Análise integrada de dados ambientais e de expressão gênica usando técnicas de mineração de dados

Heloisa Musetti Ruivo, Fernando M. Ramos

[1]Laboratório Associado de Computação e Matemática Aplicada
LAC. Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos/SP, Brasil

Abstract. *Data analysis of large experimental data sets became an important practice in life and natural sciences. This tendency requests that statistical and computational approaches begin to assume a position of great prominence within the molecular biology community [Amaratunga and Cabrera 2004]. The objective of this work is the identification of groups of genes (clustering), based in the existent similarities among their expression profiles through a computacional tool. The same tool will be used for analysis of limnology data coming from dams, and climatological data.*

Resumo. *A análise simultânea de grandes quantidades de dados experimentais tornou-se uma importante prática nas ciências da vida. Esta tendência requer cada vez mais o emprego de técnicas computacionais e estatísticas avançadas de análise e interpretação destes dados. Neste trabalho será investigado o problema da identificação de grupos de genes (clustering), com base nas semelhanças existentes entre os seus perfis de expressão através de uma ferramente computacional. A mesma ferramenta será utilizada para análise de dados limnológicos provenientes de represas, e dados climatológicos.*

1. Introdução

A análise de dados tornou-se repentinamente uma regra na ciência da vida. No começo, a ciência produzia uma quantidade limitada de dados mas a biologia tornou-se nestes últimos anos a ciência que constantemente gera esta imensidão numérica. Munidos deste grande número de dados biológicos e da informação contida neles torna-se necessário técnicas computacionais e estatísticas para análise e interpretação. Manipulá-los tornou-se então um dos maiores desafios da Bioinformática [Amaratunga and Cabrera 2004].

Na biologia molecular experimental, por exemplo, os Microarranjos (ou Microarrays - MA) são, hoje em dia, uma das tecnologias chave em estudos genômicos. Os MA permitem um monitoramento simultâneo de níveis de expressão de milhares de genes. Através da análise destes dados é possível agrupar os genes com base nas semelhanças existentes entre os seus perfis de expressão nas diversas condições analisadas. Esta análise permite diagnosticar e classificar tumores em subgrupos relevantes bem como predizer o tipo de tumor em novos pacientes que sejam portadores de tumores desconhecidos.

Avaliando a análise estatística aplicada na biologia molecular e seus resultados, utilizou-se o mesmo conceito para análise de bancos de dados com estrutura semelhante. Com isto, foi feita a aplicação da mesma ferramenta em dados ambientais e climatológicos. Os dados ambientais foram escolhidos devido à preocupação com as mudanças climáticas globais em decorrência da crescente emissão de gases de efeito estufa (GEE)

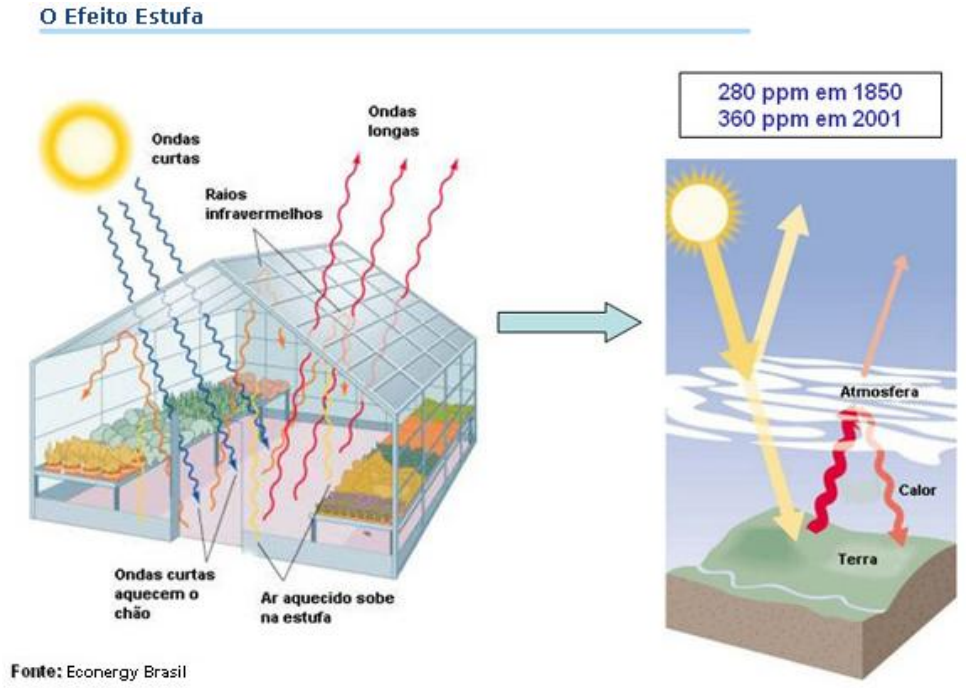

Figure 1. O Efeito Estufa

(Fig. 1). O projeto analisado é o ”Balanço de Carbono nos Reservatório de FURNAS Centrais Elétricas S.A.”.

Os dados climatológicos forma escolhidos com o objetivo de entender a seca na Amazônia em 2005 (figura 2). Segundo [J. A. Marengo 2006], a região da Amazônia passou pelo mais intenso episódio de aridez dos últimos 100 anos. A seca deixou centenas de pessoas sem alimentação devido à falta de transporte, agricultura, geração de energia por hidrelétrica e tambémn afetou diretamente e indiretamente a população que mora na corrente do Rio Amazonas.

Este impacto ecológico afetou a sustentabilidade da atividade da floresta na região, que é atualmente um ítem promissor baseado na economia regional.

Figure 2. Seca do Amazonas - 2005 (fonte: [?]).

2. Tecnologia e Resultados

Neste trabalho estudaremos em detalhe o pacote **BRB-ArrayTools** versão $3.4.0$, desenvolvido pelo *Biometric Research Branch of the Division of Cancer Treatment and Diagnosis of the National Cancer Institute*, sob a direção do Dr. Richard Simon. Trata-se de um software livre, voltado para análise de dados de MA de DNA.

BRB-ArrayTools contém utilitários para processar dados de expressão em vários experimentos, visualizá-los, agrupá-los, classificá-los, dentre outras funções. O software foi desenvolvido por estatísticos experientes em análise de dados de MA, mas possui uma interface gráfica que facilita a utilização por biólogos.

A seguir serão apresentados os resultados obtidos com a aplicação do software proposto.

2.1. Biologia

Na aplicação em biologia, foram analisados dados de expressão gênica de 27 tumores de próstata gerados com um MA contendo cerca de 4.000 spots de acordo com [E. M. Reis 2004], cujos dados permitiram a identificação de uma assinatura de expressão gênica contendo 56 genes que separa grupos de amostras de tumores de próstata em função de uma importante característica histopatológica (Grau de Gleason - GS) para o prognóstico de evolução do câncer de próstata. Tumores com GS baixo (GS $\leq$ 6) são geralmente menos agressivos, enquanto tumores com GS $\geq$ 8 são mais agressivos [E. M. Reis 2004].

Os resultados obtidos com o pacote BRB, não obstante seguirem uma outra abordagem, produziram resultados semelhantes aos de [E. M. Reis 2004]. Este resultado confirma que o software BRB permite obter resultados comparáveis com outras abordagens de análise.

2.2. Limnologia

Serão analisadas as emissões de gás carbônico e metano dos reservatórios, onde pretende-se identificar os parâmetros mais relevantes dentre os medidos pelas instituições envolvidas. Segundo [Parekh 2004], estima-se que os reservatórios absorvam 2.5% da emissão de carbono global antropogênica. No entanto, as emissão globais de carbono por reservatórios, tem sido 60% superiores às estimadas, e isto tem limitado a expectativa de vida.

Foram analisadas 166 variáveis e definidos os seguintes campos para comparação:

- Fluxo CH_4 (bolha), interface água-atmosfera,
- Fluxo CO_2 (bolha), interface água-atmosfera Interface,
- Fluxo CH_4, interface sedimento-água CO_2,
- Fluxo CO_2, interface sedimento-água CH_4.

2.3. Climatologia

Para analisar a grande seca do Amazonas ocorrida em 2005, foram utilizados dados provenientes de diversas origens.

Os índices analisados são provenientes de variabilidades climáticas ocorridas em regiões globais. Pretende-se identificar os índices mais relevantes diante da grande massa

de dados. O parâmetro analisado foi a série de vazão do rio Amazonas em local próximo à Óbidos. Os resultados forma comparadodos aos publicados em [J. A. Marengo 2006] e apresentaram uma semelhança consistente, confirmando a eficácia do software.

References

Amaratunga, D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley Interscience, USA.

E. M. Reis, e. a. (2004). Antisense intronic non-coding rna levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene*, 23:6684–6692.

J. A. Marengo, C. A. Nobre, J. T. M. D. O. G. S. O. R. O. H. C. L. M. A. I. F. B. (2006). The drought of amazonia in 2005.

Parekh, P. (2004). A preliminary review of the impact of dam reservoirs on carbon cycling.