International Protein Nomenclature Guidelines

Mission statement
These guidelines have been produced jointly by the European Bioinformatics Institute (EMBL-EBI), the National Center for Biotechnology Information (NCBI), the Protein Information Resource (PIR) and the Swiss Institute for Bioinformatics (SIB) and are intended for use by anyone who wants to name a protein to promote consistency in protein naming across databases, which aids data retrieval and improves communication.

Table of contents

1. Introduction

Consistent protein nomenclature is indispensable for communication, literature searching and entry retrieval. A good protein name is one which is unique, unambiguous, can be attributed to orthologs from other species and follows official gene nomenclature where applicable. The process of associating a name with a protein sequence has various components: sequence function identification/prediction, choosing a name and applying formatting. This document provides guidelines on naming choices and universal formatting. This does not include best practices on methods to be used for sequence function identification/prediction.

2. Formats for Protein Names

A. Language

- **Use American spelling, not British spelling**
  Examples:
  - *uncharacterized protein **not** uncharacterised protein*
  - *hemoglobin **not** haemoglobin*

- **Use protein names ending in 'in' (not 'ine')**
  *Example: maurocalcin **not** maurocalcine*

- **Avoid diacritics such as accents, umlauts etc.**
  *Example: protein spatzle 5 **not** protein spätzle 5*

- **Avoid pluralization for names based on domain and repeat content**
  *Example: ankyrin repeat-containing protein **not** ankyrin repeats-containing protein*

- **Avoid common words**
  Avoid naming proteins with common words which makes querying difficult e.g. avoid names such as 'protein IMPACT'.

- **Avoid duplication**
  Check if the proposed name for a newly discovered protein is already used for a different protein.

B. Abbreviations and symbols

- **Avoid using an abbreviation as the complete name**
  *Example: acyl carrier protein **not** ACP*

- **An abbreviation may be part of a protein name**
  Example: (3R)-hydroxymyristoyl-ACP dehydratase
  See below for a list of standard scientific abbreviations.

- **Protein name based on a protein symbol (PS) or gene symbol (GS)**
  Protein and gene symbols should use the same abbreviation. Some gene and protein symbols are easily recognized by database users in certain research communities and can be used as part of a protein name to provide specification and aid data retrieval.

    - **Prokaryote symbol guidelines**
        - A protein symbol is most commonly used in prokaryote protein names in combination with a functional protein name.
        - The first letter of a protein symbol is capitalized for prokaryotes e.g. RecA.
        - In rare occurrences when there is no functional protein name, the format "protein <PS>" may be used, not "<PS> protein".
        - *Example: recombinase RecA*

    - **Eukaryote symbol guidelines**
        - A gene symbol is commonly used in eukaryote protein names in combination with a functional protein name.
        - Capitalization conventions of gene symbols differ between organism communities and this is reflected in the casing of gene symbols used as part of eukaryotic protein names. For vertebrates, use an all

uppercase gene symbol in a protein name. For non-vertebrate eukaryotes, follow the gene casing conventions of the species in question.
    - In the case of conserved genes, if there is no known gene symbol in use in the species already, a known orthologous gene symbol from a species where the symbol was originally defined may be used.
    - In rare occurrences when there is no functional protein name, the format "protein <GS>" may be used, not "<GS> protein".
    - *Examples:*
        - Human: *tyrosine-protein kinase ABL1*
        - Mouse: *tyrosine-protein kinase ABL1*
        - *C.elegans*: *tyrosine-protein kinase abl-1*
        - *D.melanogaster*: *tyrosine-protein kinase Abl*
        - *S.cerevisiae*: *recombinase RAD51*
        - *S.pombe: recombinase rad51*

- **Prime symbol (')**
    - Use to indicate the cleavage location on a substrate and to distinguish different subunits with the same notation.
    - Use the single quote character (not the backtick) for the prime symbol.
    - *Examples:*
        - *H(+)-transporting V0 sector ATPase subunit c'*
        - *5'-nucleotidase **not** 5-prime-nucleotidase*
        - *coatomer subunit beta' **not** coatomer subunit beta-prime*

- **Chemical symbols may be part of a protein name**
    - For elements with a single valence type, use the full element name with no valence indicated.
    - For elements that have variable types of valency, use the chemical symbol for the element followed by the valence in parenthesis.
    - *Examples:*
        - *sodium/lithium-exporting P-type ATPase **not** Na(+)/Li(+)-exporting P-type ATPase*
        - *magnesium transporter **not** Mg(2+) transporter*
        - *Fe(3+)/Cu(2+)-chelate reductase **not** ferric/cupric-chelate reductase **or** **not** Fe(III)/Cu(II)-chelate reductase*

- **Standard scientific abbreviations may be part of a protein name**
    - Deoxyribonucleic acid: DNA, cDNA, dsDNA, ssDNA
    - Ribonucleic acid: dsRNA, mRNA, miRNA, piRNA, siRNA, snRNA, snoRNA, ssRNA, tRNA, tmRNA, rRNA
    - Mono-, di-, tri-nucleoside phosphates: dAMP, dCMP, dGMP, dTMP, dADP, dCDP, dGDP, dTDP, dATP, dCTP, dGTP, dTTP
    - Cofactors: FAD, FMN, NAD, NADP
    - Classes for transporters that inform about structure (e.g. ABC, MFS, RND, MATE, SMR) rather than substrate (e.g. **not** MDR)
    - *Example: rRNA methyltransferase **not** ribosomal RNA methyltransferase*

C. Punctuation

- **Slash**
  - Do not use a back slash: '\'.
  - For separating multiple domains or functions, the forward slash '/' or the word 'and' may be used.
  - *Examples:*
    - *adenylyltransferase/ADP-heptose synthase cyclohydrolase **not** adenylyltransferase\ADP-heptose synthase cyclohydrolase*
    - *WD repeat and FYVE domain-containing protein 3 **not** WD repeat\FYVE domain-containing protein 3*

- **Hyphen**
  - **Compound adjective:** a hyphen should be used to form compound modifiers (i.e. two or more words that are acting as a single modifier for a noun)
    Examples:
    - *Ras GTPase-activating protein **not** Ras GTPase activating protein*
    - *secretin-binding protein **not** secretin binding protein*
    - *pyrophosphate-dependent phosphofructokinase **not** pyrophosphate dependent phosphofructokinase*
  - **Examples of common modifiers:** activated, activating, adapting, adding, amplified, anchored, anchoring, antagonizing, associated, associating, attracting, binding, blocking, bound, branching, bridging, bundling, capping, complementing, concentrating, conjugating, containing, controlled, controlling, converting, coupled, coupling, decapping, degrading, dependent, depolymerizing, derepressing, derived, deriving, destabilizing, docking, editing, enhanced, enhancing, enriched, exposed, flanking, forming, gated, grabbing, harvesting, independent, induced, inducible, inducing, inhibited, inhibiting, insensitive, interacting, laying, like, linked, linking, metabolizing, modifying, modulating, polymerizing, potentiating, preventing, processing, promoting, recognizing, recruited, recruiting, regulated, regulating, related, released, releasing, remodeling, removing, repressing, required, requiring, resistant, responsive, rich, ripening, scaffolding, sensing, sensitive, signaling, specific, splicing, spreading, stabilized, stabilizing, stacking, stimulated, stimulating, structuring, sulfating, suppressing, trafficking, transformed, transforming, transporting
  - **More than one domain/repeat in a name:** if there is more than one domain/repeat, only use a hyphen for the last item preceding "containing", even though this violates conventional grammar.
    *Example: ankyrin repeat and SAM domain-containing protein 6 **not** ankyrin repeat- and SAM domain-containing protein 6*

- **Avoid apostrophes, periods, commas and other undesirable punctuation**
  - Remove trailing periods from names.

- Avoid use of commas except when their usage is part of accepted chemical names.
  *Example: SGT2 family TPR domain-containing protein **not** TPR repeat protein, SGT2 family*
  ***Exception** example: 3-hydroxy-16-methoxy-2,3-dihydrotabersonine N-methyltransferase*
- Avoid the semi-colon ";" or colon ":" except when it is part of an enzyme name.
  Example: *type I cuticular keratin Ha8 **not** "Keratin, type I cuticular Ha8; Hair keratin,type I Ha8; Keratin-38; K38"*
  ***Exception** example: phospholipid:diacylglycerol acyltransferase*
- Avoid the percentage sign '%'
- Avoid the at sign '@'
- Avoid the equal sign '='
  *Example: guanine nucleotide-binding protein G(t) subunit alpha-3 **not** gustducin:SUBUNIT=alpha*

- **Avoid autocorrection of protein names**
  - Data submitters should not let Microsoft Excel, Word, Outlook, or any other utility with format interpolation and spelling autocorrection touch any protein names, especially those with quotes and double-hyphens.

D. Notation

- **Use Arabic rather than Roman numerals**
  Use Arabic numbers for notation (e.g. 1, 2, 3, etc.) unless Roman numerals are a widely accepted formal nomenclature like "RNA polymerase II".
  *Example: caveolin-2 **not** caveolin-II*
  ***Exception** example: DNA-directed RNA polymerase II core subunit RPB2*

- **Specifying different members encoded by a multigene family**
  Use Arabic numbers to specify the different members encoded by a multigene family. Refrain from inventing new numbers if a notation system for protein/gene family members has been previously published.

E. Style and format

- **Capitalization**
  Use lowercase except for acronyms or proper nouns.
  Examples:
  - *proteasome core particle subunit beta 5 **not** Proteasome CORE PARTICLE subunit BETA 5*
  - *enolase **not** ENOLASE*

- **Greek letters**

- Greek letters should be written in full and entirely in lower case when indicating one of a series of proteins e.g. "alpha", "beta", “gamma”.
- In the context of steroid/fatty acid metabolism nomenclature, “Delta” should start with an upper case letter.

- **Usage of the term 'protein' in a name**
  - Avoid if not necessary, especially when the name includes terms such as "factor", "enzyme", "inhibitor" or "regulator".
  - Enzyme names commonly end with 'ase' (aminoacylase, arginase, etc). Do not append the term ‘protein’ to the enzyme name.
    Examples:
    - *Fe(3+) uptake regulator **not** Fe(3+) uptake regulator protein*
    - *ribonuclease **not** ribonuclease protein*

- **Usage of the term ‘enzyme’ in a name**
  Enzyme names commonly end with ‘ase’ (tautomerase, phosphotransferase, etc). Do not append the term ‘enzyme’ to the enzyme name.

- **Protein name based on a pathway**
  Use this format: "<Pathway> synthesis protein <GS>"
  Examples:
  - *thiamine synthesis protein ThiC*
  - *folic acid synthesis protein FOL1 **not** trifunctional dihydropteroate synthetase/dihydrohydroxymethylpterin pyrophosphokinase/dihydroneopterin aldolase FOL1*

- **Transfer enzymes**
  Transfer enzymes are often indicated with the source and destination substrate separated by a double hyphen (--).
  Example: formylmethanofuran--tetrahydromethanopterin formyltransferase

- **tRNA-charging enzymes**
  Use this format: <amino acid being attached>--tRNA (tRNA type using the three-letter amino acid code with the first letter capitalized) ligase.
  Example: tyrosine--tRNA (Tyr) ligase

- **Identifier types to avoid**
  COG ID, EC number, FOG ID, GO terms, cluster identifiers.
  Stable locus tags and stable HMM identifiers should be used only in special situations in which they point to families of proteins, and this is made clear by a qualifier in the protein name such as “family protein” or “domain-containing protein”. They should not be used for naming low copy conserved proteins. A protein name based on a locus tag (e.g. MA_1614) can never be transferred by homology, even to identical proteins, because locus tags indicate a position in one specific genome. All use of locus tags in protein names is discouraged because of the danger that simplistic annotation methods can too easily make overly specific (and therefore incorrect) assertions. The one exception is the use of a locus tag in combination with a

"family" qualifier, where the locus tag is frequently used in the literature from an annotation present in the INSDC and frequently used in comparative analyses and it is necessary to distinguish among proteins that otherwise would receive insufficiently informative names, e.g. "BB3110 family autotransporter". Names based on a Hidden Markov Model (HMM) identifier similarly may be used to improve clarity. These too must be qualified by the terms "family protein" or "domain-containing protein". See section 3B about Novel proteins of unknown function.

- **Avoid kingdom, genus or species-specific characteristics in a name**
  - Avoid expression, abundance information, disease, phenotype and anatomy-related information.
  - Avoid cellular, subcellular and environmental location. Location information is not always transferable among all organisms and should be applied conservatively.
  - Avoid molecular weight except for ribosomal proteins and well-established historical names, e.g. myosins, clathrins, dyneins.
    Exception *examples:*
    - *Eukaryotes: 60S ribosomal protein subunit L19B*
    - *Prokaryotes: 50S ribosomal protein subunit L1*
    - *myosin heavy chain 1*
  - Avoid referencing chromosomal or cytogenetic locations of the gene
    Example: methylcytosine dioxygenase TET1 ***not*** *ten-eleven translocation-1*
  - Avoid locus_tag identifiers.
  - Avoid regulatory content such as 'regulated by', 'regulates'.
  - Avoid organism names or abbreviations of species/genus/kingdom of origin or homologous species. An exception to this is adjectival organism names which can be included in rare cases where it will make a name more descriptive and less general.
    Exception *example: staphylococcal nuclease domain-containing protein 1*

F. Word usage

- **Avoid linking words and phrases**
  - Avoid the following linking words: for, or (as in name1 or name2), of, to, with.
    Example: *two-component system sensor histidine kinase* ***not*** *histidine kinase sensor of two component system*
  - Avoid the following linking phrases: also known as, together with.

- **Other phrases to avoid**
  - cell surface, cell surface protein, conserved hypothetical, hypothetical conserved, identified by, identity to, involved in, implicated in, protein domain protein, protein of unknown function, protein hypothetical, protein protein, protein putative, putative putative, questionable protein, related to, signal peptide protein, similar to, surface antigen, surface protein, unknown protein, authentic point mutation, low quality protein, C term(inal), N

term(inal), inactivated derivative, conserved uncharacterized, uncharacterized conserved

- **Terms to avoid**
  - antigen, CDS, conserved, cytoplasmic, deletion, dubious, doubtful, expressed, fragment, frame shift, frameshift, genome, homolog (unless phylogenetically determined), interrupt, KDa, K Da, likely, locus, locus_tag, novel, ORF, partial, possible, potential, predicted, probable, pseudo, pseudogene, secreted, strongly, truncat(ed), under, unique, unnamed, WGS, Xray, X-ray
  - Naming proteins as antigens is discouraged but there may be rare exceptions to match widespread community/publication usage.
    Exception example: *cellular tumor antigen p53*
  - Note that use of the term 'putative' is acceptable in certain cases - see the topic "Novel proteins of unknown function" in section 3B.

3. Choosing Protein Names

A. Sources of protein name annotation

Protein names are ideally supported by evidence from expert sources, the literature, HMMs and other protein signatures, and/or domain architectures. NCBI-RefSeq and UniProt aim to store, and publicly report, name source information of curated records which may include the expert database name, individual scientist name, PubMed ID, HMM ID, and curated domain architectures. The current rank of sources for protein naming is: a) expert sources > b) experimental reports > c) HMMs and other signatures > d) domain architectures. Note that BLAST results, FASTA headers and definition lines in database records may contain information such as organism names and other information which should not be included in a protein name. Be aware that sources of functional protein annotation listed below do not necessarily meet all the international protein nomenclature guidelines. In particular, resources may not be available to retroactively update older data.

a) Expert sources of specific and definitive names may include:

Species-specific naming authorities

- Established and maintained database authorities such as species-specific nomenclature bodies (some are listed here: http://www.uniprot.org/docs/nomlist).
- Avoid names from species-specific authorities that relate to phenotype, anatomical features or any taxon-specific characteristics. In these cases, use the widely recognized gene symbol in combination with a functional name rather than a phenotypical name. For example, 'minichromosome maintenance complex component 7' is not applicable to organisms which do not have minichromosomes so to avoid transferring such a protein name, use the gene symbol MCM7 combined with a functional name instead.
  Example: DNA replication licensing factor MCM7 ***not*** *minichromosome maintenance complex component 7*

Enzyme names from Enzyme Commission (EC)

- Strong preference to use the preferred name when it is a specific and accurate reflection of the main function of the protein and the EC name is neither too general nor too specific to apply to a group of proteins.
- In contrast, expert curators may override the EC name in certain circumstances such as when the name is not the primary function of the enzyme or they may choose an alternative EC name if the preferred EC name ends with a qualifier in parenthesis or contains two or more sets of brackets/parentheses.
    - *Examples:*
    - *ABC transporter ATP-binding protein **not** ATPase*
    - *NADP-dependent isocitrate dehydrogenase IDP3* **rather than** *isocitrate dehydrogenase (NADP(+)) IDP3*
    - *phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase* **rather than** *1-(5-phosphoribosyl)-5-((5-phosphoribosylamino)methylideneamino)imidazole-4-carboxamide isomerase*
- Keep the double hyphen '- -' used for transferases and ligases.
  Example: formylmethanofuran--tetrahydromethanopterin formyltransferase
- Use the following format for enzymes that remove or transfer phosphate groups: "<modified_residues>-protein <activity>".
  Example: tyrosine-protein phosphatase

UniProtKB/Swiss-Prot

- UniProtKB/Swiss-Prot name of an orthologous or paralogous protein, provided that it meets the guidelines in this document.

Other

- Individual scientists who specialize in a protein family.

b) Experimental reports

- A recent literature-supported name from a paper that characterized the protein function is likely the most specific and definitive name to apply (with format refinement as needed). The literature may provide a history of names over time.
- Newer more functionally specific names are preferred over older more general or biosystem-related names.

c) HMMs and other signatures

- Equivalogs are homologs that have retained a specific function from their common ancestor, whatever the evolutionary path of each protein. This stands in contrast to the definitions of orthologs (homologs from speciation events only), paralogs (homologs from duplication events), and xenologs (homologs from lateral transfer events), all of which make no assertion about function.
- An equivalog-type HMM is any HMM that asserts its member proteins share a specific function, and that supplies a descriptive protein name and other attributes for automated pipelines to use during genome annotation.
- Most TIGRFAM models are designated equivalogs, meaning they assign a specific name to proteins conserved in function from a common ancestral sequence.

Example of an equivalog type name versus a general name (see UniProtKB/SwissProt record P0A288): peptide chain release factor 1 ***versus*** *PCRF domain-containing protein*

- To apply names to proteins related to proteins named by equivalog type HMMs, use XXX-like protein or XXX family protein. These synonymous terms will carry the association that, despite obvious sequence similarity to XXX, it may or may not have the same role and function as XXX and thus it might be XXX itself, or something related. Also see the usage of "putative XXX" in section 3B about Novel proteins of unknown function.
  Example: glycine cleavage protein H-like protein for proteins of the family TIGR03077. These proteins are not bona fide glycine cleavage protein H which belong to family TIGR00528.

d) Profiles and domain architectures

- The domain architecture is defined as the sequential order of conserved domains in a protein sequence. In some cases the architecture consists of a single domain that covers the full length of the protein. Domain architecture names are usually more general than equivalog-type HMM names but provide additional protein naming evidence. A protein name based on a multi-domain architecture is more informative than a protein name based only on domain content.
  - *Example: PAS domain-containing sensor histidine kinase* (based on a multi-domain architecture).
- Protein names can be based on a single domain which does not cover the full length protein and may be associated with varied architectures.
  - *Example: PAS domain-containing protein* (general)
- Be cautious when parsing domain names. Automatically extracting a name from a domain or profile may end up being uninformative e.g. Pfam accession PF00083, Sugar_tr which results in a protein product called 'sugar'.

B. Naming procedure for specific cases

- **Multifunctional proteins**
  - Multifunctional proteins may catalyze multiple enzymatic reactions such as human protein GNE which has both epimerase and kinase activities or they may be involved in different functions such as *Arabidopsis thaliana protein ENO2* which acts as an enolase and is also involved in transcription regulation.
  - No need to list all functions.
  - If no other name is applicable, the words bifunctional or multifunctional may be used in combination with the functional names.
  - When using bifunctional, list the functions based on the order of the domains in the sequence and separate them with a forward slash.
  - In rare cases and when no other name is applicable, enzymes with more than two functions may use the format: "multifunctional protein <GS>".
  - *Examples:*

        - *bifunctional adenylyltransferase/ADP-heptose synthase cyclohydrolase*
        - *fatty acid oxidation complex subunit alpha **not** multifunctional enoyl-CoA hydratase/3-hydroxybutyryl-CoA epimerase/3-hydroxyacyl-CoA dehydrogenase*
        - *multifunctional proline degradation protein PutA **not** multifunctional DNA-binding transcriptional repressor/proline dehydrogenase/1-pyrroline-5-carboxylate dehydrogenase*

- **Naming proteins based on protein complex membership**
    - Protein complex members for well-defined multi-subunit complexes of known composition can be named according to the complex followed by the specific subunit name.
    - Use 'subunit', not 'chain' or 'component', for members of protein complexes. The exception is historical cases where 'chain' is exclusively used e.g. myosins, clathrins, dyneins.
      ***Exception** example: myosin heavy chain 1*
    - If the 'type' of subunit is known, then 'type' goes first where 'type' can be catalytic, ATP-binding, regulatory etc.
      Example: 26S proteasome non-ATPase regulatory subunit 1
    - If a subunit has a designator, then that follows the term 'subunit', e.g. subunit 1, subunit A, subunit AbcD, subunit alpha. The preference for designator use is: number > letter > gene symbol > greek letter spelled out.
      Example: *F1F0 ATP synthase subunit alpha **not** F1F0 ATP synthase alpha subunit*
    - An abbreviation may be part of a protein complex name.
      Example: (3R)-hydroxymyristoyl-ACP dehydratase
    - Avoid 'large subunit' or 'small subunit' when possible, but well-established historical names are an exception.
      ***Exception** example: 2,3-diketo-L-gulonate TRAP transporter large permease*

- **Inactive proteins**
  Inactive proteins do not refer to pseudogenes. Inactive versions of proteins refer to proteins with altered catalytic residues or inability to undergo autocatalytic cleavage, resulting in loss of expected activity. Reserve the usage of "inactive" in a protein name for such cases.
  Example: inactive glutathione hydrolase 2

- **Novel proteins of unknown function**
  Where no functional information is available, any of the following methods may be used to name a protein.

    - **where domains, repeats or motifs associated with a variety of architectures are observed:** Use the format 'xxx domain-containing protein' but avoid transferring 'xxx domain-containing protein' names based on a BLAST search. Use a protein signature search instead.
      Example: PAS domain-containing protein

- **where sequence similarity to a defined protein family is observed:** Avoid asserting the function of the family. Use a general name such as 'XXXX family protein'. Proteins given the name "XXX family protein" might be XXX itself, or something related. The name "XXX family protein" may be thought of as an unspecific and temporary name that will be replaced when more specific annotation becomes available.
  Example: flavodoxin family protein

- **where a known family protein has a predicted activity:** In general, use of the word 'putative' should be avoided. In this specific case, prefix the activity with 'putative', not the whole protein name. The term 'putative' should be located before the activity that it refers to. "putative XXX" should be used when "XXX" is considered the most likely prediction, but the reasoning used to perform the annotation carries with it enough doubt that the disclaimer is useful. The term should not be used in an automated fashion simply to mean "protein showing low-scoring homology to XXX".
  Examples:
  - *radical SAM family putative peptide maturase **not** putative radical SAM family peptide maturase **or not** radical SAM family peptide maturase, putative*
  - *putative acetylornithine deacetylase **not** predicted acetylornithine deacetylase **not** possible acetylornithine deacetylase **not** probable acetylornithine deacetylase **not** potential acetylornithine deacetylase **not** hypothetical acetylornithine deacetylase*

- **where a full length protein HMM or other signature match associated with a single architecture (equivalog type signatures) is observed:** Use the HMM name or other protein family signature name to name the protein, conforming to the rules of this document. Caution: Protein signature identifiers and the signatures themselves are not stable and may change, requiring review and renaming of proteins named using this method.
  Example: TIGR01212 family radical SAM protein

- **where no domain or motif is observed:** If a gene symbol or protein symbol has been published for this protein, use the protein <GS> or protein <PS> format. Otherwise, use the default name 'hypothetical protein' or 'uncharacterized protein' (all lowercase) with no further specifications.
  Examples:
  - *hypothetical protein **not** hypothetical protein, conserved*
  - *uncharacterized protein **not** uncharacterized protein conserved in archaea*
  - *protein XYZ1*