ProteInOn: A Web Tool for Protein Semantic Similarity

Daniel Faria, Catia Pesquita, Francisco M. Couto and André O. Falcão


Motivation

The Gene Ontology (GO) provides a structure for comparing genes/proteins on the functional level. Several measures have been used for this type of comparison (semantic similarity). The tools available for the community are few and limited in application.

Abstract

We have developed ProteInOn, an integrated web tool for computing GO-based protein semantic similarity. It implements eight distinct similarity measures, none of which were previously available online, including the novel simGIC measure developed in the context of this tool.

In addition to computing protein and GO term semantic similarity, ProteInOn combines protein interaction and annotation data, allowing the user to find proteins that interact with, or GO terms represented in a set of input proteins. The results of these queries can then be used as input for semantic similarity calculations (or other queries) providing a structure to answer more complex questions.

Interface

Simple and flexible 3-step query selection menu:

- **Input**: 2-10 input entries, proteins or GO terms
- **Query**: 3 queries for proteins, 2 queries for GO terms
- **Options**: several options for each query

Fast and easy to read results:

- semantic similarity results in all vs. all matrix
- color-coded results for easy interpretation
- term results ranked by representativity score
- results selectable for new queries
- protein and GO term ids link to source DBs

Applications

Genomics/Proteomics:

- Find out which biological process terms are better represented and more meaningful in a set of upregulated genes/proteins.
- Measure the average biological process similarity of a set of co-expressed genes.

Comparative Genomics:

- Measure the molecular function similarity between sets of homologous genes from different species.

Interactomics:

- Measure the biological process similarity between a protein and others it interacts with.
- Measure the molecular function similarity between sets of proteins that share interactors.

Semantic Similarity

ProteInOn implements eight similarity semantic measures:

- three term similarity measures (Resnik’s, Lin’s and Jiang & Conrath’s) with two distinct approaches (MICA and GraSM) plus two graph similarity measures (simUI and simGIC).
- The three term similarity measures are information content (IC) based and rely on the notion of lowest common ancestor. As they are measures for single terms, applying them to proteins requires combining the similarities between the proteins’ terms, which in ProteInOn is done with a best-match average.
- By contrast, the graph similarity measures can be applied directly to both terms and proteins. They consider similarity between two terms (or term sets) as the ratio between the intersection and the union of the graphs they define, differing only in that simUI is edge based and simGIC is IC based.
- All measures were evaluated by comparing protein semantic similarity with sequence similarity, and were found to correlate equally well. They differ only in resolution, which is the ordering criterion in ProteInOn’s options selection menu.

GO term similarity:

Resnik_{resnik}(D,E) = MAX(|IC(B), IC(C)|)

Resnik_{lin}(D,E) = AVG(|IC(D), IC(E)|)

Lin(D,E) = \( \frac{2 \times Resnik(D,E)}{IC(D) + IC(E)} \)

JiangConrath(D,E) = 1 - Resnik(D,E) - \( IC(D) \times IC(E) \)

Protein similarity:

simUI(Prot1, Prot2) = \( \frac{\text{COUNT}_{\text{Prot1,Prot2}}(F)}{8} \)

simGIC(Prot1, Prot2) = \( \frac{\text{SUM}_{\text{Prot1,Prot2}}(|IC(T)|)}{\text{SUM}_{\text{Prot1,Prot2}}(|IC(T)|)} \)

Resnik(Prot1, Prot2) = AVG(Resnik(E, F), Resnik(G, H))