

PATRIC Standard Operating Procedures for Automated Protein Curation Pipeline (APCP)

Author: Joshua Shallom and Anjan Purkayastha
Version: 1.1, October 2, 2006 11:54 AM
Supersedes version: 1.0, October 2, 2006 9:00 AM

1. Introduction

This document describes the software requirements for the Automated Protein Curation Pipeline (APCP). The APCP will automate the transfer of annotation, based on a set of defined criteria, from publicly available, curated sequences to the coding sequences (CDSs) annotated as part of PATRIC's curation process. This automation is intended to save curatorial time and resources.

2. Functional Requirements

2.1 Scope

This version of the APCP will be run only on bacterial CDSs. The procedure will be run on the output produced by running each CDS through the Protein Annotation Pipeline (PAP) [1] and the ortholog set creation program (orthoEES.pl) [2].

Only the following protein attributes will be considered for annotation transfer: Product name; EC number; Gene symbol; GO terms. Depending on the type of annotation transfer, each protein will be assigned one of three curation statuses: IC-Final; IC-Pending; IC-Provisional. CDSs with an IC-Pending or IC-Provisional status will be manually reviewed.

The APCP will also store the provenance of each annotation transfer.

2.2 Input Data

For each CDS, the following data are used as input for APCP:

1. The locus tag of the CDS
2. The isology type, score and P-value for hits to the TIGRFAM database.
3. BLAST scores for hits to the SwissProt database.
4. BLAST scores for hits to the NCBI nonredundant database (excluding SwissProt hits).
5. Scores for hits to the InterProScan database [3].
6. Scores for hits to the COGs database [4].
7. ID of the ortholog set the to which the CDS belongs.

2.3 APCP Decision Tree

The following decision tree forms the core of the annotation transfer protocol. The PAP output of each protein is evaluated against this tree. A protein exits the decision tree once it inherits a set of annotation data and is assigned a curation status.

1. If the protein has a TIGRFAM hit above the trusted/noise cut-off, then evaluate the isology type of the TIGRFam hits.
 - a. If isology type = equalog, then the protein should inherit the TIGRFam hit. The product, GO assignment, gene symbol and EC # (if applicable) fields should be filled with the corresponding data from the TIGRFam annotation. Promote the protein curation status to IC-FINAL.
 - b. If isology type = paralog, subfamily, superfamily or domain, then the protein should inherit the product name and if present, the GO assignments, EC #'s and gene symbol.
Promote protein curation status to IC-PENDING.

NOTE: Hits to PFam are not being evaluated.

NOTE: As of October 2, 2006, the following aspects of APCP have not been implemented or applied to data available from the PATRIC project

2. If the protein does not satisfy condition 1, then check the proteins BLAST results for hits against the SwissProt database.
3. If the protein has a hit to a SwissProt entry, then evaluate the identity and coverage of the hit.
 - a. If the SwissProt hit identity is $\geq 95\%$ AND length of alignment/length of query $\geq 70\%$, then the protein should inherit the SwissProt annotation. The product, gene symbol, EC# (if available) fields should be filled with the corresponding data. Proceed to step 3.a.1.
 1. If InterPro hits exist and are better than $10E-4$, then the protein inherits the associated GO assignments, else the GO term fields are left blank. Promote protein status to IC-FINAL.
 - b. If the SwissProt hit identity is $<95\%$, then the protein should inherit the SwissProt product name according to Table 1 (see Appendix II). Proceed to steps 3b.1 and 3b.2.
 1. If InterPro hits exist and are above $10E-4$ then the protein inherits GO from InterPro.
 2. If hits to the COGs database exist, then the protein inherits EC #'s from COG.

Promote the protein curation status to IC-PROVISIONAL.

- NOTE:**
1. If SwissProt gives multiple Product names, the first one should be promoted and the others promoted as the aliases.
 2. If multiple EC #'s are available, all EC's are shown and have equal weight.
 3. If multiple gene symbols are available, the one associated with the first product name, is promoted and the others promoted as the aliases.

4. If the protein does not satisfy condition 3, then check other BLAST hits.
 - a. If BLAST hit is present, then evaluate the identity of the hit. The protein should inherit the BLAST product name according to Table 1 (see Appendix II). Proceed to steps 4b.1 and 4b.2.
 1. If InterPro hits exist and are above $10E-4$, then the protein inherits GO from InterPro.
 2. If hits to the COGs database exist, then the protein inherits EC #'s from COG.

Promote the protein curation status to IC-PROVISIONAL.

5. If the protein does not satisfy condition 4a, then check ortholog sets to determine if orthologs exist.
 - a. If orthologs exist promote to "hypothetical protein, conserved".
 - b. If orthologs are absent promote to "hypothetical protein"

Promote the protein curation status to IC-Final.

3. Other Requirements

3.1 Preserving Legacy Annotation

In all cases the Refseq product name string and gene symbol should be matched with the PAP assigned product name string and the PAP assigned gene symbol. If there is no exact match then the Refseq name and gene symbol should be preserved as a searchable synonym. This will be done to ensure that end-users can search the PATRIC website for legacy annotations.

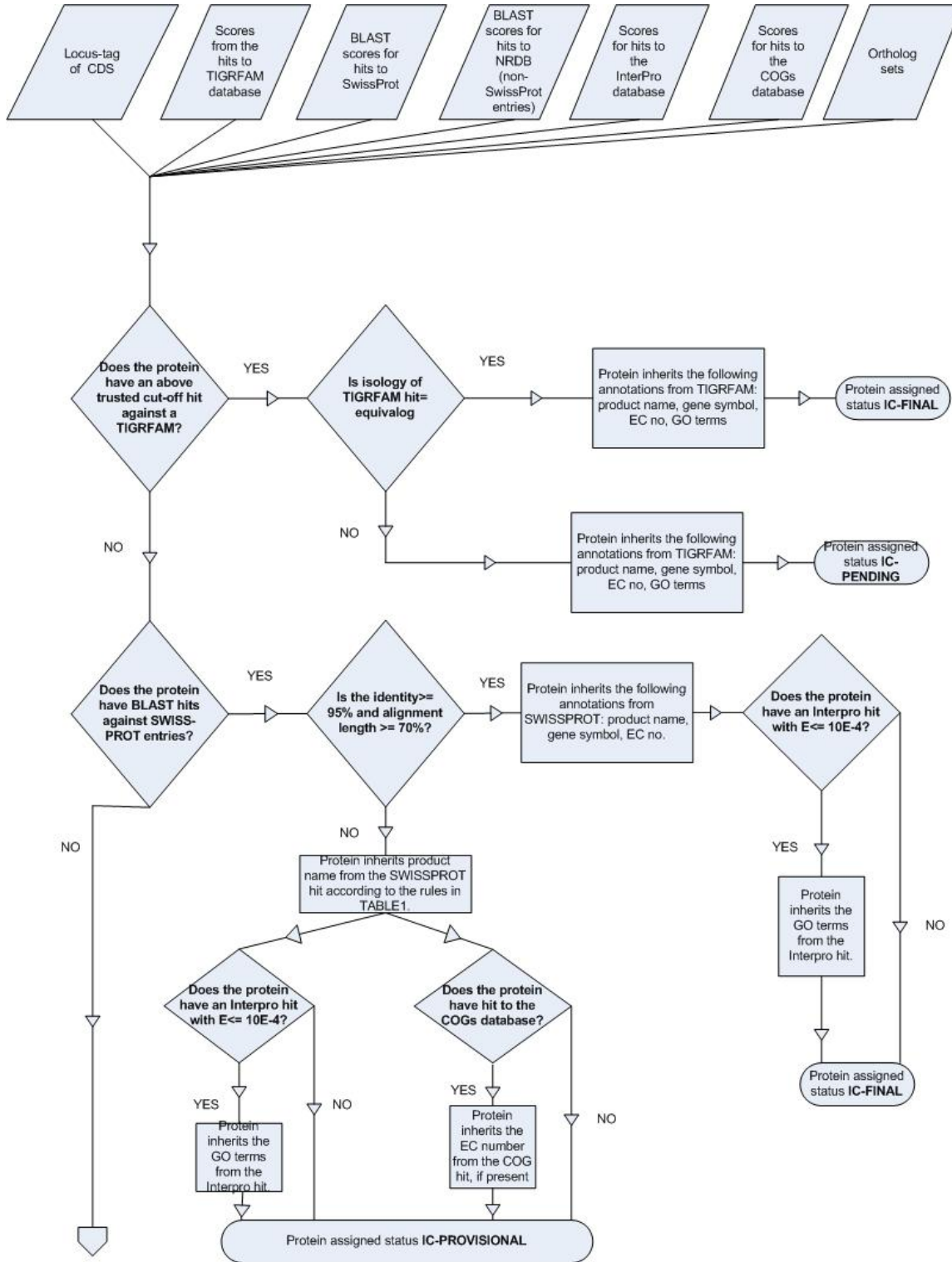
3.2 Audit-Trail for Annotation Transfer

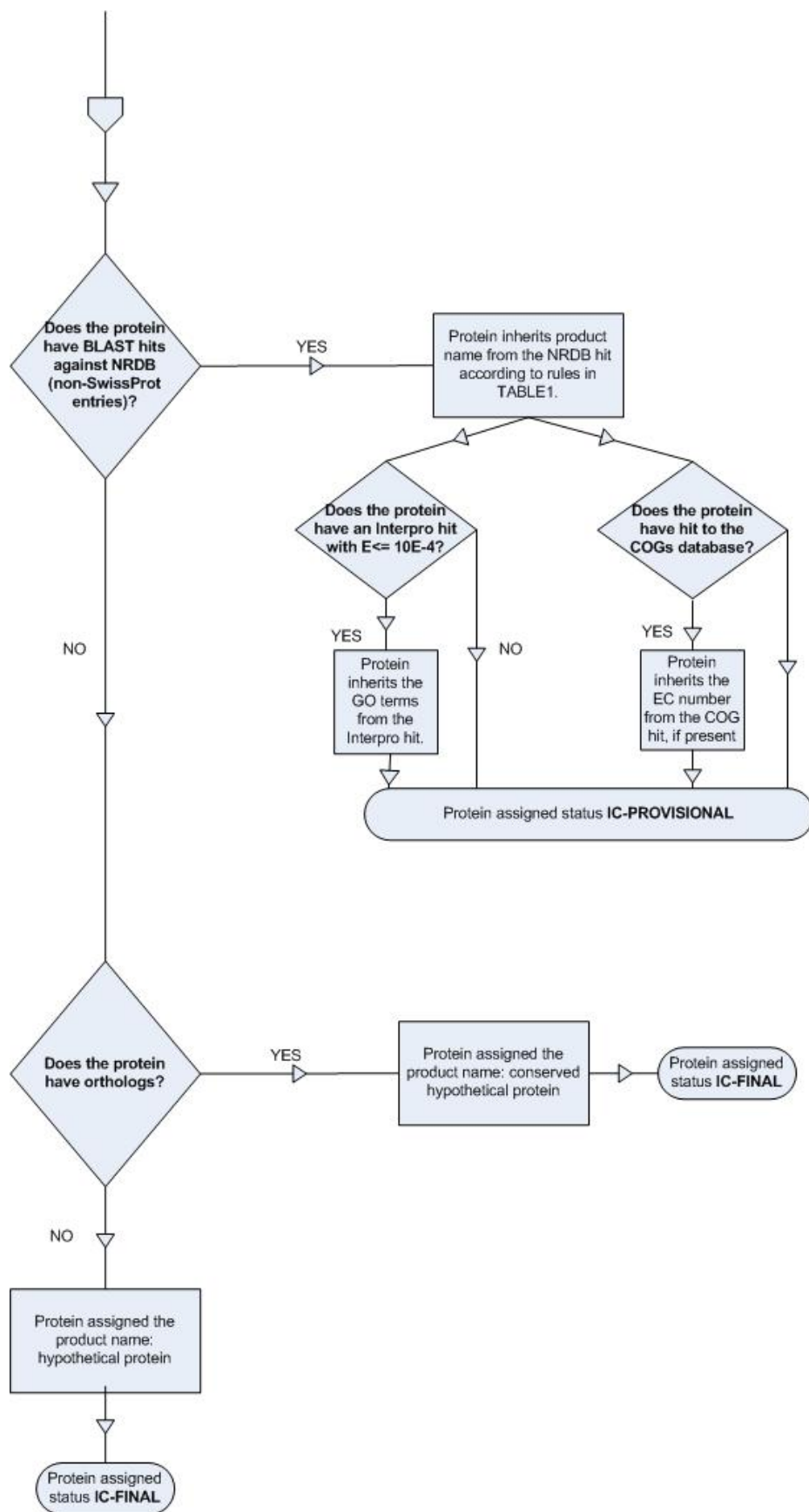
The APCP will record the provenance for each transfer of attribute information. This will include the name of the database and the unique identifier from which the annotation is inherited. For example, if the VBI protein with locus tag VBI0066RP_0011 inherits the product name "Dihydrouridine synthase" from a PFAM database entity with identifier PF01207, then both the database name (PFAM) and the unique identifier (PF01207) should be stored as the origin of transfer of the product name attribute. These should be accessible to both the curators and the end-users.

Appendices

Appendix 1

Flowchart of the APCP Decision Tree





Appendix 2

DATABASE MATCH	NAME	(COMMENTS FIELD)	% IDENTITY (length of alignment/length of Query)
Identical	*protein	(identical to ... characterization paper)	>95
Similar	*protein, putative	(similar to characterization paper)	50-94.9
Protein family	*protein family		40-49.9
Low similarity	*protein related		35.1-39.9
Conserved hypothetical protein within the same species	hypothetical protein, conserved (2 or more matches)		<35
Conserved hypothetical protein from another species	conserved hypothetical protein (1 or more matches)		<35
Predicted model, no database match	hypothetical protein		<35

Table 1. This table lays down the rules of protein product name transfer based on the percent identity between a VBI curated protein and a NRDB/SwissProt entry.