1. Purpose

This document describes the mechanism and process by which InterProScan tool is used to perform sequence analysis for bacterial and microsporidian protein sequences stored in BioHealthBase.

2. Method Description

InterProScan is a tool that combines different protein signature recognition methods into one resource. BioHealthBase InterProScan data production process utilizes both the stand-alone command-line tool and EBI web services. The results are merged and formatted for loading into BioHealthBase.

The databases used include PROSITE patterns, PROSITE profiles, PRINTS, PFAM, PRODOM, SMART, TIGRFAMs, PIR SuperFamily, SUPERFAMILY, GENE3D, PANTHER.

The methods applied include HMMPfam, HMMPanther, HMMPIR, blastprodom, coils, gene3d, HMMSmart, HMMTigr, FPRINTScan, scanregexp, profilescan, superfamily, seg, signalp, TMHMM.

Below is an example of Domain and Motif data from InterProScan as displayed in Gene Details page.

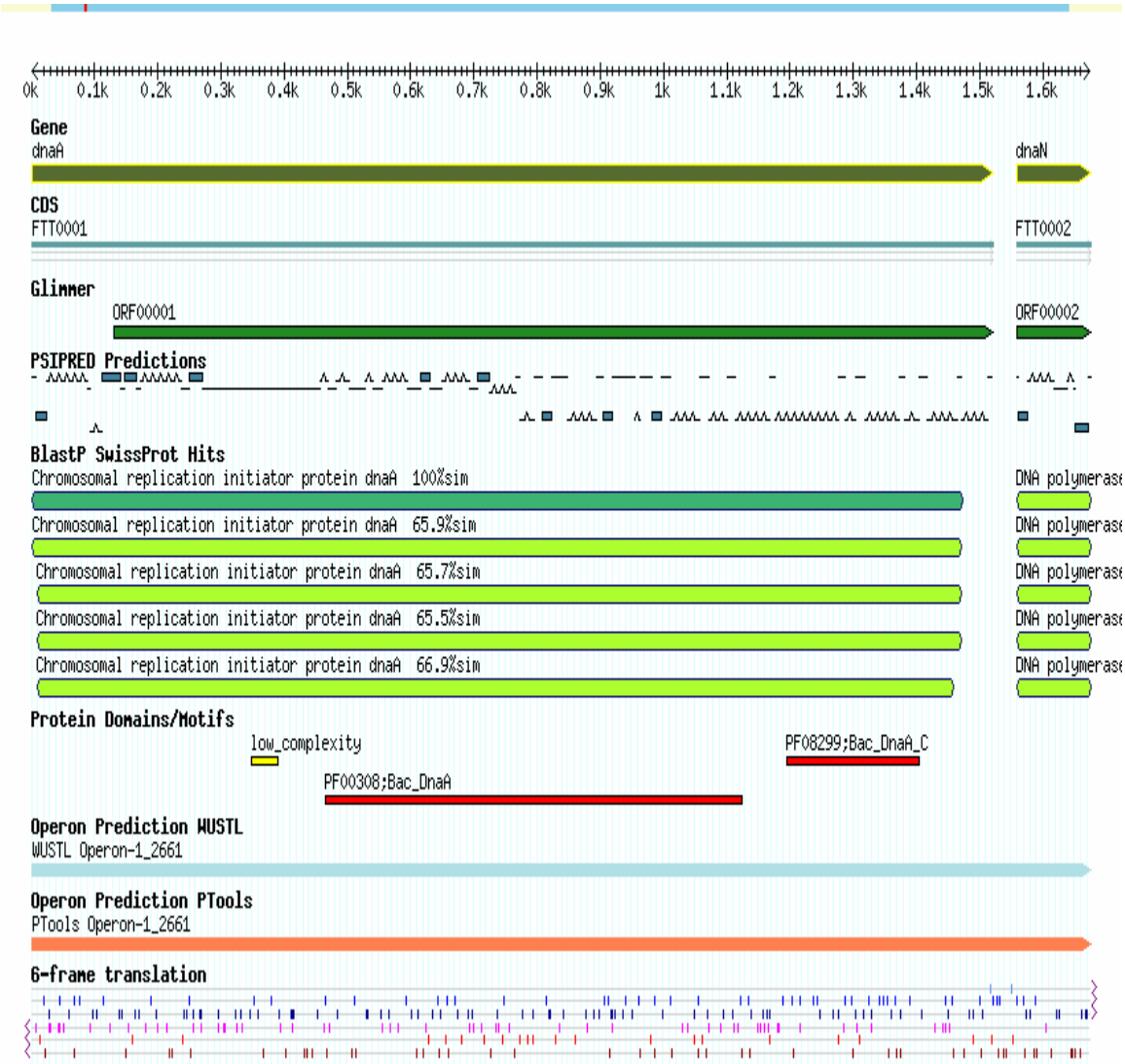
HMM/Pfam Domains Top

Accession	Name	Description	Start	End
PF00308	Bac_DnaA	Bacterial dnaA protein	156	374
PF08299	Bac_DnaA_C	Bacterial dnaA protein helix-turn-helix domain	399	468

Other Domains/Motifs Top

Domain/Motif	Start	End	Program
smart	188	314	smart
smart	399	468	smart
low_complexity	117	130	seg

Domain and Motif data from InterProScan are also displayed in the genome browser, grouped under “Protein Domains/Motifs” track, as shown in the following figure.
.



3. Input Data Preparation

The input data is a bacterial or microsporidian protein sequence in FASTA format.

4. Output Data Post-Processing and Display

The output data from InterProScan is reformatted for BioHealthBase loading. In addition, tracking information about the process is added to the final output.

5. References

1. **http://www.ebi.ac.uk/interpro/**

2. Zdobnov E.M. and Apweiler R. "InterProScan - an integration platform for the signature-recognition methods in InterPro." Bioinformatics, 2001, 17(9): 847-8.