Correlacionando genes e doenças através de caminhos metabólicos

Carla Fernandes da Silva[1,2], **Kuruvilla Joseph Abraham**[3]
Evandro Eduardo Seron Ruiz[2]

[1]Programa de Pós-Graduação em Computação Aplicada

[2]Departmento de Computação e Matemática, FFCLRP
Universidade de São Paulo.

[3]Centro Universitário Estácio de Ribeirão Preto

{carlanandess,evandro}@usp.br, abrahamkjos@gmail.com

Abstract. *One of the main challenges in science is to identify the factors that cause these diseases. Among these factors are the genes. This work will present a methodology to prioritize genes using pathways related to a complex disease. The challenge is to unveil which genes can contribute to triggering a complex disease. The goal is to develop a methodology for prediction of gene-disease through the integration of data related to genes, diseases and metabolic pathways and to eventually discover new genes associated with a disease.*

Resumo. *Um dos principais desafios da ciência é identificar os fatores que causam essas doenças, dentre estes fatores estão os genes. Neste trabalho, será apresentada uma metodologia para priorizar genes e vias metabólicas relacionados a uma doença complexa, com o desafio de descobrir quais os genes podem contribuir para desencadear uma doença complexa. O objetivo é desenvolver uma metodologia para predição de gene-doença através da integração de dados de genes-doencas-vias metabólicas, visando a descoberta de novos genes associado a doença.*

1. Introdução

Os crescentes avanços genéticos e tecnológicos têm proporcionado uma melhor compreensão das doenças humanas no âmbito geral. Muitas doenças cujas causas eram atribuídas a fatores ambientais e comportamentais, hoje sabe-se que essas doenças possuem também fatores genéticos associados. As doenças cujos fatores protagonizantes são os genéticos, aliado aos fatores ambientais e comportamentais são conhecidas como *doenças complexas* [Goh and Choi 2012]. Este trabalho aborda o desafio de propor uma metodologia para desvendar quais os genes que podem contribuir para desencadear essas doenças.

Atualmente, podemos notar um grande interesse nas pesquisas relacionadas à predição gênica, ou seja, a identificação funcional de genes em associação a uma doença. Estas pesquisas são facilitadas pela enorme quantidade e variedade de dados genéticos que estão disponíveis em bases públicas de dados. Muitas destas pesquisas consideram informações já existentes de relacionamentos de genes associados a doenças para descobrir novos relacionamentos [Goh et al. 2007, Duarte and Becker 2007, Lee et al. 2008, Lee et al. 2011, Vidal et al. 2011, Ritchie et al. 2015, Menche et al. 2015].

Diversos trabalhos na literatura como [Goh et al. 2007, Wu et al. 2008, Li and Agarwal 2009, Barabási 2007, Barrenas et al. 2009, Barabási et al. 2011, Zhou et al. 2014] utilizaram as abordagens de redes complexas para um melhor entendimento dos mecanismos que servem como base para as doenças complexas. Em um trabalho pioneiro, [Goh et al. 2007] criou uma rede de doenças humanas (*Human Disease Network* – HDN) conectando todas as doenças hereditárias que compartilham um gene causador de doença, de acordo com o banco de dados OMIM. Na rede HDN, duas doença estão conectadas se elas estão associadas a um mesmo gene. Dado que as ligações significam associação genotípica elas poderiam ser usada em conjunto com reações metabólicas.

A união da abordagem de redes gênicas juntamente com a integração de dados de diversos tipos de banco biológicos têm contribuído para diversas descobertas de novos genes para compreensão de doenças complexas. Neste trabalho, é proposta uma metodologia que integra dados sobre genes associados a doenças com informações sobre vias metabólicas para a construção de uma rede de doenças com intuito de selecionar genes que possam explicar um fenótipo da doença. Diferentemente de trabalhos relacionados está proposta inclui vias-metabólicas como fator de associação de doenças.

2. Metodologia

A metodologia proposta integra dados sobre genes, doenças e vias metabólicas visando a descoberta de novos genes associados a comorbidades. A construção da metodologia proposta neste projeto pode ser dividida em três etapas, que são:

1. A primeira etapa consiste em extrair as associações de doença-gene do banco de dados KEGG *Disease* que possuem em torno de 1.704 tipos de doenças humanas [1]. Para cada doença será criada uma lista de genes padronizada de acordo com o a nomenclatura HGNC[2]. A partir desta lista gene-doença queremos encontrar a sobreposição de genes relacionados com duas doenças ou mais doenças.
2. Na segunda etapa construíremos uma rede doença-doença a partir dos pares de diagnóstico principal e secundário presentes nas Autorizações de Internação Hospitalares (AIHs) extraídos do DATASUS. A rede doença-doença extraída das relações doenças-genes derivadas do KEGG, da etapa anterior, será confrontada com a rede doença-doença extraída das AIHs.
3. A terceira etapa compreende em criar uma rede doença-gene-via metabólica para estudar associações entre genes e doenças através de vias-metabólicas. Procuraremos diferenciar as vias metabólicas gerais das vias específicas e assim esperamos descartar os genes das vias mais gerais por elas fazerem parte de processos biológicos essenciais para a sobrevivência do indivíduo. Esta análise permitirá descobrir se existe uma correlação entre as doenças que compartilhem genes e vias metabólicas entre si. Uma visão sistemátizada da metodologia proposta neste trabalho é apresentada na Figura 1.

3. Resultados Preliminares

Até o presente momento, já conseguimos gerar a rede doença-gene que pode ser representada por um grafo bipartido com 1279 vértices correspondentes a doenças, 2692 vértices

[1]Última atualização do KEGG *Disease* em 8 janeiro de 2017.
[2]HUGO Gene Nomenclature Committee

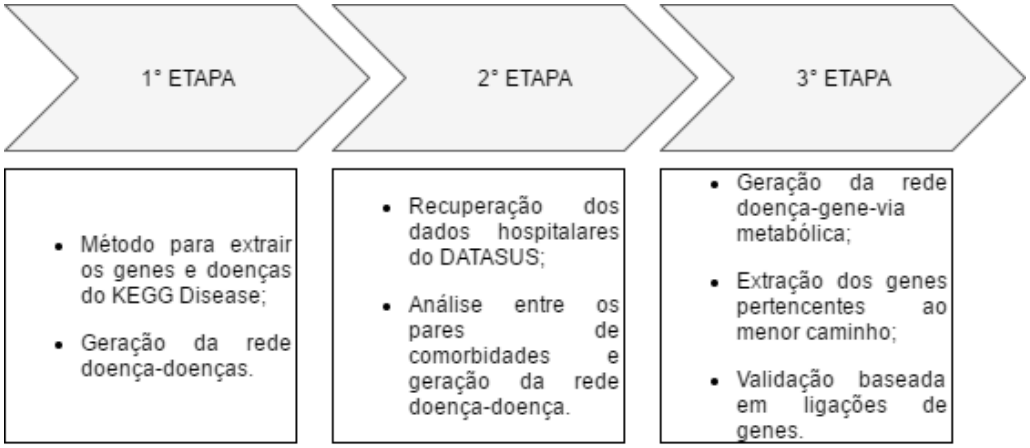

Figura 1. Pipeline de execução das etapas do projeto.

correspondentes a genes e 4432 arestas. A rede doença-gene possui 418 componentes conexos sendo que o maior componente com contém 2778 nós. A partir do grafo doença-gene foi gerada um grafo doença-doença, que possui 905 vértices (doenças) e 3480 arestas. Este grafo apresenta 44 componentes conexos sendo que o maior componente apresenta 775 doenças.

A Figura 2(a) ilustra a rede doença-doença gerada a partir do maior componente conexo e com vértices com grau $\geq$2. O grafo ilustra as relações entre 613 doenças que correspondem onde a 67,7% das doenças no agrupamento principal, e estão ligadas por 3182 genes (arestas), que representam 91,5% do mesmo agrupamento. A primeira etapa ainda não está totalmente concluída pois ainda não foi realizada a padronização dos genes e o mapeamento do código de doença do KEGG para CID-10. Pelo gráfico da Figura 2(b) é possível concluir, a partir da distribuição dos graus, que esta rede doença-doença não é uma rede aleatória, o que já se apresenta um resultado original.
Esperamos realizar a padronização via HGNC e compará-la com as relações entre as comorbidades dos registros do SUS.

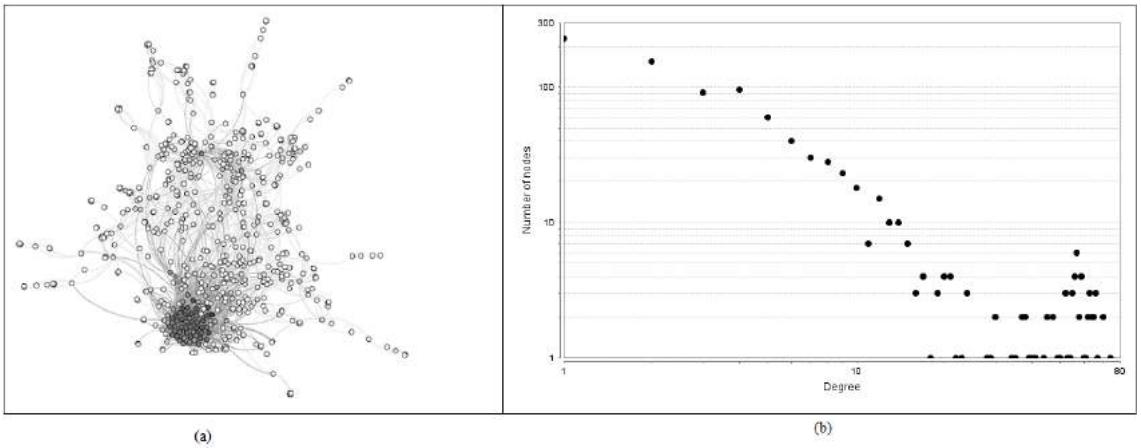

Figura 2. (a) Grafo representando as relações genéticas no KEGG entre 613 doenças, e (b)Distribuição de graus dos vértices.

Referências

Barabási, A.-L. (2007). Network Medicine — From Obesity to the "Diseasome". *NEJM*, 357:404–407.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network Medicine: A Network-based Approach to Human Disease. *Nature Reviews Genetics*, 12(1):56–68.

Barrenas, F., Chavali, S., Holme, P., Mobini, R., and Benson, M. (2009). Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS ONE*, 4(11):2–7.

Duarte, N. and Becker, S. A. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1777–1782.

Goh, K. I. and Choi, I. G. (2012). Exploring the human diseasome: The human disease network. *Briefings in Functional Genomics*, 11(6):533–542.

Goh, K.-i., Cusick, M. E., Valle, D., Childs, B., and Vidal, M. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Lee, D. S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N., and Barabasi, A. L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):9880–9885.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121.

Li, Y. and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PLoS ONE*, 4(2):e4346.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97.

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–98.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Molecular Systems Biology*, 4(189):189.

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. *Nature Communications*, 5(May).