.

Geer, L. Y., Domrachev, M., Lipman, D. J. et al. (2002). CDART: Protein Homology by Domain Architecture. Genome Research, 12, 1619-1623.

Gish, W. and States, D.J. (1993). Identification of protein coding regions by database similarity search. Nature Genetics, 3, 266-272.

Hall, B. K. (1994). The Hierarchical Basis of Comparative Biology. Homology. San Diego: Academic Press, 229–247.

Hernandez-Toro, J., Prieto, C. and De Las Rivas, J. (2007). APID2NET: unified interactome graphic analyzer. Bioinformatics, 23 (18), 2495–2497.

Jonsson, P. F., Cavanna, T., Zicha, D. and Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics, 7, 2.

Kalaev, M., Smoot, M., Ideker, T. and Sharan, R. (2008). NetworkBLAST: comparative analysis of protein networks. Bioinformatics Applications Note, 24, 594–596.

Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. R. and Ideker T. (2004). PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Research, 1 (32), W83-88.

Kent, J. W. (2002). BLAT--The BLAST-Like Alignment Tool. Genome Resource, 12, 656-664.

Kuang, R., Weston, J. Noble, W. S. and Leslie, C. (2005). Motif-based protein ranking by network propagation Bioinformatics, 21(19), 3711-3718.

McGinnis, S. and Madden. T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research, 32 (2), W20-W25.

Quan, D. N. and Bentley, W. E. (2011). Gene Network Homology in Prokaryotes Using a Similarity Search Approach: Queries of Quorum Sensing Signal Transduction. PLoS Computational Biology, 8 (8), e1002637.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13 (11), 2498-504.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. Proceedings of the National Academy of Sciences, 102, 1974-1979.

Shlomi, T., Segal, D., Ruppin, E. and Sharan, R. (2006). QPath: a method for querying pathways in a protein-protein interaction network. BMC Bioinformatics 7, 199.

Simillion, C., Janssens, K., Sterck, L. and Van de Peer, Y. (2008). i-ADHoRe 2.0: An improved tool to detect degenerated genomic homology using genomic profiles. Bioinformatics, 24 (1), 127-128.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J. and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research, 39 (1), D561-D568.

Wheeler D. L., Barrett T., Benson D. A., Bryant S. H., et al. (2006). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 35 (1), D5-D12.

Weston, A. D. and Hood, L. (2004). Systems Biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. Proteome Research, 3, 179- 196.

6. DISCUSSÃO GERAL

O desenvolvimento dos *workflows* científicos Dis2PPI e Net2Homology envolveu a criação de ferramentas para cada etapa do *workflow* (apresentadas no Anexo 3). Inicialmente, foram criadas ferramentas para recuperar os dados sobre redes PPI dos bancos de dados STRING (ver seção A3.2.1) e PDB (ver seção A3.2.2) com o intuito de compreender o funcionamento das API (*Application Program Interface*) de cada banco de dados. A API[57] do banco de dados STRING fornece as informações para montagem de uma chamada URL. O acesso ao banco de dados PDB[58] é feito através do uso de *web services* para a busca de arquivos de imagens (formato PNG de coordenadas cartesianas) proteínas. Associado a isto, foi acessado, via *web service*, o banco de dados de doenças gênicas do OMIM usando as ferramentas do NCBI Entrez (ver seção 3.3.2). Estes três programas foram usados para o desenvolvimento das etapas do *workflow* científico Dis2PPI.

Neste sentido, o desenvolvimento do Dis2PPI centrou-se em utilizar apenas uma linguagem de programação (Java) visando a integração entre as ferramentas desenvolvidas e os *workflows*, optou-se por unificar tudo em uma aplicação *desktop*. Alguns dos fatores que contribuíram para a escolha da linguagem Java é que esta tem a possibilidade de fazer chamadas remotas usando uma URL (Goodrich e Tamassia, 2007), bem como, possui uma biblioteca para acessar *web services* (Goodrich e Tamassia, 2007) desenvolvidos com a tecnologia Apache Axis[59], como é o caso do *web service* disponibilizados pela ferramenta E-Utilities do NCBI (Sayers and Wheeler, 2004). Outro fator que ajudou na escolha é o fato do programa Cytoscape também ter sido desenvolvido na linguagem Java (Shannon et al., 2003), uma vez que a primeira

[57] http://string.embl.de/help/topic/org.string-db.docs/api.html

[58] http://www.rcsb.org/pdb/software/rest.do

[59] http://ws.apache.org/axis/

versão das ferramentas desenvolvidas foi implementada na forma de um *plugin* para o Cystocape. Aliado a isto, as ferramentas para geração de *workflows* científicos (Taverna, Kepler e Vistrails) fornecem suporte a diversas linguagens de programação, tais como, Java, Phyton e Perl (Oinn et al., 2004; Altintas et. al, 2006, Davidson et al., 2007). As linguagens de programação Phyton e Perl são usadas para construir aplicações robustas, como as apresentadas neste trabalho. No entanto, estas linguagens são mais usadas para a construção de aplicações *web* e *scripts*[60][61]. A linguagem de programação Java pode ser usada para criar programas para *web*, *desktop* ou *scripts* (Goodrich e Tamassia, 2007). Para o desenvolvimento das ferramentas criadas optou-se por um programa *desktop*, bem como decidiu-se por armazenar os arquivos gerados durante a execução de cada etapa de um dos *workflows*. Aliado a necessidade de realizar chamadas URL e chamadas via *web services*, optou-se pela linguagem Java.

Os *workflows* desenvolvidos armazenam os arquivos gerados para cada execução do Psi-Blast, para cada sequência de aminoácidos consultada, para cada rede consultada no STRING (tanto os arquivos textos como as imagens) e para armazenar as informações a respeito da execução do programa NetworkBlast. Desta forma, o usuário possui todas as informações, geradas por cada etapa do *workflow*, salvas em seu computador, onde o mesmo procedimento é feito quando se usa ferramentas específicas para a geração de *workflows* científicos, tais como Kepler, Vistrails e Taverna (Oinn et al., 2004; Altintas et al., 2006; Davidson et al., 2007; Wang et al., 2012a). Estas ferramentas fornecem a funcionalidade de salvar os resultados intermediários visando que eles sejam usados em futuras execuções do *workflow* (Oinn et al., 2004; Altintas et. al, 2006, Davidson et al., 2007; Wang et al., 2012a). O programa Vistrails, além de

[60] Um *script* é um programa de computador executado em ambientes de linha de comando, tais como, o *shell* do sistema operacional Linux ou o *command.com* do sistema operacional Windows.

[61] As descrições das linguagens foram retiradas dos *sites* oficiais: http://www.python.org/about/ e http://www.perl.org/about.html. Acessados em 03/12/2012.

salvar as informações das execuções, efetua o armazenamento de todas as alterações feitas no *workflow* (Davidson et al., 2007).

Para melhorar o desempenho dos *workflows,* foi desenvolvido um banco de dados local para as informações de proteínas. Assim, cada vez que uma consulta nova é realizada, os dados obtidos são mantidos e atualizados neste banco de dados local. Quando uma consulta é refeita, a ferramenta procura por estes dados no banco de dados local, os quais são mostrados ao usuário sem a necessidade de uma nova consulta. As informações são salvas em um arquivo chamado *proteins.dat* localizado no diretório temporário do sistema operacional, o qual contém o nome da proteína, o identificador da proteína utilizado pelo banco de dados STRING, o identificador da proteína utilizado pelos bancos de dados e ferramentas do NCBI, a sequência de aminoácidos no formato FASTA de cada proteína e o nome do organismo relacionado a esta proteína.

Para o desenvolvimento do *workflow* Net2Homology foi inicialmente criada uma ferramenta na linguagem Java para a execução de alinhamento de sequências de proteínas usando o programa NCBI Psi-Blast (ver seção A3.2.3). Esta ferramenta utiliza o NCBI Entrez para a montagem da URL de conexão, bem como faz acesso ao banco de dados STRING para buscar as redes PPI, posteriormente utilizadas para a análise de homologia com os *softwares* PathBlast ou NetworkBlast.

O programa PathBlast permite alinhar e comparar duas redes de interação de proteínas entre espécies diferentes visando identificar vias e complexos metabólicos conservados ao longo de sua evolução (Kelley et al., 2004). O uso deste software é limitado à comparação de cinco proteínas através da sua versão web[62]. Outra opção de uso foi a integração do código fonte ao *workflow* Net2Homology, porém não foi obtido sucesso por causa da necessidade de criação de uma série de arquivos de configuração e

[62] http://www.pathblast.org/

de dados para a execução local deste programa. Assim, optou-se por criar mais uma etapa no *workflow* Net2Homology com o intuito do usuário testar a versão *web* limitada a cinco proteínas e definir os organismos a serem utilizados, além de pesquisar a sequência FASTA para cada proteína.

Por outro lado, o programa NetworkBlast, cujo o código-fonte foi incorporado ao *workflow* Net2Homology, permite identificar complexos de proteínas em várias redes de interação de proteínas para diferentes espécies (Kalaev, 2008; Sharan, 2005). O seu resultado final é uma rede PPI que apresenta nós evolutivamente conservados para duas redes de proteínas (Sharan, 2005; Kalaev, 2008).

Por fim, gerou-se uma ferramenta para carregar os dados de um arquivo de uma rede PPI a ser utilizada para iniciar o programa Net2Homology (ver seção A3.1). Durante a execução deste, são realizadas várias consultas de alinhamento de sequências, o que pode tornar o desempenho do *workflow* muito lento. Assim, para que o programa Net2Homology possa utilizar o algoritmo Psi-Blast, é necessário fazer duas chamadas URL ao *site* do NCBI, onde na primeira chamada devem ser fornecidos todos os dados e parâmetros da consulta, obtendo como resultado um número de protocolo. Este número é utilizado, posteriormente, para se buscar o resultado da consulta, sendo esta a segunda chamada feita ao NCBI. Infelizmente, torna-se necessário realizar várias chamadas ao *site* para obter o resultado, gerando penalidades em termos de tempo de computação. Para resolver este problema foi utilizada a técnica computacional de *threads*, onde uma *thread* é uma forma de dividir o processamento computacional (Tanembaum, 2010), ou seja, normalmente quando qualquer programa está sendo executado, somente uma *thread* está sendo executada. Desta forma, ao invés de executar apenas uma consulta ao *site* do NCBI Psi-BLAST por vez, o uso de *threads* permite que várias consultas sejam executadas simultaneamente (Tanembaum, 2010). Uma *thread*

foi criada para cada proteína sendo responsável por montar a chamada URL para executar o Psi-Blast e aguardar a resposta do *site*. Desta forma, o tempo de resposta diminui consideravelmente. A ideia de paralelizar a execução do acesso ao *site* do BLAST também foi desenvolvida na ferramenta de geração de *workflows* científicos Kepler (Wang et al., 2012a).

Para reaproveitar a rede PPI gerada pelo *workflow* Dis2PPI foi feita uma integração com o *workflow* Net2Homology (ver seção A3.3) que, então, utiliza esta rede como dados de entrada. Assim, o usuário tem a possibilidade de fazer uma prospecção de dados, visando pesquisar os processos biológicos conservados através da comparação de duas redes PPI obtidas com o *workflow* Net2Homology.

Um ponto importante do *workflow* Net2Homology é a extração das informações do relatório do Psi-Blast visando obter as redes PPI no banco de dados STRING. Um algoritmo foi desenvolvido (ver seção A3.4) para encontrar a proteína correspondente a uma sequência FASTA ou com a descrição apresentada no alinhamento. Este algoritmo utiliza três formas para encontrar a proteína-alvo usando estas informações iniciais. A primeira forma utiliza a sequência de aminoácidos para comparar com um banco de dados local de proteínas. Este banco de dados foi criado a partir da extração das sequências de proteínas para os organismos *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae* fornecidas pelo banco de dados STRING[63]. Sempre que uma busca é realizada com sucesso, o programa retorna o código de identificação da proteína utilizada pelo banco de dados STRING, sendo necessária uma nova pesquisa ao banco de dados STRING, para buscar o nome da proteína. Se a proteína não for encontrada, passa-se a utilizar a descrição das proteínas. Para tanto, dois procedimentos são realizados. O primeiro tenta identificar se há uma palavra que pode ser uma proteína

[63] Arquivo com a sequência FASTA retirado do site http://string.embl.de/newstring_download/protein.sequences.v9.0.fa.gz em 05/11/2012.

através do uso da expressão regular "([A-Z]{1}[A-Z0-9]{1,6})", onde o nome de uma proteína tem que começar por uma letra maiúscula e pode ter mais seis caracteres entre letras e números. Quando uma palavra candidata é encontrada, esta é testada junto ao banco de dados STRING para verificar se é uma proteína válida para o organismo pesquisado. Caso a resposta seja positiva, o programa retorna esta proteína Por outro lado, se a resposta é negativa, o programa continua testando a existência de mais palavras candidatas. Se não for encontrada nenhuma proteína com este procedimento, o programa passa a testar a descrição completa da proteína (as três primeiras descrições do relatório são usadas) com uma consulta ao banco de dados STRING. Se obtiver sucesso, o nome da proteína é retornada. Se nenhuma pesquisa obtiver sucesso, a proteína do organismo original é retornada como resposta. Neste caso, o programa não conseguiu encontrar uma proteína equivalente no cruzamento de espécies realizado com o Psi-Blast. Esta proteína não será levada em consideração no teste a ser realizado pelo NetworkBlast.

Em uma pesquisa feita na literatura, não foi encontrado nenhuma referência ao uso da descrição das proteínas extraídas do relatório do Psi-Blast para classificar a proteína resultante. Por outro lado, a utilização das sequências de aminoácidos das proteínas é frequentemente utilizada. Kuang et al. (2005) usou a sequência de aminoácidos para a detecção de homologia e extração de motivos. Outros trabalhos usam métodos de inteligência artificial tais como SVM (*Support Vector Machine*) e HMM (*Hidden Markov Model*) para obter a classificação das proteínas. Nanni et al. (2012) usaram as sequências de aminoácidos como dados de entrada para a matriz de substituição PSSM (*Position Specific Scoring Matrix*) e, este resultado foi usado para treinar o SVM para classificar as proteínas. Zhang et al. (2009) também usaram as sequências de aminoácidos como dados de treinamento para o SVM com o intuito de

identificar regiões homólogas remotas. Yalamanchili et al. (2012) propuseram um algoritmo neural que combina as sequências de aminoácidos como dados de entrada para um processamento usando HMM e SVM, o resultado obtido é usado para pesquisar os termos de ontologia gênica para obter a classificação das proteínas.

Durante os testes com os *workflows* científicos percebeu-se a necessidade de se gerar um mecanismo para interromper a execução do *workflow* e, posteriormente, retomar a execução a partir do ponto de parada. Tanto a parada quanto a retomada da execução são de grande utilidade para o processo de desenvolvimento e testes do *workflow*, bem como para averiguar o conjunto de dados de entrada (Freire et al., 2006; Davidson et al., 2007; Silva et al., 2010; Chirigati and Freire, 2012).

Por outro lado, a proveniência dos dados tem por objetivo salvar todas as informações utilizadas durante uma execução de um *workflow* científico (Freire et al., 2006; Davidson et al., 2007; Silva et al., 2010; Chirigati and Freire, 2012), sendo que esta documentação visa preservar e determinar a qualidade dos dados tanto quanto a interpretação e reprodução (parcial ou total) de um experimento *in silico* (Silva et al., 2010). A proveniência dos dados é representada como um conjunto de dependências entre objetos de dados (Freire et al., 2006; Davidson et al., 2007; Silva et al., 2010; Chirigati and Freire, 2012). Usando o *workflow* Net2Homology, os resultados da execução do Psi-Blast para cada proteína é uma dependência para obtenção das redes PPI do banco de dados STRING. As dependências entre os dados de objetos representam uma ordenação total de eventos (Freire et al., 2006; Davidson et al., 2007; Silva et al., 2010; Chirigati and Freire, 2012), o que pode ser facilmente percebido nas figuras que descrevem os *workflows* científicos apresentados nos Capítulos I e II. Uma das técnicas mais comuns de armazenar a proveniência dos dados em *workflows* científicos é capturar esta informação implícita através de um *log* (Davidson et al.,

2007). Um *log* é uma forma de registrar todas as operações feitas em um *software* (Sommerville, 2007). Os *workflows* Dis2PPI e Net2Homology possuem um *log* desenvolvido (ver seção A3.5), porém o uso deste não é suficiente para resolver o problema de parar e retomar a execução de um *workflow*. Para que isto aconteça, é necessário o desenvolvimento de métodos para salvar e recuperar os dados nos pontos de parada do *workflow*. Analisando o *workflow* Dis2PPI identificou-se um ponto de parada quando as proteínas candidatas são extraídas do relatório da doença consultada no OMIM e mostradas ao usuário na tela. A proveniência envolve salvar o relatório de uma doença do OMIM para que o usuário possa interromper o *workflow* neste ponto. Quando o usuário quiser reiniciar o *workflow*, o mesmo deverá selecionar um dos relatórios de doenças salvas. O *workflow* irá recomeçar no ponto de extração das palavras candidatas a proteínas. Desta maneira, é possível fazer apenas uma busca ao *site* do OMIM permitindo que o usuário possa fazer várias combinações com as proteínas para obter diferentes redes PPI.

No *workflow* Net2Homology três pontos de proveniência dos dados foram identificados. O primeiro refere-se a salvar os resultados das execuções do Psi-Blast e desenvolver um mecanismo de proveniência para que o usuário possa recuperar alinhamentos de proteínas já feitos. O *workflow* é retomado no momento de extrair as proteínas. Desta forma, o usuário poderá combinar diversas proteínas para obter diferentes redes PPI para análise de homologia. O segundo ponto é desenvolver um mecanismo de proveniência dos dados para salvar as informações das proteínas consultas no banco de dados STRING em um SGBDR (Sistema Gerenciador de Banco de Dados Relacional) (Chirigati e Freire, 2012). O banco de dados local utilizado pelos *workflows* terá as suas informações salvas em um SGBDR. Assim, a consulta é feita somente uma vez no banco de dados STRING, com isto, o número de acessos a este *site*

irá diminuir consideravelmente, aumentando o desempenho da ferramenta e diminuindo o tempo de espera do usuário para que uma etapa do *workflow* seja concluída O terceiro passo envolve armazenar os arquivos textos (ou de imagens) no computador do usuário. Neste caso, é necessário o desenvolvimento de um mecanismo para que o usuário selecione a rede PPI salva e carregue a mesma para a ferramenta. O *workflow* recomeça com o usuário acessando diretamente a etapa de análise de homologia.

Uma evolução para os *workflows* e ferramentas desenvolvidos é a criação de uma versão *web*, que tem como base as seguintes considerações: i) permitir o acesso aos *workflows* através de um navegador *web*, evitando que o usuário tenha que procurar e instalar bibliotecas necessárias para executar acesso às ferramentas mencionadas acima. Além disto, isto evita problemas com licenciamento de *software;* ii) desenvolvimento de uma *interface* visual mais interativa e mais amigável; iii) desenvolvimento de uma nova funcionalidade para proveniência dos dados; iv) utilização de um SBGDR para salvar e recuperar as informações visando diminuir o número de acessos a *web* e, para substituir o banco de dados local de proteínas; v) desenvolvimento de um acesso simultâneo ao banco de dados STRING como foi desenvolvido para acessar o BLAST; vi) evitar problemas com as versões das bibliotecas utilizadas e versões da linguagem Java; e vii) analisar a possibilidade de inclusão de um procedimento não automatizado no *workflow* Net2Homology. O usuário poderia utilizar a sua experiência para analisar o resultado do algoritmo de extração dos resultados do Psi-Blast, tendo possibilidade de alterar o vínculo encontrado entre as proteínas.

7. CONCLUSÕES

7.1 WORKFLOW DIS2PPI

1. O algoritmo Dis2PPI mostrou-se eficaz para a geração de redes interatomicas associadas a doenças monogênicas;
2. O programa desenvolvido com o algoritmo Dis2PPI permite o seu uso em diferentes sistemas operacionais e incorpora uma série de rotinas para averiguar a qualidade dos dados;
3. O programa Dis2PPI foi eficiente ao permitir aos seus usuários minerar associações entre uma doença e as suas proteínas correlatas para diferentes espécies e gerar as respectivas redes de interatomas;
4. As redes de interatomas geradas com o programa Dis2PPI apresentaram resultados positivos com a análise feita usando a biologia de sistemas para a análise de agrupamentos e de centralidade;
5. O programa Dis2PPI permitiu avaliar as vias biológicas e complexas de proteínas que podem afetar uma doença genética específica ou condição fisiológica, tal como vista para as doenças xeroderma pigmentosa e síndrome de Cockayne.
6. Os resultados demonstrados pelo programa Dis2PPI permitem aos seus usuários (pesquisadores) desenharem experimentos de validação em bancada.

7.2 WORKFLOW NET2HOMOLOGY

1. O algoritmo Net2Homology mostrou-se eficaz para a geração de redes interatomicas associadas com a comparação de proteínas de diferentes espécies;
2. O programa desenvolvido com o algoritmo Net2Homology permite o seu uso em diferentes sistemas operacionais e incorpora uma série de rotinas para averiguar a qualidade dos dados;

3. O programa Net2Homology foi eficiente, ao permitir aos seus usuários, realizar o alinhamento de sequências usando o método de similaridade para extrair as proteínas associadas a diferentes espécies;
4. O programa Net2Homology permitiu encontrar processos biológicos conservados para a doença xeroderma pigmentosa e para a enzima ciclo-oxigenase-1 em dois modelos biológicos diferentes.

7.3 FERRAMENTAS DESENVOLVIDAS

1. O desenvolvimento da ferramenta para executar o programa NCBI Psi-BLAST permite, ao pesquisador, realizar o alinhamento de sequências para uma proteína com a configuração de diferentes parâmetros. Esta ferramenta é útil para selecionar proteínas a serem utilizadas no programa Net2Homology;
2. O desenvolvimento da ferramenta para pesquisar informações sobre proteínas usando o programa de metabusca STRING permite, ao pesquisador, obter redes de interação de proteínas para diferentes organismos, bem como obter a sequência de aminoácidos de uma proteína no formato FASTA. As redes de interação de proteínas obtidas podem ser usadas como dados de entrada para o *workflow* Net2Homology, bem como, a sequência de aminoácidos pode ser utilizada como dados de entrada para executar o programa NCBI Psi-BLAST.
3. As informações de todos os alinhamentos de sequências feitos com o programa Psi-BLAST e todas as redes de interação de proteínas obtidas com o STRING podem ser recuperadas no computador do pesquisador.
4. As informações assim salvas podem ser utilizadas para validar as informações mostradas pelos *workflows* e redes de interação entre proteínas podem ser usadas como dados de entrada no programa Cytoscape para a realização de análises de biologia de sistemas.

8. PERSPECTIVAS

8.1 PROVENIÊNCIA DOS DADOS

Especialmente para o *workflow* Net2Homology é pertinente à ideia de poder retomar o *workflow* de um ponto já executado ou poder interromper e retomar a execução. Isto porque este *workflow* oferece um conjunto variado de parâmetros que podem ser combinados. Além disto, o usuário pode fazer várias combinações com as proteínas para obter diferentes redes PPI para posterior análise.

8.2 NOVA VERSÃO WEB

A ideia de uma nova versão *web* para as ferramentas desenvolvidas visa permitir o acesso em qualquer ponto da Internet, bem como evitar que o usuário tenha problemas com a instalação e configuração de bilbiotecas. Mas, principalmente, a grande vantagem seria criar uma interface nova e mais amigável, bem como desenvolver os mecanismos de proveniência dos dados para *workflows* científicos.

ANEXOS

ANEXO 1 – TOXICOLOGICAL EFFECTS OF THE MAIN CARCINOGENIC SUBSTANCES IN TOBACCO SMOKE ON HUMAN EMBRYONIC DEVELOPMENT

ANEXO 2 - PORTAL INTERGENICDB

ANEXO 3 - FERRAMENTAS DE CONSULTA A DADOS BIOLÓGICOS

ANEXO 1

TOXICOLOGICAL EFFECTS OF THE MAIN CARCINOGENIC SUBSTANCES IN TOBACCO SMOKE ON HUMAN EMBRYONIC DEVELOPMENT

Toxicological effects of the main carcinogenic substances in tobacco smoke on human embryonic development

Bruno César Feltes[1], Joice de Faria Poloni[2], Daniel Luis Notari[3] and Diego Bonatto[1*]

[1]Biotechnology Center of the Federal University of Rio Grande do Sul, Department of Molecular Biology and Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS – Brazil.

[2]Institute of Biotechnology, University of Caxias do Sul, Caxias do Sul, RS – Brazil.

[3]Computational and Information Technology Center, Universidade de Caxias do Sul, Caxias do Sul, RS – Brazil.

Short title: The effects of cigarette smoke on human embryonic development

[*] **To whom correspondence should be sent:**

Diego Bonatto
Centro de Biotecnologia da UFRGS - Sala 219
Departamento de Biologia Molecular e Biotecnologia
Universidade Federal do Rio Grande do Sul - UFRGS
Avenida Bento Gonçalves 9500 - Prédio 43421
Caixa Postal 15005
Porto Alegre – Rio Grande do Sul
BRAZIL
91509-900
Phone: (+55 51) 3308-6080
Fax: (+55 51) 3308-7309

Contract/grant sponsor: CNPq, FAPERGS, CAPES

Abstract:

The physiological and molecular effects of tobacco smoke in adult humans and the development of cancer have been well described. In contrast, how tobacco smoke affects embryonic development remains poorly described. Morphological studies of the fetuses of smoking pregnant women have shown various physical deformities induced by constant fetal exposure to tobacco components, especially nicotine. In addition, nicotine exposure decreases fetal body weight and bone/cartilage growth in addition to decreasing cranial diameter and tibia length. Unfortunately, the molecular pathways leading to these morphological anomalies are not completely understood. In this study, we applied interactome data mining tools and small compound interaction networks to elucidate possible molecular pathways associated with the effects of tobacco smoke components during embryonic development in pregnant female smokers. Our analysis showed a relationship between nicotine and 48 additional harmful substances involved in a variety of biological process that can cause abnormal proliferation, impaired cell differentiation, increased oxidative stress, and disturbed steroid synthesis. We also describe how nicotine can negatively affect retinoic acid signaling and cell differentiation through inhibition of retinoic acid receptors. In addition, nicotine causes a stress reaction and/or a pro-inflammatory response that inhibits the agonistic action of retinoic acid. Moreover, we show that the effect of cigarette smoke on the developing fetus could represent systemic and aggressive impact in the short term, causing malformations during certain stages of development. Our work provides the first approach to describing how different cigarette constituents affect a broad range of biological process in human embryonic development.

Key words: Nicotine, Retinoic Acid, Cell Differentiation, Cigarette Smoke, Embryonic Development, Systems Chemo-Biology

Introduction

There are more than 4,800 compounds present in the particulate between particle and vapor phases of cigarette smoke [1], and many of these compounds are considered to represent a human health risk [2]. Known constituents of cigarette smoke include isoprene, butadiene, polycyclic aromatic hydrocarbons (PAHs), aldehydes, metals, *N*-nitrosamines, and aromatic amines, in addition to many others [1]. Although extensive anti-tobacco public advertisements promote smoke cessation in pregnant women, a considerable number of women still smoke during their pregnancies and/or are exposed to tobacco smoke via second-hand smoke [3 to 5].

We addressed two major issues in this work. Although prenatal smoke has been previously associate with innumerable malformations during fetus growth and development and disruptions in reproductive physiology there gaps in the knowledge of how tobacco components (TCs) affect the developing embryo in pregnant women in a systemic way [3, 6, 7]. This knowledge gap is the first issue that we address. Interestingly, these abnormalities are not tissue specific or related to any unique pathway, but, rather, are systemic and connected to a broad range of birth defects [3, 5]. The secondly issue that we address relates to the fact that nicotine is the principal psychoactive constituent of tobacco, understanding its biological effects on fetal and maternal health is critical, as it may affect distinct biochemical pathways when compared to other tobacco smoke constituents. Studies concerning the morphological effects of nicotine in fetus have shown a significant decrease in bone and cartilage growth [8], reduced body mass in newborn infants, and a decrease in the biparietal-to-occipital frontal head measurement, and tibia/femur ratio [7]. One possible explanation that could link nicotine and the negative regulation of development is retinoic acid (RA) signaling. RA is an indispensable molecule involved in the regulation of gene expression and cell-cell signaling during early development [9]. RA can cross the cell membrane and bind to specific nuclear receptors, such as retinoic acid receptors (RARs) and retinoid X receptors (RXRs) [9]. Studies regarding the role of RA receptors during embryogenesis have shown that RARs are essential for the expression of HOX genes and skeletal development [10, 9]. Nicotine has been previously associated with inhibition of the RARβ gene in lung cancer, which suggests that nicotine affects RA signaling in human tissues [11]. Therefore, RA signaling is a plausible pathway through which nicotine could affect cell differentiation and cause human fetal morphological abnormalities. However, the molecular mechanisms underlying the progression or the cause of fetal abnormalities related to cigarette smoking remain unknown.

To understand these mechanisms, we performed systems chemo-biology analyses to elucidate the nature and number of proteins and modules that are associated with prenatal tobacco smoke exposure. Different protein-protein interaction (PPI) and chemical-protein interaction (CPI) networks derived from interatome projects were described. In a first analysis, we prospected and analyzed a network using a list of 95 harmful tobacco constituents [2], to elucidate how these substances could act together to influence embryonic and fetal development. In a second systems chemo-biology analysis, we prospected data on the interactome and small compounds for nicotine alone and examined how they could negatively affect cell differentiation and bone development and lead to morphological abnormalities. Furthermore, we conducted a gene ontology (GO) analyses of the major biological processes derived from the PPI and CPI networks.

A model of how selected TCs could influence embryonic development was generated. We also developed a separate model of how nicotine could affect cell

differentiation and bone development. Taken together, our systems chemo-biology data are the first to show how tobacco smoke can affect fetal and embryonic development in a systemic matter at the molecular level.

Materials and Methods

Interactome data mining and design of the chemo-biology network

To design chemo-biology interactome networks and to elucidate the interplay between development and TCs, the metasearch engines STITCH 3.1 [http://stitch.embl.de/] and STRING 9.0 [http://string-db.org/] [12, 13] were used. STITCH software allows visualization of the physical connections among different proteins and chemical compounds, whereas STRING shows protein-protein interactions. Each protein-protein or protein-chemical connection (edge) shows a degree of confidence between 0 and 1.0 (with 1.0 indicating the highest confidence). The parameters used in STITCH software were as follows: all prediction methods enabled, excluding text mining; 20 to 50 interactions; degree of confidence, medium (0.400); and a network depth equal to 1. The results gathered using these search engines were analyzed with Cytoscape 2.8.2 [14].

In addition, the GeneCards [http://www.genecards.org/] [15, 16], KEGG [http://www.genome.jp/kegg/] [17], iHop [http://www.ihop-net.org/UniPub/iHOP/] [17], PubChem [http://pubchem.ncbi.nlm.nih.gov/], ALOGPS 2.1 [http://www.vcclab.org/lab/alogps/] [19], AmiGO 1.8 [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi] [20], and Gene Expression Atlas [http://www.ebi.ac.uk/gxa/] [21] search engines were also employed using their default parameters.

To prospect protein-protein and chemical-protein interactions (PPI and CPI, respectively), we entered each TCs into STITCH program. TCs that were not present in the STITCH database (or those that did not shown any protein connections) and particularly well described components such as nitric oxide, phenol and carbon monoxide, were excluded from the analysis.

Different small CPI and PPI networks were obtained (data not shown), and these networks were further analyzed using Cytoscape 2.8.2. Each network generated by STITCH and STRING was combined into a large network using the Advanced Merge Network function, which was fully implemented in Cytoscape software.

Gene expression data for the main associated nodes of tobacco components

To determine whether mRNA sequences associated with specific proteins connected to each TC were present during development, we searched the transcriptome data from the Gene Expression Atlas [22]. We used the protein name and expression data for *Homo sapiens* embryo and fetuses for the initial inputs. The expression data indicated overexpressed and underexpressed genes (**S-Table 1**, **see Supplementary Material 1**). Gene Expression Atlas infers the expression data for specified gene by providing a list of experimental studies [22]. We considered a gene overexpressed or underexpressed based on the number of studies which states the expression state of our input.

Proteins that are only present in embryonic tissue were colored green, whereas proteins that are only present in fetal tissue were colored pink (**S-Table 1**, **see supplementary material**). The blue nodes indicate the presence of a protein in both embryonic and fetal tissue (**S-Table 1**, **see supplementary material**). Uncolored nodes (default color white), connected to a TCs were either not present in any of the selected

tissues in the initial input or were not found in the Gene Expression Atlas database (**Fig. 1**).

Solubility predictions for major tobacco component-associated CPI-PPI networks

To predict the solubility of each TC in an aqueous environment, such as in blood and plasma, we used the program ALOGPS 2.1 [http://www.vcclab.org/lab/alogps/]. ALOGPS allows simulation of the probable solubility of a given compound determined based on its structural formula or CAS number. Compounds with a solubility of less than 35 g/L [values of ALOGpS and logS (exp)] were considered lipophilic. ALOGPS 2.1 was used with its default parameters.

Module analysis of major tobacco component-associated CPI-PPI networks

The large CPI-PPI network obtained from the initial search (**Fig. 1**) was analyzed in terms of the major cluster or module composition using the program Molecular Complex Detection (MCODE) [23], which is available at http://baderlab.org/Software/MCODE. MCODE is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions according to given parameters stipulated by the researcher [23]. The parameters for cluster finding were as follows: loops included; degree cutoff 2; expansion of a cluster by one neighbor shell allowed (fluff option enabled); deletion of a single connected node from clusters (haircut option enabled); node density cutoff, 0.1; node score cutoff, 0.2; kcore, 2; and maximum network depth, 100. Each cluster generates a value of "cliquishness" (C*i*), which is the degree of connection in a given group of proteins. Thus, the higher the C*i* value, the more connected the cluster [23].

Centrality analysis of the major tobacco component-associated CPI-PPI networks

Centrality analysis was performed using the program CentiScaPe 1.2 [24]. In this analysis, the CentiScaPe algorithm evaluates each network node according to the node degree, betweenness, and closeness to establish the most "central" nodes (proteins/chemicals) within the network. Thus, the most relevant node for a determined biochemical pathway or module can be obtained and further analyzed. In general terms, the closeness analysis (1), indicate the probability that any protein/chemical compound (node in our network) is relevant to another protein/chemical compound (node) in a signaling network or its associated network [25], as determined using equation (1):

$$Clo(v) = \frac{1}{\sum w \in v^{dist(v,w)}} \quad (1)$$

where the closeness value of node *v* (*Clo(v)*) is determined by computing and totalizing the shortest paths among node *v* and all other nodes (*w*; (*dist(v,w)*) found within a network (1). The average closeness (*Clo*) score was obtained by calculating the sum of different closeness scores (Clo_i) divided by the total number of nodes analyzed (*N(v)*) (equation 2):

$$\langle Clo \rangle = \frac{\sum_i Clo_i}{N_{(v)}} \quad (2)$$

The higher the closeness value compared to the average closeness score, the higher the relevance of the protein/chemical compound to other protein nodes within the network/module. In turn, betweenness indicates the number of shortest paths that go through each node (equation 3) [26]:

$$Bet(v) = \sum_{s \neq v \neq t} \frac{\sigma_{sw}(v)}{\sigma_{sw}} \quad (3)$$

where σ_{sw} total number of the shortest paths from node *s* to node *w*, and $\sigma_{sw}(v)$ is the number of those paths that pass through the nodes. The average betweenness score (*Bet*) of the network was calculated using equation (4), where the sum of different betweenness scores (Bet_i) is divided by the total number of nodes analyzed ($N(v)$):

$$\langle Bet \rangle = \frac{\sum_i Bet_i}{N_{(v)}} \quad (4)$$

Thus, nodes with high betweenness scores compared to the average betweenness score of the network are responsible for controlling the flow of information through the network topology. The higher a node's betweenness score, the higher the probability that the node connects different modules or biological processes, such nodes are called bottlenecks nodes.

Finally, the node degree ($Deg(v)$), is a measure that indicates the number of connections (Ei) that involve a specific node (v) (equation 5):

$$Deg(v) = \sum E_i \quad (5)$$

The average node degree of a network (*Deg*) is given by equation 6, where the sum of different node degree scores (Bet_i) is divided by the total number of nodes ($N(v)$) present in the network:

$$\langle Deg \rangle = \frac{\sum_i Deg_i}{N_{(v)}} \quad (6)$$

Nodes with a high node degree are called hubs [26] and have key regulatory functions in the cell.

Gene ontology analyses of major tobacco component-associated CPI-PPI networks

The CPI-PPI modules generated by MCODE were further studied by focusing on major biology-associated processes using the, Biological Network Gene Ontology (BiNGO) 2.44 Cytoscape plugin [27] available at http://www.cytoscape.org/plugins2.php#IO_PLUGINS. The degree of functional enrichment for a given cluster and category was quantitatively assessed (*p*-value) using a hypergeometric distribution. Multiple test correction was also assessed by applying the false discovery rate (FDR) algorithm [28], which was fully implemented in BiNGO software at a significance level of $p < 0.05$. The most statistically relevant processes were taken into account when developing the interaction model.

Results and Discussion

Data prospecting and topological design of a major CPI-PPI network of different tobacco constituents

Systems chemo-biology tools allow interatome networks of high-throughput data to be design for chemical-protein or protein-protein interaction (CPI and PPI, respectively) networks. Thus, we have examined the relationship between tobacco components (TCs) and the fetus during embryonic development in human female smokers using systems chemo-biology tools. Many of the thousands of substances in

tobacco smoke are considered to represent public hazards, and some have carcinogenic potential.

We searched data of previously reported harmful tobacco smoke constituents [2]. From the listed substances, we generated 51 small CPI-PPI networks (data not shown). Compounds that were not present in the STITCH database were excluded from the analysis. We also excluded well-described compounds such as carbon monoxide, nitric oxide and phenol, because they affect a broad spectrum of proteins involved in different processes, which clouds our ability to understand the least well-described compounds.

Both STRING and STITCH add the nodes with the highest probability to be connected to a given node. Therefore, to create different CPI-PPI networks, we identify 20 to 50 additional proteins linked to each compound using only STITCH and STRING data and merged all networks, by using the Advanced Network Merge tool. This software is fully implemented in Cytoscape, and it scans each network for proteins with the same name to be merged, which generates a single large network (referred to as the "main network", **Fig. 1A**). After creating the small networks, we found that RA and receptors were present in the nicotine network. We decided to expand the nicotine network by adding a small network including RA, proteins related to RA signaling and embryonic development (**Fig. 1B**). The nicotine module was extracted from the first network to be studied independently because it showed a distinct module within the main network.

The resultant network after the nicotine module was extracted was referred as the "major CPI-PPI network" and was composed of 898 nodes and 3452 edges (**Fig. 1C**). It should be noted that, after merging each of the small CPI-PPI networks, two substances, 3-aminobiphenyl and dicyclohexyl, did not display any proteins in common with other compounds and were excluded from the analysis. Remarkably, the major CPI-PPI network did not show a wide overlap among the nodes, which indicates that the substances included in the network may have a broad influence and most likely affect different bioprocesses. This result could be correlated with the number of different morphological abnormalities observed in the fetuses of smoking pregnant women [3].

We next aimed to strengthen our understanding of our networks. We examined two types of data: (i) transcriptome data for each node directly associated with TCs to clarify whether the mRNA and, by inference, the proteins were present in the fetus (pink color), embryo (green color) or both (blue color) (**Fig. 1, S-Table 1, see Supplementary Material 1**); and (ii) solubility prediction for the TCs and how this factor may influence the developing organism, by characterizing each TC as hydrophilic or lipophilic (hydrophobic) (**Fig. 1, S-Table 2, see Supplementary Material 1**).

Interestingly, the majority of the nodes (145 of 234 total nodes) were found in different phases of human embryonic development, showing that the action of the TCs might not be restricted to one specific phase of development (from embryo to fetus) and, thus, may affect the organism over a long period. Nodes that did not show expression in either stage of development were left with uncolored (white) (**Fig. 1**, **S-Table 1**, **see Supplementary Material**).

To predict TCs solubility, we used the program ALOGPS 2.1. ALOGPS allows the probable solubility to be simulated for a given compound, as determined by its structural formula or CAS number. Among 48 TCs in our major CPI-PPI network (**Fig. 1C**), we identified 21 lipophilic compounds and 27 hydrophilic components. Of the 27 hydrophilic components, 10 are inorganic, and 17 are organic (**S-Table 2**, **see Supplementary Material**).

In addition, we used the program CentiScaPe 1.2 to examine the major CPI-PPI network for the most relevant proteins/compounds (**Figs. 1** and **2**). In a scale-free biological network, the most important nodes are the so called hub-bottlenecks (HBs) [29] because they combine the bottleneck function (nodes that control the information flow in a given network and display a betweenness score above the network average) and the hubs characteristic (nodes with a high number of connections above the average node degree value of the network). Thus, HBs are critical nodes in a biological network [29]. In our analysis, we observed 143 HB nodes, of which 30 are TCs, and 53 were marked as present in both the fetus and embryo, 17 only in the embryo, 7 only in the fetus and 36 in neither fetus nor embryo (white nodes) (**Fig. 2**). White nodes present in all of the networks are either not connected directly with the selected compound, or do not show expression in any of the selected stages of development. Thus, it is possible to observe that the majority of the most critical nodes in the large CPI-PPI network are present in both the fetus and embryo (**Fig. 2**). Because we only colored the direct nodes associated with a TC, it is clear that the TCs have a broad impact during development, acting in critical nodes that are present in both embryo and fetus. Pregnant smoking women that have a constant habit of consuming tobacco in its different formulations (e.g., through cigarette and pipes), could subject the developing embryo/fetus to a continuous level of toxic substances and thereby impair many aspects of development.

Furthermore, we sought to evaluate which TCs have the broadest effects on the major CPI-PPI network. Therefore, a closeness analysis was performed. Considering that the nodes showing the highest closeness are most relevant to the greatest number of nodes in a network [24], it can be assumed that the TCs exhibiting the highest closeness are those with the greatest systemic effect and impact the greatest number of proteins. A graph of closeness and betweenness was generated, showing that 33 TCs (from a total of 48 present TCs) have closeness value above the average closeness of the network (**S-Figure 1**, **see Supplementary Material 2**). This finding is consistent with our interpretation that TCs have a systemic effect, impacting different proteins and physiological processes.

To understand how TCs interact with their targets, we analyzed the major CPI-PPI network for modules with the program MCODE. Modules are defined as topological elements with very connected nodes. The module analyses allowed the identification of compounds that are more likely to interact with the same connected group of proteins. From these analyzes, we obtained the major TCs that affect different modules. After extraction of the nicotine subnetwork, MCODE found 22 significant modules (**Figs. 3-7**).

Once the modules were obtained (**Figs. 3-7**), a gene ontology (GO) analysis was performed using BiNGO. Biological processes that are important for the development of organisms were listed by BiNGO (**S-Table 3**, **see Supplementary Material 1**). Likewise, we performed additional GO analyses for the selected HBs (**Table 1**) and in each cluster (**S-Tables 4-25, see Supplementary Material 1**).

Clusters that present were not associated with significant GOs due to a lack of data or were highly speculative in our analysis were excluded (**S-Figure 2**, **S-Tables 13, 15, 16, 17, 18, 22** and **25, see Supplementary Material 1 and 2**).

Although the effects of tobacco smoking are well known, it is not clear whether TCs display a promiscuous or a restricted mechanism of action that affects different biological processes, such as morphogenesis. Thus, an understanding of the combined mechanisms underlying the action of TCs is necessary to evaluate their systemic effects. The major CPI-PPI network (**Fig. 1C)** indicates 48 different TCs, present in both the particle and vapor phases of tobacco formulations. Additionally, topological analysis of

the modules and centralities of the major CPI-PPI network allowed a specific data-set to be generated that indicates the influence of cigarette smoking-associated TCs on embryonic development in terms of individual and clusters of CPI-PPI modules. Furthermore, the individual modules were explored considering their associated GOs and how the different connected TCs can affect the annotated bioprocesses.

Systemic effects of tobacco smoking in human embryogenesis: redox and prostaglandin metabolic processes

The modularity data gathered from the major PPI-CPI network (**Fig. 1C**) were subjected to GO analysis. The GO analysis of clusters 1, 4, 11, 16, and 20 (**Fig. 3A-E**) revealed five main processes annotations, as indicated: (i) oxidation reduction (redox), (ii) prostaglandin metabolism, (iii) steroid biosynthesis, (iv) lipid modification, and (v) unsaturated fatty acid metabolism (**S-Tables 4, 7, 14, 19 and 23**, **see Supplementary Material 1**). Given the overlap among the different processes, these subnetworks were merged into a single network (**Fig. 3F**). It was observed that lipophilic molecules (e.g., chrysene, toluene, benz[a]anthracene, benzo[b]fluoranthene, 7H-dibenzo(c,g)carbazole, 2-naphthylamine, 4-aminobiphenyl and 5-methylchrysene; **S-Table 2**; **see Supplementary Material 1**) were observed to be most connected to the proteins annotated as being involved in redox process (**Fig. 3A**). Moreover, lipophilic molecules are related to increases in reactive oxygen species (ROS) and have been observed to cross cellular membranes and act in protein folding in the endoplasmatic reticulum, which alters the expression of genes involved in antioxidant defense, inflammation, energy metabolism, apoptosis and the cell cycle [30].

Homeostasis ROS and antioxidants has been reported to be necessary for a healthy pregnancy [30]. Tobacco consumption has been associated with redox mechanisms and the generation of oxidative stress, leading to an inflammatory response [30 to 33]. In adult lung tissues, cigarette abuse can increase the levels of superoxide radicals ($O_2^{\bullet-}$) through the activation of NADH oxidase [33], and lead to high ROS levels and the activation of pro-inflammatory factors [33]. It is important to highlight the fact that mitochondrial damage caused by ROS can result in a feed-forward process that creates further oxidative damage [30].

Within the merged network (**Fig. 3F**), two prostaglandin synthases (PTGS1 and PTGS2), which are enzymes responsible for the synthesis of prostaglandins, and two 5-lipooxygenases (ALOX5 and ALOX15B), which play a role in the synthesis of leukotriene [34], were identified. PTGSs are not only related to inflammatory responses when they are present at high levels in tissues but are also associated with normal pregnancies due to promoting adequate circulatory adaptation and regular maternal-fetal blood flow [31, 35]. In addition to the results of our GO analyses, it is known that maternal smoke diminishes prostaglandin levels, which causes low birth weight [35], most likely due to impairment on the establishment of circulatory systems between the mother and fetus.

Another important molecule is arachidonic acid (AA), which is associated with inflammatory processes and embryonic developmental problems. It has been demonstrated that AA is released from the amniochorion and oxidized by PTGSs in the classical prostaglandin cascade [31]. Moreover, AA is a precursor for leukotriene, which is formed via the action of ALOX5 and ALOX 15B, and it plays a critical role in pro-inflammatory responses [34]. Linking the generation of leukotrienes and inflammation, neutrophil granulocytes are the first responders during the acute phase of inflammation in response to environmental exposure, and they migrate to the site of exposure following leukotriene B4 signals through chemotaxis [36, 37]. NADH

oxidases are also important to achieve a successful immune response involving neutrophils because the action of neutrophils generates $O_2^{\bullet-}$ by NADH oxidase [38], demonstrating the interplay between redox mechanisms and inflammation processes.

Consistent with our interpretation, one study showed that smokers exhibit higher levels of plasmatic leukotriene B4 and, 20-carboxy-leukotriene B4 and elevated expression of 5-lypoxygenase in comparison with non-smokers [39]. In addition, arsenic (**Fig. 3B**), which is present in this module, is related to increased oxidative stress via the redox mechanisms [40].

Considering the data amassed in this module, it is possible to speculate that pro-oxidative stimulation by TCs, like those included in **Fig. 3** can generate a pro-inflammatory cascade, followed by downregulation PTGSs and increased availability of AA and leukotriene, which promotes a continuous pro-inflammatory process.

Systemic effects of tobacco smoking in human embryogenesis: steroid biosyntheis, lipid modification, and unsaturated fatty acid metabolism

One interesting aspect of the major GOs associated with the analyzed modules is their strong link with the metabolism and modification of lipids (e.g., cholesterol) and steroids (**Figs. 3A to 3F, S-Tables 4, 14, 19** and **23, see Supplementary Material 1**).

Exposure of rats to cigarette smoke condensate has been reported to cause inhibition of follicular development, resulting in a decrease in oocyte maturation due a deficiency in the levels of estradiol [6]. As with other steroids, estradiol is synthesized from cholesterol and a second hormone, pregnenolone, which is the precursor for all other steroids and, is produced in the mitochondria from cholesterol [41]. The same study indicated that cigarette smoke causes disruption in the transport of cholesterol into the mitochondria by the interfering with StAR, a protein required for the import of cholesterol [41]. The StAR protein was not present in the major CPI-PPI network (**Fig. 1C**). However, we search the data using STITCH to verify whether the protein connected to one of our clusters. Indeed, StAR was connected to POMC, a melanocortin receptor related to a broad range of processes, including inflammation and steroidogenesis. POMC was present in our module (**Fig. 3B** and **3F**) (data not shown) and observed to be expressed in both the fetus and embryo.

Another TC associated with steroidogenesis in our module (**Fig. 3C**) was isoprene, which is in turn, associated with FDPS and FDFT1 (**Fig. 3C**). The enzyme FDFT1 acts in the first step of cholesterol synthesis by dimerizing farnesyldiphosphate to form squalene, while FDPS catalyzes the production of geranyl-pyrophosphate into farnesyl-pyrophosphate, an intermediate in cholesterol and sterol biosynthesis. Both proteins are present in embryo and fetuses, and taking our data into consideration, isoprene could be related to the disruption of steroid synthesis.

It is also important to note that the some hydrophilic TCs, such as pyridine, *N*-nitrosoanabasine and urethane, are mainly connected to the CYP cluster (**Fig. 3A and 3F**). The CYP membrane proteins, which belong to the cytochrome P-450 family of proteins in the inner-membrane of the mitochondria, have substrates that include many steroids and lipids. Interestingly, *N*-nitrosoanabasine is connected to the UGT cluster (**Fig. 3A** and **3F**). UGT proteins are responsible for catalyzing the glucuronidation of estrogens and androgens, which makes these molecules more soluble and easily transportable. This process may be critical during development for the proper exposure of the fetus to maternal hormones, and this exposure essential during differentiation, body pattern formation and sex determination. Maternal smoking is also associated with defects in human testis development [42].

In summary, the systems chemo-biology data show that TCs negatively affect the synthesis of cholesterol via impacting proteins necessary for its synthesis, such as FDPS and FDFT1. The action of lipophilic molecules in the mitochondria has been observed, and hydrophilic compounds could also impair mitochondrial function [30]. Indeed, in our modules, it can be observed that some hydrophilic TCs could affect the import of cholesterol to the mitochondria by acting on transport proteins, such as StAR. We also noted that TCs could have an effect on the UGT protein cluster, down-regulating hormone solubility and transport.

Systemic effects of tobacco smoking on human embryogenesis: regulation of cell communication and cell-cell signaling

Cellular communication is of great importance for embryonic development, being essential to coordinate the different biochemical signals required to control cellular differentiation and migration. Interestingly, GO analysis of clusters 2 and 18 (**Fig. 4**) revealed two related processes: (i) regulation of cell communication and (ii) cell-cell signaling (**S-Tables 5 and 21**, **see Supplementary Material 1**).

Considering the different proteins found in cluster 2 (**Fig. 4A**), two nodes appear to be important TC targets: (i) signal transducer and activator of transcription 3 (STAT3), which is related to cell-cell signaling in stem cell cultures [43]; and (ii) colony stimulating factor receptor-beta (CSF2RB), a CSF2 receptor molecule that is important for post-blastocyst development, embryo differentiation, and implantation [44]. Epidermal Growth Factor (EGF) and its receptor EGFR were also present in this subnetwork (**Fig. 4A**). EGFR is a plasma membrane glycoprotein that is necessary for implantation and epithelial differentiation as well as for cell signal transmission during embryogenesis [43, 45, 46]. Cluster 2 also contains APP, a protein responsible for the inhibition of NOTCH, which is related to cell-cell communication [47]. Interestingly, Notch2 mutant mice exhibit defects in kidney, eye and heart development [48], similar to defects observed in the fetuses of human female smokers [3].

It should be noted that both EGF and EGFR were linked to cadmium and methylamine (**Fig. 4A**) in our systems chemo-biology data, and one study showed that cadmium (**Fig. 4A**) could alter the progesterone concentration in the placenta [49]. Other growth factors, such as nerve growth factor (NGF) and transforming growth factor, alpha (TGFA), are also present in cluster 2. It is possible that the selected constituents, cadmium and methylamine (**Fig. 4**), can play a negative role in hormone and cell-cell signaling via inhibition of NOTCH and other growth factors such as EGF, EGFR, NGF and TGFA.

Progesterone plays a pivotal role in controlling implantation, zygote division and the activation of gene expression cascades that initiate cell-cell signaling during development [10, 50, 51]. Knockout of the progesterone receptor in mice leads to an increase in proliferation and inflammation mechanisms rather than embryonic stem cell differentiation [10].

Considering the importance of the placenta for the regulation of hormone levels throughout pregnancy, one study showed that smoking is associated with spontaneous abortion due to diminishing the levels of serum estradiol and chorionic gonadotropin [52]. As discussed previously in this work, TCs play a negative role in steroid synthesis. Because many abnormalities are observed in the newborn fetuses of smoking pregnant women [3], and taking into account our data, it becomes clear that TCs are likely to act through a general mechanism altering the patterning of gene expression, such as hormone signaling. Moreover, all hormones play critical roles in the expression of homeotic genes during development, especially the HOX genes [10]. HOX genes are

expressed in a spatial-temporal manner, and any disruption or negative regulation of the expression of these genes could affect body pattern specification and cell differentiation [10]. It should be noted that in the major CPI-PPI network, 1,3-butadiene is linked with HOXD13 (**Fig. 1C**). The HOXD cluster is associated with final stages of limb morphology, whereas mutations in HOXD13 are associated with abnormal limb length [53]. Considering that tobacco abuse can lead to limb aberrations in newborns [3], the HOXD cluster should be an interesting target with respect to understanding the effects of cigarette compounds during and development. Consistent with our interpretation, when 1,3-butadiene was injected into pregnant mice at a high dose (8,000 ppm) [http://www.who.int/ipcs/publications/cicad/en/cicad30.pdf], the newborn offspring showed aberrations in the skull, spine, sternum, ribs and long bones [http://www.who.int/ipcs/publications/cicad/en/cicad30.pdf].

It should also be noted that cluster 2 (**Fig. 4A**) contains ERBB2, ERBB3 and ERBB4, which are all members of the tyrosine kinase family, and show a with similar structure to EGFR, which appears to be crucial for skeletal development [54].

Thus, a potential candidate for the mechanism underlying the impact of TCs is hormone synthesis and signaling, which subsequently affects homeobox expression and spatial-temporal coordination. Two substances (trimethylamine and ethylbenzene) in cluster 18 are connected to CHRNA4 (**Fig. 4B**), a cholinergic nervous system receptor that causes opening of plasma membrane channels, leading to ion entry. Interestingly, CHRNA4 is present in the embryo and fetus, and this protein alone is connected to YWHAH (also known as 14-3-3η) (**Fig. 4B**) [55], a protein related to cytoskeletal proteins, apoptosis and cell cycle [http://www.genecards.org]. YWHAH negatively regulates PDPK1, a master kinase that acts by phosphorylating a broad range of important kinases, including as AKTs (AKT1 to 3) and SMADs (SMADs 2 to 4 and SMAD7), which are proteins associated with chromatin remodeling, cell differentiation and cell proliferation [http://www.genecards.org]. PDPK1 also prevents the nuclear translocation of SMADs 2-4 and SMAD7.

Therefore, in addition to the role of TCs in hormone, homeobox and growth factor signaling, they might also disturb chromatin remodeling.

Systemic effects of tobacco smoking in human embryogenesis: metabolism of DNA, DNA damage stimulus, the cell cycle and chromatin organization

In the GO analysis of clusters 3, 6, 17 and 21 (**Fig. 5**), we identified two related processes: (i) RNA-splicing, and (ii) metabolism of DNA (**S-Tables 6, 9, 20 and 24**, **see Supplementary Material 1**).

CYP proteins are related to the metabolism of xenobiotics and are also responsible for the carcinogenic activation of compounds such as PAHs and aromatic amines [52]. While CYPs convert these compounds into reactive carcinogens, glutathione S-transferases (GSTs) (**Fig. 5B**) detoxify PAH carcinogens after they are converted to glutathione by GSTM1 [52]. Regarding the importance of GSTM1 in the detoxification of PAHs, one study showed that a single adenine-to-guanine substitution in the GSTP1 gene that generates an isoleucine instead of a valine in the protein can lead to the loss of its detoxification capability [52]. This loss of the detoxification capability of GSTM1 can result from the action of a combination of TCs. In our GO analysis, not only were TCs associated with the metabolism of nucleotides in four different clusters, but each cluster contained different interacting compounds, including both hydrophilic (catechol) (**Fig. 5C to 5E**) and lipophilic (chrysene and 1,3-butadiene, crotonaldehyde) (**Fig. 5B**), as well as organic (chrysene and 1,3-butadiene) (**Fig. 5B**) and inorganic (beryllium, polonium-210 and arsenic) (**Fig. 5A**, **5C**, **5D**, and **5E**).

Remarkably, in cluster 6, both substances (chrysene and 1,3-butadiene) are linked to HPRT1 (**Fig. 5B**), a hypoxanthine that is responsible for the metabolism of purines.

The applied systems biology tools also identified inosine triphosphatase protein (ITPA) in cluster 6 (**Fig. 5B**). ITPA is important for the removal of deaminated purine nucleotides in mammals [56]. $ITPA^{-/-}$ mice show perinatal lethality, and $ITPA^{-/-}$ mouse embryonic fibroblast exhibit increased chromosomal abnormalities and accumulation of single strand-breaks in nuclear DNA [57]. Thus, given the role of ITPA in embryonic lethality and chromosomal aberrations, ITPA is expected to be a good target in attempting to understand the role of TCs in DNA repair and metabolism.

Moreover, 1,3-butadiene, has been found to be linked to increased genotoxic stress due to DNA damage, through the formation of DNA-DNA cross-links at adenine and guanine nucleobases by its metabolites, 1,2,3,4-diepoxybutane and 3,4-epoxy-1,2-butanediol [58, 59]. This compound has also been associated with epigenotoxic effects caused by the loss of global DNA methylation and trimethylation of histone H3 lysines 9, and 272 and H4 lysine 20, all of which are known for their roles in regulating gene expression patterns [59].

These associations revealed that substances in cigarette smoke can act through distinct pathways to affect DNA metabolism. Regarding the last set of TCs, DNA mutations not only have a teratogenic effect but also alter gene expression.

Next, in the GO analysis of clusters 5, 8 and 9 (**Fig. 6**), we identified two related processes: (i) DNA damage stimulation, and (ii) the cell cycle (**S-Tables 8, 11 and 12**, **see Supplementary Material 1**).

In this cluster, arsenic binds directly to PLM (**Fig. 6B** and **6D**), which is a protein with functions involved in chromatin organization, cell differentiation, DNA repair, protein sequestration and post-translational modifications [http://www.genecards.org]. PML is linked to significant proteins such as p53, p300 and BING2 (**Fig. 6B** and **6D**). BING2 is a protein that travels between the nucleus and cytoplasm and is linked to the activation MDM2 ubiquitination activity [http://www.genecards.org]. Thus, TCs could play a role in p53-mediated degradation by MDM2, leading to inefficient cell cycle arrest and DNA repair. MDM2 is found in clusters 5 and 8 (**Fig. 6A** and **6B**).

Furthermore, ethylamine is connected to USP2 (**Fig. 6A** and **6D**), a protein related to myogenic differentiation during embryogenesis, and is indirectly associated with the degradation of p53 by MDM2 [http://www.genecards.org]. Increased degradation of p53 can affect cell cycle arrest in G_1/S and increase the likelihood of DNA damage in the fetus. In this sense, in lung tissues, rats carrying *p53* mutations exposed to environmental tobacco smoke showed inefficiency in inducing pro-apoptotic genes, and an increase in the expression of proliferation, immune response and inflammation genes [60]. Evidence also suggests that chronic inflammation causes by environmental toxicants is linked to many aspects of DNA damage and tumorigenesis [61], which is consistent with the previously described connection between TCs and increased inflammation.

Considering the increased proliferation and inflammatory response caused by the lack of p53, urethane is directly connected to FOS (**Fig. 6B** and **6D**), a central protein involved in proliferation, and TNF, a pro-inflammatory cytokine. Urethane is reported to affect the cell cycle during neural tube development in mouse embryogenesis [62]. In addition, another study found that urethane alters placental morphology and down-regulates cell cycle genes as well as cytokines and other growth factors [63].

Both urethane and ethylamine appear to affect the cell cycle. Thus, our data show that TCs (**Fig. 6)** could have an indirect effect on p53 by causing its degradation

through MDM2, which is stimulated by USP2 and BING2. This down-regulation of p53 is related not only to DNA damage but also to pro-inflammatory responses and increased proliferation.

In the GO analysis of cluster 7 (**Fig. 6B**), we only identified chromatin organization (**S-Table 10**, **see Supplementary Material 1**) as a major biological process.

Cluster 7 included dimethylamine, which is connected to MBD2 (**Fig. 6B**), a protein associated with regions of methylated DNA in CpG islands that can recruit histone deacetylases and DNA methyltransferases. This indicates a possible interaction of TCs with chromatin organization. Epigenetic alterations have been linked to developmental changes that have a pathological outcome in adult life [5, 64, 65]. DNA methylation is also correlated with gene silencing through polycomb repression complexes (PRC), and in lung cancer, tobacco smoking is correlated with aberrant histone methylation [66].

Although the mechanisms underlying these pathologies have yet to be elucidated, it is possible that the impact of TCs on the methylation of polycomb genes such as *EZH2*, *SUZ12* and *EED*, which are part of PRC2, can alter gene expression and silencing. PRC2 is involved in to the silencing of many HOX genes, which is critical for normal fetus development. Additionally, MBD2, the protein linked to dimethylamine, is correlated with the inactivation of sexual chromosomes and is a candidate for recruiting DNA-methyltransferases (DNMTs) to the silenced promoters of long-term repressed genes [67]. Hence, MBD2 is a good candidate for improving the understanding the molecular mechanisms underlying the impact of TCs on chromatin structure.

DNA methylation also appears to be important for the expression of insulin factors [65]. Interestingly, in our analysis, cluster 14 (**S-Figure 2, see Supplementary Material 2**) showed insulin signaling as the related GO (**S-Table 17**, **see Supplementary Material 1**).

Effect of nicotine on retinoic acid signaling, cell proliferation and differentiation

Thus, a second analysis using tools from chemo-systems biology was developed to elucidate the relationships between nicotine, RA signaling and cell differentiation in the fetus during embryonic development in female smokers. The extracted subnetwork was examined separately due the distinct module involving nicotine and its interacting proteins. RA was added to the network because we observed that many proteins connected to nicotine were related to embryonic development and RA signaling.

Thus, the data amassed allowed the design of a major CPI network associated with nicotine and RA signaling (**Fig. 7**), which revealed several proteins that related to embryonic development, stress responses, and cell proliferation.

Therefore, Cytoscape 2.8.2 was used to build a CPI network with 330 nodes and 4078 connections (**Fig. 7**). Several of the proteins in the CPI network, directly connected to nicotine, including: (i) VEGFA, a factor that induces blood vessel formation (angiogenesis) [68]; (ii) DNMT1, a DNA methyl transferase responsible for the methylation of 5' CpG islands in DNA and epigenetic regulation (**Fig. 7**) [69]; (iii), FOS and JNK1 (MAPK8), which are both inducers of cell proliferation [70, 71]; and (iv) SOD2, which is responsible for mitochondrial superoxide dismutation. In addition, many proteins involved in cellular responses to stress, DNA damage and inflammation are interconnected with nicotine in the CPI network. Similar to cytokines, such as IL1, IL2 and TNF, the proteins BAX, SOD2 and nicotinic receptors (CHRNA 2, 4-6, 9, 10, and CHRNB 2-4) (**Fig. 7**) were involved with nicotine. Nicotine binds specifically to

proteins that are related to stress and cellular proliferation (**Fig. 7**). However, many of these proteins can inhibit growth, such as the nuclear phosphoprotein FOS (c-Fos) [72].

FOS is an evolutionarily conserved transcription factor that is present in all vertebrates and is essential for mitogenic signal propagation and the induction of cell division [71]. The highest levels of c-Fos are detected near the peak of cell proliferation, and its action antagonizes the RA signaling mechanism involved in differentiation [71]. Thus, as one process is linked to differentiation and the other to proliferation, the processes are antagonistic and cannot occur simultaneously in the same cell.

The protein JNK1 also induced by nicotine and is present in our CPI network (**Fig. 7**) [70]. Once induced by nicotine, we hypothesize that the actions of FOS and JNK1 lead to increased cell proliferation due to altering RA signaling and cell differentiation. The systems biology analysis also showed that FOS is related to the PDGF (platelet-derived growth factor) protein family. The proteins SIS (PDGFB), PDGFA, PDGFC, IEGF (PDGFD), and PDGFR are known members of this family that have been observed to induce cell proliferation and rearrange the cytoskeleton and are affected by nicotine [73]. Therefore, nicotine potentially acts as a cellular proliferation inducer that antagonizes the cellular differentiation processes controlled by RA signaling (**Fig. 8**).

We also observed a connection between nicotine and JNK1 through their association with RARα in the CPI network (**Fig. 7**). JNK1 is expressed when the cell undergoes cellular stresses, such as inflammation, oxidative stress, and heat [74]. In a murine model, nicotine was found to be related to the expression of JNK1 in respiratory system tissues through nicotinic receptors and receptor kinins B1 and B2, whose stimulation, by a peptide called bradykinin, leads to increased levels of intracellular Ca^{2+} [70]. Cellular stress can activate JNK1, which phosphorylates RARα and causes its proteosomal degradation [74]. These relationships demonstrate that a stable level of circulating nicotine in a pregnant smoking woman inhibits RARs, which are transcription factors that are regulated by intracellular calcium influx and the recruitment of protein kinases.

In our CPI network, RA interacted with an extensive range of proteins, especially the Hox, PAX, FGF, p300, p53, and NANOG proteins (**Fig. 7**). This result was expected because RA is a compound with very diverse functions during embryonic development and adult life.

During embryonic development, Hox proteins are induced by RA, which is one of their most important inducers [10]. Furthermore, RA binds to different proteins responsible for chromatin remodeling, such as SMAD2, SMAD3 and p300 (**Fig. 7**). Some of the proteins have been characterized as being involved in epigenetic modification and cell differentiation, including AKT1, JNK1, OPN, VFR, BMP2, BMP4 and RARs (RARα, RARβ, RARγ, RXRA, RXRB and RXRG), which are found founded in the CPI networks connected to RA.

Nicotine has been reported to inhibit RARβ in lung cancer [11]. Thus, in response to the presence of retinoids, chicken ovalbumin upstream promoter (COUP-TF), a nuclear receptor, modulates the expression of RARβ [11]. Nicotine has been observed to be a suppressor of the COUP-TF receptor due to inducing nuclear receptor TR3 (NR4A1, nuclear receptor subfamily 4, group A, member 1), which inhibits the transactivation of COUP-TF by the RARβ promoter, preventing COUP- TF from binding in response to retinoid elements [11]. Interestingly, TR3 is stimulated by various biochemical signals, such as the release of Ca^{2+} [11]. Nicotine can stimulate the opening of Ca^{2+} ion channels in neurons by recruiting the cAMP responsive element binding (CREB1) protein [75]. Although this recruitment has mainly been observed in

neural cells, it is possible that this association could also occur in other cell types. Hence, nicotine crosses the cell membrane and causes a depolarization that could release intracellular Ca^{2+}, stimulating the up-regulation of a number of kinases, such as ERK1, p38 and JNK1 [76, 77]. Thus, one of the ways in which nicotine could promote the recruitment of TR3 would be through the action of kinases that are stimulated by the influx of Ca^{2+} and indirectly inhibit RA receptors (**Fig. 8**). Likewise, a lack of RARβ has also been described as an essential factor in the expression of p300 [78], which is involved in important cell processes, such as chromatin remodeling. Thus, it is expected that the inhibitory effect of nicotine on RARβ could cause errors in gene expression induced by p300 as well as possible chromosomal abnormalities.

The systems biology analysis performed in this study also showed that nicotine is directly connected to the protein CYS26A1 (**Fig. 7**). This protein is responsible for regulating RA levels [79] and is expressed in a spatial-temporal manner during the development of mice, mainly in the anterior segment of the embryo and in the neural crest-derived mesenchyme [79]. However, inhibition of this protein generates an accumulation of RA and leads to deformities in the embryo, such as abnormalities in the cerebellum, urogenital tract, and spinal cord [79]. The influence of nicotine on the inhibition of CYP26A1 has been identified in mammary carcinoma cells [80]. In the same study, it was shown that the inhibition of proliferation is unaffected by RA in cell cultures deficient in CYP26A1, indicating that CYP26A1 is required for this process [80]. Therefore, in addition to the inhibition of the RA receptor, the inhibition of CYP26A1 is responsible for the down-regulation of RA signaling and for inducing cellular proliferation. Thus, it is interesting to hypothesize that CYP26A1 could display the closest relationship with the Hox genes. The abnormalities observed in the fetuses of smoking women appear to act primarily during bone and skeletal development, where many Hox genes are necessary for the formation of the body pattern and tissue differentiation. This scenario was examined to some extent in a study that noted that the abnormal levels of RA caused by a lack of CYP26A1 increase the expression of HOXA1, resulting in embryonic lethality [79].

Moreover, nicotine exhibited 22 proteins in common with RA (**Table 2**). These proteins are mostly related to the immune system, stress, and cell proliferation (**Table 2**), indicating that nicotine affects RA signaling through cellular stress caused by constant tobacco use. Exposure to nicotine could cause an increase in proinflammatory signaling, leading to abnormal expression of VEGFA (vascular endothelium growth factor A) and other placental growth factors, which leads to impairment of the endovascular trophoblast and causes causing reduced uroplacental blood flow, culminating in fetal growth restriction [81] (**Fig. 8**). We observed that nicotine is directly linked with VEGFA in our analysis (**Fig. 7**).

Role of nicotine in the differentiation of bone tissue

An indirect association of nicotine with RA receptors was observed in the network via the influence of nicotine on the transcription factor JUN (**Fig. 7**). The JUN protein can be activated by the action of JNK1 during osteoblast differentiation [82]. $JNK1^{-/-}$ rats show negative regulation of transcription factors that inhibit the differentiation of osteoblasts. In addition, JNK1 was associated with interleukin 1 (IL1) and tumor necrosis factor α (TNFα), which are cytokines that increase osteoblast differentiation in cases of rheumatoid arthritis [82].

In a smoking woman, the levels of nicotine in the blood are maintained at a stable level depending on the degree of tobacco use [83]. During embryogenesis, constant levels of nicotine can affect bone development during osteoblast

differentiation, and morphological data have demonstrated a decrease in bone and cartilage growth [8] as well as a reduction of body mass in newborn fetuses and a decrease in the biparietal-to-occipital frontal head measurement [7]. Therefore it is possible to conclude that a stimulus during bone tissue differentiation might cause abnormal limb anomalies. The role of nicotine in JNK1 signaling could be one of the factors involved in this pathology. An additional impact of nicotine on bone tissue differentiation involves the relationship with the BMP2 and BSP proteins.

The BMP protein family includes the most potent osteogenic growth factors described to date [68] and is connected to nicotine (**Fig. 7**). A study in rabbits showed that treatment with nicotine affects BMP2 RNA levels and the activity of osteoblasts [68]. Similarly, the BSP protein is a glycoprotein that acts on bone mineralization, which has also been described as being inhibited by nicotine in rat osteoblast cells [84]. The effects of nicotine could go beyond these data because a number of other proteins connect to BMPs and BSP are linked to osteoblast differentiation and to bone tissue formation, such as BMP4, the vitamin D receptor (VDR) transcription factor and the zinc finger protein osterix OSX (SP7) (**Fig.7**).

Modularity analyses linking nicotine with abnormal embryonic development

Once the CPI network was generated (**Fig. 7**), we aimed to understand which major protein clusters might be present. For this purpose, MCODE was applied to analyze the highly connected regions in the network and to generate a cohesion coefficient defining a cluster [85]. The CPI main network (**Fig. 7**) showed the presence of six modules with a coefficient of cohesion greater than or equal to 3.00 (Clusters 1-6, **S-Fig. 3**, **see Supplementary Material 2**).

It was observed that nicotine appeared in clusters 1-4 (**S-Fig. 3A-D, see Supplementary Material 2**), but not associated with RA (only in **S-Fig. 3C, see Supplementary Material 2**), which exhibits many connections other than nicotine in the network. Nicotine is connected to 49 proteins with 281 connections, and RA is connected to 130 proteins with 1471 connections (**Fig. 7**). The expected result would be a prevalent modularization of RA over nicotine because it was related to many more proteins and connectors. It is possible that nicotine has a strong systemic effect, even though it nicotine has five times fewer connections than RA. From the systems biology analysis, it was observed that nicotine more readily clustered in a network focused on proteins involved in development and cellular stress.

Two main hypotheses can be drawn from these data: (i) nicotine affects the activity of proteins involved in development through cellular stress (**Fig. 7**), culminating in various morphological deformations in the embryo, and (ii) the effect of RA, although it is essential to the developing embryo and connected to more proteins than nicotine, is expressed in a spatiotemporal manner, and its role in embryonic development therefore occurs though small impacts, whereas the effects of nicotine is more massive and produces systemic damage in a short period of time. Thus, nicotine will always be present at certain levels and will be prone to increase with constant use, whereas RA levels will be regulated and restrained to specific stages of development and will not be able to overcome the constant cellular stress promoted by nicotine (**Fig. 8**). We also observed that certain clusters did not contain either nicotine or RA (**S-Figs. 3E and F, see Supplementary Material 2**). In cluster 5 (**S-Fig. 3E, see Supplementary Material 2**) we observed a prevalence of proteins linked to (i) chromatin remodeling, such as EZH2, EED, SUZ12, DNMT1, DNMT3A, DNMT3B, HDAC2, HDAC4, HDAC5, and (ii) development and differentiation, including several HOX proteins [A1, 1C (A5), 4B (D4), 2I (B1), B13, B4 and 4F (A11)], PAX1, PAX6,

NANOG, RARβ, RARγ, RXRβ, NOTCH1, CYP2B6, and CYP26A1. This last cluster was composed of 89 proteins with 300 connections and mainly consisted of developmental proteins, indicating that the network appears to be restricted to development processes, without a wide influence on other biological processes. Thus, this systems biology view also supports our hypothesis of the impact of nicotine regarding overriding the mechanisms of differentiation induced by RA.

Centrality analysis

To identify the major nodes within the CPI network (**Fig. 8**) the program CentiScaPe was used to calculate betweenness, closeness and node degree centralities. From these analyses, two graphs were generated containing the proteins that showed the highest centrality values (**S-Figs. 4** and **5, see Supplementary Material 2**). Interestingly, these nodes present a similar relevance order in both graphs. Thus, RA, nicotine, JNK1, p300, AKT1, p53 and ERK showed the highest betweenness, closeness and node degree values. As these proteins play major roles in cellular physiology, it was expected that they would exhibit higher values for the three variables. The proteins with the highest values were taken into consideration for the design of a molecular model of the effect of nicotine on embryonic development (**Fig. 8**).

As noted above, a lack of RARβ can affect the expression of p300. In the centrality analysis, it was observed that p300 appeared as the fourth protein, showing the highest values of betweenness, closeness, and the node degree (**S-Figs. 4** and **5, see Supplementary Material 2**). This scenario demonstrates that there is a major influence of p300 on the network regarding the number of connections with other proteins (92 proteins), the implications of its importance for neighboring proteins (closeness) and its relationships to clusters and bioprocesses (betweenness). Therefore, the negative regulation of this protein induced by nicotine can also lead to fetal malformations and could be a potential study target for understanding the influence of nicotine in development.

Conclusions

It is clear that fetal exposure to TCs during pregnancy increases the probability of later pathologies such as obesity and diabetes and causes morphological abnormalities. However, it is not clear how those abnormalities and pathologies arise or what the underlying mechanisms are behind them. In the present study, we showed, using systems biology tools, how the primary harmful constituents of tobacco interact with specific biological processes and affect them. Due the variety of abnormalities caused by TCs, it is clear that they utilize general mechanisms during development. Our cluster analysis results show that TCs act in many bioprocesses, including cell communication and signaling, hormone synthesis and signaling, DNA metabolism, DNA repair, and inflammation. Although these processes have wide effects on cellular and embryonic physiology, they can be disturbed by the levels of the constituents of tobacco smoke. Because these effects are complex, we developed an interaction model and divided it in five parts (**Figs. 9A** to **9E**): (A) increased inflammatory processes; (B) diminished nucleotide metabolism and increased DNA damage; (C) increased oxidative damage; (D) low hormone synthesis and signaling; and (E) negative regulation of gene expression, cell differentiation and cell signaling (**Fig. 9**). The systems model is related to low birth weight, an increased probability of abortion, morphological abnormalities (mainly in the skeletal system), low neutrophil activity and increased proliferation rates. Furthermore, our model can help improve knowledge and provide new insights

regarding how chemicals in cigarettes cause the many morphological abnormalities observed in the newborn offspring of smoking pregnant women. The role of nicotine in embryonic development has also not been well studied. The analysis performed in this study demonstrates that nicotine has an aggressive effect on cell differentiation, affecting RA signaling in the embryo, inhibiting RA receptors due to intracellular calcium influx and stimulating cell proliferation proteins that antagonize RA activity. Moreover, the accumulation of RA due to the inhibition of CYP26A1 can cause potentially teratogenic effects. Osteoblast differentiation is also affected by nicotine, via inhibiting proteins that stimulate bone tissue formation, which complements the TC model. Together, these data show that the birth defects observed in morphological studies could be caused by the negative action of nicotine on RA signaling. The networks also show that the pro-inflammatory pathway triggered by nicotine could be a factor leading to decreased body weight in the fetuses of smoking women. Finally, cluster analysis shows a systemic effect of nicotine, which could affect the network in a more aggressive and short-term way via cellular stress cascades.

References

1. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, el al. (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 21(48): 7435-7451.
2. Fowles J, Bates M (2000). The Chemical Constituents in Cigarette and Cigarette Smoke: Priorities for Harm Reduction. Available: http://www.moh.govt.nz/moh.nsf/pagescm/1003/$File/chemicalconstituentscigarettespriorities.pdf.
3. Hackshaw A, Rodeck C, Boniface S (2011) Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. Human Reprod Update 17(5): 589-604.
4. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, et al. (2009) Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. Ther Drug Monit 31(1): 14-30.
5. Morris CV, DiNieri JA, Szutorisz H, Hurd YL (2011) Molecular mechanisms of maternal cannabis and cigarette use on human neurodevelopment. Eur J Neurosci 34(10): 1574-1583.
6. Sadeu JC, Foster WG (2011) Cigarette smoke condensate exposure delays follicular development and function in a stage-dependent manner. Fertil Steril 95(7): 2410-2417.
7. Lampl M, Kuzawa CW, Jeanty P (2003) Prenatal Smoke Exposure Alters Growth in Limb Porpotions and Head Shape in Midgestation Human Fetus. Am J Hum Biol 15: 533-546.
8. Kawakita A, Sato K, Makino H, Ikegami H, Takayama S, et al. (2008) Nicotine Acts on Growth Plate Chondrocytes to Delay Skeletal Growth through the α7 Neuronal Nicotinic Acetylcholine Receptor. PLoS One 3(12): e3945.
9. Duester G (2008) Retinoic Acid Synthesis and Signaling during Early Organogenesis. Cell 134(6): 921-931.
10. Daftary GS, Taylor HS (2006) Endocrine Regulation of HOX genes. Endocr Rev 27(4): 331-355.
11. Cheng CG, Lin B, Dawson MI, Zhang XK (2002) Nicotine modulates the effects of retinoids on growth inhibition and RARβ expression in lung cancer cells. Int J Cancer 99 (2): 171-178.
12. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic. Acids Res 37: D412-D416.

13. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acid Res 28(18): 3442-3444.
14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
15. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13: 163.
16. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, et al. (2010) GeneCards Version 3: the human gene integrator Database.
17. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27-30.
18. Hoffmann R, Valencia A (2004) A Gene Network for Navigating the Literature. Nature Genetics 36: 664.
19. Tetko IV, Tanchuk VY (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J Chem Inf Comput Sci 42(5): 1136-1145.
20. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2): 288-289.
21. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A (2010) Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res 38(Database issue): D690-D698.
22. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, et al. (2012) Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res (Database issue): D1077-D1081.
23. BADER GD, HOGUE CW (2003) AN AUTOMATED METHOD FOR FINDING MOLECULAR COMPLEXES IN LARGE PROTEIN INTERACTION NETWORKS. BMC BIOINFORMATICS 4: 2.
24. Scardoni G, Petrerlini M, Laudanna C (2009) Analyzing biological network parameters with CentiScaPe. Bioinformatics 25(21): 2857-2859.
25. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Networks 27: 39-54.
26. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3(4): e59.
27. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21(16): 3448-3449.
28. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289-300.
29. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLoS Comput Biol 3(4): e59.
30. van der Toorn M, Rezayat D, Kauffman HF, Bakker SJ, Gans RO, et al. (2009) Lipid-soluble components in cigarette smoke induce mitochondrial production of reactive oxygen species in lung epithelial cells. Am J Physiol Lung Cell Mol Physiol 297(1): L109-L114.
31. Menon R, Fortunato SJ, Yu J, Milne GL, Sanchez S, et al. (2011) Cigarette smoke induces oxidative stress and apoptosis in normal term fetal membranes. Placenta 32(4): 317-322.

32. Yao H, Yang SR, Kode A, Rajendrasozhan S, Caito S, et al. (2007) Redox regulation of lung inflammation: role of NADPH oxidase and NF-kappaB signaling. Biochem Soc Trans 35(Pt 5): 1151-1155.
33. Yao H, Edirisinghe I, Yang SR, Rajendrasozhan S, Kode A, et al. (2008) Genetic ablation of NADPH oxidase enhances susceptibility to cigarette smoke-induced lung inflammation and emphysema in mice. Am J Pathol 172(5): 1222-1237.
34. Haeggström JZ, Funk CD (2011) Lipoxygenase and leukotriene pathways: biochemistry, biology, and roles in disease. Chem Rev 111(10): 5866-5898.
35. Ylikorkala O, Viinikka L (1992) The role of prostaglandins in obstetrical disorders. Baillieres Clin Obstet Gynaecol 6(4): 809-827.
36. Jacobs L, Nawrot TS, de Geus B, Meeusen R, Degraeuwe B, et al. (2010) Subclinical responses in healthy cyclists briefly exposed to traffic-related air pollution: an intervention study. Environ Health 9: 64.
37. Witko-Sarsat V, Rieu P, Descamps-Latscha B, Lesavre P, Halbwachs-Mecarelli L (2000) Neutrophils: Molecules, Functions and Pathophysiological Aspects Lab Invest 80(5): 617-653.
38. Crooks, SW, Stockley RA (1998) Molecular Focus – Leukotriene B4. Int J Biochem Cell Biol 30: 173-178.
39. Loke WM, Lam KM, Chong WL, Chew SE, Quek AM, et al. (2012) Products of 5-lipoxygenase and myeloperoxidase activities are increased in young male cigarette smokers. Free Radic Res [In Press].
40. Lantz RC, Hays AM (2006) Role of oxidative stress in arsenic-induced toxicity. Drug Metab Rev 38(4): 791-804.
41. Bose M, Whittal RM, Gairola CG, Bose HS (2008) Cigarette smoke decreases mitochondrial porin expression and steroidogenesis. Toxicol Appl Pharmacol 227(2): 284-290.
42. Fowler PA, Cassie S, Rhind SM, Brewer MJ, Collinson JM, et al. (2008) Maternal smoking during pregnancy specifically reduces human fetal desert hedgehog gene expression during testis development. J Clin Endocrinol Metab 93(2): 619-626.
43. Moledina F, Clarke G, Oskooei A, Onishi K, Günther A, et al. (2012) Predictive microfluidic control of regulatory ligand trajectories in individual pluripotent cells. Proc Natl Acad Sci U S A 109(9): 3264-3269.
44. Loureiro B, Oliveira LJ, Favoreto MG, Hansen PJ (2011) Colony-stimulating factor 2 inhibits induction of apoptosis in the bovine preimplantation embryo. Am J Reprod Immunol 65(6): 578-588.
45. Kim YJ, Lee GS, Hyun SH, Ka HH, Choi KC, et al. (2009) Uterine Expression of Epidermal Growth Factor Family During the Course of Pregnancy in Pigs. Reprod Domest Anim 44(5): 797-804.
46. Shilo BZ (2005) Regulating the dynamics of EGF receptor signaling in space and time. Development 132(18): 4017-4027.
47. Greenwald I (2012) Notch and the awesome power of genetics. Genetics 191(3): 655-669.
48. McCright B, Gao X, Shen L, Lozier J, Lan Y, et al. (2001) Defects in development of the kidney, heart and eye vasculature in mice homozygous for a hypomorphic Notch2 mutation. Development 128(4): 491-502.
49. Piasek M, Blanusa M, Kostial K, Laskey JW (2001) Placental cadmium and progesterone concentrations in cigarette smokers. Reprod Toxicol 15(6): 673-81.
50. Gallego MJ, Porayette P, Kaltcheva MM, Bowen RL, Vadakkadath Meethal S, et al. (2010) The pregnancy hormones human chorionic gonadotropin and progesterone

induce human embryonic stem cell proliferation and differentiation into neuroectodermal rosettes. Stem Cell Res Ther 1(4): 28.
51. Lee K, Jeong J, Tsai MJ, Tsai S, Lydon JP, et al. (2006) Molecular mechanisms involved in progesterone receptor regulation of uterine function. J Steroid Biochem Mol Biol 102(1-5): 41-50.
52. Pliarchopoulou K, Voutsinas G, Papaxoinis G, Florou K, Skondra M, et al. (2012) Correlation of CYP1A1, GSTP1 and GSTM1 gene polymorphisms and lung cancer risk among smokers. Oncol Lett 3(6): 1301-1306.
53. Delpretti S, Zakany J, Duboule D (2012) A function for all posterior Hoxd genes during digit development? Dev Dyn 241(4): 792-802.
54. Singh AB, Harris RC (2005) Autocrine, paracrine and juxtacrine signaling by EGFR ligands. Cell Signal 17(10): 1183-1193.
55. Jeanclos EM, Lin L, Treuil MW, Rao J, DeCoster MA, et al. (2001) The chaperone protein 14-3-3eta interacts with the nicotinic acetylcholine receptor alpha 4 subunit. Evidence for a dynamic role in subunit stabilization. J Biol Chem 276(30): 28281-28290.
56. Sakumi K, Abolhassani N, Behmanesh M, Iyama T, Tsuchimoto D, et al. (2010) ITPA protein, an enzyme that eliminates deaminated purine nucleoside triphosphates in cells. Mutat Res 703(1): 43-50.
57. Abolhassani N, Iyama T, Tsuchimoto D, Sakumi K, Ohno M, et al. (2010) NUDT16 and ITPA play a dual protective role in maintaining chromosome stability and cell growth by eliminating dIDP/IDP and dITP/ITP from nucleotide pools in mammals. Nucleic Acids Res 38(9): 2891-2903.
58. Goggin M, Sangaraju D, Walker VE, Wickliffe J, Swenberg JA, et al. (2011) Persistence and repair of bifunctional DNA adducts in tissues of laboratory animals exposed to 1,3-butadiene by inhalation. Chem Res Toxicol 24(6): 809-817.
59. Koturbash I, Scherhag A, Sorrentino J, Sexton K, Bodnar W, et al. (2011) Epigenetic mechanisms of mouse interstrain variability in genotoxicity of the environmental toxicant 1,3-butadiene. Toxicol Sci 122(2): 448-456.
60. Izzotti A, Cartiglia C, Longobardi M, Bagnasco M, Merello A, et al. (2004) Gene expression in the lung of p53 mutant mice exposed to cigarette smoke. Cancer Res 64(23): 8566-8572.
61. Kamp DW, Shacter E, Weitzman SA (2011) Chronic inflammation and cancer: the role of the mitochondria. Oncology (Williston Park) 25(5): 400-410.
62. Kauffman SL (1969) Cell proliferation in embryonic mouse neural tube following urethane exposure. Dev Biol 20(2): 146-157.
63. Sharova LV, Sharov AA, Sura P, Gogal RM, Smith BJ, et al. (2003) Maternal immune stimulation reduces both placental morphologic damage and down-regulated placental growth-factor and cell cycle gene expression caused by urethane - are these events related to reduced teratogenesis. Int Immunopharmacol 3(7): 945-955.
64. Murphy SK, Adigun A, Huang Z, Overcash F, Wang F, et al. (2012) Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. Gene 494(1): 36-43.
65. Perkins E, Murphy SK, Murtha AP, Schildkraut J, Jirtle RL, et al. (2012) Insulin-like growth factor 2/h19 methylation at birth and risk of overweight and obesity in children. J Pediatr 161(1): 31-39.
66. Hussain M, Rao M, Humphries AE, Hong JA, Liu F, et al. (2009) Tobacco Smoke Induces Polycomb-Mediated Repression of Dickkopf-1 in Lung Cancer Cells. Cancer Res 69(8): 3570-3578.

67. Matarazzo MR, De Bonis ML, Strazzullo M, Cerase A, Ferraro M, et al. (2007) Multiple binding of methyl-CpG and polycomb proteins in long-term gene silencing events. J Cell Physiol 210(3): 711-719.
68. Ma L, Zheng LW, Sham MH, Cheung LK (2010) Uncoupled Angiogenesis and Osteogenesis in Nicotine-Compromised Bone Healing. J Bone Miner Res 26(6): 1305-1313.
69. Lopatina N, Haskell JF, Andrews LG, Poole JC, Saldanha S, et al. (2002) Differential maintenance and de novo methylating activity by three DNA methyltransferases in aging and immortalized fibroblasts. J Cell Biochem 84(2): 324-334.
70. Xu Y, Zhang Y, Cardell LO (2010) Nicotine enhances murine airway contractile responses to kinin receptor agonists via activation of JNK- and PDE4-related intracellular pathways. Respir Res 11: 13.
71. León Y, Sanchez JA, Miner C, Ariza-McNaughton L, Represa JJ, et al. (1995) Developmental regulation of Fos-protein during proliferative growth of the otic vesicle and its relation to differentiation induced by retinoic acid. Dev Biol 167(1): 75-86.
72. Ichino N, Ishiguro H, Yamada K, Nishii K, Sawada H, et al. (1999) Nicotine withdrawal up-regulates c-Fos transcription in pheomochromocytoma cells. Neurosci Res 35(1): 63-69.
73. Cucina A, Sapienza P, Borrelli V, Corvino V, Foresi G, et al. (2000) Nicotine Reorganizes Cytoskeleton of Vascular Endothelial Cell through Platelet-Derived Growth Factor BB. J Surg Res 92(2): 233-238.
74. Srinivas H, Juroske DM, Kalyankrishna S, Cody D, Price RE, et al. (2005) c-Jun N-terminal kinase contributes to aberrant retinoid signaling in lung cancer cells by phosphorylating and inducing proteasomal degradation of retinoic acid receptor alpha. Mol Cell Biol 25(3): 1054-1069.
75. Nakayama H, Numakawa T, Ikeuchi T, Hatanaka H (2001) Nicotine-induced phosphorylation of extracellular signal-regulated protein kinase and CREB in PC12h cells.J Neurochem 79(3): 489-498.
76. Tang K, Wu H, Mahata SK, O'Connor DT (1998) A crucial role for the mitogen-activated protein kinase pathway in nicotinic cholinergic signaling to secretory protein transcription in pheochromocytoma cells. Mol Pharmacol 54(1): 59-69.
77. ZHAO Z, REECE EA (2005) NICOTINE-INDUCED EMBRYONIC MALFORMATIONS MEDIATED BY APOPTOSIS FROM INCREASING INTRACELLULAR CALCIUM AND OXIDATIVE STRESS. BIRTH DEFECTS RES B DEV REPROD TOXICOL 74(5): 383-391.
78. Dietze EC, Caldwell LE, Marcom K, Collins SJ, Yee L, et al. (2002) Retinoids and retinoic acid receptors regulate growth arrest and apoptosis in human mammary epithelial cells and modulate expression of CBP/p300. Microsc Res Tech 59(1): 23-40.
79. Han BC, Xia HF, Sun J, Yang Y, Peng JP (2010) Retinoic acid-metabolizing enzyme cytochrome P450 26a1 (cyp26a1) is essential for implantation - functional study of its role in early pregnancy. J Cell Physiol 223(2): 471-479.
80. Osanai M, Lee GH (2011) Nicotine-mediated suppression of the retinoic acid metabolizing enzyme CYP26A1 limits the oncogenic potential of breast cancer. Cancer Sci 102(6): 1158-1163.
81. Feltes BC, de Faria Poloni J, Bonatto D (2011) The developmental aging and origins of health and disease hypotheses explained by different protein networks. Biogerontology 12(4): 293-308.

82. David P, Sabapathy K, Hoffmann O, Idarraga MH, Wagner EF (2002) JNK1 modulates osteoclastogenesis through both c-Jun phosphorylation-dependent and -independent mechanisms. J Cell Sci 115(22): 4317-4325.
83. Yildiz D (2004) Nicotine, its metabolism and an overview of its biological effects. Toxicon 43: 619-632.
84. Nakayama Y, Mezawa M, Araki S, Sasaki Y, Wang S, et al. (2009) Nicotine suppresses bone sialoprotein gene expression. J Periodont Res 44: 657-663.
85. ALBERT R (2005) SCALE-FREE NETWORKS IN CELL BIOLOGY. J CELL SCI 118(21): 4947-4957.

Figure Legends

Figure 1: A binary network of chemical-protein and protein-protein interactions (CPI-PPI network) generated by the program Cytoscape 2.8.2. (A) The main network, showing 49 known substances present in tobacco, 1177 nodes (49 substances, 1128 proteins) and 7522 edges (connections). Proteins were colored to identify the tissue in which they were present: (i) pink indicates fetal tissue; (ii) green, embryonic tissue; and (iii) dark blue, both fetal and embryonic tissues. In addition, each substance was colored according to its solubility: (i) yellow indicates lipophilic and (ii) light blue, hydrophilic. We observed that nicotine resided in a module apart from the major network (A). Therefore, we separated it from the major CPI-PPI network and colored its module purple. (B) The nicotine subnetwork is shown separately, from the major CPI-PPI network. It contained proteins related to retinoic acid signaling and retinoic acid (lipophilic molecule). (C) The final major CPI-PPI network after the nicotine module was extracted.

Figure 2: HBs found in the major CPI-PPI network. Betweenness and node degrees were assessed using the program CentiScaPe. Among the 143 HBs, 53 proteins are present in both tissues, reinforcing the idea of a prolonged effect of CS on embryonic development.

Figure 3: Cluster analysis of the major CPI-PPI network indicating clusters 1, 4, 11, 16 and 20. Cluster 1 (A) is composed of 83 nodes and 565 edges, with C*i* = 6,843. The associated hydrophilic constituents are urethane, *N*-nitrosoanabasine, *N*-methylpyrrolidine and pyridine. The lipophilic constituents are toluene, 4-aminobiphenyl, 5-methylcrysene, benz[a]anthracene, chrysene, benzo[b]fluoranthene, 7H-dibenzo[c,g]carbazole and 2-naphthylamine. Related GO terms: oxidation reduction and unsaturated fatty acid metabolic processes. Cluster 4 (B) is composed of 90 nodes and 411 edges, with C*i* = 4,567. The associated hydrophilic compounds are urethane, *N*-nitrosoanabasine, *N*-methylpyrrolidine, arsenic, selenium and cadmium, and the lipophilic compounds are acrolein, crotonaldehyde, toluene, xylene, ethylbenzene, benz[a]anthracene, styrene and 2-naphthylamine. Related GO term: oxidation reduction. Cluster 11 (C) is composed of 23 nodes and 74 edges, with C*i* = 3,217. Only the lipophilic compound isoprene is present in this cluster. Related GO term: steroid biosynthetic processes. Cluster 16 (D) is composed of 15 nodes and 36 edges, with C*i* = 2,400. The associated hydrophilic compound is butyraldehyde. Related GO term: lipid modification. Cluster 20 (E) is composed of 29 nodes and 65 edges, with C*i* = 2,241. The associated lipophilic compounds are acrolein, 2-naphthylamine, 1,3-butadiene, cyclopentane and 4-aminobiphenyl. Related GO term: prostaglandin metabolic processes and unsaturated fatty acid metabolic processes. A merge of clusters 1, 4 and 20 (F). Clusters 11 and 16 did not show any proteins overlapping protein with any other cluster.

Figure 4: Cluster analysis of the major CPI-PPI network and the modules related to cell-cell signaling. Cluster 2 (A) is composed of 73 nodes and 354 edges, with C*i* = 4,849. Cluster 2 contains the two hydrophilic substances cadmium and methylamine. Related GO term: regulation of cell communication. Cluster 18 (B) is composed of 13 nodes and 30 edges, with C*i* = 2, 304. Cluster 18 contains one hydrophilic compound, trimethylamine, and one lipophilic compound, ethylbenzene. Related GO term: Cell-cell signaling.

Figure 5: A merge of clusters 3, 6, 17 and 21. In (A), cluster 3 is composed of 14 nodes and 65 edges, with C*i* = 4,643. The associated hydrophilic components are urethane, beryllium and polonium-210. Related GO terms: RNA-splicing and nucleobase, nucleoside, nucleotide and nucleic acid metabolic processes. Cluster 6 (B) is composed of 24 nodes and 102 edges, with C*i* = 4,250. The associated lipophilic constituents are 1,3-butadiene and chrysene. Related GO term: nucleobase, nucleoside and nucleotide metabolic processes. Cluster 17 (C) is composed of 35 nodes and 83 edges, with C*i* = 2,371. The hydrophilic constituents present include catechol and arsenic. Related GO term: Regulation of nucleobase, nucleoside nucleotide and nucleic acid metabolic process. Cluster 21 (D) is composed of 22 nodes and 49 edges, with C*i* = 2,227. The associated hydrophilic constituent is beryllium, and the lipophilic constituent is crotonaldehyde. Related GO term: regulation of DNA metabolic processes. The union of clusters 3, 17 and 21 (E). Cluster 6 did not show proteins overlapping with the others.

Figure 6: Subnetworks derived from the merge of clusters 5, 8 and 9. In (A), Cluster 5 is composed of 69 nodes and 315 edges, with C*i* = 4,565. The associated hydrophilic components are vinyl acetate and ethylamine. Related GO terms: response to DNA-damage stimulus and cell cycle. Cluster 7 (B) is composed of 13 nodes and 54 edges with C*i* = 4,154. The associated hydrophilic compound is dimethylamine. Related GO term: chromatin organization. Cluster 8 (C) is composed of 85 nodes and 338 edges with C*i* = 3,976. The associated lipophilic constituents are acrolein and benzo[b]fluoranthene, whereas the hydrophilic constituents are urethane and arsenic. Related GO terms: DNA-damage stimulus and regulation of cell cycle. Cluster 9 (D) is composed of 16 nodes and 58 edges, with C*i* = 3,625. The hydrophilic constituent present is trimethylamine. Related GO term: cell cycle processes. The union of clusters 5, 8 and 9 (F). Clusters 7 did not show any proteins overlapping with any others cluster.

Figure 7: A model of the interaction from a systemic view showing how TCs affect development. In (A), we show that increasing TC levels generate a pro-inflammatory cascade by increasing the levels of PTGS1 and PTGS2. PTGSs are associated with inflammatory responses and are essential for normal pregnancy. Disturbances in PTGS expression could cause impairments in fetal development. Prostaglandins also generate arachidonic acid, a precursor of leukotriene. TCs are connected to ALOX5 and ALOX15B, which are proteins involved in the synthesis of leukotriene, a molecule that plays pivotal roles in pro-inflammatory responses. The consequence of (A) is low birth weight in newborn infants, abortions and increased proliferation. Moreover, in (B), TCs are linked to BING2 and USP2, which are proteins related to increased activity of MDM2. This MDM2 mediated up-regulation can rapidly down-regulate p53 protein, leaving the cell more susceptible to DNA damage. TCs also down-regulate HPRT1, diminishing purine metabolism. This system exhibit a relationship with increased proliferation. The system in (C) shows that TCs are associated with the generation of superoxides due to up-regulating NADH oxidase, which increases ROS levels and consequently oxidative stress. Increased oxidative stress is known to be related to birth defects. In addition, system (C) is associated with low birth weight and low neutrophil activity. Moreover, (D) shows the relationship between TCs and low hormone synthesis

and signaling. Exposure to TCs could have a negative effect on androgen and estrogen solubility due to acting on the UGT cluster. TCs could also be associated with low levels of cholesterol synthesis due to increasing the levels of CYPs and diminishing the levels of FDFT1 and FDPS, which are two enzymes related to cholesterol synthesis. Low cholesterol availability would decrease general hormone synthesis. In addition, TCs affect the transport of cholesterol to the mitochondria by acting on the membrane protein StAR. Finally, in (E), the system shows the relationship between TCs and decreased global gene expression and cellular differentiation and signaling. The activities of the TCs would increase the levels of MBD2, a methylation enzyme. DNA methylation is related to gene silencing. We postulate that TCs could affect the PRC2 complex via its methylation and disturb gene expression, including that of HOX genes. TCs could also have a negative effect on gene expression by increasing YWHAH levels, which would decrease the levels of the master kinase PDPK1 and is linked to AKT activation and SMAD nuclear translocation. In addition, NOTCH signaling could be affected through the action of TCs on APP activation.

Figure 8: A binary network of the interactions between chemical compounds and proteins generated by the program Cytoscape 2.8.2, which contained 330 proteins and 4078 connections. Nicotine appears in the network as the green node, and RA appears as the blue node. White nodes are connected to both compounds are proteins. A) A sub-network containing 49 nodes and 281 edges and showing the proteins with direct connections with nicotine. B) A sub-network containing 130 nodes and 1471 edges and showing the proteins that make direct connections with RA.

Figure 9: A molecular model illustrating how nicotine affects differentiation in the embryo. In the first part of the model (I), it can be observed that by the generating cellular stress, nicotine promotes the recruitment of JNK1 and TR3 through the influx of intracellular Ca^{2+}. JNK1 promotes the inhibition of RARα. Moreover, the Ca^{2+} influx leads to the recruitment of TR3, which inhibits COUP-TF and prevents the expression of RARβ. Through the recruitment of FOS and PFGFs by nicotine, an increase in cell proliferation occurs, which causes a decrease in differentiation and RA signaling. Finally, nicotine promotes the inhibition of CYP26A1, which generates an accumulation of RA in the cell and an increase in cell proliferation. In the second part of the model (II), the inhibition of BMP2 and BSP is promoted by nicotine, which results in the negative regulation of bone mineralization and skeletal development. In addition, nicotine also promotes a proinflammatory reaction that recruits VEGF and placental growth factors, which leads to an impairment of the endovascular trophoblast and results in a fetal growth restriction.

Table 1: Major bioprocesses associated with the Hubs-bottlenecks subnetwork.

GO-ID	GO	*p*-value	Corr *p*-value	*k**	n#	Proteins
55114	Oxidation Reduction	4.4×10^{-16}	9×10^{-14}	26	645	CYP3A5;CYP1B1;PTGS2;CYP2C19;CYP2B6;PGD;PTGS1;ALDH3A2;AKR1C3;GSR;GPX1;FMO1;GPX4;HMOX1;GPX3;CAT;NQO1;HADH;CYP1A1;CYP2C8;MAOB;IDO1;CYP2E1;CYP1A2;DECR1;SOD2;LDLCQ3;ALDH7A1;G6PD;AKR1B1;TXN;ALDH2;MPO
48545	Regulation of Steroid Hormone Stimuli	1.3×10^{-13}	1.8×10^{-11}	18	225	TNF;PTGS2;MAP2K1;RELA;PTGS1;MAOB;BRCA1;MAPK1;FOS;CCND1;CDKN1A;EP300;HMOX1;GPX4;GPX3;ALDH2;INSR;NGF
42127	Regulation of Cell Proliferation	1.3×10^{-12}	1.5×10^{-10}	30	848	CSF2;TNF;GNAI2;PTGS2;PTGS1;TAC1;GPX1;INS;HMOX1;IFNG;SHC1;EGF;INSR;MYC;EGFR;KAT2B;IL8;MAP2K1;RELA;TP53;IDO1;STAT1;MBD2;BRCA1;SOD2;MAPK1;CDKN1A;CCND1;IL2;NGF
42981	Regulation of Apoptosis	9.3×10^{-12}	9.4×10^{-10}	29	282	CSF2;TNF;PTGS2;MMP9;GPX1;APP;INS;ALB;HMOX1;SOS1;IFNG;CAT;NQO1;EGFR;RELA;GRIN2A;TP53;IDO1;STAT1;BRCA1;SOD2;MAPK1;CDKN1A;F2;MPO;PDCD6;GSTP1;IL2;NGF
10646	Regulation of Cell Communication	6×10^{-10}	2.6×10^{-8}	31	1154	CSF2;TNF;GNAI2;PTGS2;CD8A;GRB2;TAC1;GPX1;APP;INS;SOS1;HMOX1;IFNG;CHRNA4;SHC1;CAT;EGF;INSR;AGAP2;EGFR;MAP2K1;RELA;MAOB;GRIN2A;TP53;MBD2;PTPN11;LAP3;CCND1;IL2;NGF

* = Number of nodes for a given GO in the network;

= Total number of proteins for a given GO annotation.

Table 2: The relationships between of common proteins, nicotine and RA and their specific biochemical functions. These data were obtained from the GeneCards (http://www.genecards.org) and iHop (http://www.ihop-net.org/UniPub/iHOP/) databases.

Protein	Biological Function	Role
VEGFA	Growth factor	Crucial role in angiogenesis, vasculogenesis and endothelial growth
TGFB1	Cytokine	Acts in differentiation, proliferation, adhesion, and migration; also a potent stimulator of bone growth
ALPP	Alkaline phosphatase	Expressed in the placenta, which is involved in circulation during pregnancy
HSF	Transcription factor	Activated under conditions of heat or other cellular stresses
FGF2	Growth factor	Involved in tumor growth, development of the nervous system, cell differentiation and angiogenesis
TH	Hydroxylase	Hydroxylase that function in the physiology of adrenergic neurons
FOS	Nuclear phosphoprotein	Nuclear phosphoprotein that participates in cell differentiation, proliferation and apoptosis
IL10	Cytokine	Involved in the immune response against pathogens and in the inflammatory response; also related to the intestinal immune system
DNMT1	Methyltransferase	DNA methylation and the establishment of methylation patterns
IL1	Cytokine	Involved in the immune response to against pathogens and the inflammatory response
AKT1	Kinase	Involved in tumor formation, angiogenesis and insulin regulation
BAX	Transcription factor	Pro-apoptotic protein
ALPL	Alkaline phosphatase	Mineralization of bone matrix
BCL1 (IL5)	Cytokine	Involved in the immune response against pathogens and the inflammatory response
NOS2A	Nitric oxide synthase	Produces nitric oxide (NO)
PPARG	Proliferator peroxisome receptor	Regulator of adipocyte differentiation and glucose homeostasis
K60 (IL8)	Chemokine	Involved in the inflammatory response; angiogenesis inducer

Table 2 *Continued.*

Proteins	Biological Function	Role
CREB1	Transcription factor	Controls circadian rhythm, acts as tumor suppressor and the expression of various genes involved in cell survival
PLAU	Protease	Involved in extracellular matrix degradation and possibly tumorigenesis
IL2	Cytokine	Essential in the proliferation of T-cells of the immune system. Stimulates the production of B-cells, monocytes and natural killer cells
KDR (VEGFR)	Growth factor	Plays a crucial role in vasculogenesis and angiogenesis
RARB	Retinoic Acid Receptor	Involved in cell differentiation, cell growth arrest, and signaling and transcription of target genes

Figure 1: CPI network of 49 compounds of TC developed by the program Cytoscape.

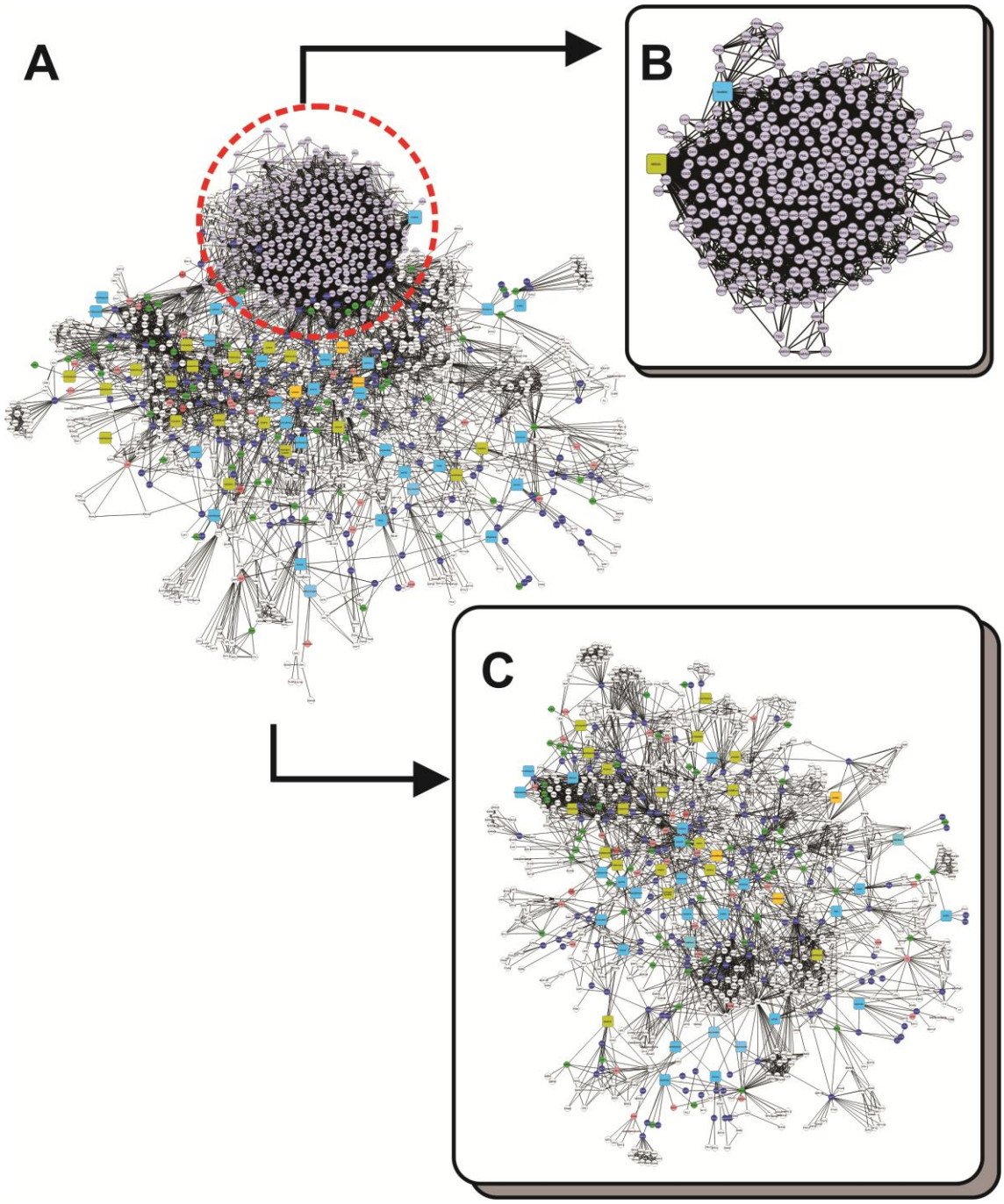

A
B
C

Figure 2: Hubs-bottlenecks present in the main network.

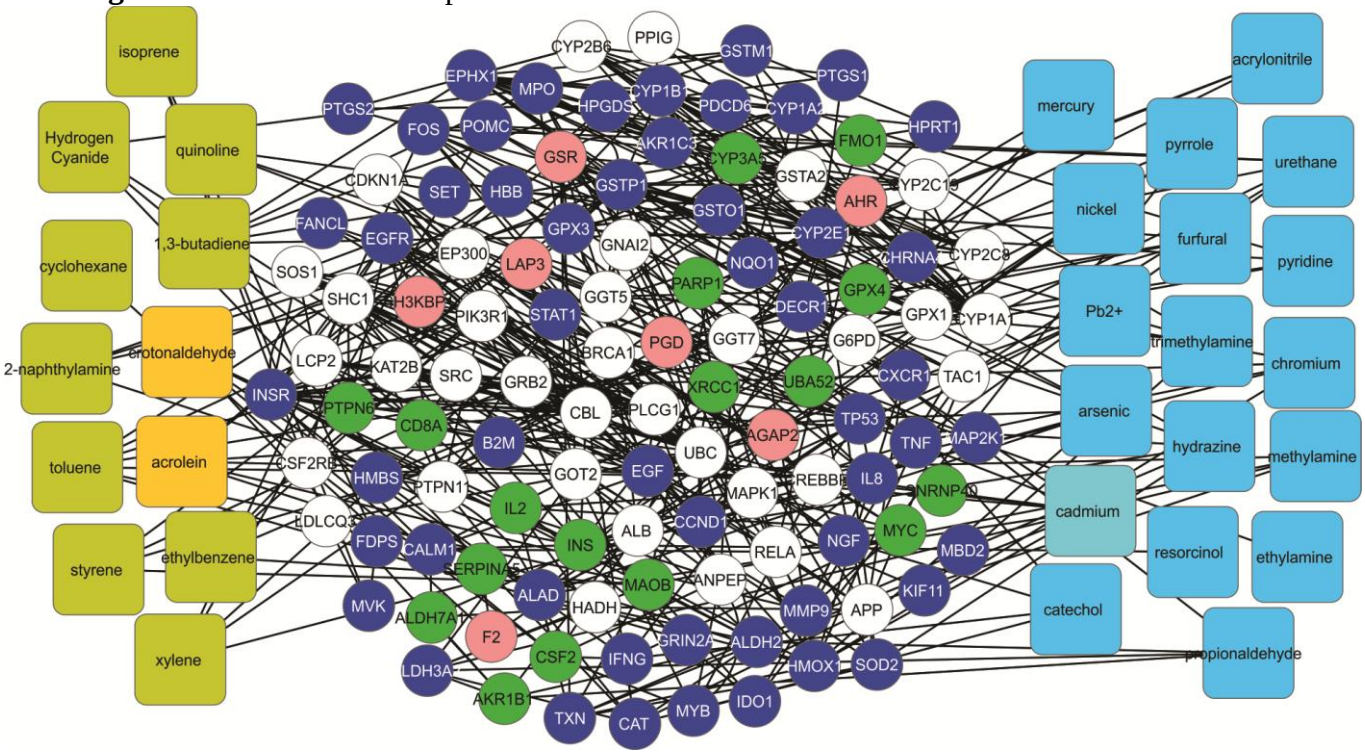


Figure 3: Union of clusters 1, 4, 11, 16 and 20.

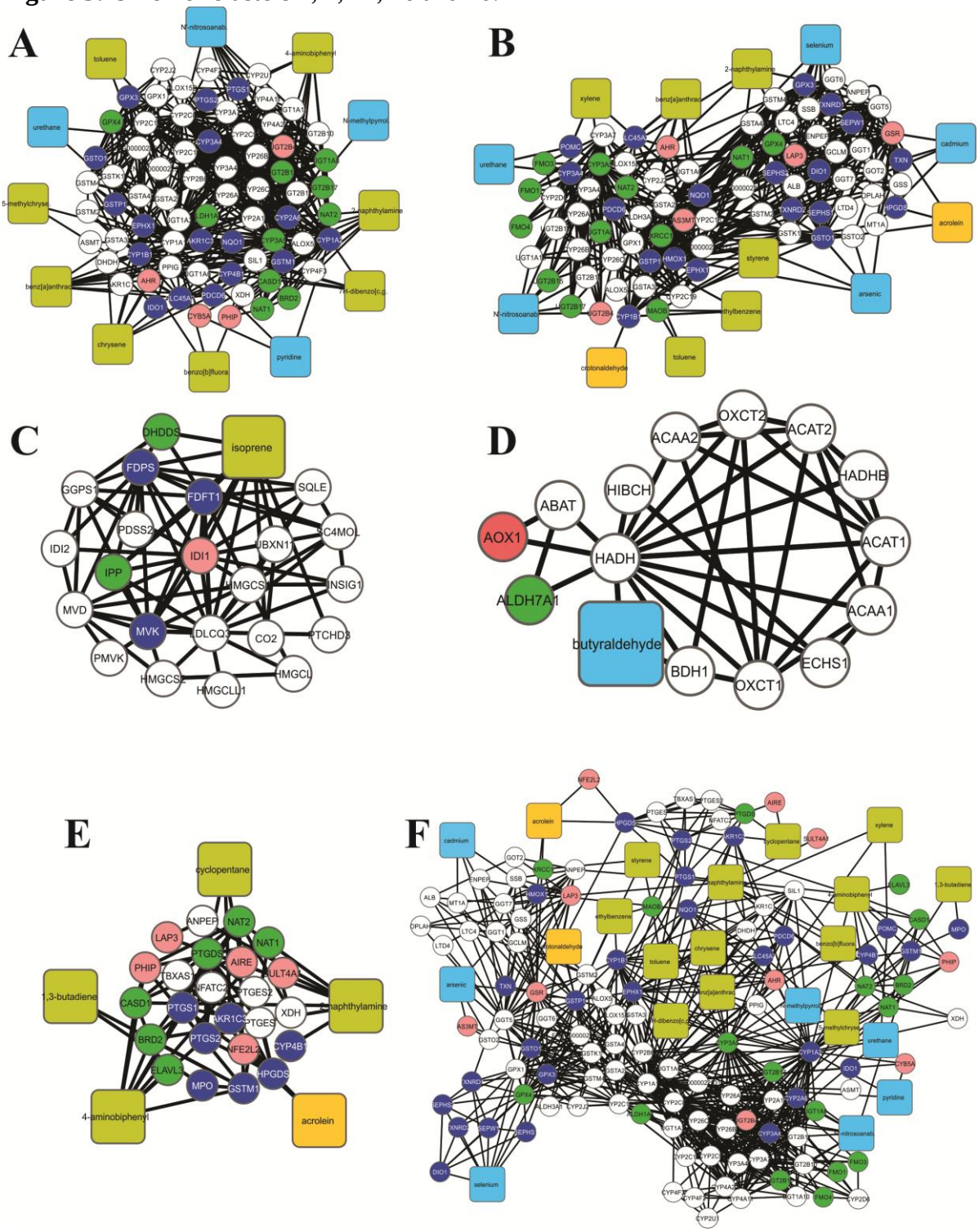


Figure 4: Cluster 2 and 18.

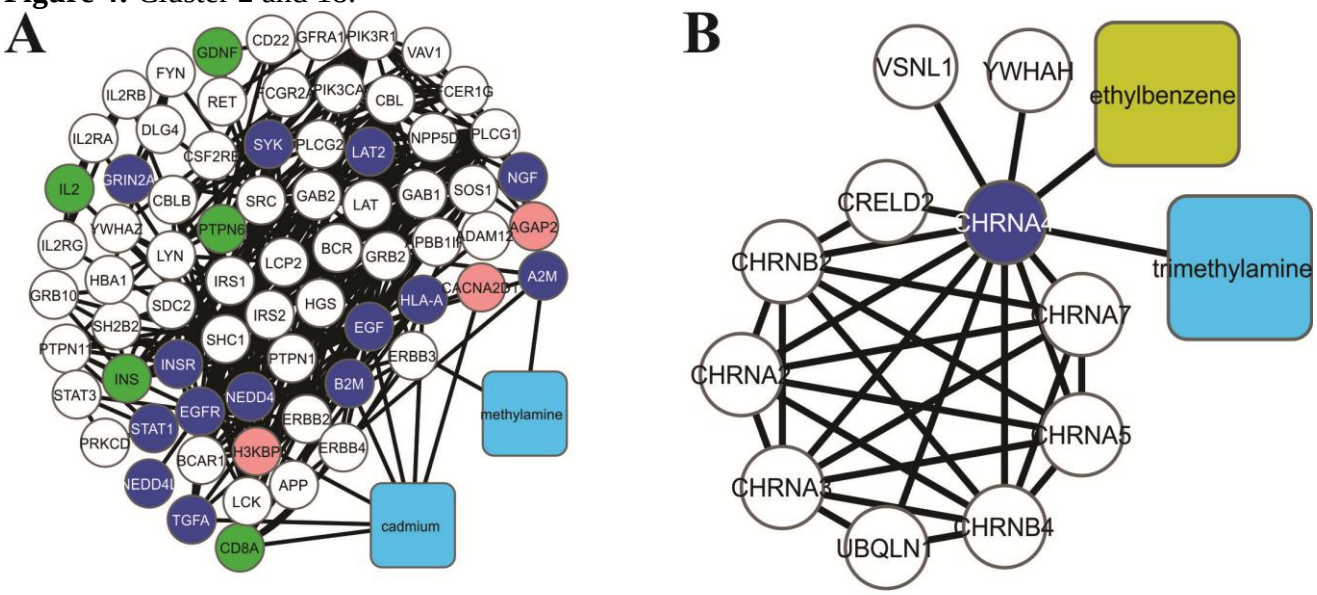


Figure 5: Clusters 3, 6, 17 and 21.

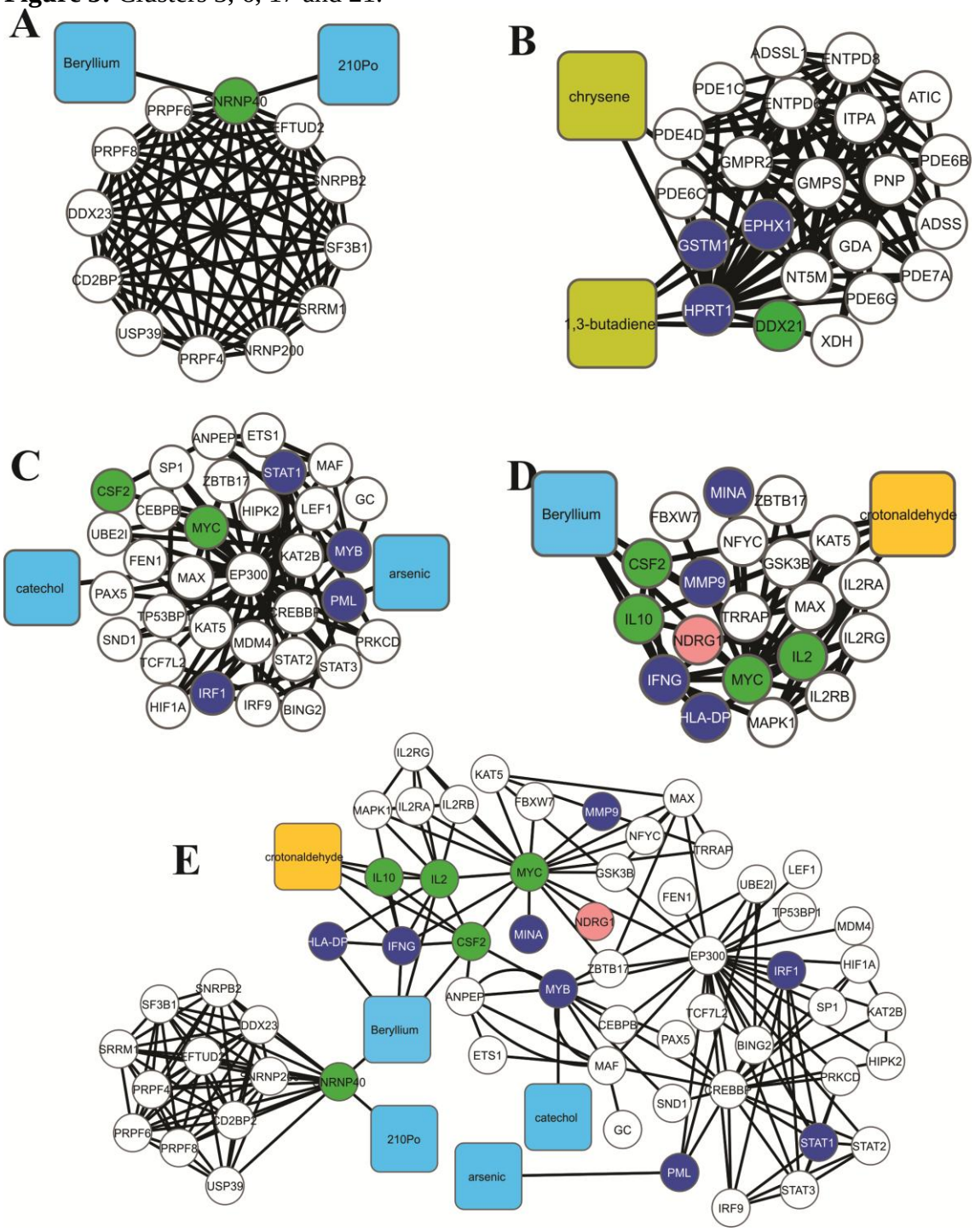


Figure 6: Union of clusters 5, 7, 8 and 9

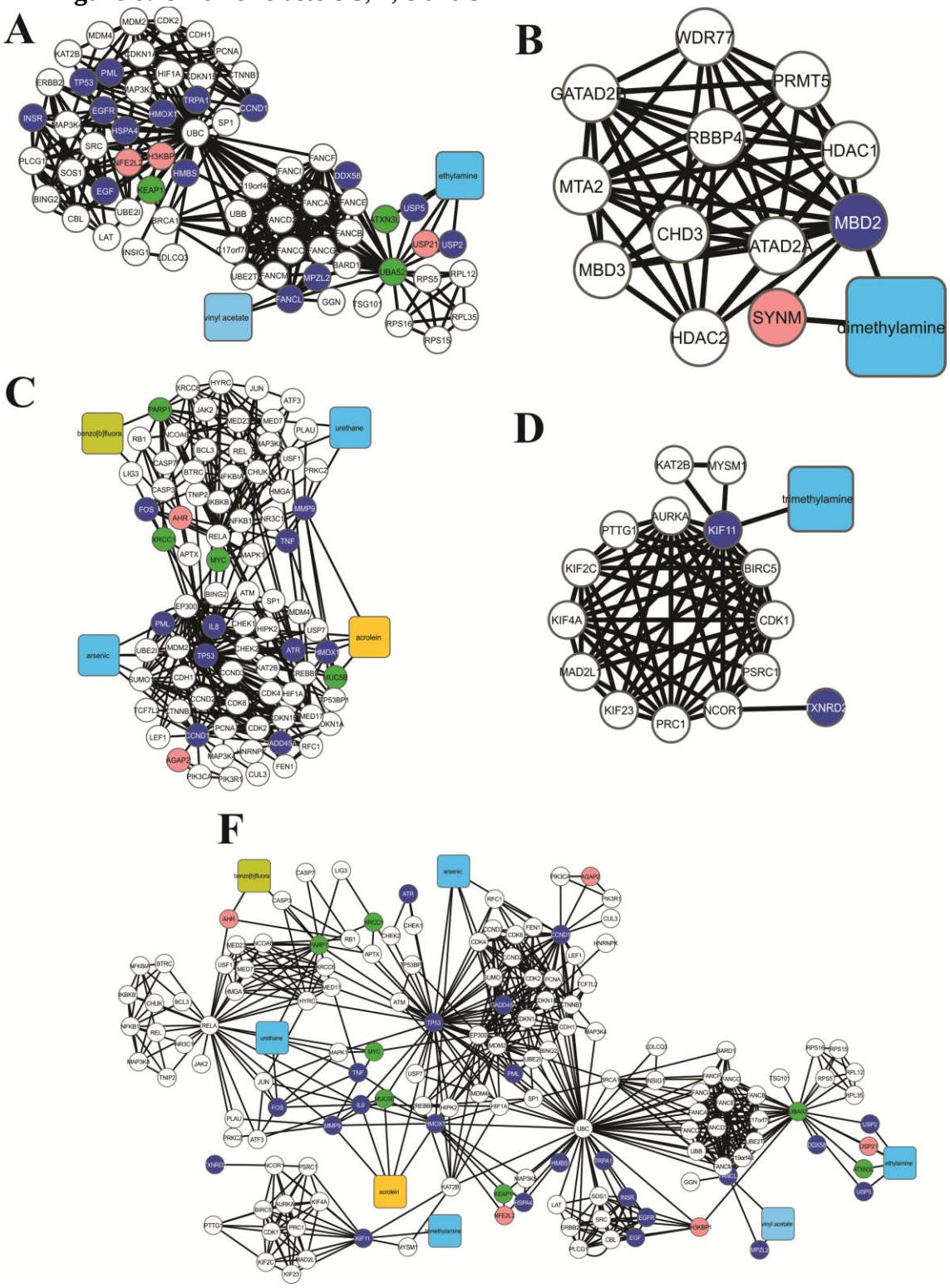


Figure 7: Model of interaction of how cigarette compounds affect the developmental process.

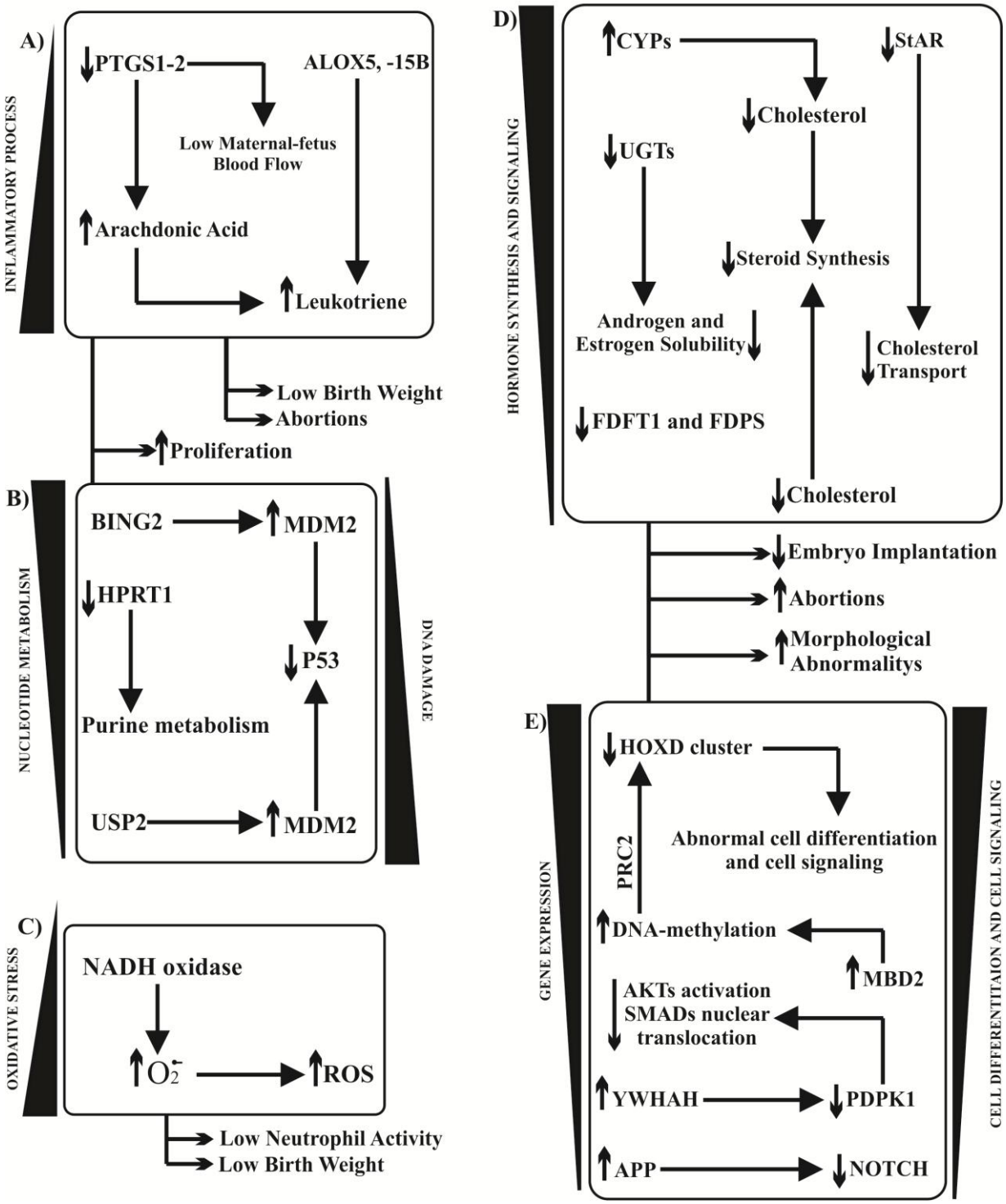


Figure 8: Main network of nicotine, with two sub-networks.

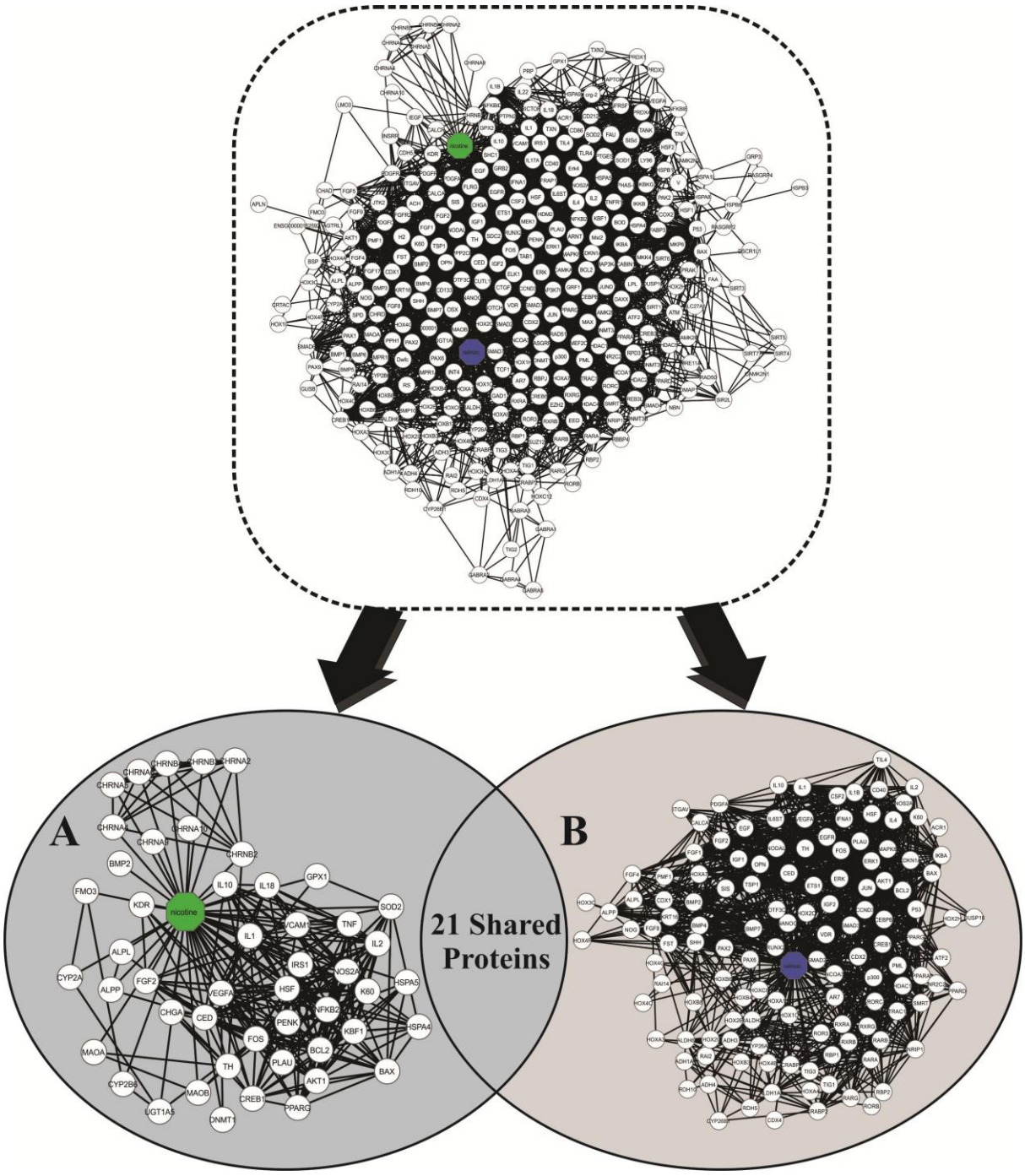
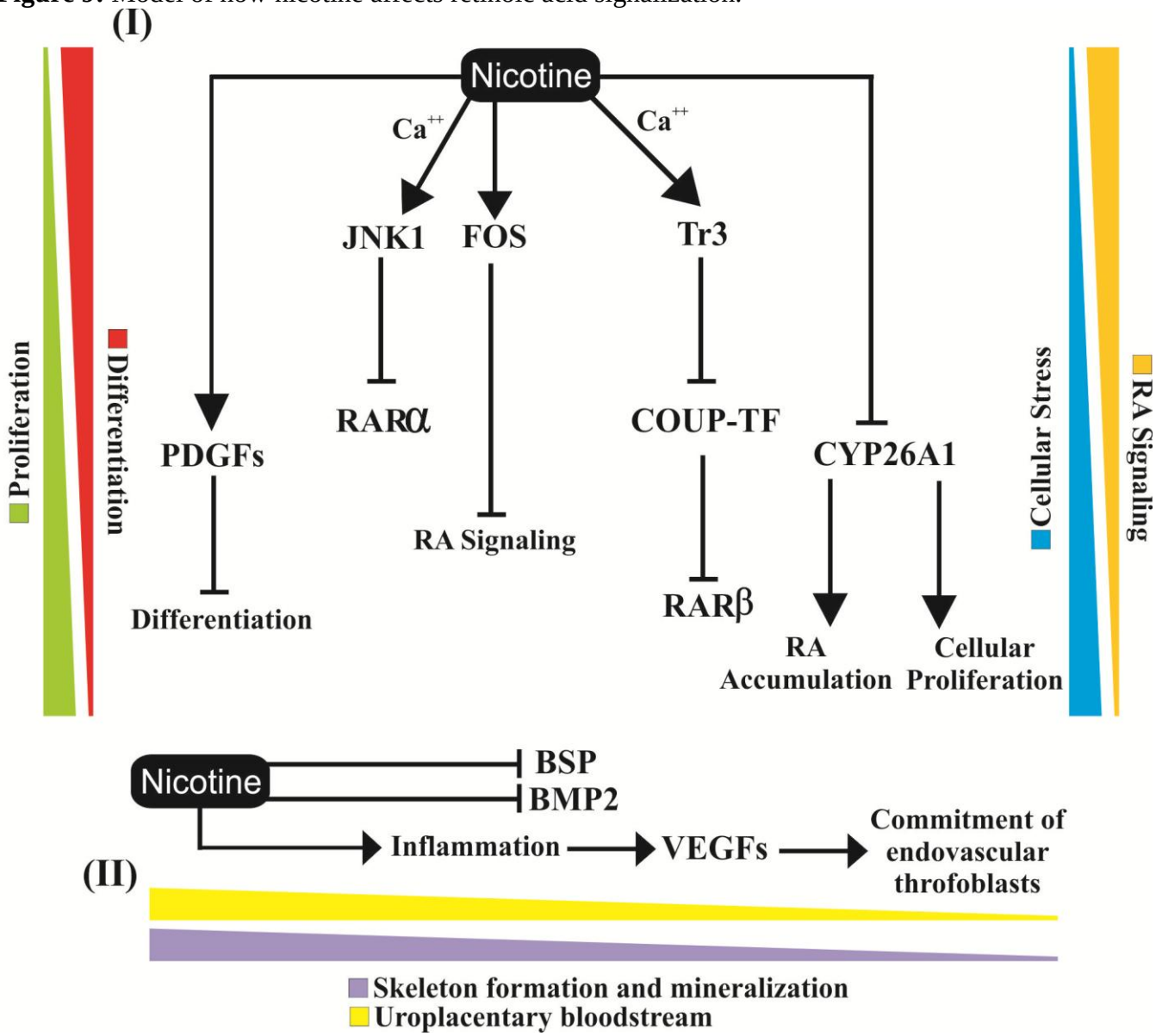


Figure 9: Model of how nicotine affects retinoic acid signalization.



ANEXO 2 - PORTAL INTERGENICDB

A2. PORTAL INTERGENICDB

O objetivo do IntergenicDB é ser um portal de consultas para regiões promotoras (Davanzo, 2010; Picolloto, 2012) e, para armazenar as informações deste portal, foi desenvolvido um banco de dados, chamado PromotoresDB que contêm as informações de sequências promotoras de DNA de organismos procariontes (Molin, 2009; Avila e Silva, 2011).

A motivação para o desenvolvimento deste portal é que, atualmente, não há um banco de dados disponível na comunidade científica que atenda aos requisitos solicitados pelos pesquisadores do núcleo de pesquisa de Bioinformática da Universidade de Caxias do Sul (Avila e Silva, 2011). Os requisitos conceituais do banco de dados PromotoresDB são (Molin, 2009; Davanzo, 2010; Avila e Silva, 2011): i) vincular uma região promotora a um organismo; ii) obter informações sobre os genes; iii) armazenar informações de uma região intergênica; iv) descobrir e armazenar as informações sobre os métodos de predição utilizados na predição de sequências promotoras; e, v) catalogar as publicações (artigos, trabalhos, etc.) relacionadas ao estudo de promotores. As próximas seções apresentam o banco de dados PromotoresDB e as funcionalidades do portal IntergenicDB.

A2.1 BANCO DE DADOS PROMOTORESDB

Os requisitos iniciais do banco de dados foram coletados por Molin (2009) a partir de pesquisas em artigos científicos e, em bancos de dados existem a respeito deste assunto junto ao CMR[64] (*Comprehensive Microbial Resourse*), EcoCyc[65], bancos de dados do NCBI e o RegulonDB. O CMR é um banco de dados que contém anotações robustas de todos os genomas microbianos e fornece informações sobre regiões

[64] http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi

[65] http://ecocyc.org/

específicas (genes, promotores e regiões intergênicas; Peterson et al., 2001). O EcoCyc é um banco de dados científico com diferentes categorias de informações sobre a bactéria *Escherichia coli* (Keseler et al. 2011). Os bancos de dados do NCBI e o RegulonDB foram explicados na seção 3.2 da Revisão Bibliográfica Geral. As informações armazenadas no banco de dados PromotoresDB são as seguintes (Molin, 2009):

• Cada região promotora está ligada a um organismo, que possui um nome, um reino, uma família e um tipo de molécula;

• Cada gene possui um nome, um símbolo, um número de identificação de início e fim, uma função e uma quantidade percentual de CG.

• Cada região intergênica possui um número para a posição inicial e final na sequência, um tamanho, sua sequência de nucleotídeos e o tipo de fita que pertence.

• A região intergênica de teste é uma região que contém 60 nucleotídeos e é uma fonte de dados usada para predição que define se a sequência é ou não candidata a ser promotora.

• Um método preditor é um algoritmo usado na predição das chances de uma região ser ou não promotor, e possui um nome, um percentual de uma cadeia analisada e uma descrição de suas características de funcionamento.

• Uma publicação é um documento relacionado à área de estudo de promotores, que possui um autor, um título, uma data de publicação, um instituto ao qual está vinculado, uma cidade e um país, um endereço *web* e um espaço para informações adicionais.

Um banco de dados precisa ter um modelo conceitual para a criação da sua estrutura física de armazenamento (Heuser, 2008; Elmari e Navathe, 2004). Na ciência da computação utiliza-se a técnica de modelagem conceitual para representar o conjunto

de dados a serem armazenados para um *software* (Heuser, 2008; Elmari e Navathe, 2004). O Diagrama Entidade-Relacionamento (DER) é utilizado como modelo conceitual para representar a estrutura do banco de dados de forma visual (Heuser, 2008; Elmari e Navathe, 2004). O DER possui um conjunto de entidades que representam objetos do mundo real e, cada entidade possui um conjunto de atributos que representam as características de uma entidade (Heuser, 2008; Elmari e Navathe, 2004). O modelo conceitual do banco de dados PromotoresDB revisado é mostrado na Figura 1 (Molin, 2009; Davanzo, 2010; Avila e Silva, 2011; Picolloto, 2012).

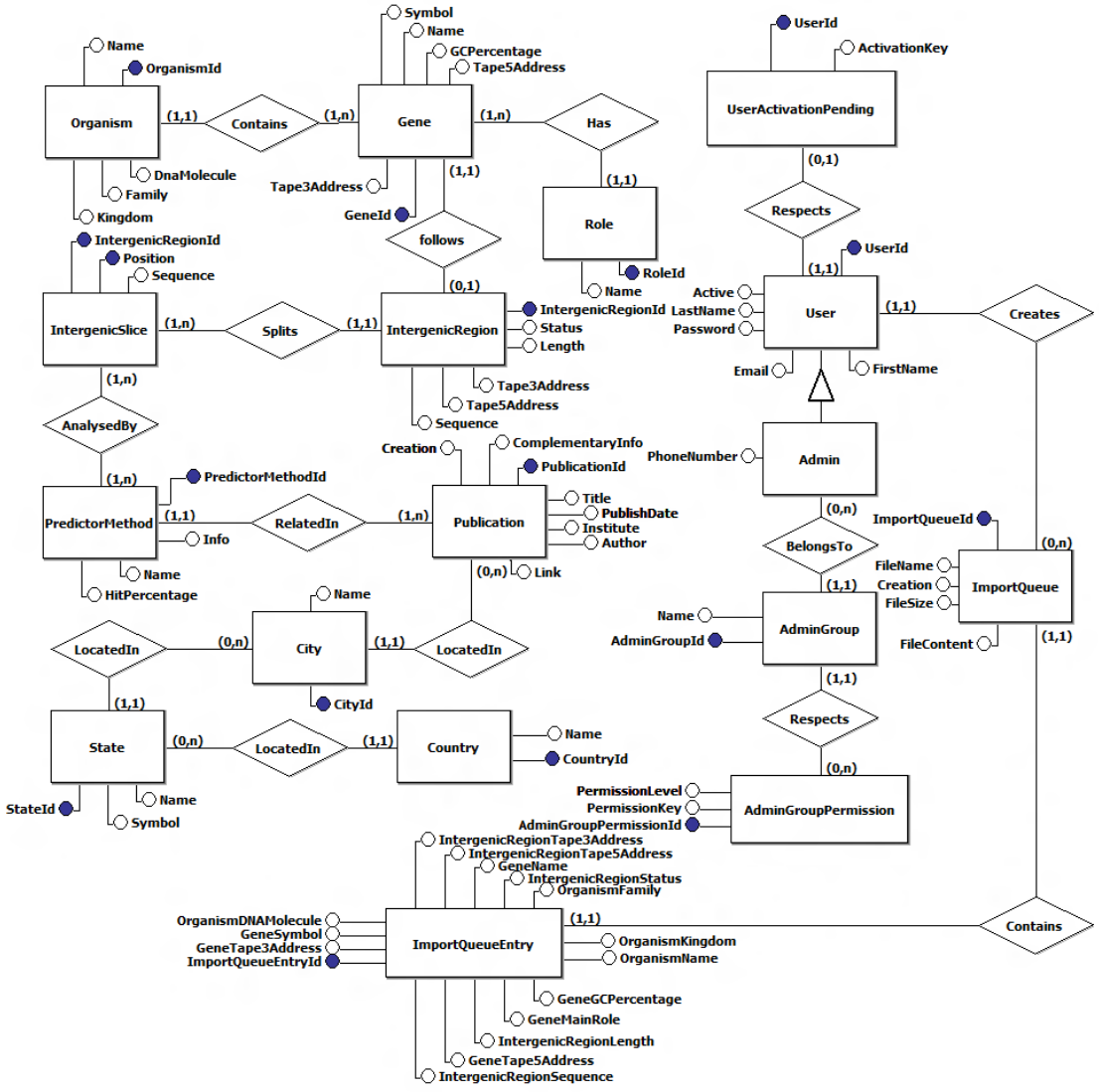


Figura 1: Modelo conceitual do banco de dados PromotoresDB que representa as informações a serem consultadas e atualizadas no Portal IntergenicDB.

As entidades definidas para o armazenamento das informações do banco de dados PromotoresDB são (Molin, 2009; Davanzo, 2010; Avila e Silva, 2011; Picolloto, 2012):

- *Organism*: informações dos organismos biológicos.
- *Gene*: informações dos genes para cada organismo biológico.
- *Role*: informações sobre que papel o gene desempenha.
- *IntergenicRegion* regiões intergênicas de cada genes.
- *IntergenicSlice*: pedaços de regiões intergênicas dos genes.
- *PredictorMethod*: métodos de predição utilizados para obter os dados das regiões intergênicas.
- *Publication*: informações sobre as publicações referentes aos métodos de predição utilizados para obter as regiões intergênicas.
- *City*, *State* e *Country*: informações da cidade, estado e país onde foram publicados os métodos de predição utilizados.
- *User* e *Admin*: informações sobre os usuários e administradores do portal.
- *AdminGroup*, *AdmimGroupPermission* e *UserActivationPending* : informações sobre as permissões de cada usuário no portal.
- *ImportQueue* e *ImportQueueEntry*: informações temporárias sobre novas informações de predição de genes que serão incorporadas ao IntergenicDB após validação de um usuário administrador.

A2.3 FUNCIONALIDADES DO PORTAL

Os usuários do portal são classificados em administradores e usuários comuns sendo que é necessária uma autenticação de um usuário para ter acesso aos serviços do portal, conforme pode ser visto na Figura 2.

Os administradores terão acesso a todas as funcionalidades do portal e, com acesso exclusivo para cadastrar novos usuários e permitir a inserção de novos dados no banco de dados PromotoresDB; enquanto que o usuário comum poderá realizar todas as consultas disponíveis e solicitar a inserção de novos dados (Picolloto, 2012).

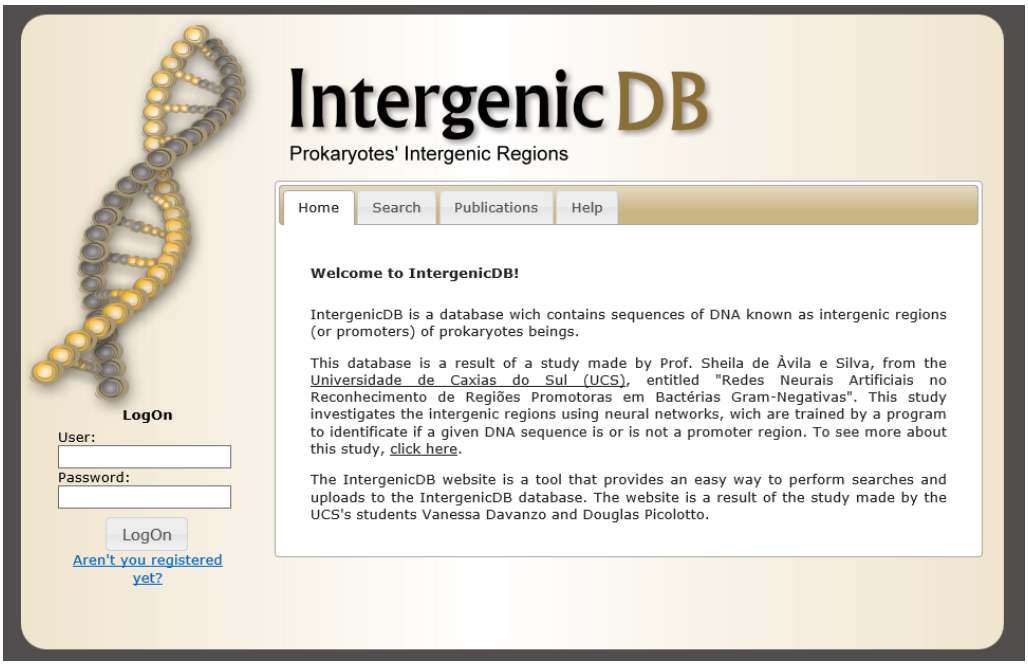


Figura 2: Tela inicial do Portal IntergenicDB para fazer autenticação de usuários.

A Figura 3 mostra a tela com as funcionalidades que podem ser manipuladas pelo usuário administrador. Na primeira parte mostra o acesso para manipular os usuários e grupos de usuários. A segunda parte mostra o acesso para manipular as informações de biologia molecular do banco de dados. E, a terceira parte permite ao administrador manipular as informações de localização do usuário, publicações a respeito das informações inseridas no banco de dados e controlar as solicitações de inserção de novos dados.

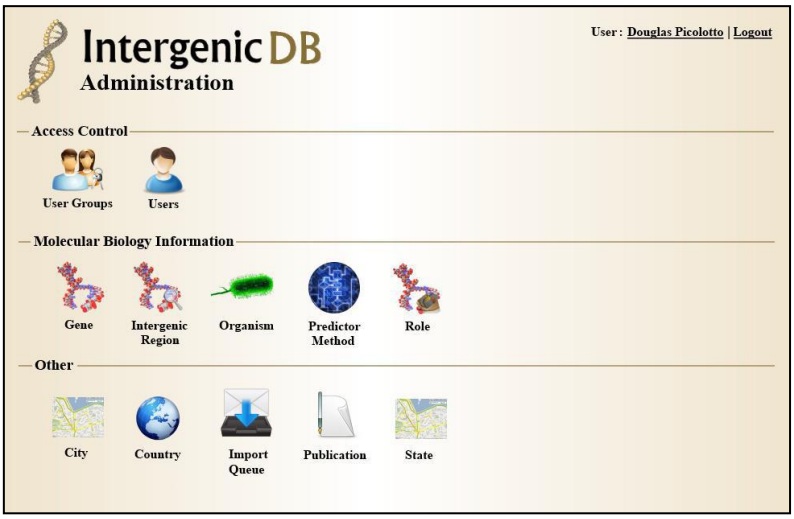


Figura 3: Tela com as funcionalidades do administrador do Portal IntergenicDB.

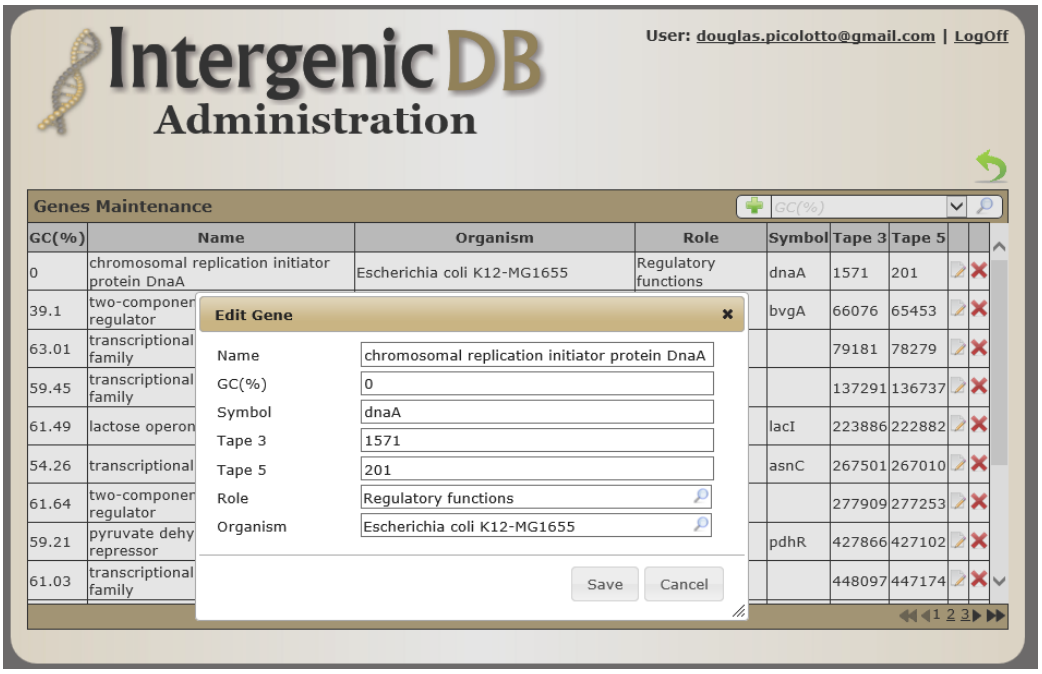


Figura 4: As telas para cadastro de informações no portal IntergenicDB

A funcionalidade de manutenção dos dados é de responsabilidade dos usuários administradores. Esta funcionalidade permite atualizar todos os dados armazenados no banco de dados mostrado na Figura 1. As telas de cadastro para todas as informações do portal são semelhantes, por isto, a Figura 4 mostra a tela de cadastro de informações de genes. O administrador insere um gene usando a tela de edição (tela em destaque) com as informações de GC, nome, organismo, papel, símbolo e informações da posição nas

fitas. Na parte de cima à direita, há um botão (+) para inserção de um novo gene e uma caixa de texto para pesquisa. Além disto, ao final de cada linha há um botão para edição e um botão para exclusão de um gene.

Um usuário poderá solicitar a inserção de novos dados no portal. Isto será feito usando um modo supervisionado onde os dados importados serão armazenados em uma base de dados temporária para não comprometer a integridade e veracidade dos dados existentes. O processo de importação é dividido em quatro etapas: a) etapa de *upload* onde o usuário envia o arquivo a ser importado para análise; b): etapa de análise onde um usuário responsável analisa os dados importados; c) etapa de importação onde o usuário responsável submete os dados importados ao banco de dados através da tela de administração do IntergenicDB; e, d) etapa de notificação onde o sistema notifica o usuário que submeteu o arquivo se a importação foi concluída com sucesso ou não.

A principal funcionalidade do portal é permitir a realização de consultas (Figura 5). O usuário pode consultar utilizando como elementos de pesquisa: organismos, genes, regiões intergênicas e métodos de predição. O usuário pode consultar cada uma das opções individualmente ou combinadas. O portal terá, também, uma área de postagem portal onde possam ser visualizados documentos relacionados a promotores através da aba “News”. A Figura 6 mostra um exemplo de uma consulta realizada no portal para a bactéria Escherichia coli.

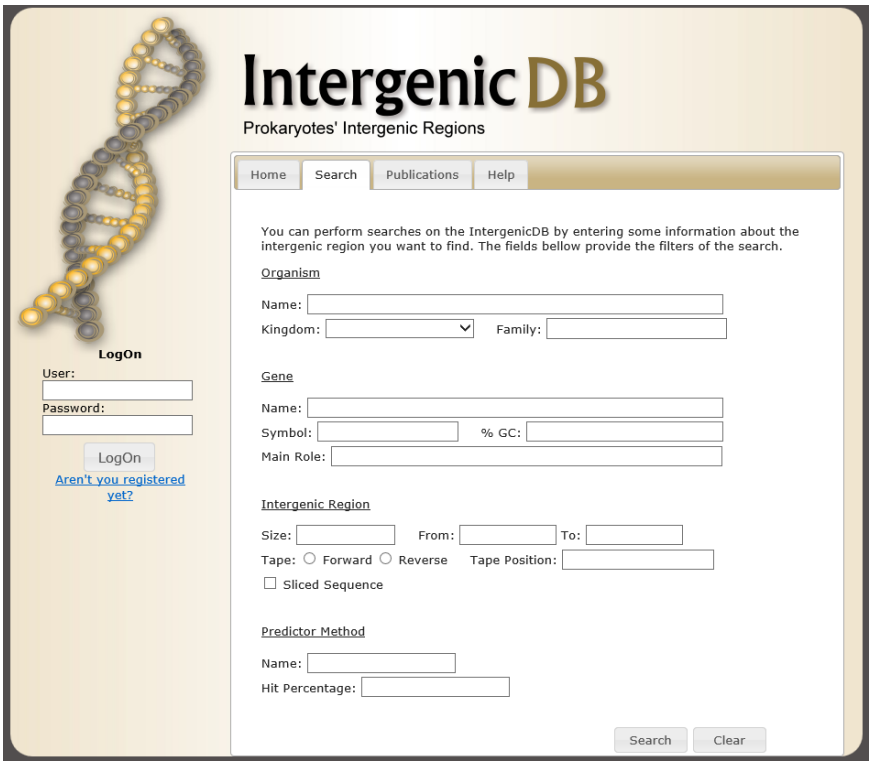


Figura 5: Tela de consulta do portal com as opções de consulta sobre organismos, genes, regiões intergênicas e métodos de predição.

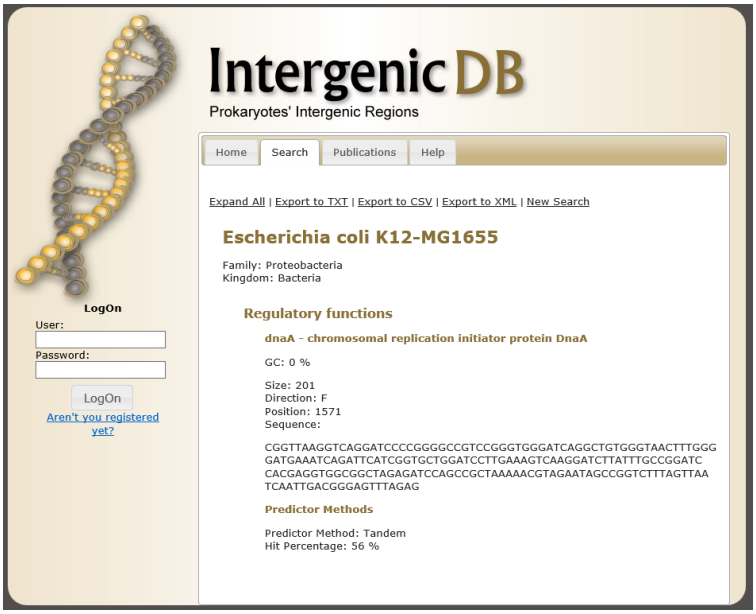


Figura 6: Resultado de uma consulta realizada no portal para a bactéria *Escherichia coli*.

ANEXO 3

DESENVOLVIMENTO DE FERRAMENTAS DE ACESSO A BANCOS DE DADOS BIOLÓGICOS

A3. DESENVOLVIMENTO DE FERRAMENTAS DE ACESSO A BANCOS DE DADOS BIOLÓGICOS

A construção de ferramentas auxiliares foi necessária para a execução dos *workflows* científicos Dis2PPI e Net2Homology, visando acessar os bancos de dados e ferramentas de bioinformática. As ferramentas construídas foram as seguintes:

i) Carregar uma rede de interação de proteínas no formato do banco de dados STRING;

ii) Criar mecanismos de consulta ao banco de dados STRING para obter redes de interação de proteínas na forma de arquivos textos e imagens, bem como, obter identificadores, descrições e a sequência de aminoácidos (formato FASTA) de proteínas;

iii) Criar mecanismos de consulta para acesso ao banco de dados PDB para a busca de imagens tridimensionais de proteínas;

iv) Desenvolver uma interface de execução para o algoritmo de alinhamento de sequências BLAST do NCBI;

v) Prospecção de dados de doenças monogênicas para a análise de homologia;

vi) Algoritmo para extração de informações do Psi-Blast e geração de redes ppi usando o banco de dados string;

vii) Configuração do sistema para uso dos workflows; e,

viii) Registro dos comandos executados pela ferramenta (*log*) e avisos aos usuários.

As próximas seções apresentam informações sobre o uso e implementação destas ferramentas.

A3.1 CARREGAR UMA REDE PPI

Para a execução do *workflow* Net2Homology é necessário fornecer como entrada uma rede de interação de proteínas. Uma forma de obter esta rede é com o resultado da execução do workflow Di2PPI. Outra forma é carregar uma rede de interação de proteínas (Figura 1) a partir de um arquivo texto (1A). No momento que esta opção é selecionada, o usuário é informado sobre o formato da estrutura do arquivo a ser lido (1B). Além disto, o usuário poderá as proteínas carregadas (1C) para fazer consultas aos bancos de dados STRING e PDB, bem como, para realizar um alinhamento de sequências.

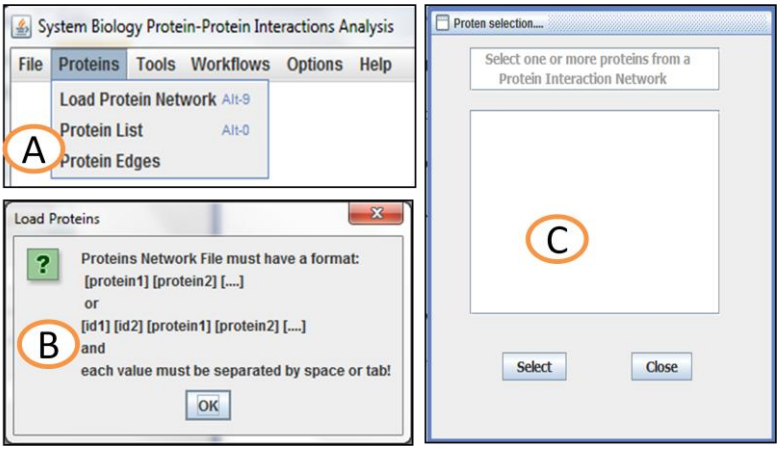


Figura 1: Carregar uma rede de interação de proteínas para o *workflow* Net2Homology.

A3.2 ACESSO AOS BANCOS DE DADOS E ALINHAMENTO DE SEQUÊNCIAS

Para o entendimento dos acessos aos bancos de dados STRING e PDB, bem como, a execução do alinhamento de proteínas usando o BLAST foram implementados uma interface de consulta para cada uma destas funcionalidades.

A3.2.1 CONSULTA AO BANCO DE DADOS STRING

A consulta ao banco de dados STRING (Figura 2) é feita através da informação de uma ou mais proteínas (uma por linha; 2A) ou através do uso da tela de seleção (2D) apresentada na Figura 1C. É necessário informar ainda a espécie (ou organismo, 2B) e o formato do arquivo de retorno (2C). Após a informação dos dados necessários, o usuário executa a consulta (2E) e, o resultado é mostrado no navegador *web* configurado na ferramenta.

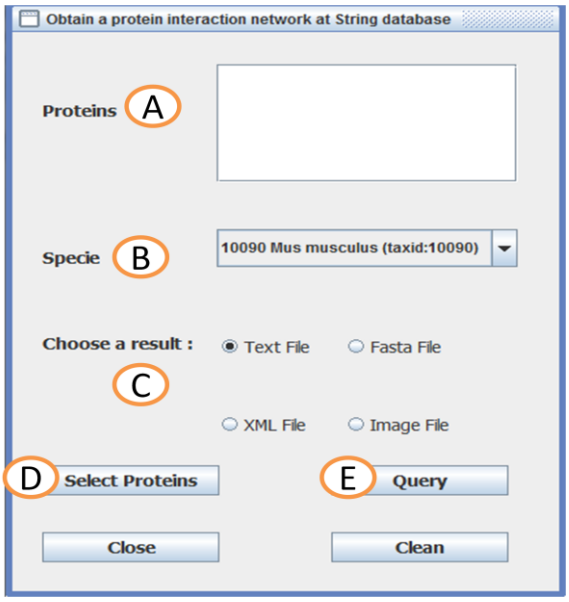


Figura 2: Tela de consulta ao banco de dados.

Para o acesso ao banco de dados STRING é montado uma URL para executar a consulta como, por exemplo:

http://string-db.org/api/psi-mi-tab/interactionsList?identifiers=10090.ENSMUSP00000020268.

Na URL é informado o *site* do banco de dados (string-db.org), o tipo de acesso (*api*), o tipo de consulta (neste caso, buscar uma rede no formato de arquivo texto: psi-mi-tab) e a lista de proteínas (o código do STRING para a proteína P53 é ENSMUSP00000020268) e a sua espécie (10090 é o código para a espécie *Mus musculus*).

A3.2.2 CONSULTA AO BANCO DE DADOS PDB

A consulta ao banco de dados PDB (Figura 3) é feita através da informação de uma proteína (não aceita mais de uma proteína; 3A) ou através do uso da tela de seleção (3B) apresentada na Figura 1C. Após a informação dos dados necessários, o usuário executa a consulta (3C) e o resultado (uma ou mais imagens da estrutura tridimensional de uma proteína) é mostrado no navegador *web* configurado na ferramenta. O acesso ao banco de dados PDB é feito através de um serviço (*web service*) fornecido pelo próprio site.

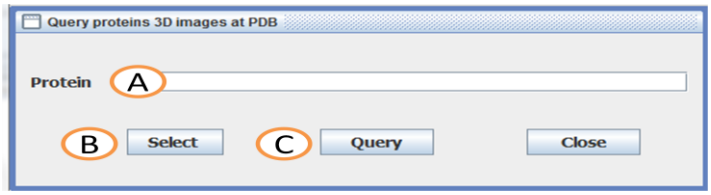


Figura 3: Tela de consulta ao banco de dados.

A3.2.3 ALINHAMENTO DE SEQUÊNCIAS DE PROTEÍNAS USANDO BLAST

A execução de uma consulta usando o algoritmo BLAST do NCBI é mostrada na Figura 4.

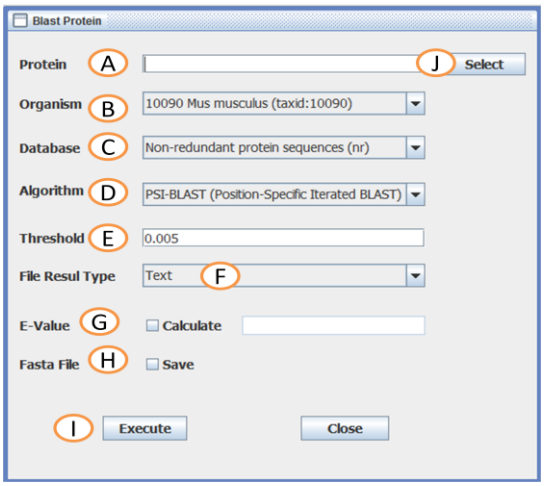


Figura 4: Tela de execução de um alinhamento de sequências usando o NCBI PSI-BLAST do NCBI.

Inicialmente é necessário informar uma proteína (4A) ou escolher através da tela de seleção (4J) apresentada na Figura 1C. Além da proteína, é necessário informar o

organismo (4B), o banco de dados do NCBI a ser usado (4C), o algoritmo de alinhamento de sequências (4D), o valor do limite aceitável (4E) e, o tipo do arquivo do resultado (4F). Após a informação dos dados necessários, o usuário executa a consulta (4I), o resultado do valor do *e-value* é mostrado na tela (4G) e, se a opção de salvar as informações do arquivo FASTA (4H) estiver marcada, as informações são salvas na pasta configurada. O resultado do alinhamento de sequências é mostrado no navegador *web* configurado na ferramenta. O acesso ao algoritmo BLAST é feito usando as ferramentas do Entrez do NCBI explicadas nas seções 2.3.1 a 2.3.3.

A3.3 PROSPECÇÃO DE DADOS DE DOENÇAS MONOGÊNICAS PARA A ANÁLISE DE HOMOLOGIA

A integração entre os *workflows* desenvolvidos foi feita para permitir que uma rede PPI resultante do *workflow* Dis2PPI fosse utilizada como dados de entrada para o *workflow* Net2Homology. A Figura 5 mostra a tela, acrescentada ao *workflow* Dis2PPI, com as proteínas (5A) extraídas da rede PPI resultante da pesquisa de uma doença gênica. Na sequência, o usuário pode selecionar todas as proteínas (5B) ou marcar as proteínas de seu interesse (5C). Para finalizar esta etapa, o usuário deve pressionar o botão (5D) para prosseguir, sendo que a execução do *workflow* seguirá a partir da tela inicial do *workflow* Net2Homology.

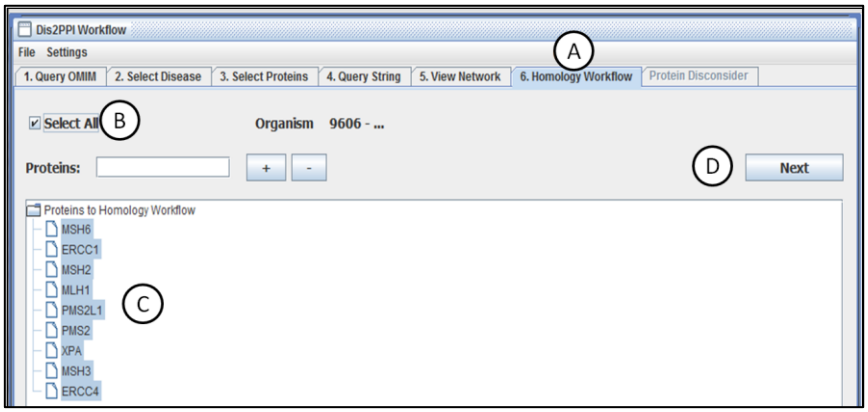


Figura 5: Tela de integração entre os workflows Dis2PPI e Net2Homology.

A3.4 ALGORITMO PARA EXTRAÇÃO DE INFORMAÇÕES DO PSI-BLAST E GERAÇÃO DE REDES PPI USANDO O BANCO DE DADOS STRING

O terceiro passo do *workflow* Net2Homology envolve preparar os dados para a análise de homologia. Para esta preparação dos dados, é necessário extrair informações do relatório de execução do alinhamento de proteínas feito com o Psi-Blast (Figura 6). As informações relevantes, para este trabalho, são o *e-value* (6A), a descrição de proteínas (6B) e a sequência de aminoácidos (formato FASTA) dos alinhamentos realizados (6C). Estas informações são utilizadas para obter duas redes PPI para os organismos definidos durante uma execução deste *workflow*.

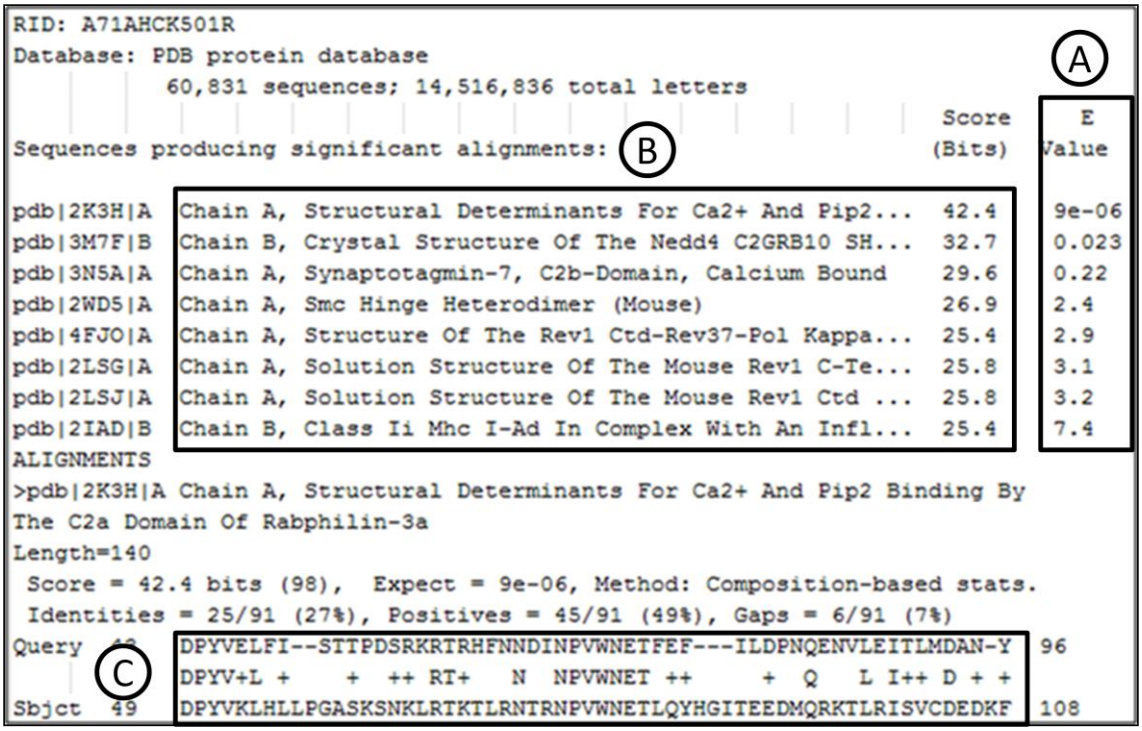


Figura 6: Relatório da execução do Psi-Blast para a proteína PLA2G4A e para o organismo *Mus musculus*.

Um algoritmo foi desenvolvido para extrair as informações do relatório do Psi-Blast e pesquisar as redes PPI no banco de dados STRING (Figura 7). A parte mais importante desta extração (7A) é encontrar a proteína correspondente a uma sequência fasta ou com a descrição apresentada no alinhamento. Este algoritmo utiliza três formas de encontrar a proteína alvo usando estas informações. A primeira forma utiliza a

sequência de aminoácidos (7C) para comparar com um banco de dados local de aminoácidos (7B).

```
input: a set of proteins, source and target organism
(A) for each protein selected {
        execute Psi-Blast for target organism
        extract e-value using first result
        extract fasta sequence using first result
        extract protein description using three results
}
(B) for each fasta sequence found {
        query fasta sequence to a local database (
        compare protein fasta sequence extracted to fasta sequence storage on local database
        if matches return protein name and jump to step (D)
 }
(C) for each protein blast result {
        for each protein description {
            extract words candidate to be a protein using regular expression
            query string database to know if exists this protein candidate
            if matches return protein name found
            else continue until has more words or finish when all descriptions were used
        }
       if failed {
              for each protein description {
                  query string database using entire description
                  if matches return protein name found
              }
        }
       if failed or protein description contains protein source name {
                  return protein name source
              }
        }
}
(D) for each source and target protein names {
        query string to obtain protein string identification (id)
}
(E) Query string database to obtain a PPI network from source and target organism using proteins ids
```

Figura 7: Algoritmo de extração das informações relevantes do relatório de execução do Psi-Blast.

Se a proteína não for encontrada, passa-se a utilizar a descrição das proteínas (Figura 6B). Neste caso, dois procedimentos são realizados conforme mostrado na Figura 7C. O primeiro tenta identificar se há uma palavra que pode ser uma proteína através do uso da expressão regular "([A-Z]{1}[A-Z0-9]{1,6})", onde o nome de uma proteína tem que começar por uma letra maiúscula e pode ter mais seis caracteres entre letras e números. Quando uma palavra candidata é encontrada, esta é testada junto ao

banco de dados STRING para verificar se é uma proteína válida para o organismo pesquisado. Se for, o programa retorna esta proteína, caso contrário, o programa continua a testar se há mais palavras candidatas. Se não for encontrada nenhuma proteína com este procedimento, o programa passa a testar a descrição completa da proteína (as três primeiras descrições do relatório são usadas) com uma consulta ao banco de dados STRING. Se obtiver sucesso, o nome da proteína é retornada. Se nenhuma pesquisa obtiver sucesso, a proteína do organismo original é retornada como resposta. Neste caso, o programa não conseguiu encontrar uma proteína equivalente no cruzamento de espécies realizado com o Psi-Blast. Esta proteína não será levada em consideração no teste a ser realizado pelo NetworkBlast.

Após cada uma das proteínas serem identificadas (7D), uma pesquisa é realizada no banco de dados STRING para obter o seu código de identificação (por exemplo, ENSP00000356425 para a proteína UCHL5 do organismo *Homo sapiens*). Estes códigos identificadores são utilizados para buscar as redes PPI para os organismos pesquisados (7E) no banco de dados STRING.

A3.5 CONFIGURAÇÕES DA FERRAMENTA

Para executar os *workflows* e as consultas aos bancos de dados biológicos é necessário fazer algumas configurações iniciais (Figura 8). O usuário precisa instalar dois *softwares* em sua máquina: um navegador *web* e o *software* Cytoscape. O navegador *web* será utilizado para mostrar os arquivos de texto e imagens consultados no banco de dados STRING, as imagens obtidas de consulta ao banco de dados PDB, os arquivos FASTA consultados do NCBI, o resultado de um alinhamento de sequências, bem como, irá mostrar os arquivos de configuração e o relatório de execução da ferramenta NetworkBlast do *workflow* Net2Homology. O *software* Cytoscape será

utilizado para mostrar a rede de interação de proteínas obtida ao final da execução do *workflows* Dis2PPi e Net2Homology.

O arquivo executável do navegador escolhido deve ser escolhido (8A), bem como, o arquivo executável do Cytoscape (8B). Para a seleção destes arquivos é aberta uma tela de navegação padrão para as pastas do sistema operacional.

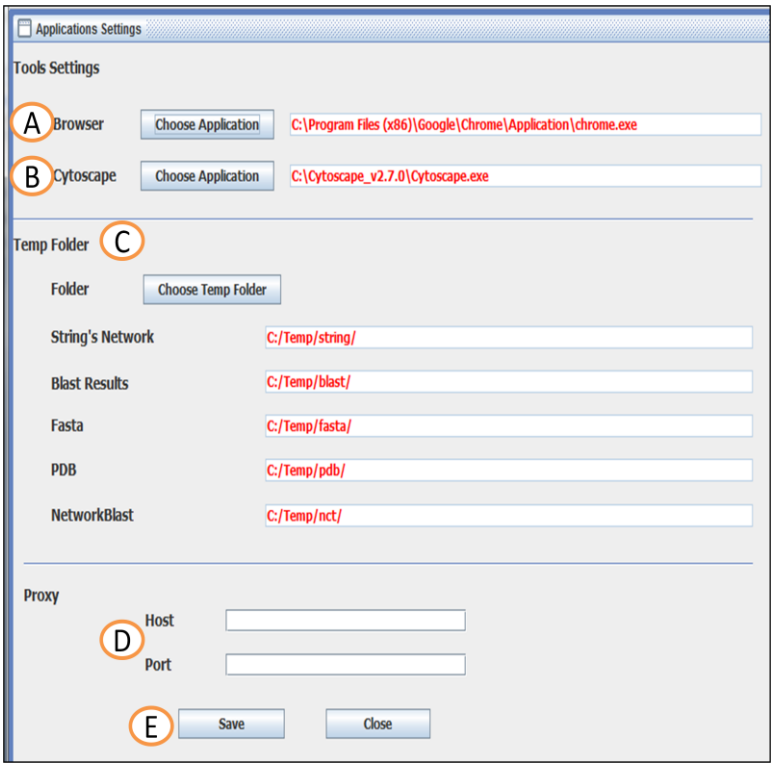


Figura 8: Tela de configuração para uso dos *workflows* científicos.

Além da seleção dos *softwares* acima, esta ferramenta necessita criar um conjunto de pastas para armazenar os arquivos resultantes da execução de ferramentas, bem como, armazenar arquivos temporários. Para tanto, o usuário deve selecionar uma pasta (8C) e a ferramenta irá criar automaticamente as pastas necessárias. Para a seleção da pasta é aberta uma tela de navegação padrão para as pastas do sistema operacional. Além disto, é possível configurar o acesso a Internet através do uso de um *proxy* informando o nome do servidor (ou endereço IP (*Internet Protocol*)) e a porta de acesso (8D). Após as configurações feitas, a ferramenta irá salvar estas informações, criar as pastas e configurar o *proxy* (8E).

A ferramenta possui, também, um registro dos principais comandos executados (Figura 9). Estas informações abrangem a configuração, acesso e respostas das ferramentas de bioinformáticas usadas durante a execução desta ferramenta.

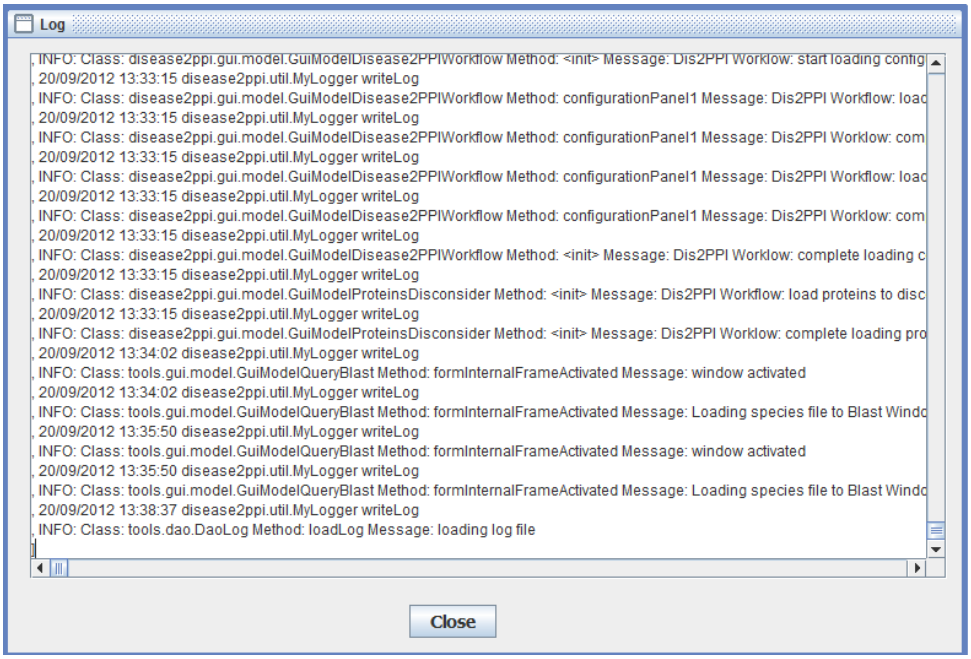


Figura 9: Tela de *log* da ferramenta com informações dos comandos usados para configurar e acessar as ferramentas de bioinformática.

A ferramenta possui uma tela que mostra erros e avisos ocorridos com a execução das funcionalidades desta ferramenta. As informações mostradas (Figura 10) compreendem informações detalhadas de implementação, tais como, nome da classe (10A) e métodos (10B) da linguagem Java, bem como, a mensagem gerada pela aplicação (10C) e a mensagem devolvida pela execução de um método (10D) . Além disto, mostra todas as chamadas de métodos executadas da linguagem Java (10E) no momento que ocorreu o erro. Pode ser que mais de um erro ou aviso se gerado ao mesmo tempo, neste caso, o usuário pode navegar pelos erros avançando ou retrocedendo neles (10F).

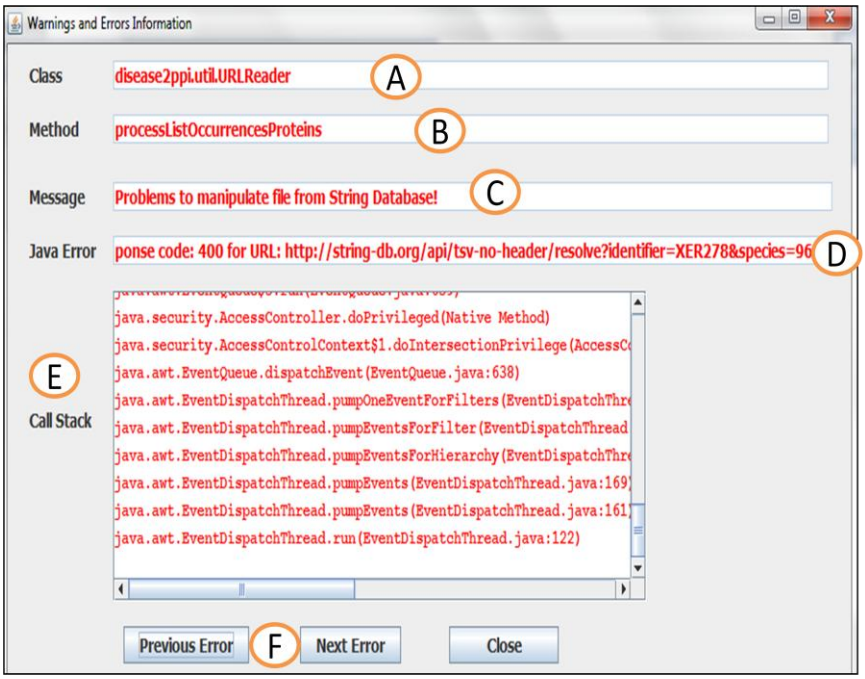


Figura 10: Tela que informa erros de execução da ferramenta ou avisos de problemas ocorridos como acesso a Internet para acessar as ferramentas de bioinformática.

Os erros de execução mais comuns de ocorre são: i) problemas na comunicação com a Internet (erro 400); ii) falta de configuração de um *proxy*; iii) problema de execução interno da ferramenta; iv) obtenção de um identificador inválido de uma proteína ao acessar o banco de dados STRING; e, v) conexão fechada na execução de um alinhamento de sequências acessando o NCBI através das ferramentas do ENTREZ.

9. REFERÊNCIAS COMPLEMENTARES

Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analyzing networks in cell biology. Briefings in Bioinformatics, 7, 3, 243-255.

Albert, R. (2005). Scale-free networks in cell biology. Journal of Cell Sciense, 118, 4947–57.

Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. Fundamentos da Biologia Molecular: Uma introdução à Biologia Molecular da Celula. ArtMed, 1999.

Altintas, I. et al. (2004). Kepler: an extensible system for design and execution of scientific workflows. In Proceedings of Scientific and Statistical Database Management, 21–23 June 2004, Santorini Island, Greece. IEEE Computer Society, pp. 423–424.

Altintas, I., Barney, O. and Jaeger-Frank, E. (2006). Provenance collection support in the Kepler scientific workflow system, in: Proceedings of International Provenance and Annotation Workshop, 118–132.

Altschul, S.F. et al. (1990). Basic Local Alignment Search Tool – BLAST. Journal Molecular Biology, 215, 403-440.

Altschul, S. F. et al. (1997). Gapped blast and Psi-Blast: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389–3402.

Amberger, J., Bocchini, C. A., Scott, A. F. and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Research, 37(suppl 1): D793-D796.

Artimo, A., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., Castro, E., Duvaud, S., Flegel, V., Fortier, A. Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R. Moretti, S., Mostaguir, K., Redaschi, N.,

Rossier, G., Xenarios, I. and Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research, 40, W597-W603.

Avila e Silva, S. Redes neurais artificiais aplicadas no reconhecimento de regiões promotoras em bactérias Gram-negativas. Tese de Doutorado, Programa de Pós-Graduação em Biotecnologia da Universidade de Caxias do Sul, 2011.

Bader, G. D. and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4, 2.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004). Swiss-Prot: Juggling between evolution and stability. Briefings in Bioinformatics, 5(1), 39-55.

Barabási, A. L. and Albert, R, (1999). Emergence of Scaling in Random Networks. Science, 286, 509.

Barabási, A. L. and Albert, R. (2002). Statistical physics of complex networks. Reviews of Modern Physics 74, 47.

Barabási, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews. Genetics 5, 101-113 .

Barabási, A. L. (2009). Scale-Free Networks: A Decade and Beyond. Science, 325 (5939), 412-413.

Barabási, A. L., Gulbahce, N. and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews, Genetics, 12, 56-68.

Barnes, M. R. e Gray, I. C. Bioinformatics for geneticists. Wiley, 2003.

Batagelj, V. and Mrvar, A. (2002). Pajek— Analysis and Visualization of Large Networks. Lecture Notes in Computer Science, 2265/2002, 8-11.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2011). GenBank. Nucleic Acids Research, 39, D32–37.

Berman, H. M. 2000. Protein Data Bank Annual Report. RCSB Project. Available at Internet at http://www.pdb.org/pdb/general_information/news_publications/annual_reports/annual_report_year_2000.pdf .

Bolton, E., Wang, Y., Thiessen, P. A. and Bryant, S. H. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN Annual Reports in Computational Chemistry, American Chemical Society, 4.

Bruin. J. S., Deelder, A. M. and Palmblad, M. (2012). Scientific Workflow Management in Proteomics. Molecular & Cellular Proteomics, 11, M111.010595.

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T. and Vo, H. T. (2006). VisTrails: visualization meets data management. In Proceedings of the 2006 ACM SIGMOD international conference on Management of data (SIGMOD '06). ACM, New York, NY, USA, 745-747.

Chirigati, F. and Freire, J. (2012). Towards Integrating Workflow and Database Provenance. Springer, Provenance and Annotation of Data and Processes, 7525, 11-23.

Claeys, M., Storms, V., Sun, H., Michoel, T. and Marchal, K. MotifSuite: workflow for probabilistic motif detection and assessment. Bioinformatics,28(14), 1931-2.

Cohen, J. Bioinformatics—An Introduction for Computer Scientists. Brandeis University, 122–158 (2004).

Davanzo, V. F. Desenvolvimento de consultas para um banco de dados de regiões promotoras. Centro de Computação e Tecnologia da Informação, Universidade de Caxias do Sul: Caxias do Sul: 2010. Bacharelado em Ciência da Computação - Trabalho de Conclusão de Curso.

Davidson, S., Boulakia, S., Eyal, A., Ludäscher, B., McPhillips, T., Bowers, S., Anand, M., Freire, J. (2007). Provenance in Scientific Workflow Systems. IEEE Data Engineering Bulletin, 30(4), 44–50.

Edgar, E., Domrachev, M. and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30 (1), 207-210.

Elmari, Ramez and Navathe, Shamkant. Fundamentals of Database Systems. 4th ed. Pearson Education, 2004.

Everett, M. G. and Borgatti, S. P. (2012). Categorical attribute based centrality: E–I and G–Fcentrality. Social Networks, 10.1016/j.socnet.2012.06.002.

Freire, J., Silva, C. T., Callahan, S. P., Santos, E., Scheidegger, C. E. and Vo, H. T. (2006). Managing rapidly-evolving scientific workflows. In International Provenance and Annotation Workshop (IPAW), LNCS 4145, 10–18.

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W. and Bryant, S. H. (2010) The NCBI BioSystems database. Nucleic Acids Research, 38, D492-6.

Gibbons, A. Algorithm graph teory. Cambridge University Press, 1994.

Gish, W. and States, D.J. 1993. Identification of protein coding regions by database similarity search. Nature Genet. 3:266-272.

Goecks, J., Nekrutenko, A., Taylor, J. and The Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 11, R86.

Goodrich, M. T. e Tamassia, R. Estrutura de dados e algoritmos em Java. 4. ed., Porto Alegre: Boodman, 2007.

Heuser, C. Projeto de Banco de Dados. Porto Alegre: Artmed, 2008, Série de Livros Didáticos Informática UFGRS, número 4.

Jensen L. J., Kuhn M., Stark M., Chaffron S., Creevey C., Muller J., Doerks T., Julien P., Roth A., Simonovic M., Bork P., von Mering C. STRING 8 - A Global View on Proteins and their Functional Interactions in 630 Organisms. PubMed, 2009.

Jones, N. C. and Pevzner, P. A. An introduction to bioinformatics algorithms. Massachusetts Institute of Technology. MIT Press books, 2004.

Kalaev, M.; Smoot, M.; Ideker, T. and Sharan, R. 2008. NetworkBLAST: comparative analysis of protein networks. Bioinformatics Applications Note: 24, 4, 594–596.

Karsenti, E. (2012). Towards an 'Oceans Systems Biology'. Molecular Systems Biology, 8, 575.

Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. R. and Ideker T. (2004). PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Resourch, 1, 32, W83-8.

Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A.G., Mackie, A., Paulsen, I., Gunsalus, R.P., and Karp, P.D. (2011). EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Research, 39, D583-90.

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L. and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research, 37 (1), D387-392

Kingsbury, D. T. Computacional Biology. ACM Computing Surveys, Vol. 28, No. 1, 101-103, (1996).

Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciufo, S., Fedorov, B., Kiryutin, B., O'Neill, K., Resch, W., Resenchuk, S., Schafer, S., Tolstoy, L. and Ratusova, T. (2009). The national center for biotechnology information's protein clusters database. Nucleic Acids Research, 37, D216–D223.

Köster, J. and Rahmann, S. (2012). Snakemake – A scalable bioinformatics workflow engine. Bioinformatics, 28 (21).

Kuang, R., IE, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y. and Leslie, C. (2005). Profiled-based string kernels for remote homology detection and motif extraction. Journal of Bioinformatics and Computational Biology, 3 (03), 527-550.

Kundaje, A., Xin, X., Lan, C., Lianoglou, S., Zhou, M., et al. (2008). A Predictive Model of the Oxygen and Heme Regulatory Network in Yeast. PLoS Compututational Biology, 4 (11), e1000224.

Larkin, M. A., Blackshieldsm G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.

Linke, B., Giegerich, R. and Goesmann, A. (2011). Conveyor: a workflow engine for bioinformatic analyses. Bioinformatics 27(7), 903 - 911.

Lehninger, A. L.; Cox, M. M.; Nelson, D. L. Principles of biochemistry. 4. ed. New York: Worth, 2005.

Lesk, A. M. Introdução à Bioinformática. 2.ed. Porto Alegre: Artmed, 2008.

Maere, S., Heymans, K. and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21(16), 3448-3449.

Meyer, L. A. V. C., Rössle. S. C., Bisch, P. M. and Mattoso, M. (2004). Parallelism in Bioinformatics Workflows. Springer Berlin, High Performance Computing for Computational Science (VECPAR), Lecture Notes in Computer Science, v. 3402, 233-257.

Molin, A. F. Prototipagem de um Banco de Dados de Promotores como Base para um Portal de Serviços. Centro de Computação e Tecnologia da Informação, Universidade de Caxias do Sul: Caxias do Sul: 2009. Bacharelado em Ciência da Computação - Trabalho de Conclusão de Curso.

Murray, R. K., Bender, D. A. and Botham, K. M. (2012). Computational Biology. Harper's Illustrated Biochemistry, 29e.

Nanni, L., Lumini, A. and Brahnam, S. (2012). An empirical study on the matrix-based protein representations and their combination with sequence-based approaches. Amino Acids, 1-15.

O'Malley, M. A. and Dupre, J. (2005). Fundamental issues in system biology. Bioessays, 27: 1270–1276.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. and Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20, 3045-3054

Papadopoulos, J. S. and Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics, 1, 23(9), 1073-9.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci., Ser. A 5, 17-61.

Pearson, W. and Lipman, D. (1988). Improved tools for biological sequence comparison (amino acid/nucleic acid/data base searches/local similarity). Proceedings of the National Academy of Sciences, 85, 2444-2448.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K. and White, O. (2001). The Comprehensive Microbial Resource. Nucleic Acids Research, 1, 29(1), 123-5.

Picolloto, D. Implementação de novas funcionalidades para o Portal IntergenicDB. Centro de Computação e Tecnologia da Informação, Universidade de Caxias do Sul: Caxias do Sul: 2012. Bacharelado em Sistemas de Informação - Trabalho de Conclusão de Curso.

Ptitsyn, A. and Moroz. L. L. (2012). Computational workflow for analysis of gain and loss of genes in distantly related genomes. BMC Bioinformatics, 13 (15), S5.

Rahn, J. J., Adair, G. M. and Nairn, R. S. (2010). Multiple roles of ERCC1-XPF in mammalian interstrand crosslink repair. Environmental and Molecular Mutagenesis, 51(6), 567-81.

Rosvall, M. and Sneppen, K. (2003). Modeling Dynamics of Information Networks. Physical Review Letters, 91, 178701.

Rotmistrovsky, K., Jang, W. and Schuler, G. D. (2004). A web server for performing electronic PCR. Nucleic Acids Research, 1, 32, W108-12.

Sayers, E. and Wheeler, D. (2004). Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils). In: NCBI Short Courses [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2004. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1058/.

Scardoni, G., Petterlini, M. and Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe. Bioinformatics, 25 (21), 2857-59.

Shaffer. C. A. Data Structures and Algorithm Analysis. Department of Computer Science, Virginia Tech, Blacksburg. Dover Publications, 2011.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research,13(11):2498-504.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. and Ideker, T. 2005. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A, 102, 1974-1979.

Siegal, M. L., Promislow, D. E. L. and Bergman, A. Functional and evolutionary inference in gene networks: does topology matter? Genetica, 10, v.129, n.1, p.83103, 2007.

Silva, F. N., Cavalcanti, M. C., Dávila, A. M. R. (2006). In Services: Data Management for In Silico Workflows. IEEE: Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06), 72, 206 - 210.

Silva, C., Anderson, E., Santos, E. and Freire, J. (2010). Using VisTrails and Provenance for Teaching Scientific Visualization. Proceedings of the Eurographics Education Program, 2010.

Sommerville, I. Engenharia de Software. 8.ed. São Paulo: Pearson Addison-Wesley, 2007.

Tanembaum, A. S. Sistemas Operacionais Modernos. São Paulo: Prentice-Hall, 2010, 3º ed.

The Gene Ontology Consortium, Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, 25, 1, 25–29.

Vizcaino, J. A., Cote, R., Reisinger, F., Barsnes, H., Foster, J. M., et al. (2010). The Proteomics Identifications database: 2010 update. Nucleic Acids Research, 38, D736–742.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., Bolton, E., Gindulyte, A. and Bryant, S. H. (2012). PubChem's BioAssay Database. Nucleic Acids Research, 40(1), D400-12.

Wang, J., Crawl, D. and Altintas, I. (2012a). A Framework for Distributed Data-Parallel Execution in the Kepler Scientific Workflow System. Procedia Computer Science, 9, 1620–29.

WfMC - Workflow Management Coalition, The Workflow Handbook 2004, Fischer,L.(ed.). Disponível em: <http://www.wfmc.org/information/handbook04.htm>.

Weston, A. D. and Hood, L. (2004). Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. Journal of Proteome Research, 3 (2), 179–196.

Wheeler D. L., Barrett T., Benson D. A., Bryant S. H., Canese K., Chetvernin V., Church D. M., DiCuccio M., Edgar R., Federhen S., Geer L. Y., Kapustin Y., Khovayko O., Landsman D., Lipman D. J., Madden T., Maglott D. R., Ostell J., Miller V., Pruitt K. D., Schuler G. D., Sequeira E., Sherry S. T., Sirotkin K., Souvorov A., Starchenko G., Tatusov R. L., Tatusova T. A., Wagner L.and Yaschenko E. (2006).

Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 35 (suppl 1), D5-D12.

Wright, J. and Wagner, A. (2008). The Systems Biology Research Tool: evolvable open-source software. BMC Systems Biology, 2, 55.

Yalamanchili, H. K., Xiao, Q. and Wang, J. (2012). A novel neural response algorithm for protein function prediction. BMC Systems Biology, 6 (1), S19.

Yu, H., Kim, P. M., Sprecher, S., Trifonov, V. and Gerstein, M. (2007). The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLoS Computational Biology, 3, 4, e59.

Zaha, A. Biologia Molecular Básica. Mercado Aberto, 1996.

Zhang, Z., Kochhar, S. and Grigorov, M. G. (2009). Descriptor-based protein remote homology identification. Protein Science, 14 (2), 431–444.