Protein Function Prediction in UniProt with the Comparison of Structural Domain Arrangements

**INTRODUCTION**

**METHODS**

- **Data Preparation**
  - DA generation
  - Classification

- **Generation of DAs**
  - Using InterProScan results on UniProtKB proteins

- **Grouping**
  - proteins under shared DAs & separation of learning and test sets

- **Mining**
  - functional annotations from source databases (for learning set)

- **Classification**
  - each annotation term has its own class → use of class-specific parameters

**RESULTS**

- DAAC: a new approach in the field of automatic functional annotation of UniProtKB/TrEMBL proteins especially to capture the complex relationship between the distant multi-domain proteins
- Novel in: (i) DA alignment as the basis of similarity, (ii) supervised multi-label classification where each class represents a unique functional term, (iii) InterPro as the domain annotation resource

**CONCLUSIONS**

- This approach is proposed as complementary to the conventional sequence based function prediction methods
- The system is planned to be implemented to work as a part of the UniProt Automatic Annotation Pipeline to increase the coverage and the quality of the functional predictions
- We plan to establish the method to predict other types of annotations as well: recommended protein names, sub-cellular locations, keywords, comments and features.

**Schematic representation of DAAC:**

- DA generation
- Classification
- Mining
- **Definition of domain architecture (DA):**
  - concatination of the InterPro DoDs of the domains on the protein sequence

**RESULTS**

- DAAC is trained for GO term and EC number prediction
- The method is cross-validated on the UniProtKB/Swiss-Prot (ground-truth data) and applied on non-reviewed proteins in UniProtKB/TrEMBL
- The # of unique DAs is nearly 1/10 of the # of proteins: (for UniProtKB/SwissProt # of entries: 546,238; # of DAs: 58,834)

**CONCLUSIONS**

- **Data Preparation**
  - DA generation
  - Classification

- **Generation of DAs**
  - Using InterProScan results on UniProtKB proteins

- **Grouping**
  - proteins under shared DAs & separation of learning and test sets

- **Mining**
  - functional annotations from source databases (for learning set)

- **Classification**
  - each annotation term has its own class → use of class-specific parameters

**CONCLUSIONS**

- This approach is proposed as complementary to the conventional sequence based function prediction methods
- The system is planned to be implemented to work as a part of the UniProt Automatic Annotation Pipeline to increase the coverage and the quality of the functional predictions
- We plan to establish the method to predict other types of annotations as well: recommended protein names, sub-cellular locations, keywords, comments and features.