Protcomp-B

Program for Identification of sub-cellular localization of bacterial proteins.

Protcomp-B combines several methods of protein localization prediction - Linear Discriminant Function-based prediction; direct comparison with bases 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 and transmembrane segments. It means that the program treats correctly complete sequences only, containing signal sequences, anchors, and other functional peptides, if any.

For Gramm-positive bacteria proteins three locations are discriminated: Cytoplasmic, Membrane and Extracellular (Secreted).

For Gramm-negative bacteria proteins five locations are discriminated: Cytoplasmic, Membrane (Outer and Inner), Periplasmic and Extracellular (Secreted).

If bacteria type is not defined locations for Gramm-negative bacteria are discriminated.

Output sample for complete version:

```
ProtComp Version 3. Identifying sub-cellular location Bacterial (Gramm
negative)

Seq name: Test sequence 330
Significant similarity in Location DB -  Location:Membrane
Database sequence: AC=P55569 Location:Membrane  DE  PROBABLE ABC TRANSPORTER
PERMEASE PROTEIN Y4MJ.
Score=16110, Sequence length=333, Alignment length=330
Predicted by LDA staff - Inner Membrane with score    1.4
******** Signal 1-25 is found
******** Transmembrane segments are found: .+59:157-..-174:199+..+225:327+.
Integral Prediction of protein location: Inner Membrane with score    7.0
Location weights:     LocDB / PotLocDB /     LDA     / Pentamers / Integral
 Cytoplasmic           0.00 /     0.00 /        0.02 /      0.00 /     0.02
 Membrane          16110.00 /  4010.00 /        1.42 /      1.51 /     6.95
 Periplasmic           0.00 /     0.00 /       -0.65 /      0.00 /    -0.65
 Secreted              0.00 /     0.00 /        0.08 /      0.03 /     0.10
```

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.

LDA are scores have been assigned by Linear discriminant functions.

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

Integral are final scores as combinations of previous 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 functions, 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 LDA 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 LDA and other 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.