



DESIGN & IMPLEMENTATION OF A PIPELINE FOR HIGH-THROUGHPUT ENZYME FUNCTION PREDICTION

Seth Johnson

November 21st, 2006

Department of Bioinformatics, GMU.

BHSAI, DoD.



- INTRODUCTION
 - ABOUT ENZYMES
 - Importance of Enzymes
 - Classification of Enzymes
 - CHALLENGES IN PROTEIN FUNCTION ANNOTATION
 - Lack of Unified Pipeline
- BACKGROUND & RELATED WORK
 - OVERVIEW OF EXISTING APPROACHES
 - Homology-Based Prediction
 - Other Sources of Function Classification
 - NEED FOR HIGH-THROUGHPUT (HT) ANNOTATION
 - Disparity Between Number of Genes & Their Annotation
 - NOVEL **HT** ENZYME FUNCTION PREDICTION PIPELINE
 - Overlapping Families
 - Enzyme Family Definition
 - Enzyme Function Prediction Sensitivity
 - Multiple Enzyme Function Prediction
 - Combined Sequence Homology & Functionally Discriminative Site Search



- **APPROACH**
 - **DATABASES**
 - The Gene Ontology (GO) Database
 - CSA Web Scraper Software
 - Incorporating Functionally Determining Sites & Active Sites
- **SOFTWARE IMPLEMENTATION**
 - DETAILS OF HIGH PERFORMANCE COMPUTING
 - SERIAL & PARALLEL EXECUTION TIME
 - GENERATED DATABASES
- **VALIDATION**
 - ENZYME VALIDATION SET
 - RESULTS
 - 3-digit EC Families
 - Discussion of 3-digit Results
 - 4-digit EC Families
 - Discussion of 4-digit Results
- **CONCLUSIONS & FUTURE WORK**
 - COMBINATION WITH OTHER TOOLS
 - COMBINATION WITH PATHWAYS



INTRODUCTION

General Information About Subject



Importance of Enzymes

- Enzyme is an Organic Catalyst
 - Speed up Chemical Reactions
 - May Lead to Different Products
- Exhibit Much Higher Level of Specificity
 - Reactions
 - Substrates
- Statistics About Enzymes
 - 4,000 Reactions Catalyzed
 - 30% of Eukaryotic Genome



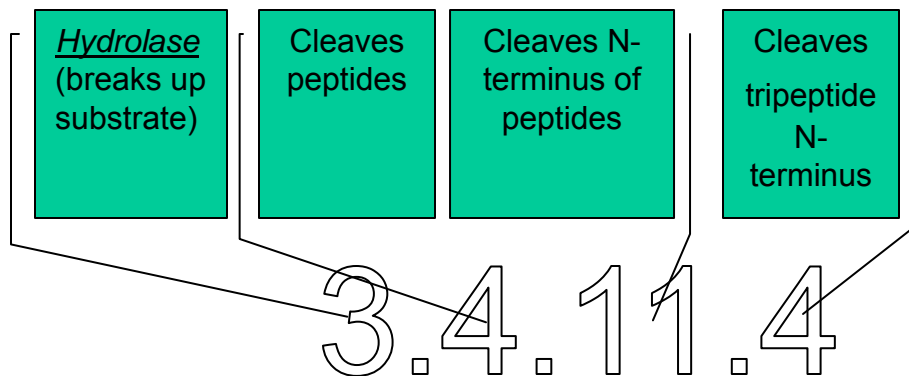
Classification of Enzymes

Named According to Catalyzed Reaction

Suffix -ase added

- **Substrate**
 - lactase
- **Type of Reaction**
 - DNA Polymerase
- **EC Classification**
 - Sometimes Cannot Distinguish One Enzyme from Another
 - Represent Progressively Finer Classification

Group	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
EC 1 Oxidoreductases	To catalyze oxidation /reduction reactions; transfer of H and O atoms or electrons from one substance to another	$AH + B \rightarrow A + BH$ (reduced) $A + O \rightarrow AO$ (oxidized)	Dehydrogenase , oxidase
EC 2 Transferases	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	$AB + C \rightarrow A + BC$	Transaminase , kinase
EC 3 Hydrolases	Formation of two products from a substrate by hydrolysis	$AB + H_2O \rightarrow AOH + BH$	Lipase , amylase , peptidase
EC 4 Lyases	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	$RCO_2COOH \rightarrow RCO_2H + CO_2$	
EC 5 Isomerases	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	$AB \rightarrow BA$	Isomerase , mutase
EC 6 Ligases	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	$X + Y + ATP \rightarrow XY + ADP + Pi$	Synthetase





Challenges in Protein Function Annotation

- Rapid Advances in Sequencing
 - Inability to Determine Function Experimentally
- Complex and Vague Relationships
 - Different Avenues of Approach
- Erroneous Data
 - Need Correction
 - Manual Curation
- Software Inadequacies
 - Modification of Existing Algorithms
 - Integration of Tools into Packages
 - Maintenance
 - Adaptation for High Performance

Need For High Throughput Function Prediction

DoD BHSI System:

1. Homology Search Method
 - High Sensitivity
2. Enzyme Profile Database (EPD)
 - RPS-Blast Based
3. Functionally Determining Sites
 - CSA
 - Computed from EPD



Background & Related Work

Current State of Technology & Science



Overview of Existing Approaches

Homology Based Prediction

Evolutionary Related Proteins Share Function

1. Sequence Based Methods
 - EFICAz
 - PRIAM
2. Structure Based Methods
 - CSA
 - SFLD

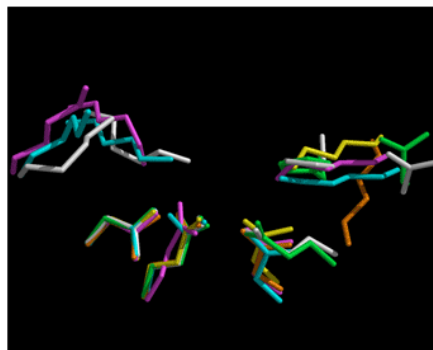
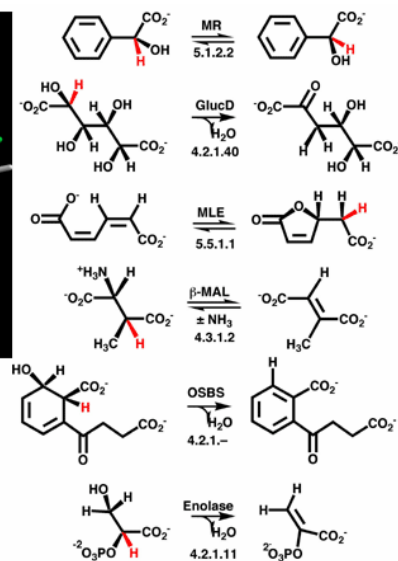


Figure 1. **Above:** Superimposition of active sites of some divergently related members of the ES. **Right:** Some chemical reactions performed by divergent members of the ES. The proton abstracted to initiate each reaction is shown in red.





Other Source of Function Classification

- Gene Ontology (GO)
 1. Ontology
 1. Molecular Function
 2. Role in Biological Processes
 3. Localization
 2. Annotation
- Clusters of Orthologous Groups (COG)

Framework for Functional and Evolutionary Analysis

 - 66 Complete Genomes
 - 14 Phylogenetic Lineages
 - 4872 Clusters



Need for High Throughput Annotation

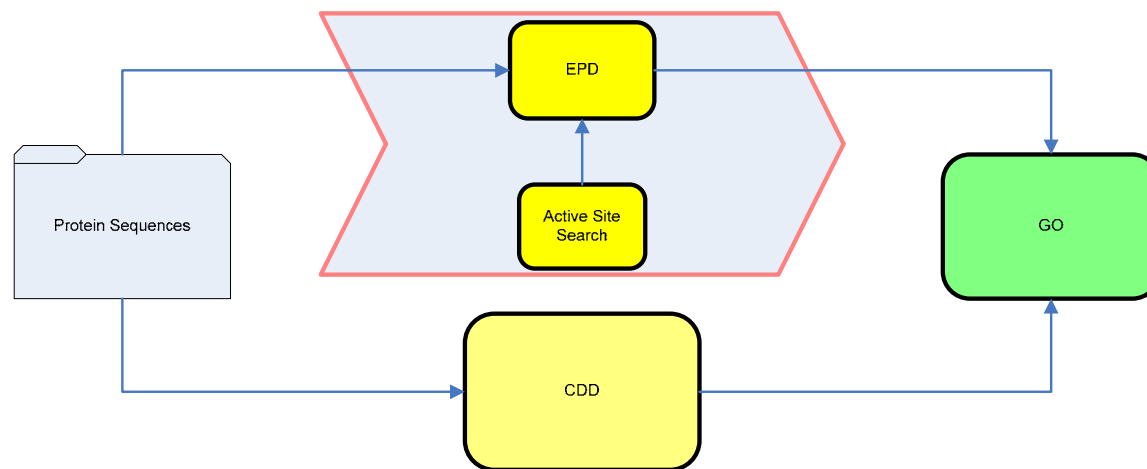
Widening Gap Between Sequencing & Annotation

- Sequencing
 - 96 Capillaries
 - 2,000,000 bases a day sequenced
- Annotation
 - Structure Based Methods
 - (-) 3-D Structure using X-ray Crystallography
 - (+) Protein Structure Initiative (PSI)
 - Sequence Based Methods
 - (+) On-the-fly Solution Coupled with Sequencing
 - (+) Bypasses Expensive Purification & Crystallization
 - (-) Enrichment from Protein Interaction, Expression, etc.



Novel HT Enzyme Function Prediction Pipeline

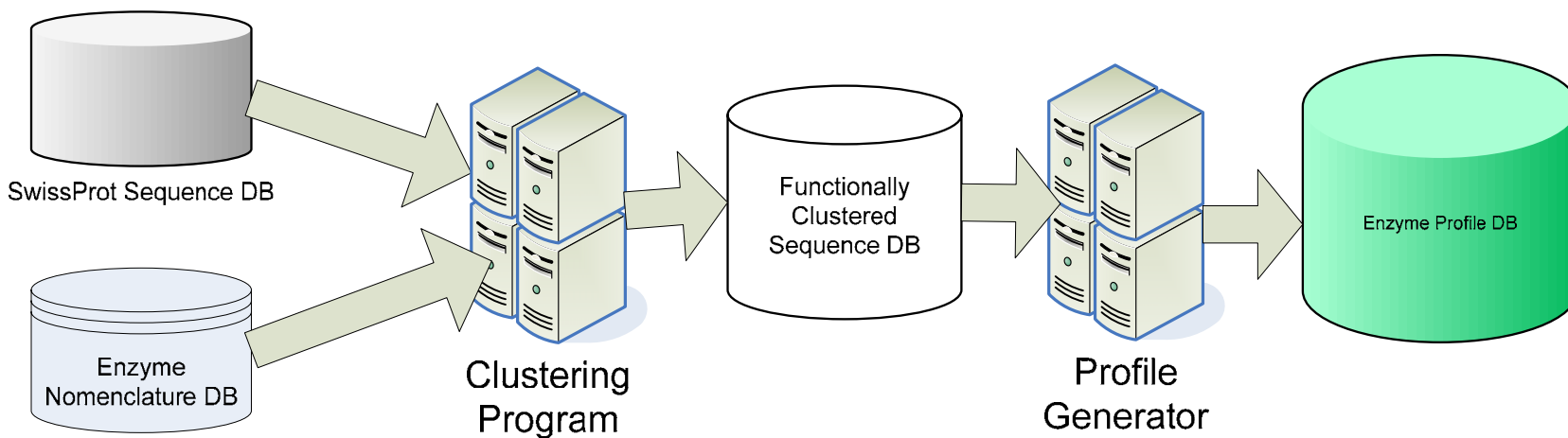
- CDD
 - Expands Scope to Other Functions
 - Contains SMART, Pfam, COG
 - Groups Related by Common Descent
- RPS-Blast
 - Position-Specific Score Matrix
 - Flexible Display Option
 - Pairwise Alignments
 - Multiple Alignments





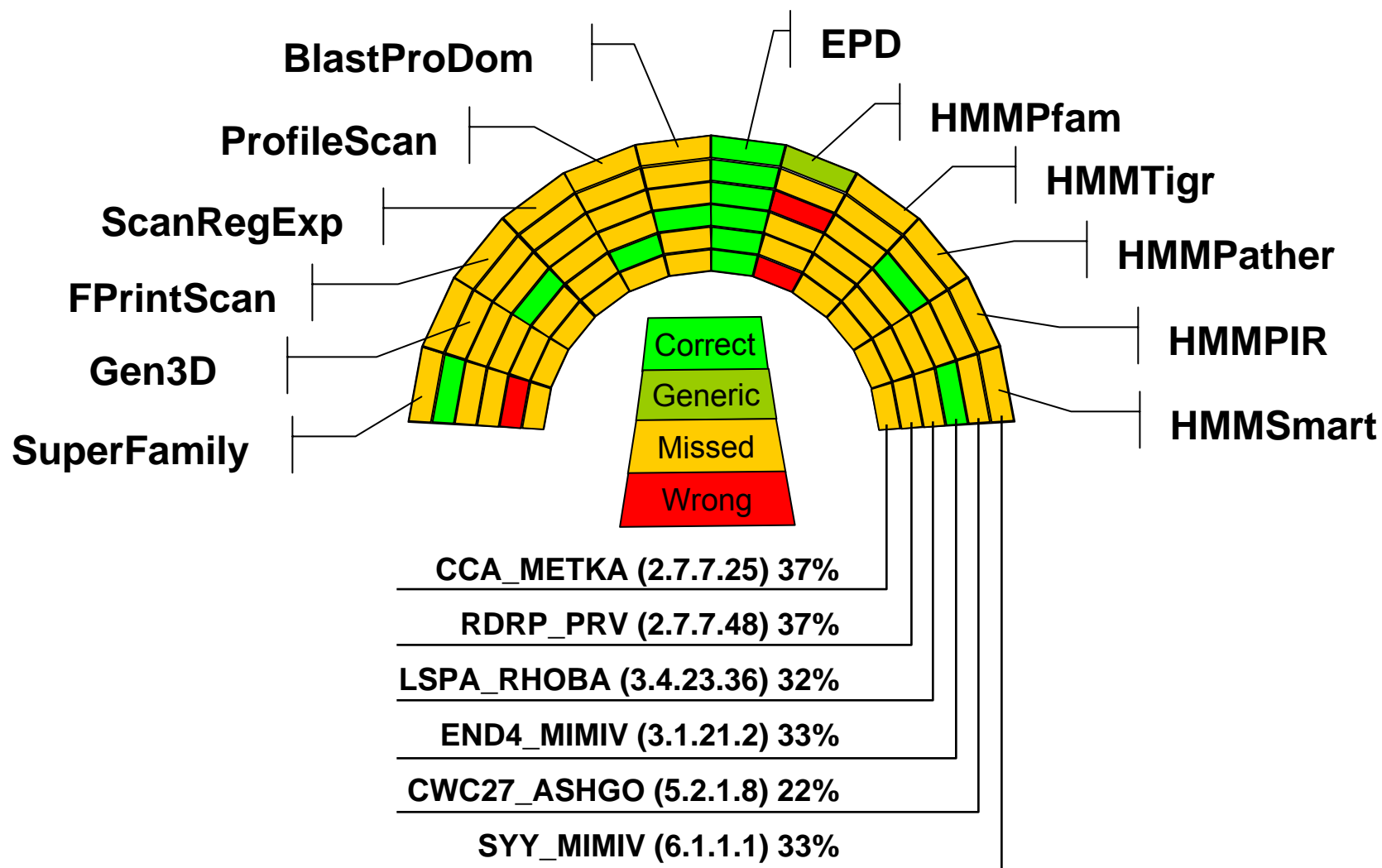
Novel HT Enzyme Function Prediction Pipeline

- Multiple Enzyme Function Prediction
 - Based on ENZYME rel. 37
 - Allow Prediction of Nearly All 3- and 4-digit EC Numbers
 - 190 3-digit Families
 - 1,747 4-digit Families
 - No Predictions for Families with < 3 Members





Novel HT Enzyme Function Prediction Pipeline





Novel HT Enzyme Function Prediction Pipeline

Combined Sequence Homology & FDS

- Improvement Upon Pure Homology Methods
- 3-digit Families
 - Functionally Discriminative Sites (FDS)
 - Conserved Sites in Multiple Sequence Alignment of 3-d Families

CLUSTAL W (1.83) multiple sequence alignment

```

GLDA_ECOL6  --MDRIIQSPGKYIQGADVINRLGEYLKPLAERWLUGDKFULGFAQSTUEKSFKDAGLU
GLDA_PSEPU  --MDRAIQSPGKYVQGADALQRLGDYKPLADSWLUIADKFULGFEDTIRQSLSKAGLA
GLDA_BACST  MAAERUFISPAKYVQGKNUITKIANYLEGIGNKTUUIADEIUVKTIAGHTIUNELKGNIA
           :* : **:** ** :. : :.***: :. : :*.:** ** * : :. ....

GLDA_ECOL6  UEIAPFGGECSQNEIDRLRGI AETAQC GAILG GGGKTLDTAKALAHFMGUPVAIAPTIA
GLDA_PSEPU  MDIVAFNGECSQGEVDRLCQLATQNGRSAIUGI GGGKTLDTAKAVAUFFQKUPVAVAPTIA
GLDA_BACST  AEEUVFSGEASRNEVERIANIARKAEAAIUIG GGGKTLDTAKAVADELDAVIUIUPTAA
           : . *.*.*.:**:**: * * . :*: *****:* . :.*** *

GLDA_ECOL6  STDAPCSALSUIYDDEGEFDRYLLPNPNMVIUDTKI VAGAPARLLAAGIGDALATWFE
GLDA_PSEPU  STDAPCSALSULYDDEGEFDRYLMPTNPALVUUDTAIVARAPARLLAAGIGDALATWFE
GLDA_BACST  STDAPTSALSUIYSDDGUFESYRFYKKNPDLULUDTKI IANAPRLLASGIADALATWUE
           ***** **:**:**:* * : * : .** :** ** * * **:**:**:**:**

GLDA_ECOL6  ARACRSRGATTHAGGKCTQAALALAE LCVNTLLEEGEKAMLAEQH UUTPALERIUEANT
GLDA_PSEPU  ARAASRSSAATHAGGPATQTALNLARFCYDTLLEEGEKAMLAUQAQ UUTPALERIUEANT
GLDA_BACST  ARSUIKSGGKTHAGGIPTIAEAIAEKCEQTLFKYGLAYESUKAK UUTPALEUUEANT
           **: :*.. ***** * * :* . * **:**: * * :. : ***** :****

GLDA_ECOL6  YLSGV GFESGGLAAAHVHNGLTAIP-DAHYY HGEKVAFGTLQLULENAPEVEIETVA
GLDA_PSEPU  YLSGV GFESGGVAAAHVHNGLTAVA-ETHHFY HGEKVAFGULQLALENASNAEMQEUM
GLDA_BACST  LLSGL GFESGGLAAAHVHNGETALEGEIHHL HGEKVAFGTLQLALEHSQQEIERYI
           ***: *****: : ** *****: . *.:

GLDA_ECOL6  ALSHAUGLPITLAQLDIKEDUPAKMRIA EAEACA EGETIHNMPGGATPDQVYALLVADQ
GLDA_PSEPU  SLCHAUGLPITLAQLDITEDIPKMRVAELACAPGETIHNMPGGUTVEQVYCALLVADQ
GLDA_BACST  ELYLSLDLPUTLEDIKLDASREDILKVAKAATAEGETIHNMFN-UTADDUADAIFAADQ
           * :.***:** :. :. : . : ** : * * ***** . * :** **:**:**

GLDA_ECOL6  YGQRFLQWE-
GLDA_PSEPU  LGQHLEF---
GLDA_BACST  YAKAYKEHRK
           . : : :

```



Novel HT Enzyme Function Prediction Pipeline

Combined Sequence Homology & FDS

- Improvement Upon Pure Homology Methods
- 4-digit Families
 - Enzyme Active Sites
 - CSA
 - SwissProt

CLUSTAL W (1.83) multiple sequence alignment

```

GOX_TALFL      MUSVFLSTLLLAATUQAYLPAQQIDUQSSLLSDPSKVAGKTYDYIAGGGLTGLTVAAK
GOX_PENAG      -----YLPAQQIDUQSSLLSDPSKVAGKTYDYIAGGGLTGLTVAAK
                *****

GOX_TALFL      LTENPKIKULVIEKGFYESNDGAIIEDPNAYGQIFGTTUDQNYLTUPLINNRTNNIKAGK
GOX_PENAG      LTENPKIKULVIEKGFYESNDGAIIEDPNAYGQIFGTTUDQNYLTUPLINNRTNNIKAGK
                *****

GOX_TALFL      GLGGSTLINGDSWTRPDKUQIDSWEKUFGMEGWNWDSMFEYMKKAEARPTAAQLAAGH
GOX_PENAG      GLGGSTLINGDSWTRPDKUQIDSWEKUFGMEGWNWDSMFEYMKKAEARPTAAQLAAGH
                *****

GOX_TALFL      YFNATCHGTNGTUQSGARDNGQPWSPIMKALMNTUSALGUPUQQDFLCGHPRGUSMIHNN
GOX_PENAG      SFNATCHGTNGTUQSGARDNGQPWSPIMKALMNTUSALGUPUQQDFLCGHPRGUSMIHNN
                *****

GOX_TALFL      UDENQURUDAARAWLLPSYQRPNLEILTGQVUGKULFKQTASGPQAVGUNFGTNKAUNFD
GOX_PENAG      LDENQURUDAARAWLLPNYQRSNLEILTGQVUGKULFKQTASGPQAVGUNFGTNKAUNFD
                :*****

GOX_TALFL      UFAKHEULLAAGSAISPLILEYSGIGLKSULDQANUTQLLDLPUGINMQDQTTTUSRA
GOX_PENAG      UFAKHEULLAAGSAISPLILEYSGIGLKSULDQANUTQLLDLPUGINMQDQTTTUSRA
                *****

GOX_TALFL      SAAGAGQGQAVFFANFTETFGDYAPQARELLNTKLDQWAEETUARGGFHNUTALKUQYEN
GOX_PENAG      SSAGAGQGQAVFFANFTETFGDYAPQARDLLNTKLDQWAEETUARGGFHNUTALKUQYEN
                * :*****

GOX_TALFL      YRNWLLDEDVAFELFMDTEGKINFDLWDLIPFTRGSVHILSSDPYLVQFANDPKFFLNE
GOX_PENAG      YRNWLLDEDVAFELFMDTEGKINFDLWDLIPFTRGSVHILSSDPYLVQFANDPKFFLNE
                *****

GOX_TALFL      FDLLGQAAASKLARDLTSQGANKYFAGETLPGVNLQENATLSQMSDYLVQNFQRPNVAU
GOX_PENAG      FDLLGQAAASKLARDLTSQGANKYFAGETLPGVNLQENATLSQMSDYLVQNFQRPNVAU
                *****

GOX_TALFL      SSCMMSRELGGUVDATAKUYGTQGLRUIDGSIPTQSSUHTIFYGHALKVADAILDD
GOX_PENAG      SSCMMSRELGGUVDATAKUYGTQGLRUIDGSIPTQSSUHTIFYGHALKVADAILDD
                *****

GOX_TALFL      YAKSA
GOX_PENAG      YAKSA
                *****

```



Find Annotated Site: PDB code: Search Swiss-Prot code: Search EC number: Search

CSA entry for 12as

[Original Entry](#)

Title:	Ligase		
Compound:	Asparagine synthetase		
Mutant:	Yes		
UniProt/Swiss-Prot:	P00963-ASNA_ECOLI	EC Class:	6.3.1.1
Other CSA Entries:	Homologues of 12as Entries for UniProt/Swiss-Prot: P00963 Entries for EC: 6.3.1.1	Other Databases:	PDB entry: 12as PDBsum entry: 12as UniProt/Swiss-Prot: P00963 IntEnz entry: 6.3.1.1 KEGG entry: 6.3.1.1 MACiE mechanism: 12as EzCatDB entry: S00413

Literature Report:

Mechanism: A two step reaction is postulated in which the amino acid is activated by ATP forming an aminoacyl-adenylate intermediate, to which an ammonia molecule is added.

Sites:

[Catalytic Site \(Get help with this section\)](#)

Found by: [Literature reference \(Structural analysis and templates exist for the 12as family\)](#)

Residue	Chain	Number	UniProt number	Functional part
ASP	A	46	46	Sidechain
ARG	A	100	100	Sidechain
GLN	A	116	116	Sidechain

[Catalytic Site \(Get help with this section\)](#)

Found by: [Literature reference \(Structural analysis and templates exist for the 12as family\)](#)

Residue	Chain	Number	UniProt number	Functional part
ASP	B	46	46	Sidechain
ARG	B	100	100	Sidechain
GLN	B	116	116	Sidechain

Use the check-boxes to select site(s) to view on the 3D structure in RasMol, and press



APPROACH

Algorithms & Logic



The Gene Ontology (GO) Cross-Reference Database

- Common Reference Point Between Databases
 - Saves Time
 - Allows Automation
- Three Organizing Principles
 1. Cellular Component
 2. Biological Process
 3. Molecular Function
- ...2GO Mappings
 - Manually Created Text Files From Different Sources
 - Parse
 - Import into RDBMS

```
external system identifier: external system term name/id > GO:GO term name ; GO:id
```



```
!version: $Revision: 1.38 $
!date: $Date: 2005/07/27 09:42:21 $
!Mapping of GO function_ontology "enzymes" to Enzyme Commission Numbers.
!original mapping by Michael Ashburner, Cambridge.
!This version parsed from function.ontology on 2005/07/13 09:07:00
!by Daniel Barrell, EBI, Hinxton
!
EC:1.1.1.1 > GO:alcohol dehydrogenase activity ; GO:0004022
EC:1.1.1.10 > GO:L-xylulose reductase activity ; GO:0050038
EC:1.1.1.100 > GO:3-oxoacyl-[acyl-carrier protein] reductase activity ; GO:0004316
EC:1.1.1.101 > GO:acylglycerone-phosphate reductase activity ; GO:0000140
EC:1.1.1.102 > GO:3-dehydrosphinganine reductase activity ; GO:0047560
EC:1.1.1.103 > GO:L-threonine 3-dehydrogenase activity ; GO:0008743
EC:1.1.1.104 > GO:4-oxoproline reductase activity ; GO:0016617
```



Anatomy of GO Mappings Database

Five External Databases Mapped

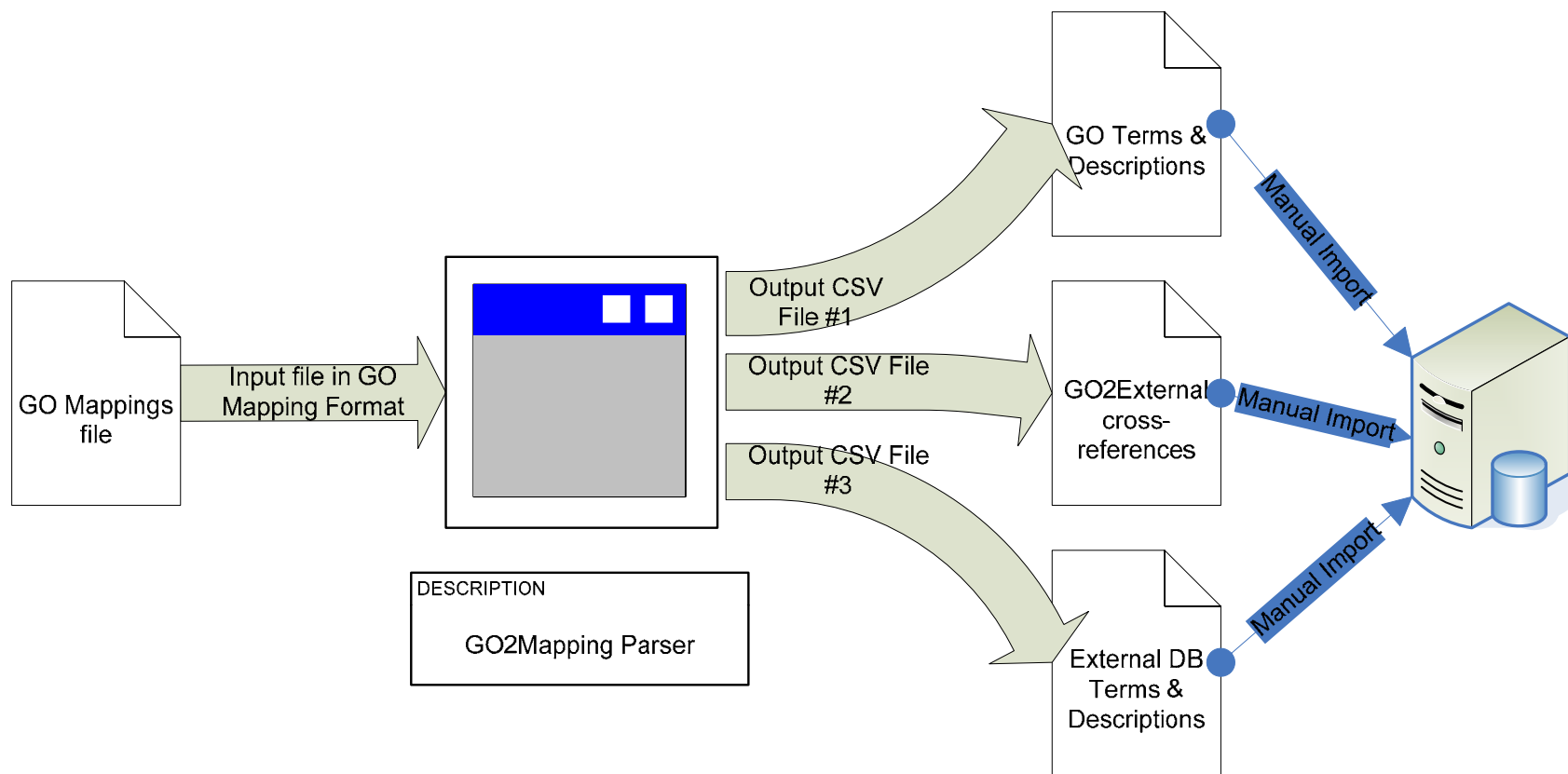
1. Pfam
 - HMM
 - MSA
2. COG
 - Complete Genomes
 - Represent Major Phylogenetic Lineages
 - Individual Paralogs From At Least 3 Lineages
3. PROSITE
 - Biologically Significant Sites, Patterns, and Profiles
 - Protein Family ID
4. ProDom
 - Generated Automatically From SWISS-PROT & TrEMBL
5. EC
 - Common Names of Enzymes
 - EC Numbers



- Parser Overview
 - Perl Language
 - 1 Command Line Parameter (Name of the Input File)
 - Automatic Identification of Mapping File Type
- Output Details (3rd Normal Form)
 1. GO Terms & Corresponding Descriptions
 2. Cross-Reference (GO ↔ External DB Term)
 3. External DB Term w/ Corresponding Description
- RDBMS Platform Details
 - 13 Tables
 - MySQL 5.0
 - SQLyog 5.02

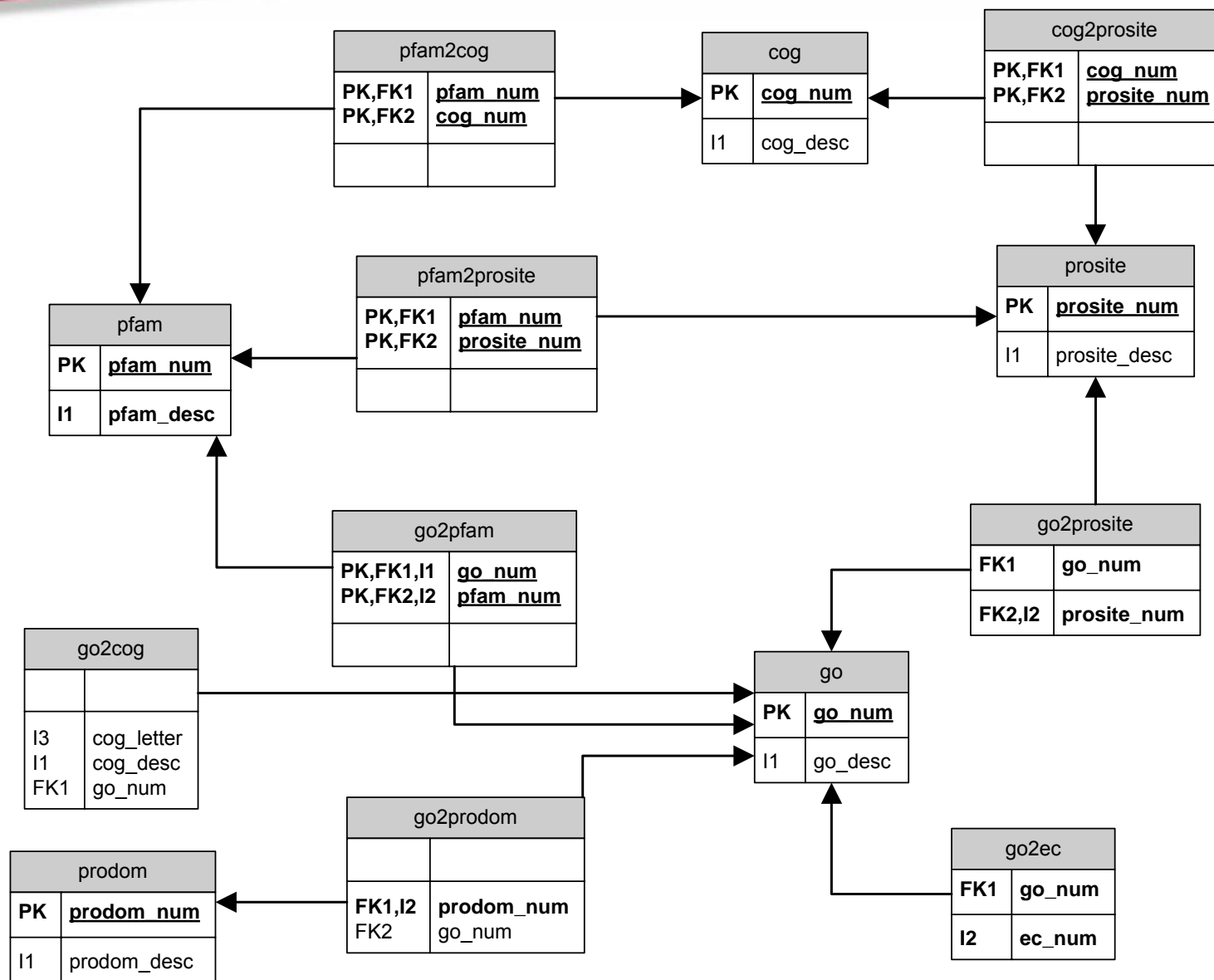


GO Mapping File Parser Diagram





GO Mappings Database Schema





- Lightweight Implementation of DB Cross-references
 - User Friendly
 - No Extraneous Information
 - Extremely Fast
 - All In One Place

<http://binf.vsevolod.com/GO.php>

- GUI Details
 - Centralized Navigation Menu Bar
 - Search Query Window
 - Tables of Results (“Synonyms”)



GO Cross-Reference Database





Reasons for Developing CSA Web Scraper

- CSA Provides Active Site Profiles for 4-digit Families

Problems with Acquisition of Data

- CSA Database Unavailable for Download
- EMBL Does Not Provide Custom Query Results
 - Browsing Only
- Cannot Be Incorporated Into Pipeline In Existing Form



- Screen Scraping
 - Technique by which software extracts information from Computer Screen



- Difference from Parsing
 - Input Intended for 'Human Consumption' Instead of Machine Interpretation
- Web Scraping
 - Works on Underlying Web Document Object Model
 - HTML
 - JavaScript



Obtaining Catalytic Sites

Step1: Obtain Complete List of PDB with Active Site Annotation

- Derived From Literature
- http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/CSA/CSA_Browse.pl

EMBL-EBI
European Bioinformatics Institute

EBI Home About EBI Groups Services Toolbox Databases Downl
Catalytic Site Atlas Version 2.1.2

Find Annotated Site: PDB code: Search Swiss-Prot code: Search EC number:

Browse literature based entries

Note that this list includes only entries derived directly from the literature. There are many more entries inferred by sequence comparison.
Click on a column heading to sort by that column.

PDB code	Name	EC number	CATH code
12as	Asparagine synthetase	6.03.01.0001	3.30.930.10
132l	Lysozyme	3.02.01.0017	1.10.530.10
135l	Lysozyme	3.02.01.0017	1.10.530.10
13pk	3-phosphoglycerate kinase	2.07.02.0003	3.40.50.1260 3.40.50.1270
1a05	3-isopropylmalate dehydrogenase	1.01.01.0085	3.40.718.10
1a0i	Dna ligase	6.05.01.0001	2.40.50.140 3.30.1490.70 3.30.470.30
1a0j	Trypsin	3.04.21.0004	2.40.10.10



Most Web Content on the Internet is Organized in HTML Tables

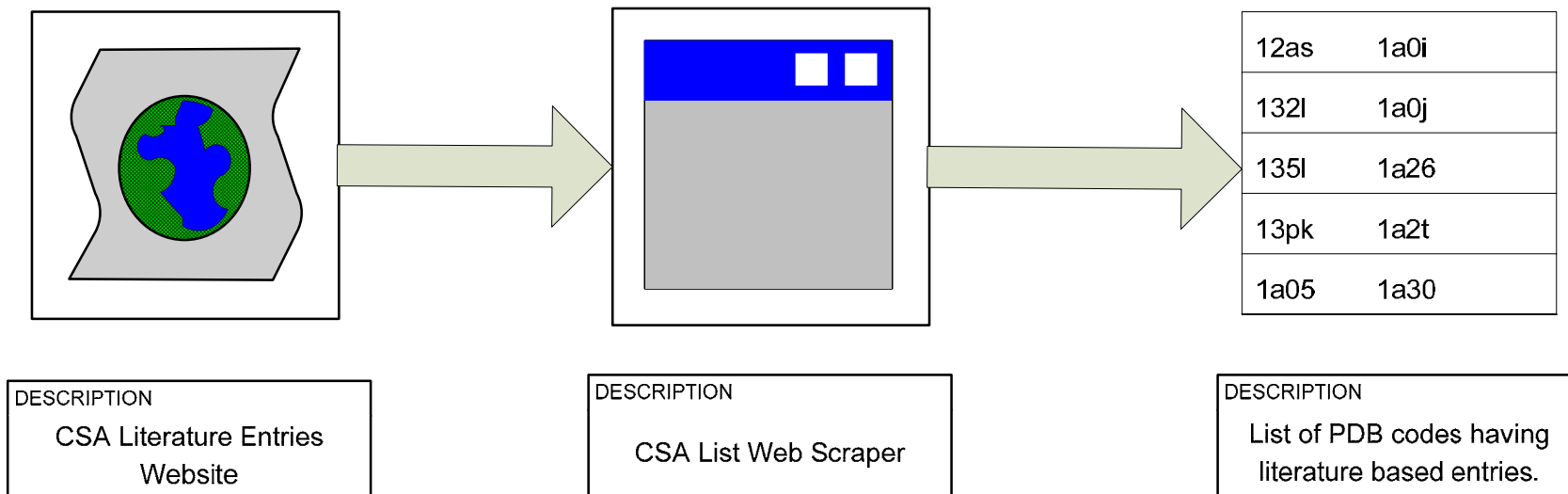
- **HTML::TableExtract**

- Extracts content from tables in HTML
- <http://www.mojotoad.com/sisk/projects/HTML-TableExtract/tables.html>

Table (0,0)			
0,0:1,0		0,0:1,1	
Table (1,0)			
East	Central	West	
1,0:2,0	1,0:2,1	1,0:2,2	
1,0:3,0		1,0:3,2	
1,0:4,0		1,0:4,2	
1,0:5,0	1,0:5,1	1,0:5,2	
0,0:2,0		0,0:2,1	
Table (1,2)			
Left	Right		
1,2:2,0	1,2:2,1		
Table (2,0)		Table (2,1)	
Pacific	Atlantic	Lefty	Righty
2,0:2,0	2,0:2,1	2,1:2,0	2,1:2,1
2,0:3,0	2,0:3,1	2,1:3,0	2,1:3,1
1,2:3,0	1,2:3,1		
1,2:4,0	1,2:4,1		
1,2:5,0	1,2:5,1		
0,0:2,1		Table (1,3)	
Pacific	Plains	Atlantic	
1,3:2,0	1,3:2,1	1,3:2,2	
	1,3:3,1	1,3:3,2	
1,3:4,0		1,3:4,2	
1,3:5,0		1,3:5,2	



Obtaining Catalytic Sites





Step 2: Analyze Data for Each PDB Containing Literature Based Sites

- Follow Link by Attaching PDB Notation to the URL
- Process Data



Obtaining Catalytic Sites

Step 2: Analyze Data for Each PDB Containing Literature Based Sites

- **HTML::TreeBuilder**

- Parser that builds an HTML syntax tree



Table (0,0)			
0,0:1,0			
Table (1,0)			
East	Central	West	
1,0:2,0	1,0:2,1	1,0:2,2	
1,0:3,0		1,0:3,2	
1,0:4,0		1,0:4,2	
1,0:5,0	1,0:5,1	1,0:5,2	
0,0:2,0			
Table (1,2)			
Left		Right	
1,2:2,0		1,2:2,1	
Table (2,0)		Table (2,1)	
Pacific	Atlantic	Lefty	Righty
2,0:2,0	2,0:2,1	2,1:2,0	2,1:2,1
2,0:3,0	2,0:3,1	2,1:3,0	2,1:3,1
1,2:3,0		1,2:3,1	
1,2:4,0		1,2:4,1	
1,2:5,0		1,2:5,1	
0,0:1,1			
Table (1,1)			
Left	Middle	Right	
1,1:2,0	1,1:2,1	1,1:2,2	
1,1:3,0	1,1:3,1	1,1:3,2	
1,1:4,0	1,1:4,1	1,1:4,2	
1,1:5,0	1,1:5,1	1,1:5,2	
0,0:2,1			
Table (1,3)			
Pacific	Plains	Atlantic	
1,3:2,0	1,3:2,1	1,3:2,2	
	1,3:3,1	1,3:3,2	
1,3:4,0		1,3:4,2	
1,3:5,0		1,3:5,2	

EMBL-EBI
European Bioinformatics Institute

EBI Home About EBI Groups Services Toolbox Databases Downloads Submissions

Catalytic Site Atlas Version 2.1.9

Find Annotated Site: PDB code: Search Swiss-Prot code: Search EC number: Search

CSA entry for 12as

Original Entry

Title: Ligase
Compound: Asparagine synthetase
Mutant: Yes
UniProt/Swiss-Prot: P00963-ASNA_ECOLI
EC Class: 6.3.1.1

Other CSA Entries: Homologues of 12as
Entries for UniProt/Swiss-Prot: P00963
Entries for EC: 6.3.1.1

Other Databases: PDB entry: 12as
PDBsum entry: 12as
UniProt/Swiss-Prot: P00963
IntEnz entry: 6.3.1.1

KEGG entry: 6.3.1.1
MACIE mechanism: 12as
EzCatDB entry: S00413

Literature Report:

Mechanism: A two step reaction is postulated in which the amino acid is activated by ATP forming an aminoacyl-adenylate intermediate, to which an ammonia molecule is added.

Sites:

Catalytic Site (Get help with this section)

Found by: Literature reference (Structural analysis and templates exist for the 12as family)

Residue	Chain	Number	UniProt number	Functional part
ASP	A	46	46	Sidechain
ARG	A	100	100	Sidechain
GLN	A	116	116	Sidechain

Catalytic Site (Get help with this section)

Found by: Literature reference (Structural analysis and templates exist for the 12as family)

Residue	Chain	Number	UniProt number	Functional part
ASP	B	46	46	Sidechain
ARG	B	100	100	Sidechain
GLN	B	116	116	Sidechain

Use the check-boxes to select site(s) to view on the 3D structure in RasMol, and press



CSA Web Scraper Logic Flow

CSA entry for 12as Original Entry

Title: Ligase
Compound: Asparagine synthetase
Mutant: Yes
UniProt/Swiss-Prot: P00963-ASNA_ECOLI
EC Class: 6.3.1.1

Other CSA Entries: Homologues of 12as
Entries for UniProt/Swiss-Prot: P00963
Entries for EC: 6.3.1.1

Other Databases: PDB entry: 12as
PDBsum entry: 12as
UniProt/Swiss-Prot: P00963
IntEnz entry: 6.3.1.1

KEGG entry: 6.3.1.1
MACIE mechanism: 12as
ExCatDB entry: S00413

Literature Report:

Mechanism: A two step reaction is postulated in which the amino acid is activated by ATP forming an aminoacyl-adenylate intermediate, to which an ammonia molecule is added.

Sites:

Catalytic Site (Get help with this section)

Found by: Literature reference (Structural analysis and templates exist for the 12as family)

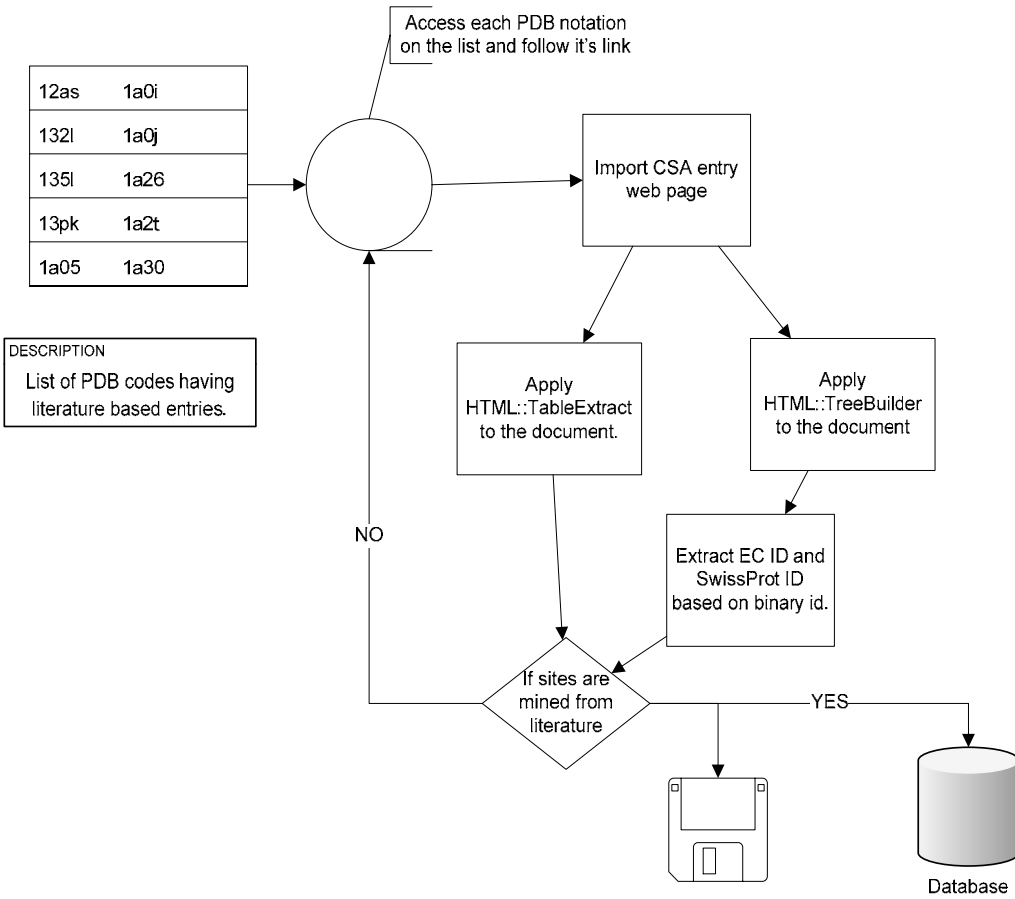
Residue	Chain	Number	UniProt number	Functional part
ASP	A	46	46	Sidechain
ARG	A	100	100	Sidechain
GLN	A	116	116	Sidechain

Catalytic Site (Get help with this section)

Found by: Literature reference (Structural analysis and templates exist for the 12as family)

Residue	Chain	Number	UniProt number	Functional part
ASP	B	46	46	Sidechain
ARG	B	100	100	Sidechain
GLN	B	116	116	Sidechain

Use the check-boxes to select site(s) to view on the 3D structure in RasMol, and press





Incorporating FDS Into Pipeline

- Combining Pipeline Components

- EPD Predictions
- FDS
- Active Sites

- Preparation for Validation

- 3-digit EC Predictions

- FDS Details

- Stored in Individual Files (.fis)
- Profiles Contain Accuracy, Sensitivity, and Specificity Information

- EPD Details

- One Continuous File
- EC Predictions
- Profile Names
- Thresholds (red)
- Scores

```
FAMILY> DHB7_HUMAN
```

```
ACC> 1.000000
```

```
SEN> 0.791667
```

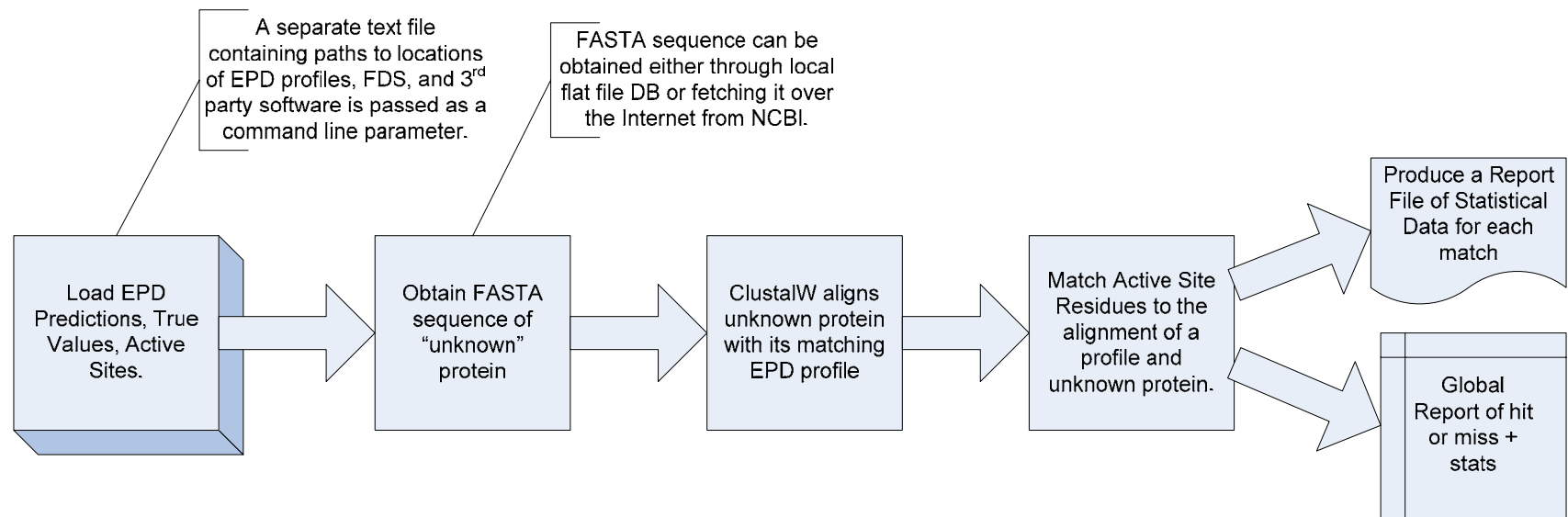
```
SPEC> 1.000000
```

```
pos res freq sc-t sc-ho sc-he
POS> 91 T 1.000 2.340 1.822 1.732
POS> 173 N 1.000 1.248 1.822 0.378
POS> 172 N 0.958 1.090 1.520 0.901
POS> 225 N 0.958 0.838 1.520 0.403
POS> 276 Y 0.958 0.720 1.520 0.044
POS> 293 E 0.875 0.683 1.027 0.987
POS> 293 D
POS> 255 S 0.958 0.610 1.520 -0.473
POS> 252 N 0.875 0.436 1.166 0.277
```

```
>>>>
3BHS_CANFA + 0 1.1.1 3BHS1_MESAU 207 1746
3BHS_CANFA + 4577 5.3.3 3BHS1_MOUSE 146 1625
3BHS_CANFA + 4578 5.3.3 3BHS_UACCA 152 1595
3BHS_CANFA + 30 1.1.1 DFRA_PETHY 312 1264
3BHS_CANFA + 17 1.1.1 ALD2_SPOSA 315 1261
3BHS_CANFA + 29 1.1.1 DFRA_ARATH 315 1261
3BHS_CANFA + 4593 5.3.3 TYRP2_HUMAN 964 1041
3BHS_CANFA + 62 1.1.1 FCL_CRIGR 231 609
3BHS_CANFA + 145 1.1.1 RFB0_RHISN 152 260
3BHS_CANFA + 784 1.6.5 NUEM_BOVIN 72 75
3BHS_CANFA + 1992 2.7.1 M3K4_MOUSE 60 65
```



Pipeline Logic Flow





Pipeline Output

MATCH> 3BHS1_MESAU
 EC> 1.1.1
 NORM> 7.87439609722514
 NAME> 3BHS4_RAT
 POS> 74 F + F
 POS> 389 T + T
 STA> 2 2

MATCH> 3BHS1_MOUSE
 EC> 5.3.3
 NORM> 10.6712328036217
 NAME> 3BHS4_RAT
 POS> 106 T + T
 POS> 164 E + E
 POS> 268 D + D d =
 POS> 342 Y + Y
 POS> 358 W + W
 STA> 5 5

MATCH> 3BHS_VACCA
 EC> 5.3.3
 NORM> 9.95394730293456
 NAME> 3BHS4_RAT
 POS> 165 E + E
 POS> 307 S + S
 POS> 308 K + K h
 STA> 3 3

...

MATCH> FCL_CRIGR
 EC> 1.1.1
 NORM> 2.10389609478833
 ACC> 0.295455 < 0.80!!!!
 STA> N\A N\A

MATCH> NUEM_BOVIN
 EC> 1.6.5
 NORM> 0.152777775655864
 NAME> 3BHS4_RAT
 POS> 57 V - T
 POS> 123 A - S
 POS> 321 S - R d =
 STA> 0 3

MATCH> RFPD_RHISM
 EC> 1.1.1
 NORM> 0.901315783543975
 NAME> 3BHS4_RAT
 POS> 389 T + T
 STA> 1 1

MATCH> RFPD_SHIFL
 EC> 1.1.1
 NORM> 0.117088607224403
 NAME> 3BHS4_RAT
 POS> 96 D + D d =
 POS> 139 N + N
 POS> 168 G - F
 POS> 192 Y + Y
 POS> 196 K + K h
 POS> 200 E + E
 STA> 5 6

EC_NUMBER	SITE_HITS	TOTAL_SITES	PCNT_MATCH	EC_HITS	TOTAL_NORM	AVG_NORM	MAX_NORM	V
5.3.3	0008	0008	1.0000	0002	20.6252	10.3126	10.8712	1
1.6.5	0000	0003	0.0000	0001	0.1528	0.1528	0.1528	0
1.1.1	0026	0027	0.9630	0007	20.1180	2.8740	7.8744	1

QUERY> 3BHS4_RAT TRUE> 1.1.1, 5.3.3.

QUERY_NAME	EC_NUMBR	SITE_HITS	TOTAL_SITES	PCNT_MATCH	EC_HITS	TOTAL_NORM	AVG_NORM	MAX_NORM	V	TRUE_MATCH
1A11_ORYSA	2.6.1 0155	0178	0.8708	0021	27.3183	1.3009	2.0127	0	4.4.1.	
1A11_ORYSA	4.4.1 0019	0020	0.9500	0004	1.9058	0.4764	0.6567	1	4.4.1.	
1A11_ORYSA	6.1.1 0002	0008	0.2500	0002	0.1244	0.0622	0.0893	0	4.4.1.	
>>>>										
1A19_ARATH	2.6.1 0155	0178	0.8708	0021	26.6684	1.2699	1.8861	0	4.4.1.	
1A19_ARATH	4.4.1 0022	0023	0.9565	0005	2.8674	0.5735	0.8004	1	4.4.1.	
>>>>										
1A1D_BURMA	4.2.1 0000	0006	0.0000	0002	0.4081	0.2040	0.2772	0	3.5.99.	
1A1D_BURMA	3.5.99	0084	0.0084	1.0000	0004	1.7666	0.4417	0.4461	1	3.5.99.
>>>>										
1A1D_BURPS	4.2.1 0000	0006	0.0000	0002	0.4081	0.2040	0.2772	0	3.5.99.	
1A1D_BURPS	3.5.99	0084	0.0084	1.0000	0004	1.7666	0.4417	0.4461	1	3.5.99.
>>>>										
1A1D_CRYNE	4.2.1 0001	0003	0.3333	0001	0.1287	0.1287	0.1287	0	3.5.99.	
1A1D_CRYNE	3.5.99	0072	0.0084	0.8571	0004	0.1248	0.0312	0.0335	1	3.5.99.
>>>>										
3BHS1_HUMAN	5.3.3 0008	0008	1.0000	0002	19.3558	9.6779	10.0137	1	1.1.1.	5.3.3.
3BHS1_HUMAN	1.1.1 0021	0021	1.0000	0006	18.3446	3.0574	7.2657	1	1.1.1.	5.3.3.
>>>>										
3BHS3_MOUSE	5.3.3 0008	0008	1.0000	0002	20.3150	10.1575	10.4795	1	1.1.1.	5.3.3.
3BHS3_MOUSE	1.6.5 0000	0003	0.0000	0001	0.0694	0.0694	0.0694	0	1.1.1.	5.3.3.
3BHS3_MOUSE	1.1.1 0021	0021	1.0000	0006	19.1758	3.1960	7.6522	1	1.1.1.	5.3.3.
>>>>										
3BHS4_RAT	5.3.3 0008	0008	1.0000	0002	20.6252	10.3126	10.6712	1	1.1.1.	5.3.3.
3BHS4_RAT	1.6.5 0000	0003	0.0000	0001	0.1528	0.1528	0.1528	0	1.1.1.	5.3.3.
3BHS4_RAT	1.1.1 0026	0027	0.9630	0007	20.1180	2.8740	7.8744	1	1.1.1.	5.3.3.
>>>>										



Incorporating Active Sites Into Pipeline

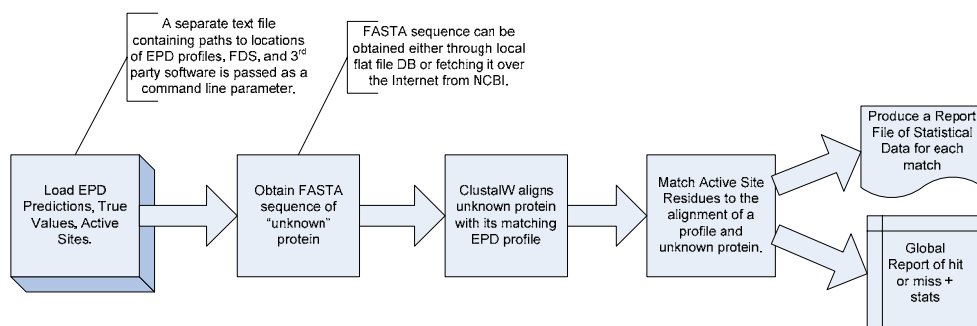
FAMILY> DHB7_HUMAN

- Preparation for Validation

- 2. 4-digit EC Predictions

- Similar to 3-d Predictions, except:
 - Translation of Active Site Positions to ClustalW Alignment Slice
 - Less Details Available

```
pos res fr  
POS> 91 T  
POS> 173 N  
POS> 172 N  
POS> 225 N  
POS> 276 Y  
POS> 293 E  
POS> 293 D  
POS> 255 S  
POS> 252 N
```



```
>>>>  
3BHS_CANFA + 0 1.1.1 3BHS1_MESAU 207 1746  
3BHS_CANFA + 4577 5.3.3 3BHS_MOUSE 146 1625  
3BHS_CANFA + 4578 5.3.3 3BHS_UACCA 152 1595  
3BHS_CANFA + 30 1.1.1 DFRA_PETHY 312 1264  
3BHS_CANFA + 17 1.1.1 ALD2_SPOSA 315 1261  
3BHS_CANFA + 29 1.1.1 DFRA_ARATH 315 1261  
3BHS_CANFA + 4593 5.3.3 TYRP2_HUMAN 964 1041  
3BHS_CANFA + 62 1.1.1 FCL_CRIGR 231 609  
3BHS_CANFA + 145 1.1.1 RFB_D_RHISN 152 260  
3BHS_CANFA + 784 1.6.5 NUEM_BOVIN 72 75  
3BHS_CANFA + 1992 2.7.1 M3K4_MOUSE 60 65
```



SOFTWARE IMPLEMENTATION

Platform & Execution Details



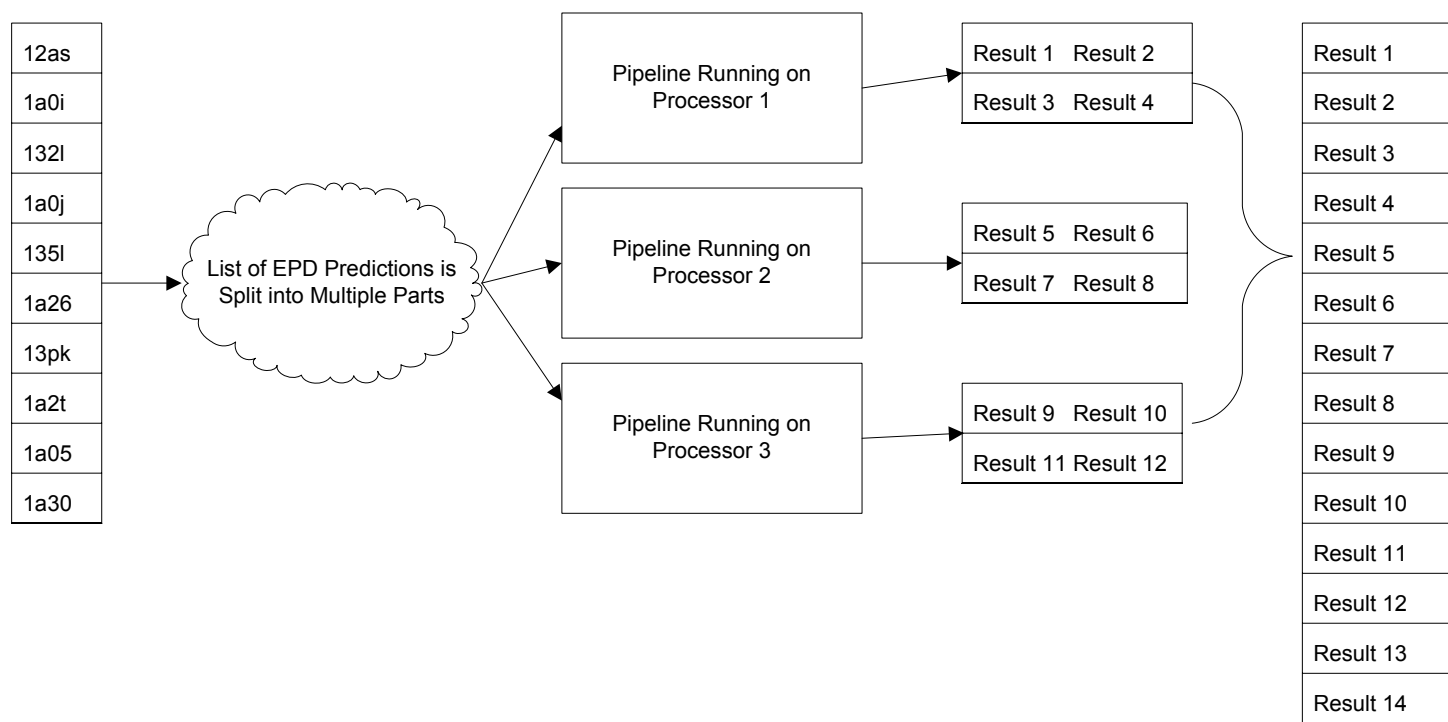
The results presented here are based on the prototype and do not reflect the performance of the improved pipeline that will be used at BHS AI.

1. Intel Pentium 4 3.0 GHz 1MB L2 800 MHz front-side bus CPU, 1.00 GB of RAM, MS Windows XP Professional @ home.
2. 4 processor Intel Xeon 2.8 GHz 2x1MB L2 800 MHz front-side bus CPU, 2.00 GB of RAM, MS Windows Server 2003 @ BHS AI.
3. 8 processor 64-bit Sun UltraSparc IIe 500MHz 256 KB L2 CPU, 2.00 GB of RAM, Sun Solaris 9 @ ExonHit Therapeutics, Inc.



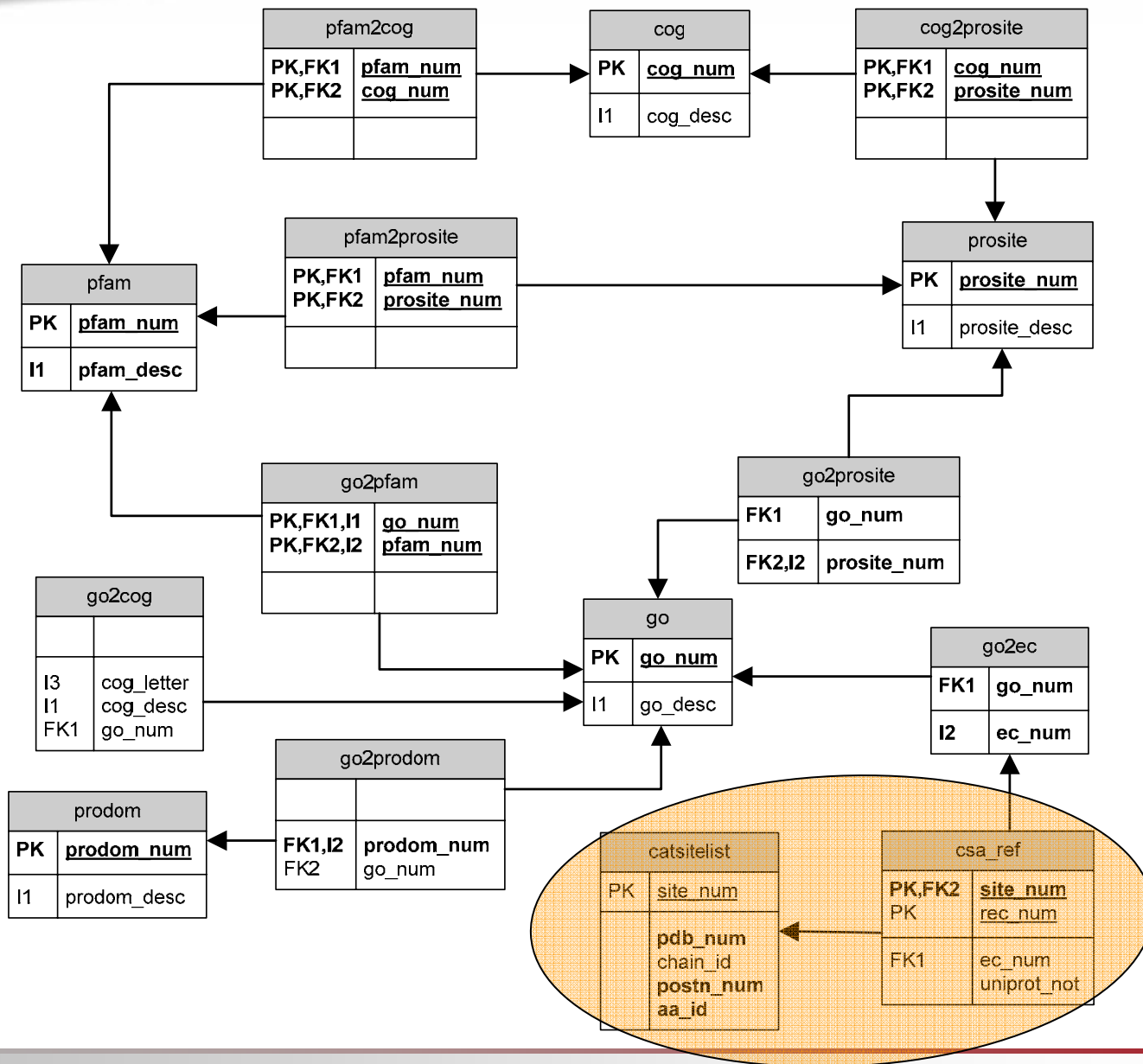
Serial & Parallel Execution Time

- Serial Execution
 - Execution Depends on Clock Speed
 - No Improvement Using #2 & #3
- Parallel Execution
 - Dividing List of Unknowns Between Processors
 - More Efficient To Estimate FDS Conservation Right After EPD Prediction (not explored)





Generated Databases





VALIDATION

Results & Discussion



Enzyme Validation Set

- 13,239 Proteins
 - Known EC Number
 - Not Used In Profile Generation



- Accuracy
 - Fraction of True Positives (TP) Obtained From All Predictions

$$Accuracy = \frac{TP}{TP + FP}$$

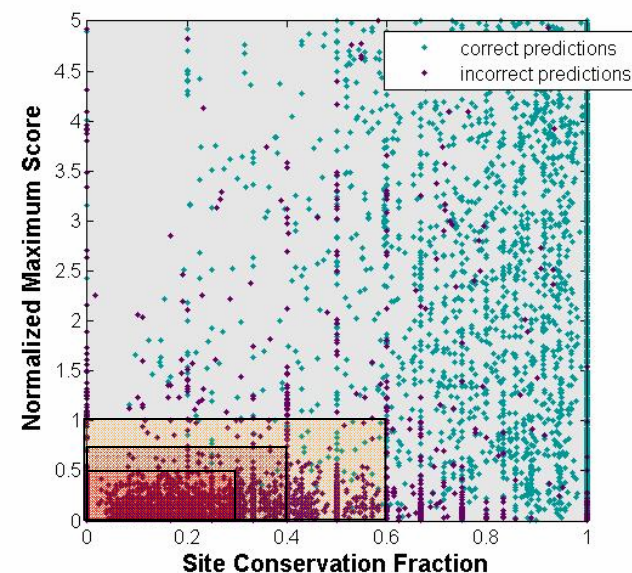
- Coverage
 - Fraction of a Number of Predictions Obtained From Total Number of Proteins Tested

$$Coverage = \frac{\text{Number Proteins with Predicted Function}}{\text{Total Number of Tested Proteins}}$$



Results (3-digit Families)

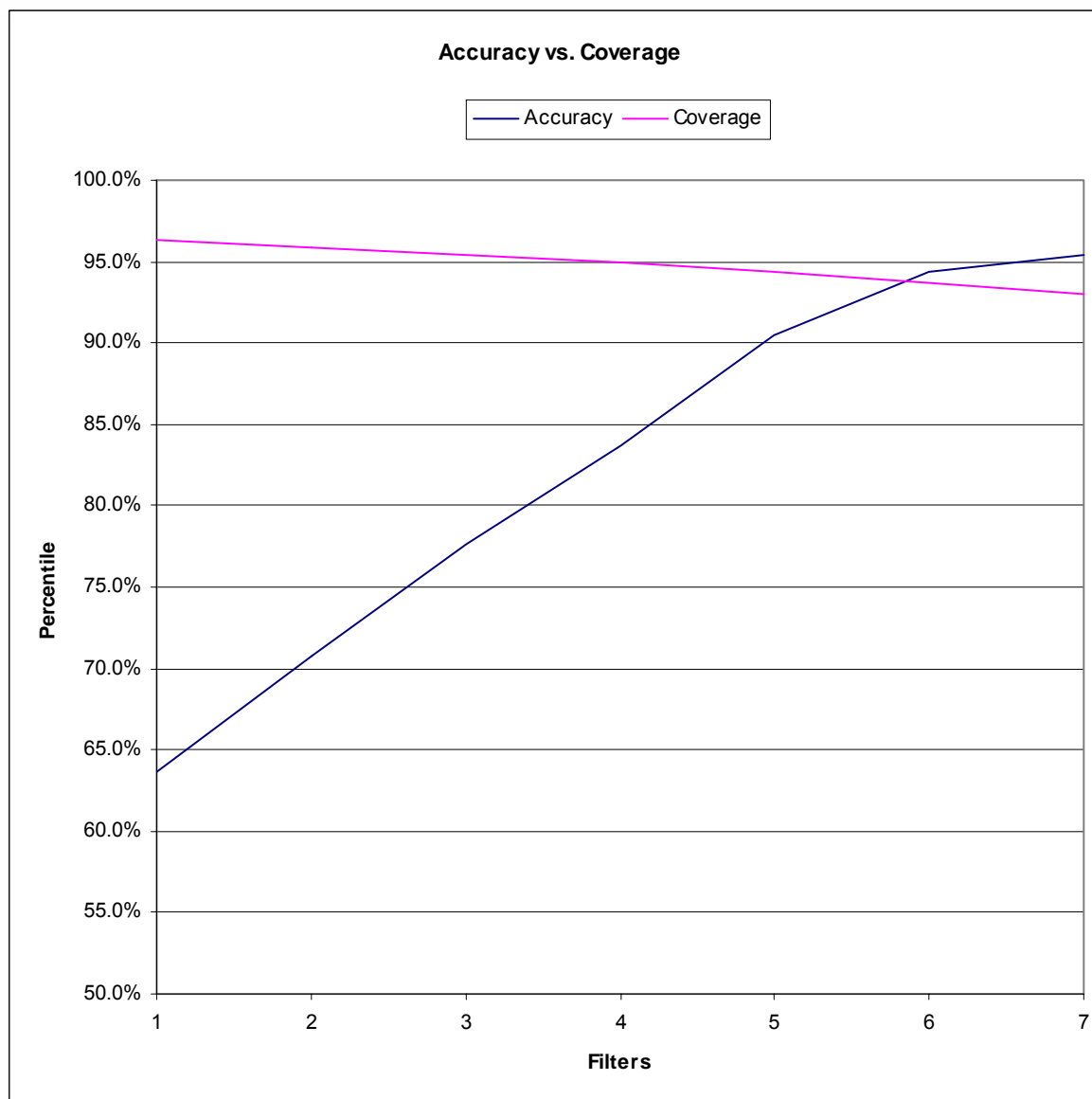
- EPD Coverage
 - 12,749
 - 96.3% Coverage
- Total Number of EC Number Predictions
 - 20,211
 - Overlapping Families Produce Multiple Predictions Per Unknown



TP	FP	Accuracy	Coverage	Filter
12,874	7,337	63.7%	96.3%	NONE
12,855	5,298	70.8%	95.9%	NMS>0.1 + SCF>0.1
12,842	3,712	77.6%	95.4%	NMS>0.2 + SCF>0.2
12,818	2,491	83.7%	95.0%	NMS>0.3 + SCF>0.3
12,787	1,341	90.5%	94.4%	NMS>0.5 + SCF>0.4
12,706	753	94.4%	93.7%	NMS>0.7 + SCF>0.5
12,618	612	95.4%	93.0%	NMS>0.9 + SCF>0.6



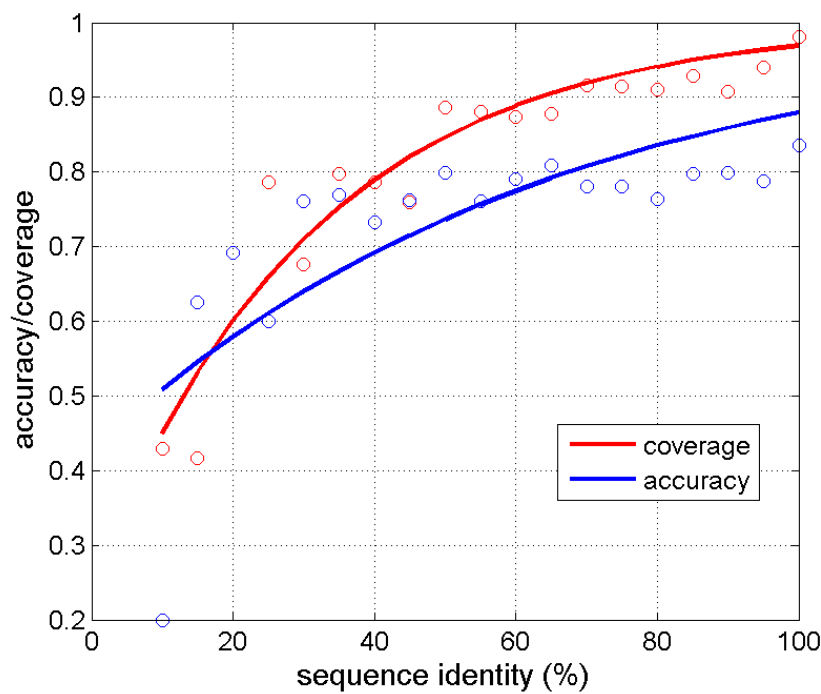
Accuracy vs. Coverage (3-digit)





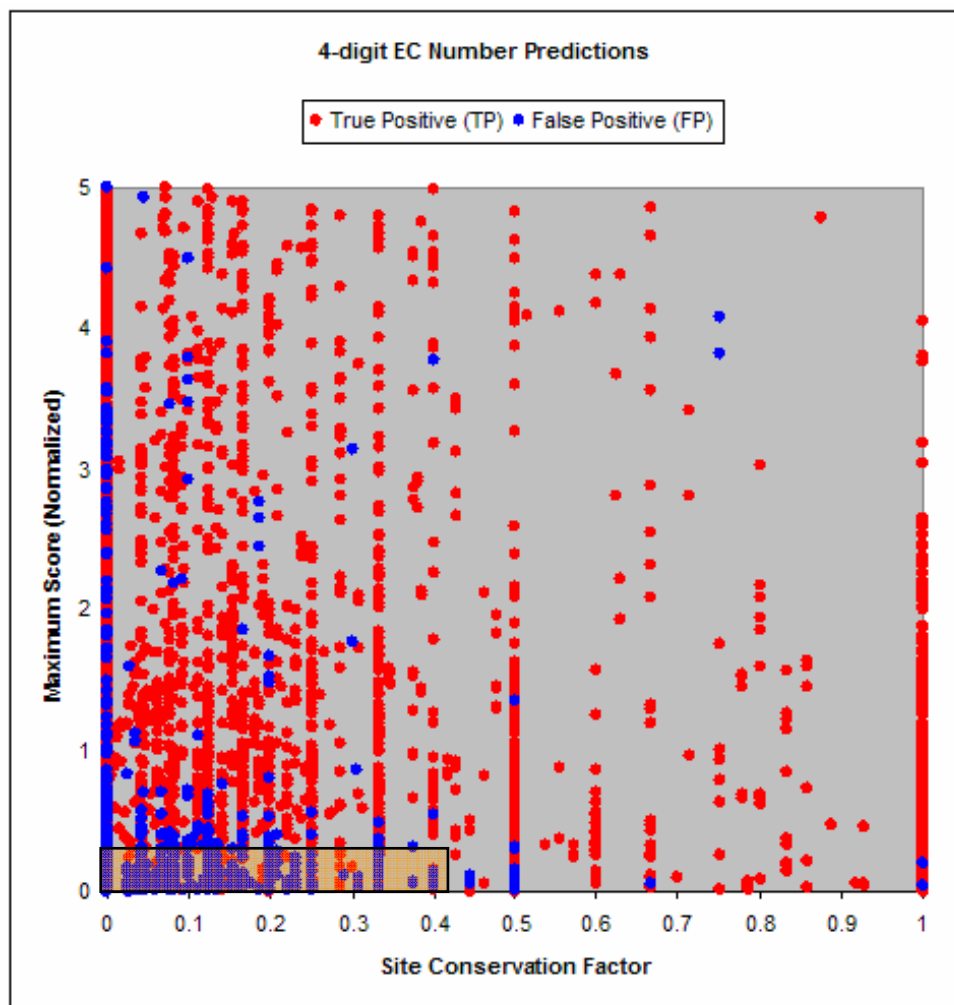
Discussion of 3-digit Results

- Trade-off Between Accuracy & Coverage as a Function of Sequence Identity
 - Use of FDS Filtering Need Further Investigation
- 60% Sequence Identity
 - Accuracy 80%
 - Coverage 90%





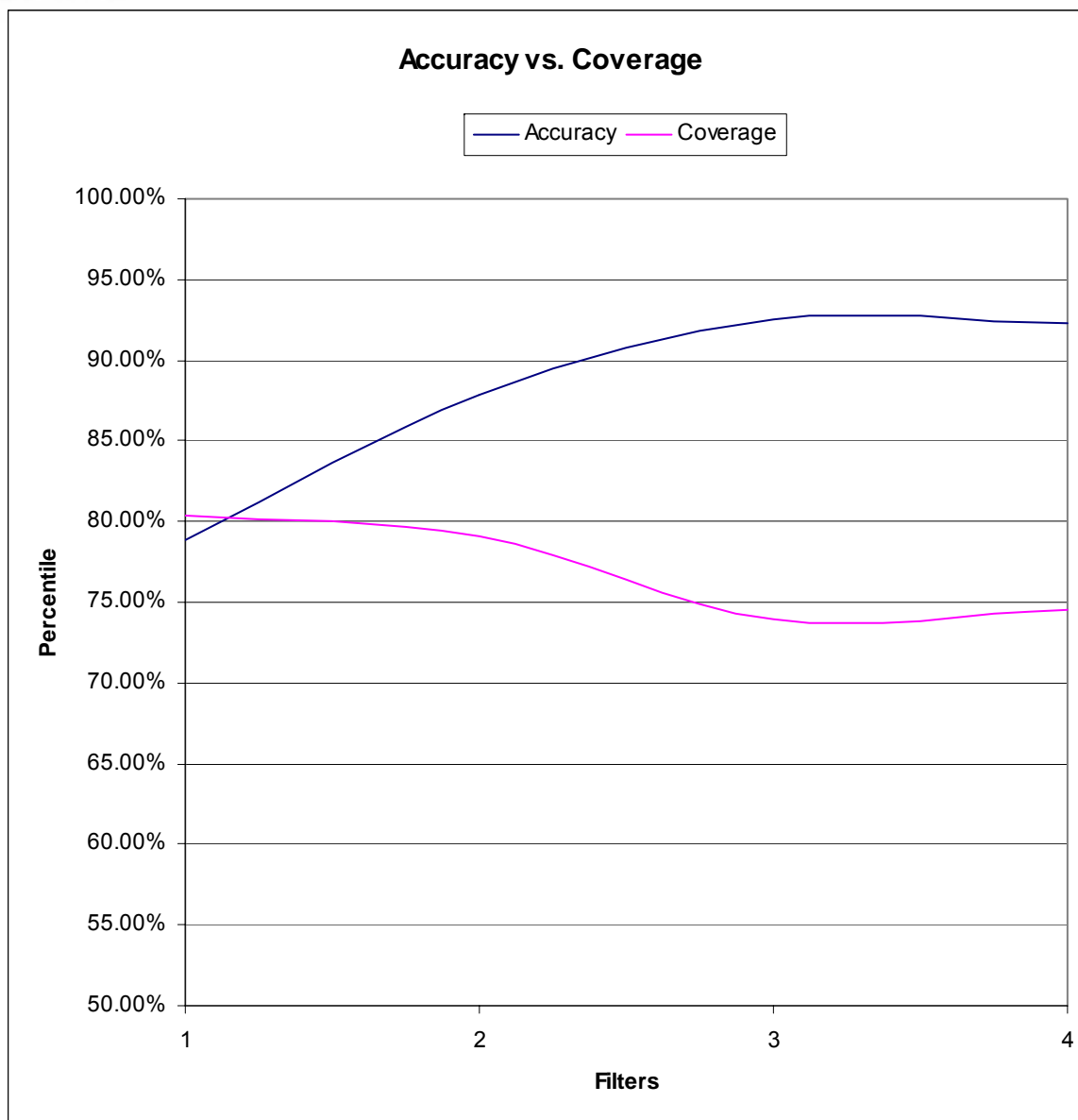
Results (4-digit Families)



TP	FP	Accuracy	Coverage	Filter
10,572	2,821	78.9%	80.4%	NONE
10,504	1,465	87.8%	79.1%	EPD>0
9,860	796	92.5%	73.9%	EPD>0 * NSM>0.1
9,953	830	92.3%	74.5%	EPD>0 * (NMS>0.1 + SCF>0.4)



Accuracy vs. Coverage (4-digit)





Discussion of 4-digit Results

- Accurate Estimation of 4-digit EC Numbers is Difficult
 - Substrate Specificity
 - Active Site Information is Scarce
 - High Probability of Incorrect Alignment
 - FDS Approach Not Used Due to Small Subfamily Sizes
- Effect of Active Site Data
 - Small Effect on Accuracy
 - High Average Conservation Indicative of Reliable Prediction
 - Out of 514 Predictions With Average Conservation Fraction > 50%, only 6 Incorrect
 - 99% Reliability
- Qualitative Regions of Reliability
 1. High: $NMS > 1$ & $SCF > 0.5$
 2. Low: All Other Admissible Cases
 3. Unreliable Predictions: Average $SCF < 0.4$ & $NMS < 0.5$



CONCLUSIONS AND FUTURE WORK

Improvement Strategies



Summary of Research Tasks

- Enzyme Function Prediction Pipeline
 - 3-digit Predictions
 - 4-digit Predictions
- GO Cross-Reference Database
 - Common Reference
 - Active Site & FDR
- Pipeline Performance Assessed



- Pipeline Performance
 - Can Provide High Accuracy With Custom Filters
 - Accuracy Comes at Expense of Coverage
- Additional Dimension (FDS & CSA Sites)
 - Varying Effect
 - Increase Accuracy When Present For All Subfamilies
 - 3-digit
 - No Effect When Scarce
 - 4-digit
 - Combination of Thresholds Can Improve Accuracy



Main Concentration is on Improving 4-digit Predictions

- Combination With Other Tools
 - InterPro
 - “Hooks” Present in GO Mappings Database
- Combination With Pathways
 - Combining Related Pathways
 - Analyze Structure & Homology of Catalyzing Enzymes



The End

Questions?
Comments?