# Appendix: Cross Species Expression Analysis using a Dirichlet Process Mixture Model with Latent Matchings

**Hai-Son Le**
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
hple@cs.cmu.edu

**Ziv Bar-Joseph**
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
zivbj@cs.cmu.edu

## 1 Posterior distribution of $\mu_Y, \Lambda_Y$:

$$\kappa_{Yk} = \kappa_{Y0} + n_Y \qquad m_Y = \frac{1}{\kappa_Y}(\kappa_{Y0} m_{Y0} + n_Y \overline{y})$$

$$S_Y^{-1} = S_{Y0}^{-1} + V_Y + \frac{\kappa_{Y0} n_Y}{\kappa_{Y0} + n_Y}(\overline{y} - m_{Y0})(\overline{y} - m_{Y0})^T \qquad \nu_Y = \nu_{Y0} + n_Y$$

where $n_Y = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} \psi_{i,j,k}$, $\overline{y} = \frac{1}{n_Y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} \psi_{i,j,k} y_j$ and $V_Y = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} \psi_{i,j,k}(y_j - \overline{y})(y_j - \overline{y})^T$.

## 2 Data Normalization

As noted in the paper, the biological data was obtained from two previous publications. In both cases the authors have deposited normalized expression values in public databases (Stanford Microarray Database and Gene Expression Omnibus) and we have retrieved these datasets from these public databases. The data was indeed log2 transformed to obtain log ratios for changes vs. control (uninfected cells in both cases). These are standard procedures in microarray analysis and we would include this description in the updated paper. Parameters for BLASTN are based on the NCBI recommendations.

## 3 GO enrichment analysis for clusters inferred by DPMMLM

Table 1, 2, 3 and 4 present the GO enrichment analysis for cluster 2, 3, 4, 5 inferred by DPMMLM.

## 4 GO enrichment analysis for clusters inferred by DPMM

To test whether these differences inferred by the algorithm are biologically meaningful we compared our Dirichlet method to a method that uses deterministic assignments, as was done in the past (DPMM). Using such assignments the algorithm identified only two clusters. Neither of these clusters looked homogenous across species. Table 5 and 6 show the GO enrichment result for the two identified clusters.

| P value | Corrected P val | GO term description |
|---|---|---|
| 2.98216e-18 | <0.001 | regulation of transcription, DNA-dependent |
| 1.58764e-16 | <0.001 | ion binding |
| 1.28856e-16 | <0.001 | cation binding |
| 4.81459e-16 | <0.001 | zinc ion binding |
| 5.44663e-14 | <0.001 | regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process |
| 8.03482e-14 | <0.001 | DNA binding |
| 2.06599e-13 | <0.001 | regulation of cellular metabolic process |
| 3.71565e-13 | <0.001 | regulation of macromolecule biosynthetic process |
| 6.12201e-13 | <0.001 | transcription factor activity |
| 6.68339e-13 | <0.001 | regulation of cellular biosynthetic process |
| 1.06224e-12 | <0.001 | regulation of biosynthetic process |
| 1.45405e-10 | <0.001 | transcription regulator activity |
| 2.26834e-12 | <0.001 | regulation of metabolic process |
| 4.89572e-12 | <0.001 | intracellular |
| 7.10911e-12 | <0.001 | regulation of gene expression |
| 1.76818e-09 | <0.001 | transmembrane receptor protein tyrosine kinase activity |
| 2.17270e-09 | <0.001 | transmembrane receptor protein kinase activity |
| 8.56661e-09 | <0.001 | ephrin receptor activity |
| 1.19570e-08 | <0.001 | transmembrane receptor protein tyrosine kinase signaling pathway |
| 1.60773e-07 | <0.001 | cell differentiation |

Table 1: The GO enrichment result for cluster 2 identified by DPMMLM.

| P value | Corrected P val | GO term description |
|---|---|---|
| 2.50635e-12 | <0.001 | response to wounding |
| 2.20395e-10 | <0.001 | response to stress |
| 6.78145e-10 | <0.001 | inflammatory response |
| 7.28686e-10 | <0.001 | extracellular space |
| 3.81453e-09 | <0.001 | response to chemical stimulus |
| 6.60775e-09 | <0.001 | extracellular region part |
| 4.65677e-08 | <0.001 | response to stimulus |
| 5.76186e-08 | <0.001 | response to molecule of bacterial origin |
| 1.20234e-07 | <0.001 | fibrillar collagen |
| 1.37118e-07 | <0.001 | wound healing |
| 1.88215e-07 | <0.001 | platelet-derived growth factor binding |
| 2.26731e-07 | <0.001 | response to organic substance |
| 4.44301e-07 | <0.001 | response to lipopolysaccharide |
| 6.71932e-07 | <0.001 | cytokine activity |
| 7.39621e-07 | <0.001 | response to external stimulus |
| 1.03839e-06 | <0.001 | developmental process |
| 1.20286e-06 | 0.001 | skin morphogenesis |
| 1.66068e-06 | 0.001 | defense response |
| 1.93814e-06 | 0.002 | receptor binding |
| 2.15753e-06 | 0.003 | chemotaxis |

Table 3: The GO enrichment result for cluster 4 identified by DPMMLM.

| P value | Corrected P val | GO term description |
|---|---|---|
| 9.80701e-11 | <0.001 | protein binding |
| 1.71813e-10 | <0.001 | hydrogen ion transmembrane transporter activity |
| 8.76251e-10 | <0.001 | energy coupled proton transport, down electrochemical gradient |
| 8.76251e-10 | <0.001 | ATP synthesis coupled proton transport |
| 9.22263e-10 | <0.001 | cytoplasmic part |
| 1.28684e-09 | <0.001 | ribosome |
| 1.54800e-09 | <0.001 | monovalent inorganic cation transmembrane transporter activity |
| 2.13295e-09 | <0.001 | translational elongation |
| 5.21561e-09 | <0.001 | proton transport |
| 7.45508e-09 | <0.001 | hydrogen transport |
| 9.14840e-09 | <0.001 | ion transmembrane transport |
| 1.62098e-08 | <0.001 | structural constituent of ribosome |
| 2.19145e-08 | <0.001 | inorganic cation transmembrane transporter activity |
| 2.42190e-08 | <0.001 | translation |
| 3.20020e-08 | <0.001 | melanosome |
| 3.20020e-08 | <0.001 | pigment granule |
| 4.99963e-08 | <0.001 | proton-transporting two-sector ATPase complex, proton-transporting domain |
| 7.22247e-08 | <0.001 | intracellular part |
| 5.13700e-08 | <0.001 | cellular process |
| 2.66531e-07 | 0.001 | ATP biosynthetic process |

Table 2: The GO enrichment result for cluster 3 identified by DPMMLM.

| P value | Corrected P val | GO term description |
|---|---|---|
| 3.32796e-05 | 0.048 | regulation of receptor biosynthetic process |
| 2.84188e-06 | 0.007 | positive regulation of receptor biosynthetic process |

Table 4: The GO enrichment result for cluster 5 identified by DPMMLM.

## 5 Discussion

If we use a corrected p-value cutoff of 0.001 we see more than 2 fold increase in the number of significant GO categories that are identified by the DPMMLM method compared to the DPMM method (76 vs. 37 with 19 in the overlap). These include some of the most important categories for these experiments as discussed in the main text (for example, apoptosis).

| P value | Corrected P val | GO term description |
|---|---|---|
| 5.66587e-13 | <0.001 | protein binding |
| 1.76867e-10 | <0.001 | transcription factor activity |
| 2.23754e-09 | <0.001 | cellular process |
| 5.84677e-09 | <0.001 | biological process |
| 9.88355e-09 | <0.001 | regulation of metabolic process |
| 1.17922e-08 | <0.001 | regulation of cellular metabolic process |
| 3.15919e-08 | <0.001 | energy coupled proton transport, down electrochemical gradient |
| 3.15919e-08 | <0.001 | ATP synthesis coupled proton transport |
| 3.92189e-08 | <0.001 | regulation of cellular process |
| 7.46999e-08 | <0.001 | regulation of biological process |

Table 5: The GO enrichment result for cluster 1 identified by DPMM.

| P value | Corrected P val | GO term description |
|---|---|---|
| 7.03858e-16 | <0.001 | response to stress |
| 6.86006e-13 | <0.001 | inflammatory response |
| 1.64180e-10 | <0.001 | immune system process |
| 2.74539e-10 | <0.001 | immune response |
| 9.16024e-10 | <0.001 | response to molecule of bacterial origin |
| 4.69286e-09 | <0.001 | cytokine activity |
| 2.77866e-08 | <0.001 | receptor binding |
| 3.16405e-08 | <0.001 | transmembrane receptor protein tyrosine kinase activity |
| 3.46386e-08 | <0.001 | cytokine receptor binding |
| 4.01148e-08 | <0.001 | extracellular space |
| 4.87151e-09 | <0.001 | defense response |
| 5.40508e-09 | <0.001 | enzyme linked receptor protein signaling pathway |
| 7.83185e-08 | <0.001 | G-protein-coupled receptor binding |
| 8.13866e-09 | <0.001 | chemotaxis |
| 8.96811e-09 | <0.001 | cell communication |
| 1.38699e-07 | <0.001 | protein binding |
| 1.69750e-07 | <0.001 | binding |
| 2.24898e-07 | <0.001 | regulation of transcription, DNA-dependent |
| 3.01156e-07 | <0.001 | receptor signaling protein activity |
| 3.54444e-07 | <0.001 | regulation of immune system process |

Table 6: The GO enrichment result for cluster 2 identified by DPMM.