Mineração de Dados em Biologia Molecular

Sobre o curso

Docente: André C. P. L. F. de Carvalho
PAE: Kemilly Dearo

Por que esse curso?

- Quantidade crescente de dados biológicos
  - Anima
    - Permitiram e vão permitir várias descobertas relevantes
      - Extração de conhecimento dos dados
  - Preocupa
    - Como extrair conhecimento desses dados
    - Falta de mão de obra especializada

Por que esse curso?

- Dados biológicos
  - Contêm conhecimento relevante sobre vários processos biológicos
    - Análise de expressão gênica
      - Baseada em *microarray* e *RNA-seq*
    - Funções de proteínas
    - Biomarcadores em sequências
  - Ferramentas para análise de dados
    - Mineração de dados

Por que esse curso?

- Extração de conhecimento
  - Novas tecnologias para armazenar, transmitir e processar esses dados
    - Big data
  - Novas técnicas para extrair conhecimento dos dados
    - Ciência de dados (*analytics*)
      - Mineração de dados

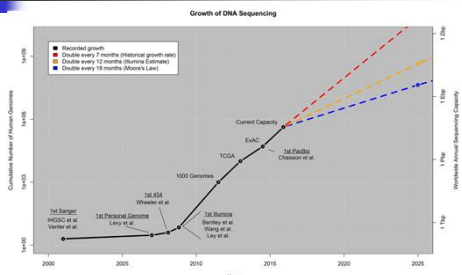
Quadro domínios para Big Data

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

Crescimento de DNAs sequenciados



http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

Objetivo do Curso

Fornecer ao aluno os principais conceitos de Mineração de Dados, enfatizando as técnicas mais utilizadas para aplicações práticas em Biologia Molecular

Ementa do curso

- Introdução a Mineração de Dados
- Preparação dos dados
- Redução de dados
- Planejamento de experimentos
- Técnicas para classificação de dados
- Técnicas para agrupamento de dados
- Análise de resultados
- Aplicações em biologia

Tópicos do Curso

- Ciência de dados
- Big Data
- Descoberta de conhecimento em bases de dados
- Análise e preparação dos dados
- Pré-processamento dos dados
- Mineração de dados
- Aplicações em Biologia Molecular

Final do curso

- Entender os principais passos da descoberta de conhecimento de conjuntos de dados biológicos
- Ser capaz de sumarizar conhecimento presente em um conjunto de dados biológicos
- Compreender o processo de preparação dos dados
- Ser capaz de produzir e analisar modelos utilizando técnicas de mineração de dados
- Ler e compreender artigos científicos sobre mineração de dados em biologia molecular

Exercícios

- Por em prática o que for visto durante o curso
  - Preparação de dados
  - Implementação
  - Realização de experimentos
  - Análise de resultados
  - Bem escrito

Material didático

- Conteúdo dos capítulos cobertos nos livros indicados
  - Ou outros livros de MD, que cubram os tópicos visto
- Os slides do curso têm tudo, menos o essencial

Projeto

- Utilizar MD para resolver problema real
  - Dados públicos
  - Detalhes a serem definidos depois

Etiqueta de aulas

- Chegar no horário da aula
- Pedir licença para entrar e sair da sala
- Usar palavras "mágicas" por favor, com licença, obrigado (a) e desculpe
- Não conversar durante a aula
- Levantar o braço para fazer perguntas e comentários
- Não ler outro material durante a aula
- Desligar celular durante a aula
- Colocar lixo no lixo
- Não copiar de colega ou site material a ser avaliado

Código de honra

- A nota de cada aluno deve ser baseado exclusivamente na avaliação de seu esforço e trabalho pessoal
  - Qualquer forma de conversa em exames, cola ou plagio em trabalhos, constitui-se em fraude
    - Punições previstas no Regime Disciplinar da USP
    - O código de honra deverá ser rigorosamente seguido sob pena de anulação de trabalhos ou prova e instalação de processo na Universidade
    - http://www.prg.usp.br/wp-content/uploads/manual_disciplinar_web2.pdf

Avaliação

- Provas:
  - Duas provas normais
  - Provas curtas, uma após cada duas aulas
- Avaliação oral
  - Alunos podem ser perguntados em aula, valendo nota
- Trabalhos:
  - Um por laboratório e um projeto final

Datas

- Prova 1: 26/09
- Prova 2: 21/11
- Projeto: 30/11
- Substitutiva: não tem
- Recuperação: só se ficar alguém

Cálculo da Média

- Sejam
  - MC = Média Chamada
  - MP = Média Aritmética das Provas
  - NT = Nota Trabalho
  - MF = Média Final
  - Se $MP \geq 5$ e $NT \geq 5$ $\rightarrow$ MF = (4MP + 3MC + 3MT) /10 ou (6MP + 4NT) / 10
  - Se $MP < 5$ ou $NT < 5$ $\rightarrow$ MF = menor valor entre MP, MC e NT

Recuperação

- Só terão direito à recuperação os alunos com **3.0 ≤ MF ≤ 5.0** e freqüência superior a 70%
- Observação:
  - Será dada a nota 0.0 (zero) para cópia parcial de programa ou prova, sendo o problema levado para a coordenação do curso

Livros para o Curso

- K. Faceli, A. Lorena, J. Gama, J. e A. de Carvalho: Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina, Editora LTC, 2011 (2nd edição).
- I. H. Witten e E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman, 2011 (terceira edição).
- V. Kumar, M. Steinbache e Pang-ning Tan, Mineração de Dados (Introduction to Data Mining), Ciência Moderna, 2009.
- M. Zaki e W. Meira Jr.: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, May 2014;

Pacotes e Conjuntos de Dados

- RapidMiner (Studio 7.1)
  - https://rapidminer.com/
- R
  - http://lancet.mit.edu/ga
- Machine Learning Data Repository UC Irvine
  - http://www.ics.uci.edu/~mlearn/ML/Repository.html

Perguntas