UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

# UMA ABORDAGEM SUPERVISIONADA SOBRE A CLASSIFICAÇÃO DA *"SACCHAROMYCES CEREVISIAE"*.

## PROPOSTA DE TRABALHO DE GRADUAÇÃO

**Aluno:** Marcelo Henrique Cavalcanti Jucá (mhcj@cin.ufpe.br)
**Orientador:** Francisco de Assis Tenório de Carvalho(fatc@cin.ufpe.br)
**Co-Orientador**: Valdir de Queiroz Balbino (vqbalbino@yahoo.com.br)

26 de Maio de 2004

# INDICE

# 1. INTRODUÇÃO

No campo computacional, uma grande quantidade de dados nem sempre está disposta de forma organizada para fornecer informações relevantes. A enorme velocidade de geração de dados compete com a rapidez de assimilação dos mesmos. Portanto, em alguns momentos, para que um determinado conjunto de dados torne-se informativo, é necessário submetê-lo a uma filtragem para torná-lo útil e compreensível em futuras análises (HAND *et al.*, 2001).

Há situações em que os dados não são fidedignos, estão incompletos, ou até mesmo não existem (KALAPANIDAS *et al.*, 2003). Por exemplo, suponha o preenchimento de um formulário para uma pesquisa interna a uma empresa que deseja traçar o perfil de seus funcionários. Se perguntas deste formulário do tipo "Qual é a quantidade de horas realmente trabalhadas por dia?" tiverem que ser respondidas diariamente, pode ocorrer de algumas respostas serem mascaradas, evitando-se a ocorrência de uma possível demissão. Porém, outras respostas podem se mostrar incompletas para perguntas do tipo *"Qual é o custo do material escolar da sua prole?*". Outras ainda, por exemplo, *"Quantas vezes você viaja a trabalho para o exterior?"*, podem nem sequer existir para determinado perfil de funcionário.

Neste contexto, Hand *et al.* (2001) afirmam que nem todos os dados são relevantes para uma análise. Deve-se, portanto, saber escolhê-los e para isto é necessária a participação de especialistas para avaliar quais são os atributos significativos em um conjunto de dados (WITTEN & FRANK, 2000).

Ainda sobre o exemplo citado anteriormente, caso os formulários da empresa fictícia sejam recolhidos após um ano de pesquisa para descobrir o perfil dos funcionários no contexto de quantidade de horas trabalhadas e custos extras para a empresa, pode ser que sejam encontradas tanto informações inúteis ("Viajam ao exterior funcionários que trabalham no mínimo sete horas por dia."), quanto informações significativas ("Funcionários que recebem menos de que dez salários mínimos aceitam fazer hora-extra.").

Percebe-se, então, que em situações como esta, faz-se necessário uma avaliação mais detalhada para evitar dados que possam mascarar, poluir ou mesmo descaracterizar o resultado de uma análise (KALAPANIDAS *et al.*, 2003).

No campo biológico, com a evolução dos tempos, cresce a quantidade de dados gerados e armazenados por instituições como o "National Center for Biotechnology Information" e "European Molecular Biology Laboratory". Este rápido crescimento do volume de dados sobre genes seqüenciados tem levado os biólogos a procurarem novas metodologias para extração de informação sobre estes dados.

O conhecimento completo ou quase completo de um genoma sugere um estudo sobre a expressão gênica de cada um dos genes que o compõe (EISEN *et al.*, 1998).

Segundo os estudos de Chu *et al.* (1998), DeRisi *et al.* (1997) e Spellman *et al.* (1998), medições do nível da expressão gênica em instantes separados por intervalos de tempo pré-determinados foram feitas através do monitoramento de determinados processos biológicos, com o uso de microarrays de cDNAs.

Os dados resultantes destes experimentos têm ajudado a identificar padrões na expressão gênica, ou seja, genes co-expressos (EISEN *et al.*, 1998; HEYER *et al.*, 1999) e avaliar os métodos utilizados para encontrar estes padrões (FILHO, 2003). Além do mais, segundo Eisen *et al.* (1998) e Spellman *et al.* (1998), genes co-expressos possuem funções relacionadas. Já Heyer *et al.* (1999) afirmam que estes padrões podem revelar informações sobre genes de sistemas regulatórios.

Percebe-se, portanto, que a Biologia aliada à Informática está conquistando uma importância cada vez maior na área das ciências aplicadas à saúde.

O termo Biotecnologia refere-se a um campo da tecnologia que envolve o processamento de material por meios biológicos (biotecnologia "tradicional", como a fermentação utilizada na preparação da cerveja e do pão) ou computacionais (biotecnologia "moderna", como a produção de vacinas através de estudos que manipulam dados biológicos como o seqüenciamento genético).

Esta área do conhecimento detém a capacidade de contribuir para a melhoria da produtividade, adicionando maior valor agregado, a produtos que pertençam a diversos setores da economia tais como as indústrias farmacêuticas, agrotóxicas, alimentícias, químicas, entre outras.

No contexto do seqüenciamento genético e dos procedimentos afins, observa-se que a Biotecnologia vem produzindo uma quantidade enorme de dados que necessitam ser avaliados.

A Informática, por sua vez, é, então, requisitada para entrar em cena oferecendo seus serviços. Especificamente a subárea da Informática conhecida por "Aprendizagem de Máquina", possibilita a identificação de padrões nos dados, os quais podem ajudar o estudo sobre os organismos seqüenciados.

# 2. OBJETIVOS

## *2.1. OBJETIVO GERAL*

Este trabalho aborda metodologias na área de Aprendizagem de Máquina e avaliará as suas utilizações na Bioinformática, mais especificamente para comparar a performance de métodos supervisionados de Aprendizagem de Máquina que identificam padrões em dados sobre a levedura ***Saccharomyces cerevisiae***.

Para tanto, o estudo realizado em Filho (2003) além de ser utilizado como incentivo, pois fez uma análise de dados nas duas áreas citadas, serviu como referencial e fonte de dados para uma abordagem supervisionada de Aprendizagem de Máquina sobre os mesmos dados utilizados pelo autor.

### *2.2. OBJETIVOS ESPECÍFICOS*

Buscar verificar a utilidade desses resultados no campo da Bioinformática.

# 3. MATERIAL E MÉTODOS

Neste trabalho, será feita uma comparação da performance de alguns esquemas de aprendizagem supervisionada de Aprendizagem de Máquina, através da análise de taxas de erro produzidas por cada um dos modelos (resultado do esquema), em cima de dados biológicos provenientes do organismo ***Saccharomyces cerevisiae***.

Na realidade, estes dados compõem um subconjunto dos utilizados por Filho (2003) que deu uma abordagem não supervisionada, para conseguir clusterizá-los.

Observa-se um caráter multidisciplinar neste trabalho pelo fato de a literatura necessária para este estudo ser de áreas distintas. Informações referentes ao organismo e aos métodos biológicos e computacionais serão exploradas para se ter um conhecimento melhor sobre os dados manipulados.

Para o processamento dos dados, Witten & Frank (2000) fornece o software escolhido para auxiliar este trabalho. O software Weka, como é conhecido, fornecerá esquemas de aprendizagem de Aprendizagem de Máquina já implementados que, a partir de uma entrada, gerarão modelos como saída. Uma das vantagens observadas, na ferramenta, foi a possibilidade de analisar informações através de dados numéricos ou gráficos, além de ser implementada em código aberto na linguagem JAVA.

No final, será feita uma comparação das taxas de erro de cada classificador utilizado.

### *3.1. POSSÍVEIS LIMITAÇÕES EM RELAÇÃO AOS OBJETIVOS.*

Este projeto será desenvolvido sob a supervisão do Dr. Francisco, Professor Adjunto do Centro de Informática (CIn) da UFPE, assistido pelo Dr. Valdir de Queiroz Balbino, Professor Visitante do Departamento de Genética desta mesma universidade. As atividades do projeto serão desenvolvidas nos Centro de Informática de Pernambuco e de Genética Molecular (CCB/UFPE), que já dispõem de toda a infra-estrutura necessária para o bom andamento das atividades previstas.

### *3.2. RELEVÂNCIA E IMPACTO DO ESTUDO*

O crescimento da genômica no Brasil resultou na produção de uma grande quantidade de informações derivadas dos vários programas de seqüênciamento genômico conduzidos nos últimos anos. Tem-se observado, no entanto, que a quantidade de pesquisadores devidamente habilitados para a análise das informações advindas dos programas de seqüenciamento ainda é bastante limitada e evidencia a necessidade de investimentos maciços na formação de recursos humanos, a exemplo do que ocorre nos países desenvolvidos.

Na presente proposta, pretende-se utilizar as bases da tecnologia da informação no desenvolvimento e implementação de um software que possa ser utilizado no estudo de organismos que tiveram seus genomas completamente seqüenciados no Brasil, a exemplo da *Xilela fastidiosa* e da *X. campestris*. Espera-se, assim, contribuir para a formação de recursos humanos na área de bioinformática, que contribuirão para a consolidação definitiva desta área de conhecimento no Estado de Pernambuco.

# 4. CRONOGRAMA

O cronograma abaixo demonstra algumas datas para as atividades chaves:

| | **Mês** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Atividade** | ***Maio*** | | | | ***Junho*** | | | ***Julho*** | | | | ***Agosto*** | | | |
| 1) Entendimento do problema abordado por Filho (2003). Assistir as aulas da disciplina Aprendizagem de Máquina. | 1 | 2 | | | | | | | | | | | | | |
| 2) Instalação e adaptação ao Weka. Assistir as aulas da disciplina Aprendizagem de Máquina. | | | 3 | 4 | | | | | | | | | | | |
| 3) Assistir as aulas da disciplina Aprendizagem de Máquina. | | | | | 5 | 6 | 7 | 8 | 9 | | | | | | |
| 4) Experimento. Assistir as aulas da disciplina Aprendizagem de Máquina. | | | | | | | | | | 10 | 11 | | | | |
| 5) Escrita da monografia. Assistir as aulas da disciplina Aprendizagem de Máquina. | | | | | | | | | | | | 12 | 13 | 15 | 16 | 17 |

# 5. REFERÊNCIAS BIBLIOGRÁFICAS


BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. Recife, Pernambuco, Brasil: Editora Saraiva, 2003.

CHO, R.; CAMPBELL, M.; WINZELER, E.; STEINMETZ, L.; CONWAY, A.; WODICKA, L.; WOLFSBERG, T.; GABRIELIAN, A.; LANDSMAN, D.; LOCKHART, J.; DAVIS, W. **A genomewide transcriptional analysis of the mitotic cell cycle**, Molecular Cell, v.2, jul. 1998. p.65-73.

CHU, S.; DELRISI, J.; EISEN, M.; MULHOLLAND, J.; BOTSTEIN, D.; BROWN, P.O.; HERSKOWITZ I. **The Transcriptional Program of Sporulation in Budding Yeast**. Science, v.282, out. 1998. p.699-705.

CYGD. Comprehensive Yeast Genome Database. **Munich Information Center for Protein Sequences**. Disponível em: <http://mips.gsf.de/genre/proj/yeast/index.jsp> Consultado em: 20 julho 2004.

DERISI, J. L.; IYER V. R.; BROWN P. O. **Exploring the metabolic and genetic control of gene expression on a genomic scale**. Science, v. 278, out. 1997. p.680-686.

EISEN, M. B.; SPELLMAN, P. T.; BROWN, P. O.; BOTSTEIN, D. **Cluster analysis and display of genome-wide expression patterns**. United States of America: Proc. of National Academy of Sciences, v. 95, dez. 1998. p.14863-14868.

FARAH, S. B. **DNA: Segredos & Mistérios**. São Paulo, Brasil: Sarvier, 1997. p.255-272.

FILHO, I. G. Costa. **Comparative Analysis of Clustering Methods for Gene Expression Data**. Recife: UFPE, 2003. Dissertação de Mestrado.

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. **Principles of Data Mining** - Adaptive Computation and Machine Learning. New Jersey, United States of America: Bradford Books, 2001. p.1-9.

HEYER, L. J.; KRUGLYAK, S.; YOOSEPH, S. **Exploring expression data:** identi?cation and analysis of coexpressed genes. Genome Research, v.9, 1999. p.1106-1115.

KALAPANIDAS, Elias; AVOURIS, Nikolaos; CRACIUN, Marian; NEAGU, Daniel. **Machine Learning algorithms:** a study on noise sensitivity. Thessaloniki, Greece: 1st Balcan Conference in Informatics, 2003.

KELLER, Frank. Introduction to Machine Learning. - Connectionist and Statistical Language Processing. **School of Informatics**. Disponível em: <http://homepages.inf.ed.ac.uk/keller/teaching/connectionism/lecture8_4up.pdf> Consultado em: 20 julho 2004.

MITCHELL, Tom M. **Machine Learning**. New York, United States of America: McGraw-Hill, 1997.

RUSSEL, Stuart J.; NORVIG, Peter. **Artificial Inteligence** – A Modern Approach. New Jersey, United States of America: Prentice-Hall, 1995. p.523-647.

SPELLMAN, P. T.; SHERLOCK, G.; ZHANG, M. Q.; IYER, V. R.; ANDERS, K.; EISEN, M. B.; BROWN, P. O.; BOTSTEIN, D.; FUTCHER, B. **Comprehensive Identi?cation of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hibridization**. Molecular Biology of the Cell, v.9, dez. 1998. p. 3273-3297.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining** – Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, CA, United States of America: Morgan Kaufmann Publishers, 2000.

Assinaturas

Francisco de Assis Tenório de Carvalho(fatc@cin.ufpe.br)
(Orientador)

Valdir de Queiroz Balbino (vqbalbino@yahoo.com.br)
(Co-orientador)

Marcelo Henrique Cavalcanti Jucá (mhcj@cin.ufpe.br)
(Aluno)