Desenvolvimento de uma Metodologia para Análise Gênica de Comorbidades a Partir da Integração de Dados Epidemiológicos

Karla F. Neto[1], Victor E. F. Ferraz[2], Domingos Alves[3], Evandro E. S. Ruiz[1]

[1]Departamento de Computação e Matemática, FFCLRP – Universidade de São Paulo
Ribeirão Preto – SP – Brasil

[2]Departamento de Genética, FMRP – Universidade de São Paulo
Ribeirão Preto – SP – Brasil

[3]Departamento de Medicina Social, FMRP – Universidade de São Paulo
Ribeirão Preto – SP – Brasil

{karla.ferraz,evandro,vferraz}@usp.br, {doquiron}@gmail.com

Abstract. *The identification of genes responsible for human diseases provides knowledge of pathological and physiological mechanisms that are essential for the development of new diagnostics and therapeutics. We know that a disease is rarely a consequence of an abnormality in a single gene, but reflects disorders in an intra and intercellular network. In this project we analyze records of comorbidities, found genes related to these diseases and evaluate whether this genes correspond to the incidence of those pairs of diseases.*

Resumo. *A identificação de genes responsáveis por doenças humanas fornece conhecimento sobre mecanismos patológicos e fisiológicos que são essenciais para o desenvolvimento de novos diagnósticos e terapias. Sabemos que uma doença é raramente uma consequência de uma anormalidade em um único gene, porém reflete desordens de uma rede intra e intercelular complexa. Neste trabalho analisamos registros de comorbidades, obtivemos os genes relacionados a estas doenças e avaliamos se tais genes correspondem a incidência causal destes pares de doenças.*

1. Introdução

A identificação de genes responsáveis por doenças humanas fornece conhecimento sobre mecanismos patológicos e fisiológicos que são essenciais para o desenvolvimento de novos diagnósticos e terapias. Dadas as interdependências funcionais entre os componentes moleculares nas células humanas, uma doença é raramente uma consequência de uma anormalidade em um único gene, porém é um reflexo de desordens de uma rede intra e intercelular complexa que liga os tecidos aos órgãos. Os métodos de bioinformática que vem sendo desenvolvidos oferecem uma plataforma para explorar sistematicamente não só a complexidade molecular de uma doença em particular, conduzindo a identificação de caminhos a doenças, mas também

as relações moleculares entre os fenótipos patológicos distintos. A demonstração de métodos bem sucedidos para priorização de genes serve para a validação de abordagens computacionais específicas utilizadas para a representação e inferência de conhecimento para o beneficio da saúde humana [Barabási 2011].

Muitos repositórios biomédicos conhecidos, como os históricos clínicos de pacientes, poderiam ser utilizados como informações fenotípicas humanas. Essas informações podem ser utilizadas em conjunto com os dados moleculares e genéticos para o auxílio a descobertas de origens moleculares de doenças. Um dado relevante que pode ser estudado a partir destes bancos é o estudo das comorbidades - a presença de uma ou mais desordens (ou doenças) em adição a uma doença ou desordem primária que o paciente apresenta [Goh 2007]. Através do estudo detalhado de doenças e suas principais comorbidades podemos tentar identificar os principais genes causadores de uma doença e com isso verificar se esses mesmos genes são também vistos como responsáveis de doenças vistas em comorbidades comuns [Lee 2008].

O objetivo deste projeto é desenvolver uma metodologia de integração de informações genéticas e fenotípicas, no caso comorbidades, para a priorização de genes. Para um estudo de caso mais detalhado, escolhemos a Fenda Palatina como doença a ser analisada, epidemiológica e geneticamente, através das suas comorbidades. Esta anomalia congênita, de etilogia multifatorial (com contribuição genética e ambiental) apresenta prevalência entre 1/700 e 1/1000 nascidos vivos [Grosen 2011]. Quase metade de todos os nascimentos com fenda palatina ocorre em crianças com outras anomalias congênitas, o que favorece nossa abordagem.

2. Metodologia

Para a análise e coleta de dados sobre comorbidades associadas a Fenda Palatina foi utilizado o Sistema de Internações Hospitalares (SIH) do DATASUS[1]. O SIH contém informações que viabilizam pagamento dos serviços hospitalares, sendo estruturado na lógica da avaliação e controle da produção. Os hospitais públicos ou conveniados do SUS enviam informações das internações, por meio da Autorização de Internação Hospitalar (AIH), para gestores municipais ou estaduais. As AIHs incluem informações de internações realizadas do país inteiro tais como valor dos procedimentos realizados, datas de entrada e saída do paciente, data de óbito, assim como diagnóstico primário e secundário, com seu respectivo código CID-10. Dessa forma, podemos analisar quais doenças ocorrem em conjunto com a Fenda Palatina (CID-10 Q351 a Q359). Para esta avaliação preliminar focamos a análise na associação com outras anomalias congênitas, centrando a investigação nos códigos do Capítulo XVII do CID 10: Malformações Congênitas, Deformidades e Anomalias Cromossômicas.

A partir dos pares de diagnóstico principal e secundário presentes nas AIHs, calculamos as frequências, com o objetivo de diferenciar o quão significante os pares são em comparação com toda a amostra de pares de doenças. Isso foi feito a partir de Tabelas de Contingências e Testes de Fisher, garantindo significância estatística. Através disso, os p-valores são calculados a partir do teste exato de Fisher e os pares são escolhidos de acordo com os valores e o valor limiar estipulado. Filtramos a lista

[1] http://www.datasus.gov.br/catalogo/sihsus.htm

impondo um valor de corte em p = 0,05 de escore de comorbidade entre as doenças A e B definidas por $cs_{AB} = \ln(\frac{Obs+1}{Expt+1}), Expt = \frac{n_a * n_b}{n_{total}}$, onde Obs é o número observado da associação dos pares A, B na amostra, e Expt é o número esperado [Roque 2011]. Dessa forma, conseguimos uma lista de pares de doenças estatisticamente significantes para análise comparativa de genes envolvidos em tais doenças.

A busca de genes é feita a partir do banco Online Mendelian Inheritance in Man (OMIM)[2]. Como não há um mapeamento direto entre os termos de doenças codificadas em CID-10 e os termos da base OMIM, é necessário um mapeamento intermediário. Esse mapeamento intermediário é feito a partir do Unified Medical Language System (UMLS) e do software MetaMap. A UMLS é uma lista exaustiva de terminologias biomédicas criada para o desenvolvimento e interoperabilidade entre sistemas computacionais especializados em biomedicina e saúde [UMLS 2012]. MetaMap é um programa altamente configurável para mapear texto biomédico ao Metatesauro do UMLS ou, equivalentemente, para descobrir conceitos do Metatesauro referidos no texto. MetaMap utiliza uma abordagem baseada no conhecimento simbólico, processamento de linguagem natural (NLP) e técnicas linguísticas computacionais [MetaMap 2012]. Dessa forma, a partir de um termo CID-10 adquirimos o termo mais relevante na base OMIM. Após esta etapa, podemos encontrar uma lista dos genes responsáveis a cada uma das doenças.

Com os genes responsáveis de cada doença, e com as comorbidades epidemiologicamente ligadas, podemos analisar se os genes responsáveis de cada doença justificam a ligação epidemiológica dos pares. Assim, para cada gene de cada doença, verificamos a pertinência da associação entre os genes correspondentes.

Construímos um diagrama em que podemos visualizar as principais etapas metodológicas deste projeto. Vide a Figura 1.

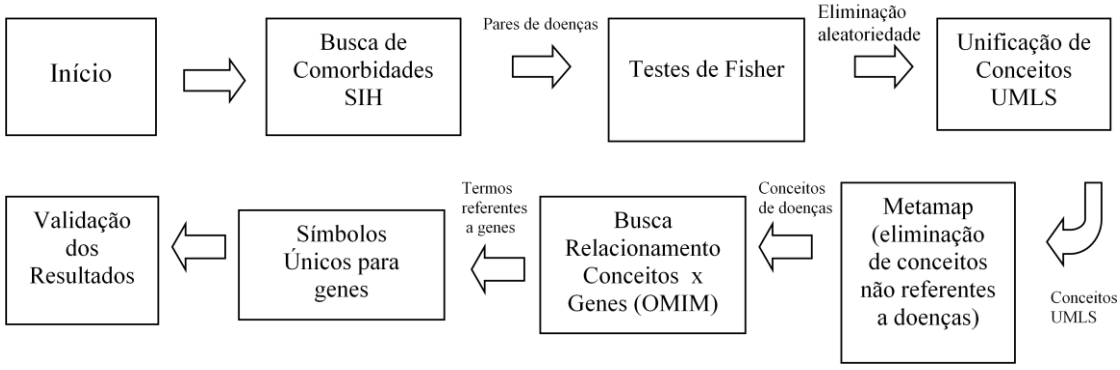

Figura 1. Principais etapas metodológicas.

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

Resultados Preliminares

A partir da busca de genes pelo OMIM e da geração das listas de genes para cada doença, podemos construir uma tabela com as referências gênicas relacionadas às comorbidades de Fenda Palatina. Alguns destes termos referem-se a termos encontrados em textos científicos escritos antes da unificação dos nomes dos genes, assim estes termos ainda precisam ser tratados e validados futuramente. Parte dessas correlações pode ser vista na Tabela 1.

Tabela 1. Conjunto de Doenças com as respectivas referências gênicas comuns a Fenda Palatina.

Doenças	Referências Gênicas
Hérnia Ventral	THAS,TAS,WBS,WMS,WS,DEL7q11,C7DELq11
Anencefalia	THAS,TAS
Pneumonia	MDLS,MDS,MDCR,DEL17p13.3,C17DELp13.3
Microcefalia	PQBP1,NPW38,SHS,MRX55,MRXS3,RENS1,MRXS8,DEL3pterp25,C3DELpterp25
Hipertensão primária	WBS,WMS,WS,DEL7q11,C7DELq11,TGFB1,DPD1,CED,MMP2,CLG4A,MONA
Hérnia Umbilical	OGS2,BBBG2,GBBB2,DEL18q,CATMANS,ZLS,GUSB,MPS7
Oligoidrâmnio	HFM,EEC1
Hidrocefalia	DEL1p36,C1DELp36,FGFR2,BEK,CFD1,JWS,THAS,TAS,HFM
Ptose	DEL3pterp25,C3DELpterp25,DEL8q13,C8DELq13,DEL17q21.31,C17DELq21.31
Carcinoma	CDH1,UVO,LCAM,ECAD,NBCCS,BCNS
Hipotireoidismo	FOXE1,FKHL15,TITF2,TTF2,DEL1p36,C1DELp36,WBS,WMS,WS,DEL7q11,C7DELq11,PTLS,ALB
Aneurisma	TGFBR2,HNPCC6,AAT3,MFS2,LDS1B,LDS2B,TGFBR1,ALK5,AAT5,LDS2A,LDS1A,MSSE,MFS1,WMS2,SSKS,GPHYSD2,MMP2,CLG4A,MONA,TGFB1,DPD1,CED
Polidactilia	GLI3,PAPA,PAPB,DEL3pterp25,C3DELpterp25
Sindactilia	FGFR2,BEK,CFD1,JWS,EEC1,GLI3,PAPA,PAPB,IRF6,VWS,LPS,PIT,PPS,OFC6

O próximo passo dessa pesquisa é a validação da hipótese de que os grupos de genes comuns aos pares de doenças justificam as comorbidades vistas nos registros hospitalares.

Referências Bibliográficas

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

Goh K-I, Cusick ME, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685–8690.

Lee DS, Park J, et al. (2008) The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A 105: 9880–9885.

Grosen D, Bille C, et al. 2011. Risk of oral clefts in twins. Epidemiol 22:313–319.

Roque FS, Jensen PB, et al. (2011) Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts.

Definição UMLS: http://www.nlm.nih.gov/research/umls/. Último acesso em 7 de abril, 2012.

Definição Metamap: http://metamap.nlm.nih.gov/.Último acesso em 7 de abril, 2012.