

# BPAP : A Computational Tool for Whole Genome Analysis and Annotation

Barrett Abel and Martin Gollery  
Department of Biochemistry  
College of Agriculture, Biotechnology and Natural Resources  
University of Nevada Reno  
Reno, Nevada 89557  
Email: [babel@unr.edu](mailto:babel@unr.edu), [mgollery@unr.edu](mailto:mgollery@unr.edu)

## Abstract

*Since the genesis of the Human Genome Project, more than 100 genomes have been sequenced, and over 1000 genomes are expected to be sequenced in the next decade. With the advent of this extensive repertoire of raw sequence information, the next major challenge for a modern researcher is to interpret this wealth of information. At the remarkable rate of progress that is being made in sequencing organisms, experimental research techniques alone cannot keep up with the influx of genomic information. As a powerful complementary approach to experimental techniques, computational sequence analysis can assist in the categorization, characterization and creation of a functional definition of protein sequences.*

*Our efforts have produced a biological protein analysis pipeline that can characterize proteins and attempts to produce working hypothesis of the functional and characteristic nature of a protein in a high throughput, automated manner with a minimum of amount of operational complexity for the user.*

## 1. Introduction

BPAP is software tool with the functional task of providing annotation of biologically relevant information from a nucleotide or proteomic sequence. Its' goal is to provide a simple yet powerful interface for the analysis and mining of genomic information while seamlessly handling the nonscientific complexities of interfacing with hardware, computational clusters, software packages, raw data, and file formats.

The design of the BPAP software package (biological protein analysis pipeline) was engineered to be an 'extensible framework' that facilitates expansion with any bioinformatics software tool approaching point and click simplicity.

## 2. Method

We are utilizing a series of techniques for the prediction of specific characteristics of proteins. Each analysis package is treated as a modularized component that can be dynamically incorporated in BPAP to facilitate extensibility and allow updating as new methods are devised and advancements or revisions to new methods are developed.

### 2.1 Approach

Included in the initial release of BPAP, we have developed or incorporated existing algorithms for the analysis of the following qualities:

- Post-Translational Modifications
- Subcellular Localization
- Associated Homology
- Related Structural Molecules / Homologs
- Significant motif(s)
- Transmembrane regions/ helices
- Physiological / chemical characteristics
- Rudimentary Fold Patterns
- Family Identification
- Globularity
- Recognizable domains
- Secondary / Super-secondary Structure

Details as to the nature of algorithms, external resources, databases and software tools incorporated in the BPAP are listed as a web resource.

### 2.2 Implementation

BPAP is composed of three individual separate software components, namely - A graphical user interface, a processing engine, and a report engine.

The graphical user interface provides the control point that the user interacts with on a personal

computer. This component is developed to be run on any desktop computer (Mac OSX, Linux and Windows) using a multiplatform GUI toolkit for uniformity.

The processing engine is a listening inet / unix socket daemon that authenticates, manages the forking and execution of individual analyses or generation of shell scripts for the intelligent submission to a processing queue. Queues currently supported are Maui, Portable Batch System (OpenPBS) and the Sun Grid Engine (SGE).

The report engine combines resulting output to attempt to interpret make intelligent summaries along with presenting the raw analyses output in a failure format. Moreover, a user can specify and filter for specific protein characteristics, which enable a user to utilize BPAP as an automated search tool.

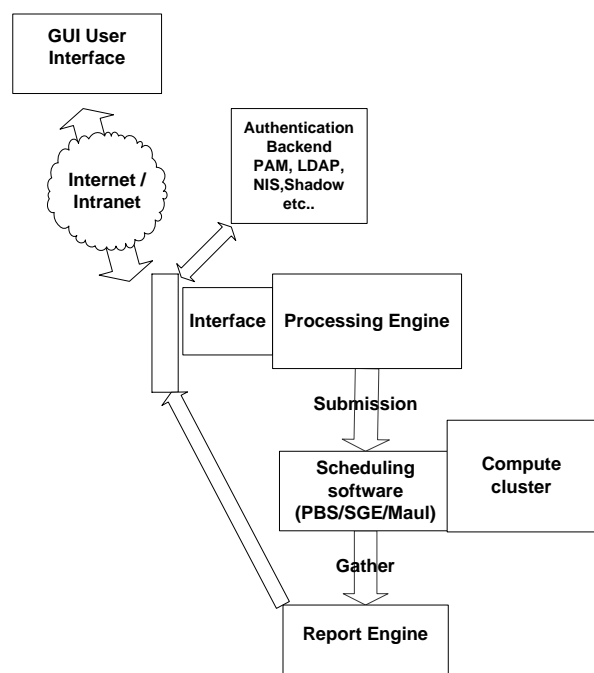


Figure 1. Operating model for BPAP

Figure 1 represents the typical high level interaction of the separate BPAP components with a user and computational hardware.

#### 4. Conclusion

We have produced the foundation of a powerful, easy to use, extensible software package that is able to make a working hypothesis of characteristics of a protein from pure protein sequence.

BPAP produces annotation results that are interpreted by computer and presented in a useful manner, more so than a conglomeration of separate analyses.

#### 5. Future Direction

It is our expectation that BPAP will be continually under development in an effort to improve functionality and become a more powerful tool in a bioinformaticists' toolbox. We use an incremental development model; development is responsive to comments, suggestion and requests in an iterative manner.

As an improvement in functionality, we are currently working on tools to allow web based submission and extraction of internet based bioinformatics tools and databases, as well as concurrent work on a web based interface to replace the functionality of the GUI module in a web environment.

The original development of BPAP was sparked by an interest in annotating the large number of hypothetical and unknown genes in the malaria proteome. When this is complete, BPAP will be expanded upon to help discover and classify late-embryogenesis abundant (LEA) proteins.

#### 6. Availability

BPAP stable binaries and complete source code will be available for Microsoft Windows, Mac OSX and Linux late summer, and will be released under standard GPL terms. Releases will be posted at <http://bioinformatics.unr.edu/BPAP> or by emailing the authors directly.

For additional detailed information and references please observe the aforementioned web resource.

#### 7. Acknowledgements

Barrett Abel gratefully acknowledges NIH-BRIN support from the National Center for Research Infrastructure (NCRR), P20 RR16464.

Martin Gollery is supported by P20 RR16464 from the National Center for Research Infrastructure through the Biomedical Research Infrastructure Network (BRIN), NSF plan genome project DBI-0321690, and NCRR COBRE.