Genomic annotation and statistical analysis of protein families and domains for functional investigation of gene lists

Masseroli M, Franceschini A, Maffezzoli A, Pinciroli F

Laboratorio di Informatica BioMedica, Dipartimento di Bioingegneria, Politecnico di Milano
piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

Protein families and domains constitute one of the most useful information to understand protein functions and to gain insight into interactions among their codifying genes. Comprehension of domain structure of proteins within completed genomes is also fundamental for better understanding the evolutionary forces and emerging functions shaping genomes. The increasing number of proteins for which domain-based annotation is available hence represents an important background for computational genome-wise analyses. To allow performing comprehensive evaluations of gene annotations sparsely available in numerous different databanks accessible via Internet, we previously developed GFINDer, a Web server that dynamically aggregates functional and phenotypic annotations of user uploaded gene lists and allows performing their statistical analysis and mining (http://www.bioinformatics.polimi.it/GFINDer/). Exploiting protein information present in Pfam and InterPro databanks, we developed and added in GFINDer new original modules specifically devoted to exploration and analysis of functional signatures of gene protein products. They allow annotating numerous user 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.

Methods

GFINDer Web system is implemented in a three-tier architecture based on a multi-database structure. In the first tier, the data tier, a MySQL DBMS manages all considered genomic annotations stored in different relational databases. In two of them, we structured protein information from Pfam and InterPro comprehensive and manually curated collections of protein families, domains and functional sites. To associate a protein characteristic with the codifying gene, we considered the protein accession numbers associated with a gene, as provided by Entrez Gene database. In order to efficiently exploit the hierarchical "parent/child" relationships that can exist between InterPro entries, relationships that describe a common ancestry between entries, we first precomputed their hierarchical trees and structured them in a GFINDer database table. Then, within GFINDer processing tier, in Javascript and Active Server Page scripts, we implemented protein functional signature categorical analyses based on controlled protein family, domain, and functional site categories. Created analysis procedures employ hypergeometric and binomial distribution tests and the Fisher's exact test to assess statistical significance of the over and under representation of categorical protein annotations in a group of user classified genes.

To interact with the MySQL DBMS server on the data tier, we used Microsoft ActiveX Data Object technology and Standard Query Language, whereas we employed Hyper Text Markup Language and Javascript to implement a Web graphic interface for the user tier, which is composed of any client computer connected to the Web server on the processing tier through an Internet/intranet communication network.

Results

In Pfam databank version 19.0 we found 8,183 protein family domain entries, and in InterPro databank release 12.0 we found 12,542 entries (8,945 protein families, 3,289 protein domains and 308 functional sites including post translational modifications, repeats, active and binding sites). Out of these entries, 3,254 (2,486 protein families, 743 protein domains and 25 functional sites) were grouped in 837 hierarchical trees of parent/child relations (574 of protein families, 252 of protein domains and 11 of functional sites). Parent/child protein family trees had a maximum of 6 levels, with an average of 414 entries per level, whereas protein domain trees had a maximum of 5

levels, with an average of 149 entries per level.
GFINDer modules developed for the exploitation of such structured data provide Protein Family & Domain Annotation, Exploration, and Statistics analyses. The Exploration Protein Families & Domains module allows to easily and graphically understand either how many and which protein families, domains, and functional sites are associated with each considered gene, or how many of the selected genes refer to each protein family, domain, or functional site. When uploaded nucleotide sequence identifiers are subdivided in classes (e.g. from clustering analysis of microarray results), the Statistics Protein Families & Domains module allows estimating relevance of Pfam or InterPro controlled annotations for the uploaded genes by highlighting protein signatures significantly more represented within user-defined classes of genes.
Thus, new GFINDer modules allow performing genomic protein function analyses that well complement previously provided phenotypic and functional evaluations in supporting better interpretation of microarray experiment results and unveil new biological knowledge about the considered genes.

Availability: http://www.bioinformatics.polimi.it/GFINDer/

Contact email: masseroli@biomed.polimi.it