PredictProtein in the cloud, the problems with

PredictProtein

In silico

experiment

transmembrane helices
transmembrane boundaries
protein domain boundaries
protein-DNA interaction sites
protein-protein interaction sites
protein globularity
homologous sequences
evolutionary profile
protein families
effects of amino acid substitutions
unstructured loops
subcellular regions
localization
signaling
nuclear localization
disulfide bridges
conformational switches
coiled coils
secondary structure
non-regular secondary structure
accessibility
β barrels
membrane proteins

Unannotated protein sequences

Annotations from prediction methods
From thesis to cloud

Your predictor
Its stable version
Versioning system

Machine image

deb (Debian/Ubuntu)
rpm (Red Hat)
tar.gz

Operating system

Databases

Other
Rost lab
methods

Server instance
Server instance
Server instance

Academic cloud
Too much data to move around

### Databases
- 15GB

### Predictions
- 4.1M predictions on 32TB; one prediction 9.5M ± 12.5M on avg

---

**Databases:**
- big 6.5GB
- big_80 2.8GB
- pfam 2.7GB
- swissprot 2.6GB

**Predictions:**
- features 5%
- blast r dbs 3%
- hssp 7%
- hssp0filter80 14%
- hssp_80 15%
- big blast 18%
- big_80 blast 19%
- saf 6%
- saf_80 13%
Time required by blast result conversion steps