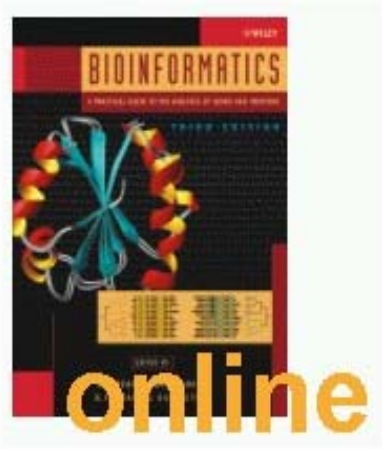
# 5. Predictive Methods Using Protein Sequences (2)

薛 佑 玲 Yow-Ling Shiue
國立中山大學生物醫學研究所
ylshiue@mail.nsysu.edu.tw




*Select a Chapter:* Chapter 8

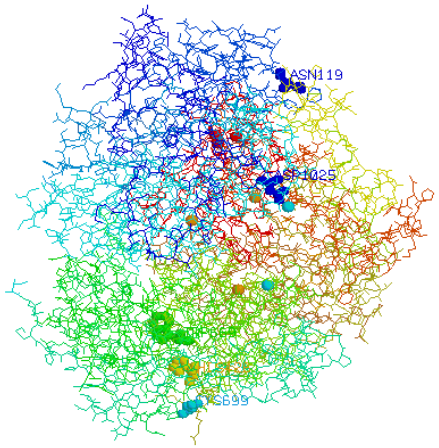
## *Chapter 8: Predictive Methods Using Protein Sequences*

- Sample Data for Problem Sets
- Internet Resources

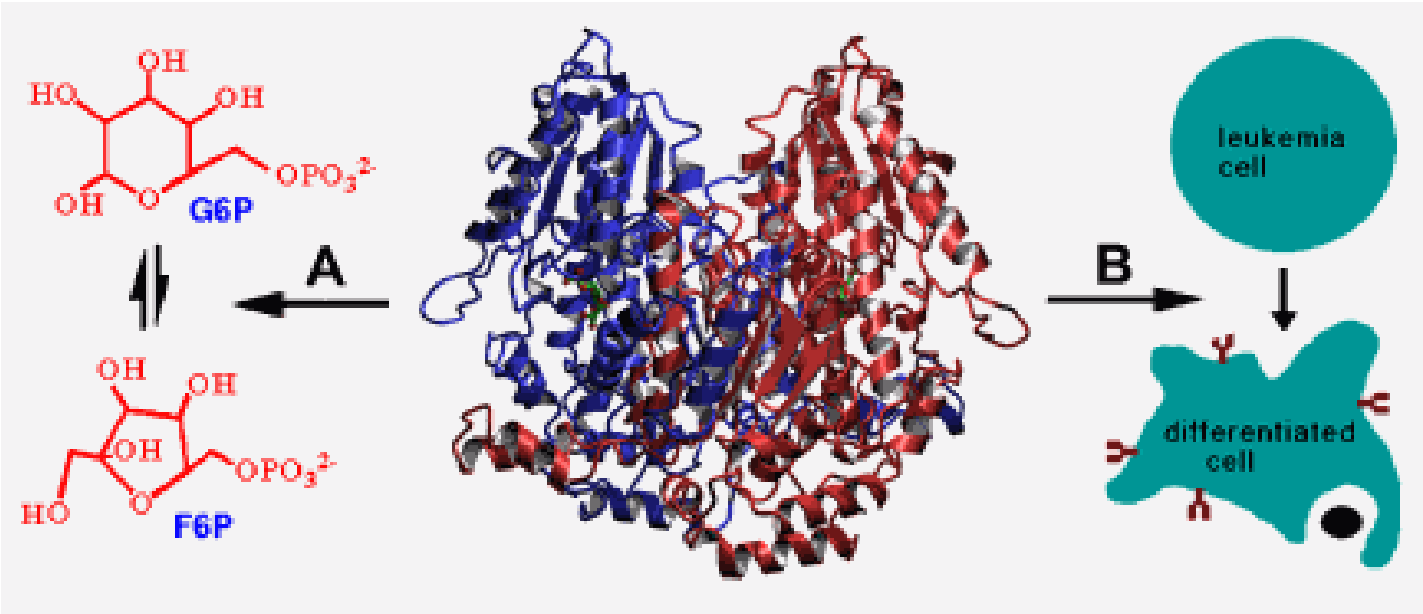
# Predicting Function

- To extract **biologically important information** from a protein sequence
  - **Structure prediction**
    - To predict **secondary structure** + solvent accessibility + transmembrane helices…(above)
  - **Function prediction**
    - Case-specific prediction
      - Did **not** result in automated tools for **function prediction**, but by
        - **Annotation transfer**
        - **Motifs & patterns**

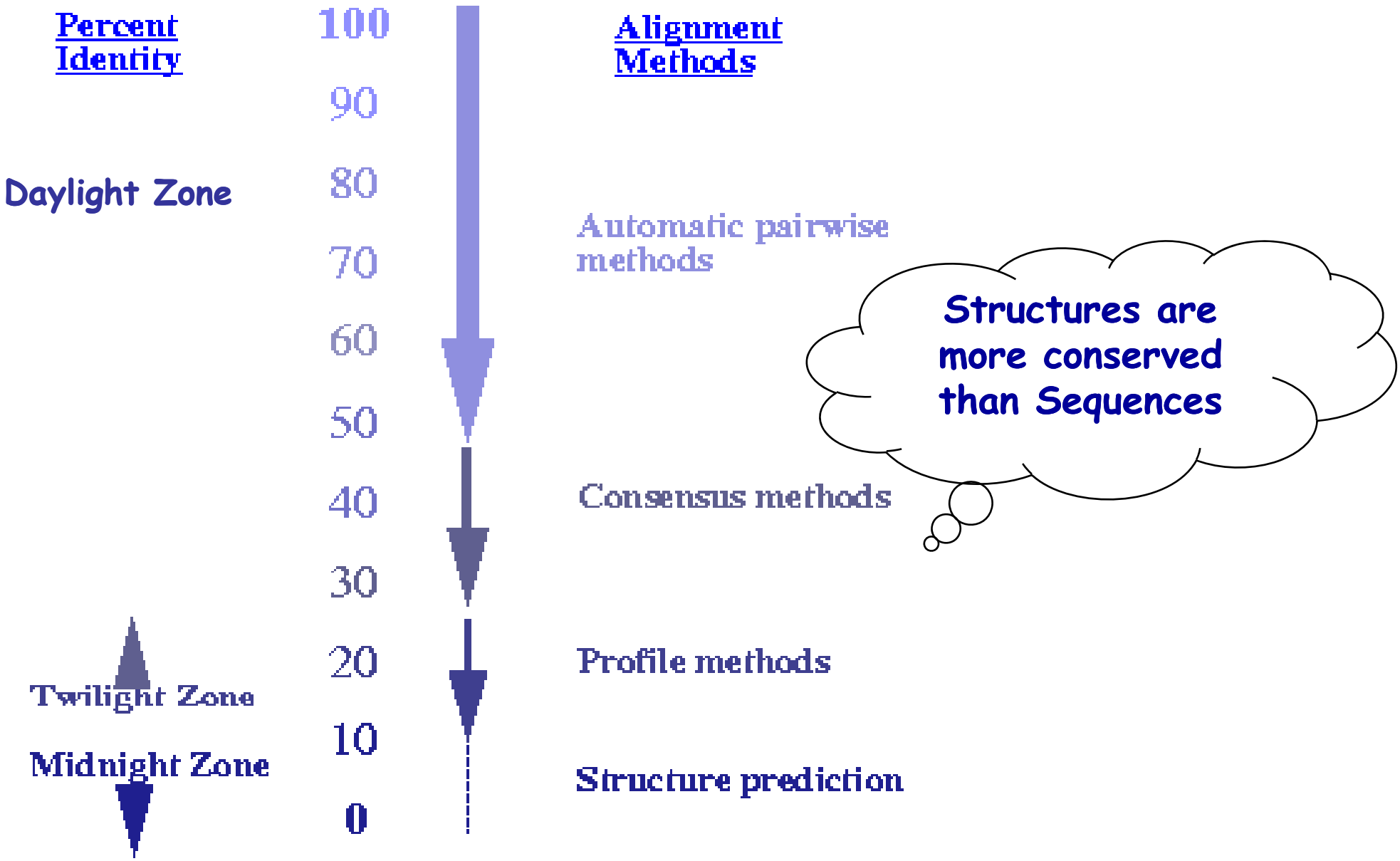
# Annotation Transfer (2)




- **Thousands** of proteins have been characterized **experimentally** in the lab
  - Their functions have been recorded **in various protein databases**
  - **Similarity at the sequence level** implies **similarity of function** (Next slide)
    - If **a newly discovered protein** bears high **similarity** to another, well-characterized one, it is reasonable to assume that they have a similar function
  - Some **notable exceptions**
    - Proteins have **different functions** depending on **their cellular location**
      - **"moonlighting proteins"** (Jeffery 2003)
        - **Not** one-gene-one-protein (follow the changes of the organism in its **physiological** & **pathological** conditions"

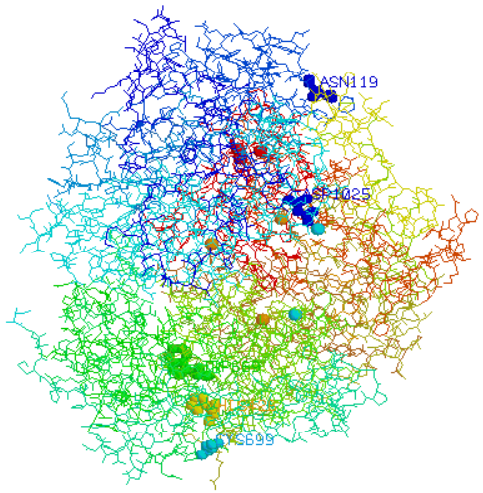



Moonlighting proteins. Moonlighting proteins have multiple, seemingly unrelated functions **not** due to gene fusions or alternative splicing. Like **PGI**, which is **a cytosolic enzyme** and an **extracellular cytokine**, dozens of other proteins have been found to **moonlight**. Connie coined the term **moonlighting proteins** and has written several review articles that develop **the idea of moonlighting proteins** and describe **additional moonlighting proteins** from the **literature**, how **they switch between functions**, how they might have evolved, and how they might **benefit the cell**. She is currently writing two additional invited articles and planning **computational studies** of the **sequences** and **structures** of known moonlighting proteins.


Percent Identity
Alignment Methods
100
90
80
70
60
50
40
30
20
10
0
Daylight Zone
Automatic pairwise methods
Structures are more conserved than Sequences
Consensus methods
Profile methods
Twilight Zone
Midnight Zone
Structure prediction

# Annotation Transfer (3)

- **Automatic** predictions, several **important points**
  - To define **a statistical threshold** of sequence similarity that permits the annotation transfer
    - This **threshold** differs from **one biological function to another**, needs to be determined *ad hoc* (specifically) for
      - **Each protein family** &
      - **Each biological function** (Rost et al. 2003)



# Motifs & Patterns (1)

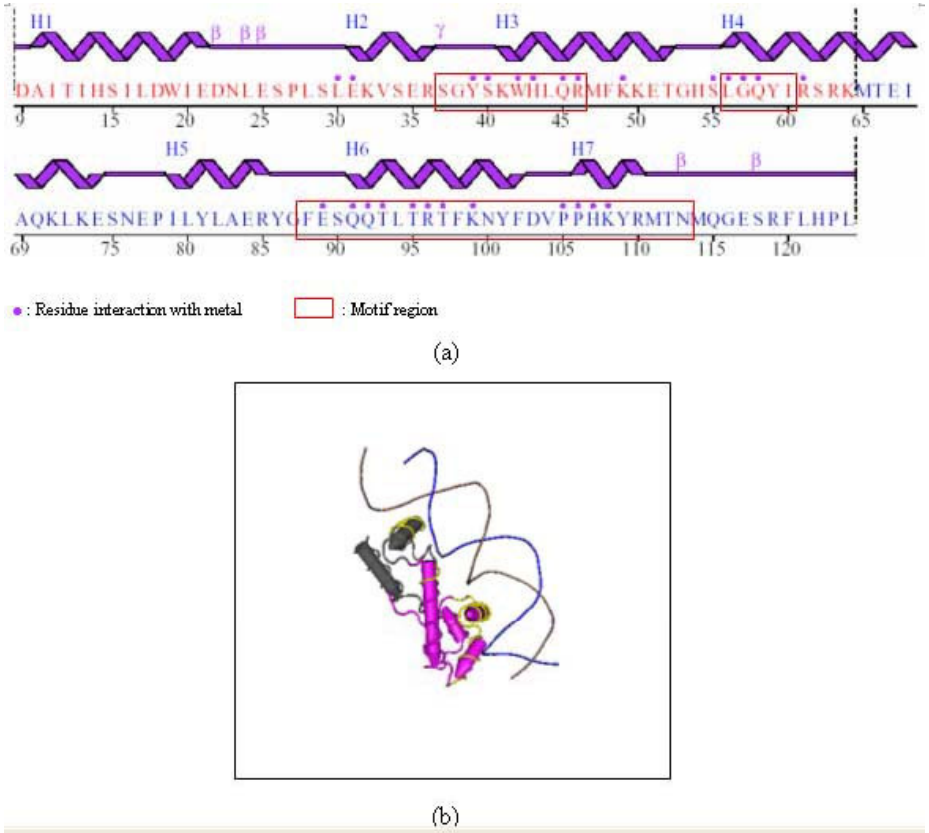
- To develop various methods **to identify functionally important residues** based on their **conservation** throughout **evolution**
  - Casari et al. 1995
- Conserved **residues** & conserved **sequence elements** , likely to be **important functionally**
  - But **no putative function** yet
- Sometimes, there is **divergence** between the **sequence** of a **newly discovered** protein & any **other previously annotated** protein ⇨ too wide to establish relatedness

# Levels of Protein Sequence and Structure Organization

| Level/ Database | Content | Example |
|---|---|---|
| Primary | Sequence | "AVILDRYFH" |
| Secondary | Motif | [AS]-[IL]2-X[DE]-R-[FYW]2-H |
| Tertiary | Domain, module | a,b,c or @, *, # |

# Motifs & Patterns (2)

- The existence of a **relatively short sequence motif** that is highly **conserved evolutionarily** & highly specific **functionally**
  - Helpful to reveal the **putative function**
    - E.g., a sequence element that appears in **many known** **DNA-binding proteins**




- **Homeodomain-like proteins**: 3 helices containing **a helix-turn-helix (HTH) DNA binding motif** (homeodomian in eukaryotes)

# Motifs & Patterns (3)

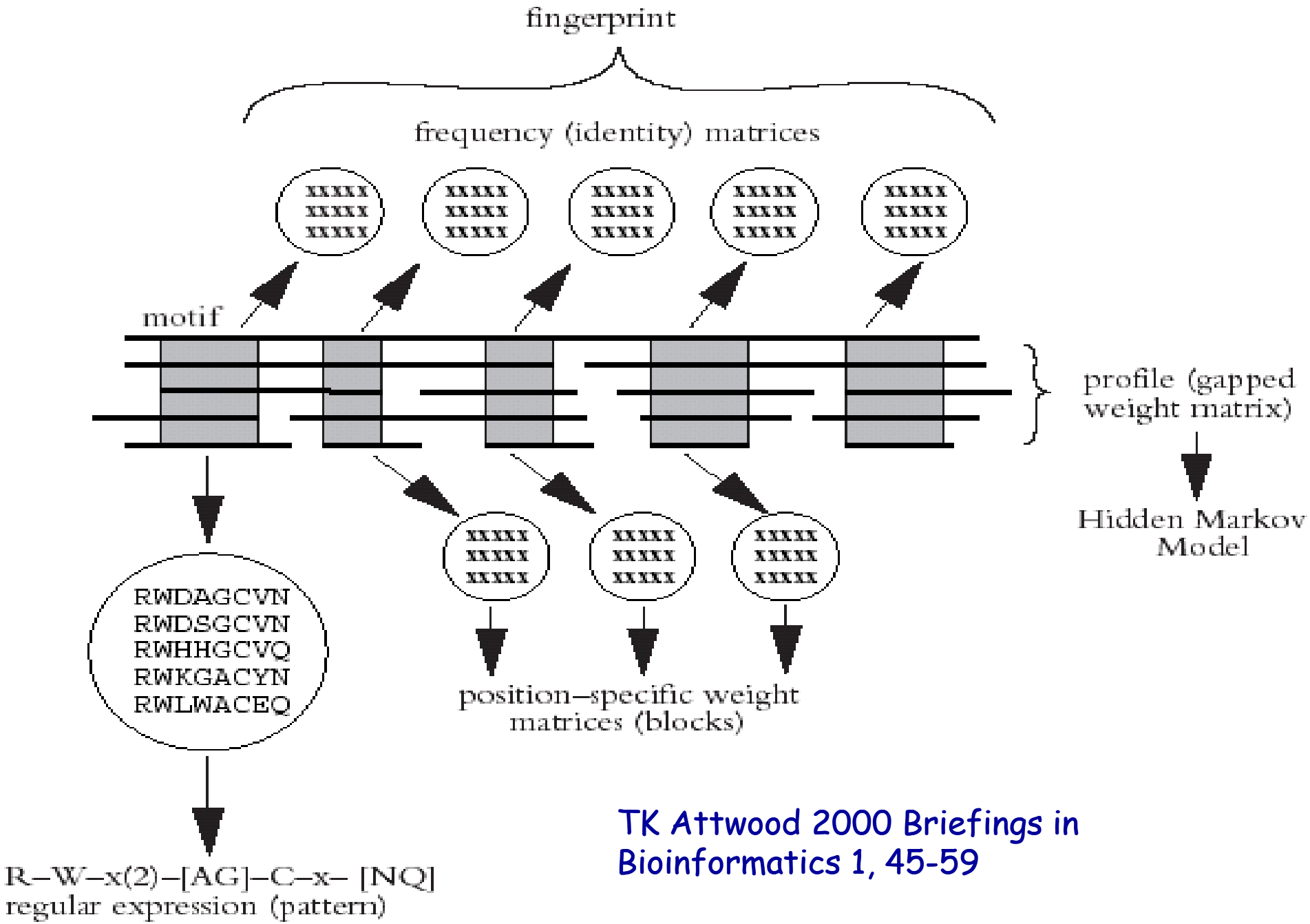
- Several **databases** offer **large libraries** of these characterized sequence motifs that have been collected either
  - **Manually** by experts or
  - Automatically by **pattern-searching algorithms**

- Many of these **libraries** include **a searching tool**
  - To submit a sequence of interest to determine whether **an interesting motif** is contained within their protein
    - Some important insight into **its structure** & **function**

# Terminology - Profile

- A **numerical** representation of a multiple sequence alignment (MSA)
  - **Intrinsic sequence information** that represents the common characteristics of that particular collection of sequences, frequently **a protein family**

- By using a profile, one is able to find **similarities** between sequences **with little or no absolute sequence identity** ⇨ to identify **distantly related proteins**




TK Attwood 2000 Briefings in Bioinformatics 1, 45-59

# Terminology – Pattern/Signature

- The common characteristics of **a protein family** or a multiple sequence alignment (MSA)
  - Simply providing **a shorthand notation** for what residues can be present at any given position
  - E.g., the pattern **[IV]-G-x-G-T-[ LIVMF] –x(2)-[GS]**
    - The first position could contain either an isoleucine or a valine, the second position could contain only a glycine…
    - {I}: residues **forbidden**
    - An **x**: any residue can appear at that position
    - x(2): two positions can be occupied by any amino acid, the number reflecting simply the length of the nonspecific run

# Methods – PROSITE (1)

- Catalog of biologically **significant sites** through the use of **motif**, **sequence profiles**, **patterns** (Falquet et al. 2002; Hulo et al. 2004)

- **Each entry in PROSITE**
  - Liking to an "annotation **document**"
    - Detailed information on the protein family that is characterized by **the profile or pattern**
    - **Annotations**
      - **Taxonomic** range
      - Domain **architecture** views
      - Known **biological** functions
      - Any **experimentally determined three-dimensional** structures
      - Relevance **references** from the literature

# Methods – PROSITE (2)

- **Each entry in PROSITE**
  - Information about how well the **pattern or profile** characterizes **the protein family** being described
    - The number of **true & false positives**, as well as false negatives
      - Which allow the users to determine the **specificity** of a given **profile** or **pattern**
  - A complete list **Swiss-Prot** sequences matching the profile or pattern

Cdc13p_yeast

USERSEQ1 (924 aa)

```
MDTLEEPECPPHKNRIFVSSSKDFEGYPSKAIVPVQFVALLTSIHLTETKCLLGFSNFERRGDQSQ
EDQYLIKLKFKDRGSERLARITISLLCQYFDIELPDLDSDSGASPTVILRDIHLERLCFSSCKALY
VSKHGNYTLFLEDIKPLDLVSVISTISTKSTNSSKHSSSELISECDLNNSLVDIFNNLIEMNRDEK
NRFKFVKLIHYDIELKKFVQDQQKVLSQKSKAAAINPFFVPNRLGIPYIESQNEFNSQLMTLNVDE
PTTDISNMGEEMHDSADPIEDSDSSTTSSTGKYFSSKSYIQSQTPERKTSVPNNWHDDDSGSKRKR
KLSFHSPNASSIRKAISYEQLSLASVGSVERLEGKIVGMNPPQFASINEFKYCTLKLYFTQLLPNV
PDKVLVPGVNCIEIVIPTRERICELFGVLNCQSDKISDILLLEKPDRISVEVERILWDNDKTASPG
MAVWSLKNISTDTQAQAQVQVPAQSSASIDPSRTRMSKMARKDPTIEFCQLGLDTFETKYITMFGM
LVSCSFDKPAFISFVFSDFTKNDIVQNYLYDRYLIDYENKLELNEGFKAIMYKNQFETFDSKLRKI
FNNGLRDLQNGRDENLSQYGIVCKMNIKVKMYNGKLNAIVRECEPVPHSQISSIASPSQCEHLRLF
YQRAFKRIGESAISRYFEEYRRFFPIHRNGSHLAKLRFDEVKHEPKKSPTTPALAEHIPDLNADVS
SFDVKFTDISSLLDSSARLPRPQQTHKSNTLYSCEGRIIAIEYHASDLCFHITNELPLLQTRGLAP
ERVLQLHIITSKNFAYFFNRSSAYLQRQPLEEKYTQLAQFLGHSFKFNITSSLTLFPDTTVALQIW
CPIECTFRELQQQLAHPKVAAAPDSGSLDCAINATVNPLRLLAAQNGVTVKKEEDNDDDAGAVPTS
```

ruler: 1 100 200 300 400 500 600 700 800 900 10

**hits by patterns:** [1 hit (by 1 pattern) on 1 sequence]

Hits by PS00141 **ASP_PROTEASE** *Eukaryotic and viral aspartyl proteases active site* :

USERSEQ1 (924 aa)

**103 - 114:** LDSDSGASPTVI

# Methods - *Pfam* (1)

- A collection of **protein domain families** (Sonnhammer et al. 1997; Bateman et al. 2004)

- **Each Pfam entry**
  - **A multiple sequence alignment (MSA)** of a protein domain or conserved region

- **Two Pfam databases**
  - **Pfam-A**
    - Manually
  - **Pfam-B**
    - Automatic clustering of **the ProDom database**

# Methods - *Pfam* (2)

- More than **80% of all Swiss-Prot** & **TrEMBL entries** are associated with **a Pfam entry**

- **Profile HMMs** derived from **the multiple sequence alignments** associated with each entry can be used to deduce whether **a new sequence of interest** can be assigned to an already-characterized family

- Particularly useful for when **similarity** can not be deduced through the use of **simple BLAST or FASTA search**

- **Keyword** & **sequence searches**


**Trusted matches - domains scoring higher than the gathering threshold (A)**

| Domain | Start | End | Bits | Evalue | Alignment | Mode |
|---|---|---|---|---|---|---|
| CDC13-DNA-bind | 504 | 686 | 442.60 | 4.5e-130 | Align | ls |

**Matches to Pfam-B**

| Domain | Start | End | Alignment |
|---|---|---|---|
| Pfam-B_137977 | 687 | 872 | Align |

**Potential matches - Domains with Evalues above the cutoff**

| Domain | Start | End | Bits | Evalue | Alignment | Mode |
|---|---|---|---|---|---|---|
| DUF560 | 12 | 24 | 11.00 | 0.0078 | Align | fs |
| UPF0227 | 41 | 69 | 4.80 | 0.36 | Align | fs |
| gp32 | 325 | 335 | 5.00 | 0.87 | Align | fs |
| ArsC | 334 | 345 | 4.70 | 0.65 | Align | fs |
| ArsC | 678 | 685 | 0.30 | 12 | Align | fs |

CDC13_YEAST

# Methods – *InterPro* (1)

- **An integrated resource** for information about **protein families, domains and functional sites**, bring together information from a number of protein domain-based resources (APweiler et al. 2000)
  - PROSITE
  - **PRINTS**
  - Pfam
  - **ProDom**

- To provide a **unified launching point** from which users can explore **protein families** further, moving from the ***InterPro*** entry to any of the sources from which *InterPro* entries are derived

# Methods – *InterPro* (2)

- **Searches**
  - Text
  - Sequence

- Example (InterPro Scan)
  - An InterPro summary page for CDC13_YEAST
    - Click on IPR001969
    - All IntroPro pages give the name of **the family** links to
      - **Signatures** that characterize members of the family
      - Information on **"children" of this protein family**
        - Used to indicate **protein family** & **subfamily** relationships

    - Links to other databases
      - **PDB, CATH, SCOP, BLOCKS and PROSITE**

# Methods – *InterPro* (3)

- **Radial taxonomy display**
  - To provide a quick overview of **the taxonomic range** in which members of this protein family are observed
- **Graphical view**
  - The presence of **discrete motifs** in proteins belonging to this family
- **References**
  - More in-depth information about the domain

# Methods – *BLOCKS* (1)

- The concept of **blocks** to identify **a family proteins**, rather than relying on the individual sequences themselves (Pietrovsky et al. 1996)
  - **Conserved motifs** from proteins **in the same family** are aligned **without introducing gaps** ⇨ **block**
    - **Block = alignment** (**not the individual sequence** themselves)
  - An individual protein can **contain one or more blocks**, corresponding to each of its **functional** & **structural motifs**
- The BLOCKS database is derived from the entries in ***InterPro***

# Methods – *BLOCKS* (2)

- **BLOCKS search**
  - The query sequence is **aligned against all blocks in the database** at all possible positions
    - For each alignment, **a score** is calculated using a **position specific scoring matrix (PSSM)**
  - Searches can be performed **optionally** against **the PRINTS database,** including information on more than **300 families** that do not have corresponding entries in BLOCKS
  - **Better way is to search both**
- **BLOCK Maker** (the same used to construct the database)
  - To submit **a set (≥3) of unaligned**, related protein sequences to **detect conserved blocks** within the sequence set

# Evaluation of Performance

- To **analyze**, not predict sequences
  - **Impossible to assess**
  - Accuracy & predictive power are **irrelevant**
  - **All complementary to each other**, the output could be **integrated** more straightforwardly

- **Pfam vs. PROSITE**
  - **Pfam:** find **a few long motifs** that can help associate the query sequence with **a well-characterized family** & predict its function

  - **PROSITE:** more short ones

# Subcellular Localization

- Annotation transfer yield no results…

- **History**
  - 2nd structure or topology
  - Structural families or folds
  - **Subcellular localization**
    - To narrow down its putative function

# Subcellular Localization – Methods (1)

- PSORT
  - Nakai & Horton 1999

- Searches for **features** characterizing that **taxonomic group** that may help deduce the subcellular localization of the protein
  - A library of **known signal peptides** & searches for them in the query sequence

  - Also checks **predicted structural features** (e.g., **topology**) that may indicate whether the protein is **soluble** or **embedded in membrane**

# Subcellular Localization – Methods (2)

- *SUBLOC*
  - Hua & Sun 2001
- Using **the amino acid composition alone**, applies support vector machine **(SVM)** to predict the subcellular locale of a protein
- **Prokaryotes**
  - Extracellular, periplasmic, or cytoplasmic
- **Eukaryotes**
  - Extracellular, mitochondrial, cytoplasmic or nuclear

# Subcellular Localization – Methods (3)

- *TargetP*
  - Emanuelsson et al. 2000
- Signal peptides at **the N-terminal end** of a protein
- A series of machine-learning algorithms, including **neural networks** & **SVMs**, to identify **signal peptides** of three types
  - **Chloroplast** transit peptides
  - **Mitochondrial** targeting peptide
  - **Secretory** pathway signal peptide

# Subcellular Localization – Methods (4)

- *LOC3D*
  - Nair & Rost 2003

- A database of predicted subcellular localizations for eukaryotic proteins of **known three-dimensional structure (PDB)**, three underlying methods
  - **PredictNLS**
    - Searching for a known **nuclear localization signal**
  - **LOChom**
    - Using homology to determine localization
  - **LOC3D**
    - A neural network-based prediction method

# Subcellular Localization – *LOCkey*

- A related service of *LOC3D*
  - Using **keywords** in **Swiss-Prot annotation** to predict the subcellular localization

- A **comprehensive** coverage of the methods & approaches available for the prediction of subcellular localization

# Evaluation of Performance

- **No** continuous, large-scale system that assesses and compares the performance of the different available methods

- Using **a large test set** & applying **the same statistical** analysis to all the methods
  - Nair & Rost (2003) found **a large variability in the accuracy** of the methods

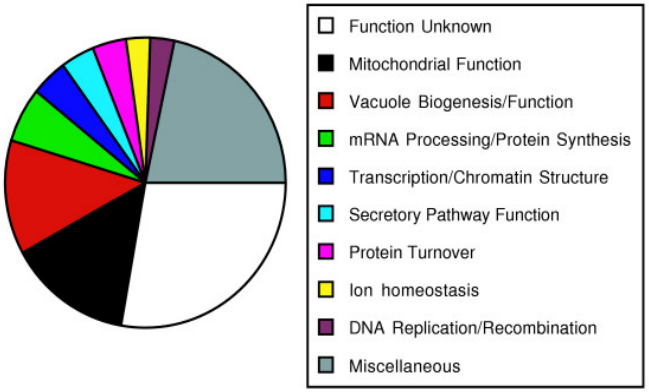
  - Best: extracellular >80%; cytoplasmic proteins > 50%; nuclear > 70%; mitochondrial > 60%

# Functional Class (1)

- The prediction of **subcellular localization** is part of the process of reducing the ambiguous term ***function*** to one that is better define

- The attempt to break down **the notion of function** into a series of **well-defined categories**
  - It is facilitated by **the definition** of **a set of functional classes** to which a protein can be assigned
  - **Software** were developed with the goal of assigning proteins to each one of the functional classes

# Functional Class (2)




- **Example**
  - A scheme of functional classes to annotate E. coli (Riley 1993; next slide)
  - **Some genome projects** adopted this general scheme, eventually making this the most widely used **terminology for the assignment of functional class**

| ENERGY | INFORMATION | COMMUNICATION |
|---|---|---|
| | | |
| METABOLISM | DNA & RNA | SIGNAL |
| AminoAcidBiosynthesis. | Replication. | Regulatory Functions. |
| Biosynthesis of Cofactors,<br><br>Prosthetic grps & Carriers. | Transcription | CellDivision. |
| Central Intermediary Metabolism. | TRANSLATION | Cell Killing. |
| Energy Metabolism. | Translation. | TRANSPORT |
| Fatty Acid andPhospholipid Biosynthesis. | PROTEINS | Transport and Binding Proteins. |
| Purines, Pyrimidines, Nucleosides, Nucleotides. | Chaperones. | ENVIRONMENT |
| | Detoxification. | Adaptation to Atypical Conditions and Others. |
| | Protein & Peptide Secretion and Transformation. | |

Equivalence between the classification of E. coli proteins *(Riley, 1993)* and the three- and eight- categories classification (STRUCTURAL proteins are not included in this schema)

# Prediction Methods - *EUCLID*

- Tamames et al. 1998
  - A basic **machine-learning algorithm** that learns, based on a **manually curated training set**, which combination of keywords is mostly likely to indicate that the protein belong to a certain functional type
    - **Keywords** in **Swiss-Prot** to assign a protein to one of Riley's functional class
    - >90% of the cases, the functional type that was determined by the automated method was identical to the one that was assigned to it **by a human expert**

  - *EUCLID* requires that some annotation **(Swiss-Prot keywords)** already be assigned to the sequence
    - Not really a method for prediction strictly from sequences

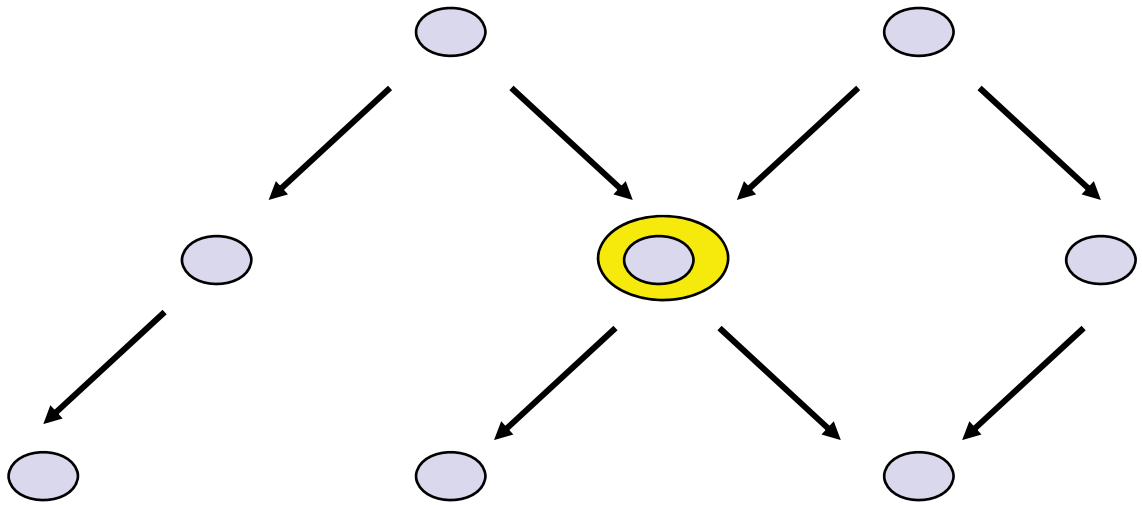
# Prediction Methods – *ProtFun* (1)

- Jensen et al. 2003

- A recent & promising step toward **the prediction of function from sequence**

- **Gene Ontology** (GO; Ashburner et al. 2000) in attempt to build a controlled vocabulary to describe genes & gene products
  - Each protein can be assigned to a certain molecular function
    - **Molecular function**
    - **Biological process**
    - **Cellular component**

# DAG Structure

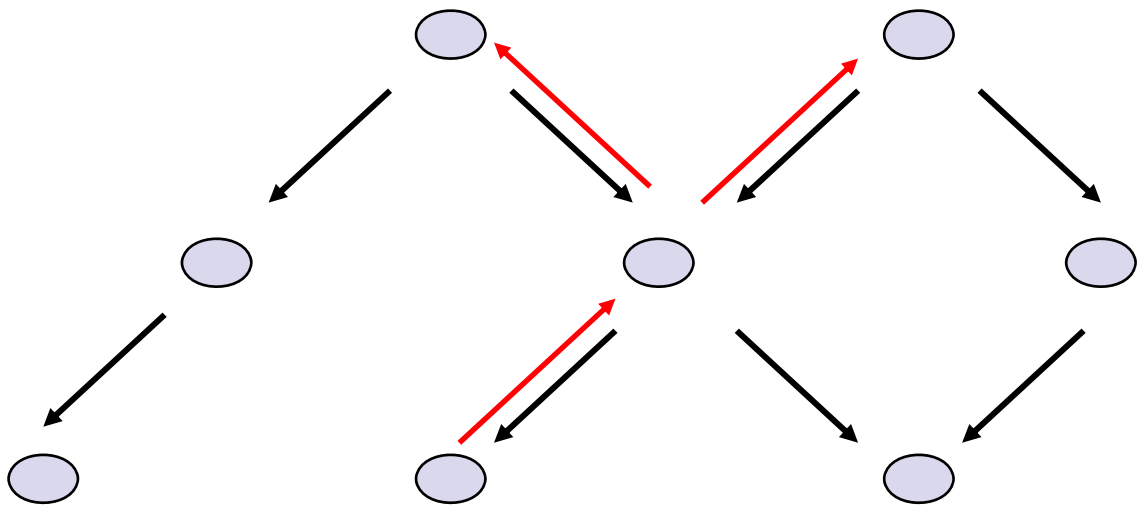


- Directed acyclic (環式) graph
  - Each child may have **one or more parents**

# Relationship Types

- **Is-a**
  - Subclass; a **is** one type of b
- **Part-of**
  - Physical part of (component)
  - Sub-process of (process)

# The True Path Rule



- Every **path** from a **node** back to the **root** must be **biologically accurate**

the Gene Ontology website - Microsoft Internet Explorer

05-17-2005 05:01:56

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛 媒體

網址(D) http://www.geneontology.org/ 連結

http://www.geneontology.org/

# GENE ONTOLOGY

OBO open biomedical ontologies

Open menus
Site Map
Home New FAQ

Downloads
Ontologies
Annotations
Database
Mappings to GO
GO Tools

Documentation

About GO
Terms Of Use
Contact GO
Report Errors

Archives

**Search** GO
go
GO terms
gene or protein name

**Search** Site
go

*What is the Gene Ontology?* | *Download the Ontologies*

The goal of the Gene Ontology project is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms to describe gene product attributes. GO is one of the controlled vocabularies of the Open Biomedical Ontologies.

- Submit new GO term suggestions via the Curator Requests Tracker at SourceForge. Help with new term submission is available.
- Send comments and questions to go@geneontology.org.

**Search GO**
GO!
GO term gene or protein name

This search uses the browser AmiGO. See the GO tools section for other GO browsers.

## What's New?

- We are pleased to announce the 2005 GO Users Meeting will be held as part of the MGED 8 meeting in Bergen, Norway, in mid September. See the meeting information page for more on the meeting and to submit your abstract. *(posted May 2, 2005)*

網際網路

# GO Evidence Code

From reviews or introductions

- IDA
  - Inferred from Direct Assay
- IMP
  - Inferred from Mutant Phenotype
- IGI
  - Inferred from Genetic Interaction
- IPI
  - Inferred from Physical Interaction
- IEP
  - Inferred from Expression Pattern

- TAS
  - Traceable Author Statement
- NAS
  - Non-traceable Author Statement
- IC
  - Inferred by Curator
- ISS
  - Inferred from **Sequence** or **Structural** Similarity

- IEA
  - Inferred from Electronic Annotation
- ND
  - Not Determined

From Primary Literature

Automated

# Prediction Methods – *ProtFun* (2)

- Currently, there are hundreds of GO categories
- ***ProtFun*** focuses on 347 of them & uses **complex systems of neural networks** to predict **the GO functional classification** of a protein from its sequence
  - Impressive accuracy >90%
- **Current coverage is only partial** & many of the query proteins are returned without any prediction at all

```
>Sequence

# Functional category                      Prob     Odds
  Amino_acid_biosynthesis                  0.158    7.190
  Biosynthesis_of_cofactors                0.087    1.203
  Cell_envelope                            0.065    1.061
  Cellular_processes                       0.027    0.373
  Central_intermediary_metabolism     => 0.241    3.830
  Energy_metabolism                        0.023    0.259
  Fatty_acid_metabolism                    0.019    1.436
  Purines_and_pyrimidines                  0.431    1.775
  Regulatory_functions                     0.334    2.075
  Replication_and_transcription            0.352    1.313
  Translation                              0.063    1.425
  Transport_and_binding                    0.165    0.402

# Enzyme/nonenzyme                         Prob     Odds
  Enzyme                              => 0.439    1.534
  Nonenzyme                                0.561    0.786

# Enzyme class                             Prob     Odds
  Oxidoreductase (EC 1.-.-.-)              0.024    0.114
  Transferase    (EC 2.-.-.-)              0.195    0.564
  Hydrolase      (EC 3.-.-.-)              0.160    0.505
  Lyase          (EC 4.-.-.-)              0.020    0.430
  Isomerase      (EC 5.-.-.-)              0.012    0.366
  Ligase         (EC 6.-.-.-)        => 0.118    2.318

# Gene Ontology category                   Prob     Odds
  Signal_transducer                        0.063    0.293
  Receptor                                 0.006    0.033
```

CDC13_YEAST