

Statistics Regarding Relevant Information Obtained from the Swiss Prot Referenced Data

From a list of UniProt accession numbers, we retrieved a collection of protein data from the Swiss Prot Database. It was necessary to find a simple solution to quickly analyze the protein data; it was impossible to succinctly analyze the information by hand due to quantity so a computational approach was instituted.

The goal of the computational approach was to calculate the distribution of ‘topics’ which appeared in the ‘comment blocks’ of protein information retrieved from Swiss Prot. Every protein in Swiss Prot can potentially have comment blocks of information and each comment block has topic associated with it in a header¹. Topic is the standard UniProt terminology and denotes what kind of protein information is being presented. A topic can have values such as “Function”, “Biophysicalchemical Properties” or “Pathway” among others². Each topic can also be associated with a status flag. The status flags serves to warn of information that is not obtained through experimental results and has in some way been inferred by the curator. The value that the status flag can take on may void the information in a topic depending on the needs of the researcher. To address this, the goal of the computation approach was altered to include the calculation of the status flag distribution among the topics. Prior to continuing, please refer to Figure 1 below for a clear example of topics, status flags and comment blocks.

Comments

- **FUNCTION** May play a role in lipid exchange and transport throughout the body. May participate in reverse cholesterol transport from peripheral cells to the liver (By Similarity);
- **SUBCELLULAR LOCATION** Secreted (Potential);
- **ALTERNATIVE PRODUCTS** 3 named isoforms [FASTA] produced by alternative splicing.

Topic

Status Flag

Figure 1 Identifying topics and status flags in Swiss Prot comments. An example from the Online version of Swiss Prot from a protein with the accession number: Q9BPW4. Function, Subcellular Location and Alternative Products are the topics followed by Potential and By Similarity in the lighter gray as the status flags.

The status flags can have three possible values: Potential, By Similarity and Probable. The values that the status flags can take are called non-experimental qualifiers, because as already mentioned, the curators have inferred information in the topic that is not proved by an experiment. Where a status flag is absent, the information in that topic should be conclusive by experimental results. Depending on the degree of certainty required the presence of a status flag could suggest the information is unreliable to an experimenter looking for information on these proteins. The UniProt User Manual provides definitions for each of the non-experimental qualifiers. Rather than paraphrase the information and risk losing exact meaning of the non-experimental qualifiers, the definitions below (in a different font) are taken from the Uniprot User Manual.

- The term '**Potential**' indicates that there is some logical or conclusive evidence that the given annotation could apply. This non-experimental qualifier is often used to present the results from protein sequence analysis tools, which are only annotated, if the result makes sense in the context of a given protein. A typical example is the annotation of N-glycosylation sites in the entries of non-cytoplasmic domains or proteins.
- The term '**Probable**' is stronger than the qualifier 'Potential' and there must be at least some experimental evidence, which indicates, that the given information is expected to be found in the natural environment of a protein.
- '**By similarity**' is added to facts that were proven for a protein or part of it, and which is then transferred to other protein family members within

a certain taxonomic range, dependent on the biological event or characteristic. Non-experimental qualifiers are also assigned to biologically important sites found within conserved domains e.g. active sites within an enzymatic domain or disulfide bonds that stabilize the structure of extracellular modules.

Table 1 below show the results of the calculating the distribution of both topics and status flags in the collection for protein data.

Table 1. Summary of topic and status flag distribution

Number refers to the total number of times this topic or non-experimental qualifier occurred in the protein data collection .

Topics		Status Flags	
Topic	Number	Non-experimental Qualifier	Number
Allergen	123	-	-
Alternative Products	111	-	-
Biophysicalchemical properties	4	-	-
Catalytic Activity	46	-	-
Caution	17	-	-
Cofactor	19	By Similarity	11
Developmental Stage	10	-	-
Disease	106	-	-
Domain	36	By Similarity	6
Enzyme Regulation	7	By Similarity	2
		Potential	4
Function	208	By Similarity	42
		Probable	1
Induction	9	-	-
Interaction	199	-	-
Mass Spectrometry	2	-	-
Miscellaneous	18	By Similarity	1
Pathway	14	-	-
Pharmaceutical	1	-	-
Polymorphism	6	-	-
PTM	57	By Similarity	11
		Probable	1
Online Information	63	-	-
Sequence Caution	37	-	-
Similarity	400	-	-
		Potential	39
Subcellular Location	196	By Similarity	33
		Probable	22
		Potential	2
Subunit	123	By Similarity	20
Tissue Specificity	148	-	-

Looking at only the status flags numbers in Table 1, it is possible to see that not ALL topics in the set of protein data have status flags associated with them. For those topics that do have status flags, the status flags are not present in every instance of the topic. For example, the Function topic occurs 208 times and of those instances, 47 have status flags associated with them, meaning that the remaining 161 occurrences of the function topic contain information that is experimentally valid. Additionally, 42 of the 47 instances have the 'probable' non-experimental qualifier implying that those 42 can in some way be experimentally validated based on the definition of the probable non-experimental qualifier. If protein function is what the researcher is most interested in this may be a good result.

From the varying number of topic distribution, it is correctly inferred that not all proteins carry information about all topics and that some proteins carry information about a topic more than once. The next natural task was to see how these topics are further distributed among the individual proteins. The results are shown below.

On Average, there are 6 topics per protein.

The median number of topics is 5

The mode, that is most frequently occurring, number topics per protein is 5.

Using Quartile Analysis:

The most number of topics in a protein (with repetition of topics) is 22.

The least number of topics is 0 (but there are pub med references). There are five cases that have no mention of topics.

25% percentile : 3 topics

50% percentile: 5 topics

75% percentile : 8 topics

In this particular experiment, we were most interested in knowing the distribution of Function and Pathway topics. Based on information gather from the results, we were sufficiently happy with the amount of experimentally valid information that could be obtained. However in a general sense, this computational method can be used to discover what other topics are in a protein data collection that could serve a useful a purpose. Also being able to quickly peruse the usability of topics of interest could aid in text mining or meditations on further experimental or computational steps.

-
- ¹ Please note, the from a technical stand point header may not be the correct term. For example, in the XML version of Swiss Prot, the topic is listed in the comment tag as a type, while in the flat file version of Swiss Prot, it is denoted be the - !- symbol. While this information is not relevant to someone who is not a programmer, it is worthwhile to mention.
- ² It is possible to find all the topics and their definitions online in the UniProt User manual.