Protein Function Prediction

Jianlin Cheng, PhD

Department of Computer Science
Informatics Institute

2009

References

- Slides and documents at: www.geneontology.org
- R. Rentzsch and C.A. Orengo, Trends Biotechnology, 2009.

Widely Used Systems for Protein Function Definition

- Enzyme Commission (EC), Transporter Classification (TC)
- Riley scheme: assign prokaryotic gene products to cellular processes
- The MIPS Functional Catalogue (FUNCAT): extension of Riley to all three kinds of life
- Kyoto Enclyclopedia of Genes and Genomes (KEGG)
- **Gene Ontology (GO)**: molecular function, biological process, and cellular component.

Gene Ontology widely adopted

FlyBase

AgBase

A Comparative mapping resource

GRAMENE

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

UniProt
the universal protein resource

The Wellcome Trust Sanger Institute

InterPro

NCBI

AstraZeneca

www.geneontology.org

Gene Ontology

- Biological process ontology

Which process is a gene product involved in?

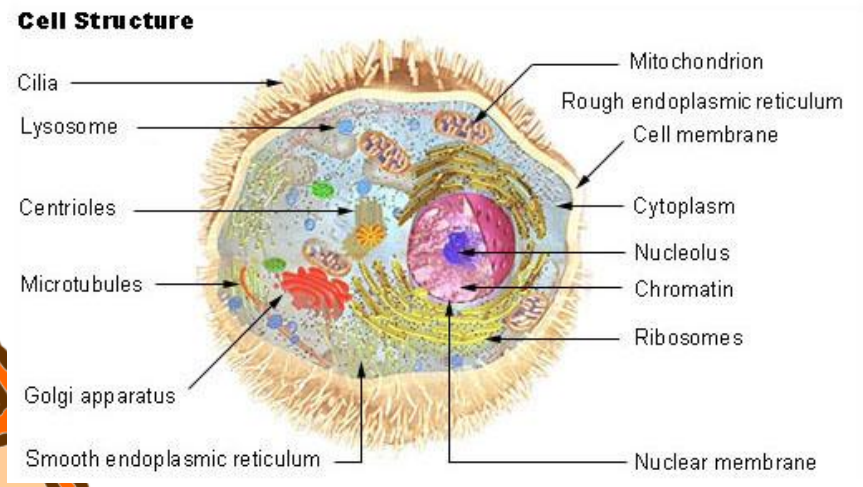
- Molecular function ontology

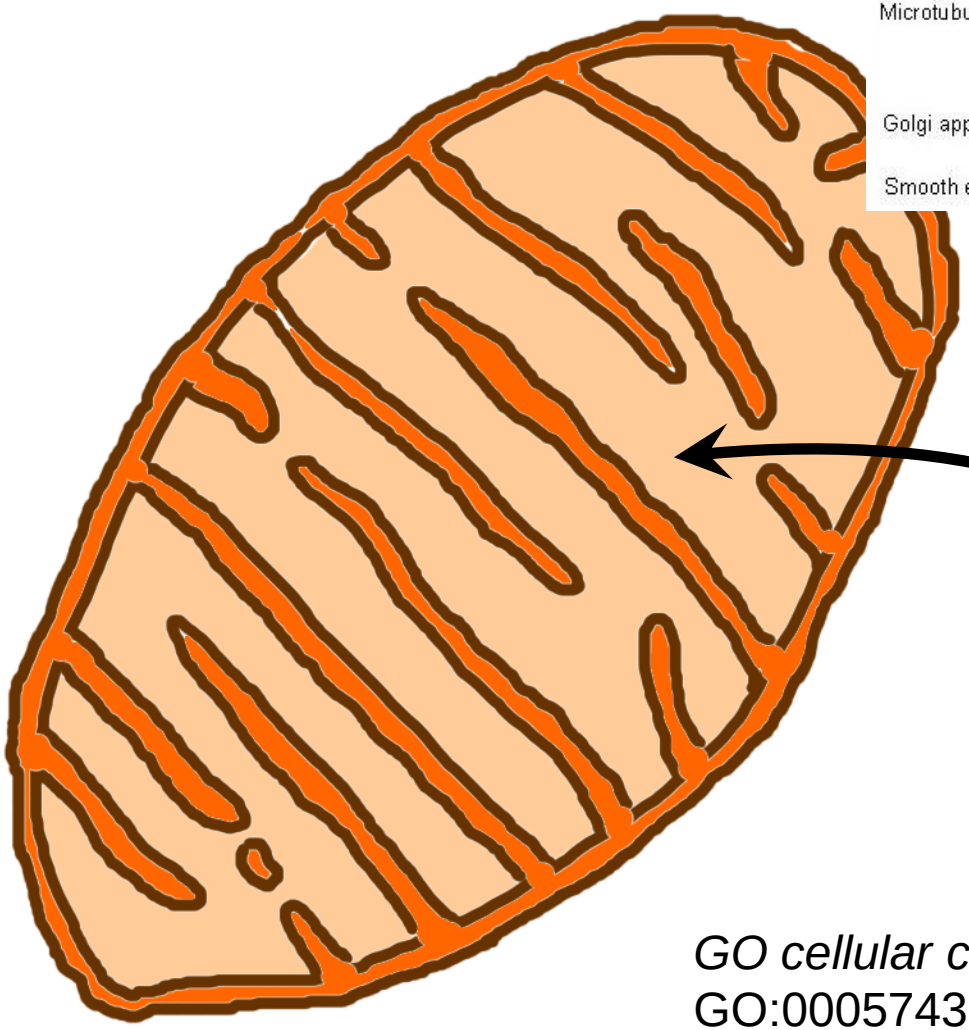
Which molecular function does a gene product have?

- Cellular component ontology

Where does a gene product act?

Where is it?





Mitochondrial p450

GO cellular component term:
GO:0005743

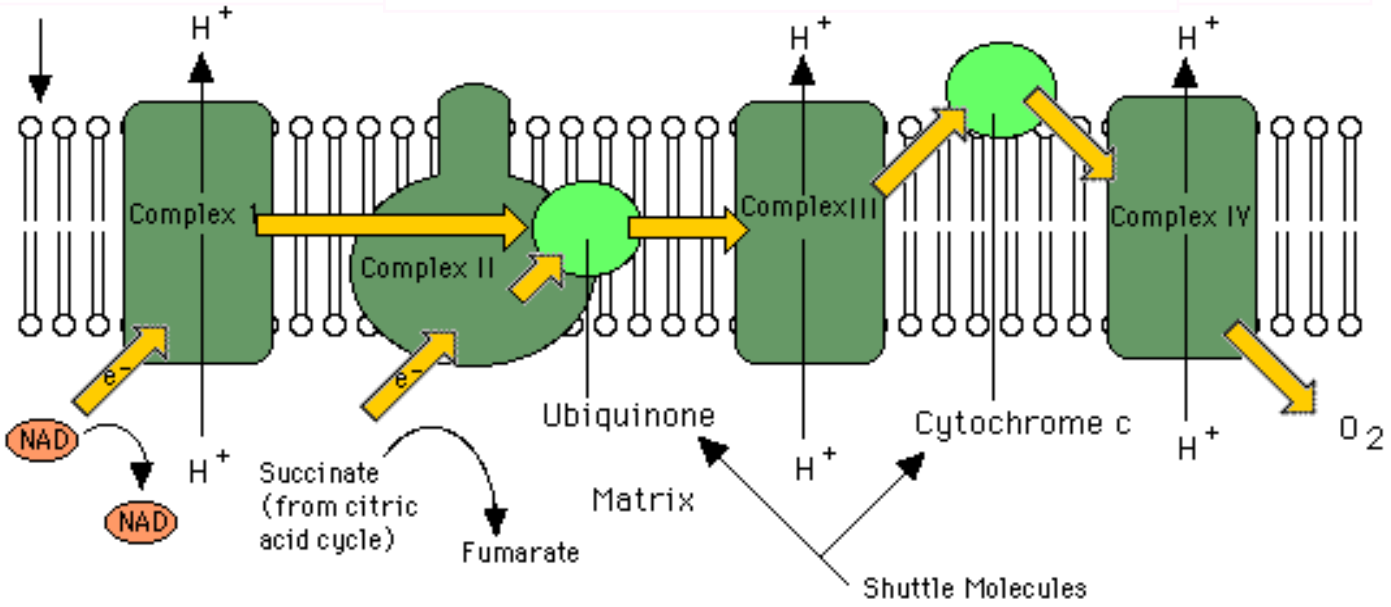
What does it do?

substrate + O_2 = CO_2 +H_20 product

monooxygenase activity

GO molecular function term:
GO:0004497

Which process is this?



electron transport

GO biological process term:
GO:0006118

http://ntri.tamuk.edu/cell/mitochondrion/krebpic.html

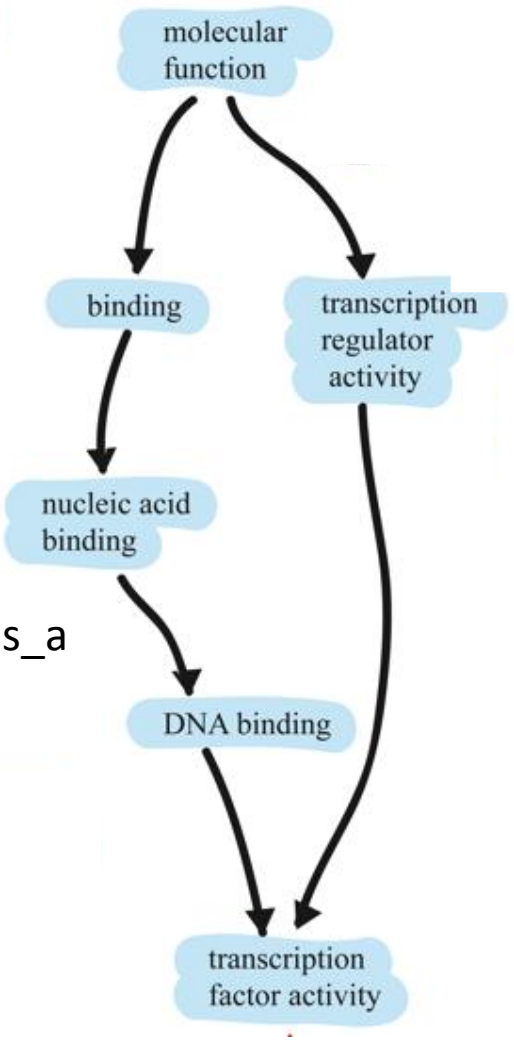
Molecular function ontology

Nucleic acid binding is a type of binding.

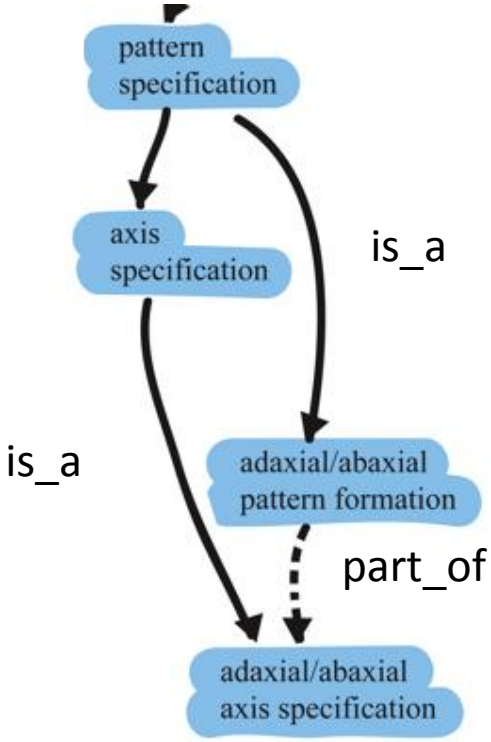
is_a

DNA binding is a type of nucleic acid binding.

is_a



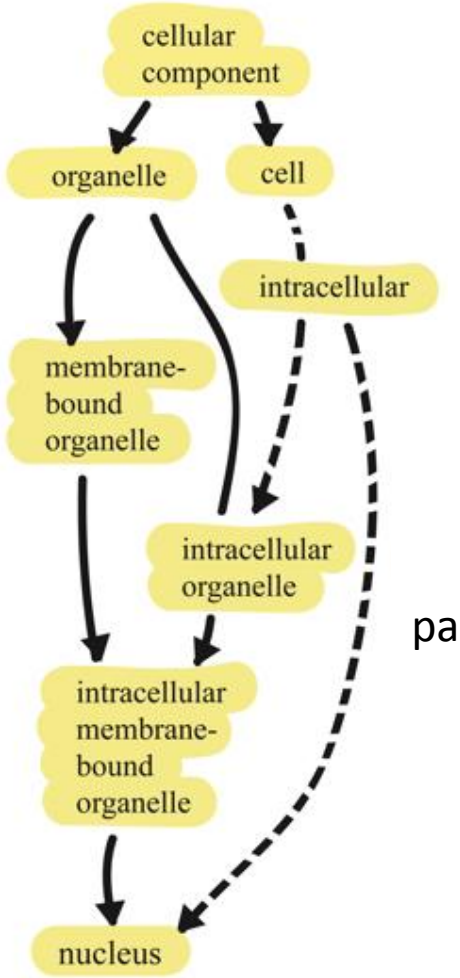
Biological process ontology



Adaxial/abaxial pattern formation is a type of pattern specification.

Adaxial/abaxial pattern specification is a part of adaxial/abaxial pattern formation.

Cellular component ontology



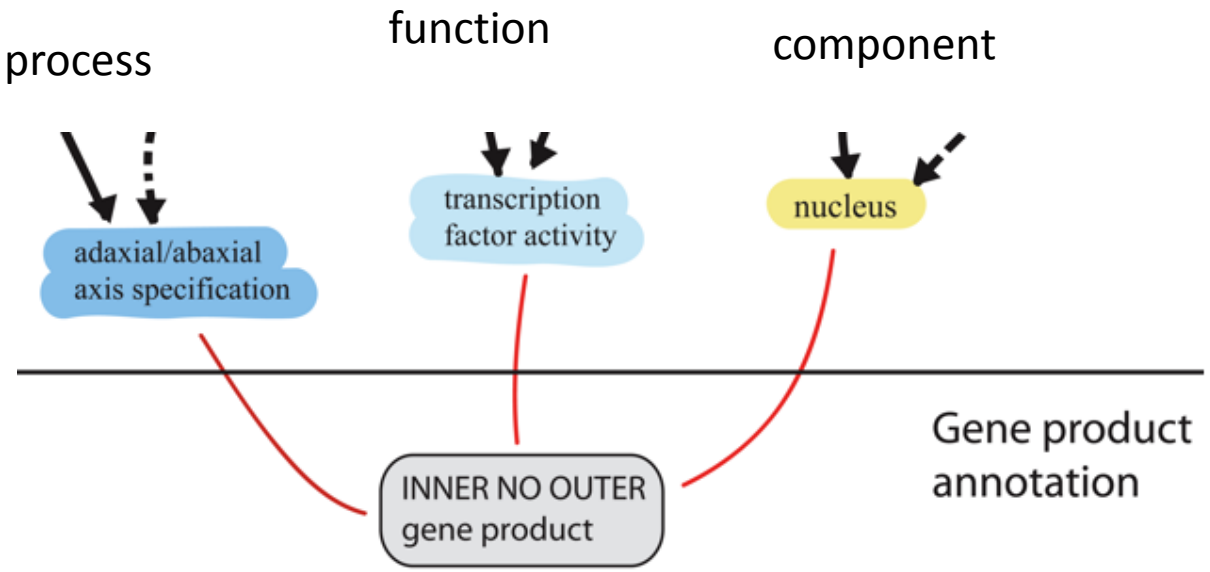
is_a

membrane-bound organelle is a type of organelle

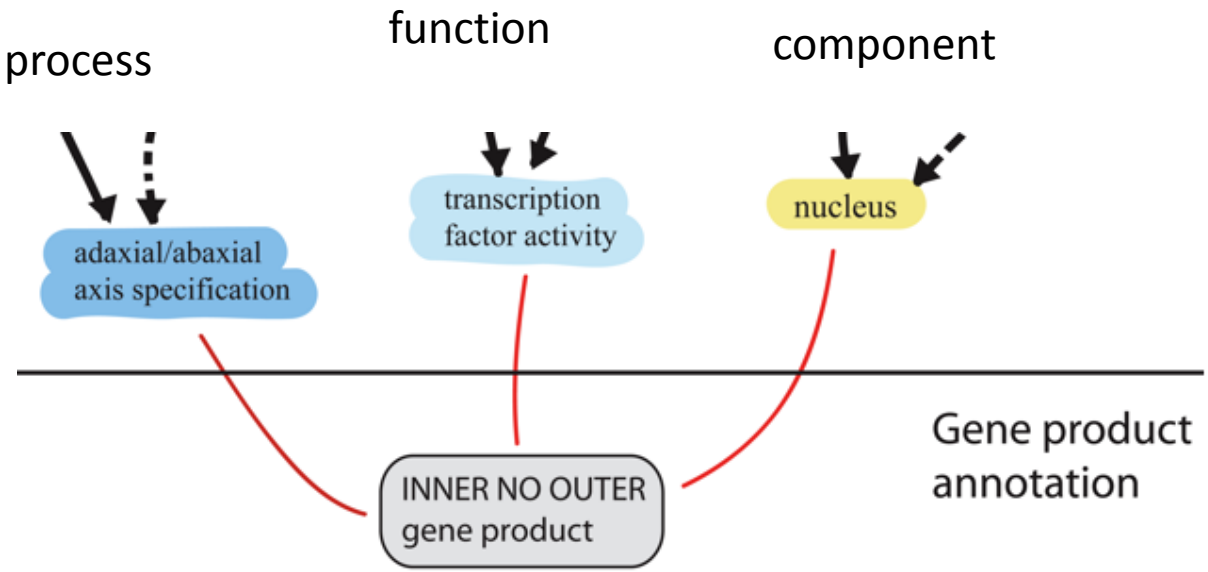
part_of

nucleus is part of the intracellular domain

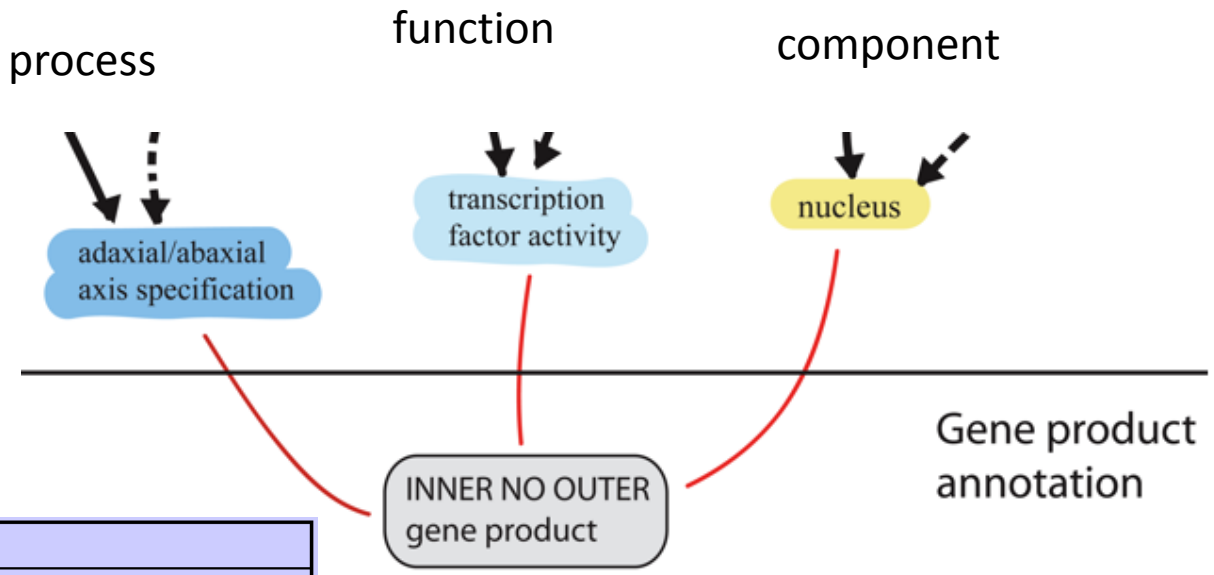
Categorizing gene products is called ‘annotation’.



The gene product *inner no outer* is involved in adaxial/abaxial axis specification.



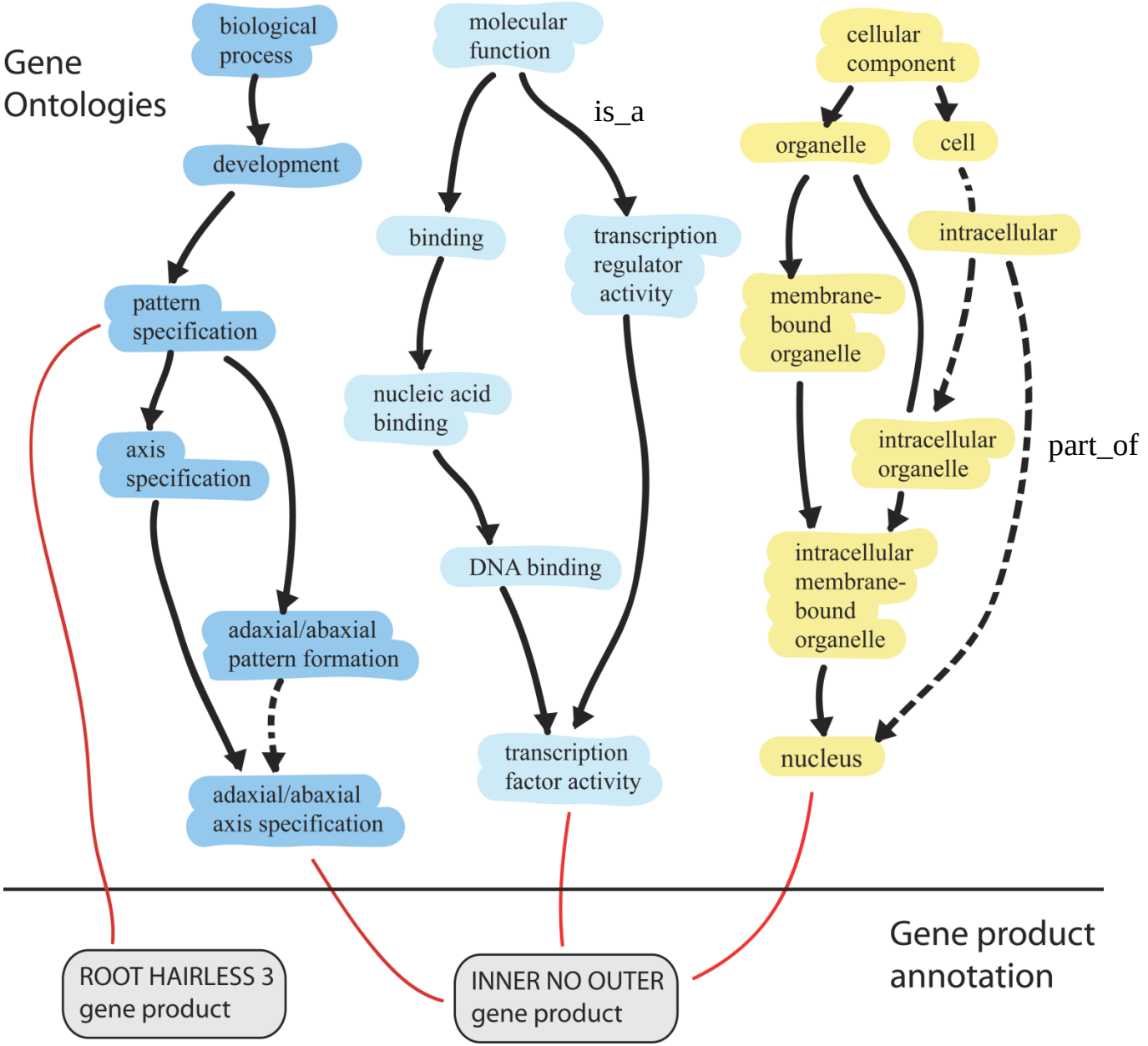
The gene product *inner no outer* has transcription factor activity.



Code	Definition
IEA	Inferred from **E**lectronic **A**nnotation
IDA	Inferred from **D**irect **A**ssay
IEP	Inferred from **E**xpression **P**attern
IGI	Inferred from **G**enetic **I**nteraction
IMP	Inferred from **M**utant **P**henotype
IPI	Inferred from **P**hysical **I**nteraction
ISS	Inferred from **S**equence **S**imilarity
TAS	**T**raceable **A**uthor **S**tatement
NAS	**N**on-traceable **A**uthor **S**tatement
RCA	**R**eviewed **C**omputational **A**nalysis
IC	Inferred from **C**urator
ND	**N**o **D**ata

The gene product *inner no outer* is active in the nucleus.

Clark *et al.*, 2005



Fun: Biological Process

courtship behavior

The Gene Ontology
is like a dictionary

Each
concept has:

- a name
- a definition
- an ID number
- Parent nodes

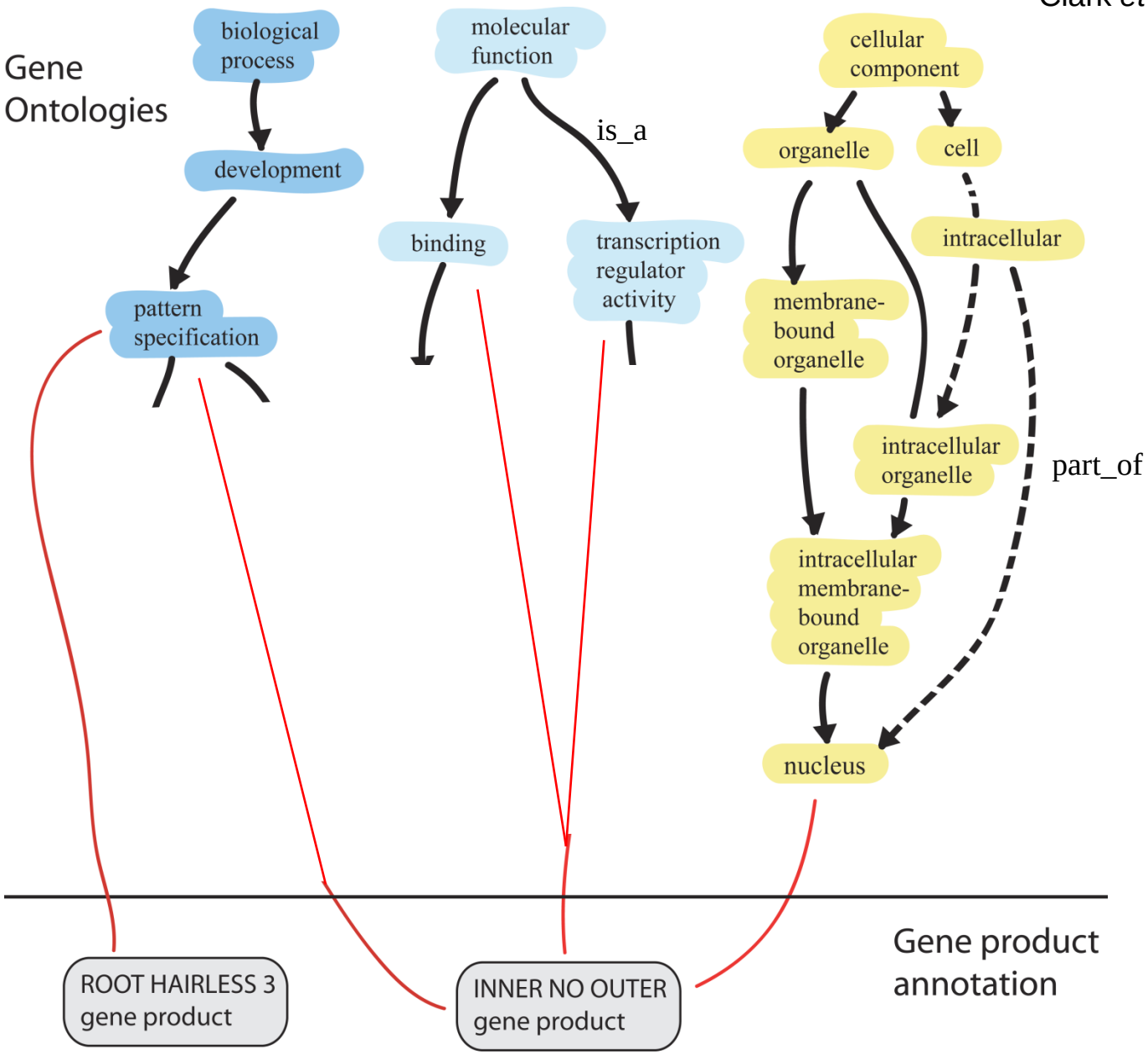
term: transcription initiation

id: GO:0006352

definition: Processes involved in the assembly of the RNA polymerase complex at the promoter region of a DNA template resulting in the subsequent synthesis of RNA from that promoter.

Parent nodes: GO:0002221, is-a

Clark *et al.*, 2005



Current State of Function of Model Genome Annotation

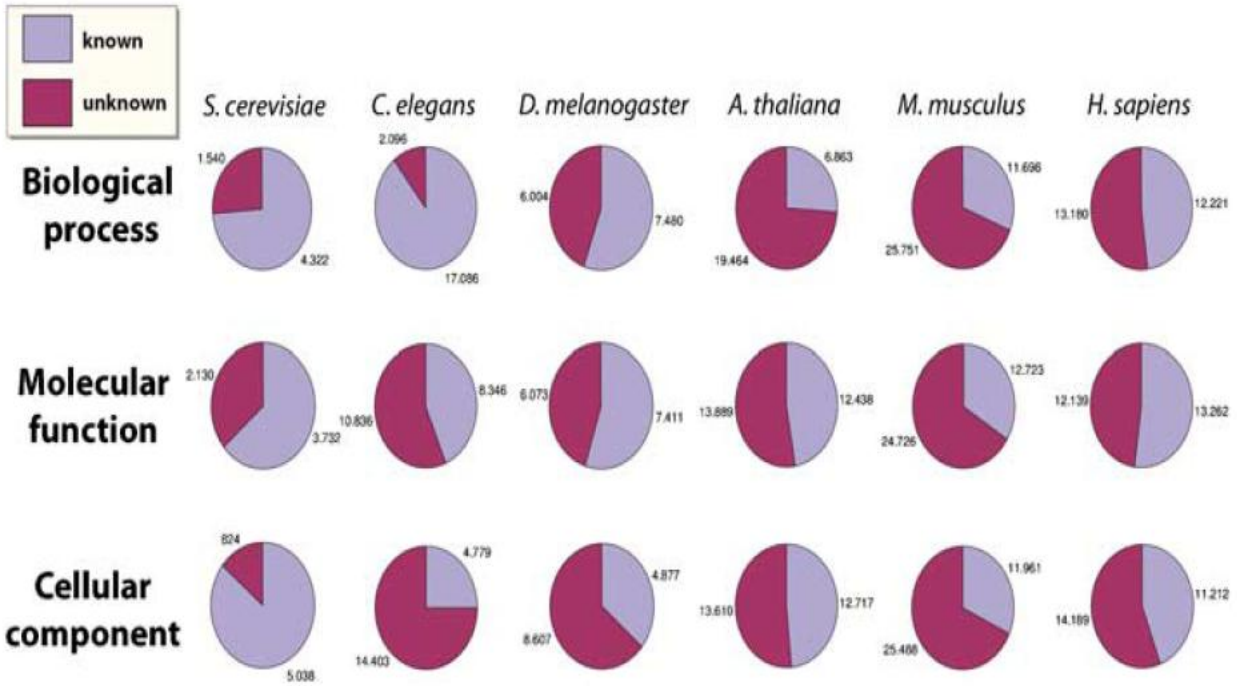
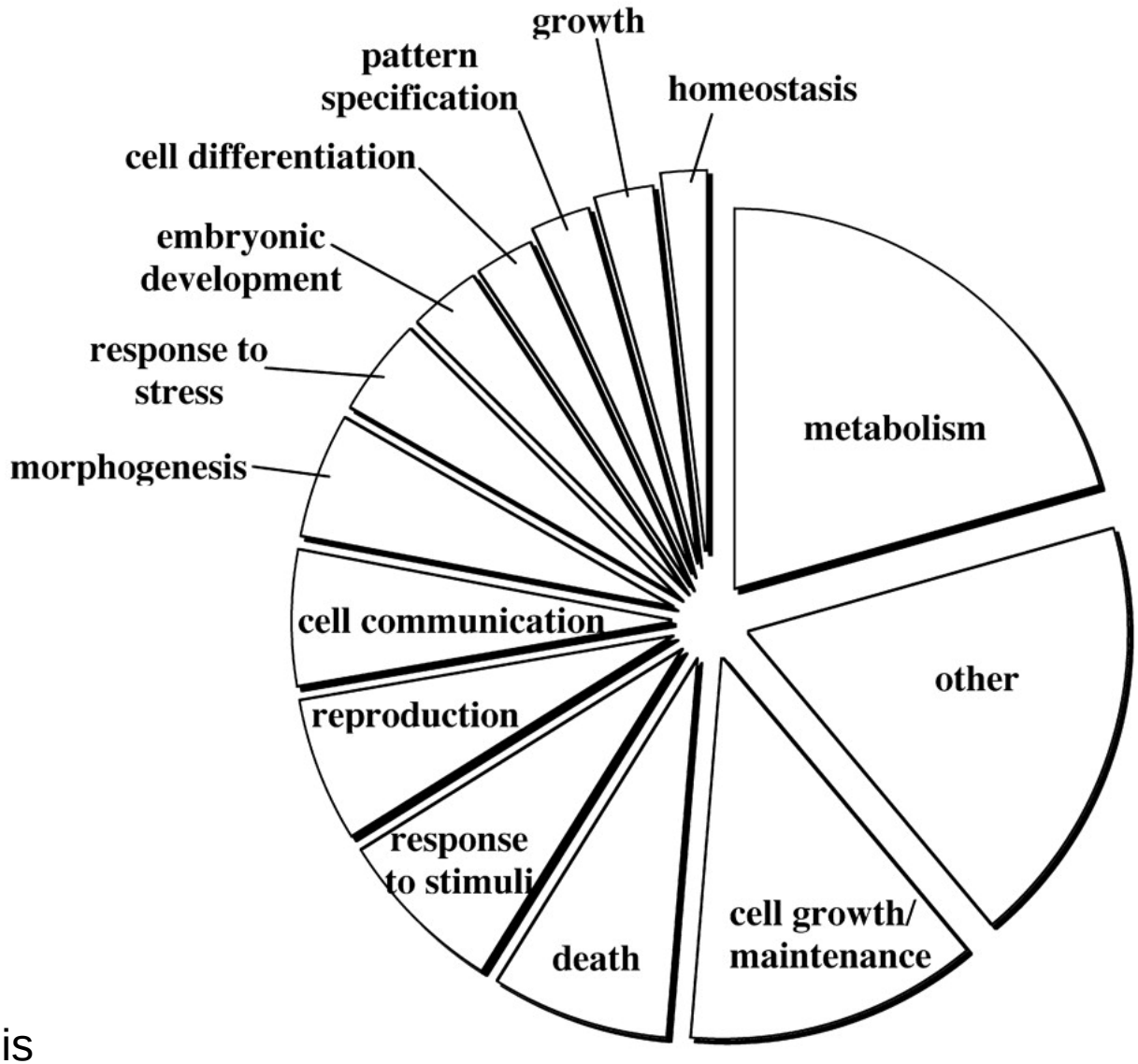


Figure 1 Extent of annotation of proteins in model species. For each species, the charts give the fractions and numbers of annotated and unannotated proteins, according to the three ontologies of the GO annotation. The numbers are based on the Entrez Gene and the WormBase databases as of September 2006.

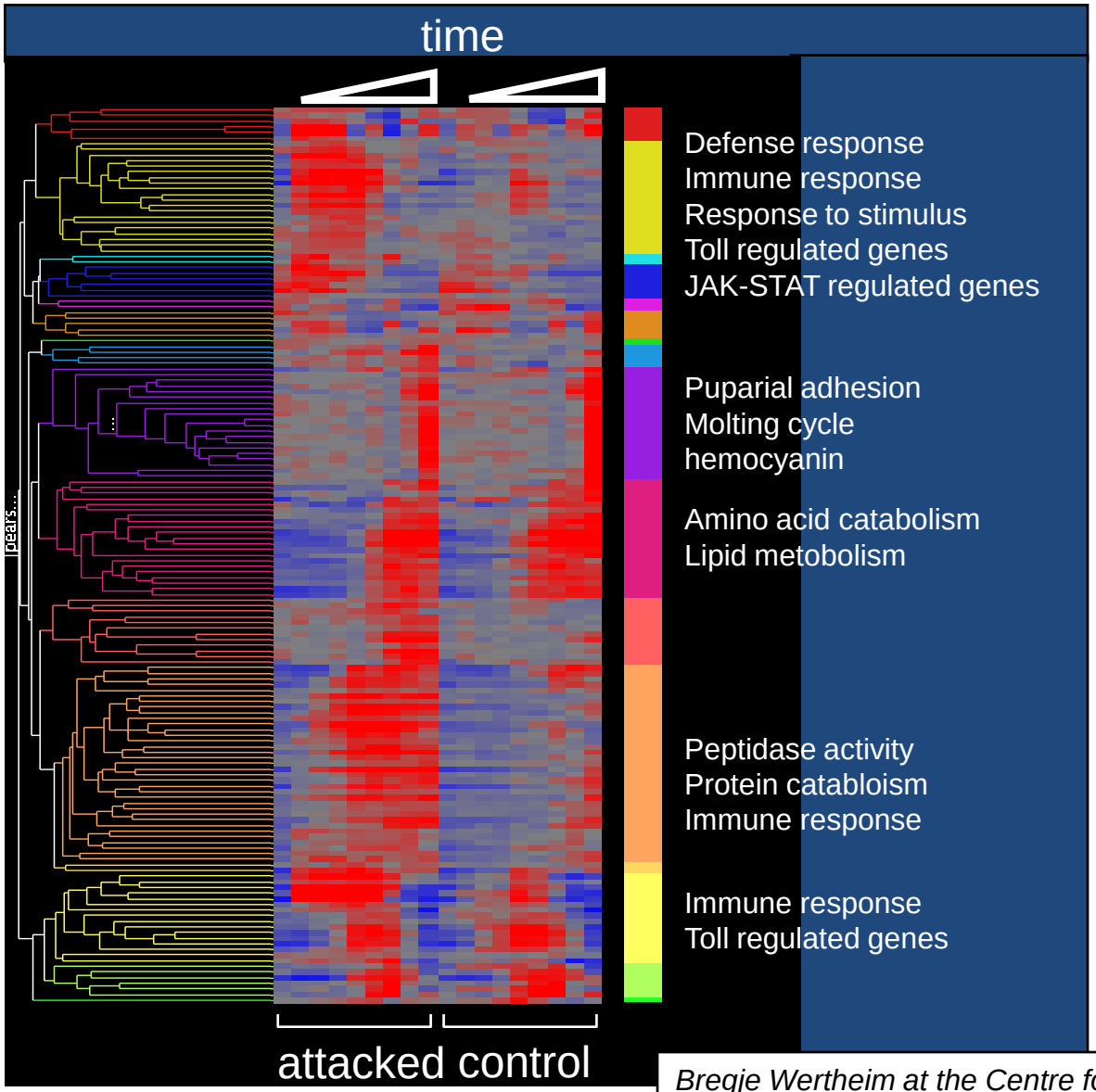
Sharan et al., Molecular Systems Biology, 2007



Whole genome analysis
(J. D. Munkvold *et al*., 2004)

…analysis of high-throughput data according to GO

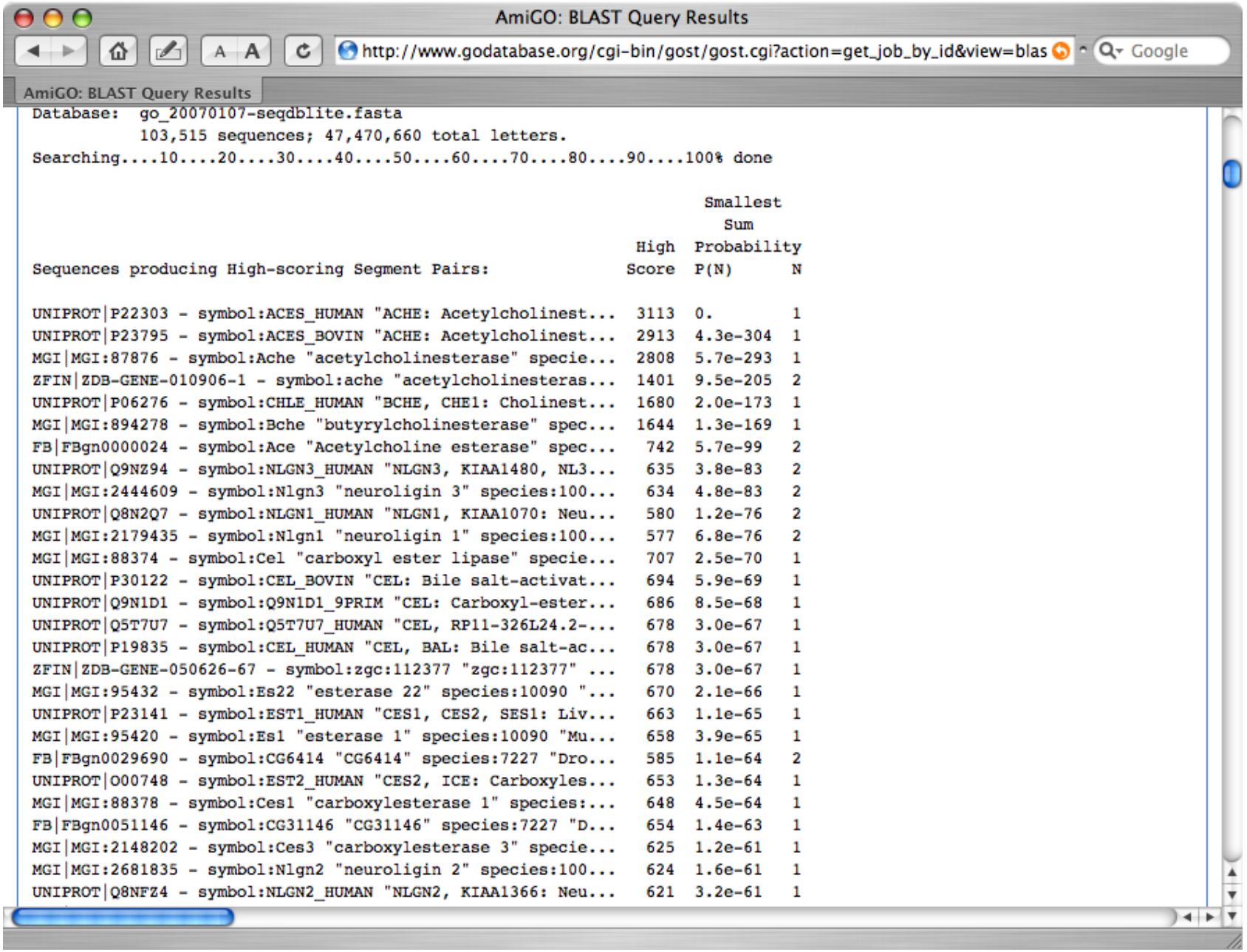
MicroArray data analysis



Bregje Wertheim at the Centre for Evolutionary Genomics, Department of Biology, UCL and Eugene Schuster Group, EBI.

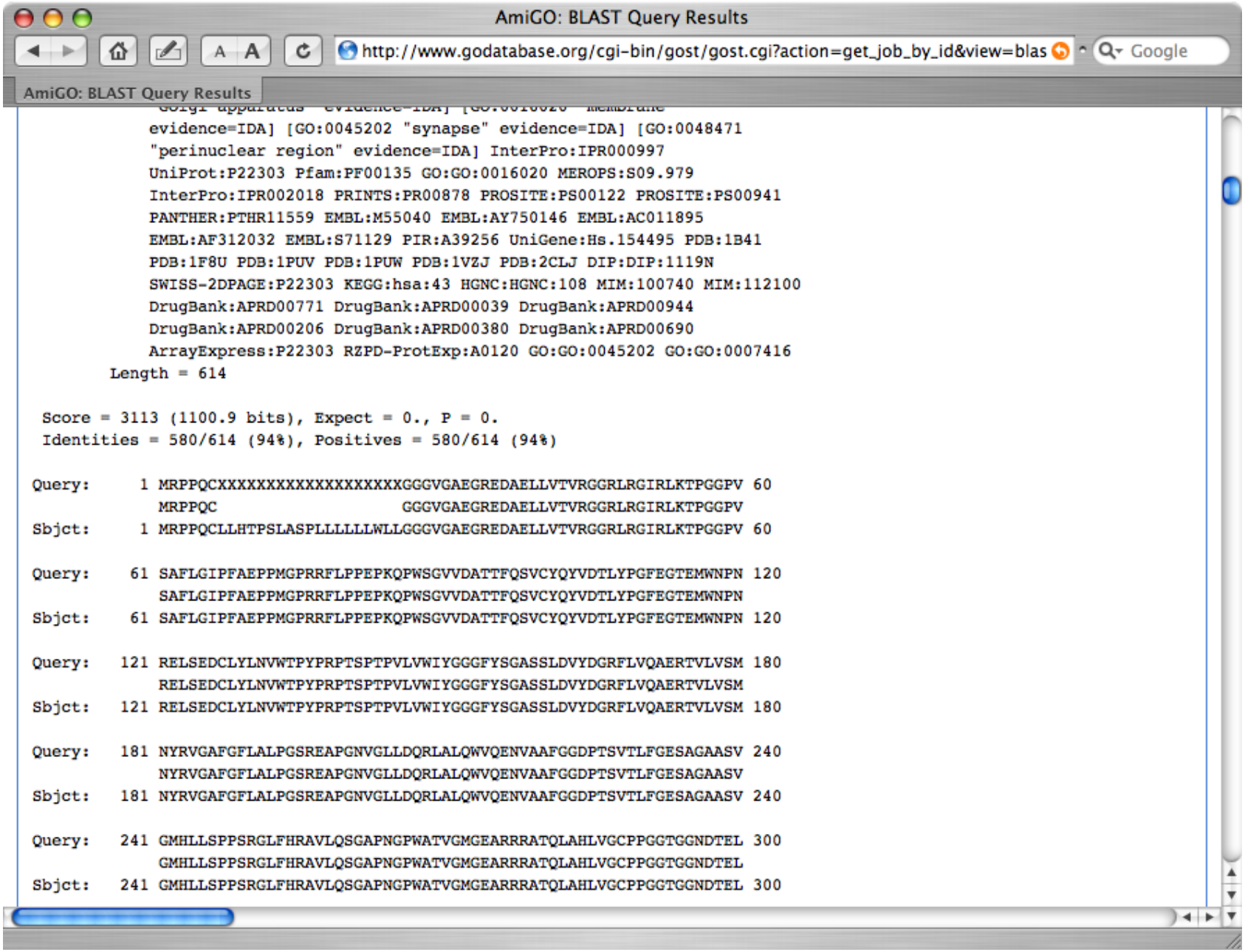
Simple Function Prediction

- The easiest way to infer the molecular function of an uncharacterized sequence is by finding an obvious (highly sequence-similar) and well-characterized homologue.
- BLAST (sequence-sequence local alignment tool)
- PSI-BLAST (profile-sequence local alignment tool)
- Problem: many proteins do not have obvious homologs



Database: go_20070107-seqdblite.fasta
103,515 sequences; 47,470,660 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
UNIPROT|P22303 - symbol:ACES_HUMAN "ACHE: Acetylcholinest...	3113	0.	1
UNIPROT|P23795 - symbol:ACES_BOVIN "ACHE: Acetylcholinest...	2913	4.3e-304	1
MGI|MGI:87876 - symbol:Ache "acetylcholinesterase" specie...	2808	5.7e-293	1
ZFIN|ZDB-GENE-010906-1 - symbol:ache "acetylcholinesteras...	1401	9.5e-205	2
UNIPROT|P06276 - symbol:CHLE_HUMAN "BCHE, CHE1: Cholinest...	1680	2.0e-173	1
MGI|MGI:894278 - symbol:Bche "butyrylcholinesterase" spec...	1644	1.3e-169	1
FB|FBgn0000024 - symbol:Ace "Acetylcholine esterase" spec...	742	5.7e-99	2
UNIPROT|Q9NZ94 - symbol:NLGN3_HUMAN "NLGN3, KIAA1480, NL3...	635	3.8e-83	2
MGI|MGI:2444609 - symbol:Nlgn3 "neuroligin 3" species:100...	634	4.8e-83	2
UNIPROT|Q8N2Q7 - symbol:NLGN1_HUMAN "NLGN1, KIAA1070: Neu...	580	1.2e-76	2
MGI|MGI:2179435 - symbol:Nlgn1 "neuroligin 1" species:100...	577	6.8e-76	2
MGI|MGI:88374 - symbol:Cel "carboxyl ester lipase" specie...	707	2.5e-70	1
UNIPROT|P30122 - symbol:CEL_BOVIN "CEL: Bile salt-activat...	694	5.9e-69	1
UNIPROT|Q9N1D1 - symbol:Q9N1D1_9PRIM "CEL: Carboxyl-ester...	686	8.5e-68	1
UNIPROT|Q5T7U7 - symbol:Q5T7U7_HUMAN "CEL, RP11-326L24.2-...	678	3.0e-67	1
UNIPROT|P19835 - symbol:CEL_HUMAN "CEL, BAL: Bile salt-ac...	678	3.0e-67	1
ZFIN|ZDB-GENE-050626-67 - symbol:zgc:112377 "zgc:112377" ...	678	3.0e-67	1
MGI|MGI:95432 - symbol:Es22 "esterase 22" species:10090 "...	670	2.1e-66	1
UNIPROT|P23141 - symbol:EST1_HUMAN "CES1, CES2, SES1: Liv...	663	1.1e-65	1
MGI|MGI:95420 - symbol:Es1 "esterase 1" species:10090 "Mu...	658	3.9e-65	1
FB|FBgn0029690 - symbol:CG6414 "CG6414" species:7227 "Dro...	585	1.1e-64	2
UNIPROT|O00748 - symbol:EST2_HUMAN "CES2, ICE: Carboxyles...	653	1.3e-64	1
MGI|MGI:88378 - symbol:Ces1 "carboxylesterase 1" species:...	648	4.5e-64	1
FB|FBgn0051146 - symbol:CG31146 "CG31146" species:7227 "D...	654	1.4e-63	1
MGI|MGI:2148202 - symbol:Ces3 "carboxylesterase 3" specie...	625	1.2e-61	1
MGI|MGI:2681835 - symbol:Nlgn2 "neuroligin 2" species:100...	624	1.6e-61	1
UNIPROT|Q8NFZ4 - symbol:NLGN2_HUMAN "NLGN2, KIAA1366: Neu...	621	3.2e-61	1


AmiGO: BLAST Query Results
http://www.godatabase.org/cgi-bin/gost/gost.cgi?action=get_job_by_id&view=blas
Google
AmiGO: BLAST Query Results
evidence=IDA] [GO:0045202 "synapse" evidence=IDA] [GO:0048471
"perinuclear region" evidence=IDA] InterPro:IPR000997
UniProt:P22303 Pfam:PF00135 GO:GO:0016020 MEROPS:S09.979
InterPro:IPR002018 PRINTS:PR00878 PROSITE:PS00122 PROSITE:PS00941
PANTHER:PTHR11559 EMBL:M55040 EMBL:AY750146 EMBL:AC011895
EMBL:AF312032 EMBL:S71129 PIR:A39256 UniGene:Hs.154495 PDB:1B41
PDB:1F8U PDB:1PUV PDB:1PUW PDB:1VZJ PDB:2CLJ DIP:DIP:1119N
SWISS-2DPAGE:P22303 KEGG:hsa:43 HGNC:HGNC:108 MIM:100740 MIM:112100
DrugBank:APRD00771 DrugBank:APRD00039 DrugBank:APRD00944
DrugBank:APRD00206 DrugBank:APRD00380 DrugBank:APRD00690
ArrayExpress:P22303 RZPD-ProtExp:A0120 GO:GO:0045202 GO:GO:0007416
Length = 614
Score = 3113 (1100.9 bits), Expect = 0., P = 0.
Identities = 580/614 (94%), Positives = 580/614 (94%)
Query: 1 MRPPQCXXXXXXXXXXXXXXXXXGGGVGAEGREDAELLVTVRGGRLRGIRLKTPGGPV 60
MRPPQC GGGVGAEGREDAELLVTVRGGRLRGIRLKTPGGPV
Sbjct: 1 MRPPQCLLHTPSLASPLLLLLLWLLGGGVGAEGREDAELLVTVRGGRLRGIRLKTPGGPV 60
Query: 61 SAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYPGFEGTEMWNPN 120
SAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYPGFEGTEMWNPN
Sbjct: 61 SAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYPGFEGTEMWNPN 120
Query: 121 RELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSM 180
RELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSM
Sbjct: 121 RELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSM 180
Query: 181 NYRVGAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASV 240
NYRVGAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASV
Sbjct: 181 NYRVGAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASV 240
Query: 241 GMHLLSPPSRGLFHRAVLQSGAPNGPWATVGMGEARRRATQLAHLVGCPPGGTGGNDTEL 300
GMHLLSPPSRGLFHRAVLQSGAPNGPWATVGMGEARRRATQLAHLVGCPPGGTGGNDTEL
Sbjct: 241 GMHLLSPPSRGLFHRAVLQSGAPNGPWATVGMGEARRRATQLAHLVGCPPGGTGGNDTEL 300

Integrative Approaches

- Similarity grouping
- Phylogenomics
- Sequence patterns
- Sequence clustering
- Machine learning
- Network approach
- Results: at least coarse functional characterization

Similarity Group Methods

Idea: Similarly, the sequences found in a similarity search will usually share some annotated functions – some GO terms will be significantly enriched over others

PFP Method

- Sequence hit retrieved by a PSI-BLAST search
- Associated GO terms are scored according to the alignment expectation value (E-value) provided by PSI-BLAST.
- The scores for terms associated to several sequence hits are combined by summation. This scoring system ranks GO terms according to both (1) their frequency of association to similar sequences and (2) the degree of similarity those sequences share with the query.
- A GO term, fa, is scored as follows:

Hawkins et al., Proteins, 2008

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{N\text{func}(i)} \left((-\log(E_{\text{value}(i)}) + b)\delta_{f_j, f_a} \right)$$

- where s(fa) is the final score assigned to the GO term, fa; N is the number of the similar sequences retrieved by PSI-BLAST, Nfunc(i) is the number of GO terms assigned to sequence j, Evalue(i) is the E-value given to the sequence i, and fj is a GO term assigned to the sequence i. delta(fj, fa) returns 1 when fj equals to fa, and 0 otherwise.
- E-value threshold is set to 125.

Function Association Matrix

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{N\text{func}(i)} \left((-\log(E_{\text{value}(i)}) + b) P(f_a | f_j) \right),$$

$$P(f_a | f_j) = \frac{c(f_a, f_j) + \varepsilon}{c(f_j) + \mu \cdot \varepsilon'}$$

The Function Association Matrix, describes the probability that two GO terms are associated to the same sequence based on the frequency at which they co-occur in UniProt sequences. This allows the FAM to associate function annotations from different GO categories, for example, the biological process ‘‘positive regulation of transcription, DNA-dependent’’ is strongly associated with the molecular function ‘‘DNA binding activity’’ (P(0045893|0003677) = 0.455).

Phylogenomic Apporach

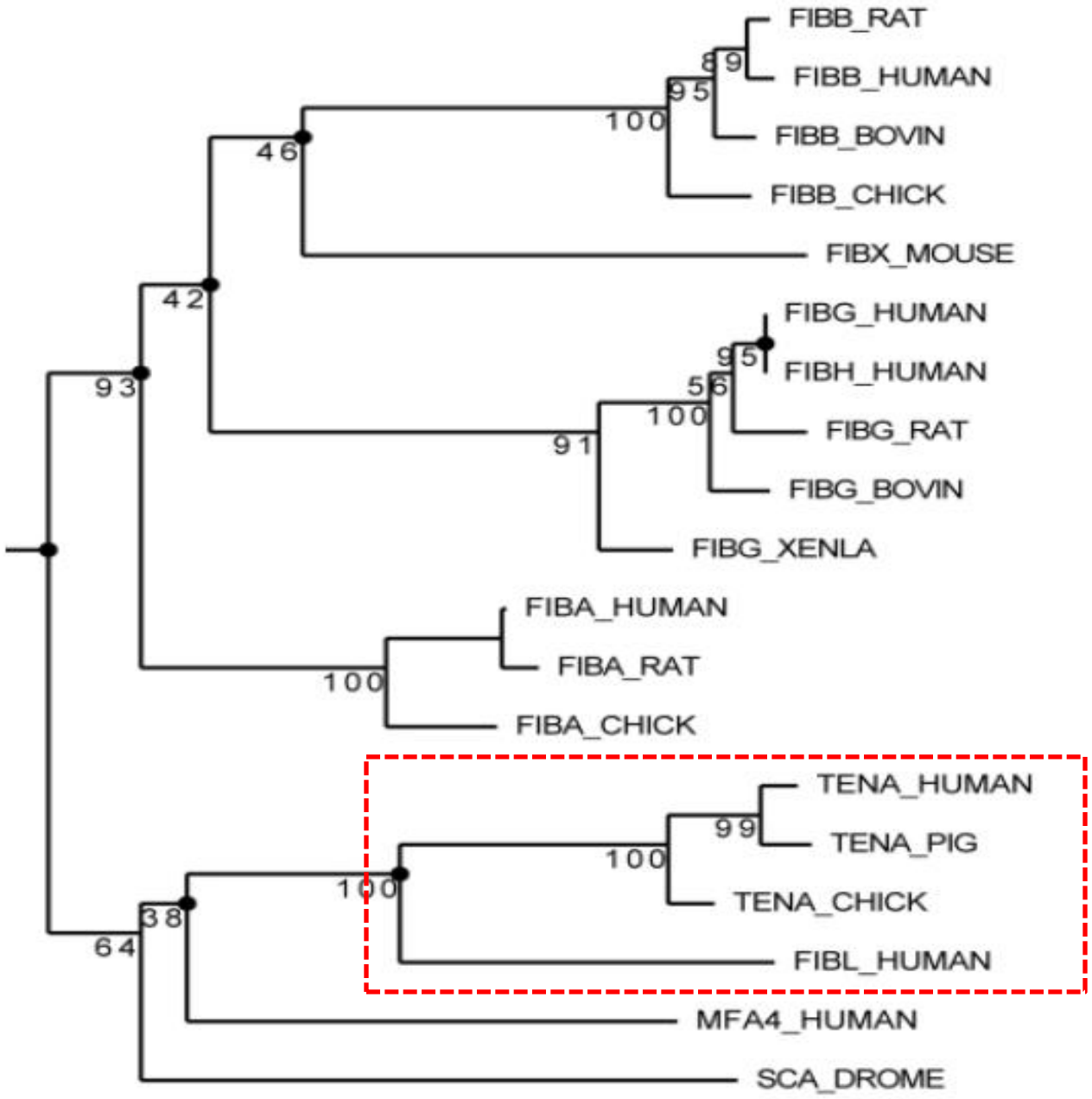
- The accuracy of annotation transfer can be increased further by taking the evolutionary relationships within protein families into account.
- This addresses the difference between orthologous and paralogous relative of a query sequence (i.e. between relatives by speciation and relatives by gene duplication)

- A "**duplication event**" captures a single instance of a gene duplicating into divergent copies of that gene within a single genome;
- a "**speciation event**" captures a single instance of a gene in an ancestral species evolving into divergent copies of a gene in distinct genomes of different species.

Which event more likely preserves function?

Steps

- Find all homologues of the query sequence and align them
- Build a phylogenetic tree and reconcile this tree (make all bifurcations in the tree as either duplication or speciation)
- Transfer functions (primarily) from orthologues



Zmasek and Eddy, Bioinformatics, 2001

SIFTER

1. Given a query protein, we find a Pfam family of a homologous domain, and extract the multiple sequence alignment from the Pfam database
2. Build a rooted phylogenetic tree with PAUP version, using parsimony with the BLOSUM50 matrix
3. Apply Forester version 1.92 to estimate the location of the duplication events at the internal nodes of the phylogeny by reconciling the topological differences between a reference species tree (taken from the Pfam database) and the protein tree.
4. Infer function mainly from orthologs and consider GO annotation evidences

Engelhardt et al., Plot Comput. Biology, 2005

Pros and Cons

- Similarity group methods are ready for whole-genome application (relatively fast)
- Show moderate precision levels, whereas phylogenomic methods are significantly slower but can provide higher precision

Pattern-Based Methods

- Classify proteins by locally conserved sequence patterns, which often indicate the functions of the whole protein (e.g. active site motifs)

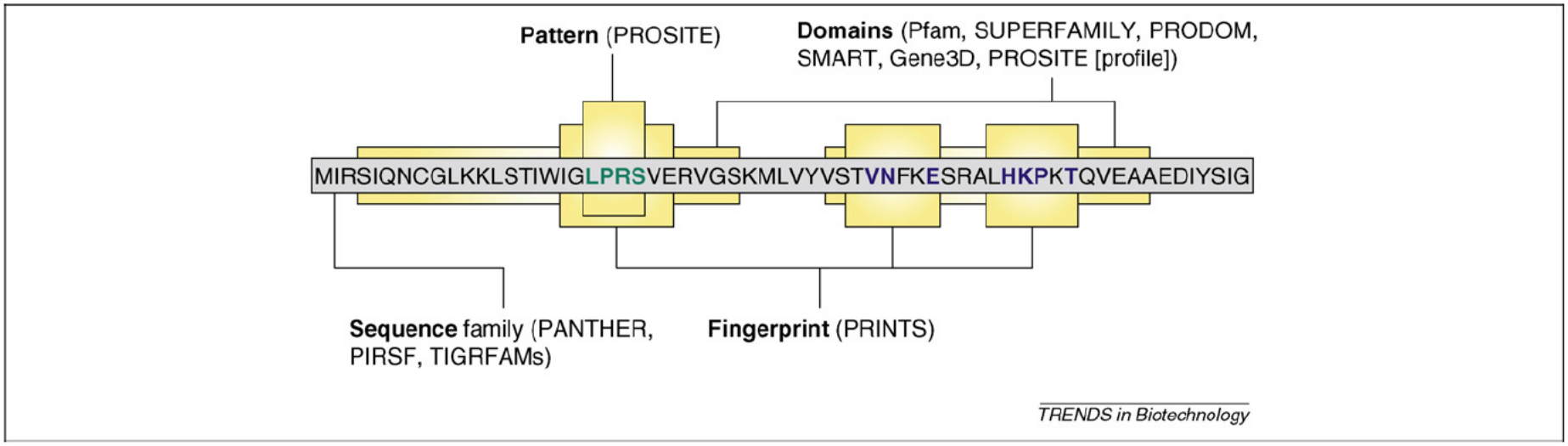


Figure 1. Conserved sequence patterns are tied to protein function. This illustration shows an example sequence, with active site (green) and cofactor-binding site (blue) residues highlighted. The different InterPro member databases (in brackets) group protein sequences into families, based on conserved short patterns, fingerprints (discontinuous patterns), domains or overall similarity.

- **InterPro**: the best gateway to pattern-based functional annotations, which collates patterns at all levels into hierarchically arranged database entries.
- **InterProScan** server is a meta-tool, which scans the query sequence against ten core member databases, from which the output is collected and presented in a simple, non-redundant manner.
- **PROSITE** scan query sequences against short, position-specific residue profiles that are characteristic of individual protein families
- **PRINTS** follows a similar principle but uses discontinuous profiles (“fingerprints”)

InterPro members

- Pfam, SUPERFAMILY, PRODOM, SMART, Gene3D, PANTHER, PIRSF, and TIGRFAMs.
- PRODOM automatically clusters evolutionary conserved sequence segments, based on recursive PSI-BLAST searches of UniProtKB.
- All others use *hidden Markov models* (HMMs), generated from multiple sequence alignments, to represent sequence families

Pfam and SMART

- Pfam focuses on the functional aspect of the “domain” defintion. Classifying sequences into a large number of relatively small (functionally conserved) families.
- SMART consists of a considerably smaller but completely manually curated set of families

SUPERFAMILY and Gene3D

- Based on structural classifications, assigning sequences to the domain families defined in the Structural Classification of Proteins (SCOP) and CATH databases
- These families are usually much bigger (less functionally conserved) than, those in Pfam – they often contain very remote homologues, only detectable by patterns of structural conservation

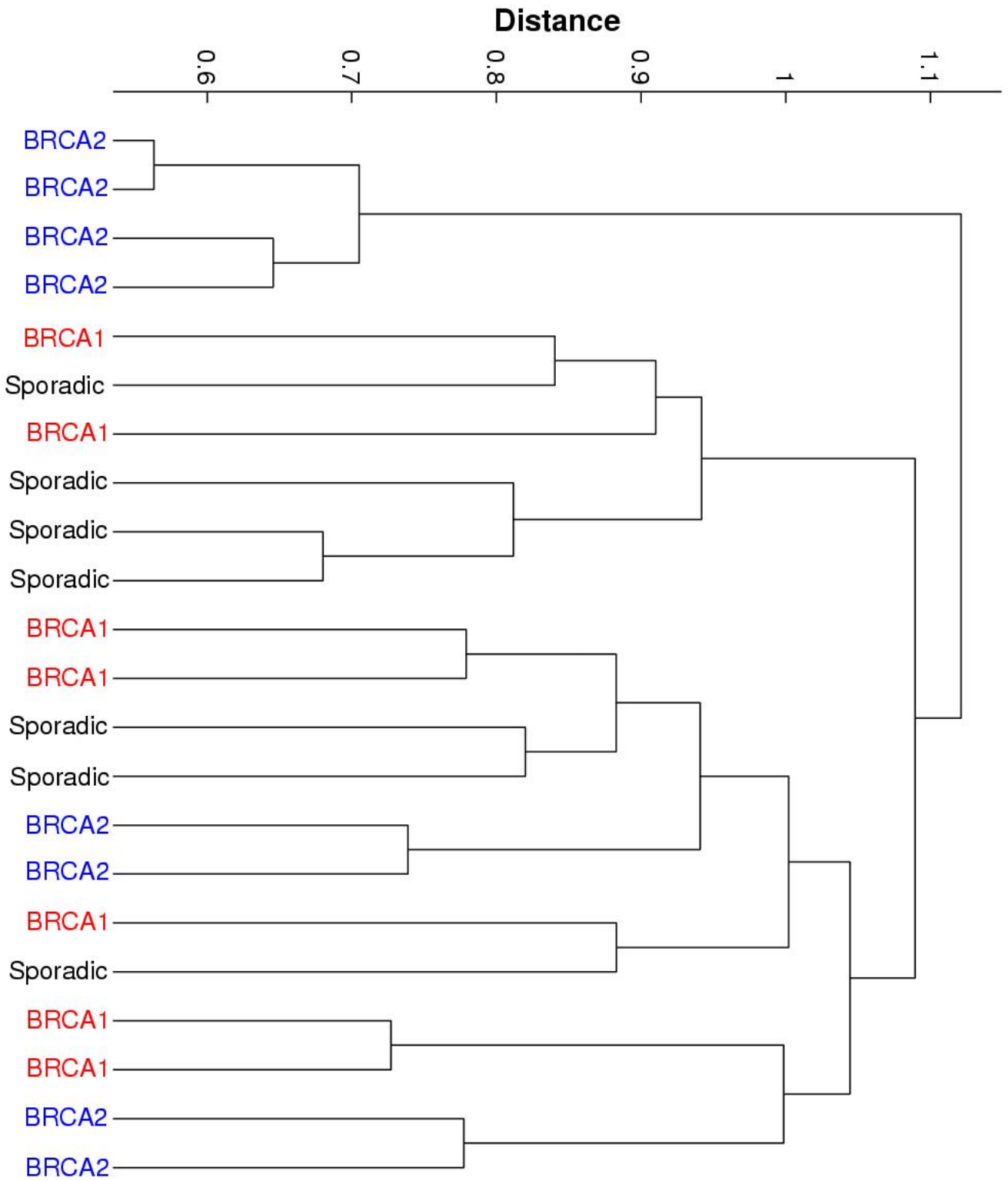
Clustering Approaches

- Cluster the known sequence space, whereby uncharacterized sequences can be functionally annotated by virtue of their clustering with characterized sequences
- Clustering based on sequence similarity (homologues)
- Clustering based on function similarity

ProtoNet

- Sequence similarity clustering-based method
- Annotation transfer within ProtoNet clusters, based on the predominant functions assigned to known members are effective
- Hierarchical clustering of over one million proteins in SwissProt.
- Clustering process is based on an all-against-all BLAST. E-score is used to perform a continuous bottom-up clustering process by joining the two most similar protein clusters at each step
- Filter clusters by function similarity
- Assess the function of novel protein sequences, by finding the best matching cluster for the new sequences

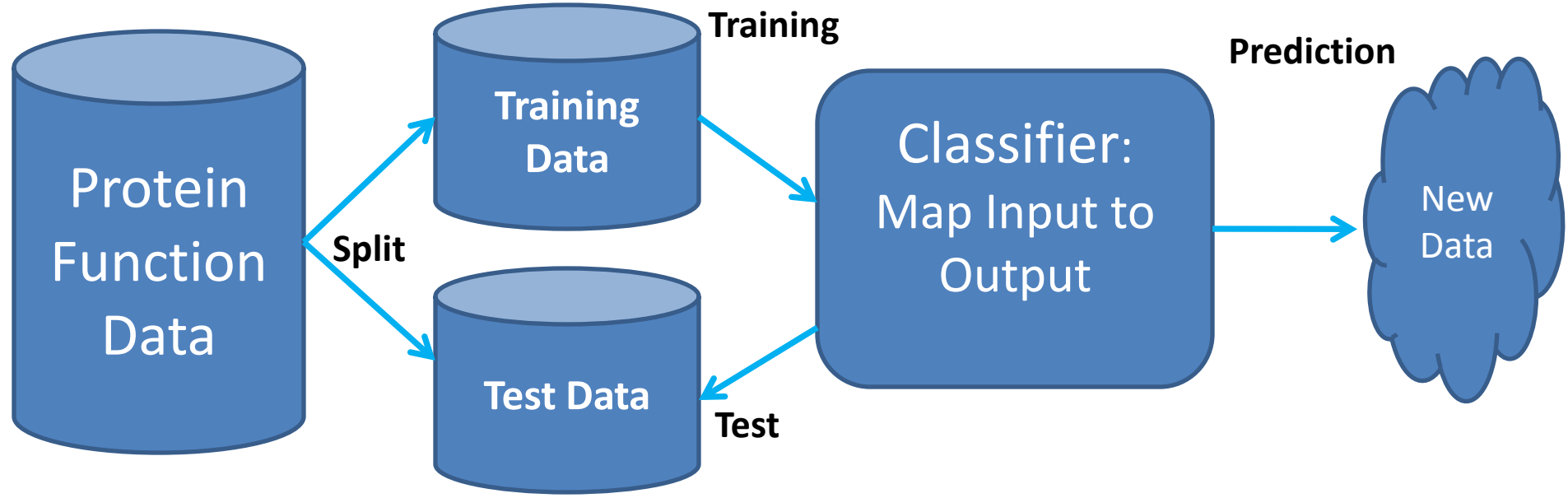
Kaplan et al., Nucleic Acids Research, 2005


Distance
0.6
0.7
0.8
0.9
1
1.1
BRCA2
BRCA2
BRCA2
BRCA2
BRCA1
Sporadic
BRCA1
Sporadic
Sporadic
Sporadic
BRCA1
BRCA1
Sporadic
Sporadic
BRCA2
BRCA2
BRCA1
Sporadic
BRCA1
BRCA1
BRCA2
BRCA2

Machine Learning Methods

- Learn a relationship between characteristic combinations of sequence features function categories in a training set of known sequences.
- Support vector machines
- Neural networks

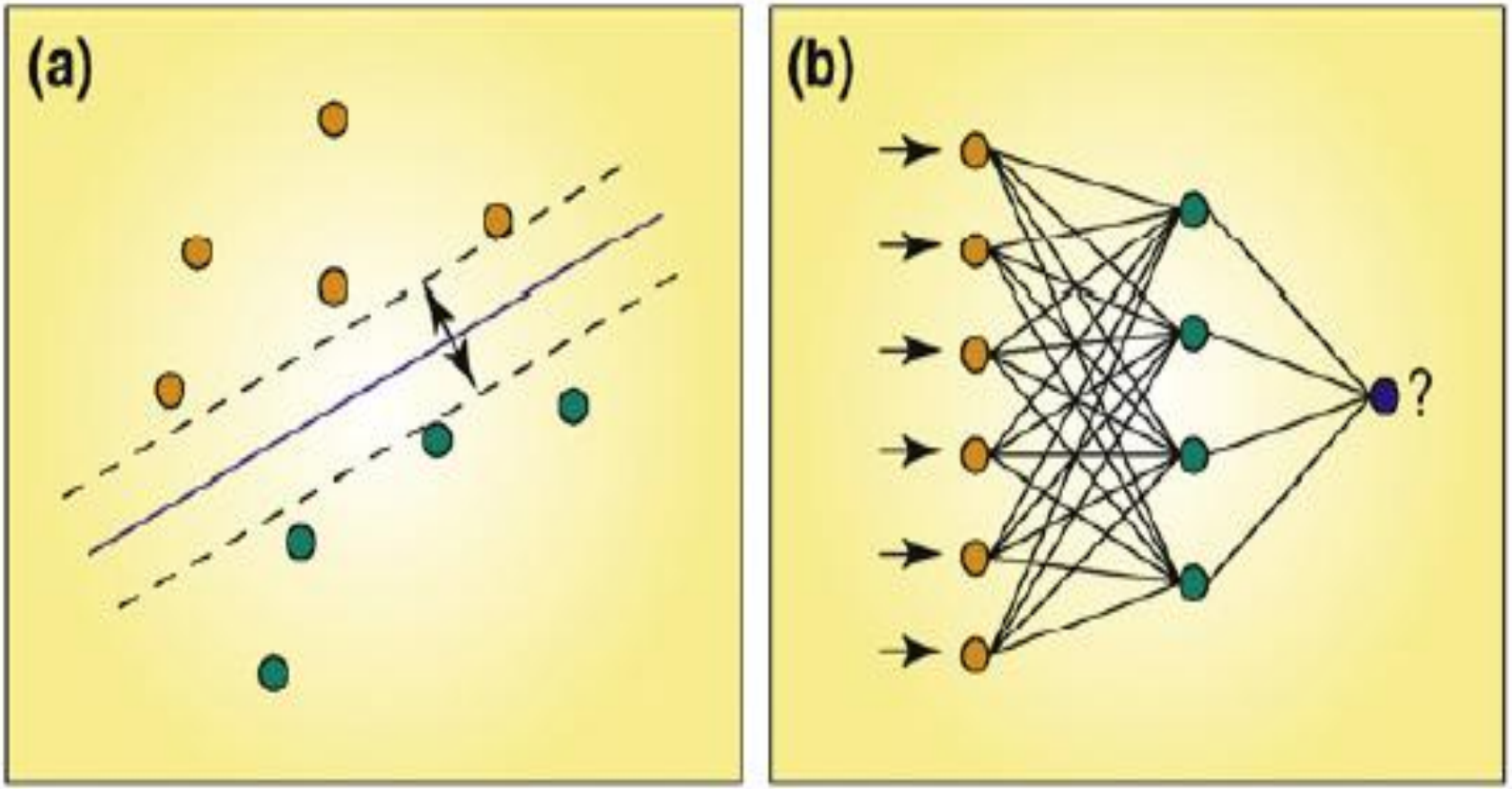
Data Driven Machine Learning Approach



Input: sequence features
Output: function category

Training: Build a classifier
Test: Test the model

Key idea: Learn from known data and Generalize to unseen data


(a)
(b)
?

ProtFun

- Assign eukaryotic query sequence to: (i) one of 14 GO categories; (ii) one of 12 “cellular roles” of the Riley scheme; (iii) an Enzyme Commission class (if an enzyme).
- Input featues: hydrophobicity, post-translational modification, subcellular location signals, secondary structure composition, putative transmembrane parts

Jensen et al., Bioinformatics, 2003

Work Flow of Sequence-Based Function Prediction Methods

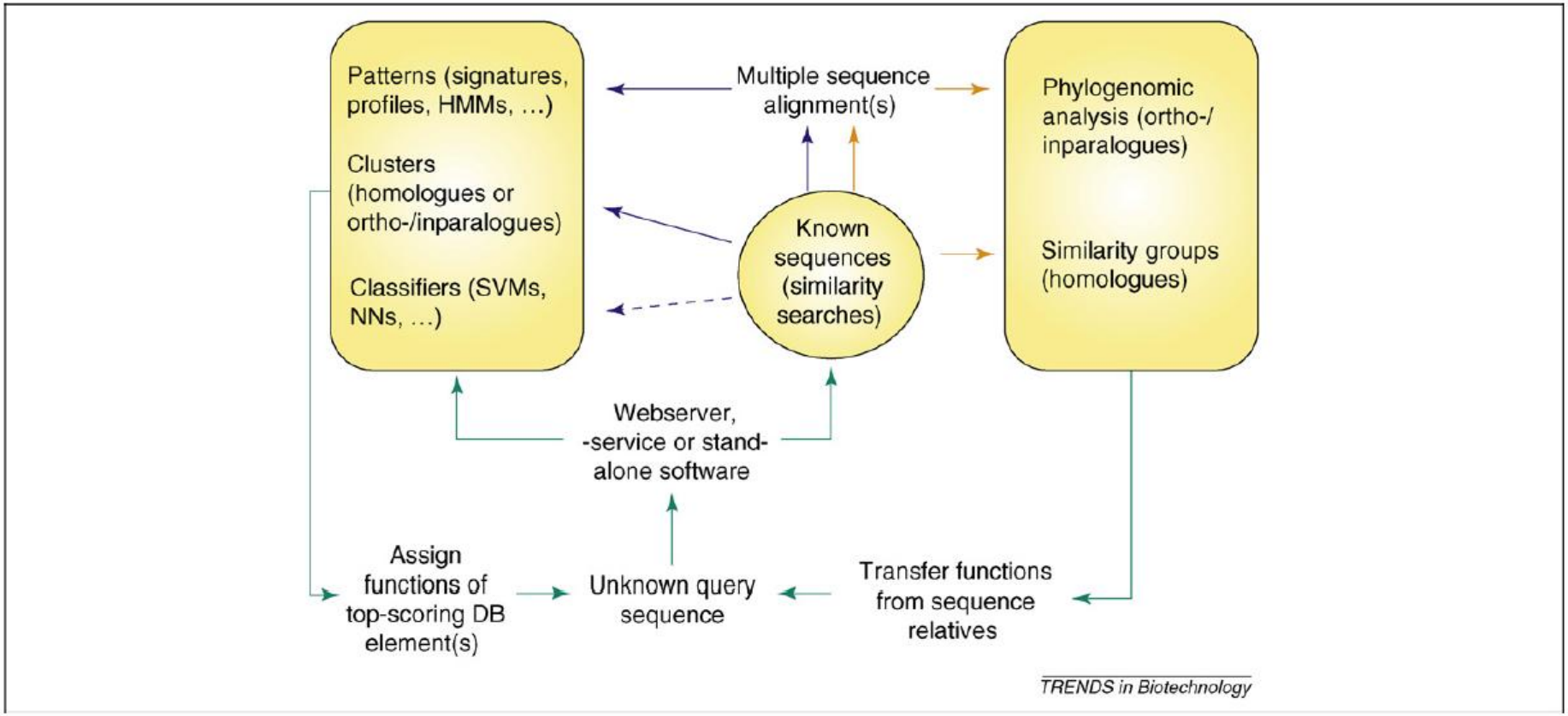


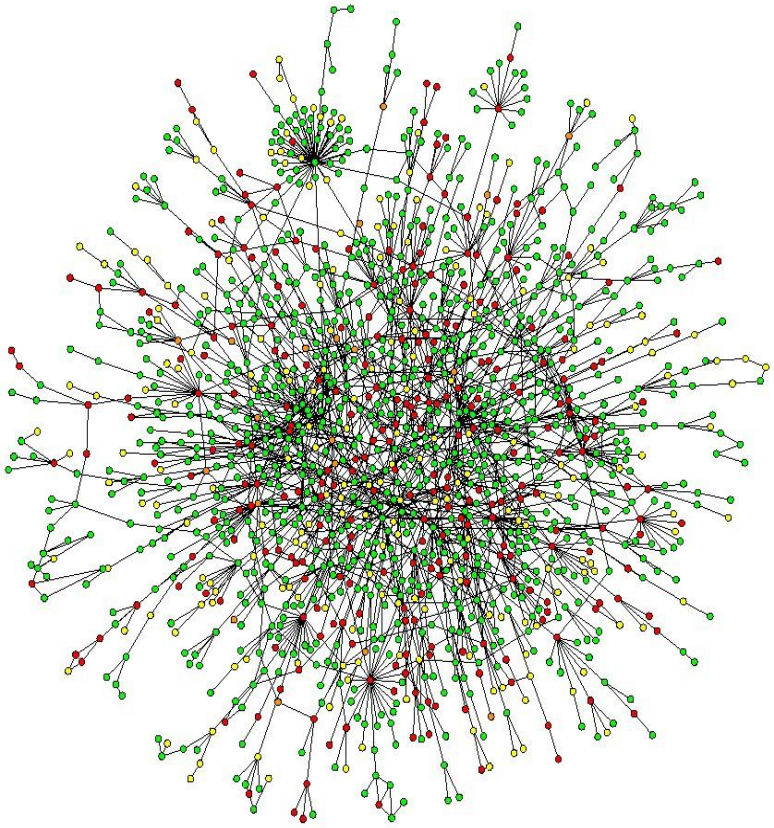
Figure 2. Current approaches in molecular function prediction. The superposed workflows of the different discussed approaches highlight their common dependency on (sufficient and high-quality annotated) sequence data. Pattern-based, clustering and ML methods build pre-computed databases (shown by blue arrows, the dashed arrow indicates no sequence comparisons are conducted), which are later scanned against individual queries. Similarity group and phylogenomic resources perform one-off similarity searches (orange arrows). Green arrows represent parts of the workflow shared by all methods.

Method Selection

- A sensible approach to molecular function prediction ‘when BLAST fails’ is to try finding consensus between these methods.
- With respect to this, the development and maintenance of a meta-server for sequence-based function prediction, querying several of the discussed resources would be incredibly beneficial to the community.

Network-Based Approach

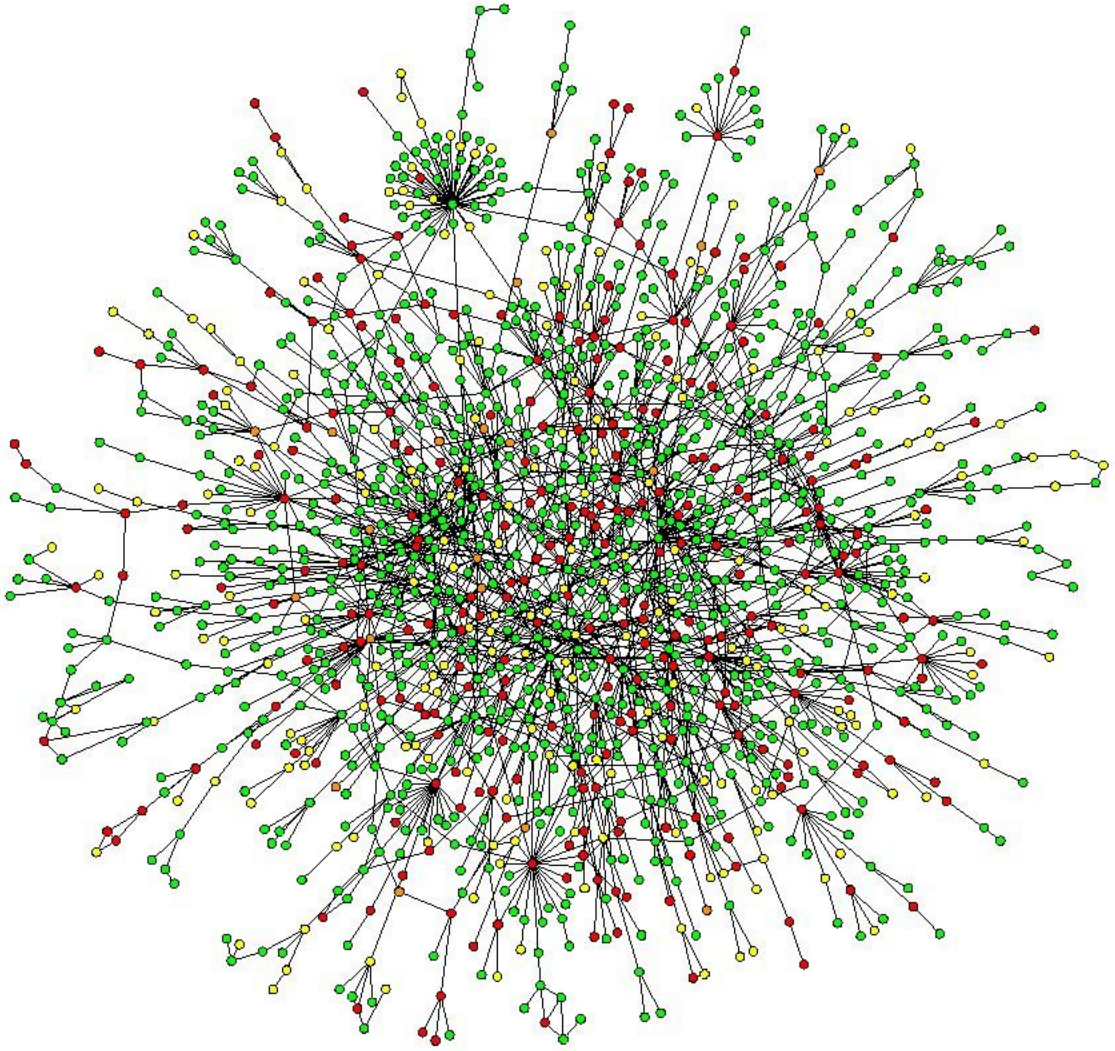
- In theory, once a network (*protein-protein interaction network*) is constructed, further methods can propagate annotations to uncharacterized members.
- Functional Linkage Networks – integration of different types of data



Function Association in Networks

The functional associations detected by these methods can be tight or loose, ranging from direct physical interaction, as opposed to co-occurrence in the same biochemical pathway, to merely being involved in the same cellular process or influencing the same phenotypical trait.

Network-Based Annotation Transfer



Assumptions and Observations

- Closer that two nodes are in the network, the more functionally similar they will be in terms of cellular pathway or process as opposed to molecular function
- Non-neighboring proteins with similar network connectivity patterns can have similar molecular functions (as members of the same complex)

Local Neighbor Methods

- Early network-based annotation methods simply inherited the function(s) most commonly observed among the direct neighbors of an uncharacterized node (“majority rule”)
- Performances increases when wider local neighborhood is taken into account and only statistically enriched functions are transferred
- The predictive power of local methods is still limited, most obviously when interaction and/or annotation are sparse

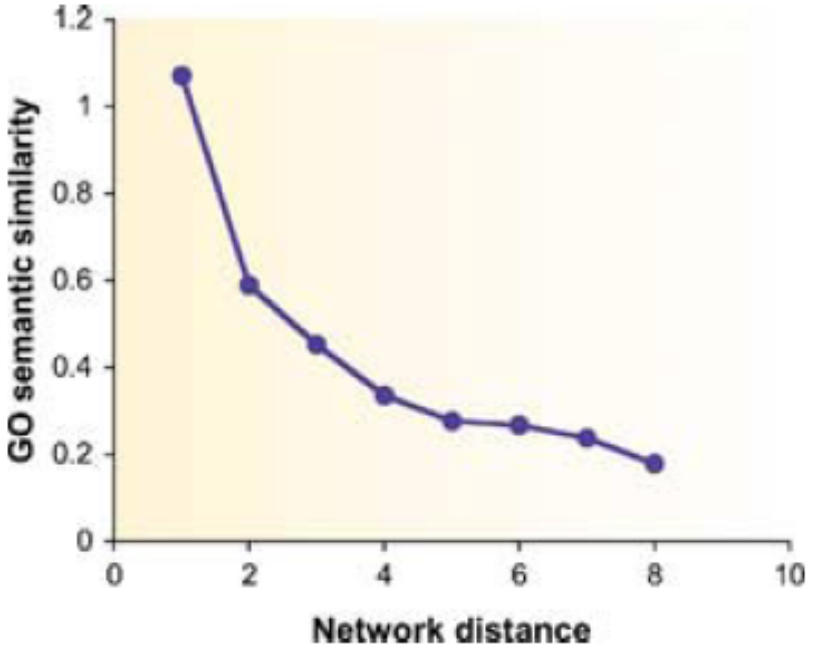


Figure 3 Correlation between protein functional distance and network distance. *X*-axis: distance in the network. *Y*-axis: average functional similarity of protein pairs that lie at the specified distance. The functional similarity of two proteins is measured using the semantic similarity of their GO categories (Lord *et al*, 2003).

Sharan et al., Molecular Systems Biology, 2007

Global Topology Methods

- Take global network topology into account
- Use graph-theory and different types of iterative stochastic approximation to identify clusters / modules
- Global methods can significantly boost performance

Module-Assisted Approaches

- Subdivide the entire network into smaller, coherent modules, within which functions are then transferred independently
- Transfer algorithm is usually one of the local ones described above
- Module detection highly influences the final annotation accuracy.
- MCODE – molecular complex detection

MCODE Method

- Stands for Molecular Complex Detection – detects densely connected regions that may represents molecular complexes
- Using the Biomolecular Interaction Network Database (BIND), collected 15,143 yeast protein-protein interactions among 4,825 proteins (about 75% of the yeast genome)
- Identify vertices with high degree (connections)
- Takes as input the vertex weighted graph, seeds a complex with the highest degree / weighted vertex and recursively moves outward from the seed vertex, including vertices in the complex whose degree / weight in the sub-graph is above a given threshold, which is a given percentage away from the weight of the seed vertex.

Bader et al., BMC Bioinformatics, 2003

Direct Neighbors VS Module Assisted

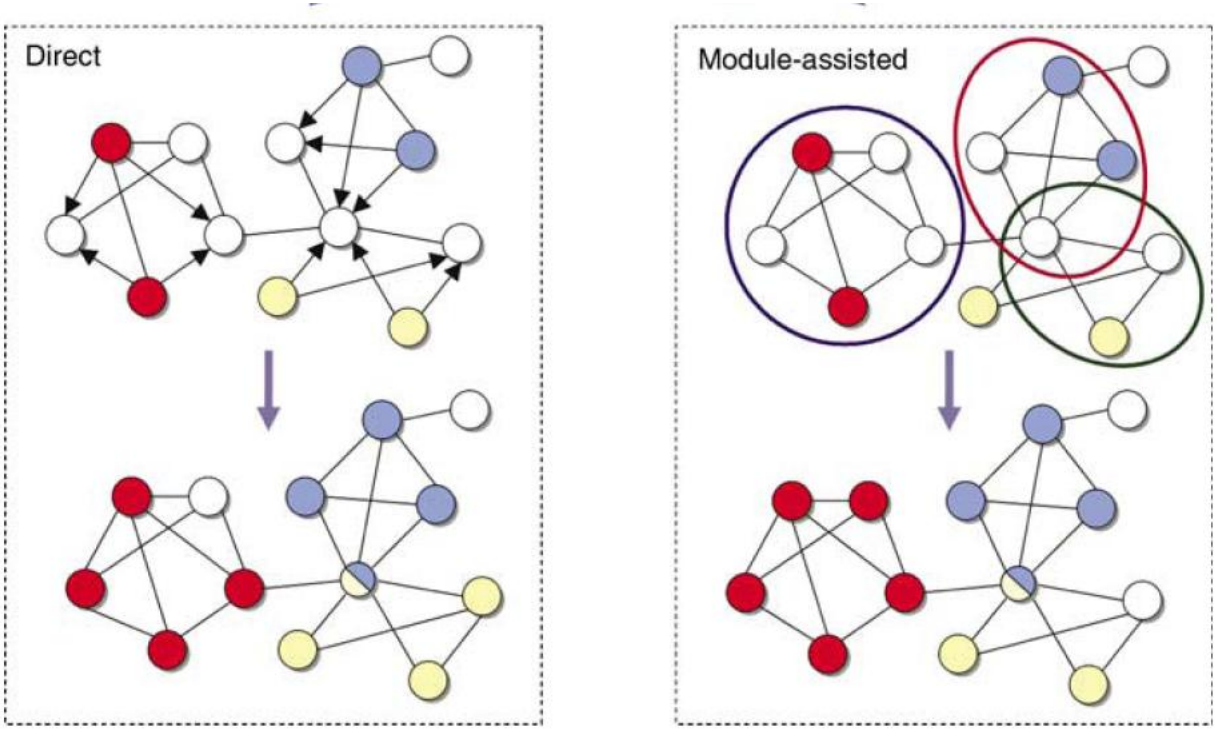


Figure 2 Direct versus module-assisted approaches for functional annotation. The scheme shows a network in which the functions of some proteins are known (top), where each function is indicated by a different color. Unannotated proteins are in white. In the direct methods (left), these proteins are assigned a color that is unusually prevalent among their neighbors. The direction of the edges indicates the influence of the annotated proteins on the unannotated ones. In the module-assisted methods (right), modules are first identified based on their density. Then, within each module, unannotated proteins are assigned a function that is unusually prevalent in the module. In both methods, proteins may be assigned with several functions.

Functional Linkage Networks

- Motivation: Individual PPI datasets are sparse and unreliable
- Integration of different types of interaction data into FLNs is a promising approach
- FLN edge weights are integrated interaction probability values (e.g. vote).

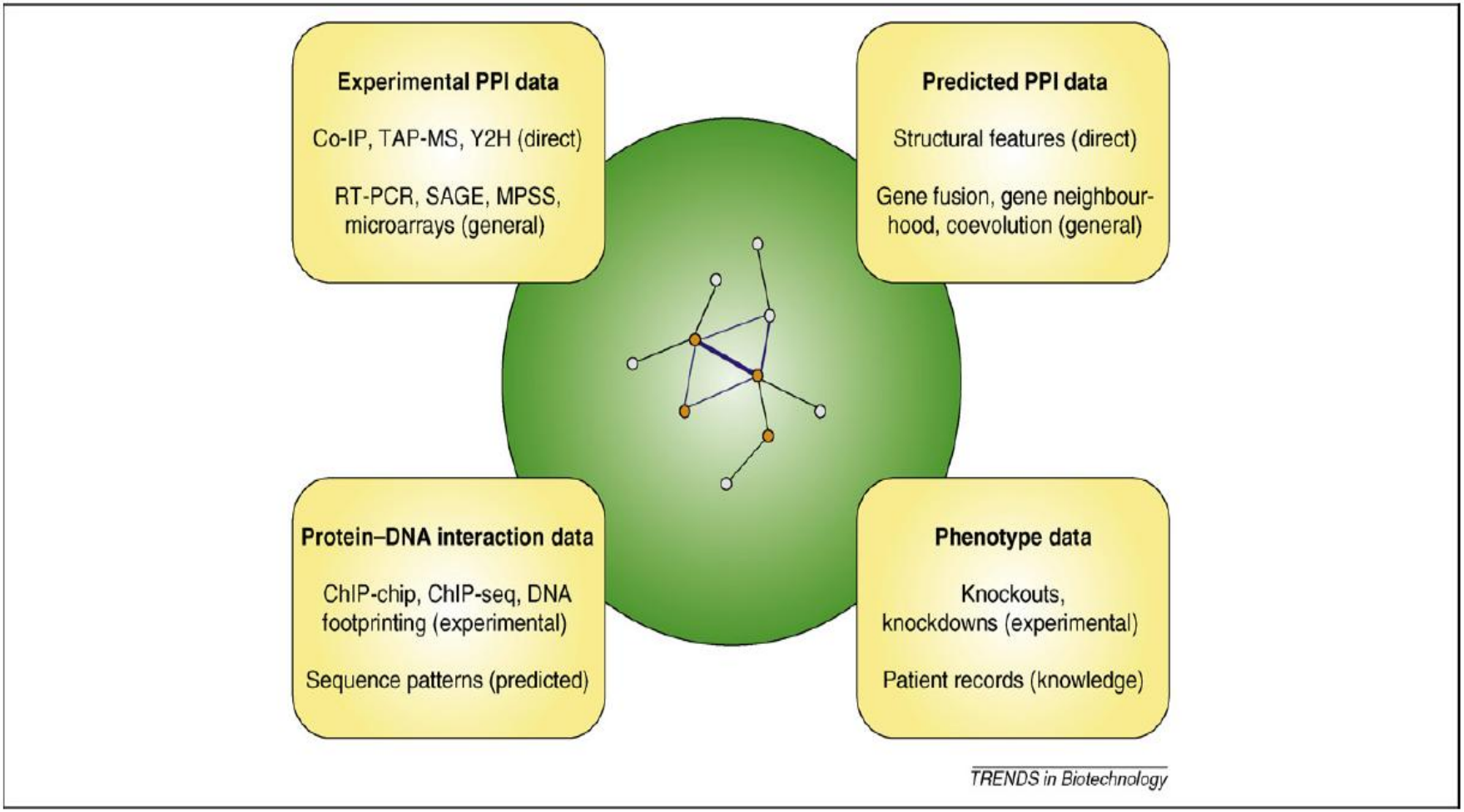


Figure 3. The concept of functional linkage networks (FLNs). FLNs can integrate multiple datasets of experimentally derived and/or predicted interactions, each containing direct (physical) or general (both physical and indirect) interactions. Each interaction (edges) is associated to an integrated probability value (edge thickness). Topological and graph-theoretic measures applied to PPI networks can equally be applied to FLNs (the blue edges here, for example, denote a clique of nodes). As with PPIs, annotations can potentially be transferred from functionally characterized (orange) to uncharacterized (grey) nodes.

GeneFAS

- A function network based on protein interaction data, microarray gene expression data, protein complex data, protein sequence data, and protein localization data
- Bayesian Function Inference of the probability that two genes have the same function (S: same function, Mr: gene expression coefficient)

$$p(S \mid M_r) = \frac{p(M_r \mid S)p(S)}{p(M_r)}$$

Chen and Xu, Nucleic Acids Research, 2004

Tools and Resources

Table 1. Resources used in protein function annotation, in order of appearance throughout the text

Method	Resource[a]	Server	Seq. queries[b]	Comments
Similarity group methods	GOtcha [9]	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php	✓	Target DB: 16 genomes
	PFP [10]	http://dragon.bio.purdue.edu/pfp/	✓	Target DB: 18 genomes
	GOsling [11]	https://www.sapac.edu.au/gosling/	✓	Target DB: UniProtKB GO sequences (2006)
Phylogenomics	SIFTER [15]	http://sifter.berkeley.edu/	n/a	Download only (uses Pfam)
	AFAWE [17]	http://bioinfo.mpiz-koeln.mpg.de/afawe/	✓	Meta-tool including SIFTER
		http://www.myexperiment.org/workflows/95/	n/a	AFAWE workflow (uses RefSeq)
Pattern/profile methods	InterProScan [20]	http://www.ebi.ac.uk/tools/interproscan/	✓	DB composition: meta-tool, queries 10 pattern-based resources (see below)
	PROSITE [21]	http://www.expasy.ch/prosite/	✓	DB composition: >1500 patterns/profiles
	PRINTS [22]	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	-	DB composition: >1900 fingerprints
	Pfam [16]	http://pfam.sanger.ac.uk/	✓	DB composition: >10 000 domain families
	SUPERFAMILY [23]	http://supfam.cs.bris.ac.uk/superfamily/	✓	DB composition: SCOP domains in 62 genomes
	PRODOM [24]	http://prodom.prabi.fr/prodom/current/html/home.php	✓	DB composition: >730 000 domain families
	SMART [25]	http://smart.embl-heidelberg.de/	✓	DB composition: >500 domain families
	Gene3D [26]	http://gene3d.biochem.ucl.ac.uk/gene3d/	✓	DB composition: CATH domains in 527 genomes
	PANTHER [27]	http://www.pantherdb.org/	✓	DB composition: >24 000 protein families
	PIRSF [28]	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	✓	DB composition: >4500 protein families
	TIGRFAMs [29]	http://www.tigr.org/TIGRFAMs/	✓	DB composition: >3600 protein families
	SCOP [30]	http://scop.mrc-lmb.cam.ac.uk/scop/	–	DB composition: >1700 domain families
	CATH [31]	http://www.cathdb.info/	✓	DB composition: >2000 domain families
	CatFam [35]	http://www.bhsai.org/downloads/catfam.tar.gz	n/a	DB composition: not stated, download only
	EFICAz [36]	http://cssb.biology.gatech.edu/skolnick/webservice/EFICAz2/index.html	✓	DB composition: 2354 enzyme families
	PRIAM [37]	http://bioinfo.genotoul.fr/priam/REL_JUL06/index_jul06.html	✓	DB composition: 2368 enzyme families

Tools and Resources

Clustering approaches	**Homologues**			
	ProtoNet [38]	http://www.protonet.cs.huji.ac.il/	✓	Clustered DB: current UniProtKB
	CluSTr [41]	http://www.ebi.ac.uk/clustr/	-	Clustered DB: current UniProtKB and IPI
	Ortho- and inparalogues			
	eggNOG [43]	http://eggnog.embl.de/	✓	Clustered DB: 373 genomes
	COGs [46]	http://www.ncbi.nlm.nih.gov/COG/	✓	Clustered DB: 66 genomes
	KOGs [46]	http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi	✓	Clustered DB: 7 genomes
	InParanoid [44]	http://inparanoid.sbc.su.se/cgi-bin/index.cgi	✓	Clustered DB: 35 genomes
	MultiParanoid [47]	http://multiparanoid.sbc.su.se/index.html	-	Clustered DB: uses InParanoid, download only
	OrthoMCL [45]	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi	✓	Clustered DB: 87 genomes
ML methods	ProtFun [50]	http://www.cbs.dtu.dk/services/ProtFun/	✓	Functional categories: 32 (14 GO terms, 1st l. ECs, etc.)
	SVM-Prot [51]	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi	✓	Functional categories: 130 (all 2nd l. ECs and TCs, etc.)
	ffPred [52]	http://bioinf.cs.ucl.ac.uk/ffpred/	✓	Functional categories: 197 (197 GO terms)
	EzyPred [53]	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/	✓	Functional categories: 49 (49 2nd l. ECs)
Network-based approaches	**Network module detection**			
	MCODE [76]	http://baderlab.org/Software/MCODE	n/a	Cytoscape plugin and source code
	MCL [48]	http://www.micans.org/mcl/	n/a	Explanation and source code
	Cytoscape	http://chianti.ucsd.edu/cyto_web/plugins/pluginjardownload.php?id=175	n/a	Cytoscape plugin using MCL
		http://www.cytoscape.org/	n/a	Network visualization software
	Functional linkage networks			
	STRING [79]	http://string.embl.de/	✓	DB of PPIs in 630 genomes
	VisANT [80]	http://visant.bu.edu/	-	DB of PPIs in 108 genomes
	VIRGO [83]	http://whipple.cs.vt.edu/virgo/welcome.cgi	n/a	Gene expression data as input

Abbreviations: DB, database; MCL, Markov Clustering; MCODE, Molecular Complex Detection; ML, machine learning; n/a, not available; PPI, protein–protein interaction.
[a]This covers actively maintained resources but is not guaranteed to be exhaustive. Some are not directly aimed at function prediction; the main text explains how they contribute to it. All servers were tested, database statistics refer to the current releases (11/2008).
[b]Indicates whether a server or database can be queried directly with a sequence. 'n/a' here means 'not applicable' (to the method, i.e. sequence queries would make no sense), whereas the dash (-) means it could (or should) have this option but does not.

Conclusion

- The combination of sequence- and network-based function transfer approaches is promising
- Complementary nature: while sequence (and structural) similarity can provide a safe basis for molecular function transfer, interactions hints at the pathways and the processes in which uncharacterized proteins participate
- We didn’t talk about structure-based function prediction

Reading Assignment (select one paper to review)

- I. Friedberg. Automated function prediction – the genomic challenge. Brief. In Bioinformatics, 2006
- Sharan, R. et al. (2007) Network-based prediction of protein function. Mol. Syst. Biol. 3, 88
- Gherardini, P.F. and Helmer-Citterich, M. (2008) Structure-based function prediction: approaches and applications. Brief. Funct. Genomic. Proteomic. 7, 291–302