Automated Functional Annotation and Necessary Tools

Functional Annotation Overview

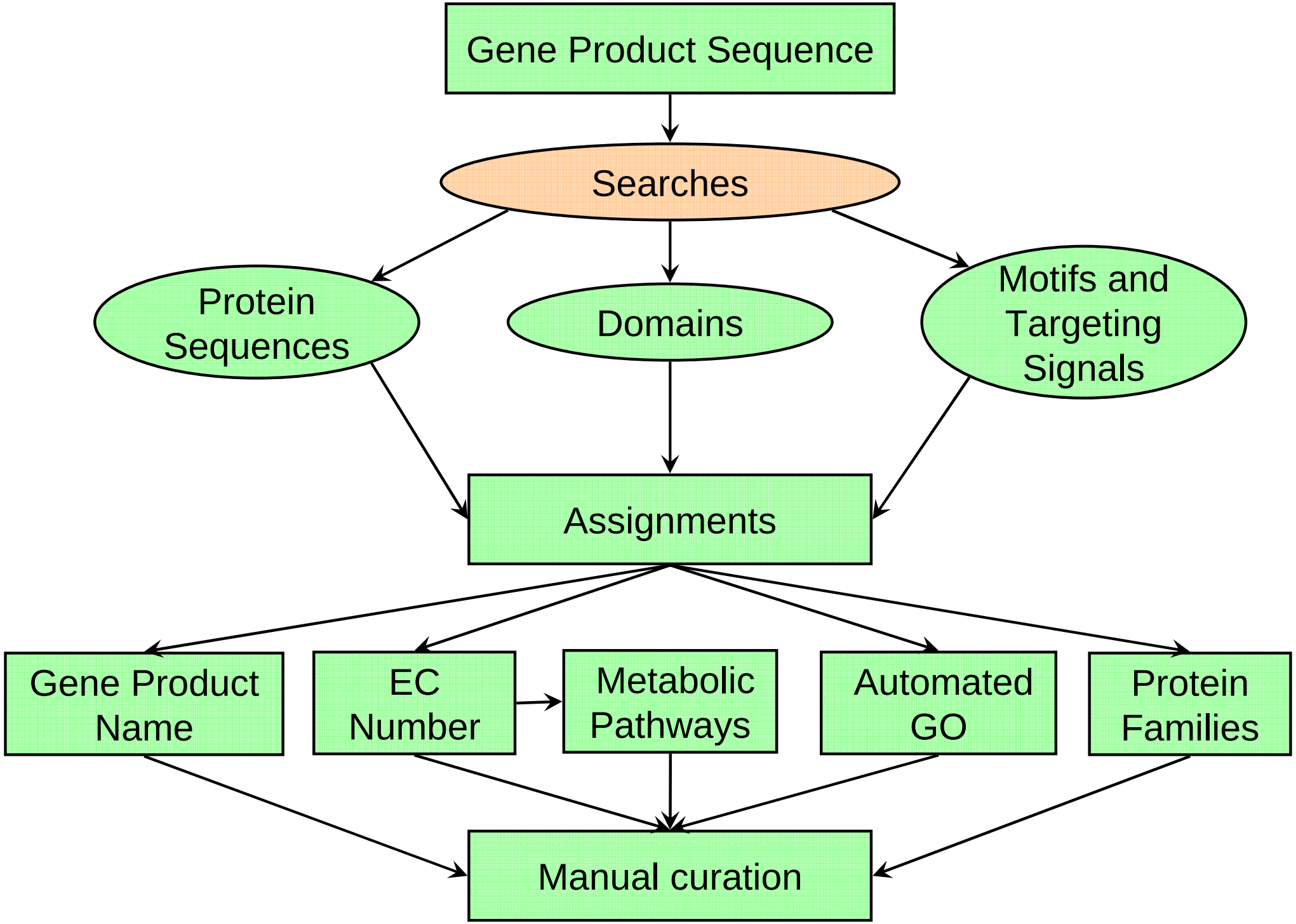
- What is functional annotation?
- Steps we take to annotate genes.
- Software tools used for functional annotation.

Functional Annotation and its Goals

- Functional annotation in genomics is about classifying and attributing the identified structural elements.
- Goals
  - Assigning names of gene products.
  - Interpreting functions of genes within an organism if possible.
  - Classifying the proteome into protein families.
  - Identifying the enzymes and assigning the EC numbers in an organism.
  - Assigning Gene Ontology (GO) terms for genes.
  - Generating Metabolic pathways for an organism.
  - Important to know what software tools we need and how to use them for optimum result.

Steps in Functional Annotation

- Analyze the gene structure for accuracy.
- Extract the gene product sequence.
- Search the sequences through various software tools of different algorithms against different database sources.
- Optimize the parameters of the tools for efficient annotation.
- Critically evaluate the computationally derived annotation.
- Maintain and display the annotation data.

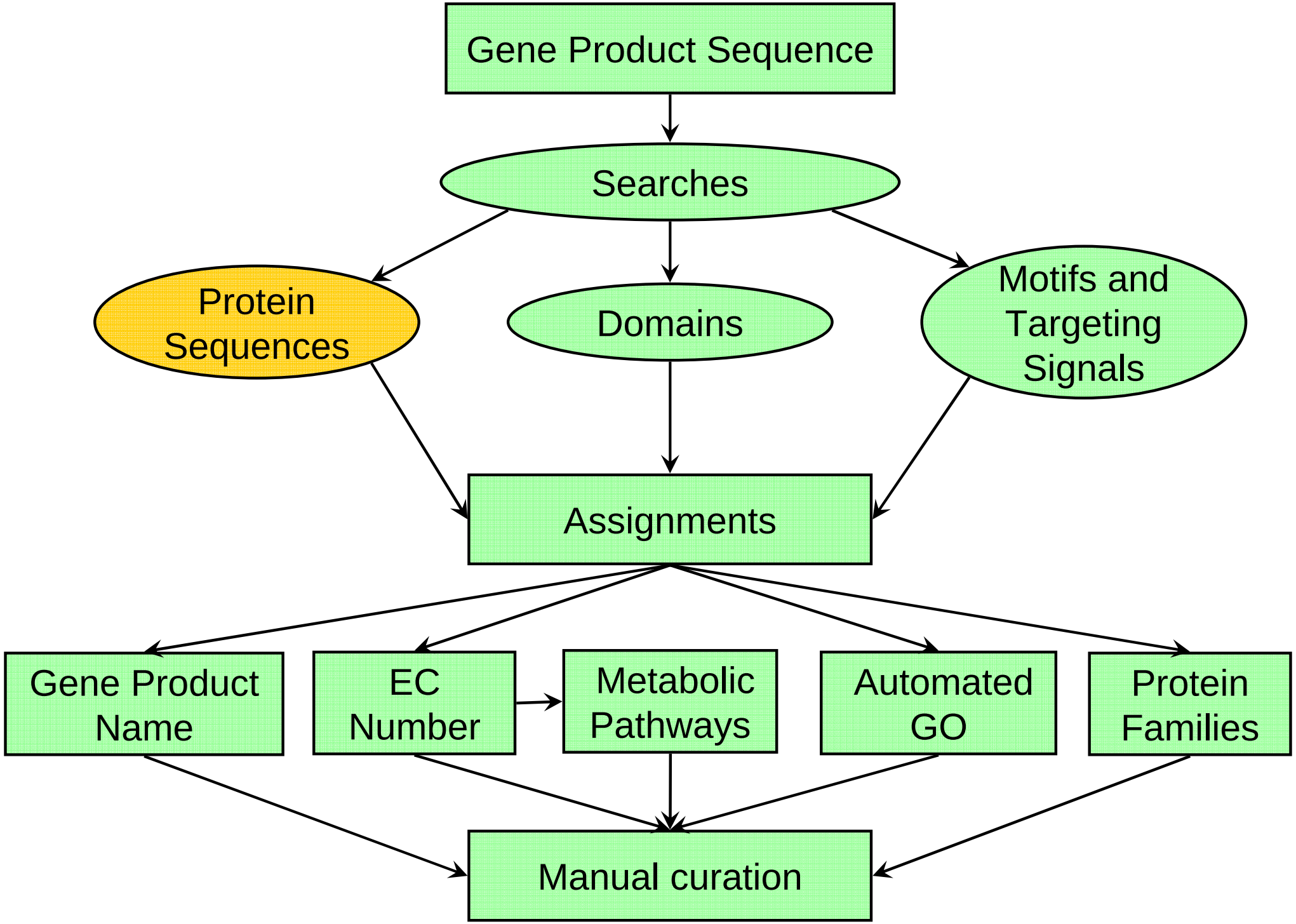

Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

Basic Searches to Run

- BLAST (nucleotide or protein homology)
- TmHMM (Transmembrane domains)
- SignalP (signal peptide cleavage sites)
- TargetP (subcellular location)
- HMMer or SAM (searches using statistical descriptions)
  - Pfam (database of protein families and HMMs)
  - TIGRFAMS (protein family based HMMs)
  - SCOP (Structural domains)
- CDD (NCBI's Conserved Domain Database)
- Prosite (biologically significant sites, patterns and profiles)
- Interpro (protein families, domains and functional sites)
- Others as needed

Automated Searches Through Pipeline

- The searches are run as a pipeline using Ergatis, a workflow system.
- The searches can be run as a pipeline by stringing the different searches together using perl scripts.
- Ergatis: http://ergatis.sourceforge.net/


Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

Primary Search Tool

- BLAST - **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Finds regions of local similarity between sequences by pair-wise alignment.
  - Compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.
  - Used to infer functional and evolutionary relationships between sequences.

Blast Programs

- Different blast programs are available for the BLAST software:
  - **blastp**: compares an amino acid query sequence against a protein sequence database.
  - **blastn**: compares a nucleotide query sequence against a nucleotide sequence database.
  - **blastx**: compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
  - **tblastn**: compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
  - **tblastx**: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. It is computationally intensive.
- Which BLAST program to use depends on the query and reference sequences.

Web BLAST

- NCBI Blast http://www.ncbi.nlm.nih.gov/blast/
- WU blast http://genome.wustl.edu/tools/blast/
- Uniprot-swissprot blast http://www.expasy.org/srs5/
- JCVI Projects http://www.tigr.org/tdb/euk/ (follow links to each database)
- The Gene Indices http://compbio.dfci.harvard.edu/tgi/
- Sanger projects http://www.sanger.ac.uk/DataSearch/
- TAIR - http://www.arabidopsis.org/Blast/index.jsp
- plasmoDB- http://www.plasmodb.org/plasmo/
- toxoDB: http://www.toxodb.org/toxo
- Flybase- http://flybase.bio.indiana.edu/
- VectorBase- http://www.vectorbase.org/index.php

Command-Line BLAST

- Downloading the program (executables and the databases)
  - WU BLAST 2.0 (http://blast.wustl.edu/)
    - Options:
      - blastp
      - blastn
      - blastx …
    - Formatting the database - “xdformat –n”
    - New version of WU blast: AB-BLAST (http://www.advbiocomp.com/)
  - NCBI BLAST (http://www.ncbi.nlm.nih.gov/blast/download.shtml)
    - Formatting the database - blastall –p blastp(n/x)
    - Formatting the database - “formatdb db_name”

Blast Parameters

- Database: -d
  - Multi-fasta format
  - Indexed database
    - Wu-blast:
      - "xdformat –n"
    - NCBI blast: "formatdb db_name"
  - Database commonly used:
    - AllGroup.nraa
    - Kingdom or species specific (protein, nucleotide, and genome)
    - customized
- Query sequences: -i
  - Single fasta (or multi-fasta) file
  - Clean unknown sequence characters

BLAST Parameters

- -e: e value, significance of a match
  - value = the chance of being a random hit (match by chance)
  - Default = 10 (NCBI blastall 2.2.15)
  - Commonly use 1e-20 for nucleotide; 1e-5 for peptides
  - What affects the e value:
    - Short sequences or domain hits may have a higher e value even when the hits are significant.
    - Database size

E Value Example

```
Query=  JCVI_SCAF_1101667073875_103_656_orientation=reverse
        (184 letters)

Database:  /local/scratch/pandaadm/panda_run/syncpanda/032107/group/BacterialGr
           oup/BacterialGroup.niaa
           1,410,935 sequences; 450,201,350 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)       N

GB|ABI87029.1|115281512|CP000440 methyltransferase FkbM f...   442  4.1e-41    1
SP|Q28JB2|Y4284_JANSC Putative methyltransferase Jann_428...   278  9.8e-24    1


Query=  JCVI_SCAF_1101667073875_103_656_orientation=reverse
        (184 letters)

Database:  /local/scratch/pandaadm/panda_run/syncpanda/032107/group/AllGroup/Al
           lGroup.niaa
           3,493,351 sequences; 1,215,183,409 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)       N

RF|YP_773363.1|115351524|NC_008390 methyltransferase FkbM...   442  1.0e-40    1
SP|Q28JB2|Y4284_JANSC Putative methyltransferase Jann_428...   278  2.4e-23    1
```

E-Value Example

```
Query=  JCVI_SCAF_1101667073875_103_656_orientation=reverse
        (159 letters)

Database:  /local/scratch/pandaadm/panda_run/syncpanda/032107/group/BacterialGr
           oup/BacterialGroup.niaa
           1,410,935 sequences; 450,201,350 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                  Smallest
                                                                    Sum
                                                          High  Probability
Sequences producing High-scoring Segment Pairs:           Score   P(N)       N

GB|ABI87029.1|115281512|CP000440 methyltransferase FkbM f...   385   4.5e-35   1
SP|Q28JB2|Y4284_JANSC Putative methyltransferase Jann_428...   222   8.4e-18   1


Query=  JCVI_SCAF_1101667073875_103_656_orientation=reverse
        (159 letters)

Database:  /local/scratch/pandaadm/panda_run/syncpanda/032107/group/AllGroup/Al
           lGroup.niaa
           3,493,351 sequences; 1,215,183,409 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                  Smallest
                                                                    Sum
                                                          High  Probability
Sequences producing High-scoring Segment Pairs:           Score   P(N)       N

RF|YP_773363.1|115351524|NC_008390 methyltransferase FkbM...   385   1.1e-34   1
SP|Q28JB2|Y4284_JANSC Putative methyltransferase Jann_428...   222   2.1e-17   1
```

BLAST Parameters

- -F: mask low complexity sequences (DUST with blastn or SEG with others).
  - Low complexity sequence hits may mislead the users to give false functional assignment.
- -M: Matrix. Default = BLOSUM62
  - BLOSUM62 is among the best of the available matrices for detecting weak protein similarities.
  - The BLOSUM matrix assigns a probability score for each position in an alignment that is based on the frequency with which that substitution is known to occur among consensus blocks within related proteins.
  - Other supported options include PAM30, PAM70, BLOSUM80, and BLOSUM45.

BLAST Parameters

- -v: number of one-line descriptions
- -b: number of alignments to show
- -W: default is 11 for blastn, 3 for other programs.
- -m: 8 tabular output

BLAST Command

- A typical and simple blast command at unix prompt

>blastp -d /path/AllGroup.niaa -i /path/query_seq.fasta -v 3 -b 3 -e 1e-20 -o blastp_query_vs_db.out

-d: AllGroup.niaa as db

-i: query_seq.fasta as query

-v: one-line description for top 3 hits

-b: pair-wise alignment for top 3 hits

-e: 1e-20

-o: out file named “blastp_query_vs_db.out”

BLAST Tabular Output

- To get the tabular file for each blast hit above the cutoff

  >btab blast_output_file (NCBI blast)

  example: >btab blastp.out

  >BPbtab blast_output_file (wu-blast)

⇒ “blastp.out.btab” will be automatically generated

- Columns in the btab file:

```
13:07 Thu_Jun14: pfafoo_dir :>more pfa_pexel_part2.out.btab
Plasmodium_falciparum_3D7|MAL10|PF10_0003|Pf/1-345   179  BLASTP  pexel_pfa1.pep  2296.m00364  1   147  1   147  85.7  85.7  633  248  0 null  345  5e-67
Plasmodium_falciparum_3D7|MAL10|PF10_0004|Pf/1-371   179  BLASTP  pexel_pfa1.pep  2296.m00363  1   170  1   170  100.0 100.0 901  351  0 null  371  5e-98
Plasmodium_falciparum_3D7|MAL10|PF10_0005|Pf/1-407   179  BLASTP  pexel_pfa1.pep  2296.m00362  1   179  1   179  87.7  87.7  819  320  0 null  407  1e-88
Plasmodium_falciparum_3D7|MAL10|PF10_0006|Pf/1-362   179  BLASTP  pexel_pfa1.pep  2296.m00361  1   179  1   179  54.2  54.2  429  169  0 null  361  2e-43
Plasmodium_falciparum_3D7|MAL10|PF10_0008|Pf/1-98    98   BLASTP  pexel_pfa1.pep  2296.m00359  1   98   1   98   79.6  79.6  390  154  0 null  98   2e-39
Plasmodium_falciparum_3D7|MAL10|PF10_0014|Pf/1-194   179  BLASTP  pexel_pfa1.pep  2296.m00354  23  179  23  179  91.7  91.7  758  296  0 null  193  2e-81
Plasmodium_falciparum_3D7|MAL10|PF10_0017|Pf/1-284   179  BLASTP  pexel_pfa1.pep  2296.m00351  1   179  1   173  78.2  78.2  657  257  0 null  278  9e-70
Plasmodium_falciparum_3D7|MAL10|PF10_0018|Pf/1-921   179  BLASTP  pexel_pfa1.pep  2296.m00350  1   179  1   179  68.2  68.2  594  233  0 null  921  2e-62
Plasmodium_falciparum_3D7|MAL10|PF10_0020|Pf/1-763   179  BLASTP  pexel_pfa1.pep  2296.m00348  1   179  1   179  100.0 100.0 927  361  0 null  763  1e-101
Plasmodium_falciparum_3D7|MAL10|PF10_0021|Pf/1-266   179  BLASTP  pexel_pfa1.pep  2296.m00347  1   179  1   179  89.4  89.4  833  325  0 null  266  4e-90
Plasmodium_falciparum_3D7|MAL10|PF10_0023|Pf/1-276   179  BLASTP  pexel_pfa1.pep  2296.m00345  1   179  1   179  93.3  93.3  868  338  0 null  276  3e-94
Plasmodium_falciparum_3D7|MAL10|PF10_0024|Pf/1-469   179  BLASTP  pexel_pfa1.pep  2296.m00344  1   179  1   179  94.4  94.4  889  347  0 null  469  1e-96
Plasmodium_falciparum_3D7|MAL10|PF10_0025|Pf/1-631   179  BLASTP  pexel_pfa1.pep  2296.m00343  18  179  18  179  92.6  92.6  760  297  0 null  631  1e-81
Plasmodium_falciparum_3D7|MAL10|PF10_0027|Pf/1-427   179  BLASTP  pexel_pfa1.pep  2296.m00392  1   179  1   179  100.0 100.0 963  375  0 null  427  1e-105
```

1-5: query identifier, query length, Program, database, db_subject identifier,

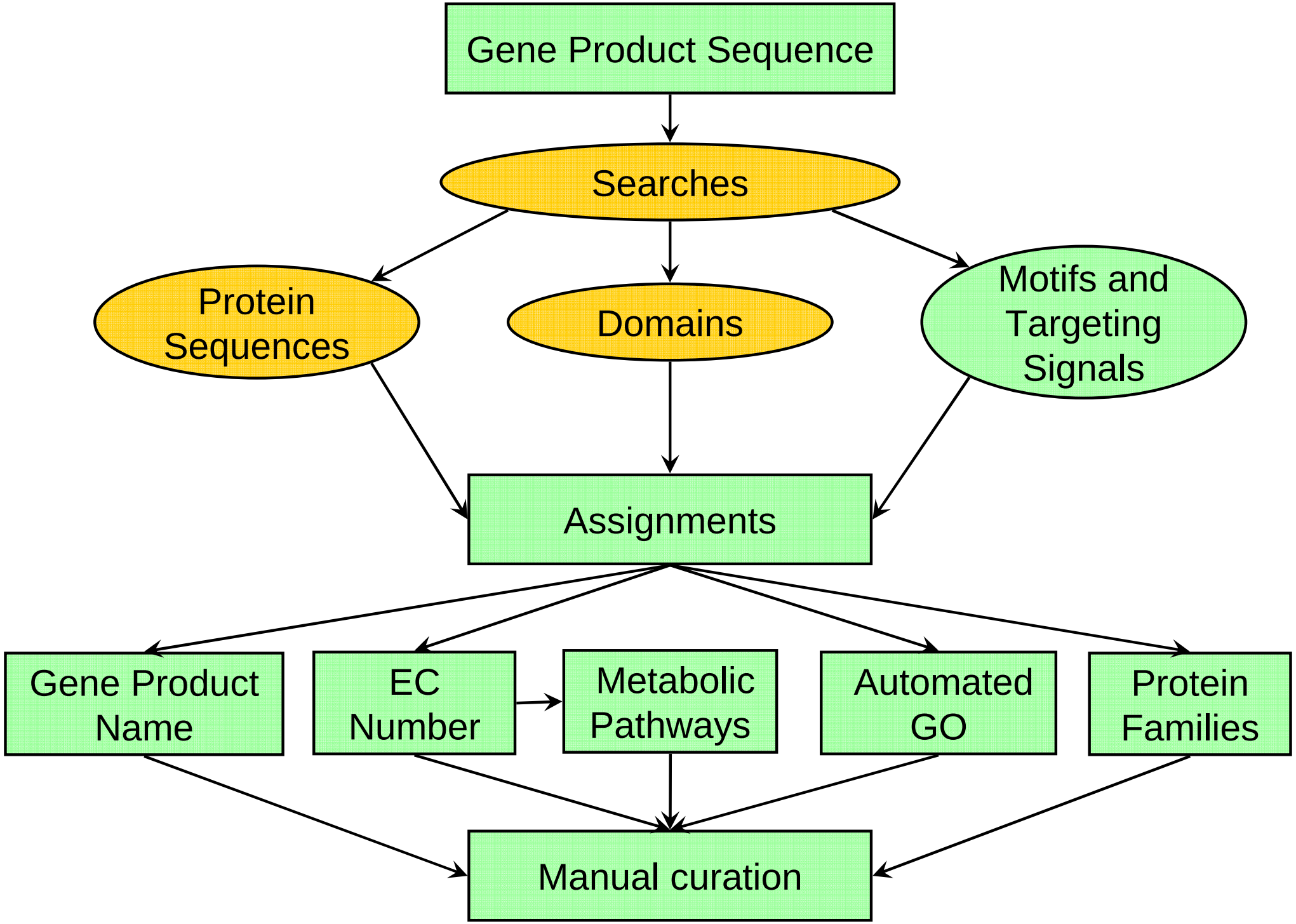
6-9: query hits coordinates, subject hits coordinates

10-12: identities, positives (similarity), blast score

13-14: position in blast output file itself

15-16: frame, strand

17-18: subject length, e-values


Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

Domain Searches

- Tools to use for domain searches:
  - HMMER, HMMPFAM: http://hmmer.janelia.org/
  - SAM (Sequence Alignment and Modeling System) http://compbio.soe.ucsc.edu/sam.html
- Databases of domains to search:
  - Pfam: http://www.sanger.ac.uk/Software/Pfam/
  - TIGRFAMs: http://www.jcvi.org/cms/research/projects/tigrfams/overview/
  - SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/

Pfam HMM Searches

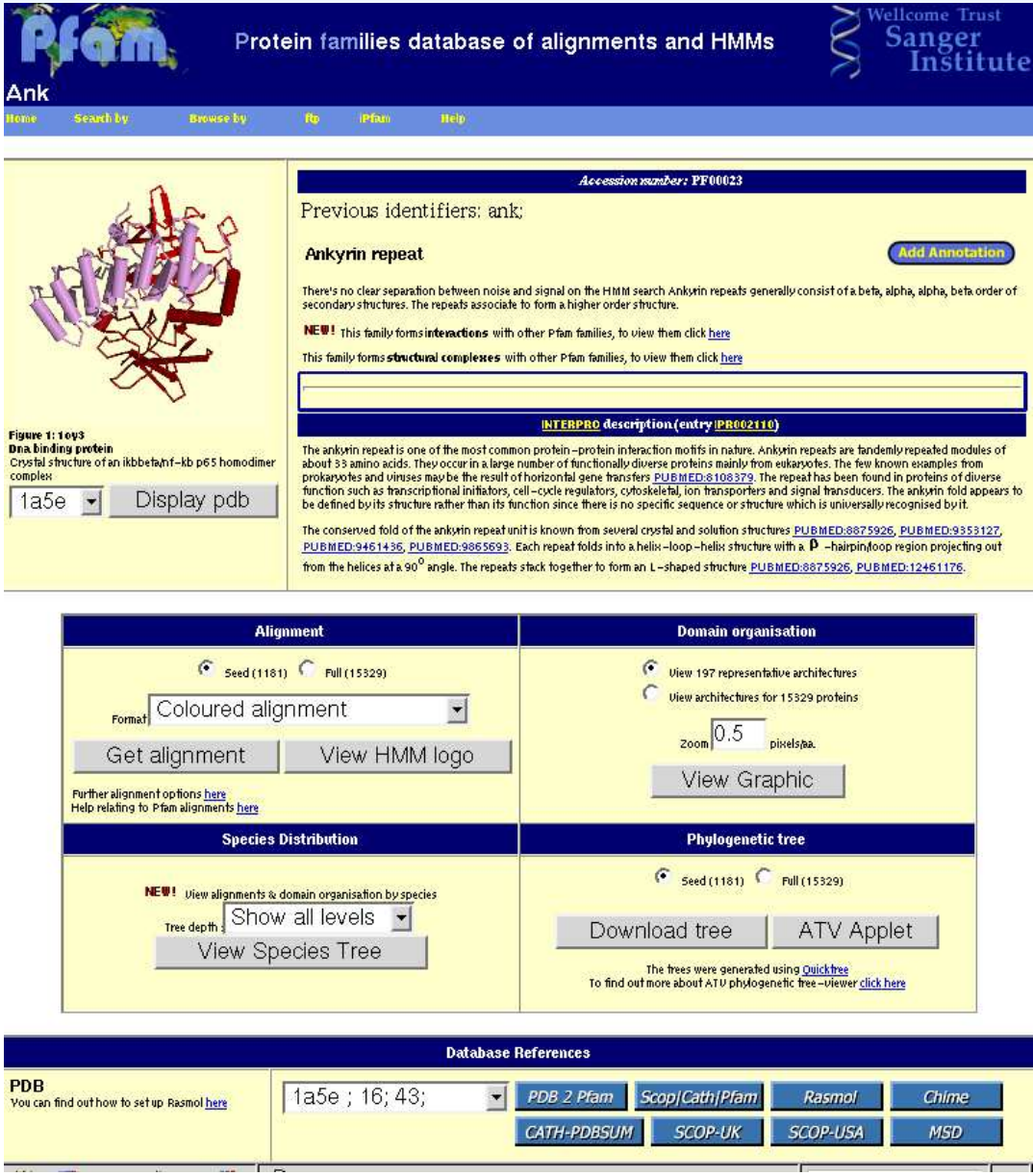
- **HMMer** is a software suite for making and using Hidden Markov Models (HMMs) of biological sequences.
- Hidden Markov models (HMMs) are statistical models produced from a multiple sequence alignment of a sequence family which should be the functionally significant segments of the sequence.
- Protein sequences are searched against the Pfam HMM databases to identify the proteins with important domain conservation above significant score.

Pfam

http://www.sanger.ac.uk/Software/Pfam/

For each family in Pfam you can:

- Look at multiple alignments.
- View protein domain architectures.
- Examine species distribution.
- Follow links to other databases.
- View known protein structures.



TIGRfams

http://www.jcvi.org/cms/research/projects/tigrfams/overview/

TIGRFAMs: A collection of protein families featuring curated multiple sequence alignments, Hidden Markov Models (HMMs) and associated information designed to support the automated functional identification of proteins by sequence homology.

You can use the TIGRfam page to see

- The curated seed alignment for each TIGRFAM.
- The full alignment of all family members.
- The cutoff scores for inclusion in each of the TIGRfams.

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

tigr fams
tigr protein families

TIGR FAMs Page | Text Search | HMM Search | Download | CMR Home | Help

HMM Profile Page
Accession #: TIGR00262 Name: trpA

Both TIGRFAMs and Pfams are displayed on this page. TIGRFAMs and Pfams are based on Hidden Mark
Models or HMMs. An HMM is a statistical model for any system that can be represented as a succession of
transitions between discrete states. Scores are reported both in bits of information and as an E-value. See
below for more information on this TIGRFAM or Pfam and its HMM.

trpA Information: See below for detailed information on this family, including the cutoff score for inclusion i
this family and the average score of genes/proteins in this family. To view all genes with the same EC
number, click on the **EC Number** link. To view more information on the Role Category for this family, click o
the **Role Category** link.

Accession:	TIGR00262	**Name:**	trpA	**Isology Type:**	equivalog
Common Name:	tryptophan synthase, alpha subunit				
Noise Cutoff:	-5.00	**Trusted Cutoff:**	2.50	**Avg. Score:**	391.56 +/- 24.74
EC Number:	4.2.1.20	**HMM length:**	262		
Relationship:	**InterPro assignment:** IPR002028				
Role Category:	**Mainrole:** Amino acid biosynthesis			**Subrole:**	Aromatic amino acid family
Gene Ontology (GO) Terms:	GO:0000162	process	tryptophan biosynthesis		
	GO:0004834	function	tryptophan synthase activity		
Author(s):	Loftus BJ	**Created:**	Apr 20 1999 2:09PM	**Last Modified:**	Sep 23 2003 4:23PM

Command-line HMM Searches

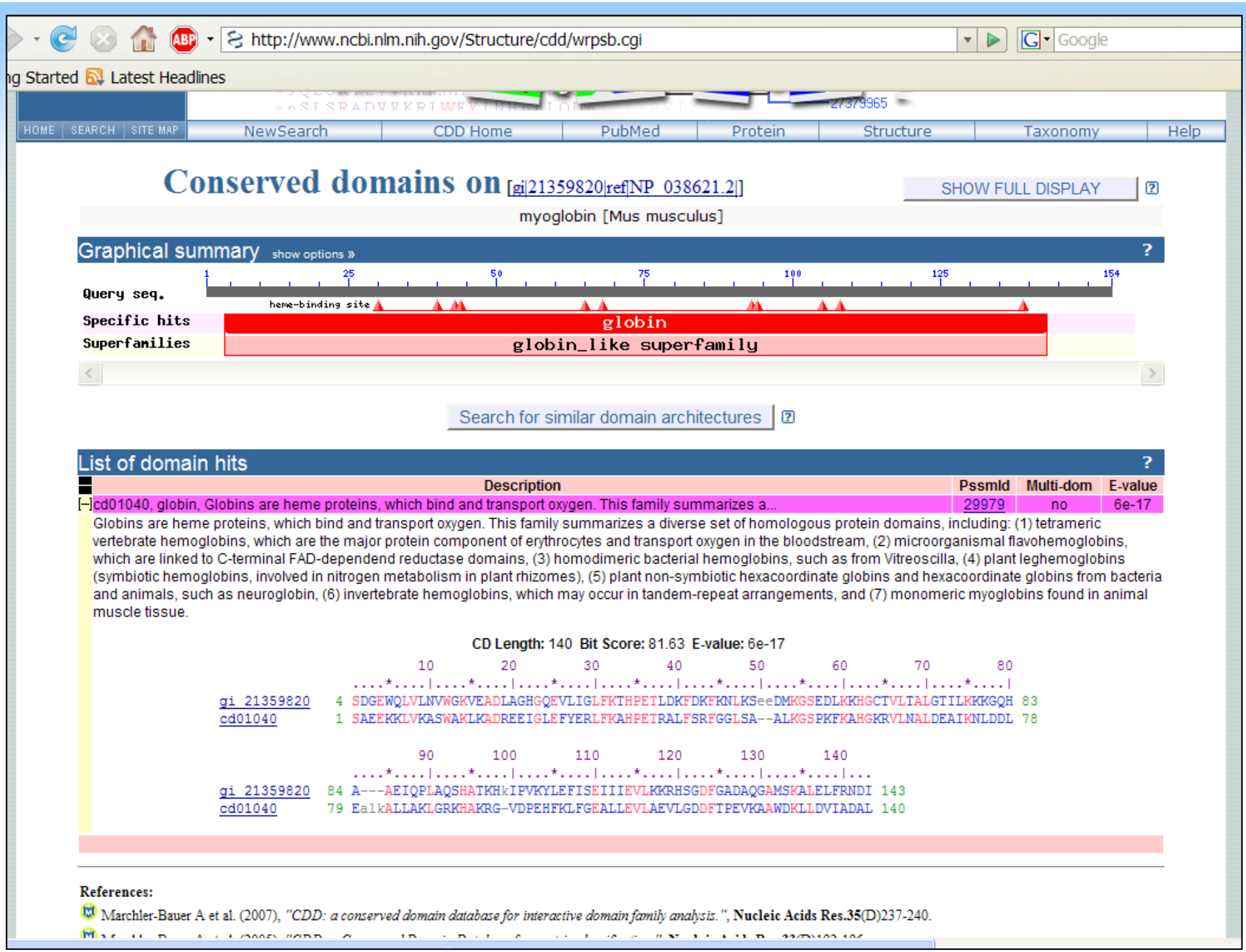
- Download HMMER software package which includes the hmm search program.
- Download Pfam and TIGRFAM databases.
- Process the protein sequences against the desired dataset of HMMs.
- Command:
  - Hmmpfam query_file(protein sequences) database_file(HMM) > output_file
  - Easy to generate tab delimited file from the hmm output.

Conserved Domain Database (CDD)

- NCBI's Conserved Domain Database(CDD) is a collection of multiple sequence alignments for ancient domains and full-length proteins.
- CDD contains protein domain models imported from outside sources, such as Pfam, SMART, COGs (Clusters of Orthologous Groups of proteins), PRK (PRotein K(c)lusters), and are curated at NCBI.
- Models are linked to other NCBI databases, including protein sequences, known structures, taxonomy and hierarchical classification.

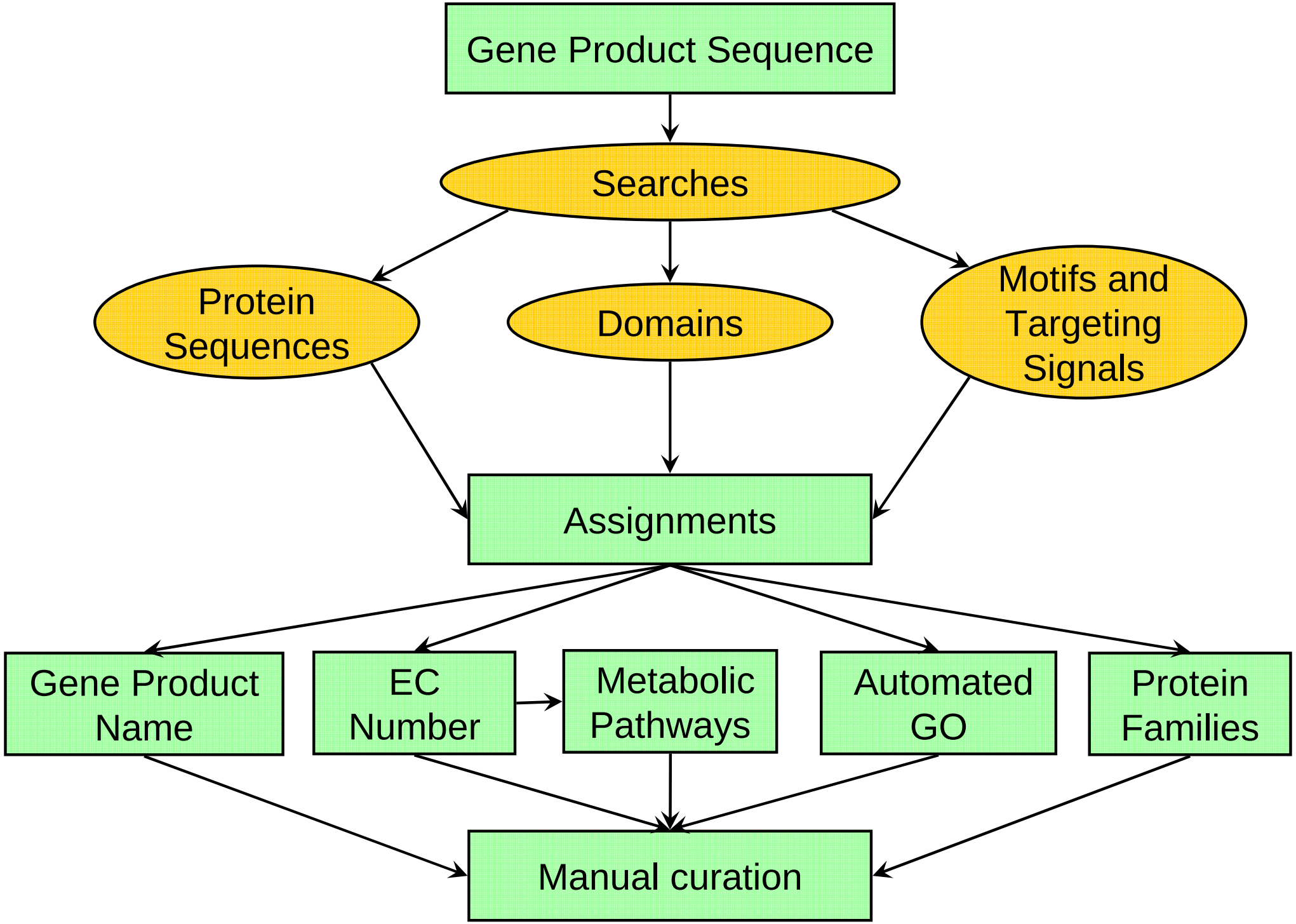
CD-Search

- The CD-Search service is a web-based tool for the detection of conserved domains in protein sequences.
- The CD-Search service uses RPS-BLAST to compare a query protein sequence against conserved domain databases.
- You can have graphical display as output.
- http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi


http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Google
g Started
Latest Headlines
HOME
SEARCH
SITE MAP
NewSearch
CDD Home
PubMed
Protein
Structure
Taxonomy
Help
Conserved domains on [gi|21359820|ref|NP_038621.2|]
SHOW FULL DISPLAY
myoglobin [Mus musculus]
Graphical summary
show options »
Query seq.
heme-binding site
Specific hits
globin
Superfamilies
globin_like superfamily
Search for similar domain architectures
List of domain hits
Description
Pssmid
Multi-dom
E-value
cd01040, globin, Globins are heme proteins, which bind and transport oxygen. This family summarizes a...
29979
no
6e-17
Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.
CD Length: 140 Bit Score: 81.63 E-value: 6e-17
gi 21359820 4 SDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSeeDMKGSEDLKKHGCTVLTALGTILKKKGQH 83
cd01040 1 SAEEKKLVKASWAKLKADREEIGLEFYERLFKAHPETRALFSRFGGLSA--ALKGSPKFKAHGKRVLNALDEAIKNLDDL 78
gi 21359820 84 A---AEIQPLAQSHATKHkIPVKYLEFISEIIIEVLKKRHSGDFGADAQGAMSKALELFRNDI 143
cd01040 79 EalkALLAKLGRKHAKRG-VDPEHFKLFGEALLEVLAEVLGDDFTPEVKAAWDKLLDVIADAL 140
References:
Marchler-Bauer A et al. (2007), "CDD: a conserved domain database for interactive domain family analysis.", Nucleic Acids Res.35(D)237-240.

Command-line CDD Search

- We can run RPS-BLAST locally on command line.
- A standalone version of RPS-BLAST can be downloaded: ftp://ftp.ncbi.nih.gov/toolbox).
- Pre-formatted search databases, which have already been processed by Formatrpsdb, are available on the CDD FTP site for download.


Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

Motif Searches

- SignalP: http://www.cbs.dtu.dk/services/SignalP
- TargetP: http://www.cbs.dtu.dk/services/TargetP
- Prosite: http://expasy.org/tools/scanprosite
- Interpro: http://ebi.ac.uk.interpro
- tmHMM: http://www.cbs.dtu.dk/services/TMHMM/

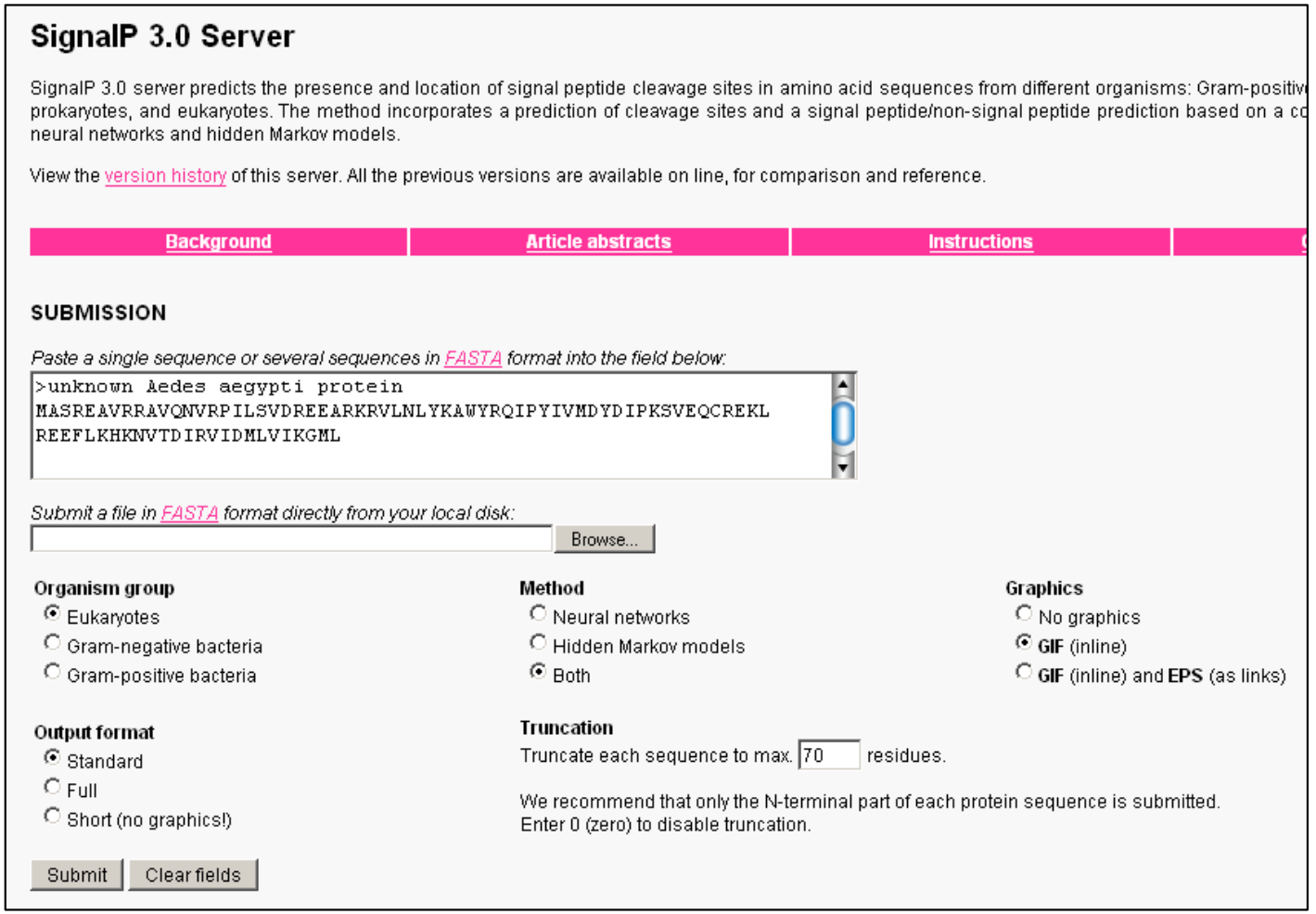
All the above programs offer the downloadable software packages and the required dataset which we can used as command-line to streamline the search result for the desired output.

SignalP

SignalP predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms.

http://www.cbs.dtu.dk/services/SignalP/

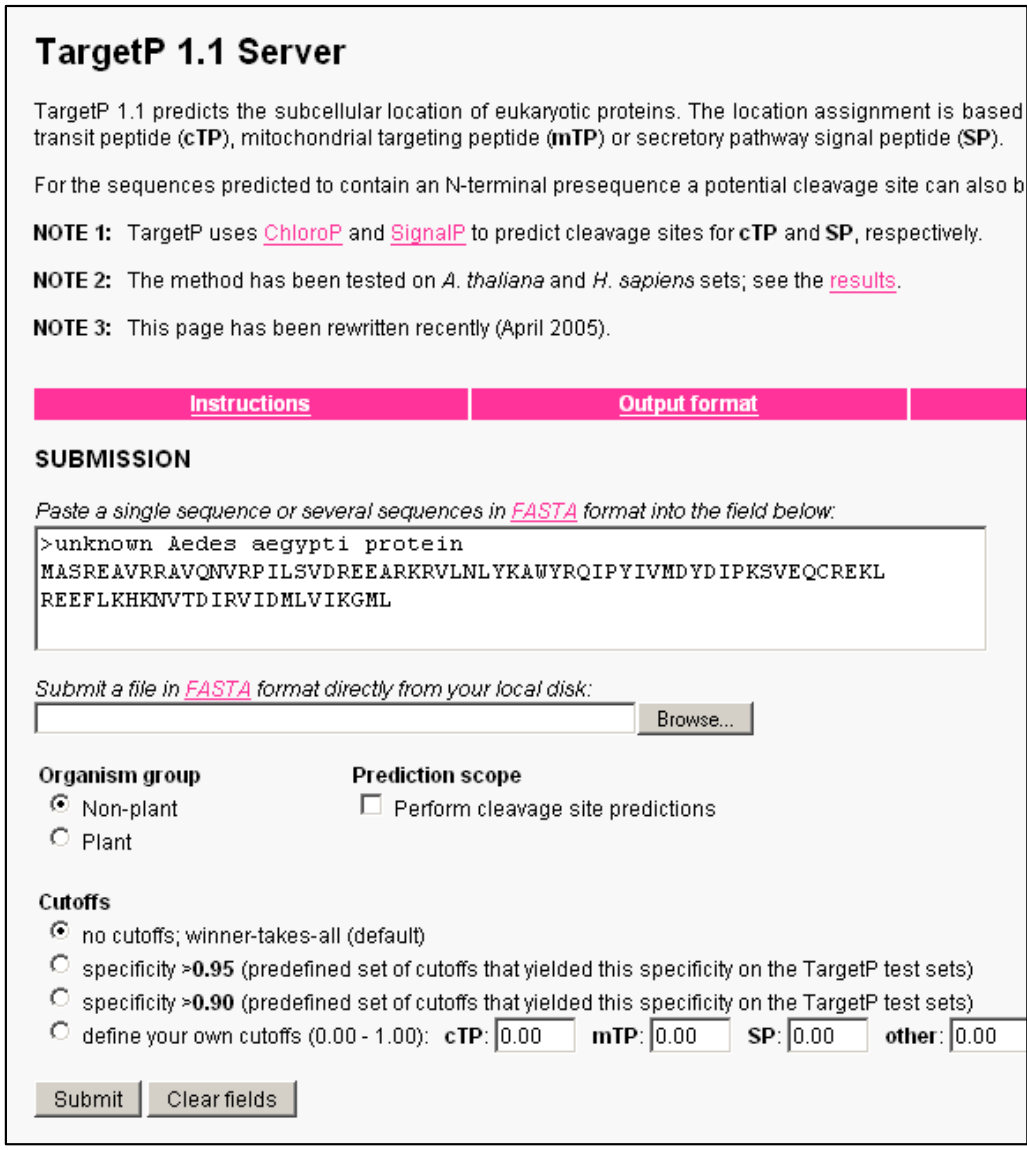
- The method is based on a combination of artificial neural networks and hidden Markov models.



TargetP

- TargetP predicts the subcellular location of eukaryotic proteins.
- The location assignment is based on the predicted presence of any of the N-terminal presequences:
  - chloroplast transit peptide (**cTP)**
  - mitochondrial targeting peptide **(mTP)**
  - secretory pathway signal peptide **(SP).**

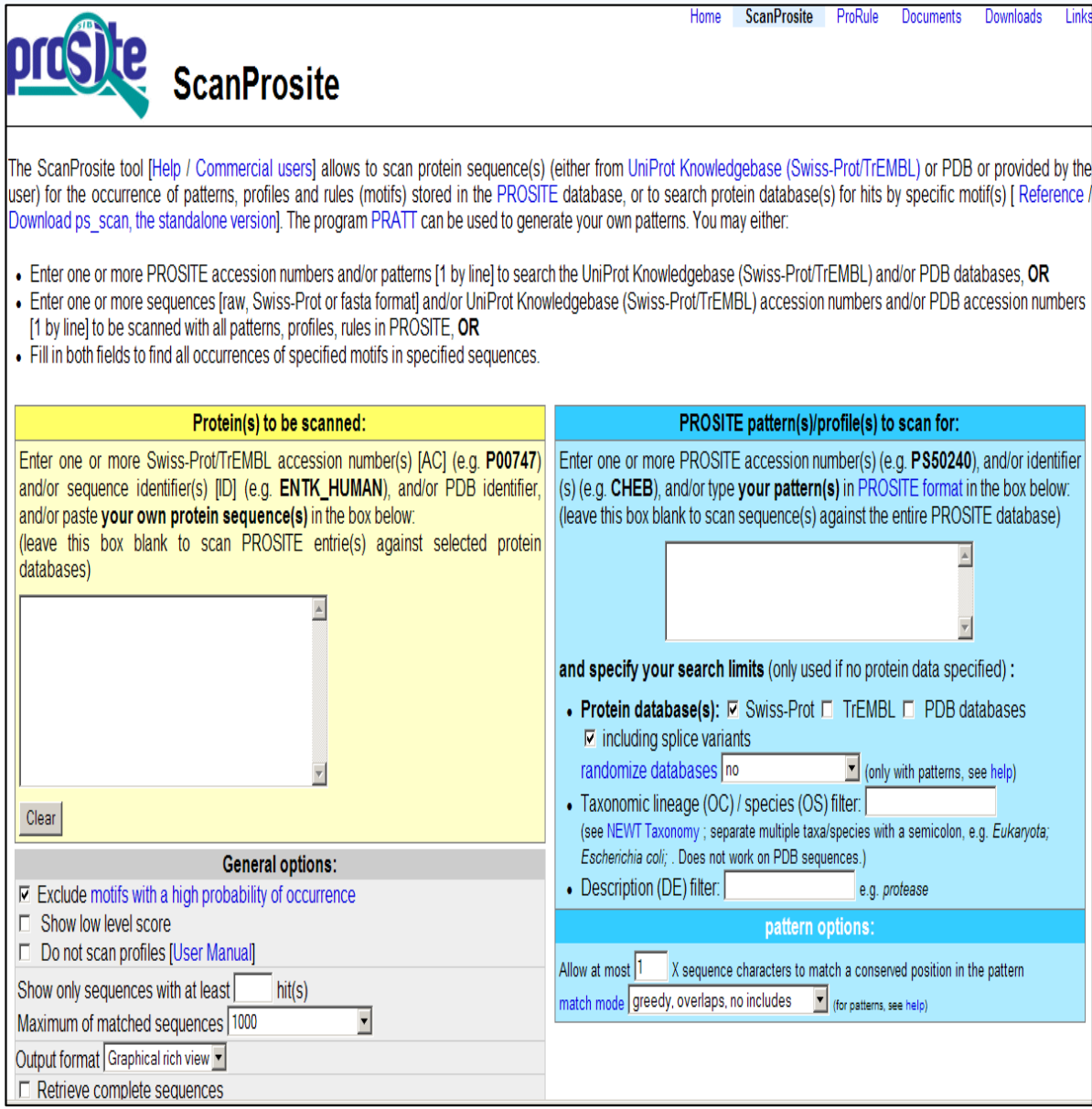
http://www.cbs.dtu.dk/services/TargetP/



Prosite

http://expasy.org/tools/scanprosite/

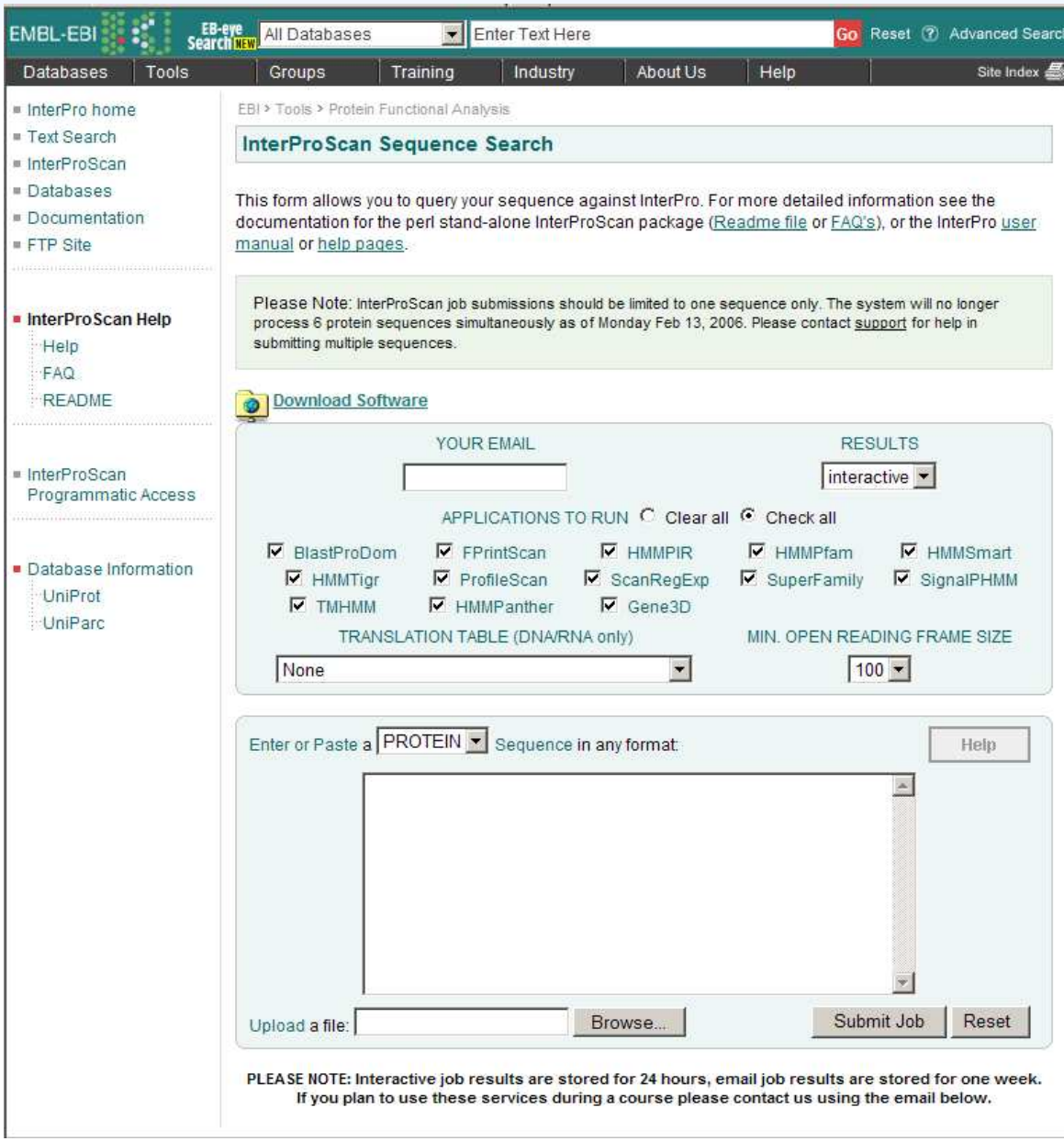
- PROSITE Database consists of protein domains, families, functional sites, as well as patterns and profiles.
- ScanProsite identifies the occurrence of patterns, profiles stored in the Prosite database.

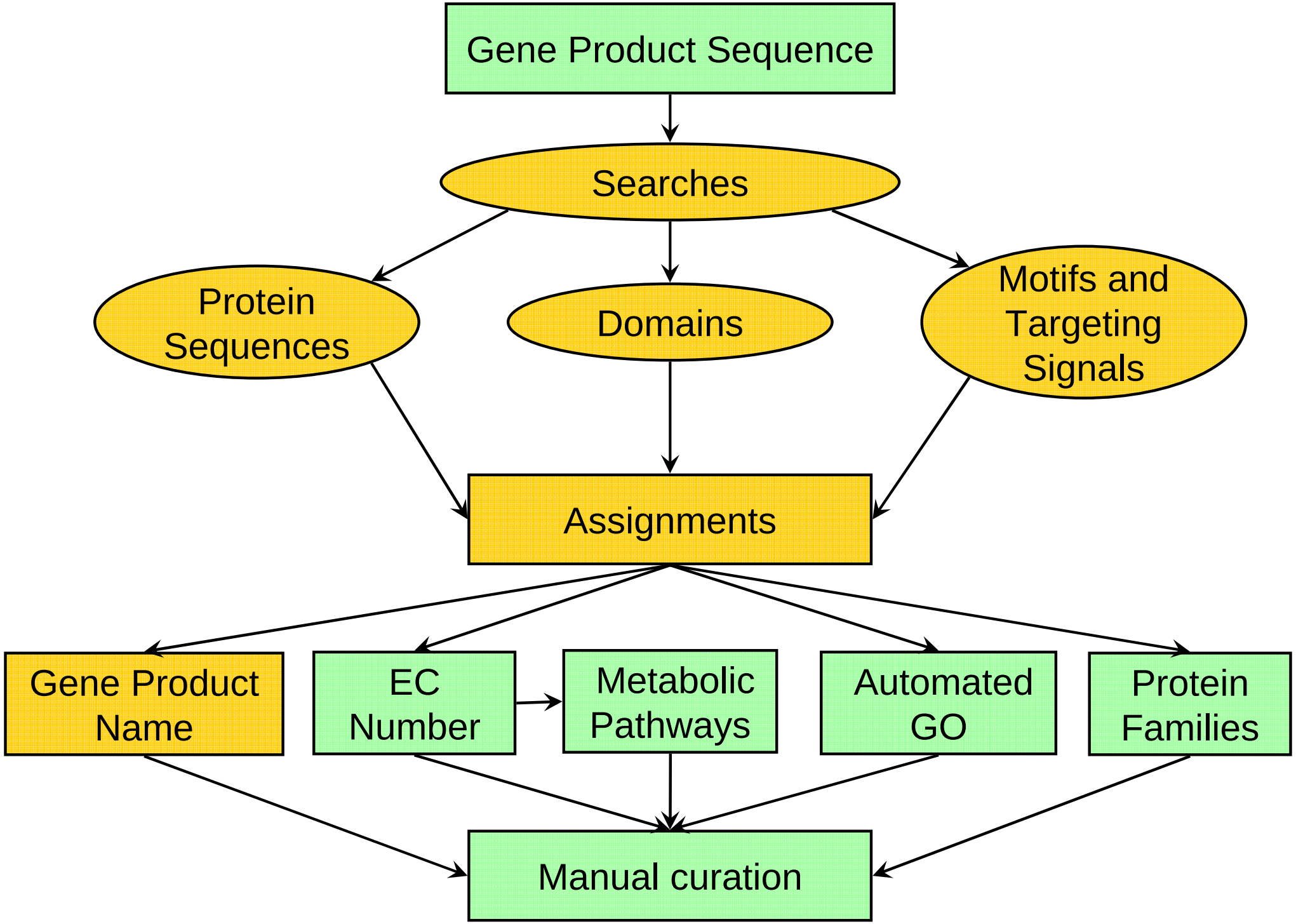


Interpro

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

- Database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.




Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

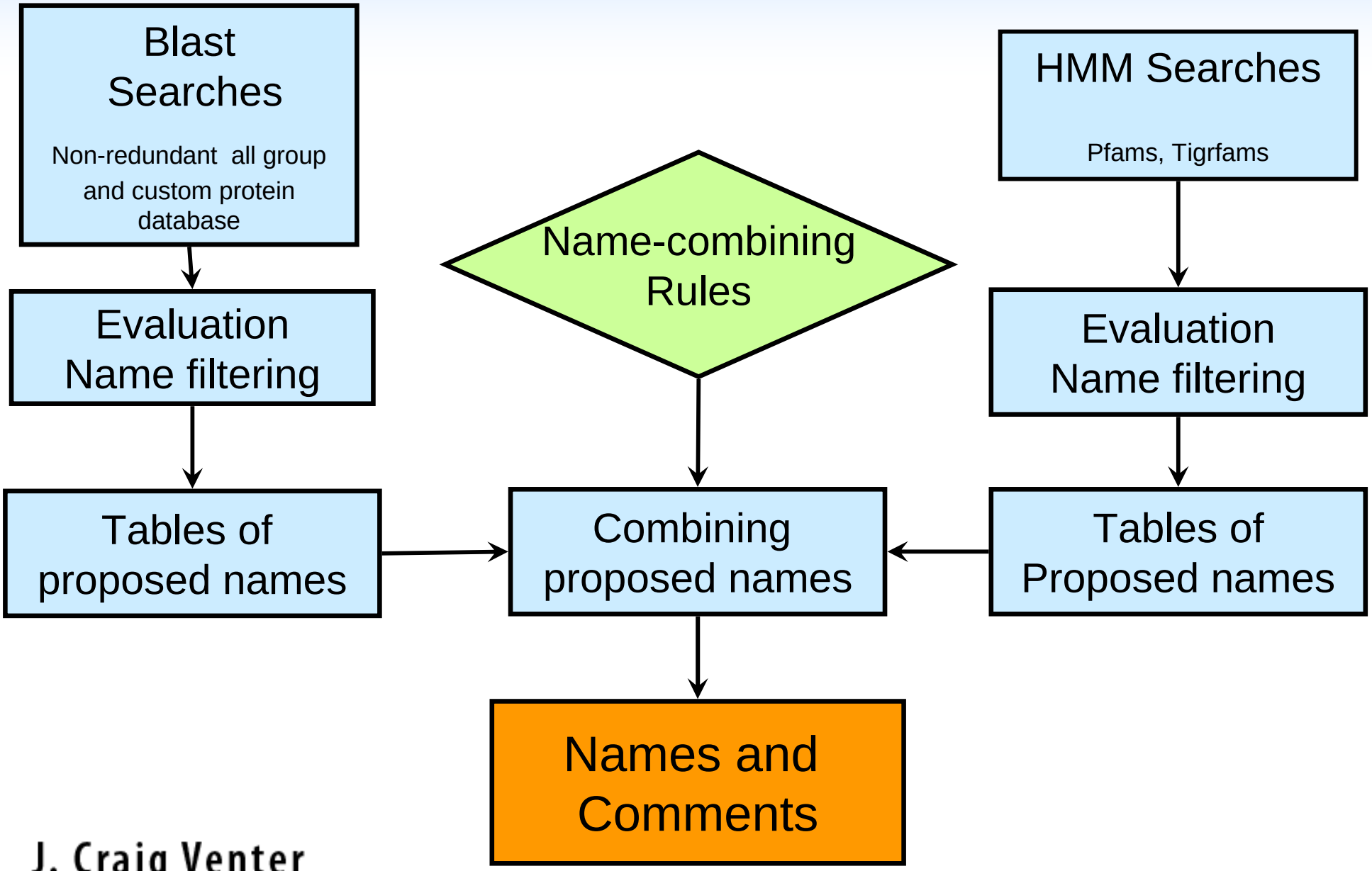
Naming the Gene Product

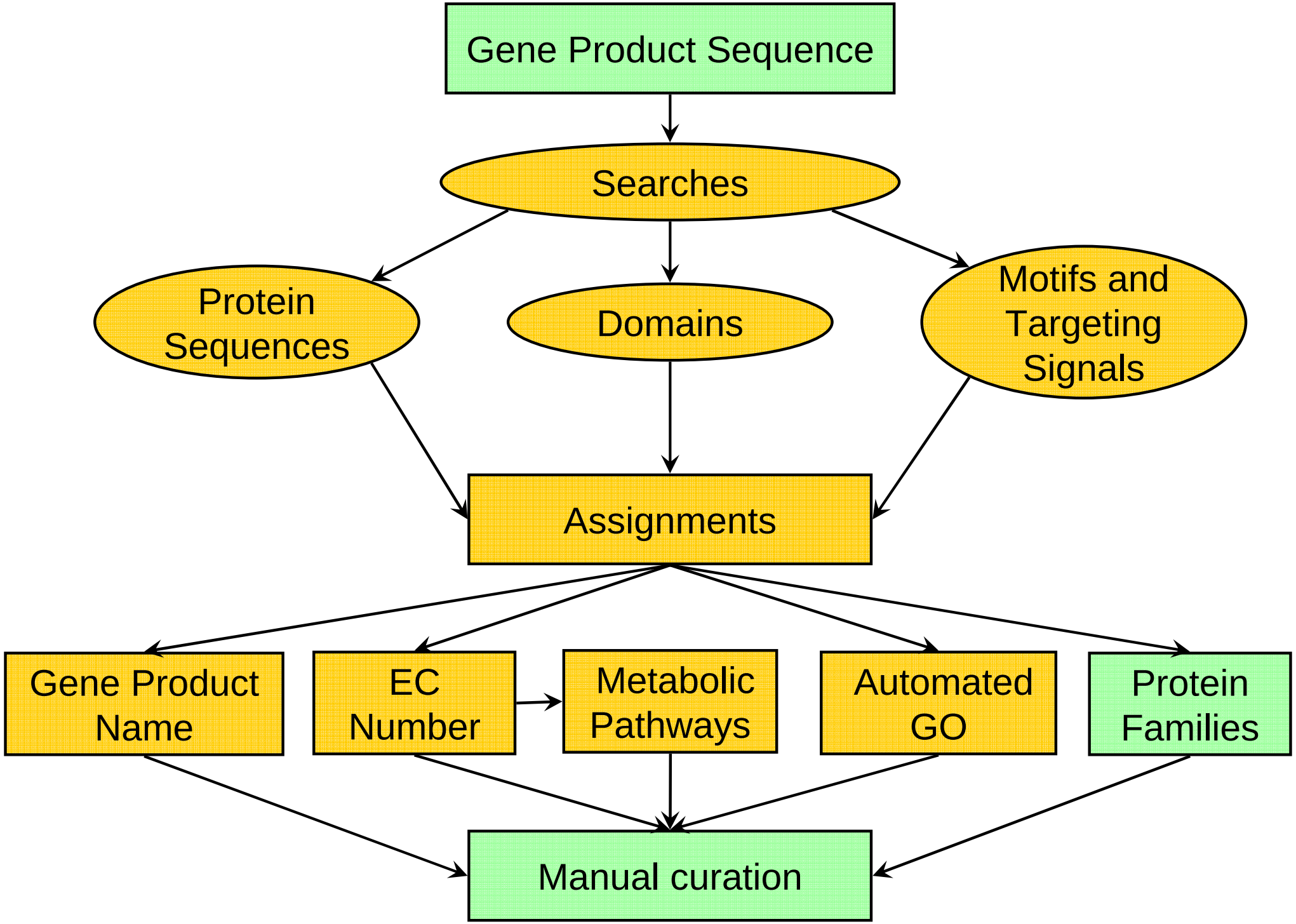
- Assigning a functional name to a gene product in an automated fashion is still one of the major challenges in eukaryotic annotation.
- Protein databases have very heterogeneous formats for sequence headers.
- The number of database with curated protein names are small.
- Many gene names are the result of circular annotation of genome projects.
- Several analysis tools provide information in comments, but these cannot be easily parsed into meaningful names.
- Critical evaluation of names is fundamental for functional annotation.

Potential Sources for Names and/or Comments

- Curated protein database which has experimental evidence (literature).
- Blast searches against protein databases (e.g., NR, UniProt-SwissProt, custom databases).
- Searches of structural and functional domains through HMM profiles (e.g., Pfam, TIGRfam).
- EC Number assignments for enzymes.
- Motifs identified as SignalP, TargetP, tmHMMs, and others.

Auto-naming Strategy




Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

Assigning Enzyme Commission (EC) Number

- EC classification scheme is a hierarchical numerical classification based on the chemical reactions enzymes catalyze. http://www.chem.qmul.ac.uk/iubmb/enzyme/
- Every enzyme code consists of four numbers separated by periods. Ex.- EC 1.1.1.1- alcohol dehydrogenase
- EC numbers may be assigned computationally.
- There are many available tools and methods for predicting EC numbers and pathways.
- Common problems:
  - The computational method may not be specific for assigning EC number to the enzymes. It may be accurate to decide an enzyme family for a gene rather than a specific enzyme. To be precise, the fourth number (Ex. 1.1.1-) is often left blank.

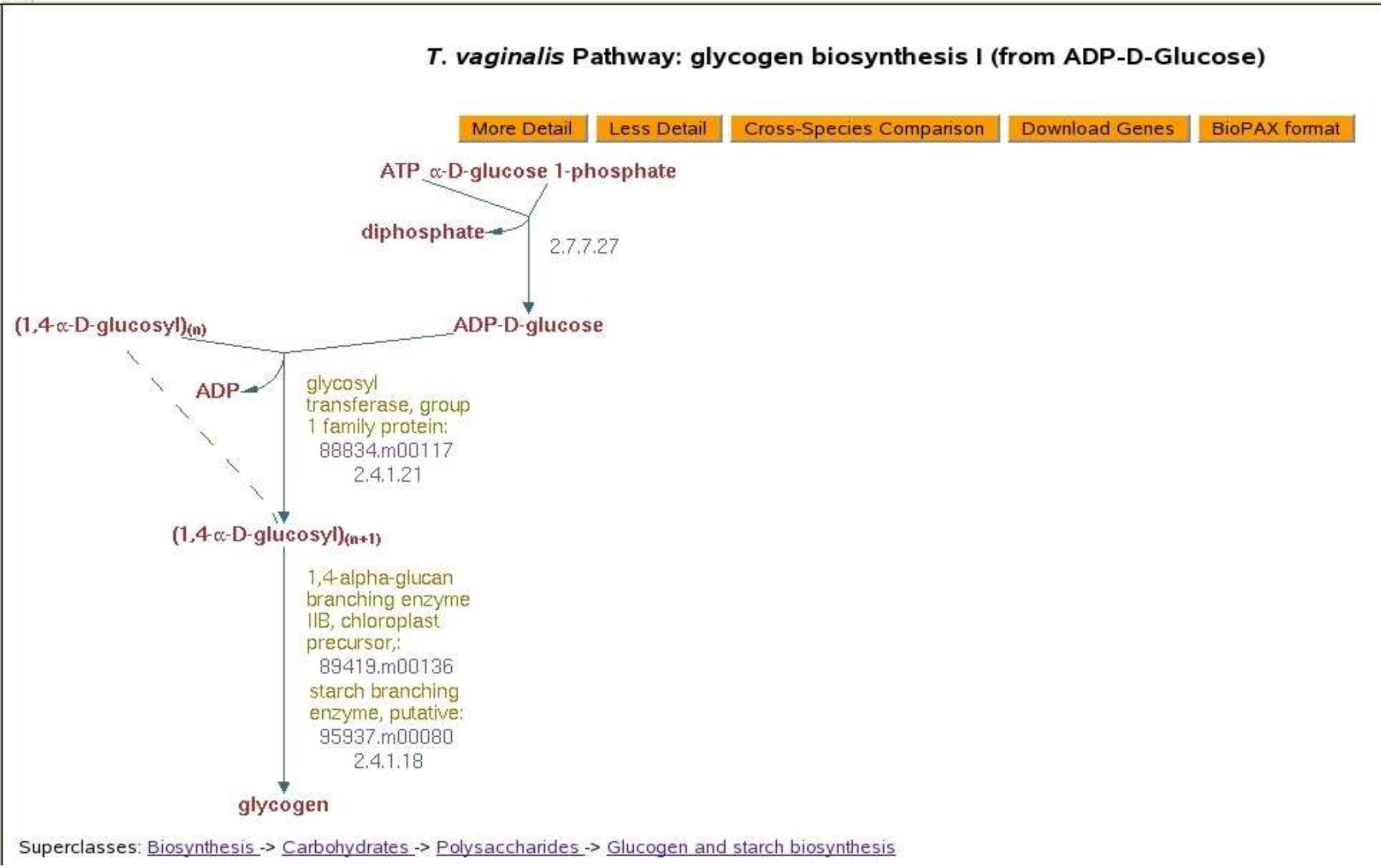
Tools for EC Number Assignment

- Priam profiles: Profiles are built from the sequences available in the ENZYME database based on position-specific score matrices (PSSMs) automatically tailored for each enzyme entry. (http://genopole.toulouse.inra.fr/bioinfo/priam)
- Pfam domains: HMM profiles built from the cross species multiple enzyme sequences.
  - Difficult to assign EC numbers on multi-domain proteins or for bi-functional enzymes.
- Map the EC numbers from the enzymes of an evolutionarily closely related organism based on the homology.

View Pathways

- Graphical interface for users to visualize the substrates, final products and steps in a completed pathway catalyzed by an enzyme (gene).
  - KEGG: http://www.genome.jp/kegg/tool/search_pathway.html
  - MetaCyc: http://metacyc.org
  - Pathway Tools: http://bioinformatics.ai.sri.com/ptools

An Example of Pathway Interface

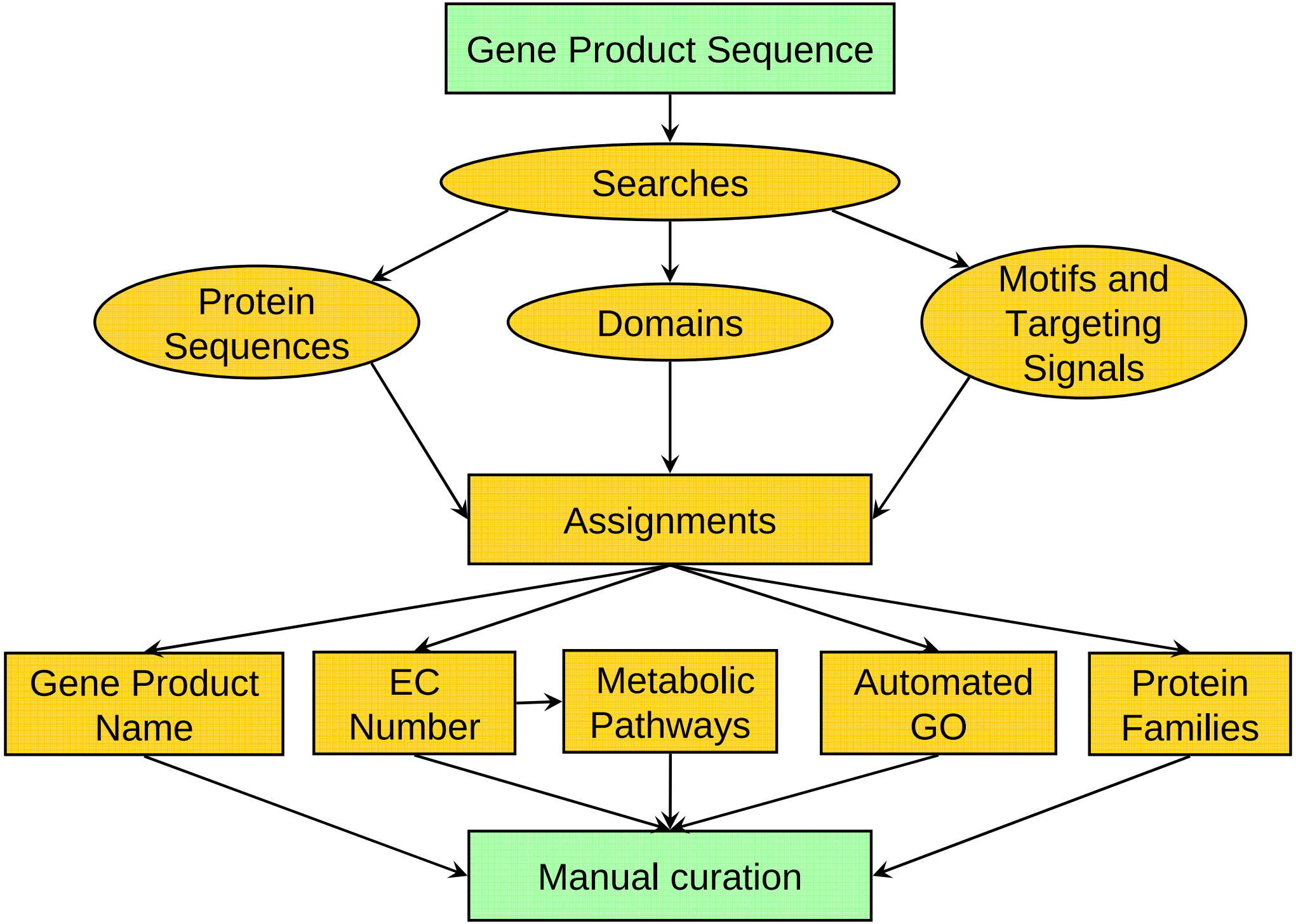


Three Types of GO Terms

- Gene Ontology (Gene Ontology Consortium™) is a method used to structure biological knowledge using a dynamic controlled vocabulary across organisms.
- Molecular function
- Biological process
- Cellular component

Automated Gene Ontology (GO) Assignments

- Automated GO is primarily done by homology searches using different scripts.
- GO terms are often transferred based on orthology from a closely related organism which has manually curated GO assignments.
- Mapping of GO terms from other sources as Interpro, EC numbers, and Pfam domains which have significant similarity, is also a valid computational method.


Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

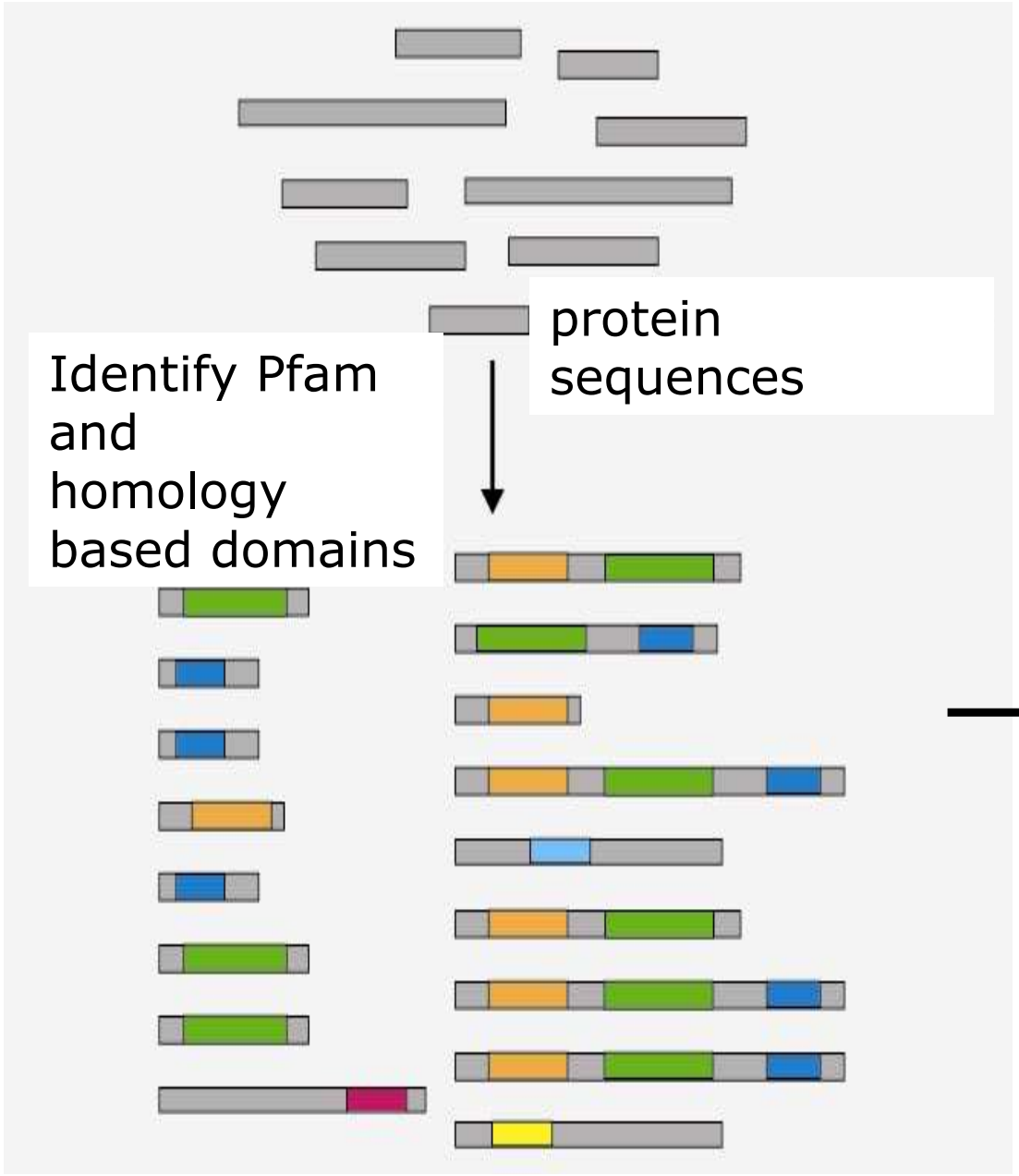
Why Compute Protein Families?

- To group proteins by probable function.
- To identify possible gene structure problems.
- To identify evolutionary relationships between protein families.
- To ensure consistency of annotation.

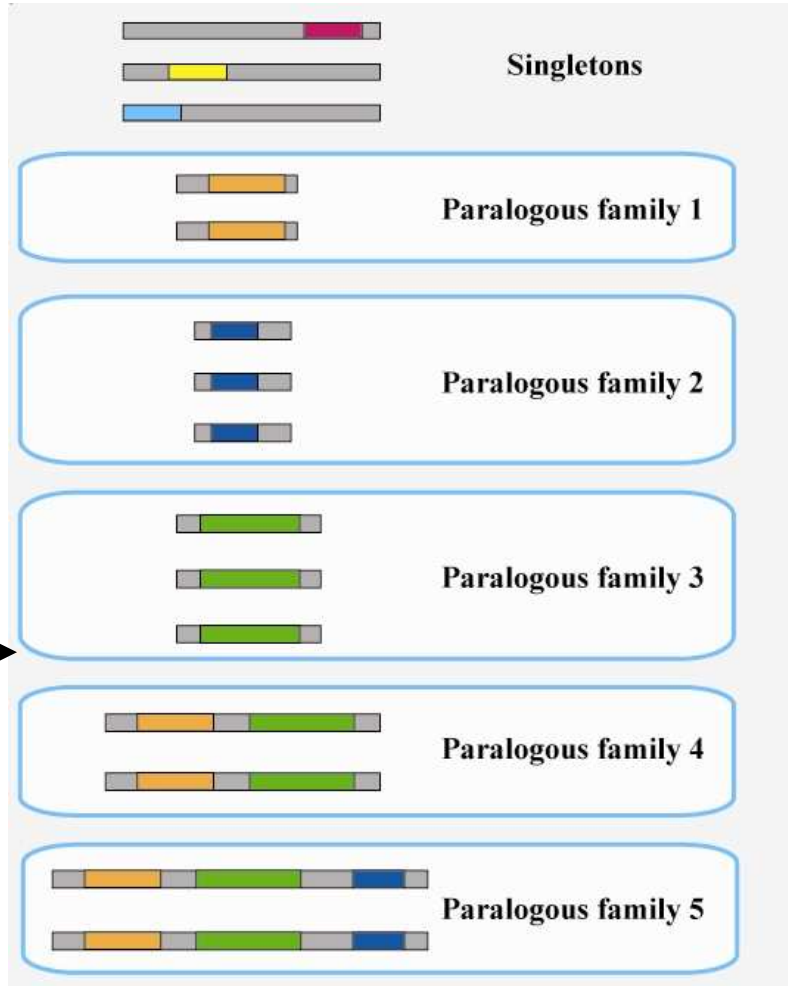
JCVI Domain Based Protein Families

- Steps in generation protein Families:
  - Identify Pfam/TIGRfam domains.
  - Identify homology-based domain-like regions from the proteome by all verses all blast. searches with a significant identity and length cut off.
  - Organize the proteins into families based on similar domain composition.

Families Based on Domain Composition



protein sequences containing Pfam and/or blastP based domains

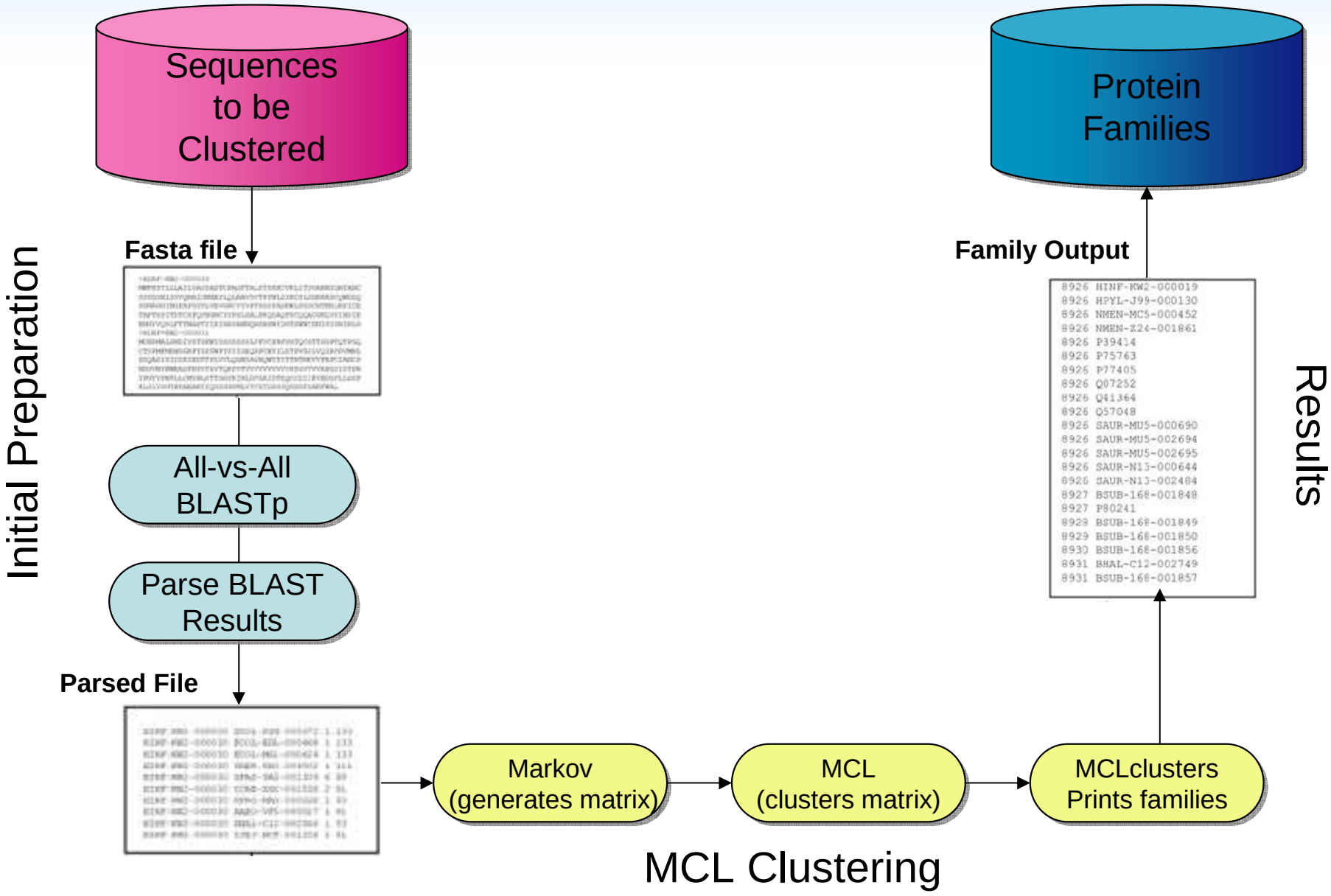


Families grouped based on type and number of domains

MCL Protein Clustering

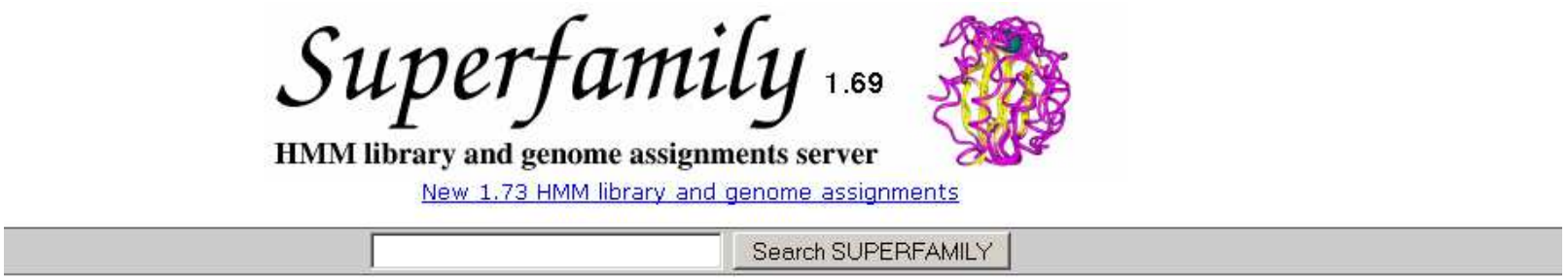
- Method for clustering proteins into related groups, which are termed ‘protein families’.
- Uses Markov clustering method, which is achieved by analyzing similarity patterns between proteins in a given dataset.

MCL

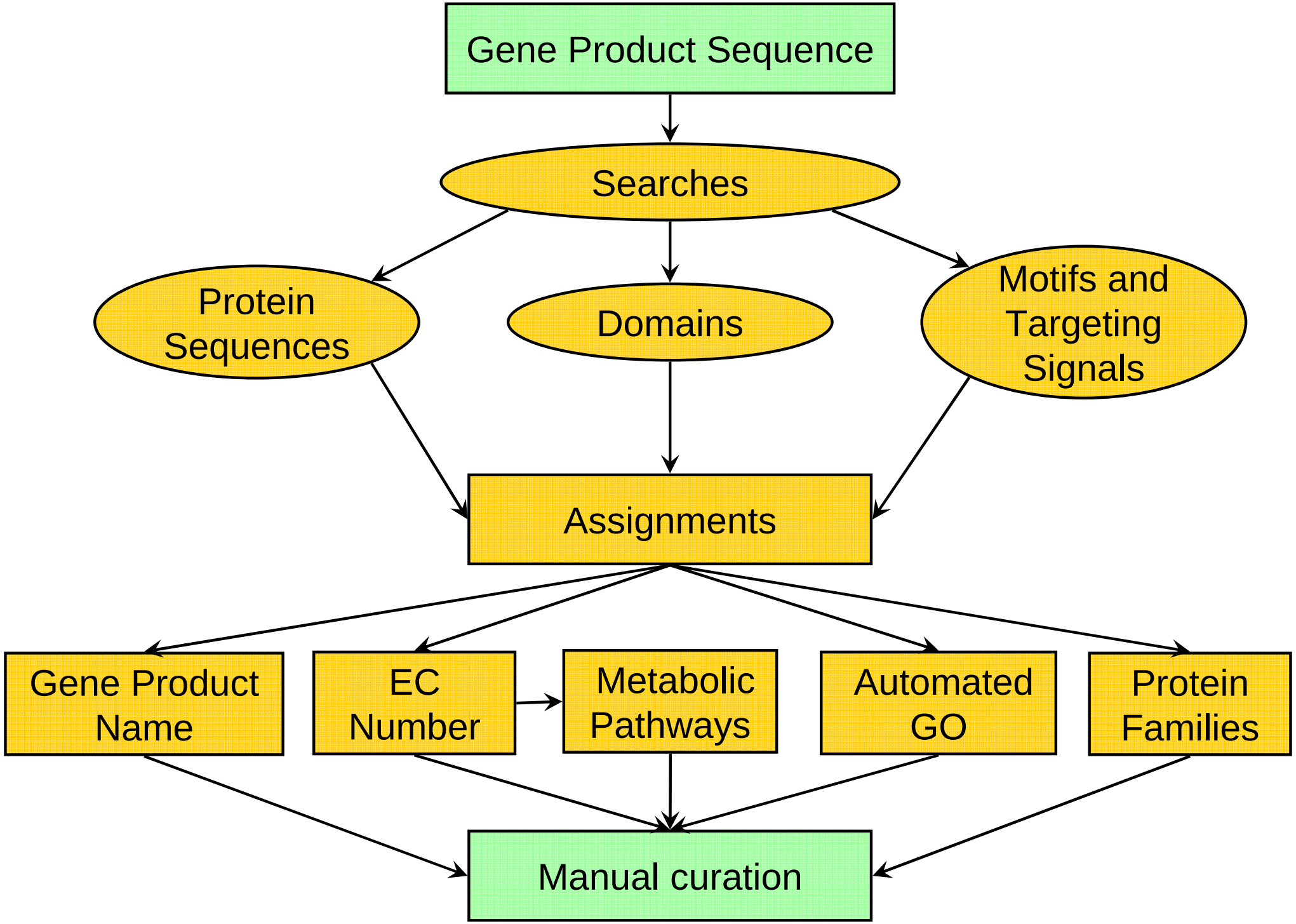

Sequences to be Clustered
Protein Families
Fasta file
Family Output
8926 HINF-RW2-000019
8926 HPYL-J99-000130
8926 NMEN-MC5-000452
8926 NMEN-Z24-001861
8926 P39414
8926 P75763
8926 P77405
8926 Q07252
8926 Q41364
8926 Q57048
8926 SAUR-MU5-000690
8926 SAUR-MU5-002694
8926 SAUR-MU5-002695
8926 SAUR-N13-000644
8926 SAUR-N13-002484
8927 BSUB-168-001848
8927 P80241
8929 BSUB-168-001849
8929 BSUB-168-001850
8930 BSUB-168-001856
8931 BHAL-C12-002749
8931 BSUB-168-001857
All-vs-All BLASTp
Parse BLAST Results
Parsed File
Markov (generates matrix)
MCL (clusters matrix)
MCLclusters Prints families
Initial Preparation
Results
MCL Clustering

Superfamily

http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/



- A superfamily groups together domains of different families which have a common evolutionary ancestor based on structural, functional and evolutionary data.
- SUPERFAMILY domain assignments are generated using an expert curated set of profile HMMs. All models and structural assignments are available for browsing and download. Sophisticated tools are provided for the analysis of superfamily (and family) domain assignments.
- The SCOP domains used as seed sequences to build the SUPERFAMILY models are available for download.


Gene Product Sequence
Searches
Protein Sequences
Domains
Motifs and Targeting Signals
Assignments
Gene Product Name
EC Number
Metabolic Pathways
Automated GO
Protein Families
Manual curation

Summary of Automated Functional Annotation

- High Throughput automated functional annotation is done through the series of scripts using different software tools.
- Pipelines are configured differently for each project, depending on the requirements.
- There is always a hands-on component in eukaryotic functional annotation for the optimum performance of the tools.
- Critically evaluate the computationally derived annotation through manual curation.