# The EST Annotation Machine and the Keyword Clustering Machine: web-based resources for preliminary functional characterization of High-Troughput Gene Expression data

A. Guffanti1, J. F. Reid2, G. Simon1

1 IFOM ? Istituto FIRC di Oncologia Molecolare - Via Adamello,16 - 20139 Milano, Italia
2 LNCIB ? AREA Science Park ? Padriciano, 99 ? 34102 Trieste, Italia

The completion of the sequencing of the human genome and the emerging high-throughput tools such as cDNA microarrays and oligonucleotide arrays are leading the way to exponential biological data acquisition. The interpretation challenge is to extract relevant information from this large amount of data. A growing variety of statistical analysis approaches are available to identify clusters of genes that share common expression characteristics, but provide no information regarding the biological similarities of genes within clusters. We propose two integrated data mining and annotation tools for helping in the functional annotation of genes which show interesting expression relationships. These tools are not intended for providing a definite answer to the problem of associating a function to every gene under investigation, but rather as a framework or a guide for giving a shape to otherwise unstructured data.

The EST Annotation Machine (http://bio.ifom-firc.it/EST_MACHINE/index.html) is a tool for batch retrieval of the 'functional annotation' of thousands of EST or cDNA sequences which belong to UniGene clusters, such as immobilized cDNAs on arrays. Information about cluster ID, gene name, cluster title, chromosome band mapping, representative EST, LocusLink identifier, tissue expression pattern, chromosome mapping and protein similarities is retrieved directly from the UniGene/UniEST database. Annotation on protein similarities is retrieved dynamically from the sequence databases. Keywords are retrieved from Swissprot or Swissprot + PIR proteins similar to the given EST sequence.

Starting from a list of sequence identifiers followed by a list of tab-separated keywords (such the output of the EST Annotation Machine) the Keyword Clustering Machine (http://bio.ifom-firc.it/KW_CLUST/index.html) uses a hierarchical clustering which produces a dendrogram of sequence identifiers sharing common keywords. By comparing the keywords of sequence identifiers belonging to the same nodes in the tree, the researcher will readily identify the common keywords which could provide insight on the possible common function of these sequences. One could also compare these "keyword profiles" with gene expression profiles and see whether they will overlap up to a certain extent.