Protcomp-PL

Program for Identification of sub-cellular localization of Eukaryotic proteins: Plants

Protcomp combines several methods of protein localization prediction - neural networks-based prediction; direct comparison with updated base of homologous proteins of known localization; comparisons of pentamer distributions calculated for query and DB sequences; prediction of certain functional peptide sequences, such as signal peptides, signal-anchors, GPI-anchors, transit peptides of mitochondria and chloroplasts and transmembrane segments; and search for certain localization-specific motifs. It means that the program treats correctly complete sequences only, containing signal sequences, anchors, and other functional peptides, if any. The program includes separately trained recognizers for animal/fungal and plant proteins, which dramatically improves recognition accuracy. The following table provides approximate prediction accuracy for each compartment of animal/fungal proteins. Testing was performed on a samples of proteins of known localization (~200 in each localization), which were NOT included in training samples for the programs.

Compartment	Percent predicted correctly		
	ver. 4	ver. 5	ver. 6
Nucleus	80	88	91
Plasma Membrane	80	87	100
Extracellular	69	83	86
Cytoplasm	46	63	88
Mitochondria	76	82	89
Endoplasmic Reticulum	67	83	89
Peroxisome	95	97	91
Lysosome	69	91	100
Golgi	57	77	91

Output sample for complete version:

```
Seq name: Q7M1E7 Location:Extracellular (Secreted)  DE  Polygalacturonase
precursor (PG) 514
Significant similarity in Location DB -  Location:Extracellular (Secreted)
Database  sequence:  AC=P35336  Location:Extracellular  (Secreted)    DE
Polygalacturonase precursor (EC 3.
Score=7765, Sequence length=467, Alignment length=335
Predicted by Neural Nets - Extracellular (Secreted) with score    2.7
******** Signal 1-49 is found
Integral Prediction of protein location: Extracellular (Secreted) with score
4.4
Location weights:     LocDB / PotLocDB / Neural Nets / Pentamers / Integral
 Nuclear                0.0 /      0.0 /        0.70 /      0.08 /     0.77
 Plasma membrane        0.0 /      0.0 /        1.06 /      4.36 /     5.42
 Extracellular       7765.0 /      0.0 /        2.68 /      0.00 /     4.41
 Cytoplasmic            0.0 /      0.0 /        0.72 /      0.00 /     0.72
 Mitochondrial          0.0 /      0.0 /        0.70 /      0.00 /     0.70
 Chloroplast            0.0 /      0.0 /        0.65 /      0.00 /     0.65
 Endoplasm. retic.      0.0 /      0.0 /        1.58 /      0.00 /     1.58
 Peroxisomal            0.0 /      0.0 /        0.48 /      0.00 /     0.48
```

LocDB are scores based on query protein's homologies with proteins of known localization.

PotLocDB are scores based on homologies with proteins which locations are not experimentally known but are assumed based on strong theoretical evidence.

Neural Nets are scores have been assigned by neural networks.

Pentamers are scores based on comparisons of pentamer distributions calculated for QUERY and DB sequences.

Integral are final scores as combinations of previous four scores.

In this reduced version time and disk space consuming processes of DB search and comparisons of pentamers' distributions are abandoned. Columns "LocDB" and "PotLocDB" (results of DB search) and/or "Pentamers" (results of comparisons of pentamers' distributions) are excluded from output tables. However, one should remember, that such abandonment decreases recognition accuracy.

While interpreting output results, it must be kept in mind that:

1. Protcomp's scores *per se*, being weights of complex neural networks, do not represent probabilities of protein's location in a particular compartment.
2. Significant homology with protein of known location is a very strong indicator of query protein's location.
3. For neural networks scores, their relative values for different compartments are more important than absolute values, i.e. if the second best score is much lower than the best one, prediction is more reliable, regardless of absolute values.
4. If both neural networks and homology predictions point to the same compartment, this is very reliable prediction.

In this version comparison with base of homologous proteins of known localization as well as comparisons of pentamer distributions calculated for query and DB sequences are absent.