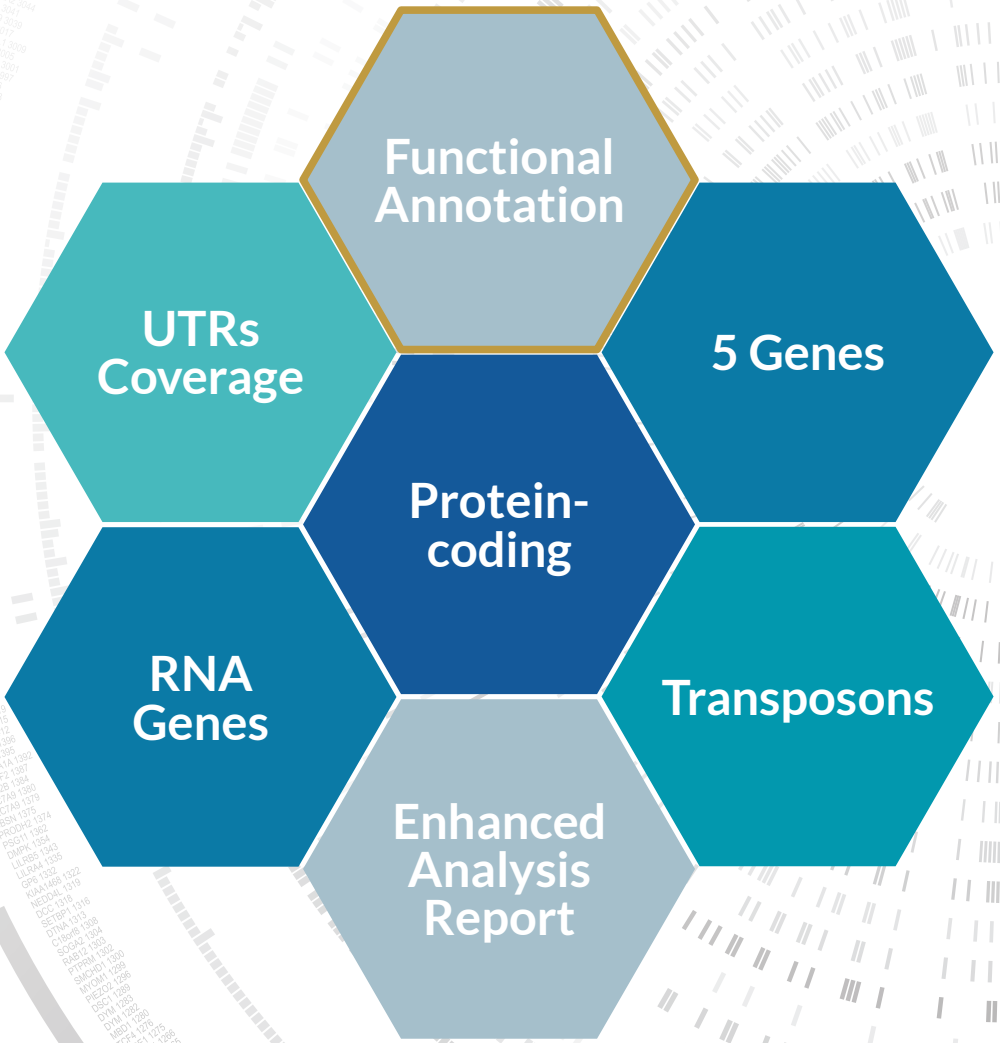
GOENOMICS

Functional annotation

assigning names, domains, EC-numbers and GO-terms to protein-coding genes, pseudogenes and transposons



Functional annotation

assigning names, domains, EC-numbers and GO-terms to protein-coding genes, pseudogenes and transposons

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

Functional annotation is the process of assigning biological meaning to identified gene sequences. It involves predicting the roles of genes, transcripts, and proteins by comparing them to known databases, identifying protein-coding regions, functional domains, motifs, and gene ontology (GO) terms. Functional annotation also includes the identification of regulatory elements, pathways, and interactions, providing insights into the biological processes, molecular functions, and cellular components associated with each gene.

Naming protein-coding genes

Protein and gene names are usually assigned based on homology to named proteins/genes. Homology is determined by comparing the predicted genes with data from comprehensive databases such as GenBank, SwissProt and UniProt or species databases such as FlyBase, TAIR, Xenbase, ZFIN and others using tools such as BLAST. However,

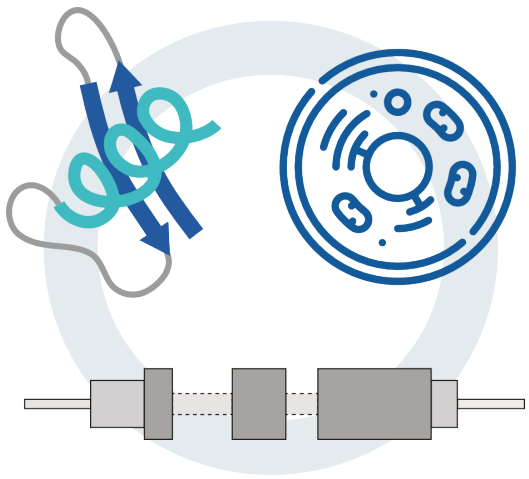


Figure 1: The functional annotation assigns information about the protein structure (e.g. protein domains, sequence motifs), the biochemical function (e.g. belonging to an enzyme class, EC number) and the molecular and cellular functions (e.g. protein homology, GO terms) to the genes.

- The naming of proteins and genes is inconsistent in the literature (orthologous genes may be named differently in human and mouse or Arabidopsis and rice, for example),
- the names were assigned in the last two decades based on the respective databases available at the time of annotation (earlier annotations might reflect old and resolved names, while later annotations might contain newly assigned names)
- the names were assigned on the basis of different databases (the best hit in the search for TAIR may be very different from the best hit in the search for SwissProt),
- the names are not assigned according to the domains defining the protein family, but according to the best (longest) hit (long coiled-coil-containing proteins often yield the best hits against the filament-forming muscle myosin heavy chain proteins and are therefore called “[muscle] myosin”, although they lack the myosin motor domain defining the family),

Version 2024.11.11

- similarly, naming by including subfamily/class/subclass/group classifiers leads to tremendous confusion and misnomers, as protein family classification requires thorough phylogenetic analysis,
- and inconsistent naming approaches lead to additions such as “hypothetical”, “potential”, “probable” and others.

Apart from the database and database version used for comparison, the results of naming protein-coding genes strongly depend on the software used for alignment and the parameters used for filtering. Depending on the number of names assigned, a software can be considered accurate or sloppy, which always depends on the user's view. When naming proteins based on protein alignment, a lower E-value seems to be more appropriate and some sequence coverage seems to be required to exclude naming proteins based on small common domain motifs, which is particularly the case for eukaryotic protein datasets with their longer sequences.

Controversy around the max-target-seq parameter in BLAST

The controversy surrounding the BLAST parameter max-target-seqs arose from confusion about its behaviour and interpretation when filtering and reporting sequence alignments. Many users assumed that max-target-seqs controls the maximum number of hits (unique sequences) reported for each query. In reality, it controls the number of high-scoring pairwise alignments (HSPs) that are considered in post-processing and do not need to directly match unique sequences. If max-target-seqs is set to a low value (e.g. 1), BLAST can stop processing after finding the first significant match without guaranteeing that it is the best match. This behaviour led to confusion as BLAST sometimes produced suboptimal results when the best match appeared after the cutoff. The misunderstanding of the parameter is often the reason for the non-reproducibility of the results and many incorrect assignments in functional annotations.

Table 1: Comparison of BLAST and DIAMOND with respect to assigning protein homologs depending on E-value. Total number of proteins in the dataset: 46,664. Search database and version: Swiss-Prot v. 2024_06. Software and version: BLAST v. 2.13+ and DIAMOND v. 2.1.9.

	BLASTp			DIAMOND ultra-sensitive mode		
E-value	default E-value	#proteins with name	#proteins with EC	default E-value	#proteins with name	#proteins with EC
1e-5		32,785	17,171		31,163	14,888
1e-3		33,465	17,737	x	31,919	15,546
1e-1		35,734	20,080		32,864	16,416
10	x	46,162	42,203		37,005	20,950

Table 2: Comparison of DIAMOND alignment modes with respect to default settings and restrictions on query and subject coverage. Total number of proteins in the dataset: 46,664. Search database and version: Swiss-Prot v. 2024_06. Software and version: DIAMOND v. 2.1.9. Settings applied in mendle-analytics are highlighted in dark turquoise.

		default settings (`--query-cover` & `--subject-cover` not applied)		`--query-cover=50` `--subject-cover=25`	
E-value	mode	with name	with EC	with name	with EC
1e-3	fast	27,337	12,573	24,147	11,077
[default]	[default]	29,784	13,819	25,833	11,815
1e-3	mid-sensitive	31,033	14,719	26,725	12,226
1e-3	sensitive	31,530	15,185	27,111	12,393
1e-3	more-sensitive	31,540	15,187	27,178	12,421
1e-3	very-sensitive	31,839	15,446	27,342	12,504
1e-3	ultra-sensitive	31,919	15,546	27,389	12,540

Identification of protein domains, families and peptide patterns

Functional domains are identified by comparison with protein signature databases that contain conserved patterns, motifs or domains associated with specific functional or structural properties of proteins. This can be done by searching individual databases or by using InterProScan, which allows simultaneous searching of many databases such as Pfam, SMART, TIGRFAMs, PROSITE and others (17 in total). Based on the matches found, InterProScan assigns functional annotations to the input sequence.

Information associated with protein names and domains

Other information is mapped indirectly. GO terms are associated with UniProt proteins and Pfam domains and are assigned to proteins via the matching proteins and domains. EC numbers (enzyme commission numbers) are associated with UniProt proteins and assigned in this way.

mendle-analytics functional annotation

The protein names are assigned to the input sequences by comparison with the latest UniProt database. We assign the name of the best match as the protein name and give the main consensus of the names of the top 20 matches as the protein family name after removing name prefixes,

Table 3: InterProScan analysis of a protein sequence dataset. The tools with most hits were regarded as reference. Search hits with other tools are given as included in the subset of proteins with Pfam and PANTHER hits or independent (not included), respectively. Software and version: InterProScan 5.72-103.0.

total proteins	46,664
Pfam	32,945
PANTHER	38,141

	included	independent
CDD	13,869	19
ProSitePatterns	7,812	12
ProSiteProfiles	15,008	111

	Included	independent
TIGRFAM	4,241	4
SFLD	289	0
SUPERFAMILY	24,733	226
Gene3D	26,182	213
Hamap	1,679	0
Coils	6,708	1,066
SMART	11,790	6
PRINTS	5,476	51
PIRSR	0	0
AntiFam	1	3
MobiDBLite	19,386	3,720
PIRSF	1,899	0

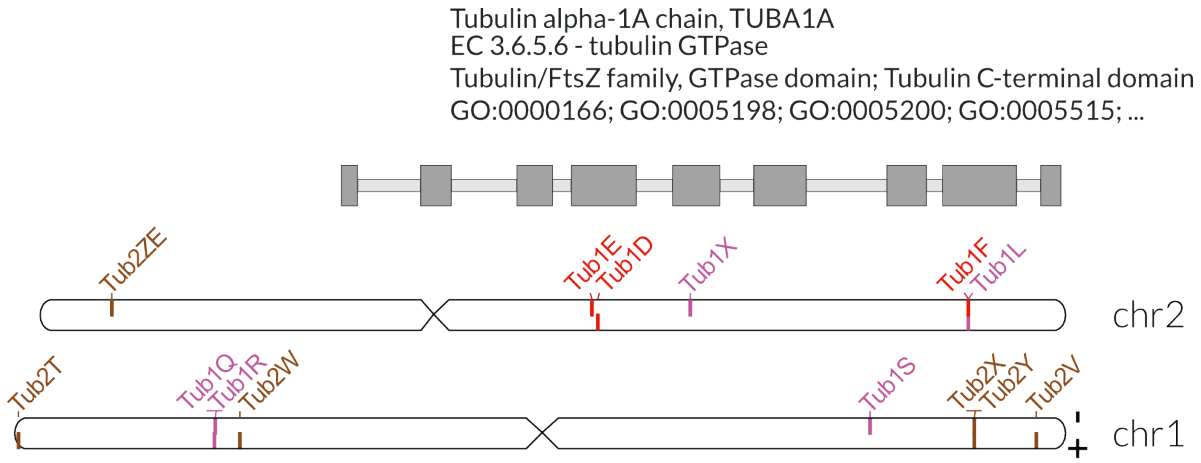


Figure 2: Functional annotation example. Annotation of tubulin genes at the scale of chromosome regions (bottom). Gene structure of the Tub1A gene and assignment of protein name, EC number, domain annotations and GO terms.

suffixes and subfamily classifiers. For the assignment of protein domains and motifs, we use InterProScan only with the Pfam, CDD and ProSitePatterns databases. The assignment of protein families via the provided protein family databases is misleading due to the outdated databases and the general problem of protein family evolution, which requires a thorough phylogenetic analysis. The prediction of coiled-coil regions by COILS has been shown to be random (Simm *et al.* 2021, Scientific Reports 11, 12439) and is therefore ignored. The prediction of protein domains by SMART and others differs from the predictions of Pfam and CDD only by minor differences in the start and end of the domains. Overlapping domain predictions from InterProScan are removed. Go terms are summarized with GOATOOLS and analyzed and plotted with custom tools.