

PATRIC: Software Requirements for the Protein Annotation Pipeline and its User Interface.

Anjan Purkayastha¹, Eric Snyder¹, Gongxin Yu¹, Mark Hance¹, Michael Czar¹, Joesph Gabbard², Deborah Hix², Oswald R. Crasta¹.

¹Virginia Bioinformatics Institute, Washington Street, Virginia Tech, Blacksburg, VA 24061.

²Usability Engineering Laboratory, Systems Research Center, 562 McBryde Hall, Virginia Tech, Blacksburg, VA 24061.

1. Introduction

1.1 Document Overview

The Pathogen Resource Integration Center (PATRIC) was established in July 2004 with a goal to provide infectious disease researchers with a centralized and comprehensive annotation of the whole-genome sequences of a select group of microbial pathogens [1]. In following this stated objective PATRIC will create a software infrastructure that will automate the annotation of large biological sequence datasets. This infrastructure, named the Genome Annotation Pipeline (GAP) (see Appendix B, Figure 1), will be comprised of a set of pre-built sequence analysis tools, a database designed to store diverse kinds of biological information and a user interface that will allow for the manual inspection and editing of the data generated by automated annotation. One component of GAP is the Protein Annotation Pipeline (PAP)- the subsystem that performs the automated annotation of the protein-coding sequences of the PATRIC-archived microbial genomes. This document sets forth the software requirements for building the Protein Annotation Pipeline and its associated user interface.

1.2 Document Organization

The PAP requirements are divided into the following sections:

1) Introduction

An overview of the biological annotation process and a system overview of PAP.

A list of the users of GAP and its subsystems, their roles and responsibilities.

2) Functional requirements

A statement of functionality of PAP that lays out the scope of the pipeline.

The workflow of PAP and a prioritized list of tools to be included.

A prioritized list of performance and storage requirements and requirements for the user interface.

3) Other requirements

Requirements for user access and database updates.

4) Appendices

1.3 Intended Audience

This document should serve as a guideline for software developers in building a robust, flexible and scalable architecture for PAP and its user interface. It should provide database analysts with adequate direction in devising solutions to store the annotation data generated by the pipeline. Finally, this document should give PATRIC curators and bioinformaticians a clear statement about the scope of PAP and the sequence analysis tools it will employ to generate the annotation data.

1.4 Document Conventions

1.4.1 Priority listing

The priority of each PAP requirement is labeled as [**H**] (high priority) or [**L**] (low priority). The high priority items must be included in the PAP version to be released as part of the Curation Infrastructure v1.0 as described in the Program Development Plan [1]. The low priority requirements may be included in a later version(s).

1.4.2 Terminology

A glossary has been included in Appendix A for the convenience of the reader. Terms that are included in the glossary appear in bold text the first time they are used.

1.5 Biological Sequence Annotation

The recent advances in DNA sequencing technologies have enabled the sequencing of the entire genomes of living organisms on an industrial scale [2]. This has resulted in a wealth of whole-genome sequence data, with the generation of new data likely to grow at an even faster pace in the foreseeable future. However, to be useful for scientific purposes biological information needs to be extracted from these raw genomic sequences. Annotation may be defined as the process of inferring structural and functional information from biological sequence data. As relying solely on “wet-lab” experiments to glean information from genomic sequences is prohibitively time-consuming and expensive, numerous computational techniques have been developed to facilitate the process of annotating biological sequences. These analytical procedures exploit *ab initio* prediction methods or sequence similarities, to previously characterized biological entities, to identify biological **features** and **attributes** [3-5].

Given the vast amounts of data that need to be analyzed for even a single organism and the repetitive nature of some aspects of the annotation process, large-scale genome annotation projects rely on biological data analysis pipelines that automate the annotation of genome sequences and the storage of the data generated. Although *in silico* methods have speeded up the annotation process, they frequently produce errors. Thus most automated sequence analyses are followed by the manual inspection and editing of the results by subject matter experts.

The end product of this annotation process- biological information inferred from genomic sequences, has become a powerful tool in the hands of researchers who use it to address diverse issues, ranging from evolution to the design of drugs, vaccines and diagnostics.

1.6 A System Overview of PAP

PAP is a subsystem of GAP and is designed to automate the annotation of the proteins encoded in the genome sequences to be analyzed by PATRIC. (see Appendix D, Figures 1 and 2). It comprises of a set of locally installed computational tools (and, where applicable, their associated databases) that generate protein sequence annotations; a database to store the data generated and a user interface. Input data is fed into PAP from a second sub-system of GAP, the Genome Sequence Analysis Pipeline (GSAP) which has been designed to identify and annotate nucleic-acid level features [6]. The protein annotations generated by PAP’s programs are manually reviewed and edited via the user interface. The annotated proteins serve as input data for two downstream sub-systems of GAP: the Metabolic Pathway tool, designed to identify the metabolic pathways the annotated microbial proteins belong to; and the Comparative Genomics Pipeline, designed to perform various comparative analyses with the annotated genes and proteins.

1.7 Users—Roles and responsibilities

There will be three main classes of users of the GAP infrastructure and its sub-components.

- **Curators:** Curators will be BS, MS or PhD-level biologists who will use their scientific training to review and edit the annotation data generated by GAP. Curators may work on-site at VBI or at off-site locations.
- **Administrators:** Administrators will be the supervisors, project managers and the PATRIC principle investigators who will manage the progress of the annotation process at PATRIC and evaluate the work of the curators.
- **Organism Experts:** Organism Experts will be scientists with several years of research experience and who are leading authorities in their fields. They will provide feedback, suggestions and expert guidance on the annotation process. All PATRIC organism experts are located at other research institutions and will be off-site users of the annotation infrastructure.

2. Functional Requirements

2.1 Scope

PAP is composed of computational tools that identify or predict the features and attributes of proteins encoded by the microbial genomes archived at PATRIC. For each protein PAP identifies and stores:

- 1) a set of physical properties: a) the theoretical molecular weight b) the theoretical isoelectric point c) the codon usage pattern and the d) hydropathy profile.

- 2) the location and types of **domain** and **motif** signatures and functional sites in the protein. Proteins sharing a common ancestor and having similar functions tend to have conserved domains and motifs.
- 3) the predicted secondary structure of the protein.
- 4) the predicted transmembrane domains of the protein.
- 5) the predicted signal peptide sequences in bacterial proteins. Two types of signal sequences will be predicted in bacterial proteins: those cleaved by SpaseI and SpaseII proteases.
- 6) the putative homologues archived in non-redundant protein database of NCBI.
- 7) the **COG** assignment of bacterial proteins and their functional categories.
- 8) the **Gene Ontology** assignment of the protein.
- 9) the Enzyme Commission number of the protein, if it is an enzyme.
- 10) the Transport Commission number of the protein, if it is a transporter.
- 11) the regions of synteny for the bacterial protein-encoding genes.

PAP, in its present version, does not identify the tertiary structure of proteins. PAP does not identify any protein functions; the curator will manually inspect the PAP data to infer protein function. In the case of viral proteins, a polyprotein is often processed into mature proteins. PAP does not identify the mature proteins. Curators will have to review the polyprotein annotation to annotate the mature proteins encoded by the polyprotein. Finally, PAP does not assign the proteins to known metabolic pathways or perform any comparative genomics with the protein sequences. These requirements will be set at a later date.

2.2 PAP workflow

2.2.1 Input data

For a given microbial genome sequence GSAP identifies a list of predicted gene intervals (PGIs) containing one gene each. A subset of these PGIs encode proteins. PAP takes as its input data the amino acid sequence of each of these encoded proteins.

A PAP run is triggered if :

- (A) GSAP finishes processing a microbial genome sequence and stores the PGI information in the PATRIC database [6]. PAP will sequentially retrieve the amino acid sequences of each encoded protein and start the execution of its pipeline tools.
- (B) A curator edits the PGI data generated by GSAP and sends the corrected PGI encoded protein to PAP.
- (C) A curator reviews the annotation data generated by a previous PAP run and decides to re-run PAP. In this case, PAP fetches and processes the amino acid sequence of the chosen protein.

2.2.2 Processing

The following is the complete list of PAP programs and the types of biological information they infer from the amino acid sequence of a protein:

Program	Features & Attributes	Priority
BioPerl modules	Molecular Weight; Isoelectric point	[H]
Codon Usage script	Codon usage pattern	[H]
Hydropathy script	Hydropathy profile	[H]
InterProScan	Domain and motif signatures; functional sites Transmembrane domains Signal peptide sequences (SPaseII cleaved sites) Gene Ontology assignments; E.C. number	[H]
BLOCKS	Motif signatures	[H]
PSIPRED	Secondary structure	[H]
MEMSTAT2	Transmembrane domains	[H]
LipoP	Signal peptide sequences (SPaseI cleavage sites)	[H]
BLAST	Homologues in the non-redundant NCBI database Homologues in the Saier's transporter database Transport Commission number	[H]
COGnitor	the COG assignment Functional category	[H]
STRING	Syntenic bacterial genes	[L]

2.2.3 Storage

PAP parses the output of each program and store the list of features and attributes along with a set of associated fields. Each feature and attribute will be stored with an associated unique id in the annotation database. PAP also proposes and stores a set of provisional features and attributes based on the pipeline results.

2.2.4 Manual Inspection

Curators inspect and edit the PAP-proposed list of features and attributes. Curators may also add new features/attributes and associated evidence to the PAP-generated annotation. All additions and edits are accompanied by an audit trail.

2.3 Performance Requirements

2.3.1 Runtime requirements

- (A) PATRIC has currently archived about 300 whole genome sequences of both viruses and bacteria. The anticipated volume of data that PAP should be able to process is *ca.* 24000 protein sequences, with increases likely to occur in the near future. [H]
- (B) PAP should support the workflow mentioned in section 2.2. [H]
- (C) PAP should be notified when GSAP finishes the nucleotide-level annotation of a microbial genome sequence. PAP should then fetch the amino acid sequence of each encoded protein and submit it, in the FASTA format, to each of its individual programs. A single file should contain information on pipeline execution: the program names and versions and their run parameters. The run parameters of each program have yet to be determined. They will be supplied at a later date. [H]
- (D) There will be situations in which the PGI data will have to be edited by the curator. In such cases the curator should be able to run PAP from the Gene Edit Page, the user interface for GSAP. [L]
- (E) We anticipate situations where the curator inspects the output of PAP and has to reanalyze the amino acid sequence of the protein. Provisions should be made for an *ad hoc* run of all or a subset of PAP programs, with a curator-specified set of parameters. [L]

(F) PAP should have provisions to parse through the output of each program, store the required fields and propose a list of features and attributes based on the generated data. (see Appendix B, for the specific fields to be stored from each program). [H]

2.3.2 Storage requirements

2.3.2.1 Storage of annotation data.

The PAP sequence analysis programs will generate annotation for the protein being processed. Scores and confidence levels will be associated with each of these biological inferences. Additional information will also be generated from each of the programs, e.g. an InterProScan output has a family or domain identifier associated with each predicted domain. Furthermore, while some features may span several amino acid residues, other features may be associated with individual amino acid residues, e.g., the secondary structure of a protein. PAP will use the output of its programs to propose a list of features and attributes.

The following are the storage requirements for the annotation data:

(A) All data generated by the PAP programs should be uneditable. However, the PAP-proposed features and attributes may be editable [H].

(B) Provisions should be made to store the output fields specified in Appendix B. [H]

(C) With each program output PAP should also store the run parameters and the date of execution. [H]

(D) The database should be able to store both multi-residue and single-residue features and attributes. [H]

(E) A protein will have multiple putative features and attributes. Each feature/attribute should be stored with an associated unique identifier and be traceable to its parent protein. [H]

2.3.2.2 Provenance

As explained in section 2.3.2.1, PAP presents the curator with a list of proposed features and attributes based on the program outputs. The provenance of each feature/attribute is the sequence analysis result(s) used as evidence to support the feature/attribute.

Requirements for the storage of feature/attribute provenance are:

(A) PAP should store the links between the proposed feature/attribute and the program, along with its output fields and run-date, that was used to generate the evidence. [H]

(B) A curator may add a new feature based on literature data or data from a sequence analysis program external to PAP. PAP should allow for the storage of these curator-added features and the relevant linked evidence. [H]

2.3.2.3 Audit and Roll-back

(A) The annotation generated by PAP will be regularly reviewed by a team of curators. PAP should have provisions to keep an audit trail for each edit made to the PAP data. This allows a history of each edit to be stored in the database and roll back to a previous version if necessary. Any feature edit or addition should be accompanied with the curator name, date and time of entry and the versions of the various programs and databases in PAP at that time. [H]

(B) The user interface should provide a mechanism to roll-back specific edits [H] or the work of a curator [L] of the manual editing done on a particular day [L].

2.4 The Protein Edit Page: the user interface to PAP

The automated sequence analysis results generated by PAP will have to be submitted to curators for review and inspection. The Protein Edit Page (PEP) will serve as the user interface to the PAP data stored in the PATRIC database. PEP should allow a curator (a) to view the annotation results and make edits if required and (b) to add additional features and attributes based on evidence generated outside the pipeline.

2.4.1 PEP: Interface layout

The Protein Edit Page will display the PAP-generated data and the PAP-proposed features and attributes in both graphic and tabular forms (see Appendix D Figure 3). PEP will be divided into five panels: (a) navigation bar (b) graphic view panel (c) evidence tables (d) similarity table and (e) curated features table.

2.4.1.1 Navigation Bar

The Navigation Bar will contain the name of the organism the protein belongs to and a navigable link to the gene encoding the protein and to the genome overview of the organism. It will display the name and user class of the current user. It will also display such basic information as: location of the gene, encoding the protein, in the genome sequence, protein length, molecular weight, isoelectric point and a PAP-generated unique identifier and a set of hyperlinks to the amino acid sequence and the coding sequence of the protein in the FASTA format.

2.4.1.2 Graphic view panel

The graphic view panel will display the amino acid sequence of the protein and the predicted features. Each feature will be displayed as a coloured bar; feature types will be colour-coded. The graphic view will also display a user-specified set of BLAST alignments and the hydropathy profile. A link, “Genome View” will provide a genomic view of the gene encoding the protein, and its neighbours. The user will be able to zoom and scroll the graphic view panel.

2.4.1.3 Evidence tables

The evidence tables will display in tabular format the annotation data generated by all the PAP programs. These data will be treated as evidence for the provisional features proposed by PAP. A curator may also add data generated outside PAP, to the evidence table. However, no PAP-generated data will be editable. The evidence table will have the following fields:

Column no.	Column Name	Contents
1	Color	Index number and color. Feature types will be colour-coded and will match the feature graphic colour in the Graphic View panel. The number is used as an identifier in the Provenance column in the Curated Features table to indicate the evidence on which the curated features are based. [H]
2	Feature	Feature type and name. [H]
3	Method	The PAP program used to generate the data. For a feature inferred from the literature, the experimental technique used to identify or characterize the feature [H]
4	PMID	PubMed Identifier for paper reporting the feature. For PAP-generated data, a reference for the program. [H]
5	Evidence	Genome Ontology{Harris, 2004 #15} Evidence Code{Butler, 2000 #37}. Computationally predicted features are coded as “IEA” – inferred from electronic annotation. Experimentally determined features have codes reflecting the methodology used, e.g. “IPI”—inferred from physical interaction. [H]
6	Start	Feature Start. The position of the first amino-acid residue. [H]
7	End	Feature End. The position of the last amino-acid residue. [H]
8	Owner	The name of the curator who entered the data. PAP-generated data are owned by PAP. [H]
9	Date	Date of creation. [H]
10	Feature/Attribute ID	Unique feature/attribute identifier [H]
11	Details	Link to detailed view of feature. This view typically displays all the fields of the output stored by PAP for that feature. [H]

2.4.1.4 Similarity table

The similarity table will display the BLAST alignments in a tabular format.

The similarity table will have the following fields:

Column no.	Column name	Contents
------------	-------------	----------

1	Database	The name of the database the protein sequence was queried against. [H]
2	Subject ID	Unique identifier associated with the subject sequence. [H]
3	Query Start	The start of the query sequence that aligns against the subject sequence. [H]
4	Query End	The end of the query sequence that aligns against the subject sequence. [H]
5	Subject Start	The start of the subject sequence that aligns against the query sequence. [H]
6	Subject End	The end of the subject sequence that aligns against the query sequence. [H]
7	Alignment length	The length of the match reported by BLAST [H]
8	Identity	Percent identity reported by BLAST [H]
9	Positives	Percent of positive hits [H]
10	Score	The normalized BLAST score. [H]
11	E-value	[H]

The similarity table will also have links to the complete BLAST report that displays significant alignments above a pre-set threshold.

2.4.1.5 Curated Features table

The curated features table will initially display an editable list of PAP-proposed features and attributes. These PAP output (displayed in the evidence table) used as evidence to generate the features and attributes will be specified in the Provenance field. A curator may also enter new features and attributes in this table. Each feature has an associated status. The following are the status values a curated feature may have:

Status	Meaning
Final	Feature has been entered by curator and approved by supervisor
Pending	Entered by curator, awaiting approval
Interim	Entered by curator on a temporary basis, to be upgraded to “Pending” at a later date
Provisional	Feature proposed by PAP and transcribed to Curated Features table to facilitate curation process. The curator can then approve them if accurate, edit if necessary or delete.

2.4.2 PEP: User interaction

A curator may use PEP to review the data generated by PAP and the PAP-proposed features and attributes. Based on the data presented in the evidence table, the curator may edit or delete the features proposed by PAP. S/he may also add new features and attributes to the existing annotation. A pull-down menu will give the curator a choice of feature/attribute type that may be added. The curator will have to enter evidence for the new feature in the evidence table before entering the feature in the curated features table. After the review and edit process the curator may change the status of the feature to pending or final.

2.5 Other Requirements

2.5.1 User access, roles and privileges

(A) Password protected accounts should be created for all allowed users of the annotation infrastructure. Provisions should be made for off-site personnel- curators and organism experts, to remotely log into GAP. [H]

(B) Administrators will measure the progress of the annotation project and the work of the curators. The PAP user interface should display summary statistics on each genome sequence and individual curator to allow the administrators to make evaluations. [H]

(C) The annotation editing privileges will differ among the user classes. These edit privilege requirements will be set at a later date. [L]

2.5.2 Updates of PAP tool-associated databases

The databases associated with the PAP programs will be periodically updated by their parent organizations to reflect additions and changes to their stored data. PAP should keep track of these database updates and update their locally installed versions. A database update should start the execution of the associated PAP

program with all the GSAP-identified protein sequences. The user interface should track changes in evidence fields for previously curated proteins. Means should be implemented to indicate to the curator a level of significance for the changes as they relate to the existing annotation. [L]

Appendix A. Glossary

Gene-an ordered sequence of nucleotides located in a defined position on a chromosome that encodes a specific functional product (i.e., a protein or RNA molecule).

Protein-a large molecule composed of one or more chains of ordered amino acid sequences that have been determined by the sequence of nucleotides in the gene coding for the protein.

Domain- Used to describe a discrete part of a protein that shares common physicochemical features, eg. hydrophobic, polar, globular, α -helical domains, or properties eg. DNA-binding domain, ATP-binding domain.

Motif-A small structural element that is recognizable in several proteins, usually refers to a smaller discrete region than a domain.

Secondary structure-The folded, coiled or twisted shape a nucleotide or amino acid chain takes on when hydrogen bonds form between adjacent parts of the molecule.

Feature-the characteristics of a protein sequence (sequence, domains, motifs) which are used as evidence in determining the functional identity of a protein.

Attribute- An annotation, property, characteristic or function that applies to the sequence as a whole. An attribute could be the molecular weight of a protein or its functional GO assignment. An Enzyme Commission number is an attribute because it applies to the whole sequence.

Gene Ontology- a controlled vocabulary used to categorize gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner

COGs- (Clusters of Orthologous Groups) Conserved families of proteins from completely sequenced genomes. Each COG contains orthologous sets of proteins from at least three phylogenetic lineages, which are assumed to have evolved from an individual ancestral protein. Orthologs, by definition, are genes which are connected through vertical evolutionary descent (the “same” gene in different species). Since orthologs typically perform the same function across organisms, delineation of families of orthologs from diverse species can aid in the functional annotation in newly-sequenced genomes.

Appendix B. Specific storage requirements for each PAP program

PROGRAM	FIELDS TO BE STORED
BioPerl Module	Molecular weight and isoelectric point
Codon Usage script	Codon usage table
Hydropathy script	Mean hydrophobicity of residue
InterProScan	InterProScan entry type and name Parent of entry Children of entry Found In Contains Database/Application name Domain ID in database Domain name in database Start of domain in query End of domain in query E-value GO IDs and associated functions
BLOCKS	BLOCK family ID Number of BLOCKS in query protein Number of BLOCKS in family Global E value Family Description Individual BLOCK IDs Start in query protein End in query protein Local E value
PSIPRED	residue structure PSIPRED confidence score
MEMSAT2	residue topology MEMSAT2 confidence score
LipoP	Prediction Class Location Score Sequence Position+2
BLAST	Database name Subject Accession number Subject name Score E value Identities Positives Transport Commission number (for Saier's database) Alignments
COGnitor	Functional Category code COG assignment Assigned function Number of clades hit COG score Alignments
STRING	To be determined

Appendix C. Future Requirements.

1) Annotating Reference Genomes and Associated Genomes

The genomes of each Pathosystem will be divided into a set of one or more Reference Genomes and a set of Associated Genomes. The Reference Genomes will undergo a thorough annotation while genes in the Associated Genomes will inherit orthology-based annotations from their Reference Genome counterparts. PAP should include tools to determine orthologs between a given Reference Genome and its Associated Genome and transfer relevant information to the Associated Genome gene.

2) Additional tools in PAP

We expect future versions of PAP to have additional capabilities. These may include additional sequence analysis tools or the display of literature data associated with each curated protein.

3) Interaction of the PAP with other downstream subsystems.

As stated in Section 1.6 the “cleaned-up” version of the PAP generated data will be fed into downstream systems for other protein sequence-based analyses. The PAP-associated database design should allow for the easy access of these data by the downstream systems.

Appendix D. Flowcharts & Diagrams

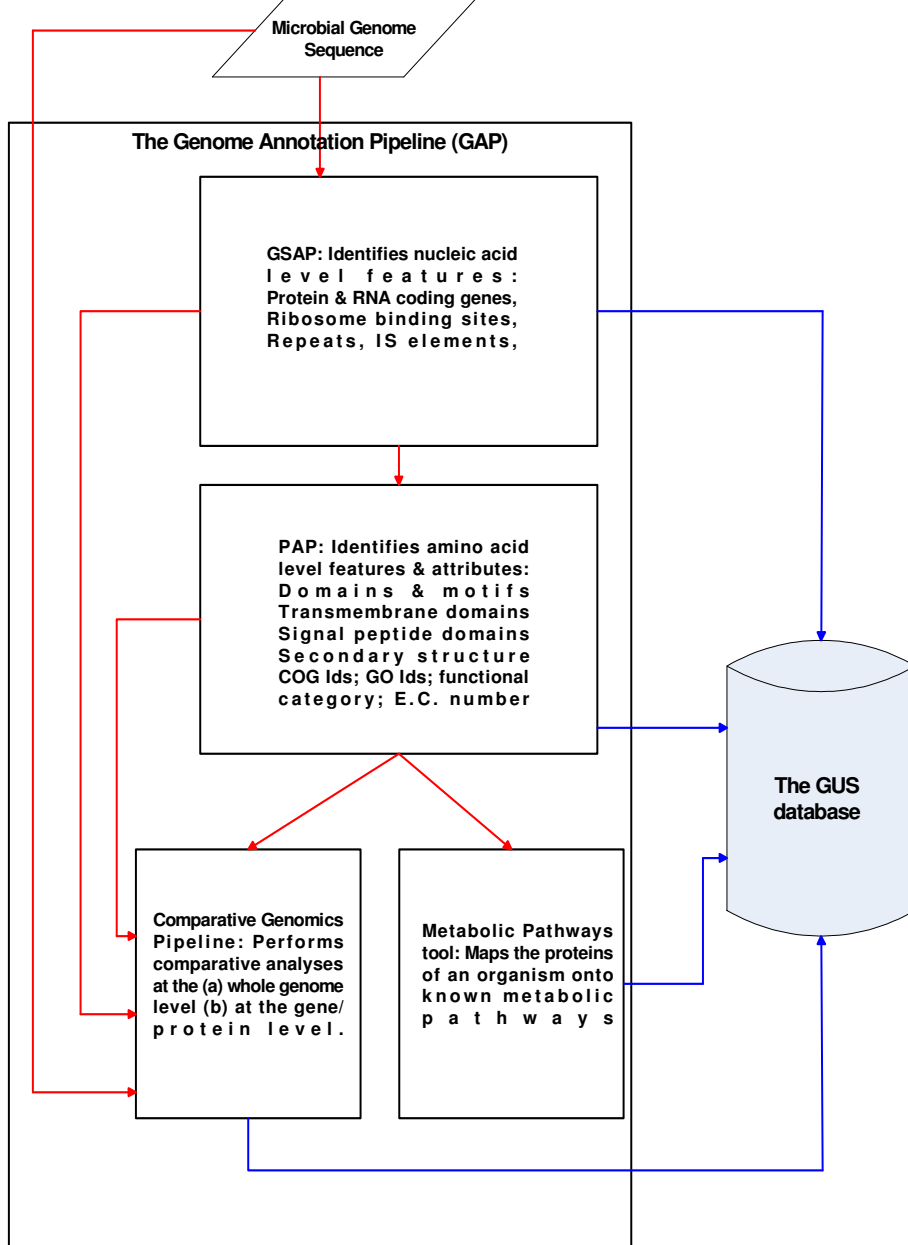


Figure 1. The Genome Annotation Pipeline. The Genome Annotation Pipeline (GAP) is the annotation infrastructure that automates the annotation of genome sequences archived at PATRIC. GAP is composed of four subcomponents: the Genome Sequence Analysis Pipeline (GSAP); the Protein Analysis Pipeline (PAP); the Metabolic Pathways tool; and the Comparative Genomics Pipeline. Red Arrows indicate the data flow into GAP and among its subcomponents; blue arrows indicate the flow of the data generated by each component into a central database that follows GUS (Genomics Unified Schema).

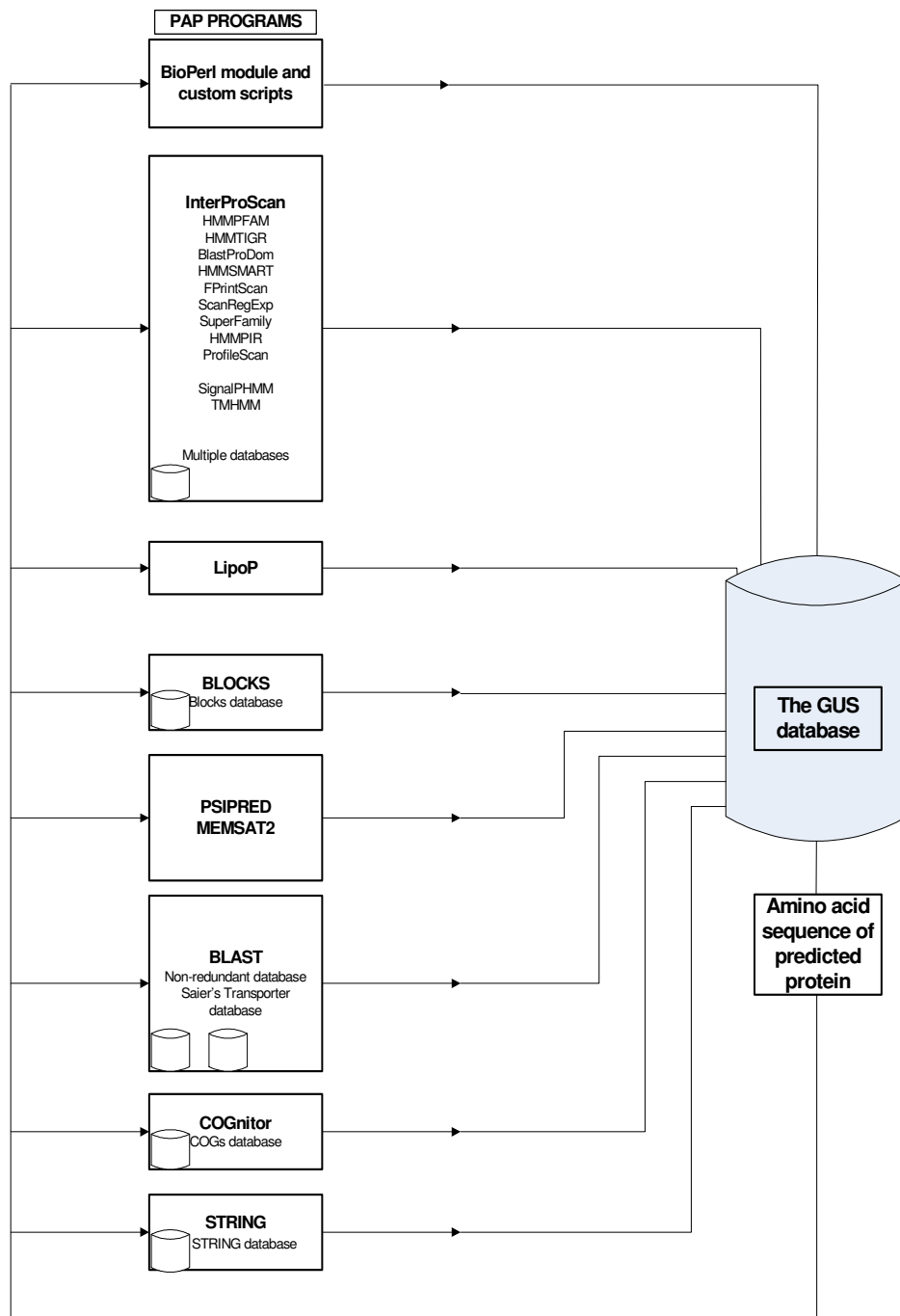


Fig. 2. The Protein Annotation Pipeline: Its components. The Protein Annotation Pipeline (PAP) is composed of eight subcomponents. (1) BioPerl and custom scripts calculate a set of physical properties of the protein; (2) InterProScan is composed of 11 domain/motif prediction programs. It has multiple databases associated with it [7]; (3) LipoP is a signal peptide identification program [8]; (4) BLOCKS is a motif identification program [9]. It is associated with the BLOCKS database; (5) PSIPRED/MEMSAT2 predict the secondary structure of a protein and its putative transmembrane domains [10]; (6) BLAST identifies homologues in the two underlying databases: the NCBI non-redundant database and the Saier's transporter database; (7) COGnitor identifies orthologous assignments [11] and (8) STRING identifies bacterial gene synteny [12].

Protein Curation Page


[Curation Summary](#) > [Brucella melitensis Genome Overview](#) > [GeneIDNO](#) > [ProteinIDNO](#) LOG OUT

Logged in as :
User class:

[Amino acid sequence](#)
[Nucleotide sequence](#)

Protein Sequence View + - □ ?

Genome View Display top 25 alignments Minimum E-value, 10e-9 Zoom:



Kyte-Doolittle Hydrophobicity Plot Sliding Window Width 90

EVIDENCE TABLES

FEATURE STATISTICS:
[Codon Usage Table](#)

DOMAINS/MOTIFS/SIGNATURES

Color	Method	PMID	Feature	Evidence	Start	End	Owner	Date	FeatureID	Details
Yellow	InterProScan		domain: PF00204.9: DNA topoisomerase	XXX	343	450	PSAP	5/2/2005	p0123	view
Cyan	InterProScan		domain: PF00986.8: DNA gyrase	XXX	500	600	PSAP	5/2/2005	p234	view
Blue	InterProScan		signature: SM00433: DNA topoisomerase II	XXX	290	650	PSAP	5/2/2005	sm0239	view
Red	SignalP		Signal peptide: XYZ signal	XXX	1	12	PSAP	5/2/2005	sp1023	view
Purple	LipoP		Signal peptide: ABC signal	XXX	1	34	PSAP	5/2/2005	lp2020	view

SECONDARY STRUCTURES

Color	Method	PMID	Feature	Evidence	Start	End	Owner	Date	FeatureID	Details
Yellow	PSIPRED		helix	XXX	1	23				view
Green			coil	XXX	24	55				view
Orange			sheet	XXX	56	100				view
Red	TMHMM		inner domain	XXX	1	23				view
Cyan			TM domain	XXX	24	55				view
Blue			outer domain	XXX	56	199				view
Green			TM domain	XXX	200	250				view
Purple			inner domain	XXX	251	279				view

MATURE PEPTIDES

Color	Method	PMID	Feature	Evidence	Start	End	Owner	Date	FeatureID	Details
Blue	Manual		mature peptide	XXX	1	23				view
Green	Manual		mature peptide	XXX	24	55				view
Purple	Manual		mature peptide	XXX	56	100				view

ATTRIBUTES

Method	PMID	Attribute	Evidence	Owner	Date	AttributeID	Details
COG		COG0004					view
GO term		XYZ					view
E.C.no.		E.C.1.2.3.4					view

First then click

SIMILARITY TABLE

Color	Method	Subject	Qstart	Q End	Sub Start	Sub End	Aln Length	Identity	Score	Evalue	Owner	Date	Feature ID	Details
Yellow	BLASTP													view
Cyan	TBLASTN													view

CURATED FEATURES

PROTEIN DOMAIN FEATURES

Color	Feature	Provenance	Evidence	Start	End	Owner	Date	FeatureID	Status	Details
Green	DNA topoisomerase domain	1,4	IC	1234	34348	AP	6/20/2005	AB1234	FINAL	edit
Blue	DNA gyrase domain	3	IC	234	3838	AP	6/20/2005	AN1230	PENDING	edit
Yellow	signal peptide	5	IC	1	34	AP	6/20/2005	S1223	PROVISIONAL	edit

PROTEIN STRUCTURAL FEATURES

Color	Feature	Provenance	Evidence	Start	End	Owner	Date	FeatureID	Status	Details
Red	Helix	6	IC	1234	34348	AP	6/20/2005	H1234	PENDING	edit
Green	Coil	6	IC	234	3838	AP	6/20/2005	AN1230	PENDING	edit
Cyan	Sheet	6	IC	1	34	AP	6/20/2005	S1223	PENDING	edit

PROTEIN ATTRIBUTES

Attribute	Attribute Name	Provenance	Evidence	Owner	Date	FeatureID	Status	Details
COG assignment	COG0040203	8	IC	AP	6/20/2005	H1234	PENDING	edit
GO assignment	XYZ	9	IC	AP	6/20/2005	AN1230	FINAL	edit
E.C. number	E.C.1.2.3.4	10	IC	AP	6/20/2005	S1223	PENDING	edit

First then click

Navigation Bar

Graphic View Panel

Evidence tables

Similarity table

Curated Features table

Fig. 3. The Protein Edit Page: the user interface of PAP. PEP and its five panels are used to present to the curator the annotation data generated for a single protein sequence. The curator can edit the PAP-proposed features/attributes and can also add additional features/attributes.

References

1. Crasta, O.R., et al., *Program Development Plan for PATRIC*. 2004.
2. Mitnik, L., et al., *Recent advances in DNA sequencing by capillary and microdevice electrophoresis*. *Electrophoresis*, 2001. **22**(19): p. 4104-17.
3. Bull, A.T., A.C. Ward, and M. Goodfellow, *Search and discovery strategies for biotechnology: the paradigm shift*. *Microbiol Mol Biol Rev*, 2000. **64**(3): p. 573-606.
4. Whisstock, J.C. and A.M. Lesk, *Prediction of protein function from protein sequence and structure*. *Q Rev Biophys*, 2003. **36**(3): p. 307-40.
5. Gabaldon, T. and M.A. Huynen, *Prediction of protein function and pathways in the genome era*. *Cell Mol Life Sci*, 2004. **61**(7-8): p. 930-44.
6. Snyder, E.E., et al., *PATRIC: Requirements for Genomic Sequence Curation*. 2005.
7. Zdobnov, E.M. and R. Apweiler, *InterProScan--an integration platform for the signature-recognition methods in InterPro*. *Bioinformatics*, 2001. **17**(9): p. 847-8.
8. Juncker, A.S., et al., *Prediction of lipoprotein signal peptides in Gram-negative bacteria*. *Protein Sci*, 2003. **12**(8): p. 1652-62.
9. Henikoff, J.G. and S. Henikoff, *Blocks database and its applications*. *Methods Enzymol*, 1996. **266**: p. 88-105.
10. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server*. *Bioinformatics*, 2000. **16**(4): p. 404-5.
11. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution*. *Nucleic Acids Res*, 2000. **28**(1): p. 33-6.
12. von Mering, C., et al., *STRING: a database of predicted functional associations between proteins*. *Nucleic Acids Res*, 2003. **31**(1): p. 258-61.

Table of Contents

1.	Introduction	1
1.1	Document Overview	1
1.2	Document Organization	1
1.3	Intended Audience	1
1.4	Document Conventions	1
1.4.1	Priority listing	1
1.4.2	Terminology	1
1.5	Biological Sequence Annotation	2
1.6	A System Overview of PAP	2
1.7	Users—Roles and responsibilities	2
2.	Functional Requirements	2
2.1	Scope	2
2.2	PAP workflow	3
2.2.1	Input data	3
2.2.2	Processing	3
2.2.3	Storage	4
2.2.4	Manual Inspection	4
2.3	Performance Requirements	4
2.3.1	Runtime requirements	4
2.3.2	Storage requirements	5
2.3.2.1	Storage of annotation data	5
2.3.2.2	Provenance	5
2.3.2.3	Audit and Roll-back	5
2.4	The Protein Edit Page: the user interface to PAP	5
2.4.3	PEP: Interface layout	6
2.4.3.1	Navigation Bar	6
2.4.3.2	Graphic view panel	6
2.4.3.3	Evidence tables	6
2.4.3.4	Similarity table	6
2.4.3.5	Curated Features table	7
2.4.4	PEP: User interaction	7
2.5	Other Requirements	7
2.5.1	User access, roles and privileges	7
2.5.2	Updates of PAP tool-associated databases	7
Appendix A.	Glossary	8
Appendix B.	Specific storage requirements for each PAP program	9
Appendix C.	Future Requirements.	10
Appendix D.	Flowcharts & Diagrams	11
References		14