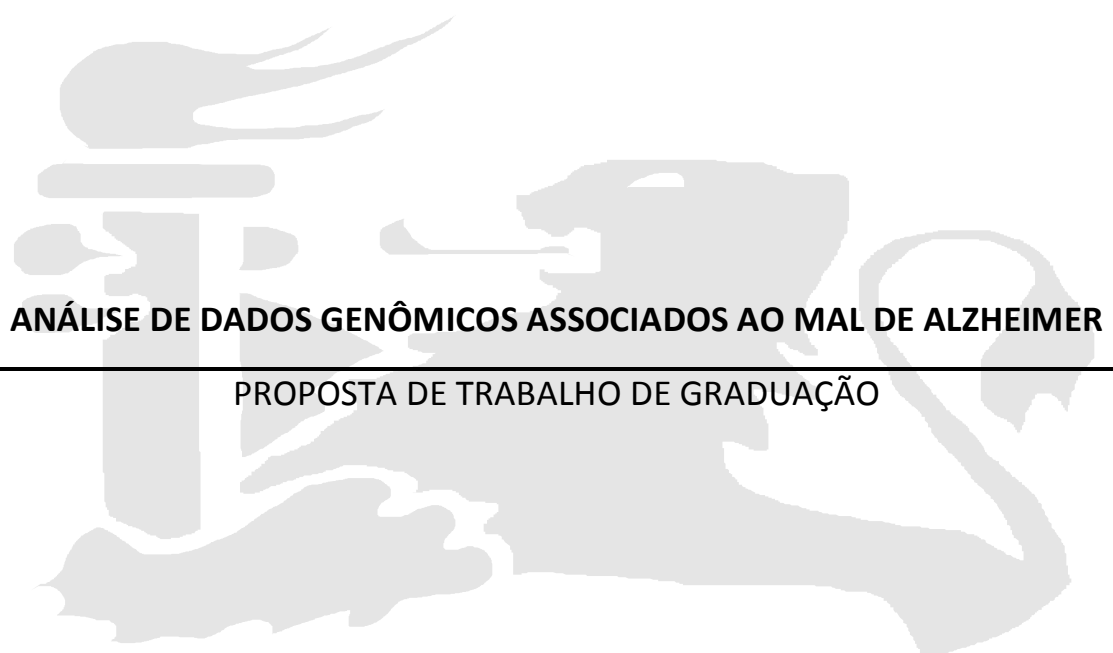


UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA



ANÁLISE DE DADOS GENÔMICOS ASSOCIADOS AO MAL DE ALZHEIMER

PROPOSTA DE TRABALHO DE GRADUAÇÃO

ALUNO: Leyla Manoella M. R. de Lima {lmmrl@cin.ufpe.br}
ORIENTADOR: Katia S. Guimarães, PhD {katiag@cin.ufpe.br}

Recife, Abril de 2011

1. Contexto

Uma meta importante da genética humana é entender como as mudanças no DNA dos indivíduos afetam o risco de eles desenvolverem determinadas doenças, para assim poder criar métodos mais efetivos de prevenção e tratamento. Dentre as alterações genéticas que influenciam na susceptibilidade a doenças, as mais comuns e melhor estudadas são os polimorfismos de base única, SNPs (do inglês *Single Nucleotide Polymorphism*), que ocorrem devido à alteração de um único nucleotídeo do DNA. Atualmente já foram identificados mais de um milhão de SNPs do genoma humano.

Para analisar como a ocorrência de SNPs e a interação entre eles influenciam o risco de doenças complexas comuns, como hipertensão, diabetes mellitus, câncer e mal de Alzheimer, uma das estratégias mais usadas atualmente são os estudos de associação que englobam todo o genoma, chamados GWAS (do inglês *Genome-Wide Association Studies*). Neste tipo de estudo, dezenas de milhares de SNPs são testados para associação com doenças em milhares de pessoas, possibilitando uma investigação profunda do genoma humano.

Os GWAS só começaram a se tornar viáveis nos últimos cinco anos, após a publicação do sequenciamento do genoma humano (International Human Genome Sequencing Consortium, 2004) e da criação do Projeto Internacional HapMap (The International HapMap Consortium, 2005). O Projeto HapMap foi concebido com o objetivo de melhor caracterizar a variação genética humana, através da criação de um banco de dados de aproximadamente 3 milhões de SNPs comuns do genoma, além de documentar suas frequências e as relações estruturais ou desequilíbrio de ligação (LD) entre os alelos [1].

Muita expectativa foi criada em torno das potencialidades desse tipo de estudo e, com o avanço tecnológico que possibilitou diminuir o custo do sequenciamento genético, eles se disseminaram. Até o fim de 2010, já haviam sido publicados 1212 GWAS englobando 210 diferentes doenças e características humanas [2]. No entanto, diversos estudos que avaliam os prós e os contras dessa abordagem [3], mostram que os GWAS ainda apresentam várias limitações, como o excessivo número de falsos positivos e a baixa herdabilidade observada nos resultados. De fato, poucos dos *loci* (posições de ocorrência dos SNPs) descobertos via GWAS estão associados com um aumento moderado ou alto do risco de doença. A maioria dos SNPs têm efeitos muito pequenos na susceptibilidade a doenças e, portanto, não podem ainda ser usados para testes genéticos.

Estes estudos de prós e contras sugerem uma forte influência de uma combinação de SNPs, que atuam em cooperação para um determinado fenótipo ou propensão a uma determinada doença. Diante da multiplicidade de SNPs e do número super-exponencial de possíveis combinações, o tratamento computacional do problema de identificar as combinações mais fortemente relacionadas considerando todas é proibitivo. Portanto, o desenvolvimento de abordagens computacionais e estatísticas eficientes para identificar as interações entre SNPs ainda permanece um grande desafio, devido à imensa quantidade de dados envolvida e à complexidade da relação genótipo-fenótipo que se pretende mapear.

2. Objetivos

Recentemente, Jiang et al. [4] mostrou que a utilização do modelo de redes Bayesianas para identificar interações gênicas a partir de dados de GWAS é bastante promissora. Ele utilizou dados de um estudo real de GWAS para o Mal de Alzheimer [5], além de dados simulados. Este trabalho tem como objetivo replicar os experimentos do artigo de Jiang *et al.* e realizar novas análises dos dados a partir dos resultados obtidos.

Primeiramente será feito um estudo detalhado do artigo e de todos os critérios de escore avaliados pelos autores. Será feito também um estudo de *background*, onde serão abordados conceitos básicos para o entendimento do artigo, como epistasia, redes bayesianas, GWAS, etc, e como modelar interações gênicas utilizando redes bayesianas de forma adequada. Depois será feita a implementação da rede bayesiana, serão selecionados os critérios de escore a serem usados com a rede. Finalmente, serão realizados experimentos usando os mesmos dados de Alzheimer utilizados por Reiman et al. [5], e será feita uma análise comparativa entre os resultados obtidos e os publicados na literatura.

3. Cronograma

O cronograma abaixo define os marcos principais no desenvolvimento deste trabalho.

Atividades	Março	Abril	Maio	Junho
Revisão bibliográfica e definição do tema.				
Leitura e entendimento inicial do artigo.				
Definição do escopo.				
Implementação da rede bayesiana.				
Preparação do relatório e da apresentação				

4. Possíveis Avaliadores

Katia Silva Guimarães

Renata Maria Cardoso Rodrigues de Souza

5. Referências

1. International HapMap Consortium. 2005. **A haplotype map of the human genome**. Nature; 2005; 437:1299–1320.
2. Moore, J. H. et al. **Bioinformatics challenges for genome-wide association studies**. Bioinformatics; 2010; 26:445–455.
3. Hindorff, L.A., Junkins, H.A., Manolio, T.A. **NHGRI Catalog of published genome-wide association studies**. (<http://www.genome.gov/gwastudies>, acessado em 03/11).
4. Jiang et al. **Learning genetic epistasis using Bayesian network scoring criteria**. BMC Bioinformatics, 2011.
5. Reiman EM, et al. **GAB2 alleles modify Alzheimer's risk in APOE carriers**, Neuron; 2007; 54:713-720.

6. Assinaturas

Katia Silva Guimarães, PhD (Orientador)

Leyla Manoella M. R. de Lima (Aluno)