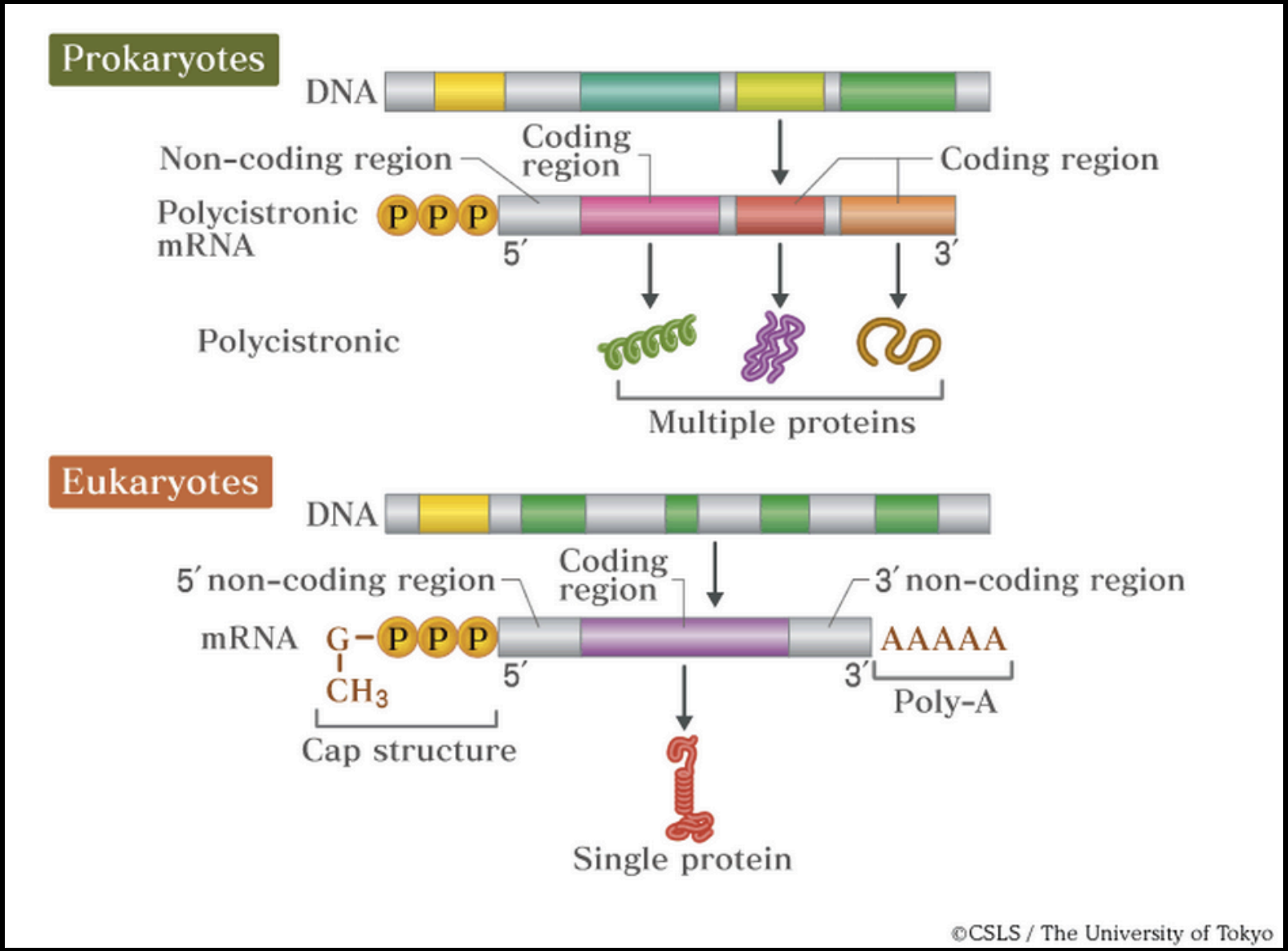
Gene prediction and functional annotation

Gene structures



Gene prediction: Eukaryotes vs prokaryotes

- Much easier in prokaryotic genomes (even in the metagenomes)

- Why?
  - Smaller genomes
  - Simple genome structure and gene structure
  - Relatively more abundant reference genomes available

Gene prediction: overall procedures and strategies

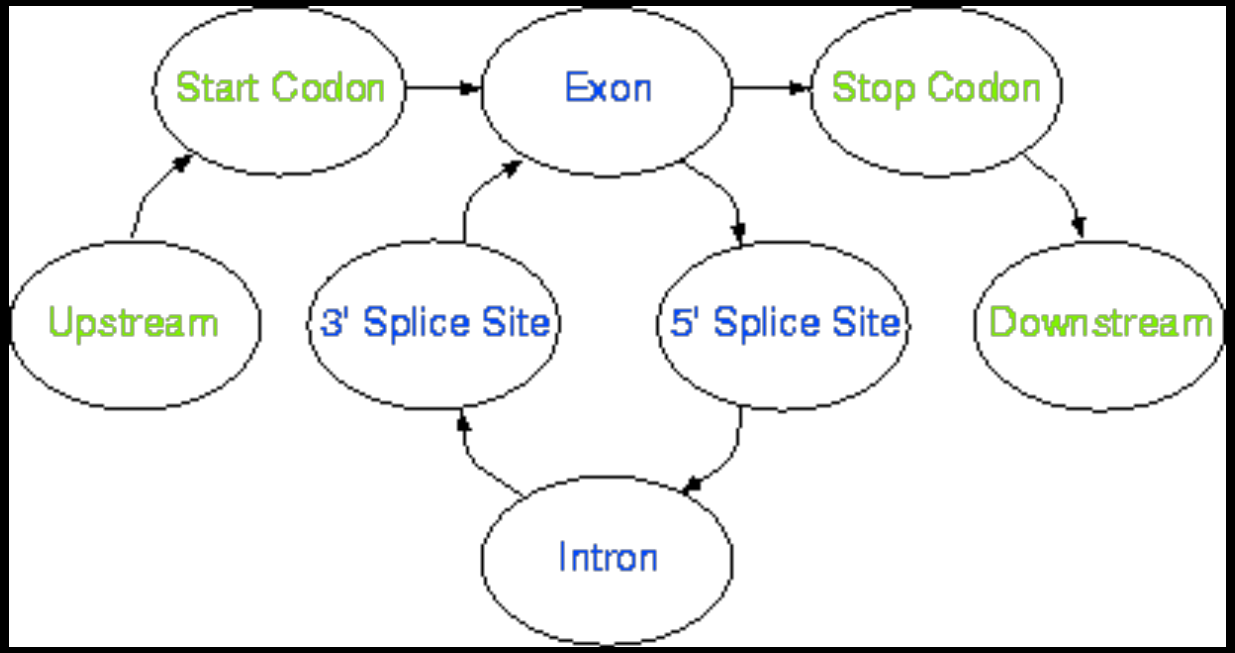
- Procedures
  - Obtain genomic sequences
  - Analyze the genomic features or translated into all 6 reading frames
  - Predict gene positions
- Strategies
  - Homology search
  - Ab initio, from the beginning
  - Evidence incorporated

more robust statistical framework

Gene prediction: commonly used models

- Models
  - Scoring based on gene structure
  - Hidden Markov model (HMM)
  - Support vector machine (SVM)
  - Artificial neural networks (ANNs)

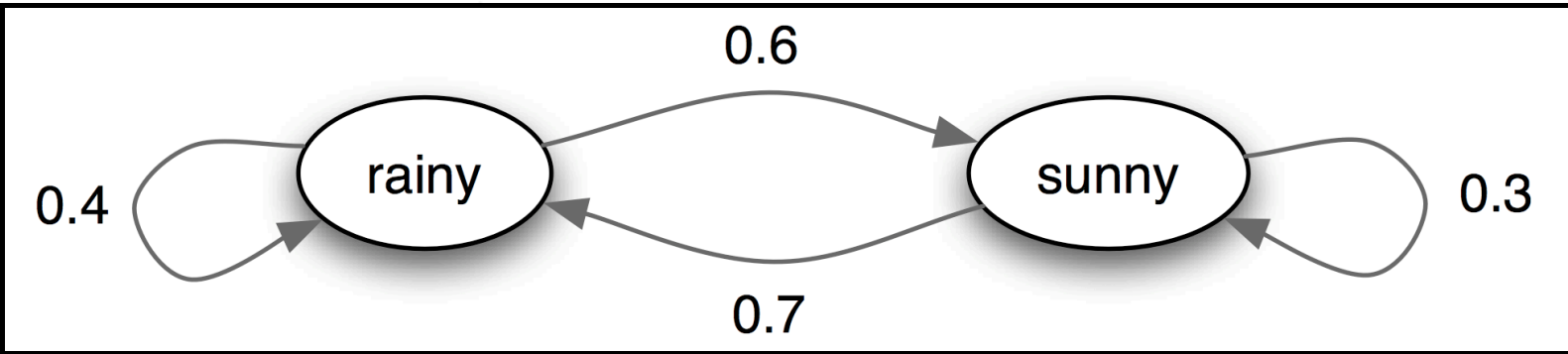
- Softwares
  - GeneMark
  - Glimmer
  - Genescan
  - …



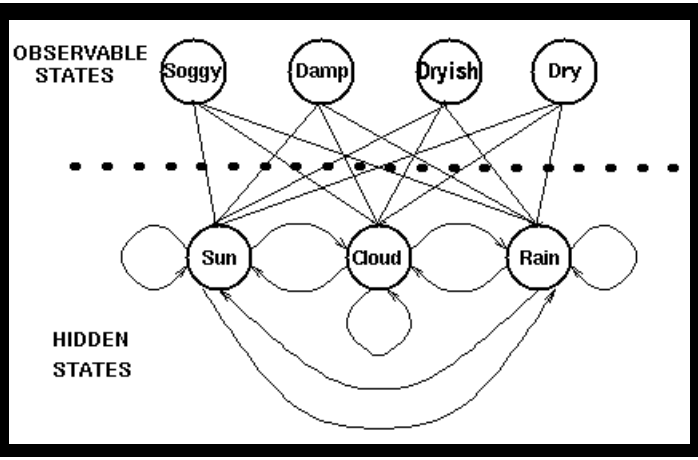
More about HMM

- Markov process
  - A discrete-time Markov chain is a sequence of random variables X_1, X_2, X_3, ... with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states

$$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x \mid X_n = x_n)$$



- HMM $\mathbf{P}\left(Y_n \in A \mid X_1 = x_1, \ldots, X_n = x_n\right) = \mathbf{P}\left(Y_n \in A \mid X_n = x_n\right)$

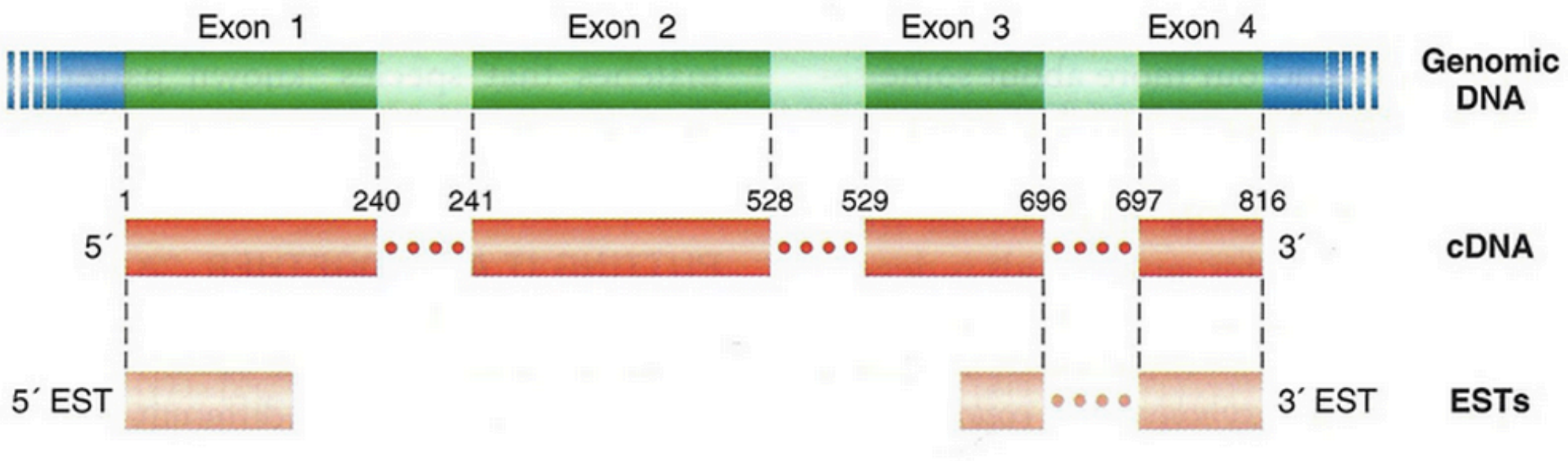


Xn is a Markov process and is not directly observable ("hidden")

$$\mathbf{P}\left(Y_n \in A \mid X_n = x_n\right)$$

Emission probability

Gene prediction with expression data



Incorporating the expression data could increase the emission probability (more evidence)

Gene prediction in metagenome assembly

- Not a tough task if the assembly is good

- Homologous search based on reads mapping to reference databases is OK but less accurate

Output of gene prediction

- GFF and GTF

GFF (**GTF as GFF2**) is a standard format for specifying features in a sequence:

```
contig-124_2    GeneMark.hmm    exon        703     2832    0       -       .       gene_id "1_g"; transcript_id "1_t";
contig-124_2    GeneMark.hmm    stop_codon  703     705     .       -       0       gene_id "1_g"; transcript_id "1_t";
contig-124_2    GeneMark.hmm    CDS         703     2832    .       -       0       gene_id "1_g"; transcript_id "1_t";
contig-124_2    GeneMark.hmm    start_codon 2830    2832    .       -       0       gene_id "1_g"; transcript_id "1_t";
contig-124_2    GeneMark.hmm    exon        3334    4254    0       +       .       gene_id "2_g"; transcript_id "2_t";
contig-124_2    GeneMark.hmm    start_codon 3334    3336    .       +       0       gene_id "2_g"; transcript_id "2_t";
contig-124_2    GeneMark.hmm    CDS         3334    4254    .       +       0       gene_id "2_g"; transcript_id "2_t";
contig-124_2    GeneMark.hmm    stop_codon  4252    4254    .       +       0       gene_id "2_g"; transcript_id "2_t";
contig-124_2    GeneMark.hmm    exon        4355    5662    0       -       .       gene_id "3_g"; transcript_id "3_t";
contig-124_2    GeneMark.hmm    stop_codon  4355    4357    .       -       0       gene_id "3_g"; transcript_id "3_t";
contig-124_2    GeneMark.hmm    CDS         4355    5662    .       -       0       gene_id "3_g"; transcript_id "3_t";
contig-124_2    GeneMark.hmm    start_codon 5660    5662    .       -       0       gene_id "3_g"; transcript_id "3_t";
contig-124_2    GeneMark.hmm    exon        6175    9090    0       -       .       gene_id "4_g"; transcript_id "4_t";
contig-124_2    GeneMark.hmm    stop_codon  6175    6177    .       -       0       gene_id "4_g"; transcript_id "4_t";
contig-124_2    GeneMark.hmm    CDS         6175    9090    .       -       0       gene_id "4_g"; transcript_id "4_t";
contig-124_2    GeneMark.hmm    start_codon 9088    9090    .       -       0       gene_id "4_g"; transcript_id "4_t";
contig-124_2    GeneMark.hmm    exon        10619   10987   0       -       .       gene_id "5_g"; transcript_id "5_t";
contig-124_2    GeneMark.hmm    stop_codon  10619   10621   .       -       0       gene_id "5_g"; transcript_id "5_t";
```

Columns are, left-to-right: (1) contig ID, (2) organism/software, (3) feature type, (4) begin coordinate, (5) end coordinate, (6) score or dot if absent, (7) strand, (8) phase, (9) grouping attribute, features of transcripts.

Functional annotation for bacterial genes

- COG (Cluster of Orthologous Groups)
- GO (Gene Ontology)
- Pfam (the Protein Families database)
- SEED subsystem
- KEGG (Kyoto Encyclopedia of Genes and Genomes)

COG annotation based on CDD search

- Conserved Domain Database (CDD)
  - RPS-blast (Reversed Position Specific BLAST) can be use to predict pre-built CDD families
    - Query sequences to profile
  - COG (Cluster of Orthologous Groups) as a module in CDD and the most frequently used system for bacteria function

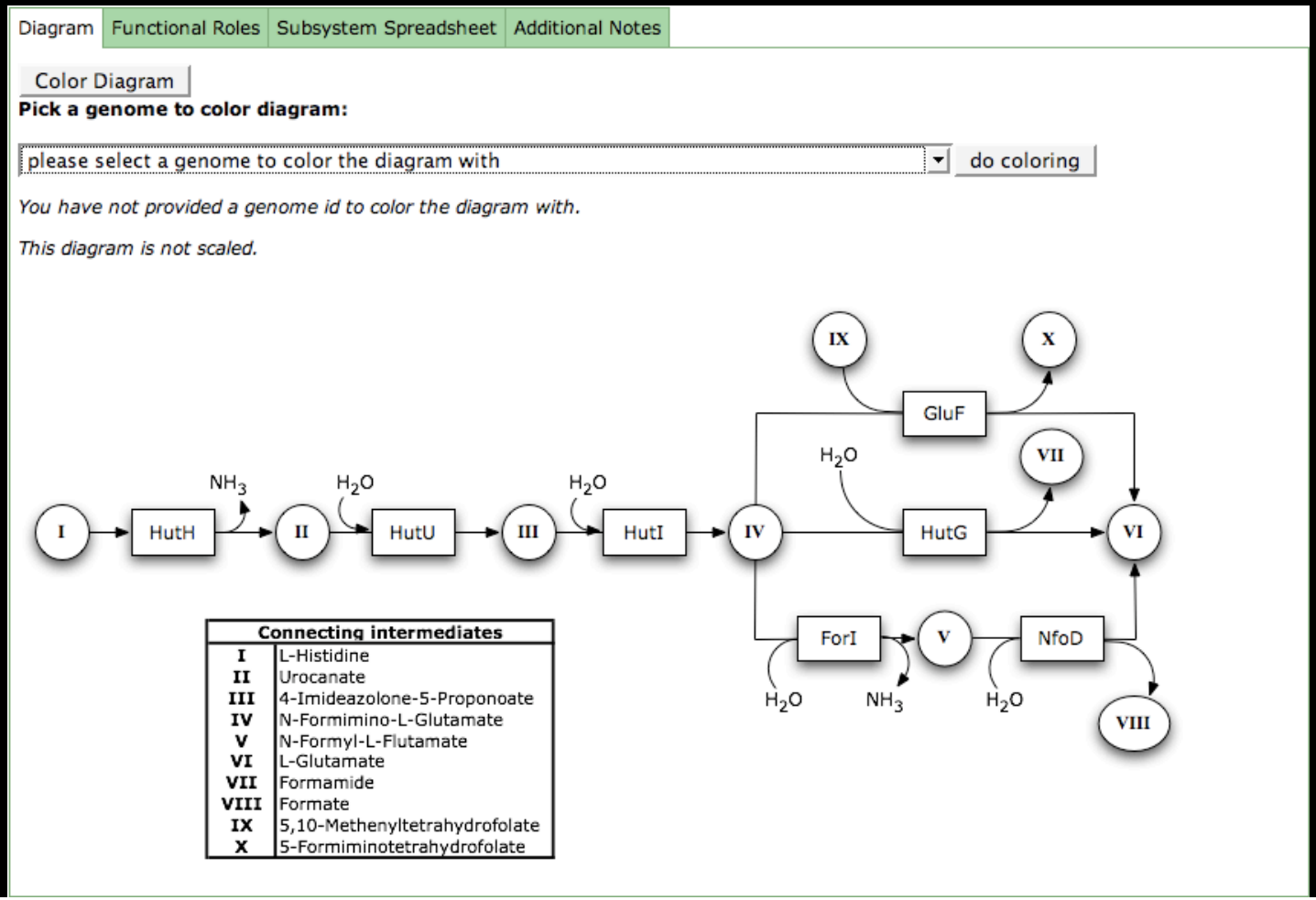
Go and Pfam annotation

- Gene Ontology (GO) Interproscan or AgBase-Goanna server
  - Cellular component
  - Molecular function
  - Biological process

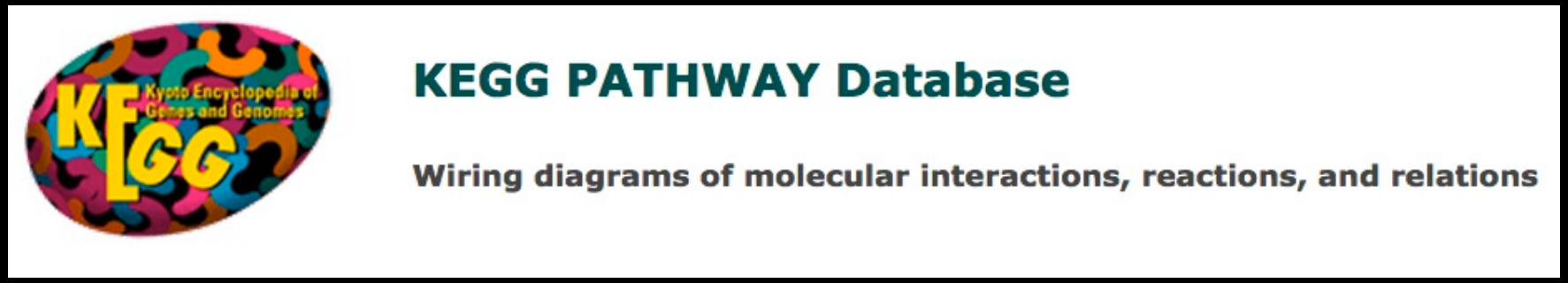
- Pfam Interproscan server, HMMER or blast to CDD
  - Largest protein family database (ortholog domains)
    - Domains are the distinct functional or structural units of a protein
  - One of the most widely accepted and used general gene familier

SEED subsystem

- A subsystem is a collection of functional elements associated to each other in a system, e.g. a metabolic pathway or a component of a cell like a secretion system.
- Figfam as unit



Annotation of KO (KEGG orthology) and KEGG pathway



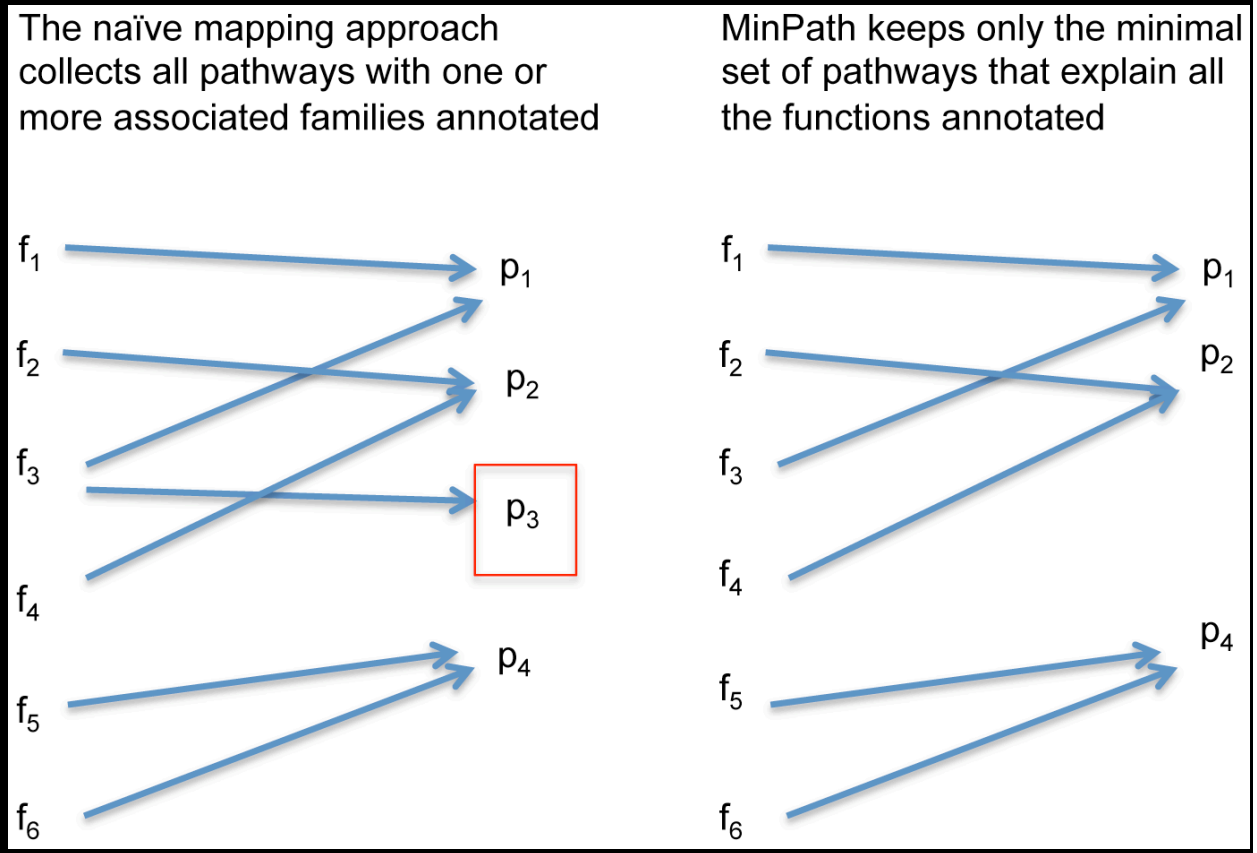
- KEGG (Kyoto Encyclopedia of Genes and Genomes)
  - Integrated database of biological systems, genetic building blocks and chemical building blocks
  - Knowledge based and manually curated
  - Standard reference (cross-species) pathways
  - annotated using ortholog based method or homology searched method (**KOBAS**)

Function and pathway annotation

- Do
  - Profile search using HMM database, including pfam
  - RPS-blast to NCBI-CDD
  - EC annotation using profile mapping
  - KO annotation and KEGG pathway mapping
- Do not
  - Map gene directly to certain functional subsets
  - Map raw read to the reference genes, unless the assembling failed

Do particular pathways exist or not? --infer pathways existence using MinPath

- Difficulty
  - Incomplete data (mainly in metagenome)
  - Pathway redundancy
- Common procedure overestimate the existing biological pathways
- Minpath
  - Parsimony approach
  - Conserved but faithful pathway estimation



Practice, I

- Ab initio gene prediction for species with references genome (yeast)
- Gene prediction for prokaryotic species
- Extract CDS sequences and translate them into protein sequences
- Gene prediction in metagenome

Practice, II

- Predict Pfam using Hmmer
- Predict CDD and COG using rpsblast
- Predict EC number for enzymes using PRIAM
- Predict KEGG Orthology (KO) using KOBAS
- Predict the existing pathways using MinPath