Genomic Functional Investigation through Statistical Analysis of Protein Families and Domains

Andrea Maffezzoli, Andrea Franceschini, Elisa Bellistri, Francesco Pinciroli, Marco Masseroli Ph.D.
BioMedical Informatics Laboratory, Bioengineering Department, Politecnico di Milano, Milan - Italy

ABSTRACT
Protein families and domains represent a very relevant resource useful to understand protein functions and interactions among their codifying genes. To perform evaluations of gene annotations sparsely available in numerous different databanks accessible via Internet, we previously developed GFINDer, a Web server that performs statistical analysis of functional and phenotypic annotations of gene lists. To exploit protein information present in Pfam and InterPro databanks, in GFINDer we integrated two new modules that allow annotating and statistically analyzing user-classified nucleotide sequence with controlled information on related protein families, domains and functional sites.

INTRODUCTION
GFINDer (http://www.bioinformatics.polimi.it/GFINDer/) Web server dynamically manages functional and phenotypic annotations and enables performing their statistical analysis and mining [1]. From protein information found in Pfam and InterPro databanks [2] we developed two new modules in GFINDer, the Exploration and Statistics Protein Families & Domains. These modules annotate numerous user-uploaded classified nucleotide sequence identifiers with controlled information on related protein families, domains, and functional sites, classifying them according to such protein annotation categories, and statistically analyzing the obtained classifications.

MATERIAL AND METHODS
GFINDer system consists of a multi-database, three-layer architecture. In the first layer, the “data layer”, a MySQL DBMS stores all genomic annotations in relational databases, including collections of protein families, domains and functional sites from Pfam and InterPro. To associate a protein characteristic with the codifying gene, we considered the protein accession numbers associated with a gene, as provided by Entrez Gene. In order to exploit the hierarchical “parent/child” relationships between InterPro entries describing common ancestry between entries, we calculated their hierarchical trees and structured them in a GFINDer database table. In GFINDer “processing layer”, we used Javascript and Active Server Page scripting technology to implement protein function categorical analysis. These analysis methods, based on controlled protein family, domain, and functional site categories, employs statistical hypergeometric and binomial distribution tests and Fisher’s exact test. Finally we implemented a Web interface as “user layer” which interact with the system and displays results of analysis performed.

RESULTS AND DISCUSSION
Based on Pfam 8,183 protein family domain entries and InterPro 12,542 entries (8,945 protein families, 3,289 protein domains and 308 functional sites), we structured 3,254 entries in 837 hierarchical trees of parent/child relations. Parent/child protein family trees show a maximum of 6 levels, with an average of 414 entries per level, and protein domain trees show a maximum of 5 levels, with an average of 149 entries per level.

GFINDer Exploration Protein Families & Domains module displays relationship between protein families, domains, or functional sites and selected gene, or how many of the selected genes refer to each protein family, domain, or functional site. The GFINDer Statistics Protein Families & Domains module allows estimating relevance of Pfam or InterPro controlled annotations for the user-selected genes by highlighting protein signatures significantly related to user-defined classes of genes.

Thus, new GFINDer modules complement previously provided phenotypic and functional evaluations in supporting better interpretation of gene list (e.g. from microarray experiment results) and help unveiling new biological and biomedical knowledge about the considered genes.

REFERENCES