

## **Functional Annotation Descriptors**

### **Protein name:**

The primary source of a protein name is the protein name assigned to an experimentally characterized gene in *Entamoeba histolytica*. If there is no homolog to an experimentally characterized gene in *Entamoeba histolytica*, then the protein name chosen is the protein name assigned in Swiss-Prot. In all other cases, the protein name is selected from the standard enzyme commission (IUBMB) name.

### **Gene symbol:**

Gene symbols are assigned based on homology to an experimentally characterized gene in *Entamoeba histolytica* where homology is deemed "identical".

### **Enzyme commission (EC) number:**

EC numbers are assigned in an automated fashion using [PRIAM](#).

### **Public comments:**

Public comments are added to each functional assignment when:

1. functional annotation is based on homology to an experimentally characterized gene in *Entamoeba histolytica*.
2. functional annotation is based solely on homology to an experimentally characterized gene of another organism and no other supporting evidence was used.

Public comments document the level of support used for each functional assignment by describing the degree of homology to an experimentally characterized gene:

- a. identical - used for a 100% identical homolog to an experimentally characterized gene in *Entamoeba histolytica*
- b. similar - used for a >50% identical homolog to an

experimentally characterized gene in any organism.

c. weak similarity - used for a <50% identical homolog to an experimentally characterized gene in any organism.

Gene Ontology (GO) terms:

GO terms are assigned in an automated fashion using PFAMs.

Levels of Functional Annotation Specificity

**Confident assignment:**

I. Homolog to an experimentally characterized gene in *Entamoeba histolytica* with strong supporting evidence. Homology must be full length (>90%) and >95% identical to an *Entamoeba histolytica* experimentally characterized gene. In these cases, the protein name chosen is the protein name assigned to the experimentally characterized *Entamoeba histolytica* homolog.

Examples:

[EHI\\_007150](#): mitochondrial-type heat shock protein 70

II. Homology to a full-length match to a highly specific (i.e., 'equivalog' isology type) HMM. Homology must be above the trusted cutoff score to that equivalog level HMM. In these cases, the protein name chosen is the protein name assigned to the HMM.

Examples:

**Function uncertain:**

Homolog to an experimentally characterized gene (includes prokaryotic organisms), but uncertainly exists. Homology should be full length (>70%) and >30% identical. In these cases, the protein name is followed by "putative".

Examples:

[EHI\\_129890](#): type A flavoprotein, putative

[EHI\\_029610](#): maltose O-acetyltransferase, putative

[EHI\\_178030](#): calcineurin catalytic subunit A, putative

Note: The protein name assigned to family\_230 is based on homology to a prokaryotic experimentally characterized gene, while family\_279 is based on homology to an *Entamoeba histolytica* experimentally characterized gene. In both cases the respective publication was used in the public comment along with a "weak similarity" designation.

**Specificity uncertain (i.e., family classification):**

Full length (>70%) homolog to members of a defined family whose functions are either known or unknown. In these cases, the appropriate family designation should be used followed by the family name abbreviated in parentheses.

Examples:

[EHI\\_162540](#): membrane-bound O-acyltransferase (MBOAT) family protein

[EHI\\_060330](#): peptidase S54 (rhomboid) family protein

**Note:** Conceptually, a protein family (or subfamily) should consist of evolutionarily-related proteins with the same function in different organisms, while evolutionarily related proteins whose functions have diverged form a superfamily. In practice these distinctions are not rigidly adhered to in many protein databases, most of whose protein families are defined by sequence relatedness without verification that all family members share the same function. In this context a defined family means: a family for which an HMM has been built, or which can be identified through the literature, found in SWISSPROT, PROSITE, or some

other similar curated database. For each instance a protein family is encountered, all supporting documentation is examined to determine if function is likely to be consistent for all members of the family before assigning a family designation.

**Limited sequence similarity (i.e., domain classification):**

Less than full length (<90% but >30%) homolog to members of a defined domain whose functions are either known or unknown. In these cases, the appropriate functional domain designation should be used followed by "domain containing protein".

**Examples:**

[EHI\\_116940](#): deoxyuridine 5'-triphosphate nucleotidohydrolase domain containing protein

**Evidence consists of uncharacterized proteins:**

I. Full length (>70%) homolog to experimentally characterized proteins in other species that are <30% identical. In these cases, the protein name assigned is "hypothetical protein, conserved" and the respective publication is used in the public comment along with a "weak similarity" designation.

**Example:**

[EHI\\_120630](#): hypothetical protein, conserved

II. Homolog to conceptual translations in other species; there are no characterized matches or other evidence to indicate true function. In these cases, the protein name assigned is "hypothetical protein, conserved".

**Example:**

[EHI\\_169260](#): hypothetical protein, conserved

III. Homolog to conceptual translations in other species; there are no characterized matches or other evidence to indicate true function other than weak domain HMM evidence. In these cases, the protein name assigned is "hypothetical protein, conserved domain containing".

Example:

[EHI\\_168500](#): hypothetical protein, conserved domain containing

[EHI\\_111830](#): hypothetical protein, conserved domain containing

**Note:** A "weak similarity" designation is used to describe weak HMM evidence in the public comment.

**No database match:**

Lacking significant similarity to any previously published genes from other species, families, or motifs. In these cases, the protein name assigned is "hypothetical protein".

Examples:

[EHI\\_106660](#): hypothetical protein

**Small signature sequences:**

Small signature sequences in proteins can be defined as conserved insert or deletions that generally constitute contiguous patterns of amino acids 10-50 residues long and are associated with a particular structure or function. Signatures can facilitate the validation and detection of domain homologies, and a further step of confidence to assign function.

Examples:

Coming soon!

### **Functional Annotation of Disrupted Reading Frames**

When a potential disrupted reading frame is noted, the underlying structural annotation is reviewed and validated, as well as the underlying sequence and assembly. If warranted, a public comment is made to document a "potential frameshift error", marking the structural annotation for further review. For projects such as *Entamoeba dispar* and *invadens* which have low coverage (~5X sequence coverage), identification of true frameshifts and pseudogenes is particularly challenging and need expert review.