

The Bioinformatics Program and SGA contributed funding to cover travel expenses for our student Vida Abedi to attend The Eighth International Conference on Machine Learning and Applications (ICMLA'09) held in Miami, Florida, December 13-15, 2009. Ms. Abedi presented a poster on “**Understanding and eliminating systemic bias in knowledge discovery from biological literature using latent semantic analysis model.**” The research project was co-authored by her advisor, Dr. Mohammed Yeasin.

### **Abstract**

*Reverse-engineering of biological systems can be improved through prior knowledge such as mining of biological literature. However, the systemic bias (i.e., gene vs. number of documents per gene) remains a challenging issue in knowledge discovery through global models such as latent semantic analysis (LSA). The effect of bias in the data can be addressed using proper encoding matrix (model) and the choice of similarity measures used for gene selection. In particular, LSA model with and without the bias term (1<sup>st</sup> eigen vector) was used for gene selection using both linear and non-linear measures. A number of empirical analyses were performed to study the efficacy of the proposed modifications. Empirical analyses suggest that i) when the systematic bias in the data-set is significant, then a non-linear similarity measure outperforms linear similarity measure and, ii) exclusion of the bias term improves the robustness of the LSA model.*