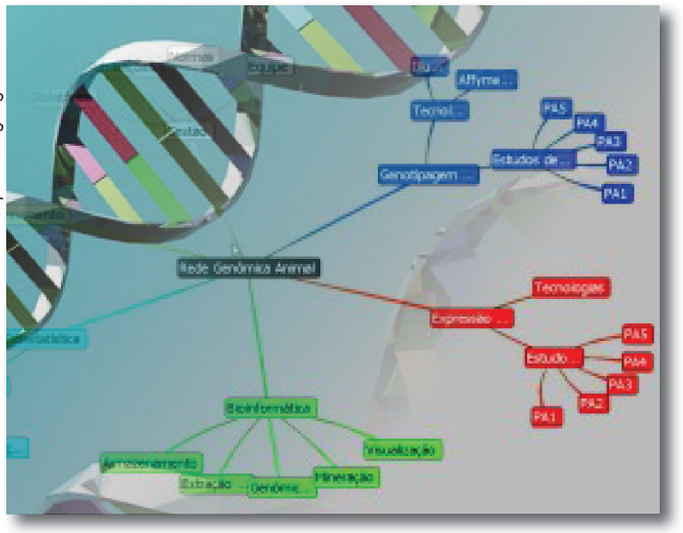
Comunicado Técnico 109

ISSN 1677-8464
Dezembro, 2011
Campinas, SP

Fotos: <http://www.google.com.br>



Análise de conjunto de genes de dados de expressão gênica na Rede Genômica Animal

Roberto Hiroshi Higa[1]
Adriana Mércia Guarantini Ibelli[2]
Fernando Flores Cardoso[3]
Luciana Correia de Almeida Regitano[4]

Introdução

Experimentos de expressão gênica por microarranjos medem a abundância de transcritos em escala genômica, constituindo-se numa importante ferramenta para a caracterização de perfis de atividade gênica entre amostras biológicas sujeitas a diferentes condições. Em particular, na agricultura, a identificação dos genes que regulam manifestações de fenótipos de interesse econômico é de grande interesse para os programas de melhoramento animal e vegetal.

No projeto “Rede Genômica Animal” (RGA) (Projeto SEG/MP1: 01.06.09.001), a análise de genes diferencialmente expressos é reralizada, conforme apresentado em (CARDOSO, 2009). Resumidamente, os pacotes R (R DEVELOPMENT CORE TEAM, 2010) Affy e Manova são utilizados para realização dos processo de controle de qualidade, correção de background, normalização, sumarização e a análise de expressão diferencial, propriamente dita.

Na análise de expressão diferencial, as variações de expressão são analisadas gene-a-gene, selecionando-se aqueles com valores de *fold-change* e significância estatística corrigida para múltiplos testes superior a limiares previamente especificados. Contudo, esse tipo de abordagem não considera as interações entre múltiplos genes envolvidos em um mesmo processo biológico, o que muitas vezes induz falsos negativos.

Para eliminar a dependência de limiares, vários métodos têm sido propostos para analisar variações de conjuntos de genes relacionados, genericamente, aqui denominados Gene Set Analysis (GSA). Esses conjuntos de genes, em geral, constituem vias ou redes gênicas e refletem o conhecimento biológico sobre as relações entre esses genes. Uma comparação entre diferentes estratégias de GSA pode ser encontrada em (SONG; BLACK, 2008).

O objetivo deste trabalho é apresentar a análise de conjuntos de genes para dados de experssão gênica, conforme utilizada no escopo da RGA. Na seção 2 é apresentado o procedimento correspondente à análise de GSA proposta por (EFRON; TIBSHIRANI, 2007), utilizado na RGA. Um exemplo completo, incluindo prin-

[1] *Doutor em Engenharia Elétrica, Pesquisador da Embrapa Informática Agropecuária, Campinas, SP, roberto@cnptia.embrapa.br*
[2] *Bolsista Capes da Universidade Federal de São Carlos, São Carlos, SP, adriana.ibelli@gmail.com*
[3] *Doutor em Bioinformática, Pesquisador da Embrapa Pecuária Sul, Bagé, RS, fcardoso@cppsul.embrapa.br*
[4] *Doutora em Genética e Melhoramento, Pesquisadora da Embrapa Pecuária Sudeste, São Carlos, SP, luciana@cppse.embrapa.br*

Embrapa

cipais comandos do correspondente *script* R é apresentado na seção 3.

Procedimento de análise de conjunto de genes

O problema abordado pela análise de conjuntos de genes consiste em detectar variações estatisticamente significativas na expressão de conjuntos de genes. Os conjuntos de genes precisam ser definidos a priori e, usualmente, utiliza-se informações de anotação de genes publicamente disponíveis, tais como GO (ASHBURNER et al., 2000), KEGG (KANEHISA; GOTO, 2000) e bioCarta (BIOCARTA, 2011). Considera-se que genes com anotações em comum definem conjuntos de genes, possivelmente com alguma sobreposição, envolvidos em um mesmo processo biológico. Os métodos propostos diferem quanto à estatística utilizada para atribuir pontuação aos conjuntos de genes e ao procedimento de cálculo da significância estatística associada com o teste de hipótese utilizado (SONG; BLACK, 2008).

O método GSA, proposto por (EFRON; TIBSHIRANI, 2007) e implementado no pacote R de mesmo nome, consiste em um refinamento do método Gene Set Enrichment Analysis (GSEA), o mais popular dos métodos de análise de conjunto de genes, proposto inicialmente por (MOOTA et al., 2003) e, posteriormente refinado por (SUBRAMANIAN et al., 2005). Especificamente, GSA é realizado por meio do seguinte procedimento:

1. Calcule a estatística z_i para cada gene (ex: para dados com duas classes, a estatística t). Seja z_s o vetor de valores z_i para os genes no conjunto de genes *S*.
2. Dado um conjunto de genes *S*, calcule estatísitca $S = s(z)$, que sumariza o conjunto de genes *S*. A estatística recomendada é a maxmean, dada por $S_{max} = max(max(z,0), - min(z,0)$ para todo z pertencente a *S*.
3. Calcule $S' = (S - mean_S)/stdev_S$, onde $mean_S$ e $stdev_S$ são a média e desvio padrão empíricos de s (padronização).
4. Calcule as permutações dos valores de saída (labels das classes para problemas com duas classes) e recompute S* para cada conjunto de dados permutado (B conjuntos permutados).
5. Calcule o p-valor: p_S = {número de vezes que um valor de S* excede o valor de S'}/B.
6. Repita os passos 2 a 5 para cada conjunto de genes.
7. Calcule a taxa de falsa descoberta (FDR) para os valores de p-valores encontrados para o pontuações dos conjunto de genes. No caso do pacote R GSA, utiliza-se o método de (BENJAMINI; HOCHBERG, 1995).

Exemplo de análise de conjunto de genes utilizando R

Para ilustrar a análise de conjuntos de genes utilizou-se um conjunto de dados de um experimento de extremos de resistência a infestação por carrapatos, analisados no âmbito da Rede Genômica Animal (IBELLI et al., 2010). Foram utilizadas amostras coletadas da pele (PE) de bovinos de 4 grupos genéticos (Nelore, Angus x Nelore, Simental x Nelore e Canchim x Nelore) previamente classificados como resistentes (R) e sensíveis (S). Os dados de expressão gênica foram obtidos utilizando-se a plataforma Affymetrix. Os pacotes R Affy e Maanova foram utilizados para realização dos processo de controle de qualidade, correção de background, normalização e sumarização (CARDOSO, 2009). Para esse conjunto de dados, nenhum gene diferencialmente expresso foi encontrado, considerando uma significância (q-valor) igual a 5%. Ao final, para a análise de conjuntos de genes, considerou-se as expressões normalizadas de todos os genes (24128) e amostras (36) que sobreviveram ao processo de controle de qualidade.

O passo inicial para realização da análise de conjuntos de genes, utilizando o pacote R GSA, é tornar suas funções disponíveis e ler o arquivo que descreve os conjuntos de genes. Os conjuntos de genes analisados foram aqueles formados pelas anotações GO (biological process) e KEGG, disponibilizados publicamente no sítio[5]

```
>library("GSA")
>geneset.obj<- GSA.read.gmt("arquivo_conjunto_genes.gmt")
```

Em seguida, executam-se os testes estatísticos utilizados por GSA, para os conjuntos de genes desejados.

[5] *Disponível em: <http://www.broadinstitute.org/gsea/>. Acesso em: 15 abr. 2010.*

A matriz dts contém os valores de expressão gênica normalizados e o parâmetro lab codifica os fenótipos, 1 para suceptível e 2 para resistente. Dentre as estatísitcas disponíveis no pacote R GSA, devido a sua robustez (EFRON; TIBSHIRANI, 2007), optou-se por utilizar a estatística maxmean.

```
>GSA.obj<-GSA(dts, lab, genenames=gnames,
     genesets=geneset.obj$genesets,
     method="maxmean", nperms=5000, ran-
     dom.seed=100, minsize=10, restand.
     basis="data", resp.type="Two class unpai-
     red")
```

Finalmente, a função GSA.listsets é utilizada para identificar os conjuntos de genes com variação de expressão entre os grupos de amostras, de acordo com um nível de significância FDR de acordo com Benjamini e Hochberg (1995) inferior a 1%.

```
>lst <-GSA.listsets(GSA.obj, geneset.
     names=(geneset.obj)$geneset.names,
     maxchar=30, FDRcut=.01)
```

O GSA apresenta os resultados em duas tabelas, uma com o conjunto de genes positivo e outra com o conjunto de genes negativo, onde negativo significa que "expressões mais baixas da maioria dos genes no conjunto de genes está correlacionada com valores maiores do fenótipo" enquanto positivo significa que "expressões mais altas da maioria dos genes em um conjunto de genes está correlacionada com valores maiores do fenótipo". Lembrando que os fenótipos foram codificados como 1 (suceptível) e 2 (resistente).

Como resultado, ao utilizar os conjuntos de genes definidos pelas anotações GO-bp, foram encontrados 9 conjuntos de genes significativos (Tabela 1).

A Tabela 2 apresenta os 8 conjuntos de genes encontrados ao se utilizar os conjuntos de genes definidos pelas vias anotadas em KEGG.

Por fim, cabe ressaltar que para esse conjunto de dados, nenhum gene diferencialmente expresso foi encontrado considerando q-valor < 5%, enquanto que utilizando a anaĺise de conjunto de genes obteve-se 9 conjuntos de genes com anotação GO-bp e 8 conjuntos de genes com anotação de vias KEGG, potencialmente implicados com o processo biológico em estudo.

Discussão

A análise de conjuntos de genes constitui uma alternativa importante às análises de expressão diferencial, caracterizadas pela análise da variação gene-a-gene. Ela incorpora conhecimento biológico prévio (conjuntos de genes relacionados) à análise estatística, o que pode favorecer a identificação de diferenças na expressão gênica (processos biológicos) difíceis de serem testados com base em análises gene-a-gene. Outra vantagem desse tipo de análise é o fato de não haver a

Tabela 1. Intervalo numérico Tabela: Conjuntos de genes definidos pela anotação GO-bp com variação sinificativa de expressão, de acordo com a estatística maxmean (GSA). Os nomes de conjuntos de genes foram limitados a 30 caracteres.

Conjunto de genes	Nome	Pontuação
76	CELL_ACTIVATION	-0.8278
255	HEMOPOIESIS	-0.6982
256	HEMOPOIETIC_OR_LYMPHOID_ORGAN_	-0.6715
305	LEUKOCYTE_DIFFERENTIATION	-1.0344
315	LYMPHOCYTE_ACTIVATION	-0.8386
316	LYMPHOCYTE_DIFFERENTIATION	-1.2676
501	POSITIVE_REGULATION_OF_IMMUNE_	-0.8885
634	REGULATION_OF_IMMUNE_SYSTEM_PR	-0.7239
642	REGULATION_OF_LYMPHOCYTE_ACTIV	-0.7597

Tabela 2. Intervalo numérico Tabela. Conjuntos de genes definidos pela anotação KEGG com variação sinificativa de expressão, de acordo com a estatística maxmean (GSA). Os nomes de conjuntos de genes foram limitados a 30 caracteres.

Conjunto de gene	Nome	Pontuação
128	HSA04012_ERBB_SIGNALING_PATHWA	0.658
169	HSA04810_REGULATION_OF_ACTIN_C	0.3831
184	HSA05120_EPITHELIAL_CELL_SIGNA	0.5945
189	HSA05212_PANCREATIC_CANCER	0.508
199	HSA05222_SMALL_CELL_LUNG_CANCE	0.6318
134	HSA04115_P53_SIGNALING_PATHWAY	0.8178
197	HSA05220_CHRONIC_MYELOID_LEUKE	0.675
195	HSA05218_MELANOM	0.5144

necessidade de utilização de um limiar de fold change, o que favorece a identificação de conjuntos de genes caracterizados por pequenas variações.

Neste trabalho abordou-se de conjuntos de genes, conforme utilizada no escopo da Rede Genômica Animal, ou seja, por meio da utilização do pacote R GSA com estatística maxmean (EFRON ; TIBSHIRANI, 2007). Contudo, cabe observar que existem diversas outras metodologias propostas na literatura, muitas delas com implementações para utilização em R (SONG; BLACK, 2008).

Espera-se, com a disseminação da análise apresentada neste trabalho, estimular outros pesquisadores interessados em realizar análise de dados de expressão gênica por microarranjos a explorar metodologias de análise de conjuntos de genes, como a aqui apresentada.

Referências

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: tool for the unification of biology: the Gene Ontology Consortium. **Nature Genetics**, New York, v. 25, n. 1, p. 25-9, 2000.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society**, London, v. 57, n.1, p.289-300, 1995.

BIOCARTA. 2011. Disponível em: <http://www.biocarta.com/>. Acesso em: 15 out. 2011.

CARDOSO, F. F. Métodos para análise de micro-arranjos de oligonucleotídeos em estudos devexpressão gênica. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 54.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 13., 2009. **Programa e resumos**... São Carlos, SP: UFSCar; Embrapa Pecuária Sudeste, 2009. p. 28.

EFRON, B.; TIBSHIRIANI, R. “On testing the significance of sets of genes”. **Annals of Applied Statistics**, Cleveland, v. 1, n. 1, p. 107-129, 2007. Disponível em: <http://www.jstor.org/stable/10.2307/4537424>. Acesso em:

IBELLI, A. M. G.; HIGA, R. H.; GIACHETTO, P. F.; YAMAGISHI, M. E. B.; OLIVEIRA, M. C. S.; CARDOSO, F. F.; ALENCAR, M. M.; REGITANO, L. C. A. Genes e vias metabólicas envolvidos nos mecanismos de resistência e susceptibilidade de bovinos infestados com carrapato Rhipicephalus microplus. In: CONGRESSO BRASILEIRO DE GENÉTICA, 56., 2010, Guarujá. **Resumos**... Ribeirão Preto: Sociedade Brasileira de Genética, 2010. p. 74.

KANEHISA, M; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research**, 2000, 28:27-30.

MOOTHA V. K.; LINDGREN, C. M.; ERIKSSON, K. F.; SUBRAMANIAN, A.; SIHA,G. S.; LEHAR, J., PUIGSERVER, P.; CARLSSON, E.; RIDDERSTRÅLE, M.; LAURILA, E.; HOUSTIS N.; DALY, M. J.; PATTERSON, N.; MESIROV, J. P.; GOLUB, T. R.; TAMAYO, P.; SPIEGELMAN, B.; LANDER, E. S.; HIRSCHHORN, J. N.; ALTSHULER, D. GROOP, L.C. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. **Nature genetics**, New York, v. 34, n.3 , p. 267-273, 2003.

R DEVELOPMENT CORE TEAM **R**: A Language and Environment for Statistical Computing. 2010. <http://www.R-project.org/.>. Acesso em: 15 out. 2011.

SONG, S.; BLACK, M. A. Microarray-based gene set analysis: a comparison of current methods. **BMC Bioinformatics**, London, v. 9, p. 502, 2008. doi:10.1.186/1471-2105-9-502.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERTET, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S.; MESIROV, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. **PNAS**, v. 102, n. 43, p. 15545-1550, Oct. 2005.

Comunicado Técnico, 109

Embrapa Informática Agropecuária
Endereço: Caixa Postal 6041 - Barão Geraldo
13083-886 - Campinas, SP
Fone: (19) 3211-5700
Fax: (19) 3211-5754
http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.com.br

1ª edição on-line - 2011

Expediente

Supervisão editorial: *Stanley Robson de Medeiros Oliveira, Neide Makiko Furukawa*

Normalização bibliográfica: *Maria Goretti Gurgel Praxedes*

Revisão de texto: *Adriana Farah Gonzalez*

Editoração eletrônica: *Neide Makiko Furukawa*

CGPE 9756