Задачи протеомики ("Белковая биоинформатика")

Ярослав Баранов
МНЛ«Компьютерные технологии», Университет ИТМО

Overview

- Role of bioinformatics/computational biology in proteomics research

- Functional annotation of proteins = assigning correct name, describing function or predicting function for a sequence
- Classification of proteins = grouping them into families of related sequences
- Annotating a family helps the annotation of its members

Sequence $\longrightarrow$ function

Bioinformatics as related to proteins

1. Sequence analysis

- Genome projects -> Gene prediction
- Protein sequence analysis
- Comparative genomics
- Protein sequence and family databases (annotation and classification)

2. Structural genomics

3. Data analysis and integration for:

- Large scale gene expression analysis
- Protein-protein interaction
- Intracellular protein localization

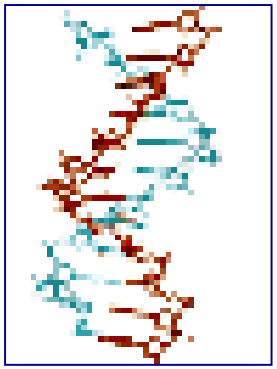
4. Integration of all data on proteins to reconstruct pathways and cellular systems, make predictions and discover new knowledge

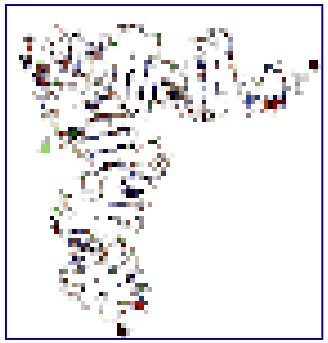
Functional Genomics and Proteomics

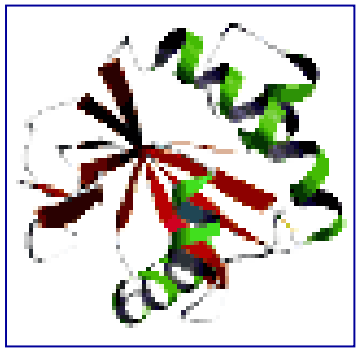
Proteomics studies biological systems based on global knowledge of protein sets (proteomes).

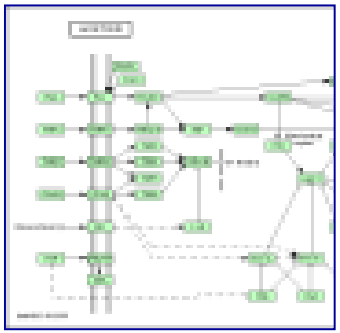
Functional genomics studies biological functions of proteins, complexes, pathways based on the analysis of genome sequences. Includes functional assignments for protein sequences.

Genome → Transcriptome → Proteome → Metabolome









Proteomics

Data: Gene expression profiling

Genome-wide analyses of gene expression (DNA Microarrays/Chips)

Data: Protein-protein interaction

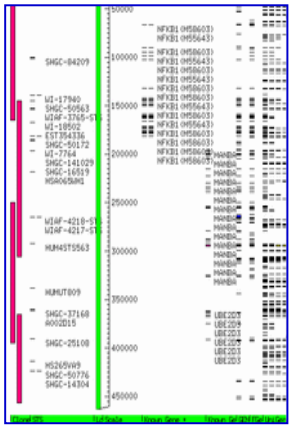
(Yeast Two-Hybrid Systems)

Data: Structural genomics

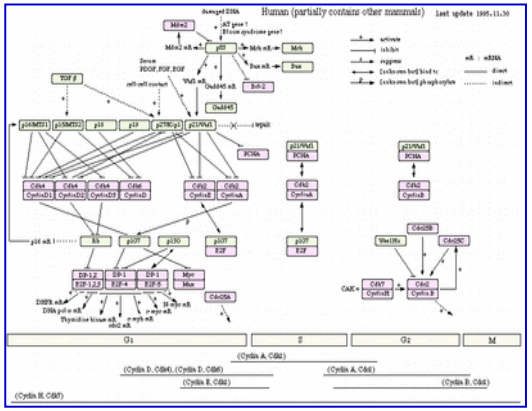
Determine 3D structures of all protein families

Data: Genome projects (Sequencing)

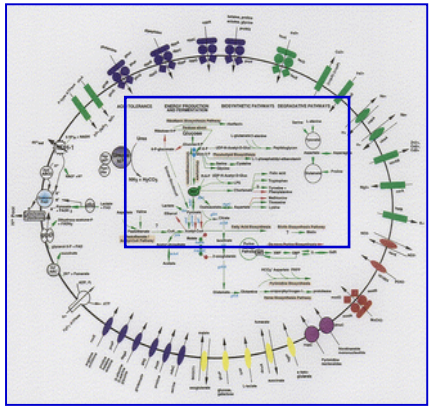
Bioinformatics and Genomics/Proteomics



Sequence,
Other Data



Pathways and Regulatory Circuits



Hypothetical Cell

Unknown
Genes

Putative Functional
Groups

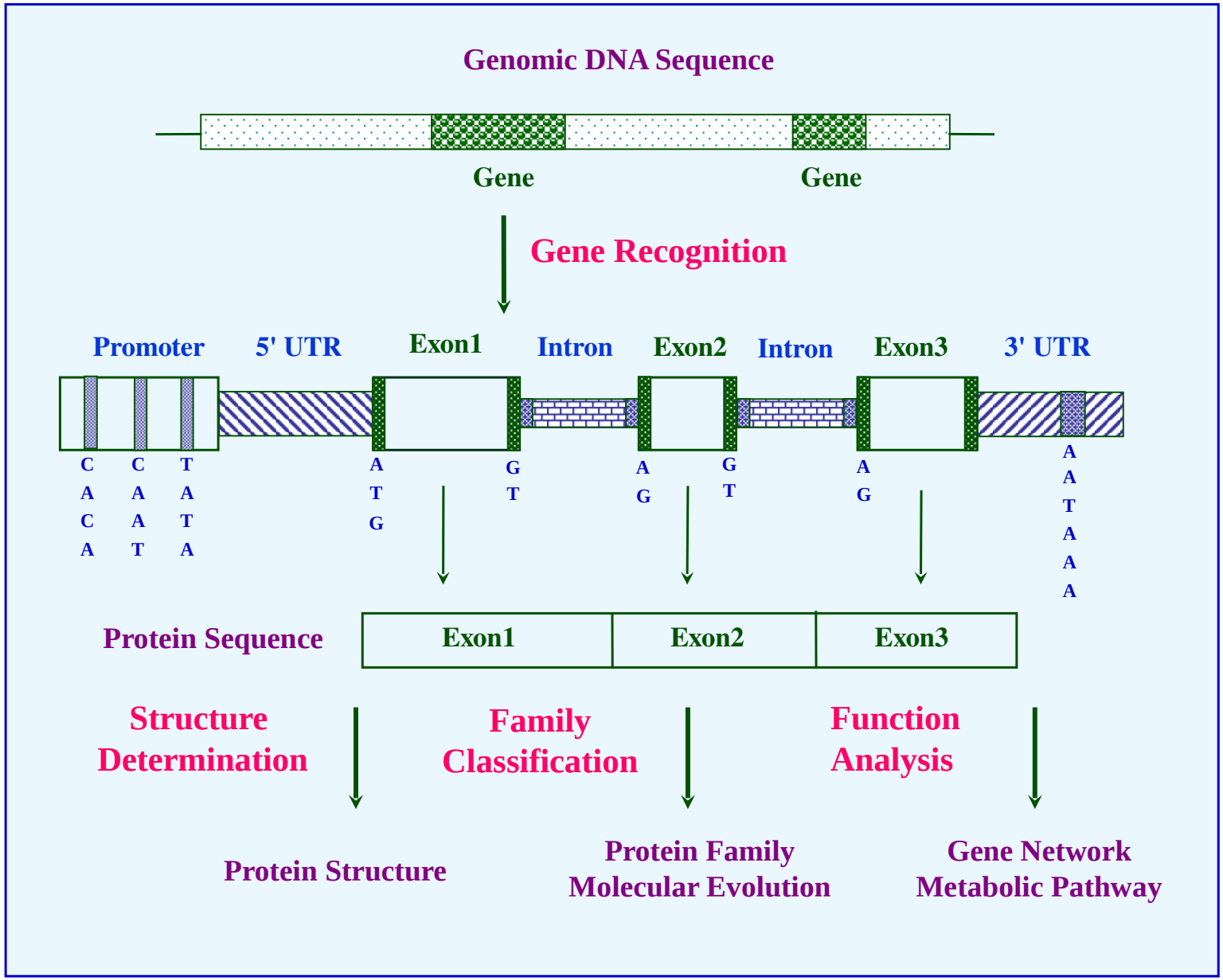
Work with protein sequence, not DNA sequence

DNA Sequence

Gene

Protein Sequence

Function



Most new proteins come from genome sequencing projects

- *Mycoplasma genitalium* - 484 proteins
- *Escherichia coli* - 4,288 proteins
- *S. cerevisiae* (yeast) - 5,932 proteins
- *C. elegans* (worm) ~ 19,000 proteins
- *Homo sapiens* ~ 30,000 proteins

... and have unknown functions

Advantages of knowing the genome sequence

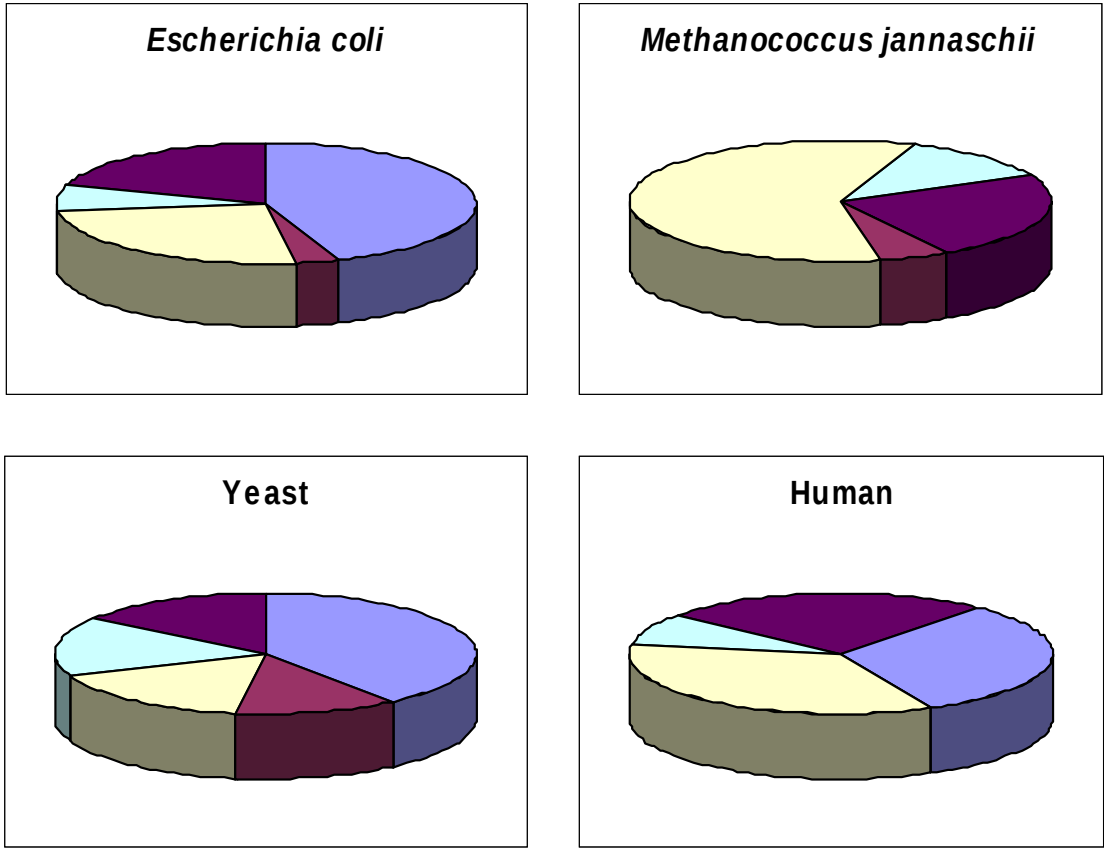
- **All encoded proteins** can be predicted and identified
- The **missing** functions can be identified and analyzed
- **Peculiarities** and **novelties** in each organism can be studied
- **Predictions** can be made and verified

The changing face of protein science

20th century	21st century
• Few well-studied proteins	• Many “hypothetical” proteins
• Mostly globular with enzymatic activity	• Various, often with no enzymatic activity
• Biased protein set	• Natural protein set

Properties of the natural protein set

- Unexpected **diversity** of even common enzymes (analogous, paralogous enzymes)
- Conservation of the **reaction chemistry**, but not the **substrate specificity**
- **Functional diversity** in closely related proteins
- Abundance of **new structures**



	E. coli	*M. jannaschii*	*S. cerevisiae*	*H. sapiens*
Characterized experimentally	2046	97	3307	10189
Characterized by similarity	1083	1025	1055	10901
Unknown, conserved	285	211	1007	2723
Unknown, no similarity	874	411	966	7965

from Koonin and Galperin, 2003, with modifications

Functional annotation of proteins
(protein sequence databases)

Automatic assignment based on sequence similarity:
gene name, protein name, function

To avoid mistakes, need human intervention (manual annotation)

Best annotated protein databases: SwissProt, PIR-1
Now part of UniProt – Universal Protein Resource

Objectives of functional analysis for different groups of proteins

- **Experimentally characterized**
  Up-to-date information, manually annotated (curated database!)
  Problems: **misinterpreted experimental results** (e.g. suppressors, cofactors)

- **“Knowns” = Characterized by similarity**
  (closely related to experimentally characterized)
  Make sure the assignment is plausible

- **Function can be predicted**
  Extract maximum possible information
  Avoid errors and overpredictions

Fill the gaps in metabolic pathways

- **“Unknowns” (conserved or unique)**
  Rank by importance

Problems in functional assignments for “knowns”

- **Low sequence complexity**
  (coiled-coil, non-globular regions)

- **Enzyme evolution:**

  Divergence in sequence and function

  Non-orthologous gene displacement:
  Convergent evolution

Functional prediction:
Dealing with “hypothetical” proteins

- **Computational analysis**

Sequence analysis of the new ORF

- **Structural analysis**
  Determination of the 3D structure

- **Mutational analysis**

- **Functional analysis**
  Expression profiling
  Tracking of cellular localization

Functional prediction: computational analysis

Cluster analysis of protein families (**family databases**)

Use of sophisticated database searches (**PSI-BLAST, HMM**)

Detailed **manual analysis** of sequence similarities

Using comparative genomics for protein analysis

Proteins (domains) with **different 3D folds are not homologous** (unrelated by origin)

Those amino acids that are **conserved in divergent proteins** within a (super)family are likely to be important for catalytic activity.

Reaction chemistry often remains **conserved** even when sequence diverges almost beyond recognition

Using comparative genomics for protein analysis

Prediction of the **3D fold** ***(if distant homologs have known structures)*** and **general biochemical function** is much easier than prediction of **exact biological** (or biochemical) **function**

Sequence analysis complements **structural comparisons** and can greatly benefit from them

Comparative analysis allows us to find subtle sequence similarities in proteins that would not have been noticed otherwise

Poorly characterized protein families: only general function can be predicted

Enzyme activity can be predicted, the substrate remains unknown (ATPases, GTPases, oxidoreductases, methyltransferases, acetyltransferases)

Helix-turn-helix motif proteins (predicted transcriptional regulators)

Membrane transporters

Functional prediction: computational analysis

Phylogenetic distribution

Wide - most likely essential

Narrow - probably clade-specific

Patchy - most intriguing, niche-specific

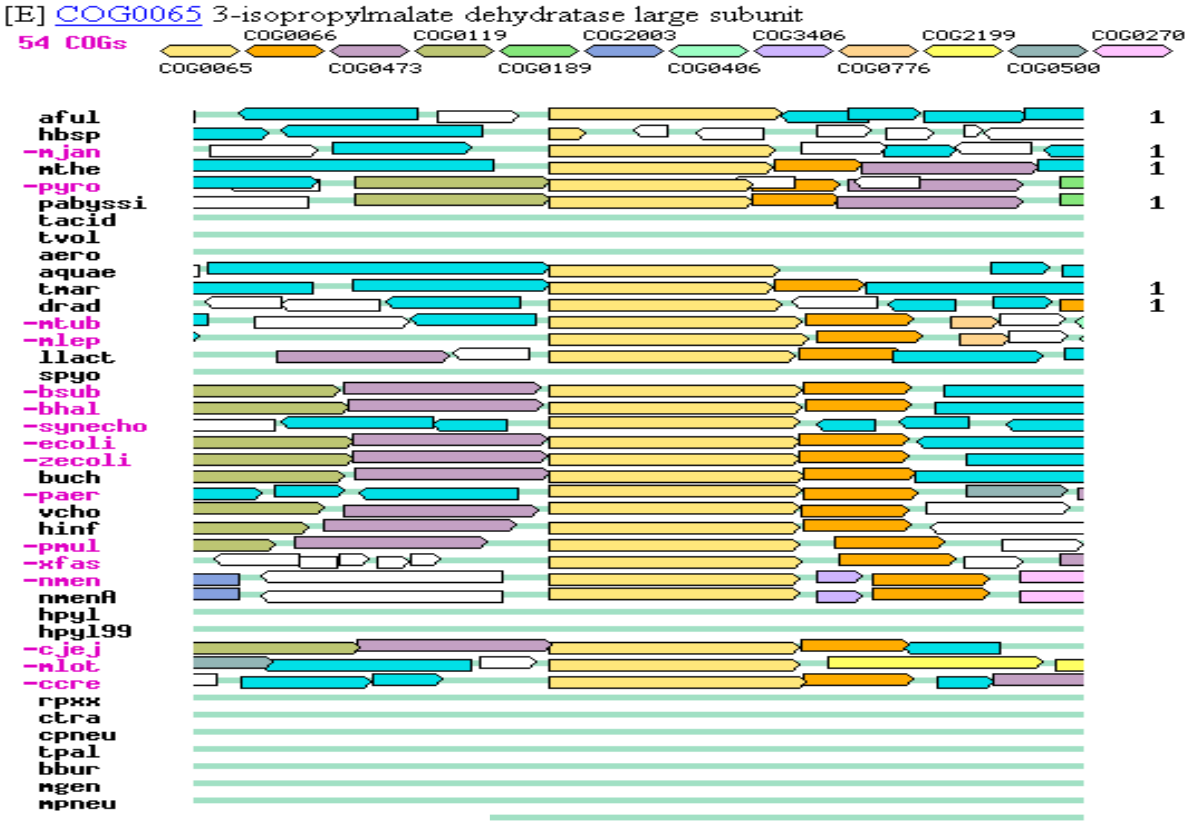
Domain association

(for multidomain proteins)

Gene neighborhood

(operon organization)

Using genome context for functional prediction



Leucine biosynthesis

ZoomIn ZoomOut

28 [E] COG0065 3-isopropylmalate dehydratase large subunit
22 [E] COG0066 3-isopropylmalate dehydratase small subunit
16 [E] COG0473 Isocitrate/isopropylmalate dehydrogenase
11 [E] COG0119 Isopropylmalate/homocitrate/citramalate synthases
2 [HJ] COG0189 Glutathione synthase/Ribosomal protein S6 modification enzyme (glutami
2 [L] COG2003 DNA repair proteins

Functional Prediction: Role of Structural Genomics

Protein Structure Initiative:
Determine 3D Structures of All Proteins

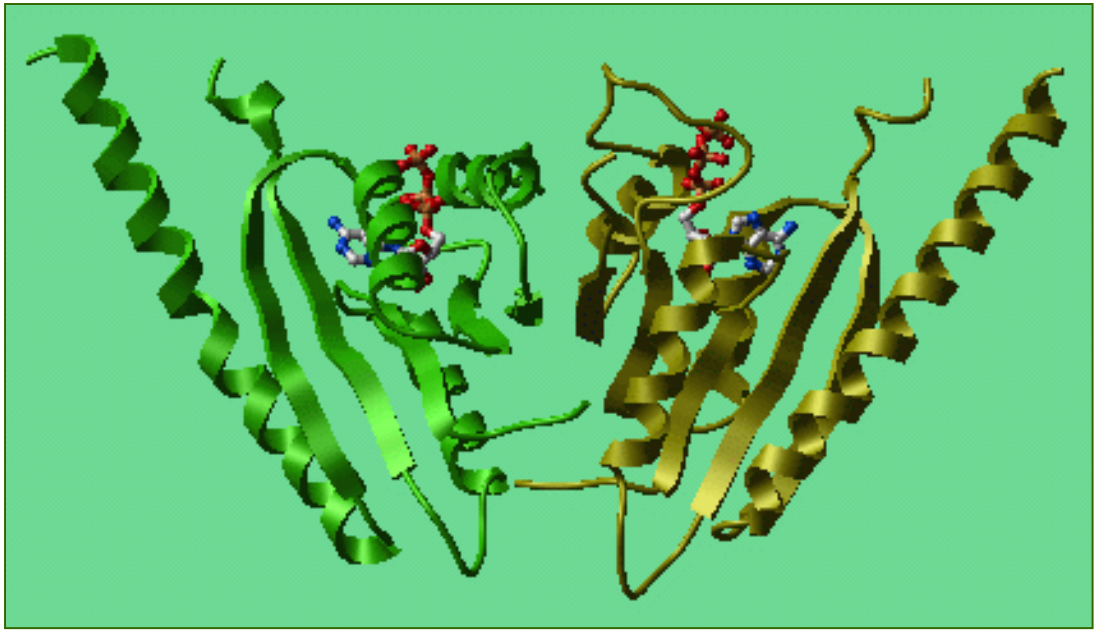
Family Classification:
Organize Protein Sequences into Families, collect families without known structures

Target Selection:
Select Family Representatives as Targets

Structure Determination:
X-Ray Crystallography or NMR Spectroscopy

Homology Modeling:
Build Models for Other Proteins by Homology
Functional prediction based on structure

Structural Genomics: Structure-Based Functional Assignments



Methanococcus jannaschii **MJ0577 (Hypothetical Protein)**

Contains bound ATP => ATPase or ATP-Mediated Molecular Switch

Confirmed by biochemical experiments

Functional prediction: problem areas

- Identification of **protein-coding regions**
- Delineation of potential function(s) for **distant paralogs**
- Identification of domains in the absence of close homologs
- Analysis of proteins with **low sequence complexity**

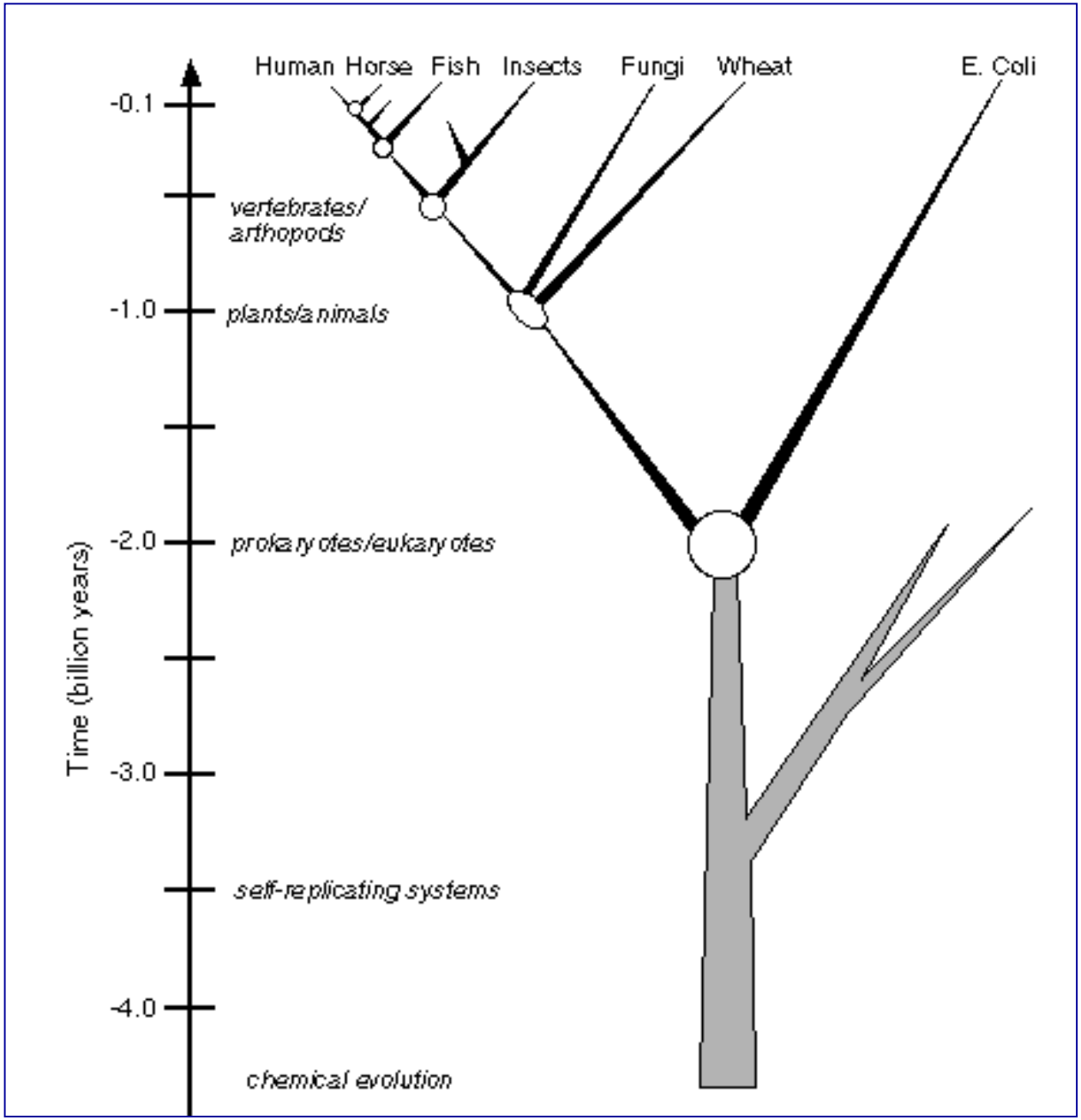
Can protein classification help?

- Protein families are real and reflect evolutionary relationships
- Function often follows along the family lines
- Therefore, matching a new protein sequence to well-annotated and curated family provides information about this new protein and helps predicting its function.
  This is more accurate than comparing the new sequence to individual proteins in a database:
  (search classification database vs search protein database)

To make annotation and functional prediction for new sequences accurate and efficient, need "natural" protein classification

Protein Evolution

- **Tree of Life & Evolution of Protein Families (Dayhoff, 1978)**
- **Can build a tree representing evolution of a protein family, based on sequences**
- **Orthologous Gene Family: Organismal and Sequence Trees Match Well**



Protein Evolution

Homolog

- **Common Ancestors**
- **Common 3D Structure**
- **Usually at least some sequence similarity (sequence motifs or more close similarity)**

Ortholog

- **Derived from Speciation**

Paralog

- **Derived from Duplication**

A

ancestor

Ax_1 Az_2

ancestor

Ax_1 Az_2 **Ax_1 Az_2**

Species 1 **Species 2**

Protein Family vs Domain

Protein domain/family

- Most proteins have « modular » structures
- Estimation: ~ 3 domains / protein

```
                                                  10        20        30        40        50        60
   Sequence ID  start end weight  ---------|---------|---------|---------|---------|---------|------
3 ⊞ EPO_HUMAN     .   .   2.41 APPRLICDSRVLERYLLEAKEAENVTMGCSEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQG
2 ⊞ EPO_RAT       .   .   2.61 APPRLICDSRVLERYILEAKEAENVTMGCAEGPRLSENITVPDTKVNFYAWKRMEVEEQAIEVWQG
3 ⊞ EPO_FELCA     .   .   2.99 APPRLICDSRVLERYILEAREAENATMGCAEGCSFSENITVPDTKVNFYAWKRMEVQQQALEVWQG
8   Consensus                8.01 APPRLICDSRVLERYILEAKEAENVTMGCAEGCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQG
1   PROSITE                       ----------------------------
```

- Domains can be defined by different methods:
  - **Pattern** (regular expression); used for very conserved domains
  - **Profiles** (weighted matrices): two-dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence

Protein Evolution: Sequence Change vs. Domain Shuffling

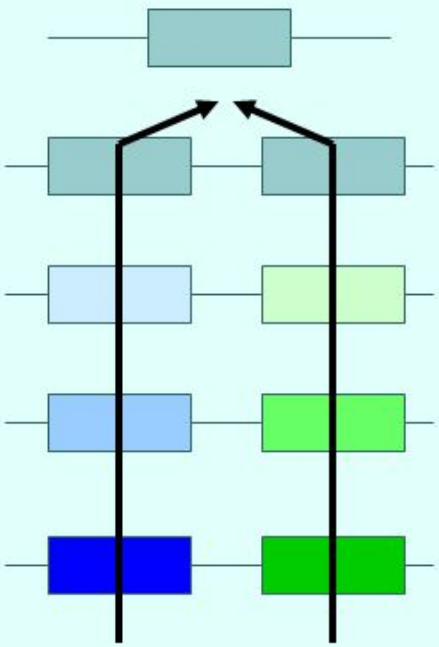
PIR
Protein Information Resource

LIGCLRNCSAVTAAAKQLAE
VTGFSN----AKTTAQHVKK

Protein Evolution

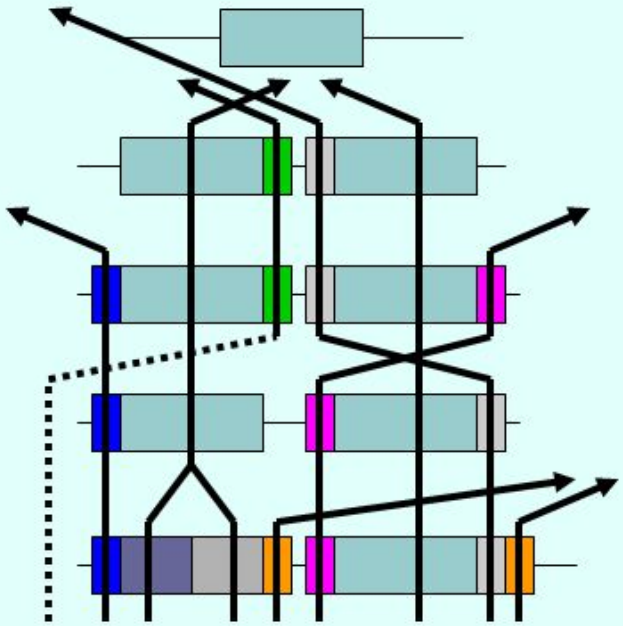
Domain: Evolutionary/Functional/Structural Unit

Sequence changes



With enough similarity, one can trace back to a common origin

Domain shuffling



What about these?

Levels of Protein Classification

Level	*Example*	*Similarity*	*Evolution*
Class	α/β	Structural elements	No relationships
Fold	TIM-Barrel	Topology of backbone	Possible monophyly
Domain Superfamily	Aldolase	Recognizable sequence similarity (motifs); basic biochemistry	Monophyletic origin
Family	Class I Aldolase	High sequence similarity (alignments); biochemical properties	Evolution by ancient duplications
Orthologous group	2-keto-3-deoxy-6-phosphogluconate aldolase	Orthology for a given set of species; biochemical activity; biological function	Traceable to a single gene in LCA
Lineage-specific expansion (LSE)	PA3131 and PA3181	Paralogy within a lineage	Recent duplication

Protein classification databases

Domain classification

Pfam

SMART

CDD

Whole protein classification

PIRSF

Mixed

TIGRFAMS

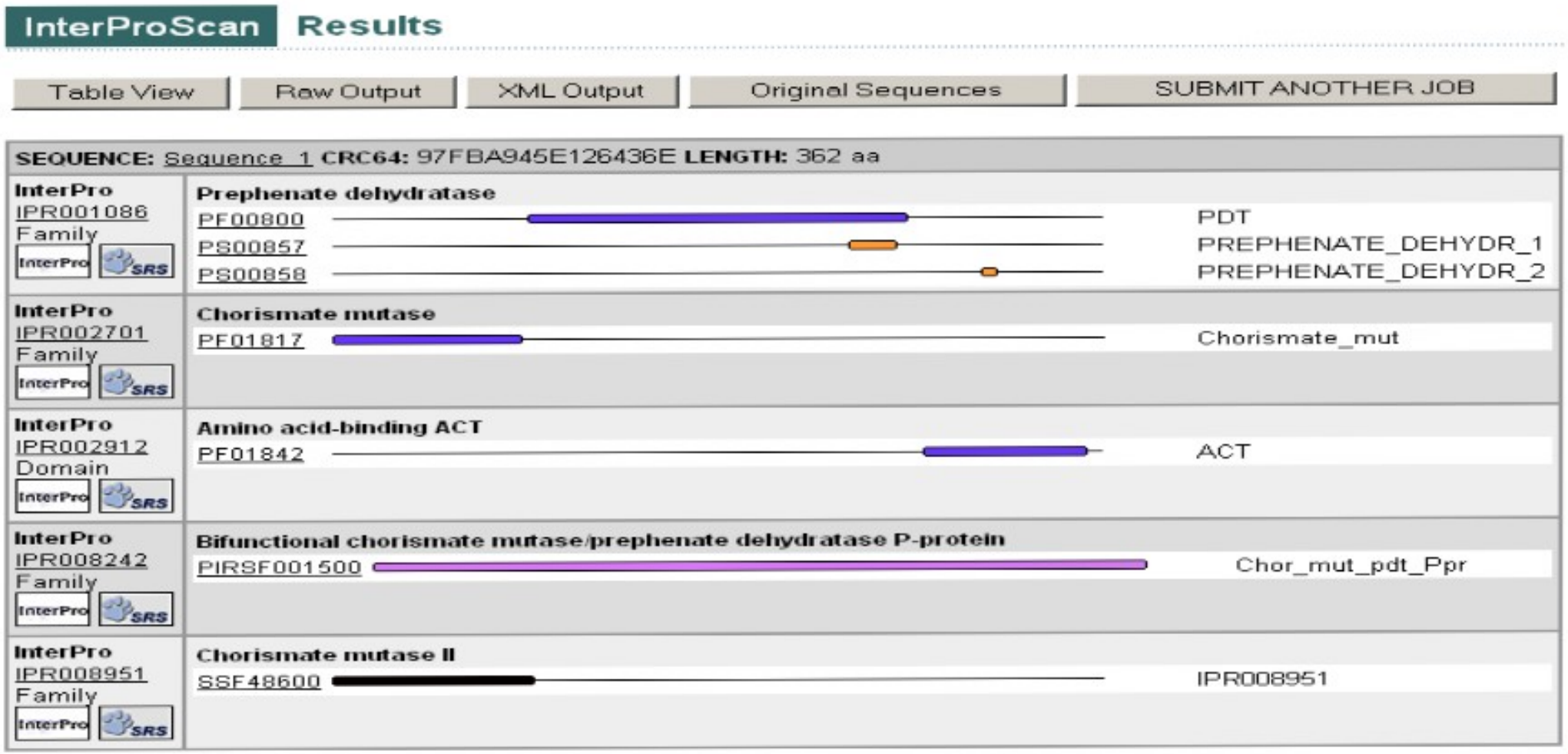
COGs

Based on structural fold

- **SCOP**

Protein family – domain – site (motif)

InterPro is an integrated resource for protein families, domains and sites.
Combines a number of databases: **PROSITE, PRINTS, Pfam, SMART, ProDom, TIGRFAMs, PIRSF**

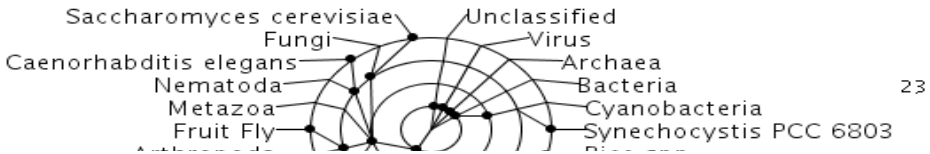


InterPro Entry

InterPro Entry Type defines the entry as a Family, Domain, Repeat, or Site
Family = protein family.

"Contains" field lists domains within this protein
"Found in" field: for domain entries, lists families which contain this domain

InterPro **Bifunctional chorismate mutase/prephenate dehydratase P-protein** [?] = help

IPR008242 **Chor_mut_pdt_Ppr**	Matches: 23 proteins View matches: [Overview][...sorted by Name][of known structure][Detailed view][Table view]
Name [?]	Bifunctional chorismate mutase/prephenate dehydratase P-protein
Signatures [?]	PIRSF001500;Chor_mut_pdt_Ppr (23 proteins)
Type [?]	Family
Dates [?]	2003-04-14 11:14:43.0 (created) 2003-04-14 11:14:43.0 (modified)
Contains [?]	IPR001086; Prephenate dehydratase IPR002701; Chorismate mutase
Process [?]	L-phenylalanine biosynthesis (GO:0009094)
Function [?]	chorismate mutase activity (GO:0004106) prephenate dehydratase activity (GO:0004664)
Component [?]	cytoplasm (GO:0005737)
Abstract [?]	The bifunctional P-protein, which plays a central role in phenylalanine biosynthesis, contains two catalytic domains (chorismate mutase and prephenate dehydratase) and a regulatory domain (ACT). It is part of the shikimate pathway, which is present only in bacteria, fungi, and plants. Chorismate mutase (CM; EC: 5.4.99.5) catalyses the rearrangement of chorismate to prephenate, the reaction at the branch point of the biosynthetic pathway leading to the three aromatic amino acids, phenylalanine, tryptophan and tyrosine (chorismic acid is the last common intermediate, and CM leads to the L-phenylalanine/L-tyrosine branch). The chorismate mutase domain of this protein belongs to the AroQ class (Prokaryotic type) [1] and has an all-helical structure. There are stand-alone versions of this domain (e.g., IPR008239), as well as fusions to other catalytic domains (prephenate dehydrogenase, IPR008244; 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase, PIRSF005994), or to regulatory domains. Prephenate dehydratase (PDT; EC: 4.2.1.51) converts prephenate to phenylpyruvate. There also exists a fusion of this domain with ACT domain alone (IPR008237), making a monofunctional PDT. In Escherichia coli P-protein, the ACT domain has been shown to be essential for phenylalanine-mediated feedback inhibition and ligand binding [2]. It is a ligand-binding regulatory domain found primarily in enzymes and regulators of amino acid and purine metabolism [3]. For additional information please see [4, 5, 6, 7, 8, 9].
Structural links [?]	PDB 1ecm CATH 1.20.59.10.1 SCOP a.130.1.1
Database links [?]	Enzyme 4.2.1.51 Enzyme 5.4.99.5
Taxonomy [?]	

Impact of protein bioinformatics and genomics

Single protein level

Discovery of new enzymes and superfamilies

Prediction of active sites and 3D structures

Pathway level

Identification of “missing” enzymes

Prediction of alternative enzyme forms

Identification of potential drug targets

Cellular metabolism level

Multisubunit protein systems

Membrane energy transducers

Cellular signaling systems

Спасибо за внимание!