Universidade Federal de Pernambuco
Centro de Informática
Graduação em Engenharia da Computação

Um estudo de desempenho das técnicas de aprendizagem de máquina na classificação de variantes genéticas quanto à patogenicidade

Proposta de Trabalho de Graduação

Aluna: Renata Correia de Andrade

Orientador: Paulo Salgado Gomes de Mattos Neto

Recife, Setembro 2016

Sumário

Contexto

O DNA (ácido desoxirribonucleico) é a estrutura que determina o funcionamento e a forma de todos os seres vivos. O ácido é composto de quatro bases químicas: adenina, guanina, citosina e timina. A disposição dessas bases na cadeia de DNA é o que define as características dos organismos. Sequenciar o DNA significa estabelecer a ordem em que essas bases estão dispostas [1].

A primeira vez que o genoma humano foi sequenciado e publicado por completo foi em 2003, como resultado do The Human Genome Project [2]. O projeto, que começou a ser idealizado na década de 80, foi iniciado apenas em 1990, levando treze anos para ser concluído. O Human Genome Project envolveu vários países do mundo e custou quase três bilhões de dólares. Só a etapa de sequenciamento chegou a custar mais de trezentos milhões [3].

Na última década, a tecnologia aplicada ao sequenciamento de DNA possibilitou uma redução drástica no preço e no tempo de sequenciamento, mudando a forma como a análise genética é aplicada. O custo para sequenciar um genoma humano completo foi de aproximadamente quatorze milhões de dólares em 2006 para menos de mil e quinhentos dólares em 2016 [3], como pode ser observado na Figura 1.

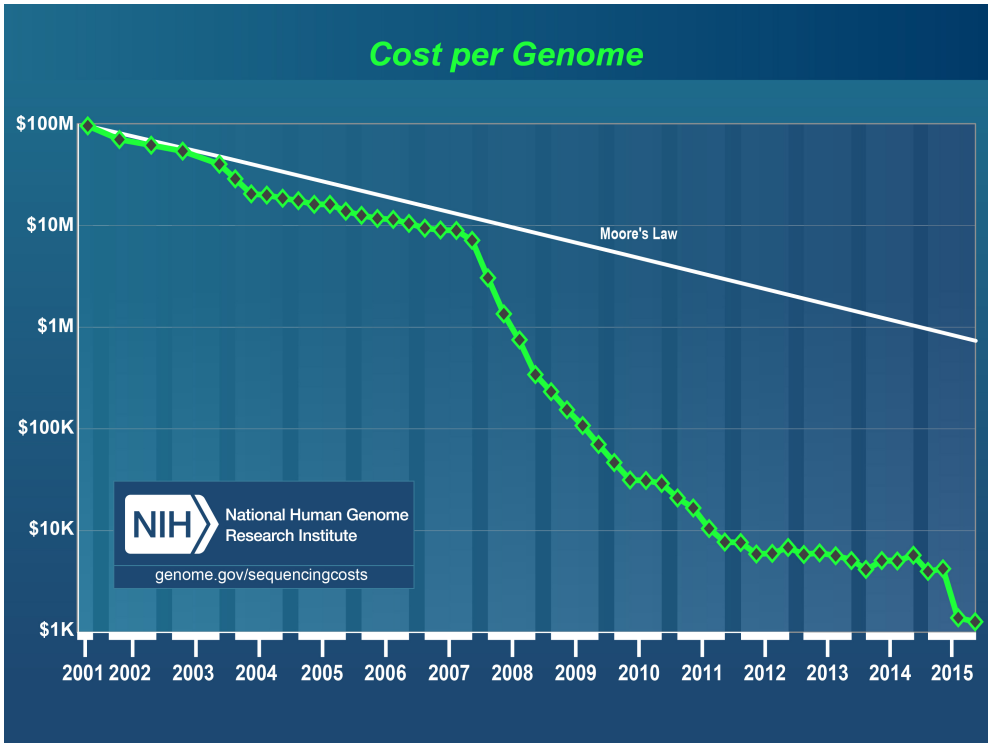

FIGURA 1 - Custo por genoma [2]

A redução no preço tornou viável a utilização da genética na medicina clínica, onde a identificação de variantes genéticas podem justificar fenótipos relacionados à doenças.

Uma variante (ou mutação) não é necessariamente prejudicial ao organismo. A heterocromia, caracterizadas por olhos de cores diferentes, é um exemplo de variante genética benigna. Por outro lado, a mutação de uma única base no gene BRCA1 pode elevar de 1.3% até 39% a chance de uma mulher desenvolver câncer de ovário [4]. Cabe então aos médicos e biomédicos que trabalham com genética, identificar, dentre as mutações encontradas, quais são realmente patogênicas e precisam de mais atenção.

O desafio está no fato de que ainda não se sabe tudo sobre os genes que compõe o genoma humano, nem o que alterações nesses genes podem causar. O homem possui cerca de 3 bilhões de pares de base no seu DNA [1] e uma infinidade de possíveis variantes. Somente o sequencialmente do exoma - parte do DNA responsável pela codificação de proteínas que e representa apenas 1,5% de todo o genoma humano[5] - pode resultar na identificação de quinze à vinte mil variantes [6]. Analisar cada uma delas individualmente inviabiliza o exame genético para aplicações clínicas. O especialista depende então de ferramentas computacionais que o auxiliem nessa tarefa.

Objetivos

O objetivo desse trabalho é investigar o desempenho das técnicas de aprendizagem de máquina quando o problema a ser resolvido é o de classificação de variantes quanto à patogenicidade.

Atualmente, algumas ferramentas que visam solucionar o problema de classificação de variantes quanto à patogenicidade estão disponíveis gratuitamente, como CADD [7] e DANN [8]. Cada uma dessas ferramentas foi treinada e testada usando bancos de dados diferentes. Muitas vezes não é especificado como esses dados foram tratados e nem quais parâmetros foram usados ema cada modelo, o que dificulta a utilização dessas ferramentas para comparar os algoritmos de aprendizagem.

A partir dos dados obtidos no ClinVar, um banco aberto de variantes, este trabalho visa dar condições iguais aos algoritmos de classificação para que o mais apropriado para o problema possa ser eleito. Selecionado o melhor modelo, seu desempenho será ainda comparado ao das ferramentas disponíveis atualmente.

Cronograma

	Setembro					Outubro				Novembro					Dezembro			
Revisão Literária	X	X	X	X	X	X	X	X	X	X	X	X	X	X				
Construção da base de Dados		X	X	X														
Seleção dos Algoritmos				X	X													
Implementação dos Modelos						X	X	X	X									
Análise dos Resultados									X	X	X	X						
Escrita do Relatório										X	X	X	X	X	X	X		

Avaliadores

Abaixo encontram-se listados os possíveis avaliadores para o trabalho descrito nesta proposta.

1. George Darmiton da Cunha Cavalcanti
2. Tsang Ing Ren

Referências Bibliográficas

[1] *DNA Sequencing Fact Sheet,* National Human Genome Research Institute.

Disponível em: <https://www.genome.gov/10001177/dna-sequencing-fact-sheet/>

Acessado em: 13 de Setembro de 2016.

[2] *An Overview of the Human Genome Project*, National Human Genome Research Institute.

Disponível em: <https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/>

Acessado em: 6 de Setembro de 2016.

[3] *The Cost of Sequencing a Human Genome*, National Human Genome Research Institute.

Disponível em: <https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/>

Acessado em: 6 de Setembro de 2016.

[4] ANTONIOU A, PHAROAH PD, NAROD S, et al. *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies*. American Journal of Human Genetics, 2003.

[5] *What is exome sequencing?*, Broad Institute of MIT and Harvard.

Disponível em <https://www.broadinstitute.org/blog/what-exome-sequencing>

Acessado em: 13 de Setembro de 2016.

[6] SITZIEI NO., KIEZUN A., SUNYAEV S. *Computational and statistical approaches to analyzing variants identified by exome sequencing*. Genome Biology, 2011.

[7] KIRCHER M, WITTEN DM, JAIN P, O'ROAK BJ, COOPER GM, SHENDURE J. *A general framework for estimating the relative pathogenicity of human genetic variants.*Nature Genetics, 2014

[8] QUANG D, CHEN Y, XIE X. *DANN: a deep learning approach for annotating the pathogenicity of genetic variants*. Bioinformatics (Oxford. Print), 2015.

Assinaturas

Referente à proposta de trabalho apresentada neste documento, com o título:

Um estudo de desempenho das técnicas de aprendizagem de máquina na classificação de variantes genéticas quanto à patogenicidade.

Recife, Setembro 2016

_____

Renata Correia de Andrade

Aluna

_____

Paulo Salgado Gomes de Mattos Neto

Orientador