SALÃO DE INICIACAO CIENTIFICA XXX SIC

15 A 19 OUTUBRO CAMPUS DO VALE

UFRGS PROPESQ

Desenvolvimento de um Banco de Dados Curado para Pesquisas em Aprendizado de Máquina

Eduardo Bassani Chandelier e Márcio Dorn

Universidade Federal do Rio Grande do Sul (UFRGS)

Introdução

O câncer é a maior causa de mortes naturais, levando a óbito mais de 8 milhões de pessoas anualmente. A geração de uma enorme quantidade de dados genômicos e transcriptômicos revolucionou a medicina do século XXI, auxiliando no entendimento de doenças como o câncer. Por outro lado, essa grande quantidade de dados trouxe um novo problema: a dificuldade de extração de dados biológicos relevantes. Neste sentido, técnicas de bioinformática tem se tornado a principal ferramenta para processamento e análise de dados ômicos. Entretanto, mesmo as técnicas padrão da bioinformática são muito custosas. Neste contexto, se encaixa o Aprendizado de Máquina *(AM)*, um conjunto de técnicas computacionais que possibilitam ao computador aprender utilizando dados como *input*, e com isso gerando resultados com menor custo computacional, quando comparado com as técnicas padrão. Um crescente esforço tem sido desenvolvido nos últimos anos na aplicação do AM para a biologia do câncer. Todavia, a maioria desses trabalhos não disponibiliza os dados utilizados, não são padronizados, acurados ou são antigos - além de não informarem como foram tratados antes das análises. Desta forma, visando otimizar, padronizar e auxiliar os estudos de AM aplicados ao estudo do câncer, é observada a necessidade de criação de um banco de dados ômico de câncer de qualidade.

Objetivo

Reunir dados ômicos de diversos tipos de câncer, curá-los, validar a qualidade dos dados usando técnicas básicas de AM e disponibilizá-los online para a comunidade científica.

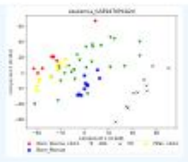
Métodos

Os dados foram coletados da plataforma NCBI-GEO (https://www.ncbi.nlm.nih.gov/gds/), utilizando um rigoroso filtro de busca e tratados com *background correction*, normalizados e manualmente curados com base na relevância das sondas. Os filtros de busca foram os seguintes: (i) exclusão de estudos que utilizam moléculas inibidoras ou terapêuticas de qualquer natureza; (ii) exclusão de estudos *knockdown* ou com modelos mutantes; (iii) estudos apenas feitos em humanos; (iv) exclusão de estudos com menos de 6 amostras de tecido para cada classe; (v) exclusão de estudos xenográficos; (vi) apenas estudos que disponibilizavam o dado bruto; e (vii) exclusão de estudos sem uma clara descrição do método.

Foram gerados 104 arquivos brutos de 14 tipos de câncer diferentes, cada arquivo contendo de 12 a 288 amostras, sendo no mínimo 6 por classe. Os dados foram empregados para gerar arquivos em formatos conhecidos e de fácil processamento por softwares de AM, são eles CSV, ARFF, GCT e CLS. Este banco estará disponível através do site (http://sbcb.inf.ufrgs.br/home) para a utilização pública.

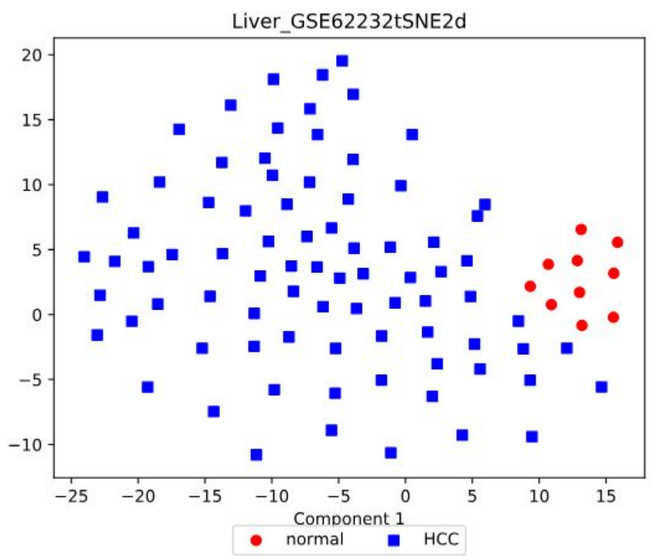
Junto aos dados estão disponíveis as descrições completas e *benchmarks* dos mesmos, para fins de validação ou comparação. Os *benchmarks* serão rodados em cada arquivo gerado. Já entre as técnicas a serem usados para verificar a qualidade dos dados estão algoritmos de Seleção, como *Principal Components* (PCA), e de classificação, como *Support Vector Machine* (SVM). Futuramente será o desenvolvimento de técnicas que utilizem o banco de dados e a comparação com os resultados do *benchmarks* presentes no mesmo. Como existe uma série de *frameworks* para criar algoritmos de AM em GPU (*Graphics Processing Unit*), devido a sua alta capacidade de processamento paralelo, será investigado quais as ferramentas cada um disponibiliza para decidir qual é o mais adequados para o problema, para então decidir quais algoritmos podem ser usados.

Resultados - Interface Web para a pesquisa

Type	GSE	#Samples	#Genes	#Classes	Downloads	PCA	t-SNE	Baseline	SVM
Leukemia	GSE9476	64	22283	5	arff gct cls csv			0.41	0.98

Este é um dos exemplos dos mais de 60 bancos de dados, nele podemos ver que o câncer examinado é a Leucemia, o código do banco, o número de amostras (ou pacientes), a quantidade de genes em cada amostras, a quantidade de classes dentro das amostras, (na maioria dos casos são duas, com ou sem câncer), os arquivos disponíveis para download, algumas imagens ilustrando os dados em 2D para auxiliar na visualização, o *baseline* do *dataset* utilizando como avaliação o algoritmo *ZeroR e o SVM,* calculado no *WEKA,* um *software* de AM que utiliza um dos formatos gerados, o ARFF. Ainda, junto aos dados no site, estarão instruções de como usar os dados, definição dos formatos dos arquivos, descrição dos métodos utilizados para visualização e *baseline*.

Visualização com Dimensão Reduzida



Antes de qualquer estudo envolvendo dados, é importante uma análise prévia dos mesmos, como a análise da distribuição das variáveis ou correlações entre elas. A imagem ao lado representa o resultado de um dos algoritmo de redução de dimensionalidade utilizados para visualização dos dados em 2D, são eles o PCA e o t-SNE.

Conclusão

Esses dados já estão sendo utilizados pelo grupo de pesquisa *Structural Bioinformatics and Computational Biology* (SBCB) do Instituto de Informática da UFRGS. A liberação dos dados *online* facilitará o seu uso interna e externamente ao grupo. Com isso todos terão acesso a dados ômicos de câncer confiáveis para pesquisas em aprendizado de máquina.
Futuramente será adicionado em uma coluna na tabela de samples com outras classificações, além do baseline, e suas respectivas fontes explicando os métodos, assim como sua acurácia. Atualmente alguns dos métodos que o grupo utiliza para classificar esse tipo de dado são SVM, Redes Neurais e Neuroevolução. Estes dados podem ser utilizados para classificação, descoberta de biomarcadores e *clustering*, entre outros.

SBCB Structural Bioinformatics and Computational Biology Lab

INF INSTITUTO DE INFORMÁTICA UFRGS

ebchandelier@inf.ufrgs.br