PROXEON

From proteomics data to biological sense in minutes

ProteinCenter™

Gaining Biological Insight

Analysis of multiple protein data sets is considered one of the biggest bottlenecks in many labs. ProteinCenter™ greatly accelerates this challenging task, but also makes you gain truly new knowledge from your experimental data.

All-in-one Protein Database

ProteinCenter is a web application with an integrated protein database containing sequence information and annotation from GenBank, RefSeq, EMBL, UniProt, Swiss-Prot, Trembl, PIR, IPI, Ensembl and many more.

data sources:

Prokaryote and eukaryote organisms

- more than 7 **million** proteins
- more than **30 million** accessions keys

This means that you no longer have to struggle with different versions of the databases. ProteinCenter handles all the tedious bookkeeping for you and makes it easy to compare datasets with accessions from different protein databases.

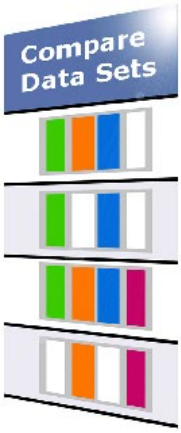
All sequence and annotation data are consolidated, and categorized to optimize the analysis of your experimental data.

Comparing Data Sets

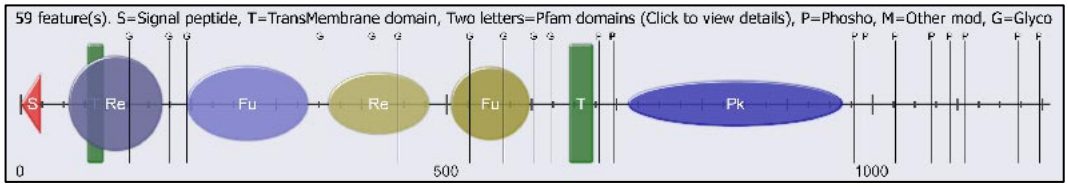
With ProteinCenter you can compare multiple data sets of thousands of proteins in minutes and get the true overlap, independently of the original database source.

Compare your own data sets with results from other scientific studies and see your data in a new context.

Data sets can be mined, compared and documented in a matter of minutes using a range of truly novel methods for handling large proteomics data sets.



Efficient Functional Analysis *Whether on single or thousands of proteins*



With ProteinCenter you will get comprehensive, biologically relevant annotations to ensure that you are not missing out on important details.

Computational Enrichment™ provides a **wealth of extra information** for each sequence based on gold standard bioinformatics tools.

ProteinCenter

Confident interpretation of proteomics data

1. Protein sequence data integration

- Integration of most public protein databases into a common data format
- Includes also the handling of outdated sequence records
- Protein similarity is calculated and neighbour information is stored
- Computational enrichment of sequences improves biological annotation by an order of magnitude over public data
- Automatic and frequent sequence updates
- Easy integration of proprietary data

2. Functional analysis of single proteins and large data sets

- Clustering of similar sequences (e.g. collapsing fragments, splice variants and alleles with full length versions to provide a better overview)
- Clustering of proteins that contain the same peptides
- Computationally enriched sequence annotation provides much more information than what is publicly available
- Fast but comprehensive overview of annotation when analyzing very large data sets. Includes e.g. GO annotation, localization, membrane regions, signal peptides, etc.
- Filtering on biologically relevant parameters for complex queries of data
- Instant BLAST (ultra fast BLAST where all matching sequences are shown with comprehensive biological annotation)
- ProteinCards (in-depth information resource for each individual sequence. Includes also information derived from similar proteins plus links to external resources and much more)
- Sequence analyses can shift from “overview” to very detailed “drill-down” interrogation on individual sequences in two-mouse-clicks (split seconds)
- Easy integration with existing tools and workflows

3. Comparison of lists of sequences

- Comparison of multiple protein datasets, using Boolean logic (the contents of lists of experimentally derived proteins can be compared regardless of which data accession number system the data originated from)
- Comparison of complete protein datasets, filtered data, clustered data
- Advanced comparisons can be based on multiple experimental observations and biological functions simultaneously

PROXEON

4. Statistics

- Easy statistical overview of what your data set contains (e.g. distribution of GO terms)

5. Report generation

- Fast reporting of key biological parameters on single or multiple proteins
- Statistical reports on large data sets
- Visually effective overviews of large data sets

6. Workspace

- Very advanced (yet simple) directory function enables sorting, selection, and management of experimental data and projects
- Data sets can be commented and stored along with analysis conclusions anytime
- Simple integration with external data generating resources and LIMS systems

7. Sequence look-up functionality

- Look-up of single or few proteins based on e.g. accession codes (regardless of format) or peptide sequences
- Makes ProteinCenter an excellent generic look-up tool for all protein related work

8. Sequence “shopping basket”

- Like a notepad this allows one-click capture of data of special interest on the fly. This feature is just way smarter than a notepad

9. Data Sharing

- Fast import and export of data
- Web-based user interface and multi-user access allow direct sharing of data
- Easy exchange of data between ProteinCenter installations

10. Simplified administration

- Web based — automatically updates sequence and annotation data
- One software application replaces multiple other applications

11. Ease of use

- Very advanced bioinformatics analysis is made available to non-experts
- Workflow oriented and highly intuitive user interface
- On-line help explains details
- Custom workflow wizard can be used to accelerate particular repetitive tasks

ProteinCenter

Examples of questions you ask or things you do, all the time:

Analysing the biological function of a single protein

- What is known about this protein?
- Does this protein contain a signal peptide, a transmembrane region, or a particular domain?
- Is this truly an unknown protein, and if so, what is known about homologous proteins?
- What is the full length sequence?
- Are there alternative splice isoforms?
- Which sequence databases include this protein?
- Does any other protein contain this peptide sequence?
- Which proteins are homologous – and which function do they have?
- Does this outdated protein really exist?

Getting the biology of a large data set:

- How many different proteins are in the set, if redundancy is removed?
- How many of my different proteins are merely due to allelic differences?
- How many different protein families are there?
- How many proteins have PTMs?
- What fraction of the data set displays a certain experimental measurement?
- How is the distribution of subcellular localization, pathways, function?
- (Combinations of the above)

Managing the analysis of large datasets

- Group related proteins in order to analyse them together
- Cluster families of uninteresting proteins – e.g. keratins
- Filter in the interesting subsets – e.g. those with a particular function or pathway or experimental measurement
- Hide non-interesting subsets

Evaluating experimental procedures

- Which purification protocol gave more membrane proteins?
- What is the distribution of long and short proteins?

Comparing two or more sets of experimental data

- What is the total set of proteins identified by combining these twenty-two data sets?
- Which proteins are different or in common between the data sets?
- Looking at these different fractions (samples, database searches etc.):
  - What are then the differences in subcellular localization, expression, pathway, diseases, molecular functions etc.?

Comparing experimental data to literature data sets

- Which proteins are different and in common?
- Are the experimental observations already reported?
- What is the difference in observed positions of e.g. PTMs?

Comparing experimental data to known annotation

- Which of my phosphor-proteins were previously known to be phosphorylated?
- Which are already known to be mitochondrial?

ProteinCenter will accelerate this kind of work by orders of magnitude

Complex analysis examples →

ProteinCenter

Examples of complex analyses that would take more than a month without ProteinCenter (if doable at all) and which can now be done in less than 5 minutes:

- Let me take the ten datasets of proteins identified from the stimulated cell line and quickly check that most proteins on each dataset are present in all datasets……and then let me take the combined set.
    - Wow! The combined set is more than 10,000 different IDs……but they actually collapse to less than 400 protein families when ProteinCenter has folded alleles, fragments, and accession number redundancy into meaningful groups.
- OK; then I need the ten datasets of proteins identified from the non-stimulated cells……also less than 400 protein families when collapsed. Excellent.
- Now, which proteins do the two combined sets have in common?
- And, which ones are different?
- OK,here are the twenty-two proteins that were up-regulated.
    - 13 of them were metabolic
    - 7 of them were involved in gene regulation
    - Let me have a closer look at these two. Aha, one of them is 95% identical to a known structural protein from rat. Boring.
    - This one doesn't look familiar? But I can see we already had it through HTS two years ago. I better give those guys a ring.
- Not bad for four minutes of data analysis…

- My nuclear protein preps have gone from 27% nuclear to 34% last time. Let me see what this prep says, once my data are in ProteinCenter?
- So. First let me cluster the two thousand accession numbers into some meaningful groups, and then…
- Well, 56% nuclear. At least that's progress.
- And the remainder are all cytosolic and ER.
- And eight of the sequences are really not annotated at all. Not in any species. Oh; they all have 95% sequence similarity with some household proteins. Boring.
- Oops. That's odd. The size distribution indicates that I'm losing all the low mass proteins somewhere
- With such a skewed data set I'm glad the analysis only took 3 minutes.
- Well, back to the lab.

- Let me cluster my data set of 3430 mitochondrial proteins to remove redundancy and ignore allelic differences for now
- Aha – around 600 proteins – well, I expected something in that ball park
- Then let me compare to this publication on 95 mitochondrial proteins involved in diseases…….
- And also compare it to this curated database of 586 mitochondrial proteins…….
- Wow, there is quite an overlap! Cool! And it is no problem that my dataset consisted of UniProt accessions, while the publication was NCBI GI's and the database IPI entries.
- Ok, here is the set of proteins that are unique to my data set.
- Some of them appear to be from a lysosome contamination. Let me subtract those from the unique set.
- Aha, the report of biological processes and subcellular component for my unique data set immediately tells me that here is probably a lot of unknown mitochondrial proteins…….but here are some that are around 80% homologous to these mitochondrial proteins. Let me park them for now.
- Here is one that must be a splice variant of this known gene, but potentially belongs to mitochondria. Let me click here to check my peptides and experimental measurement. Looks promising, let me go in depth with that one.

- So, here are my 546 proteins from the 80 pull-down experiments related to cell cycle
- Very convenient that they are ordered in 80 sets in the cluster view according to which bait was used!
- From the GO data, it immediately looks like these 8 sets must be dominated by promiscuous binders since all sorts of proteins are found.
- But, isolating proteins from these 8 sets and applying filters related to cell cycle, it looks like I also have the proteins I expected. OK!
- And, let me invert the filter to see the protein subsets that are probably incorrect. Aha, most of them are related to stress response.
- Back to the lab to redo those 8 experiments