生物信息学

功能预测与注释

吴凌云

中国科学院数学与系统科学研究院

基因组注释

- Genome annotation
- 利用生物信息学方法，对基因组各组成部分进行识别，并对其生物学功能进行注释
- 主要内容
  - 基因识别与功能注释
  - 非编码基因的识别与功能注释
  - 调控元件的识别与功能注释
  - 影响染色体结构和动力学的序列

基因的识别与功能注释

- 基因预测
- 序列搜索
- 序列motif
- 直系同源序列聚类分析（COG）
- 亚细胞定位
- 结构比较
- 蛋白质组学

序列搜索

- 假设：序列相似=同源=功能相似
- 数据库
  - NCBI-NT（非冗余核酸序列数据库）
  - NCBI-NR（非冗余蛋白质序列数据库）
  - InterPro（Swissprot）（蛋白质序列数据库）
  - KEGG
  - PDBseq（已知三维结构的蛋白质序列数据库）

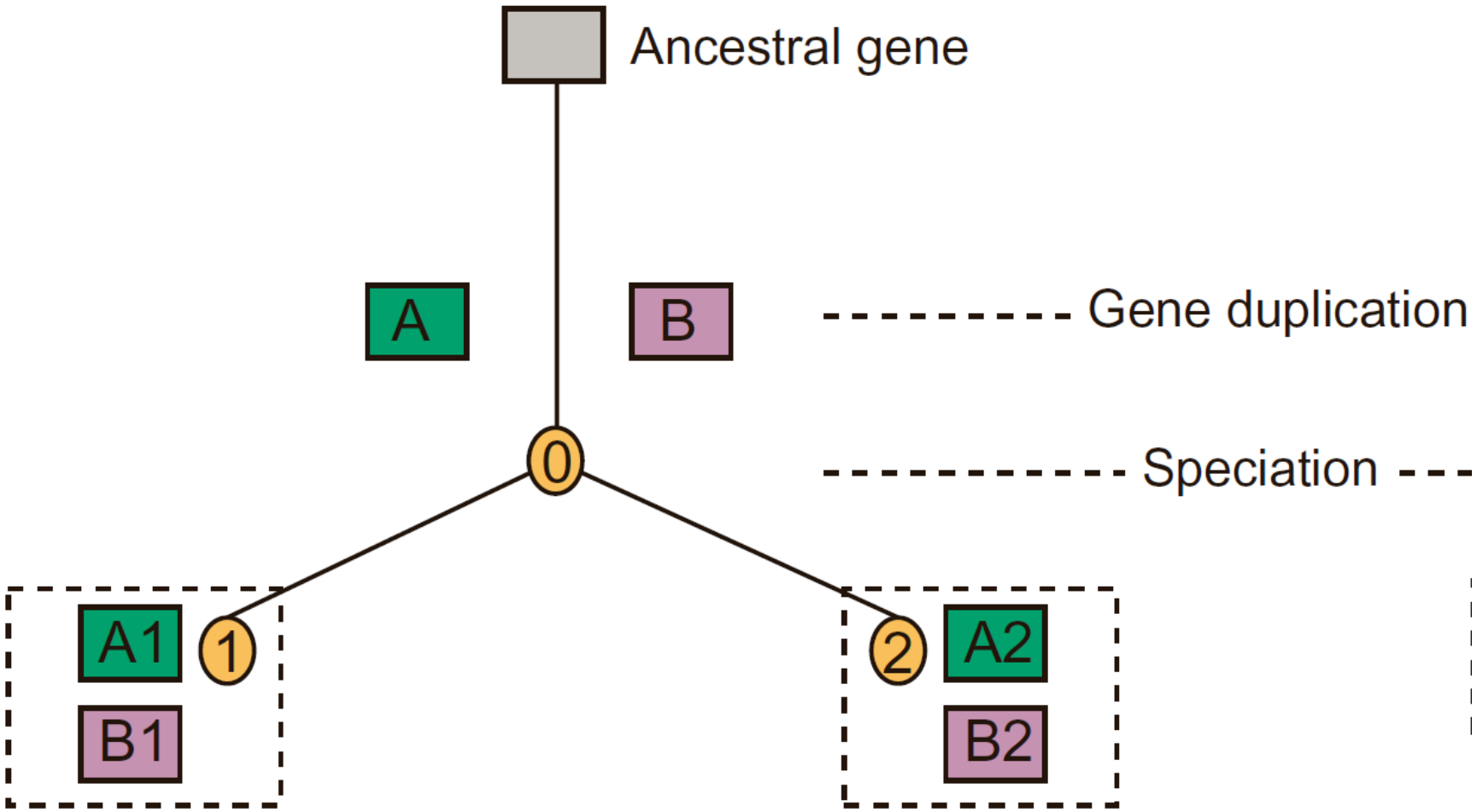
序列motif

- 查找序列上的局部特征
- 在序列同源性不明显的情况下使用
- Motif数据库构建
  - 对蛋白质家族成员进行多序列比对
- 数据库
  - Prosite

同源

- 直系同源（Ortholog）
  - 不同物种中由同一祖先进化而来的多个基因
  - 功能较一致
- 旁系同源（Paralog）
  - 同一基因组内由于基因复制而来的多个基因
  - 功能差异较大

直系同源与旁系同源


Ancestral gene
A
B
Gene duplication
0
Speciation
A1
1
B1
2
A2
B2

直系同源序列聚类分析

- 假设：直系同源=功能相似
- 数据库
  - COGs（Clusters of Orthologous Groups of proteins）
  - Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

亚细胞定位

- 假设：蛋白质的亚细胞定位与功能相关
- 通过预测亚细胞定位来预测功能

结构比较

- 假设：结构决定功能
- 预测未知基因的蛋白质结构，再通过结构比较预测其功能

蛋白质组学

- 假设：功能相关的蛋白质可能倾向于有相互作用
- 从蛋白质相互作用网络或者其他生物分子网络来预测蛋白功能

Function

- The word function within a biological context is an evolving concept and is used in many ways.
- Function can be described at many levels, ranging from biochemical function to biological processes, all the way up to the organism level.
- If only say a protein has some function, that has few meaning to biologist.

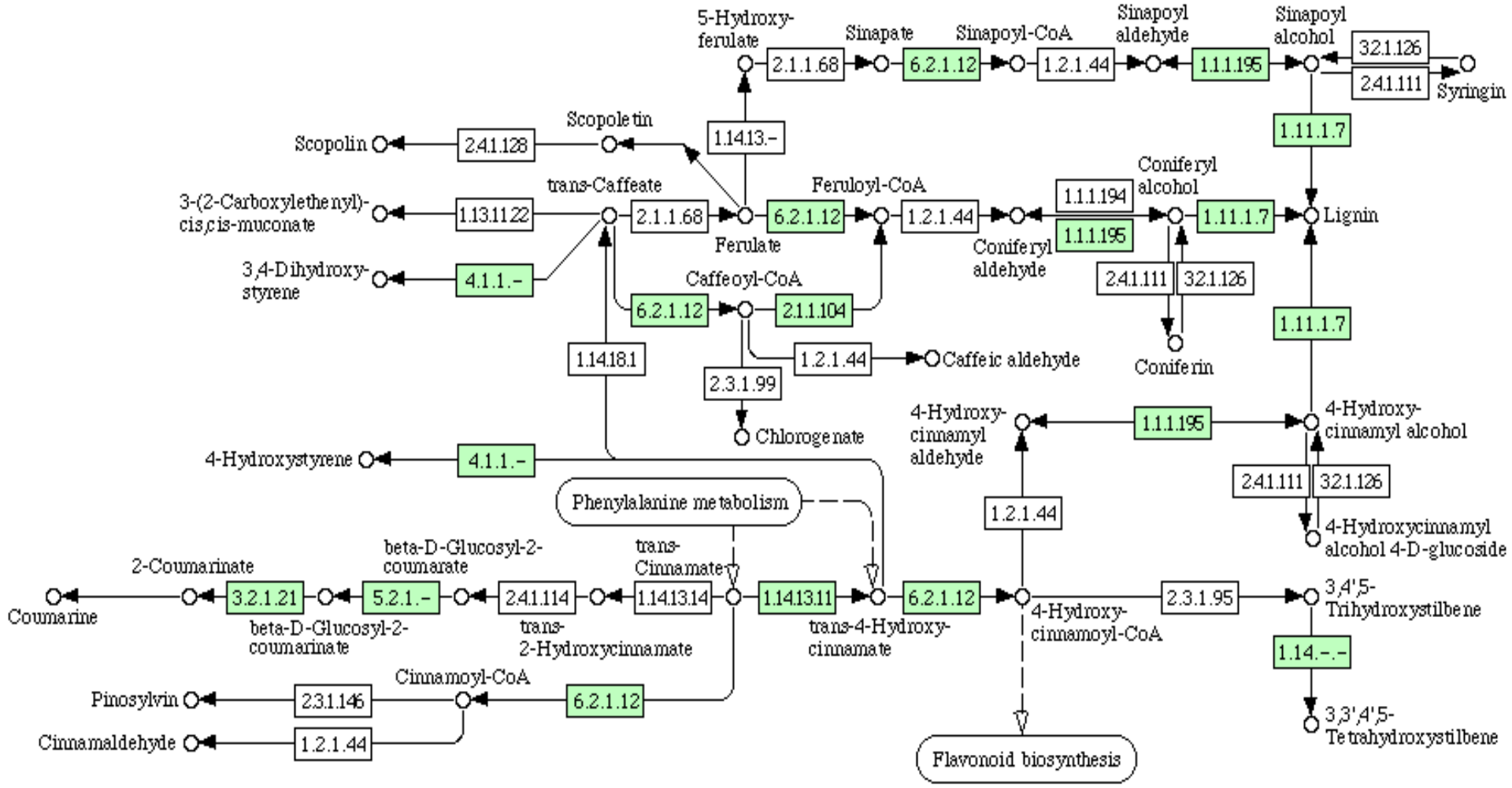
Classification of Function

- EC
  - Enzyme Commission scheme
- FunCat
  - MIPS Functional Catalogue
- GO
  - Gene Ontology

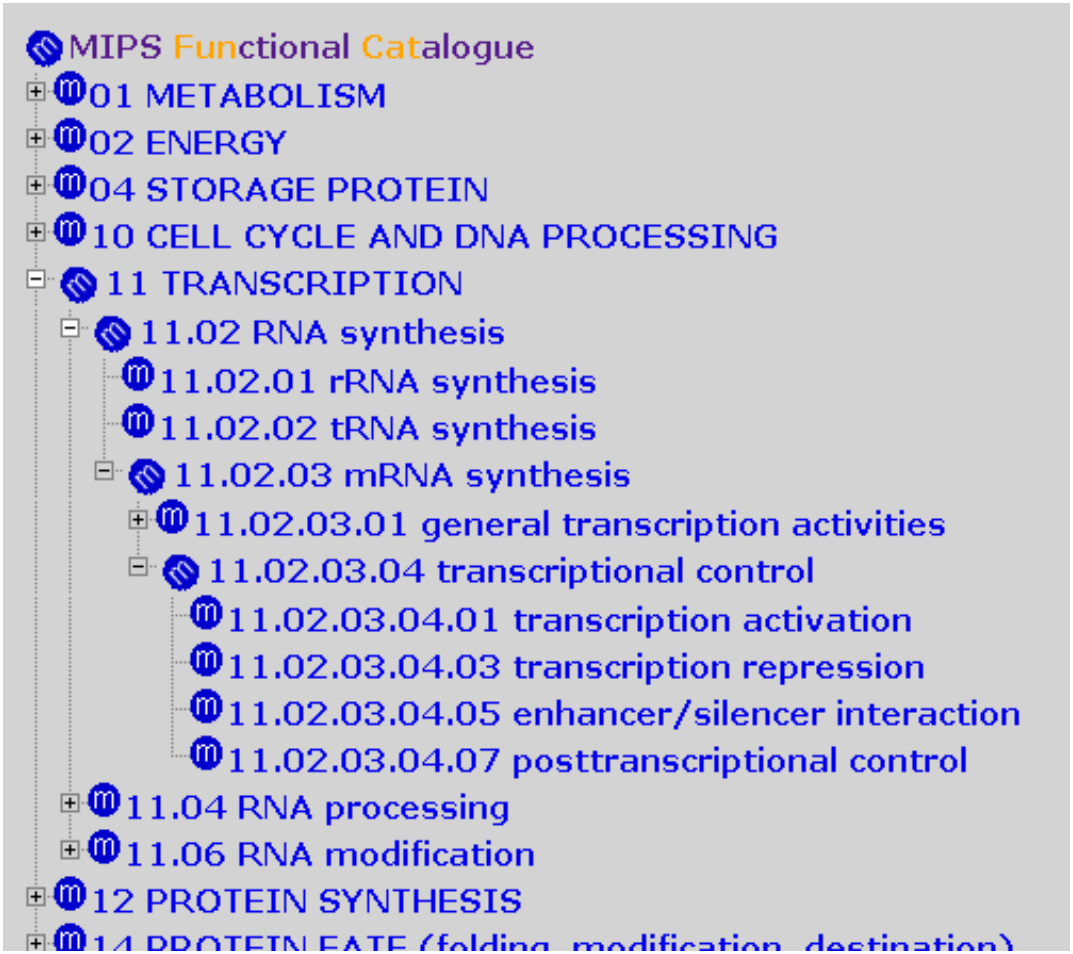
EC

- Enzyme Commission number
  - Based on the chemical reactions they catalyze.
  - Every EC number is associated with a recommended name for the respective enzyme.
  - Strictly speaking, EC numbers do not specify enzymes, but enzyme-catalyzed reactions.
  - If different enzymes (for instance from different organisms) catalyze the same reaction, then they receive the same EC number.

KEGG Patyway



MIPS FunCat


MIPS Functional Catalogue
01 METABOLISM
02 ENERGY
04 STORAGE PROTEIN
10 CELL CYCLE AND DNA PROCESSING
11 TRANSCRIPTION
11.02 RNA synthesis
11.02.01 rRNA synthesis
11.02.02 tRNA synthesis
11.02.03 mRNA synthesis
11.02.03.01 general transcription activities
11.02.03.04 transcriptional control
11.02.03.04.01 transcription activation
11.02.03.04.03 transcription repression
11.02.03.04.05 enhancer/silencer interaction
11.02.03.04.07 posttranscriptional control
11.04 RNA processing
11.06 RNA modification
12 PROTEIN SYNTHESIS

Gene Ontology (GO)

- Unify the representation of gene and gene product attributes across all species
  - Maintain and further develop its controlled vocabulary of gene and gene product attributes
  - Annotate genes and gene products, and assimilate and disseminate annotation data
  - Provide tools to facilitate access to all aspects of the data provided by the Gene Ontology project

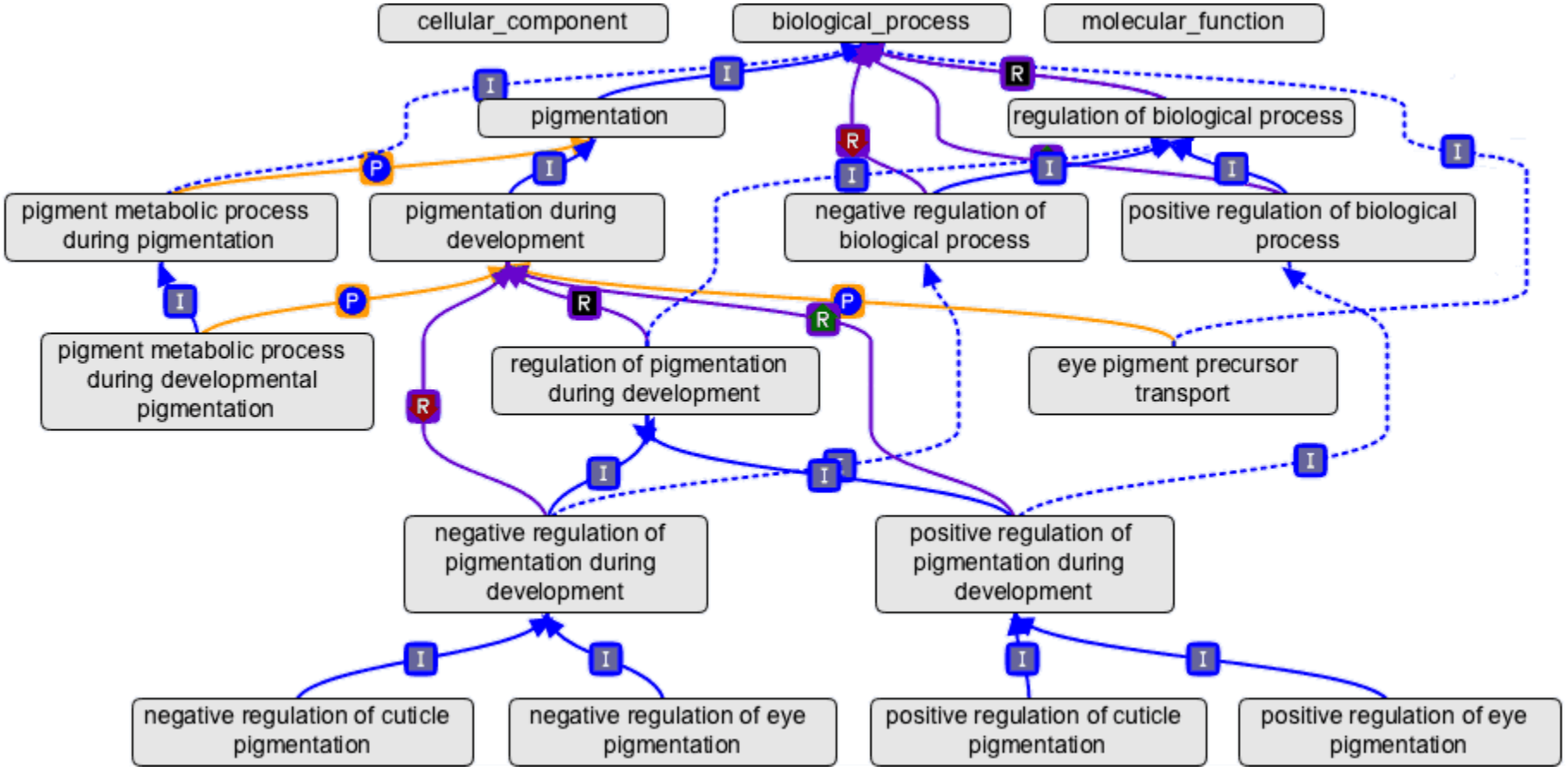
GO Domains

- Three separate GO domains
  - Molecular functions
  - Biological processes
  - Cellular components
- Each gene or gene product may
  - have more than one molecular function
  - take part in more than one biological process
  - act in more than one cellular component

Structure of GO

- Show the relation between different terms
  - One term may be a more specific description of another more general term
- Directed Acyclic Graph (DAG)
  - Similar to hierarchy
  - Allow a child node to have more than one parent

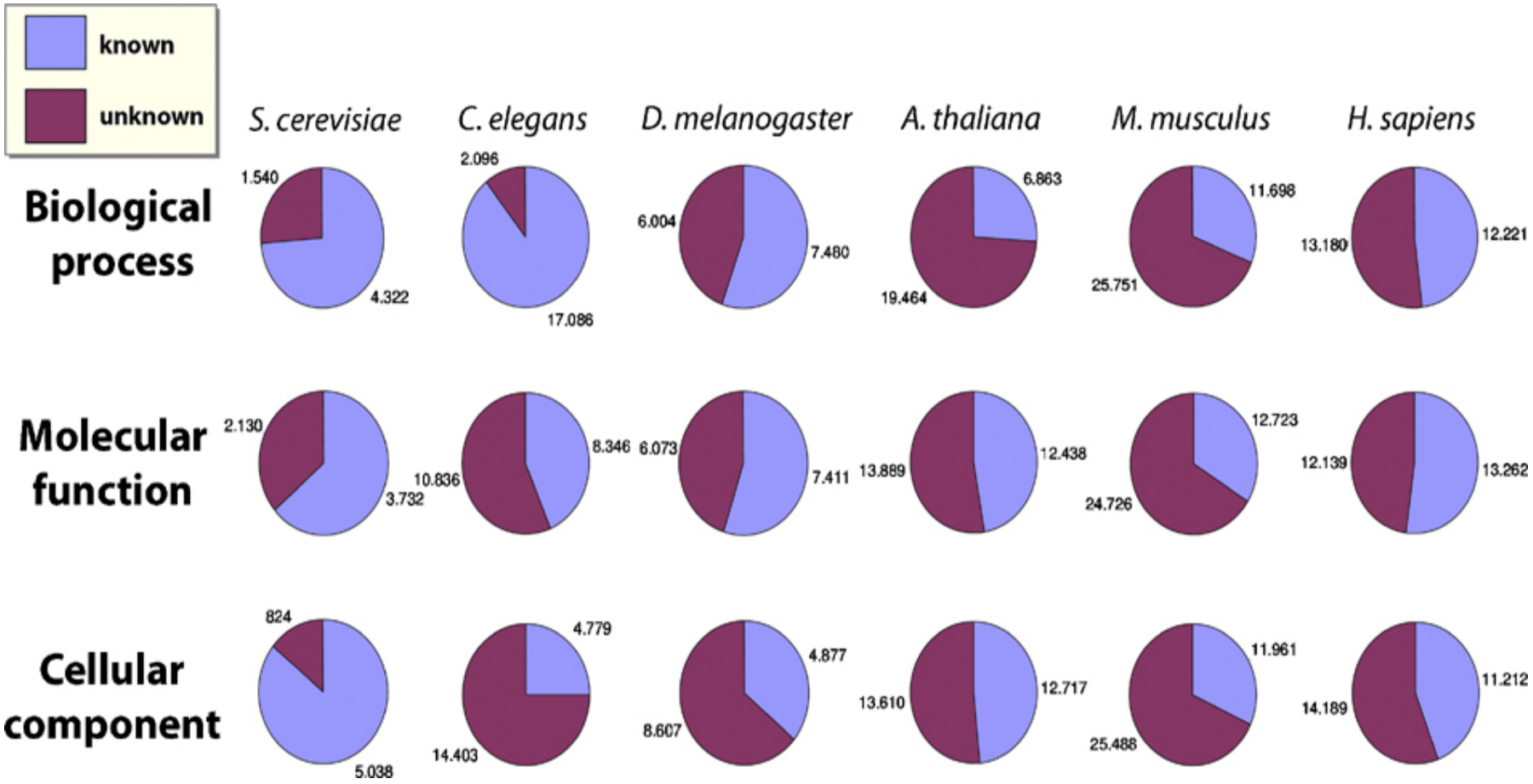
Example of GO Graph



Relations in GO

- Three relations
  - is_a (is a subtype of)
  - part of
  - Regulates, negatively regulates, positively regulates

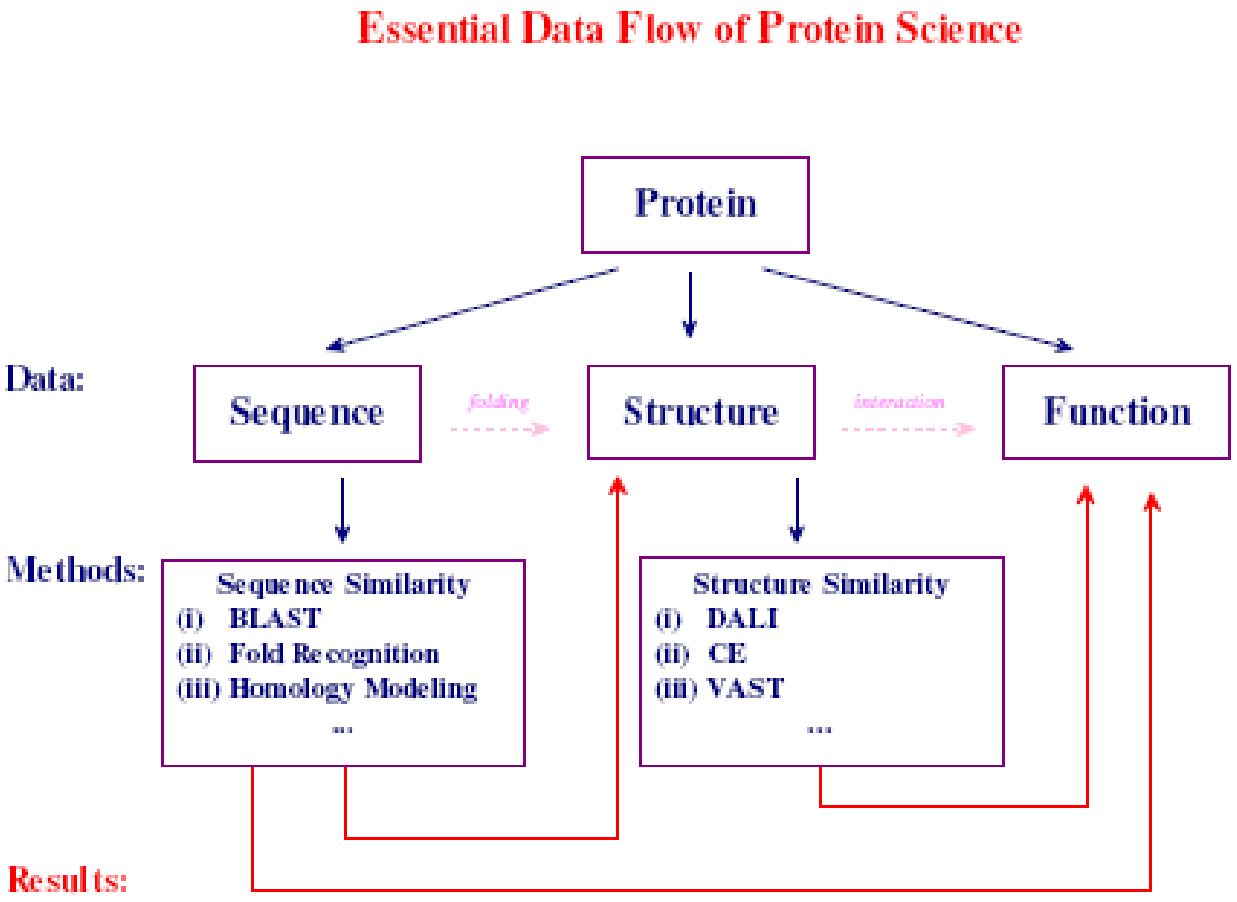
蛋白质功能注释情况


known
unknown
S. cerevisiae
C. elegans
D. melanogaster
A. thaliana
M. musculus
H. sapiens
Biological process
1.540
4.322
2.096
17.086
6.004
7.480
6.863
19.464
11.698
25.751
13.180
12.221
Molecular function
2.130
3.732
10.836
8.346
6.073
7.411
13.889
12.438
12.723
24.726
12.139
13.262
Cellular component
824
5.038
14.403
4.779
8.607
4.877
13.610
12.717
11.961
25.488
14.189
11.212

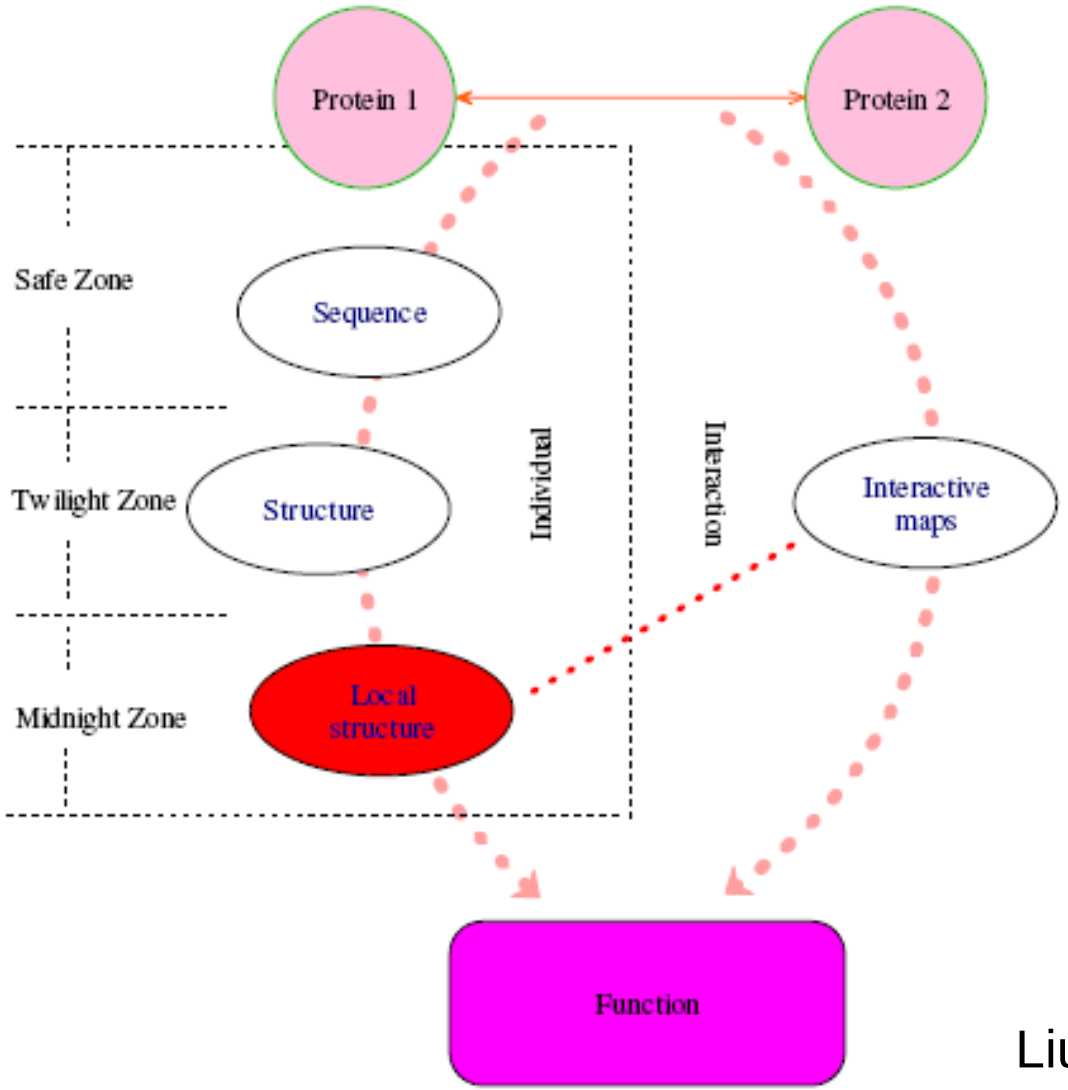
功能预测

- 数据
  - 序列
  - 整体结构
  - 局部结构
  - 蛋白相互作用
- 方法

Sequence, Structure, Function

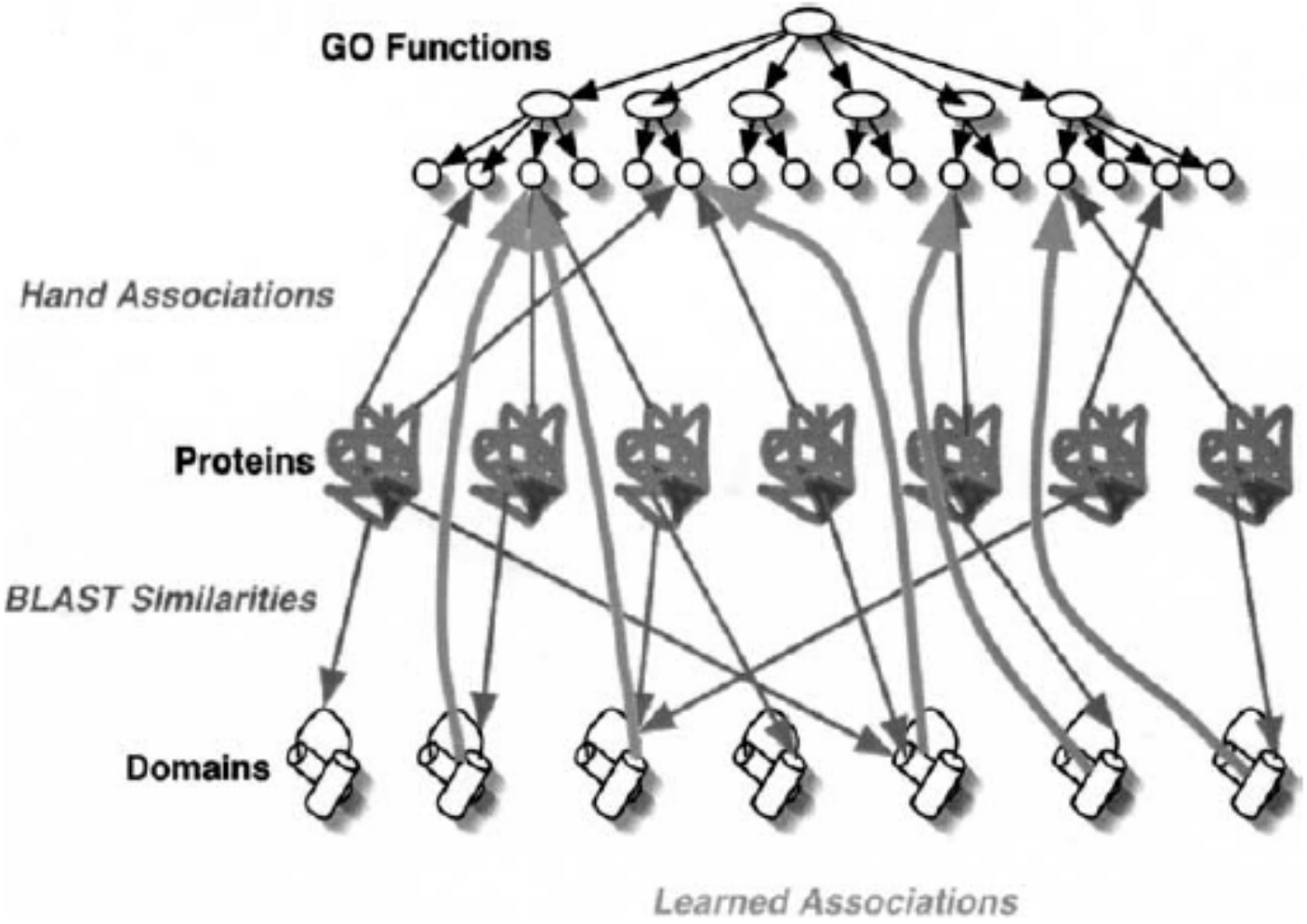


Function Annotation

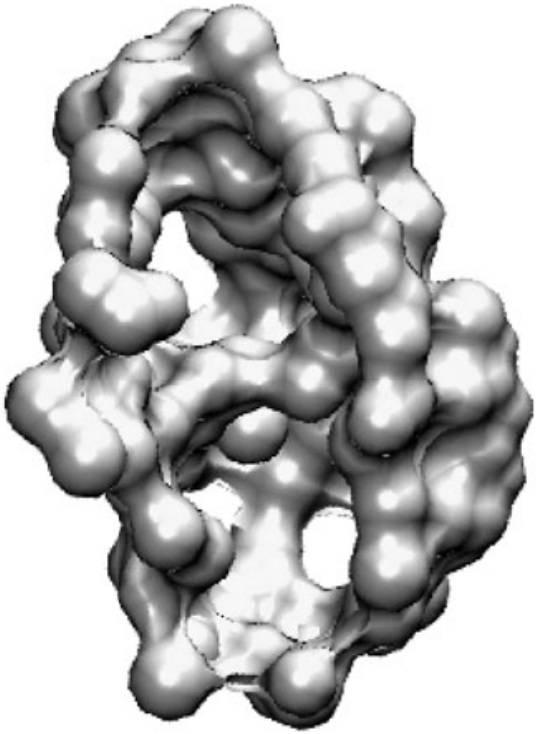


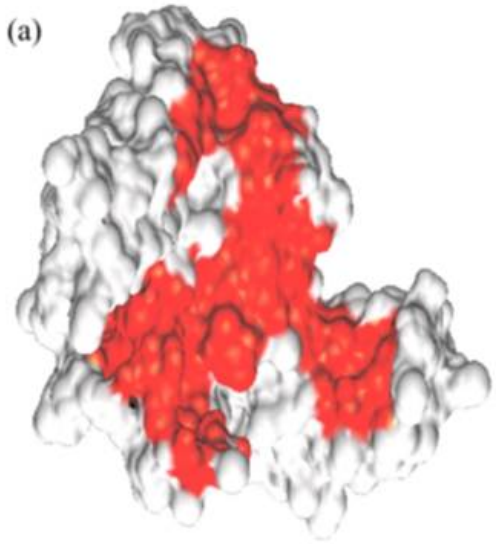
Liu *et al*., AA, 2008

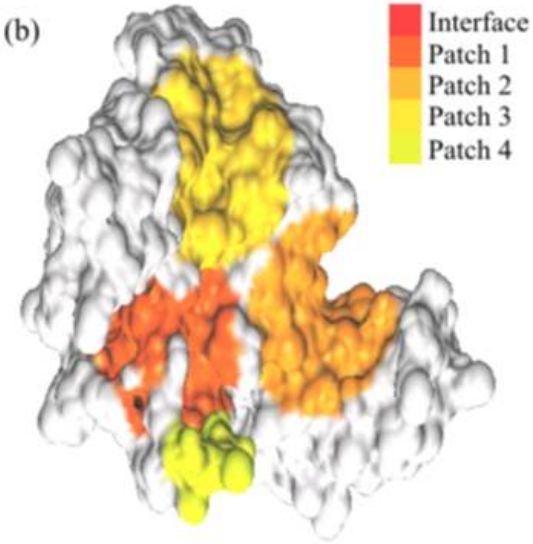
Prediction from Domain

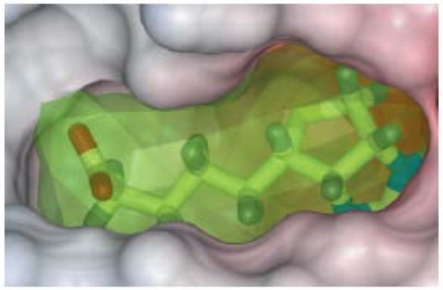


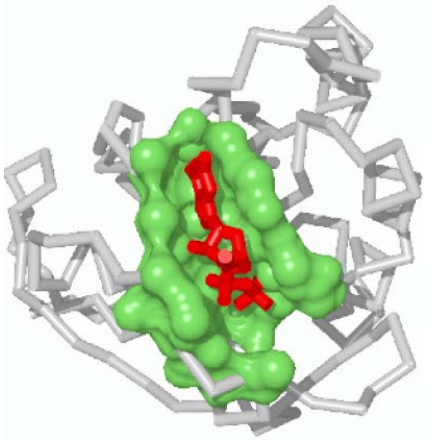
局部结构（Local Structure）



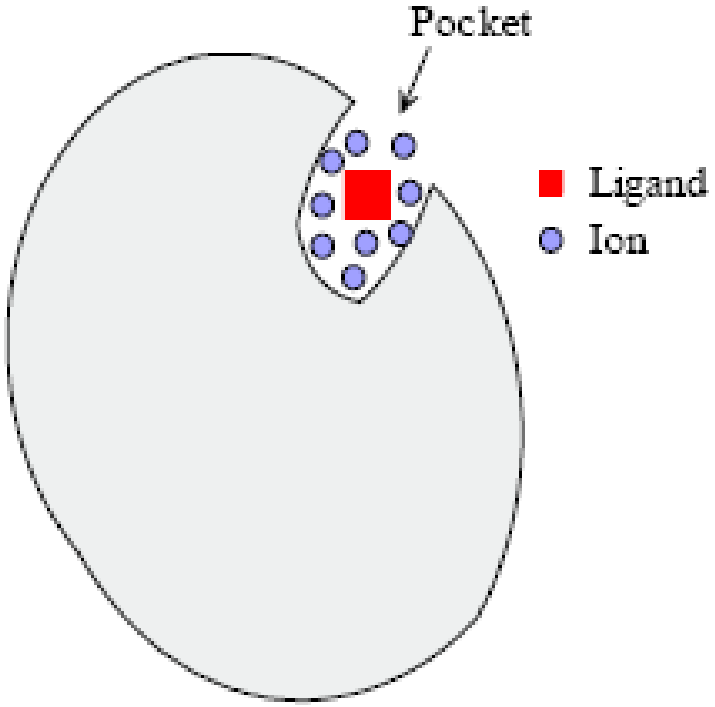




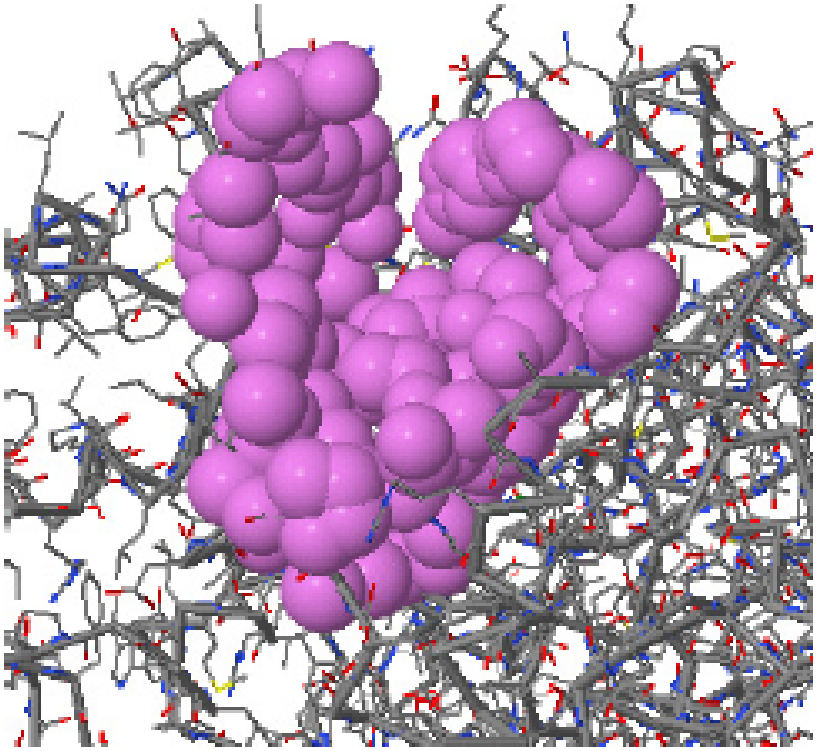




局部结构的表示



(a)



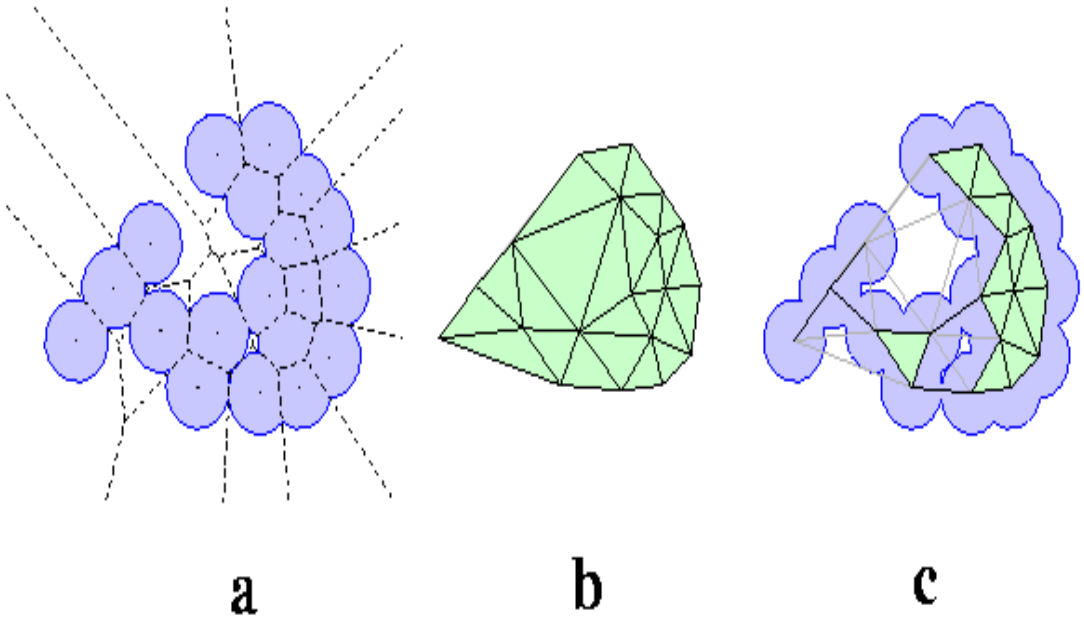
(b)

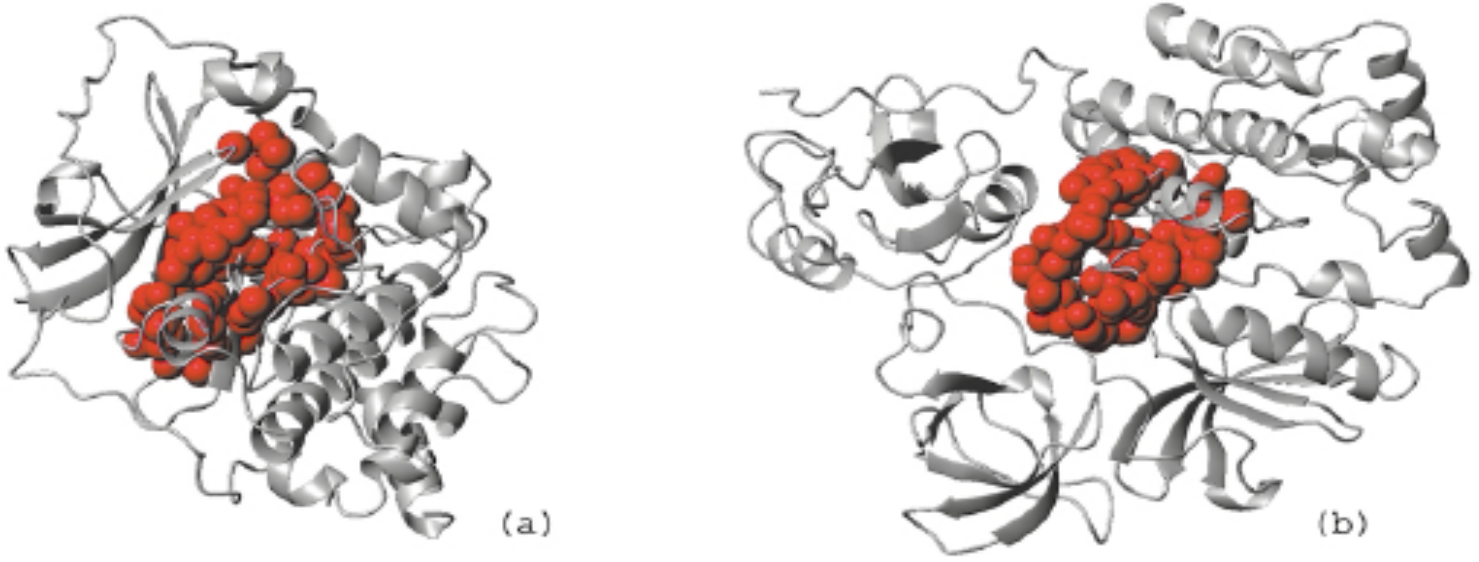
局部结构的定义

- Pocket
  - A pocket is an empty concavity on a protein surface into which solvent can gain access, i.e. these concavities have mouth openings connecting their interior with the outside bulk solution.
- Void
  - A void is an interior unoccupied space that is not accessible to the solvent probe. It has no mouth openings to the outside bulk solution.

局部结构的探测

- Computational Geometry
  - Voronoi Diagram
  - Delaunay Triangulation
  - Alpha Shape





(c)

>1cdk_A

GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHFAMKILD
KQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQI
VLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALG
VLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKDGVNDIKNHKWFATT
DWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF

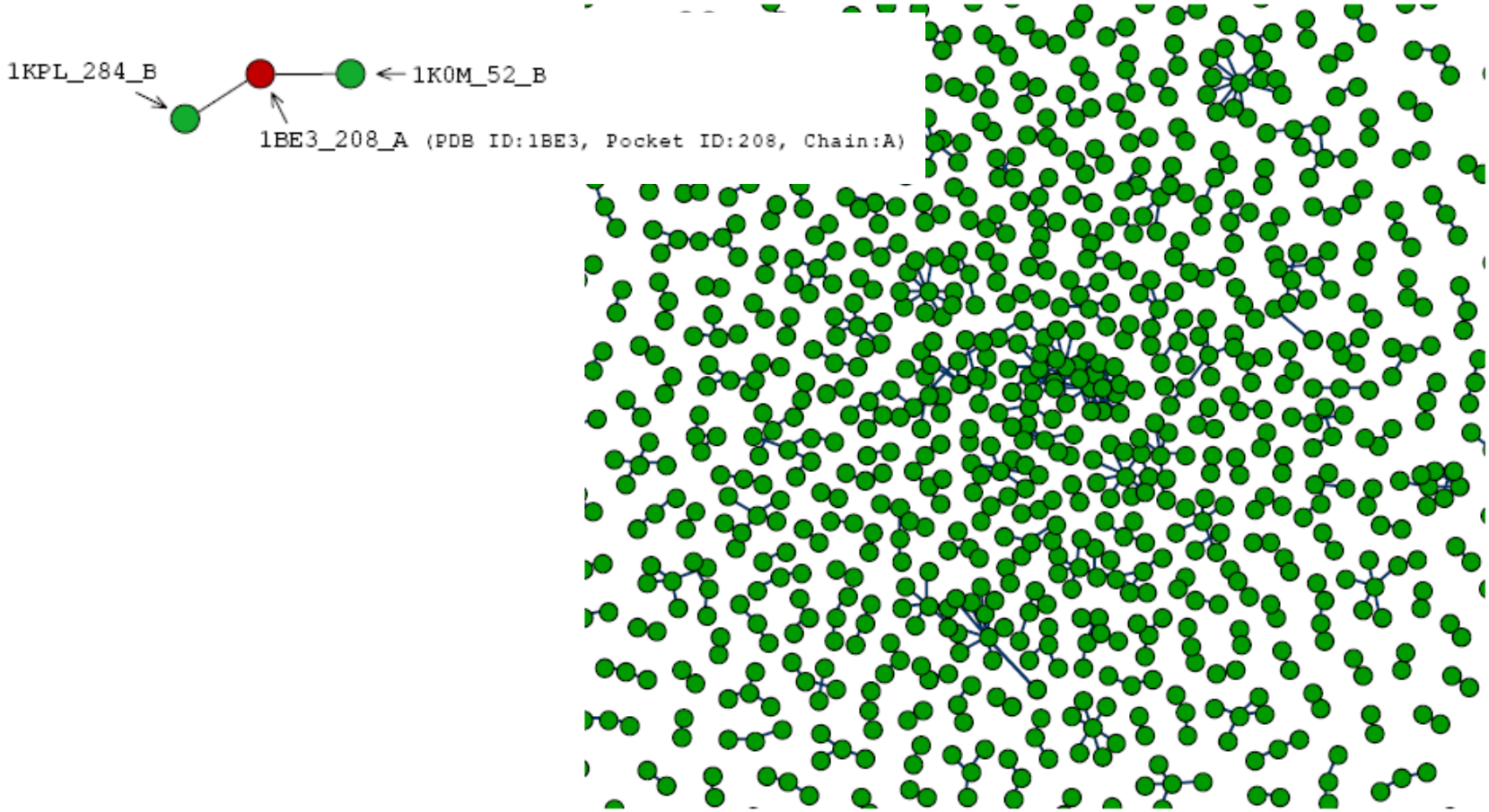
>2src_

MVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRR
ESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYS
KHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPEA
FLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMN
YVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELT
TKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQXQPGE
NL

(d)

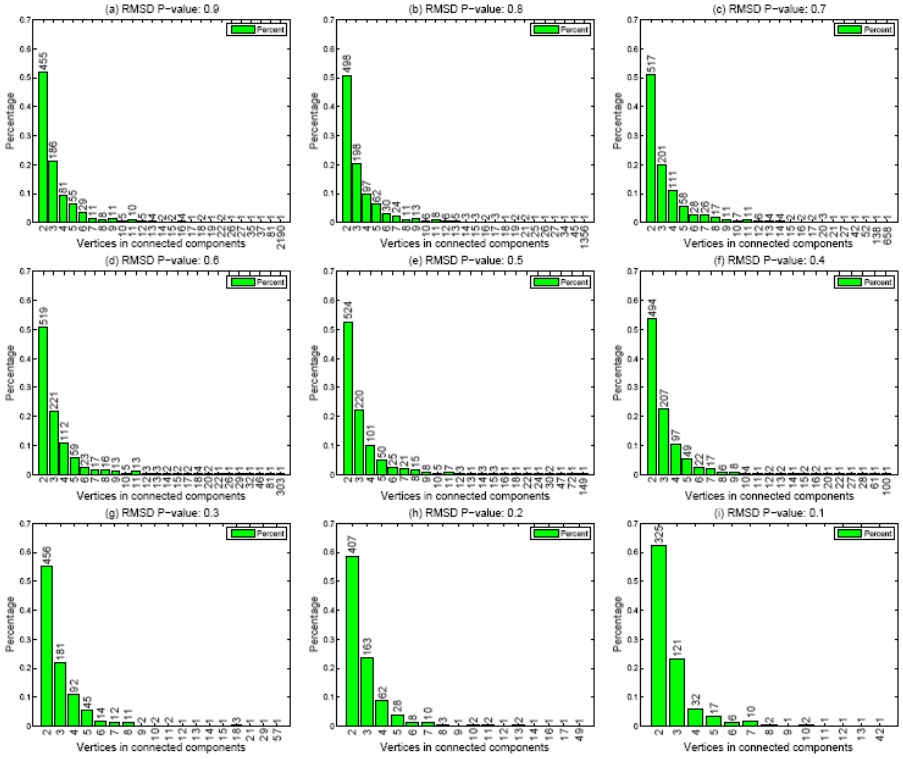
```
1cdk_104 LGTGSFGRVAKLKVLQHTELVMMEYV---EDKENLTDF
2src_51  LGQGCFGEVA-IKLMFAMVLVITEYMGSLDDRANLADF
```

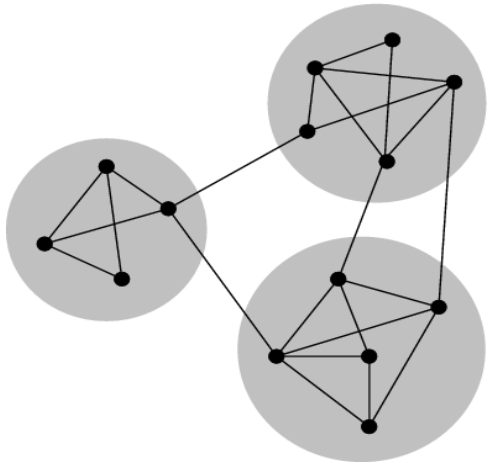
Pocket Similarity Network



Liu *et al*. PPL, 2008

Community Structure Property



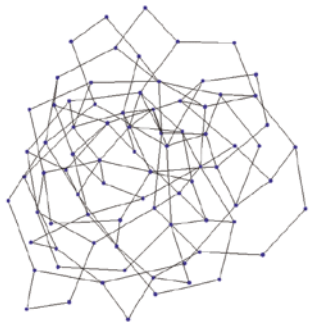


Threshold	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Node	5387	4907	4421	3957	3522	3048	2579	2018	1455
Edge	4943	4259	3681	3158	2704	2274	1854	1408	1002
Component	880	980 (480)	1016 (486)	1023 (464)	995 (435)	920 (474)	827 (469)	693 (561)	520 (563)

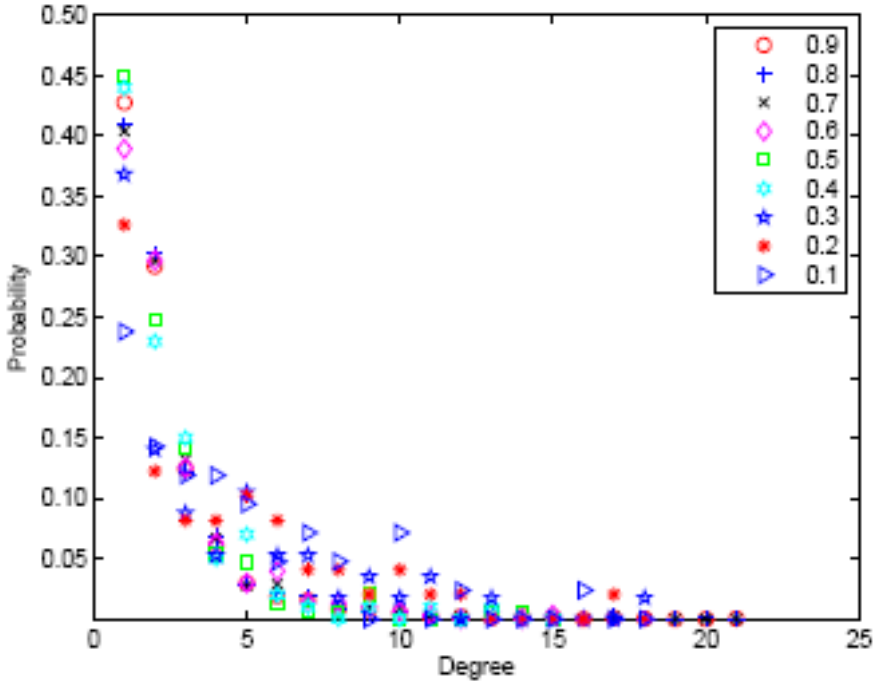
Small World Property

Threshold	Node	Edge	L	L_{random}	C	C_{random}
0.9	2190	2548	15.835	9.107	0.018	0.001
0.8	1356	1611	16.920	8.398	0.02	0.002
0.7	658	782	16.929	7.495	0.019	0.004
0.6	303	382	11.916	6.178	0.019	0.008
0.5	149	176	7.897	5.821	0.021	0.016
0.4	100	121	6.890	5.211	0.020	0.024
0.3	57	111	3.482	2.974	0.187	0.068
0.2	49	102	3.196	2.729	0.211	0.085
0.1	42	92	3.384	2.530	0.248	0.104

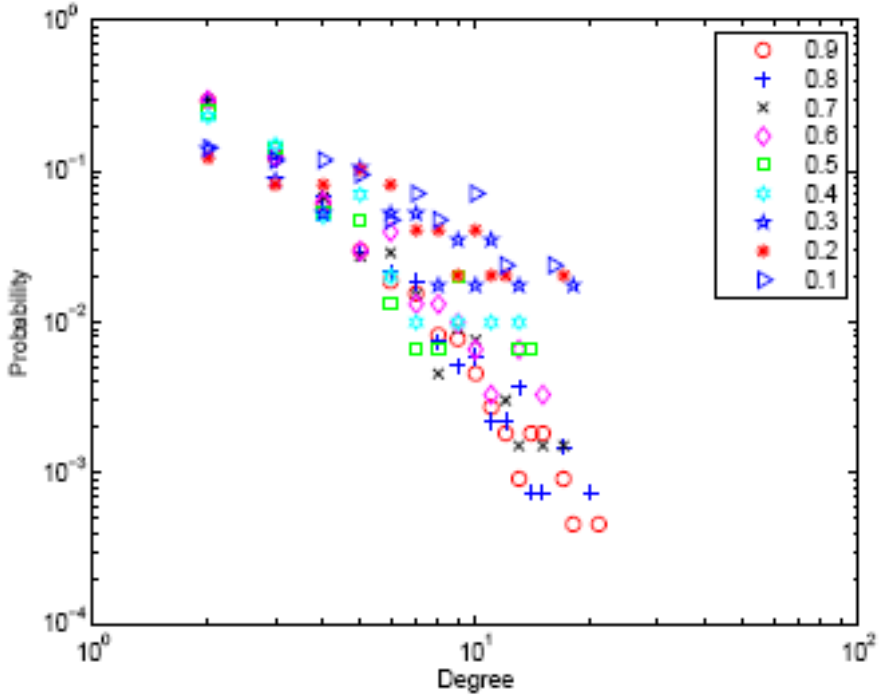
(1) L_{small_world} slightly exceeds L_{random}; (2) C_{small_world} far exceeds C_{random}



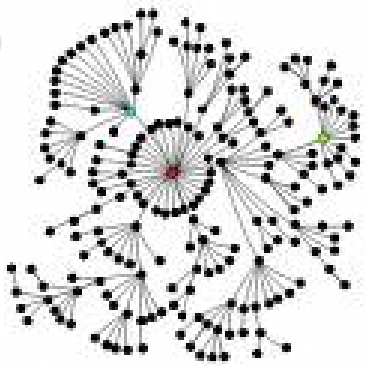
Scale Free Property

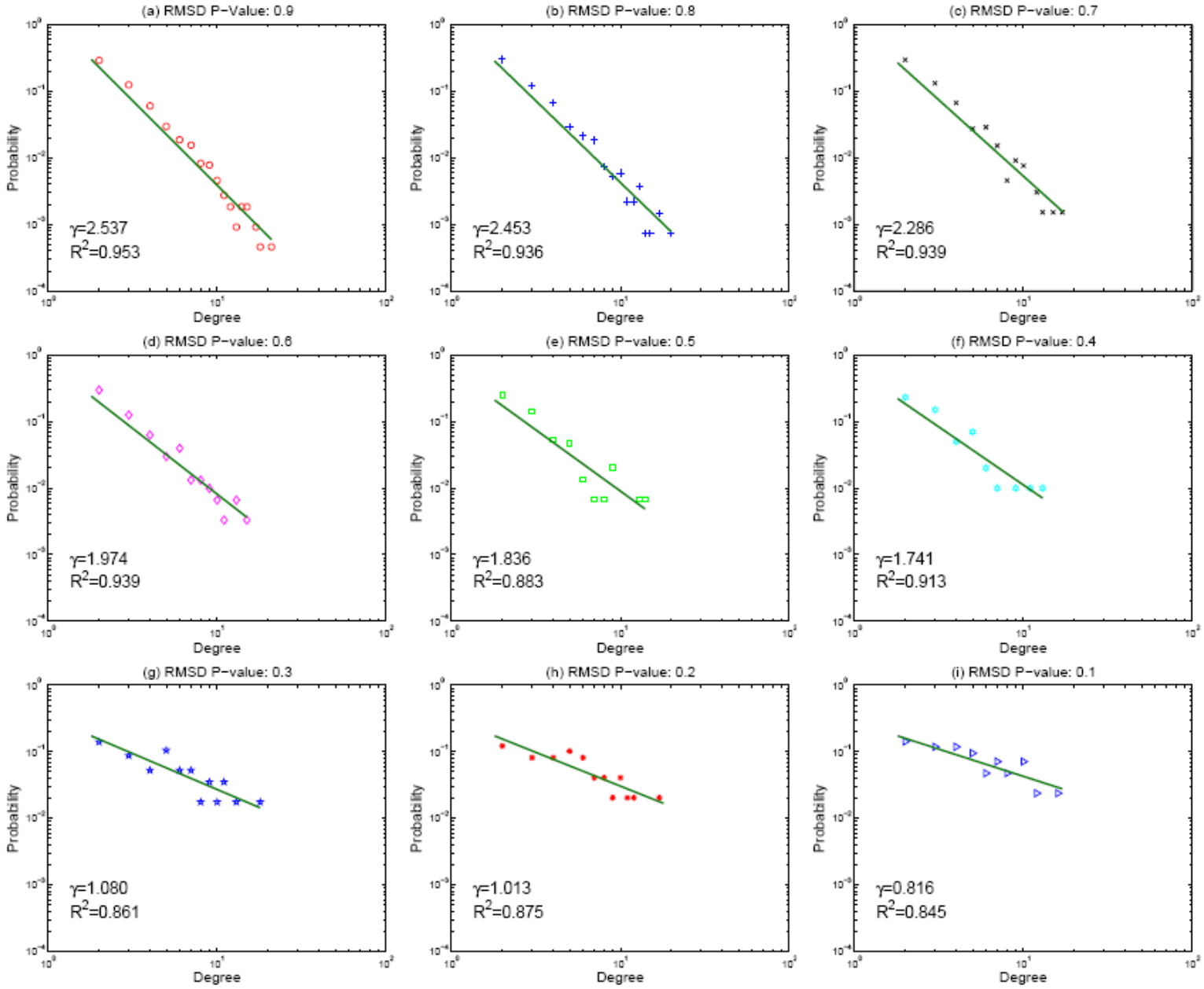


(a)



(b)




(a) RMSD P-Value: 0.9
γ=2.537
R2=0.953
(b) RMSD P-value: 0.8
γ=2.453
R2=0.936
(c) RMSD P-value: 0.7
γ=2.286
R2=0.939
(d) RMSD P-value: 0.6
γ=1.974
R2=0.939
(e) RMSD P-value: 0.5
γ=1.836
R2=0.883
(f) RMSD P-value: 0.4
γ=1.741
R2=0.913
(g) RMSD P-value: 0.3
γ=1.080
R2=0.861
(h) RMSD P-value: 0.2
γ=1.013
R2=0.875
(i) RMSD P-value: 0.1
γ=0.816
R2=0.845
Probability
Degree

Hub Pockets

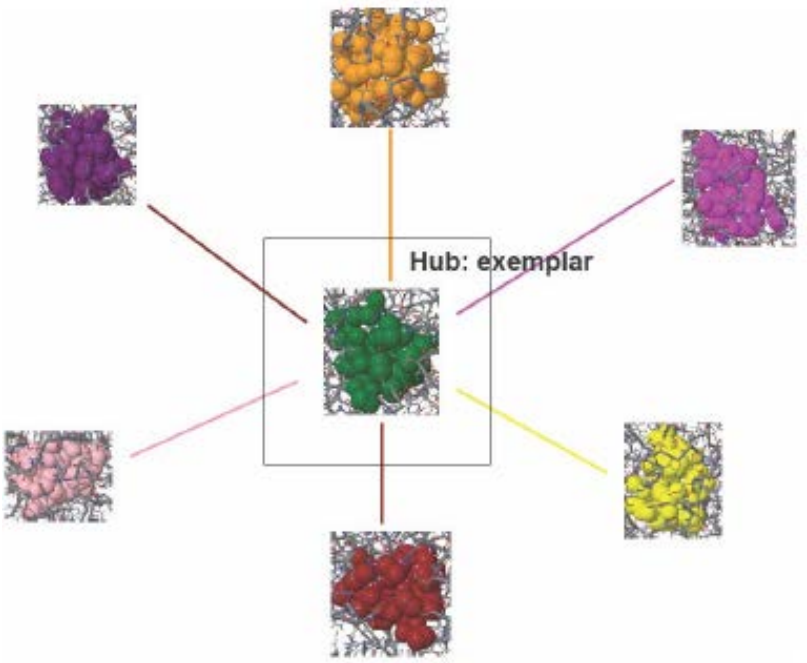
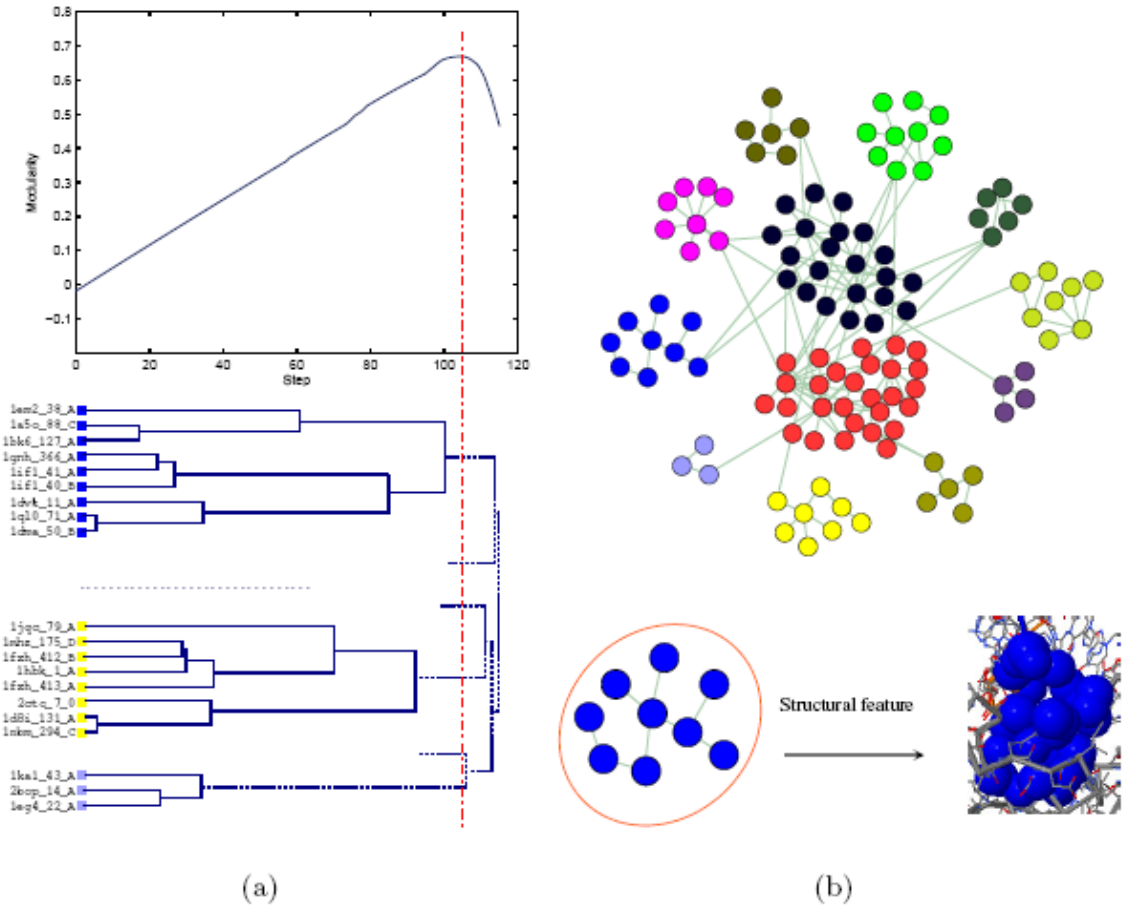


Table 3. The top 10 most highly connected hubs in the pocket similarity network.

Pocket	Protein	Residue Length	Volume	Degree
1dtd_43_A	metallocarboxypeptidase inhibitor	7	42.74	21
1a97_76_B	xanthine-guanine phospho- ribosyl transferase	39	1443.01	18
1exn_58_B	5'-exonuclease (semet labelled protein)	38	1547.7	17
2jdx_52_A	glycine amidinotransferase, deletionmutant atdeltam302	33	708.99	17
1byi_19_0	dethiobiotin synthase	15	355.77	15
2ebo_28_A	ebola virus envelope glycoprotein	14	210.61	15
1im0_37_A	outer membrane phsopholipase	65	2067.34	15
1k8u_12_A	calcium-free (or apo) human s100a6	19	787.1	15
1a4i_59_B	methylenetetrahydrofolate dehydrogense	14	308.84	14
1aym_123_A	human rhinovirus 16 coat protein	25	570.08	14

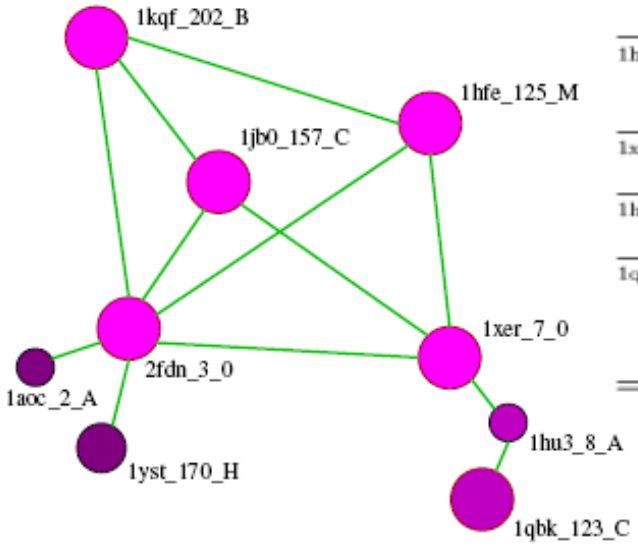
Clustering

$$Q = \sum_i (e_{ii} - a_i^2)$$



Liu *et al*. IJBRA, 2008

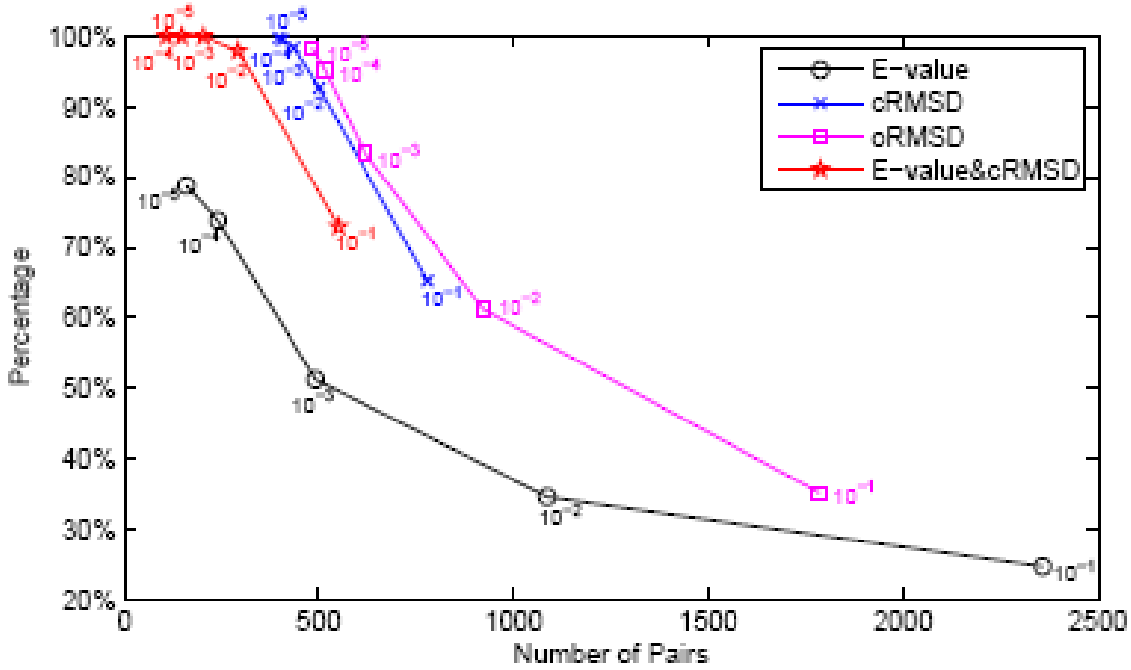
Cluster Example



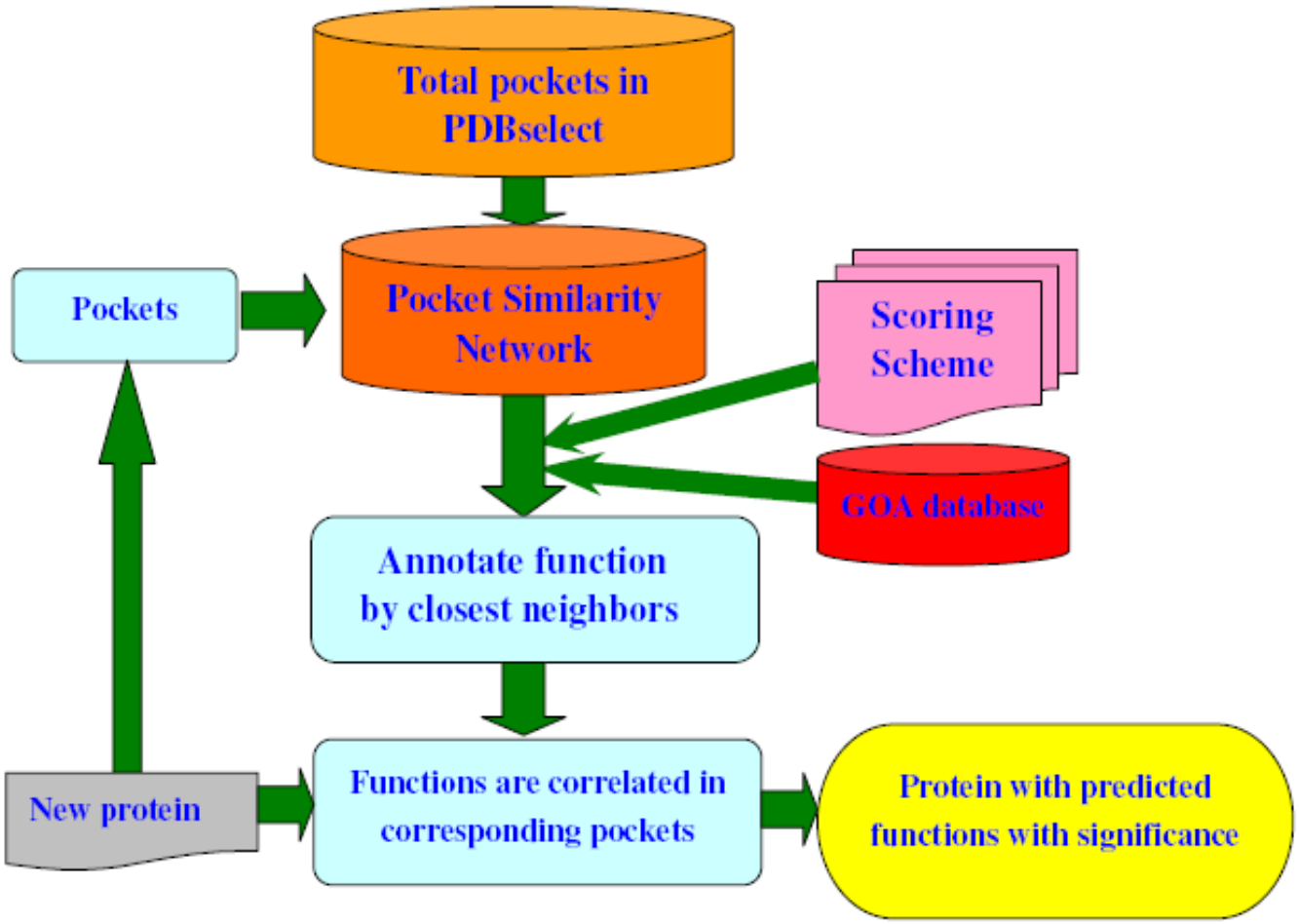
Pocket	Len	Vol	Protein	Deg	Sequence
1kqf_202_B	14	184.96	Oxidoreductase, formate dehydrogenase n from E. coli	3	HIEGGLAAEWRAKT
1jb0_157_C	14	177.16	Photosynthesis, crystal structure of photosystem i: a photosynthetic reaction center and core antenna system from cyanobacteria	3	VCPTVLCVGCKRCV
2fdn_3_0	14	173.50	Electron transport, ferredoxin from clostridium acidi-urici	6	Y*C*PVAII*C*ID*C*GA*C*
1yst_170_H	9	83.90	Photosynthetic reaction center	1	FTRASDCGA
1aoc_2_A	4	16.21	Coagulation factor, Japanese horseshoe crab coagulogen	1	CVDC
1hfe_125_M	14	191.58	Hydrogenase, a resolution structure of the Fe- only hydrogenase from desulfovibrio desulfuricans	3	V*C*PTAII*C*IN*C*GQ
1xer_7_0	14	197.04	Electron transport, structure of ferredoxin	3	V*C*PVVF*C*IF*C*MA*C*V
1hu3_8_A	5	9.50	Translation, middle domain of human eif4gii	2	QFLAN
1qbk_123_C	13	192.96	Nuclear transport protein complex structure of the karyopherin beta2-ran gppnhp nuclear transport complex	1	LV*GTGK*FIIFC*N*I

Functional Association

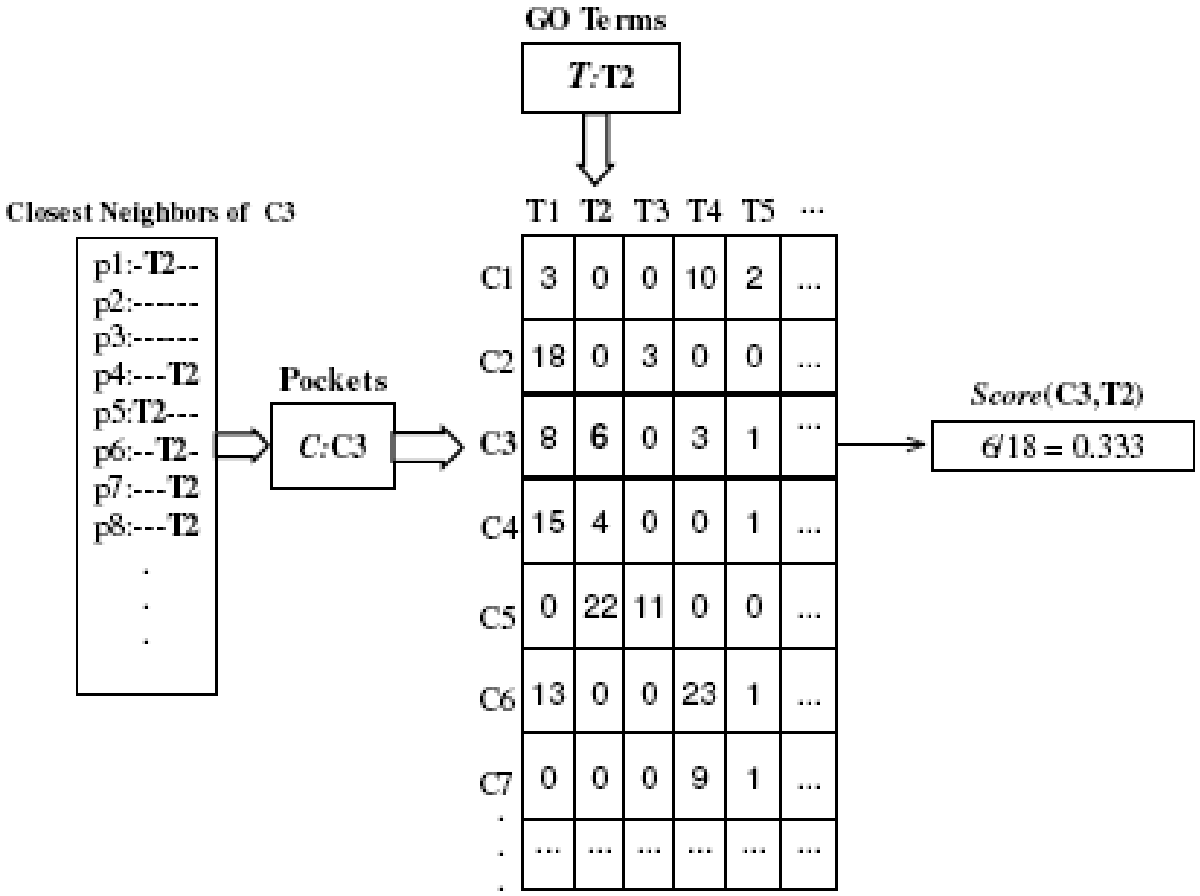
Threshold	1.0×10^{-1}	1.0×10^{-2}	1.0×10^{-3}	1.0×10^{-4}	1.0×10^{-5}
Pocket pairs	1002	602	521	481	468
GO annotated pairs	778	501	437	405	397
Similar pairs	508	464	430	403	396
Percentage	65.29%	92.61%	98.40%	99.50%	99.75%



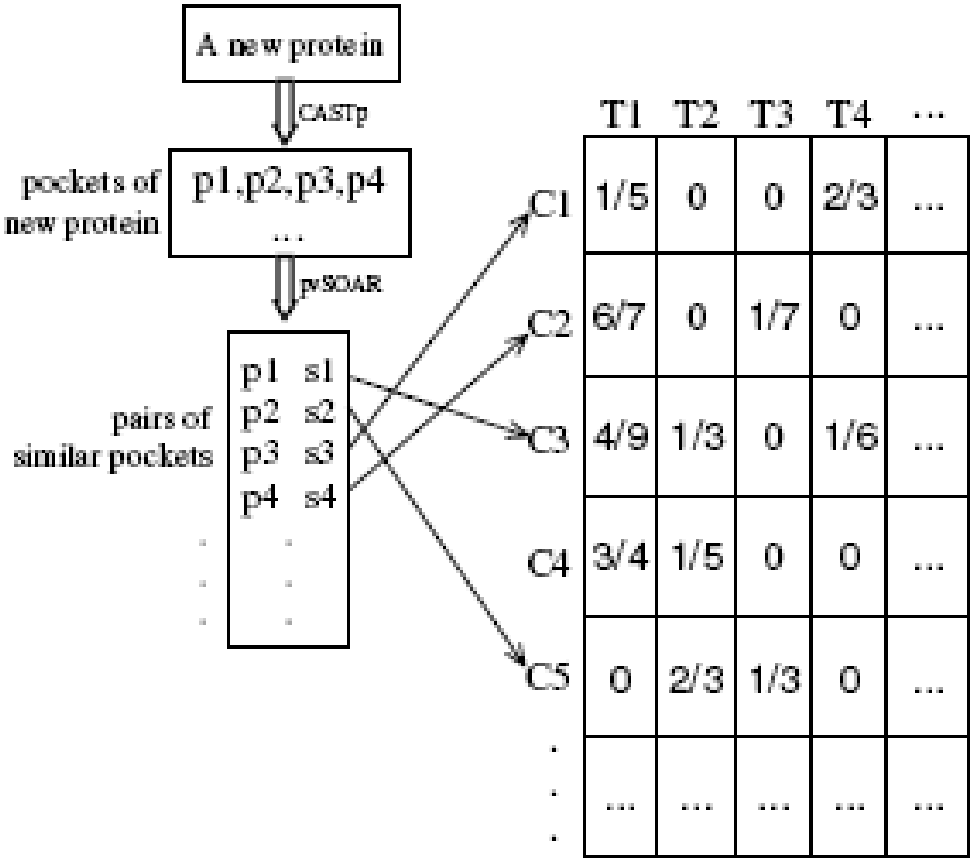
Prediction Workflow



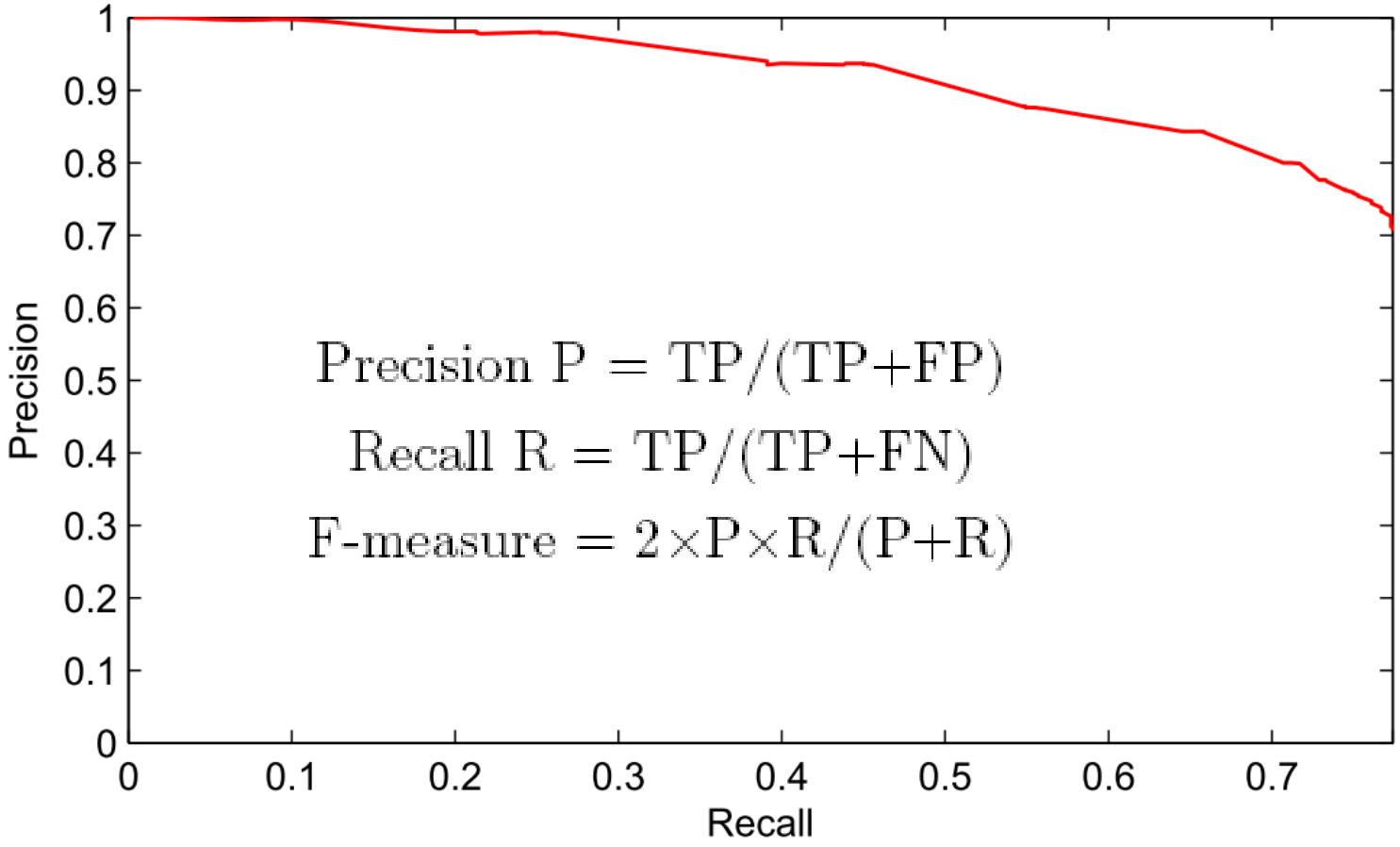
Scoring Scheme



Scoring Scheme



$$Score_{T1} = 1\times\frac{1}{5}+1\times\frac{6}{7}+1\times\frac{4}{9}+1\times 0$$

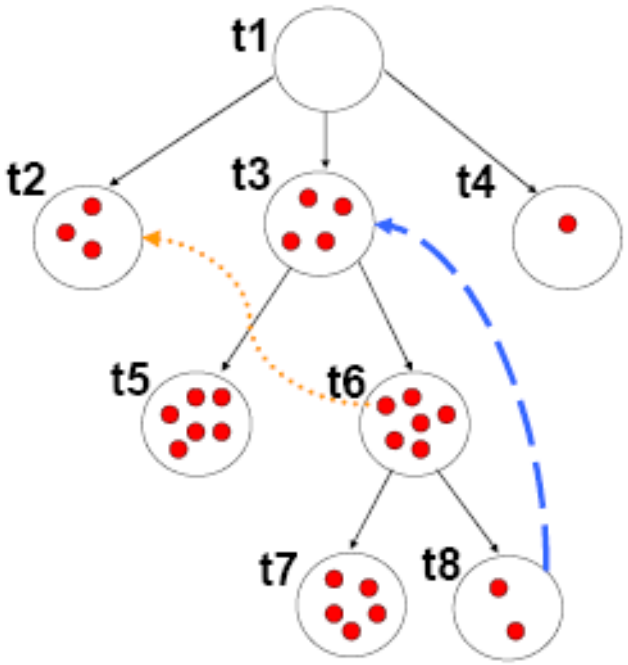

Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Precision P = TP/(TP+FP)
Recall R = TP/(TP+FN)
F-measure = 2×P×R/(P+R)

GO Specificity

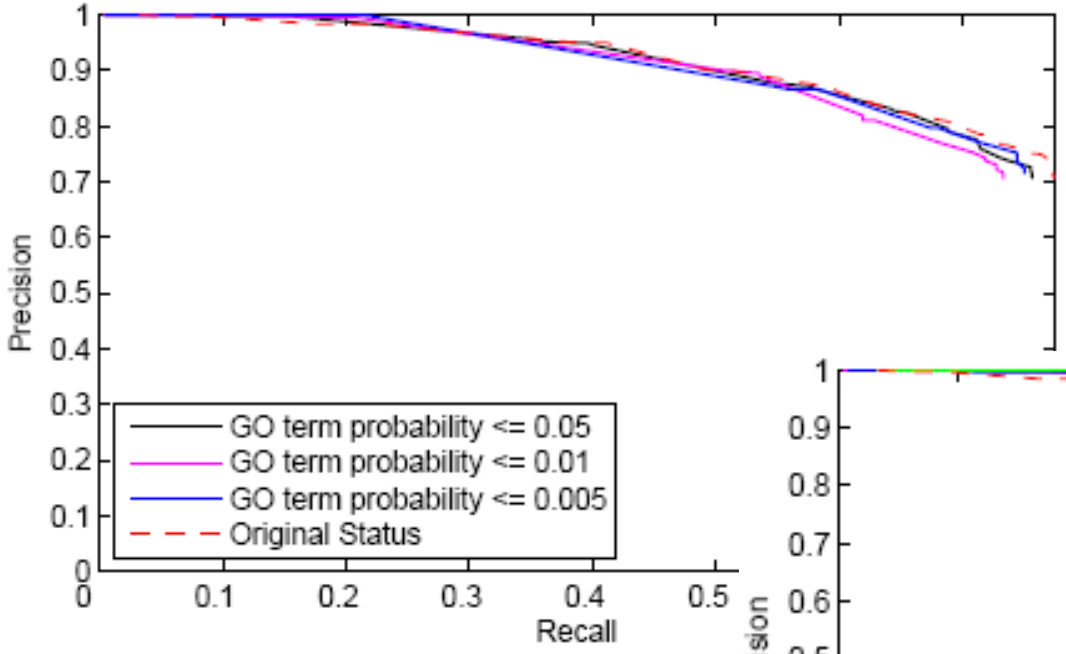
- GO term probability

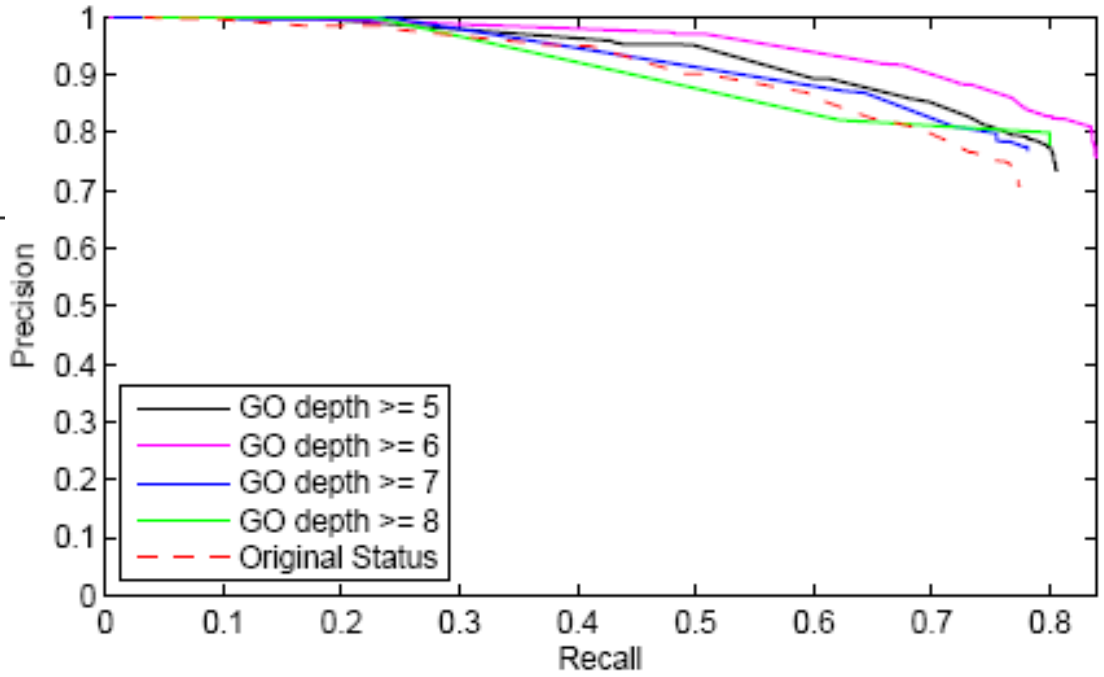
$$freq(c) = anno(c) + \sum_{h \in children(c)} freq(h)$$

- GO depth



Influence of GO Specificity



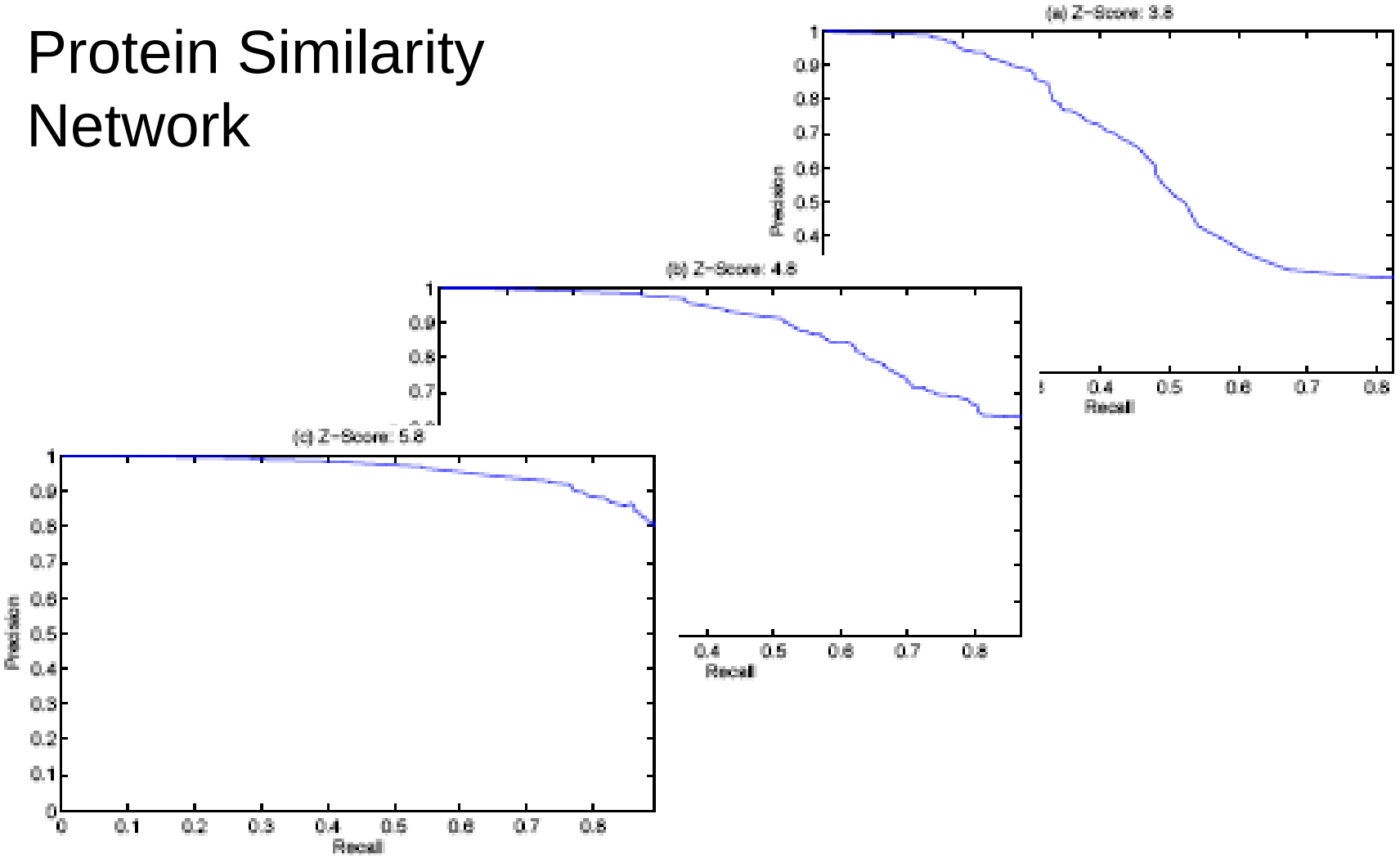


Influence of Sequence Similarity

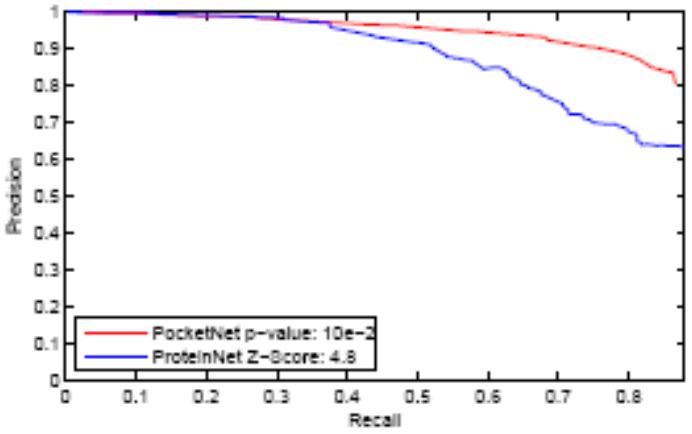
- Remove the redundancy
  - Sequence similarity
  - Multiple experiments
- PDBselect database is a subset of PDB that does not contain highly homologous sequences
- PDFselect 25
  - No two proteins have more than 25% sequence similarity

Influence of Global Structure

Protein Similarity Network

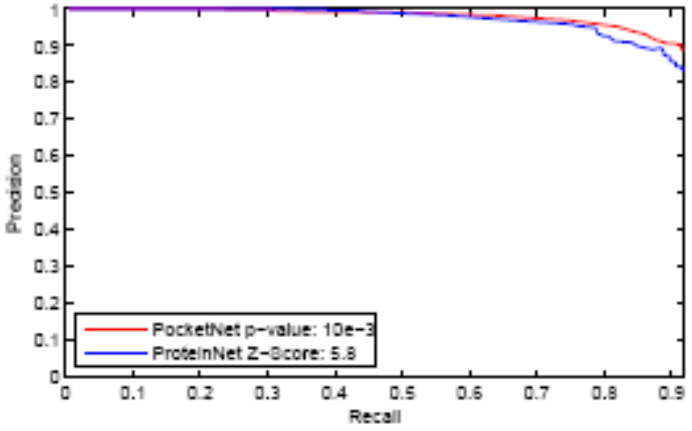


Local vs Global Structure



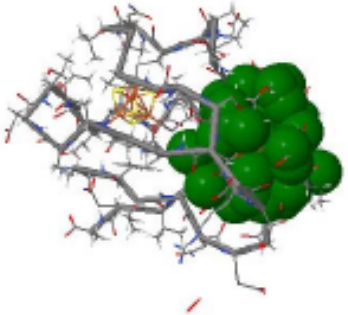
(a)

(b)

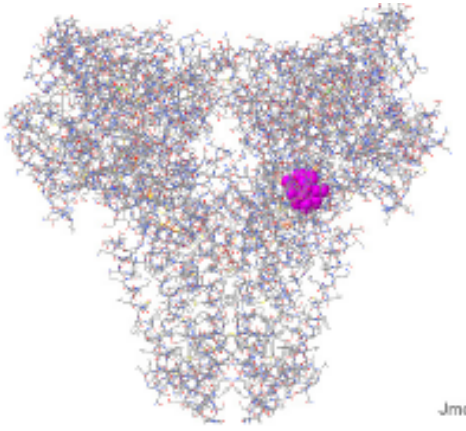


(c)

(d)

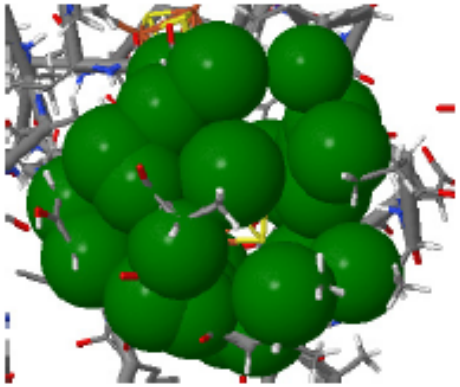


Protein: 2fdn

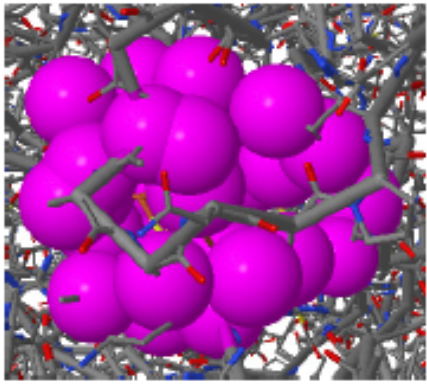


Protein: 1qla

(a)



Pocket: 2fdn_4_0



Pocket: 1qla_303_E

(b)

>2fdn: GO:0005506; GO:0009055; GO:0006118

AYVINEACISCGACEPECPVNAISSGDDRYVIDADTCIDCGACAGVCPVDAPVQA

>1qla_E: GO:0005506; GO:0009055; GO:0016491;GO:0006118; GO:0006099; GO:0016020

MGRMLTIRVFKYDPQSAVSKPHFQEYKIEEAPSMTIFIVLNMIRETYDPDLNFDFVCRAG
ICGSCGMMINGRPSLACRTLTKDFEDGVITLLPLPAFKLIKDLSVDTGNWFNGMSQRVES
WIHAQKEHDISKLEERIEPEVAQEVFELDRCIECGCCIAACGTKIMREDFVGAAGLNRVV
RFMIDPHDERTDEDYYELIGDDDGVFGCMTLLACHDVCPKNLPLQSKIAYLRRKMVSVN

EMBOSS: Identity = 19/239(7.9%), Similarity = 28/239(11.7%)
CE: RMSD = 4.9Å, Z-Score = 1.2

(c)

2fdn_4_0: ICISCGACYVCPVAP

1qla_303_E: FCIECGCCANCPKLL

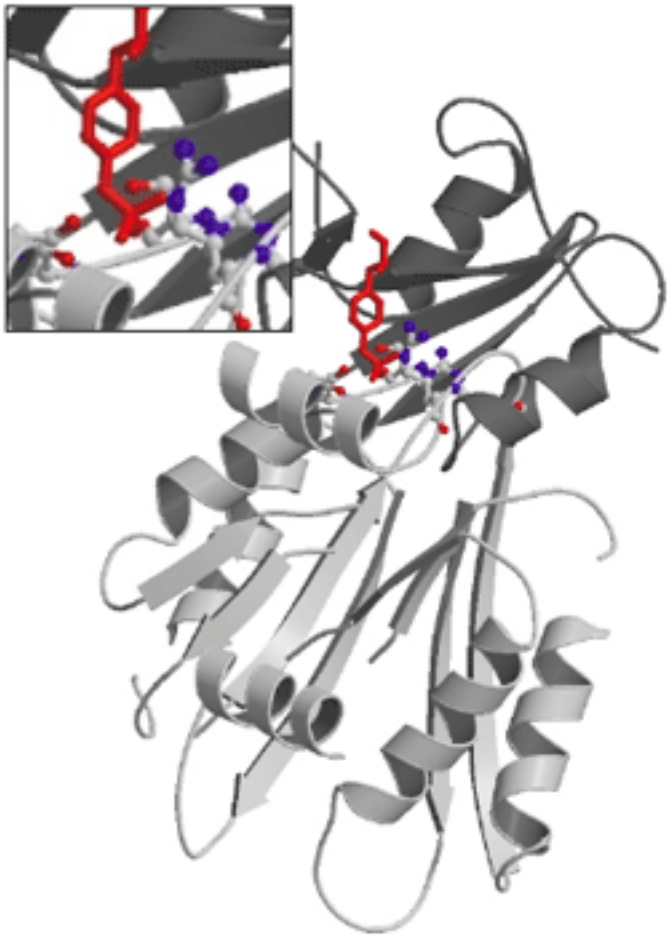
pvSOAR: cRMSD = 0.414Å, p-value = 4.054e-08

(d)

Table 14: The detailed description to the GO terms which are annotated to protein 2fdn and 1qla_E. The italic GO terms are the common GO annotations between the two proteins.

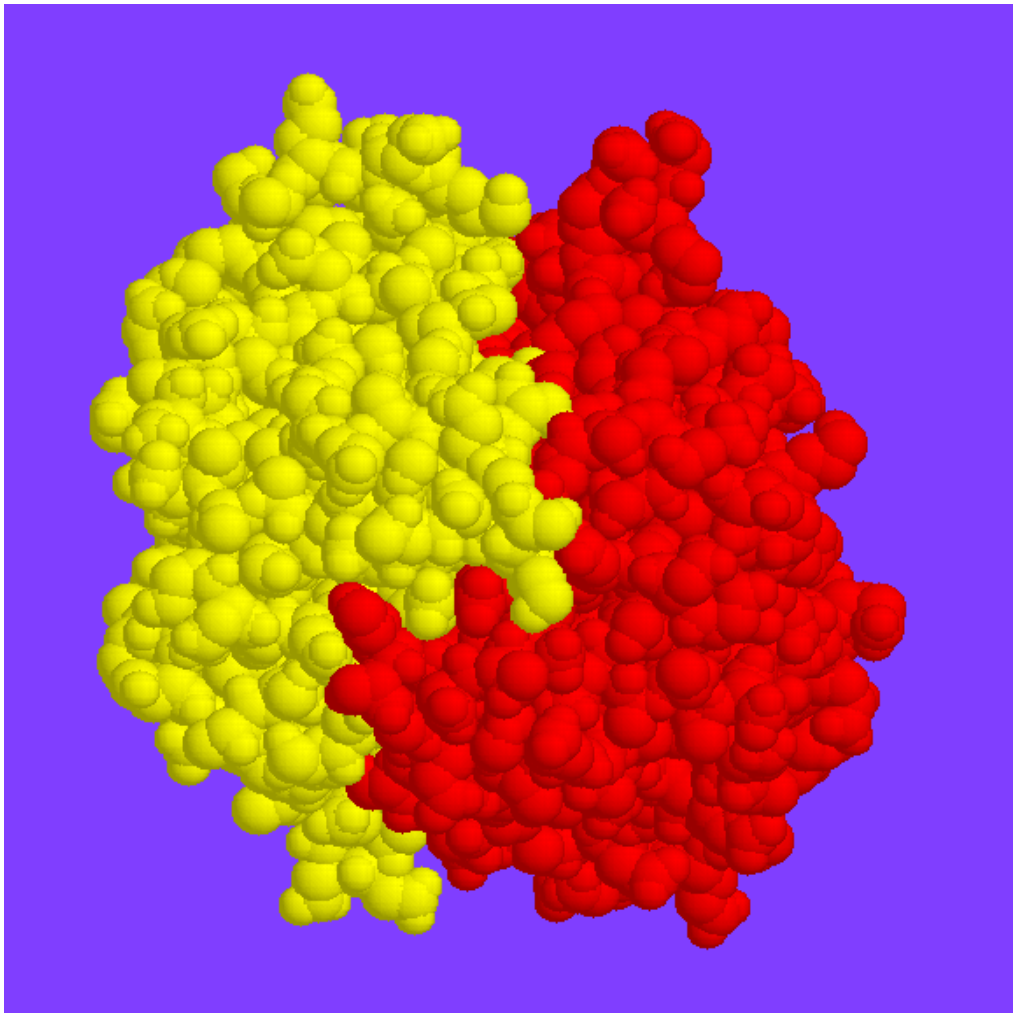
GO term	Ontology	Probability	Depth	description
GO:0005506	F	0.064	6	iron ion binding
GO:0009055	F	0.071	4	electron carrier activity
GO:0006118	P	0.139	4	electron transport
GO:0016491	F	0.179	3	oxidoreductase activity
GO:0006099	P	0.007	7	tricarboxylic acid cycle
GO:0016020	C	0.363	4	membrane

Functional Structure Motif



蛋白相互作用

Physical interaction

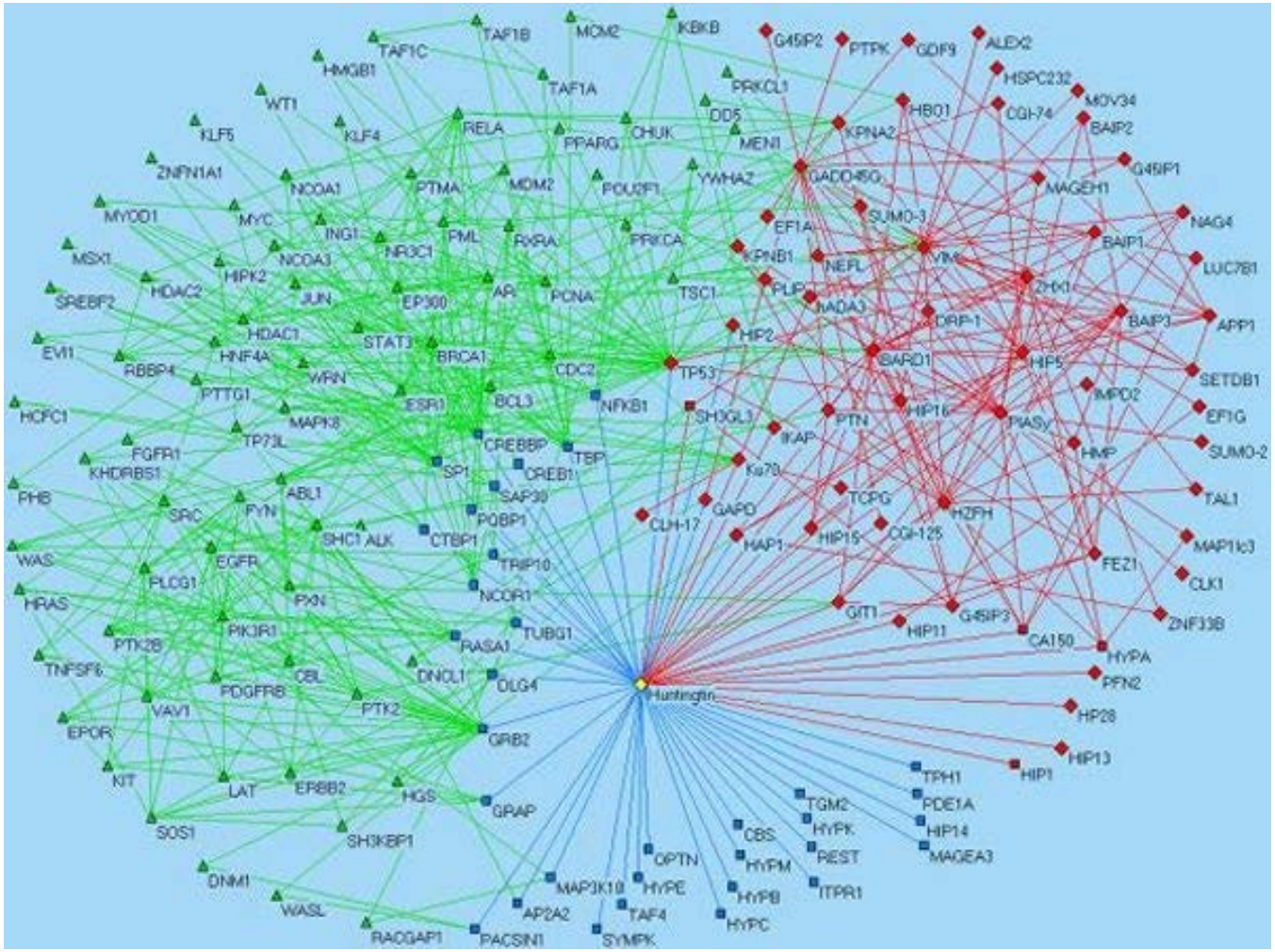


蛋白相互作用网络

无向图

节点：蛋白质

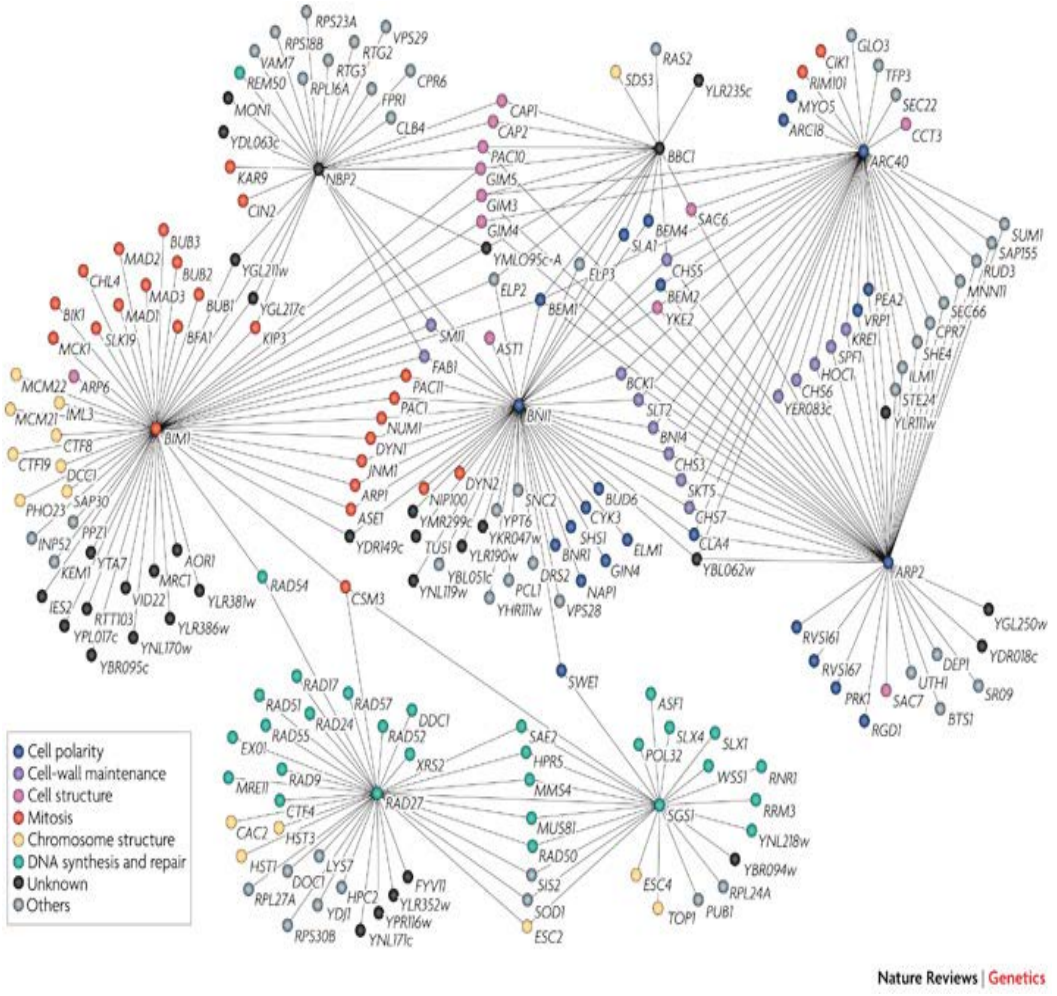
边：相互作用



遗传相互作用

Genetic Interaction

two mutations
have a combined
effect not exhibited
by either mutation
alone



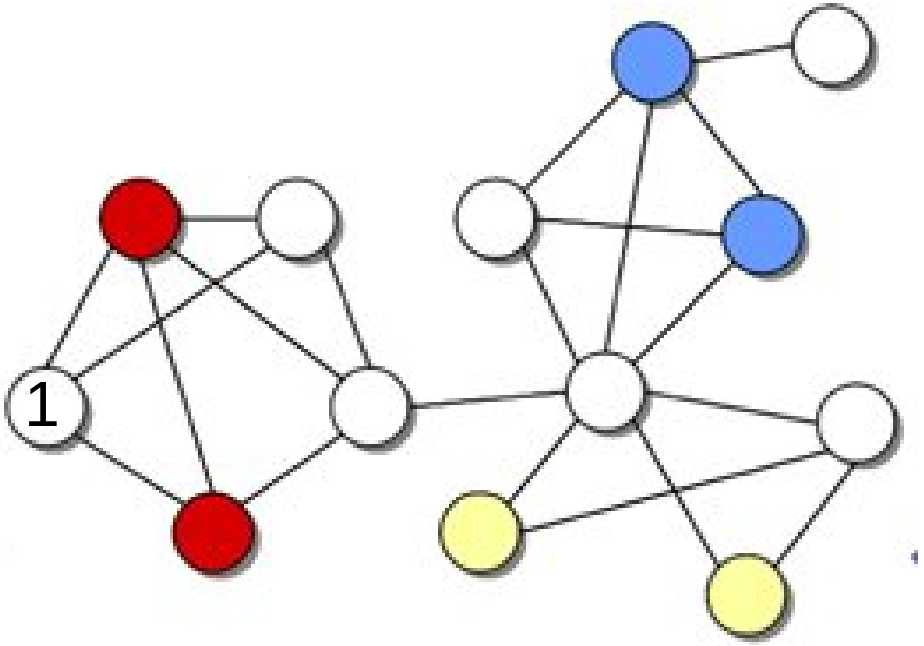
PPI & Protein annotation

Biological process

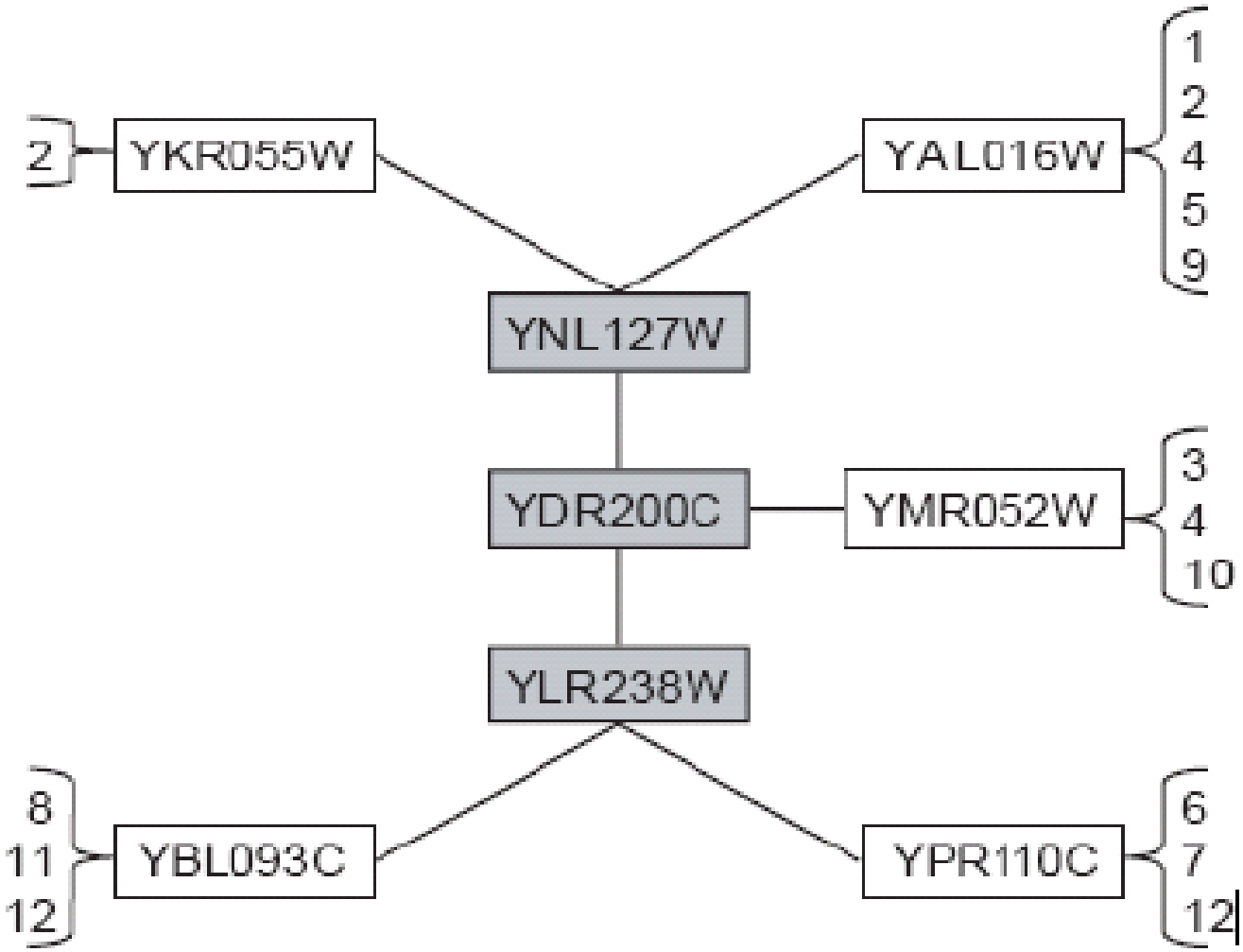
Molecular function

Cellular component

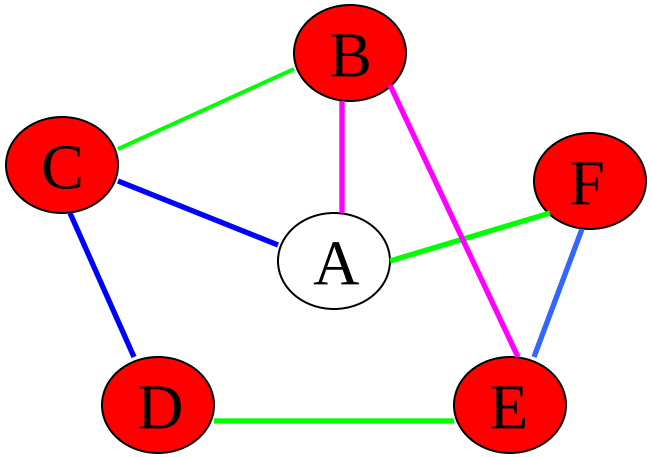
Un-annotated protein



例子


2
YKR055W
YAL016W
1
2
4
5
9
YNL127W
YDR200C
YMR052W
3
4
10
YLR238W
8
11
12
YBL093C
YPR110C
6
7
12

概率估计



B Protein binary interaction

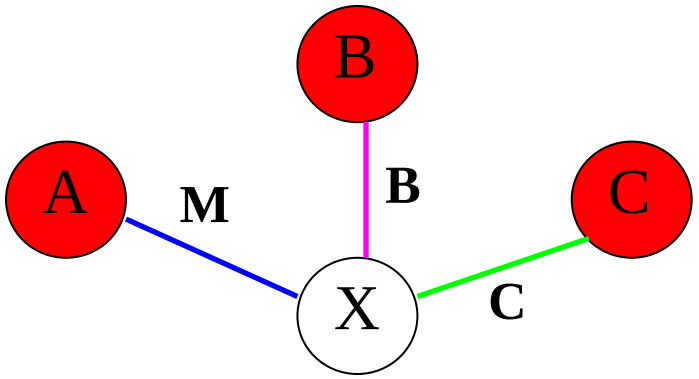
C Protein complex interaction

M Microarray expression correlated

P(S|M) = ? P(S|B) = ? P(S|C) = ?

S: the function similarity between protein X and Y

功能赋值



Protein binary interaction

Protein complex interaction

Microarray coexpressed

A, B and C are the all and only function-known proteins that interact with protein X

A, B and C ⟶ {Function i, i = 1…n}

Given the known probabilities P(S|M), P(S|B) and P(S|C)
How to assign a function F_i to for uncharacterized protein *X* ?

Neighbor Counting

- Assign *k* functions to the unannotated protein with *k* largest frequencies in its neighbors
- Disadvantage
  - Do not consider the frequency of functions in all proteins

Chi-square Method

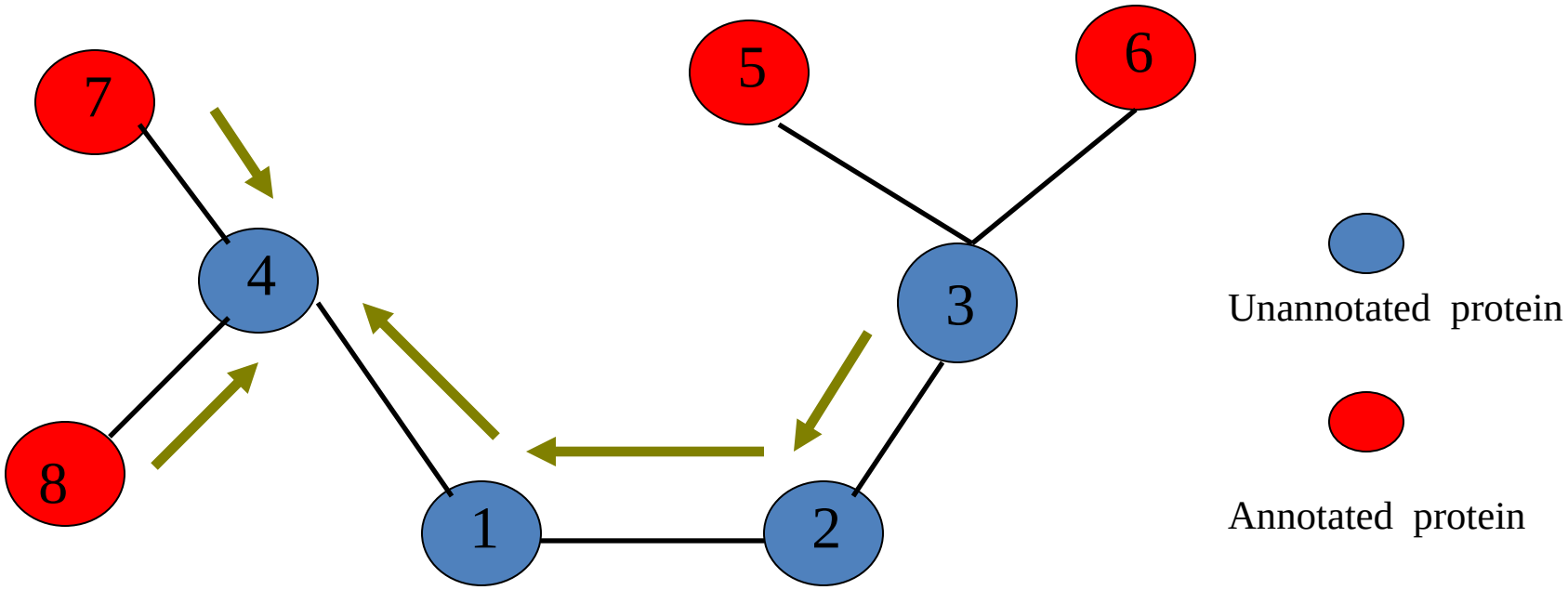
$$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)}$$

For a protein P_i, let $n_i(j)$ be the number of proteins interacting with P_i and having function F_j.

Let $e_i(j) = \#\text{Nei(i)} \times \pi_j$ be the expected number of proteins in Nei(i) having function F_j, where #Nei(i) is the number of proteins in Nei(i).

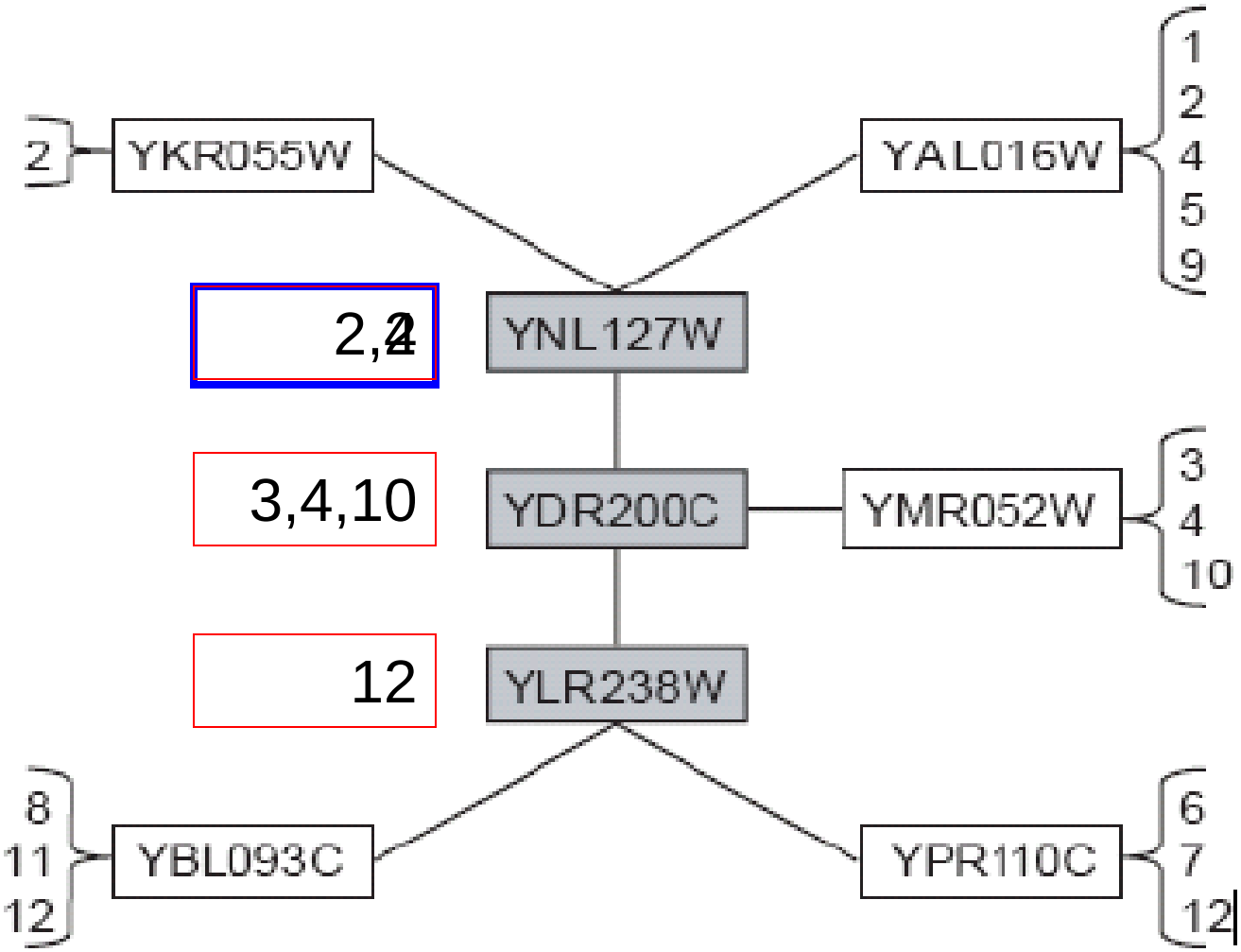
For a fixed k, the authors assigned an unannotated protein with k functions having the top k chi-square statistics.

循环迭代法

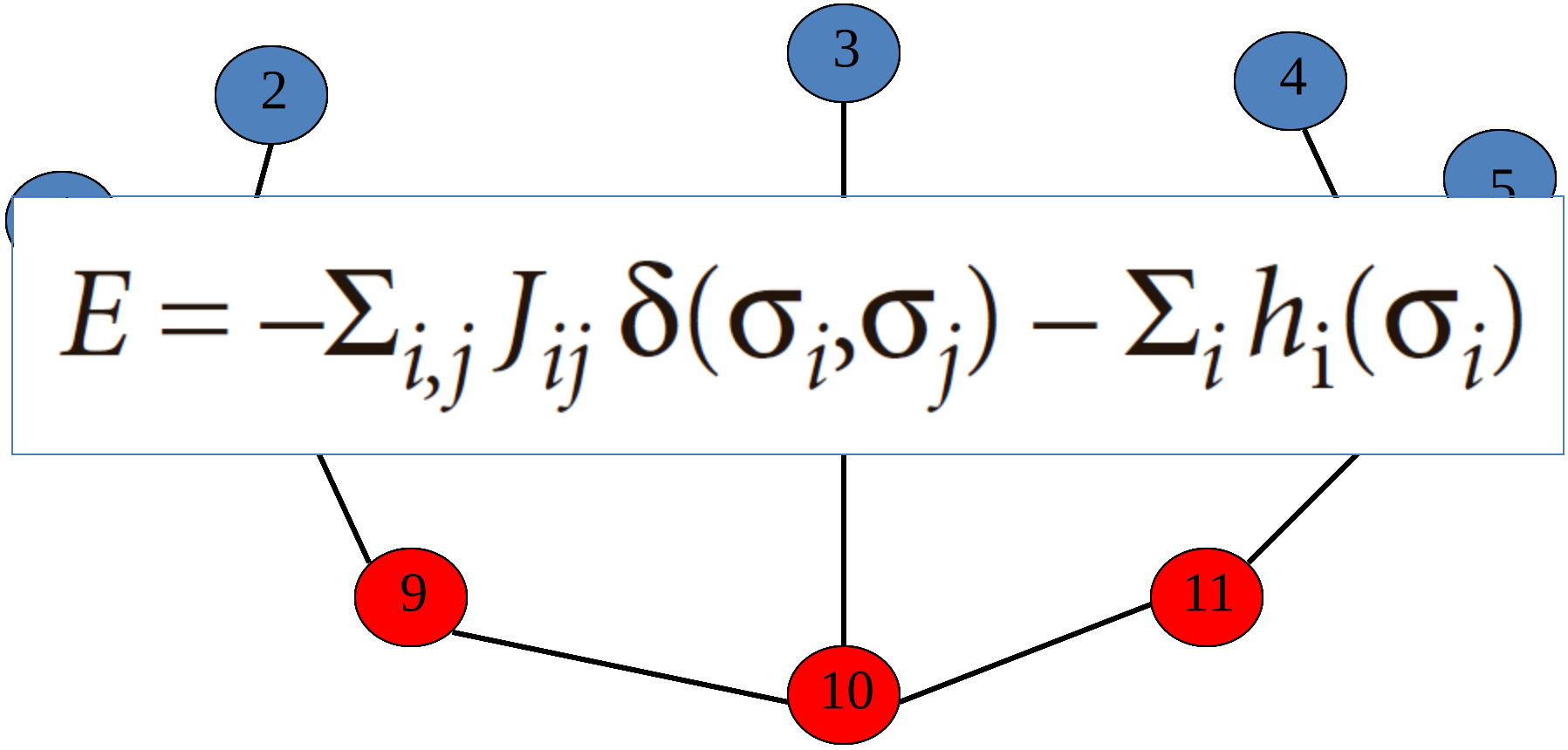


Function assignment in a self-consistent and iterative way

例子


2
YKR055W
YAL016W
1
2
4
5
9
2,4
YNL127W
3,4,10
YDR200C
YMR052W
3
4
10
12
YLR238W
8
11
12
YBL093C
YPR110C
6
7
12

模拟退火法



$$E = -\Sigma_{i,j} J_{ij}\, \delta(\sigma_i,\sigma_j) - \Sigma_i h_i(\sigma_i)$$

Annotated Hypothetical State 1 State 0

Markov Random Field

$$\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i} = \left(\frac{\pi}{1-\pi}\right)^{N_1}(1-\pi)^{N},$$

$$U(x) = -\alpha N_1 - \beta N_{10} - \gamma N_{11} - N_{00}$$

$$N_{11} = \sum_{(i,j)\in S} x_i x_j$$

$$N_{10} = \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i$$

$$N_{00} = \sum_{(i,j)\in S} (1-x_i)(1-x_j)$$

$$\Pr(X \mid \theta) = \frac{1}{Z(\theta)} \exp(-U(x))$$

Deng, *et al*. JCB, 2003.

二次规划模型

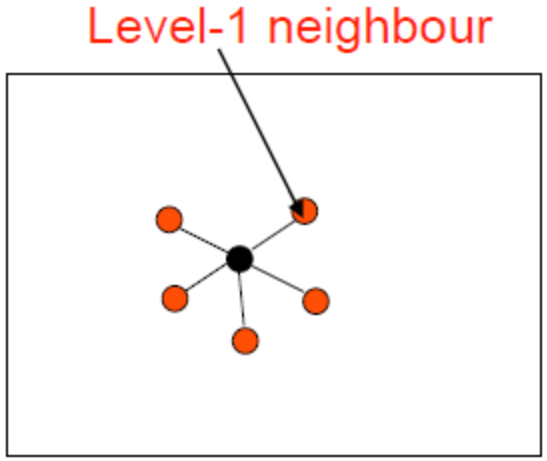
Given a graph $G = (V, E)$ with weighted edges and a set of nodes $S = \{s_{n+1}, s_{n+2}, \cdots, s_{n+m}\} \subset V$ and its corresponding binary vectors representing function annotation information: $f_1, f_2, \cdots, f_m$, where $f_i \in R^d$ and if $s_i \in F_j$, $f_i^j = 1$, else $f_i^j = 0$. To find the corresponding binary vector of nodes in $V \backslash S$, subject to the stronger related-link weight mean the more similar function information. Let

$$\max \quad \frac{1}{2}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv} x_u^T x_v + \frac{1}{2}\sum_{u=1}^{n}\sum_{w=1}^{m} b_{uw} x_u^T f_w - \lambda \sum_{u=1}^{n} x_u^T \cdot 1$$

$$s.t. \qquad x_u^j \in \{0, 1\}, \quad j = 1, 2, \cdots d.$$

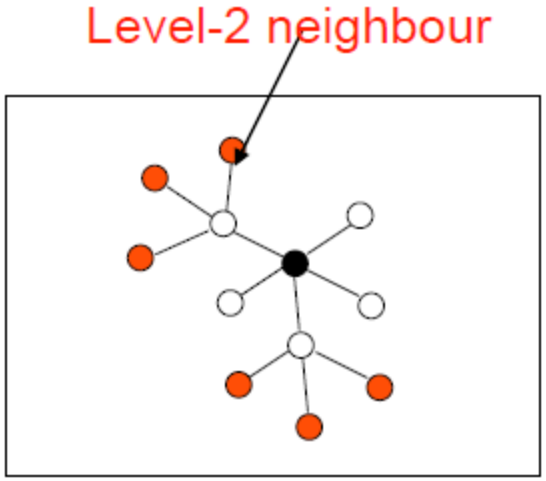
Direct Functional Association

- Direct functional association
  - Interacting partners are likely to share functions
  - Proteins from the same pathways are likely to interact

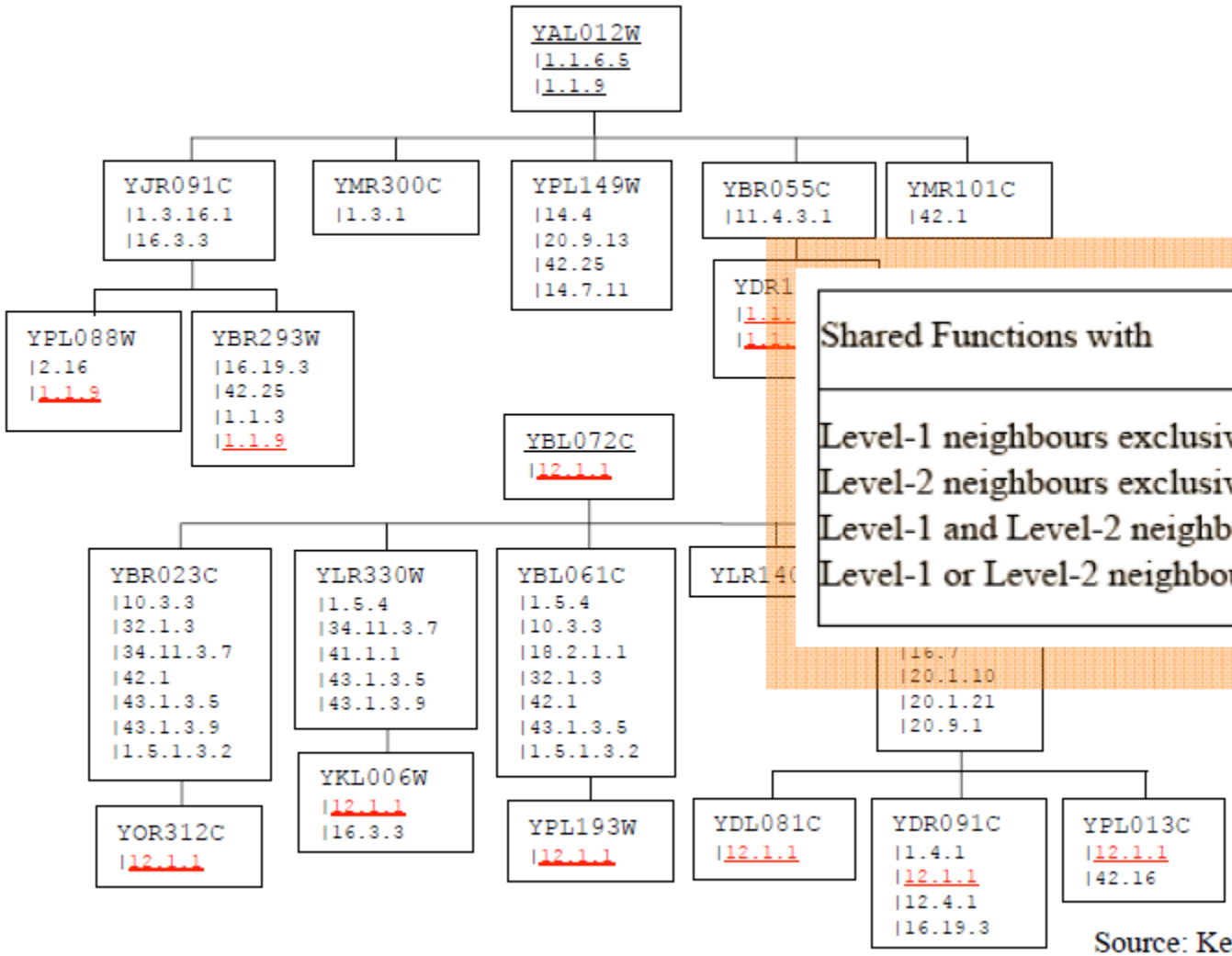


Indirect Functional Association

- Indirect functional association
  - Proteins that share interaction partners may likely to share functions
  - Proteins that have common properties are likely to bind the same proteins



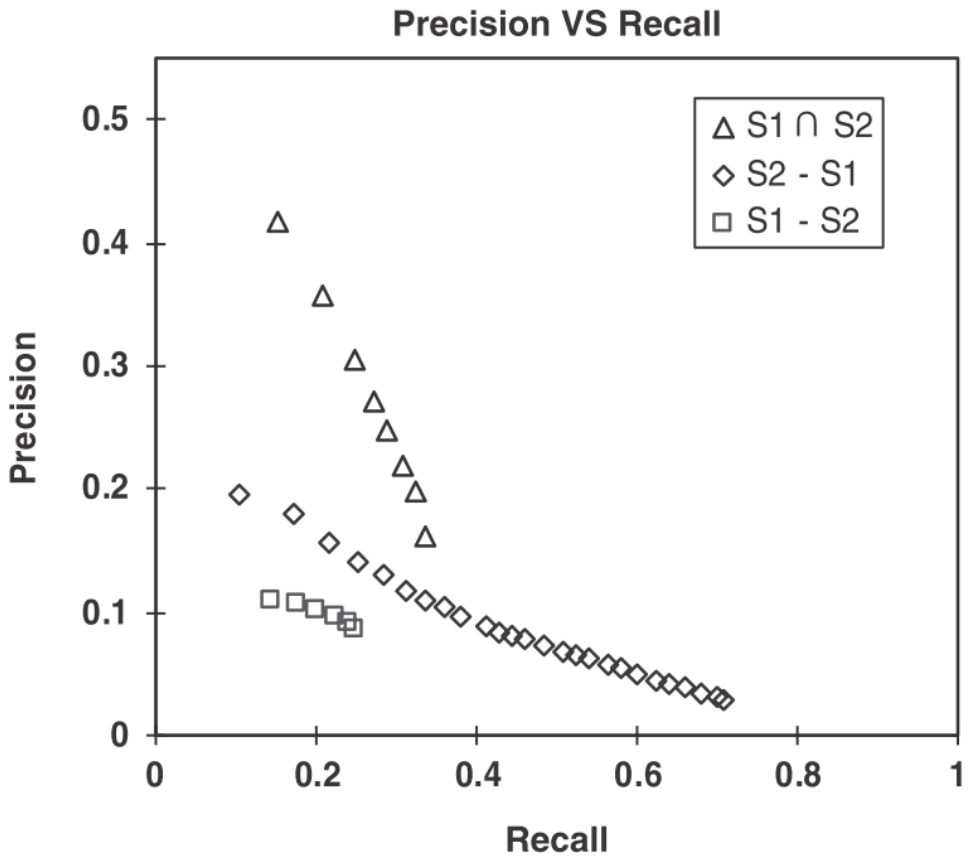
Functional Association



Shared Functions with	Fraction
Level-1 neighbours exclusively	0.016338
Level-2 neighbours exclusively	0.226574
Level-1 and Level-2 neighbours	0.463960
Level-1 or Level-2 neighbours	0.706872

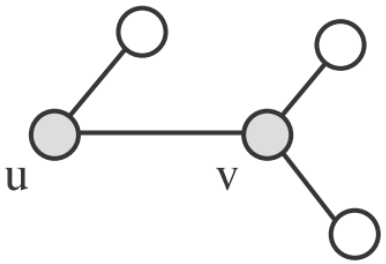
Source: Kenny Chua

Test by Neighbor Counting



Czekanowski-Dice Distance

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$



$|N_u \Delta N_v| = 3$
$|N_u \cap N_v| = 2$
$|N_u \cup N_v| = 5$

CD-Distance(u, v)
= 3 / (5+2)
= 0.429

Brun, *et al*. Genome Biology, 2003.

Functional Similarity

$$S_{\mathrm{FS}}(u,v) = \frac{2\,|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v| + \lambda_{u,v}} \times \frac{2\,|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v| + \lambda_{v,u}}$$

$$S_{FS}(u,v) = \frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in(N_u-N_v)} r_{u,w} + \sum_{w\in(N_u\cap N_v)} r_{u,w}(1-r_{v,w})\right) + 2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w} + \lambda_{u,v}} \times \frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in(N_v-N_u)} r_{v,w} + \sum_{w\in(N_u\cap N_v)} r_{v,w}(1-r_{u,w})\right) + 2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w} + \lambda_{v,u}}$$

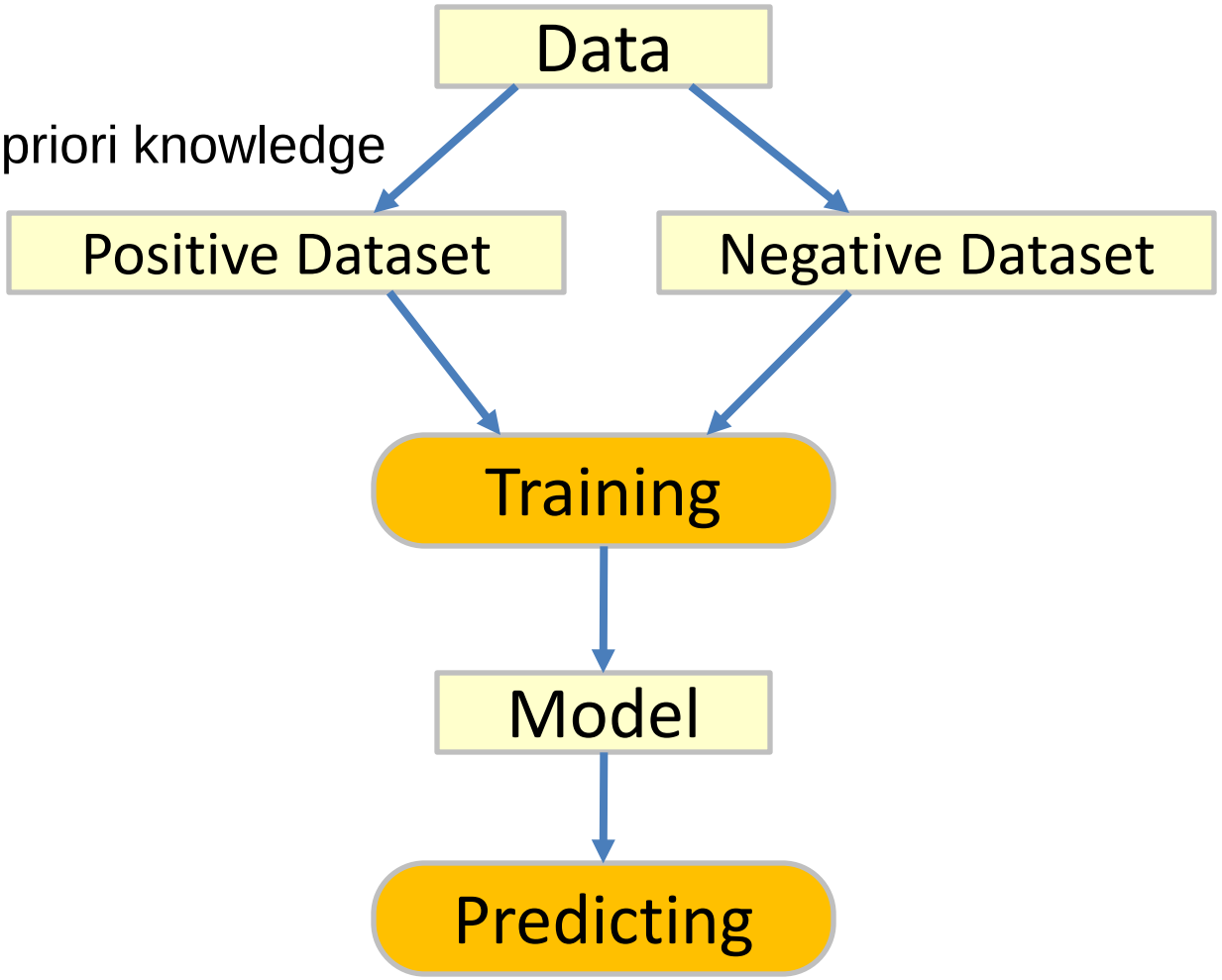
Chua, *et al*. Bioinformatics, 2006

Post-Translation Modification

- Regulation of activity
  - modification may turn activity on
  - modification may turn activity off
  - modification may generate a different function
- Protein-protein interaction
  - modification site may be a binding interface
- Subcellular localization
  - modification site may be a targeting signal
  - modification may be a membrane anchor
- Aging
  - modification may identify the protein for degradation
  - modification may target a protein to be scavenged

Machine Learning

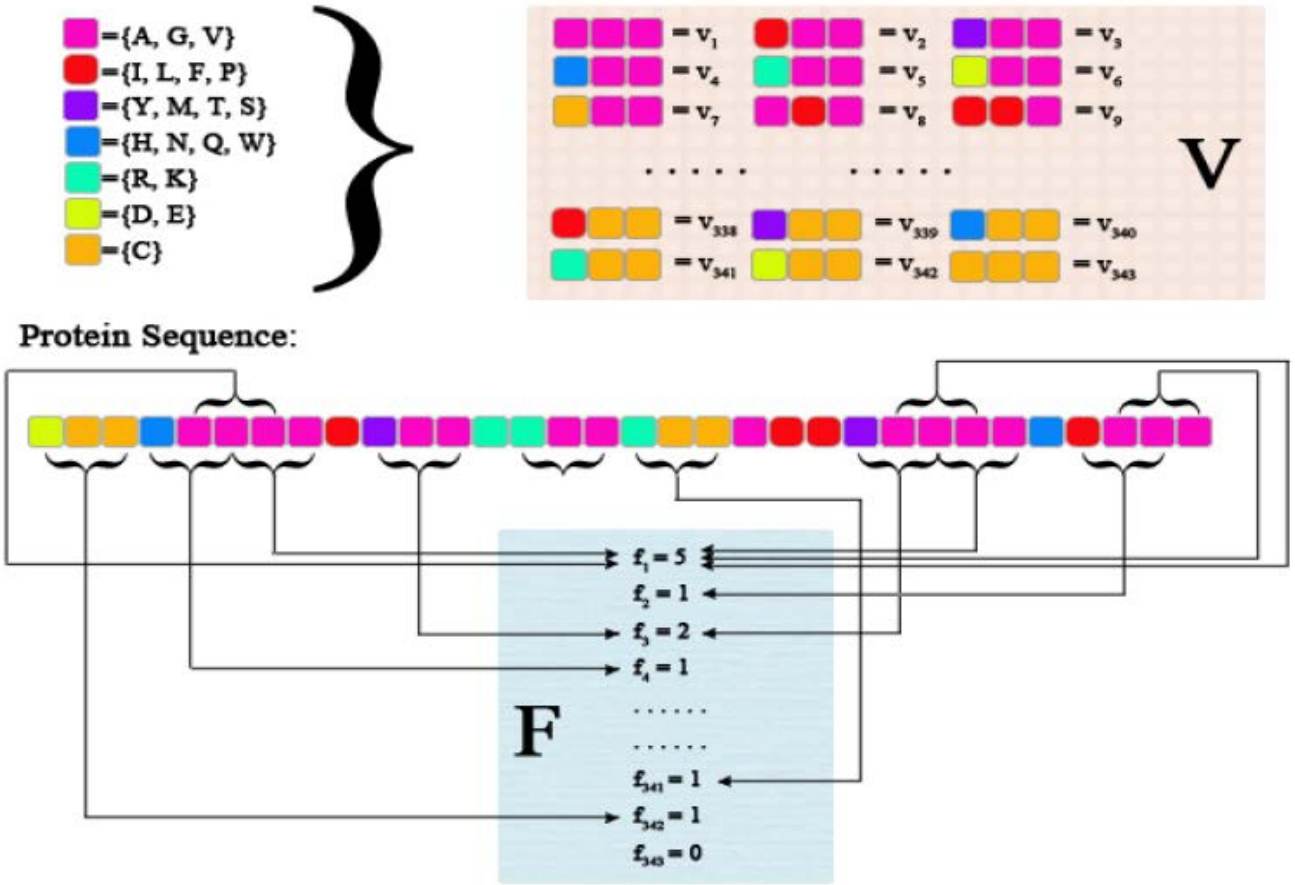
Divide data using a priori knowledge



Positional Weighted Matrix

Pos.	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	0.23	0.01	0.11	0.01	0.08	0.09	0	0.01	0.01	0.01	0.04	0.05	0.04
R	0.14	0.02	0.02	0.02	0.1	0	0	0	0	0	0.03	0	0.02
N	0.07	0.01	0.08	0.04	0.02	0.02	0	0.01	0.02	0.01	0	0.04	0.02
D	0.08	0.25	0.1	0.24	0.13	0.58	0	0.13	0.16	0.09	0.15	0.35	0.07
C	0	0	0	0	0.01	0	0	0	0	0	0	0.01	0.03
G	0.01	0.03	0.17	0.02	0.02	0.01	0	0.15	0.27	0.04	0.01	0.02	0.13
E	0.12	0.2	0.18	0.17	0.14	0.09	0	0.09	0.07	0.09	0.12	0.07	0.05
Q	0.03	0.01	0.02	0.18	0.17	0.01	0	0.02	0.08	0.02	0.09	0.04	0.02
H	0.01	0.08	0.01	0.07	0.02	0.02	0	0	0.03	0.01	0.02	0.03	0.02
I	0.02	0.09	0	0.02	0.02	0.09	0	0.08	0.01	0.02	0.02	0	0.01
L	0.05	0.02	0.04	0.02	0.05	0.01	0	0.01	0.02	0.04	0.04	0.04	0.01
K	0.02	0.03	0	0	0	0	0	0.01	0	0.02	0.02	0.04	0.08
M	0.03	0.02	0.03	0	0	0	0	0.08	0.01	0.13	0.28	0	0.01
F	0.02	0.02	0.02	0.04	0.02	0.01	0	0.01	0.02	0.02	0	0.13	0.28
P	0.05	0.03	0.02	0.01	0.04	0	0	0.05	0.02	0.03	0.05	0.04	0.07
S	0.05	0.04	0.09	0.03	0	0.01	0	0.02	0.02	0.03	0.02	0.03	0.04
T	0.02	0.04	0.05	0.05	0.04	0	0	0.25	0.02	0.08	0.01	0.03	0.05
W	0.01	0	0	0	0	0	0	0	0.11	0.28	0	0.01	0
Y	0	0.07	0.03	0.05	0.1	0.07	1	0.05	0.11	0.05	0.05	0.04	0.01
V	0.02	0.02	0.01	0	0.02	0	0	0.02	0.01	0.01	0.02	0	0.02

Conjoint Triad Method



Position Specific Dipeptide Propensity

Pos.	-3	-2	-1	0	1	2	3
AA							
AR							
AN							
AD							
AC							
…							
VW							
VY							
VV							

$$z_{i,j} = \frac{a_{i,j}^{+} - a_{i,j}^{-}}{s_{i,j}^{-}}$$