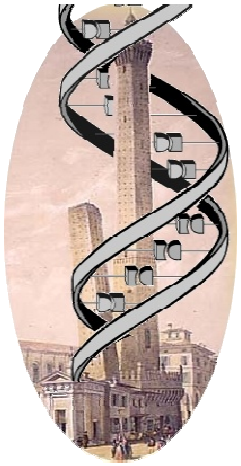
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Large scale annotation of proteins with labelling methods

Rita Casadio

BIOCOMPUTING GROUP
University of Bologna, Italy

Bologna Computational Biology Network

WetLab | CompuLab | SME Small Medium Eterprises

IRBBC

The "omic" revolution

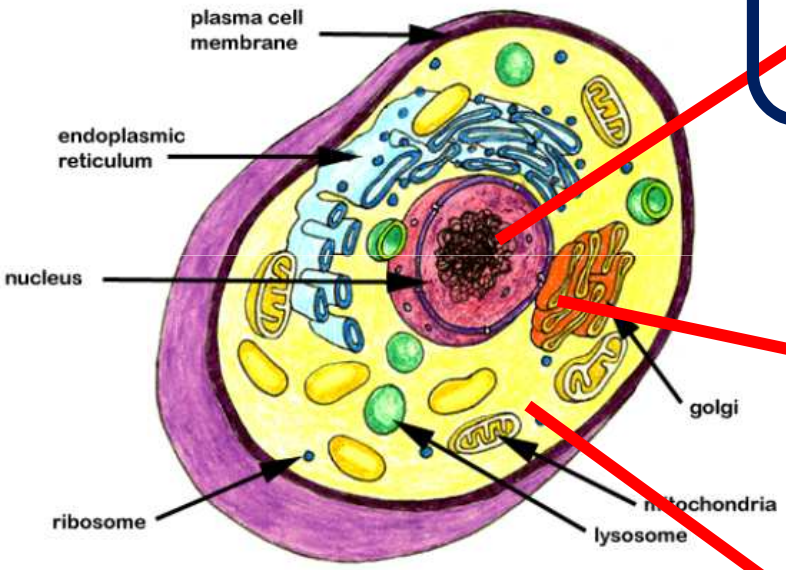
The analysis of the components of a living organism in its entirety

GA EVANS, Nature Biotechnology 2000, 18, 127

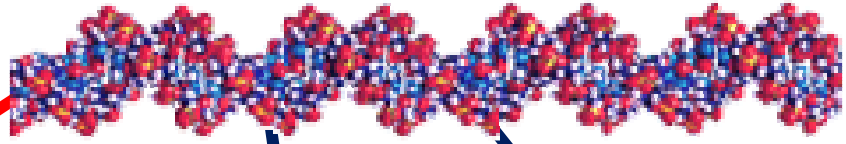
The ingredients of biological complexity at the cell level

From genes to proteins and their interaction

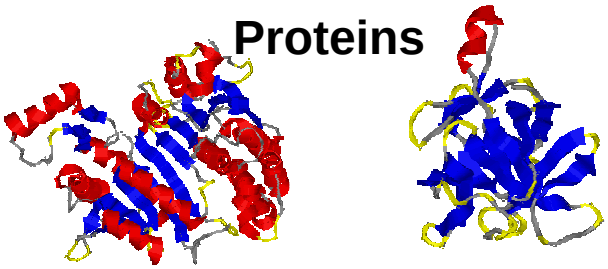
The cell



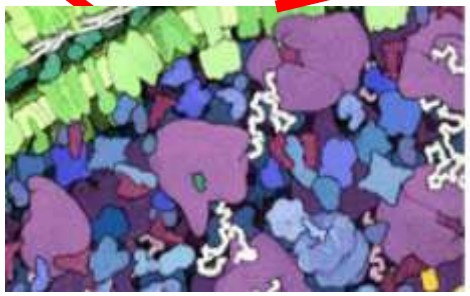
DNA & Genes



Proteins



Macromolecular crawding

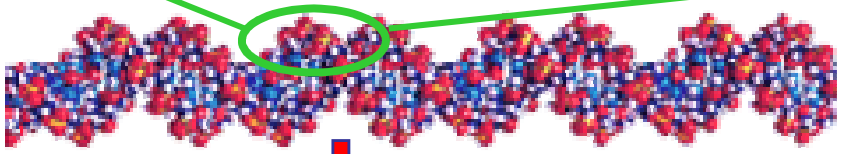


Molecular Interactions and Functions

Deep sequencing (NGS) techniques and the Protein Universe

A,T,C,G

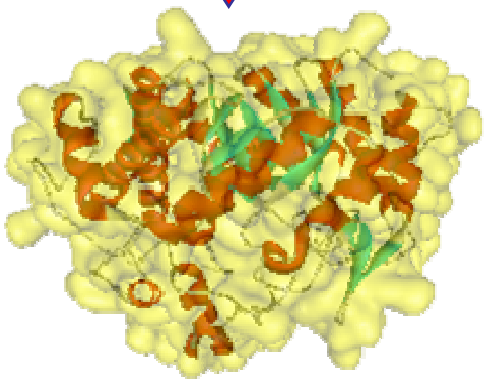
cctgttgatggcgacagggactgtatgctgatctatgctgatgcatgcatgctgactactgatgtgggggctat



From genes…

```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
```

A,C,D,E,F,G,H,I,K,L
M,N,P,Q,R,S,T,V,Y,W



…to Proteins

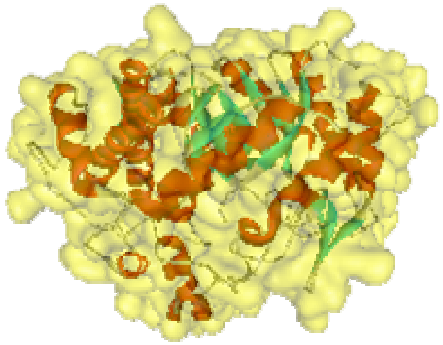
The Data Bases of Biological Sequences and Structures

GenBank: 135,440,990 sequences
126,551,501,141 nucleotides

```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

UniProt/Tremble:
16,504,022 sequences
5,357,406,695 residues

UniProt/SwissProt:
531,473 sequences
188,463,640 residues



PDB: 73,373 structures
membrane proteins <2%

≅43 HGE!

Update:
September 2011

Dealing with genomic data….

Scott D. Kahn
Science 331, 728 (2011)

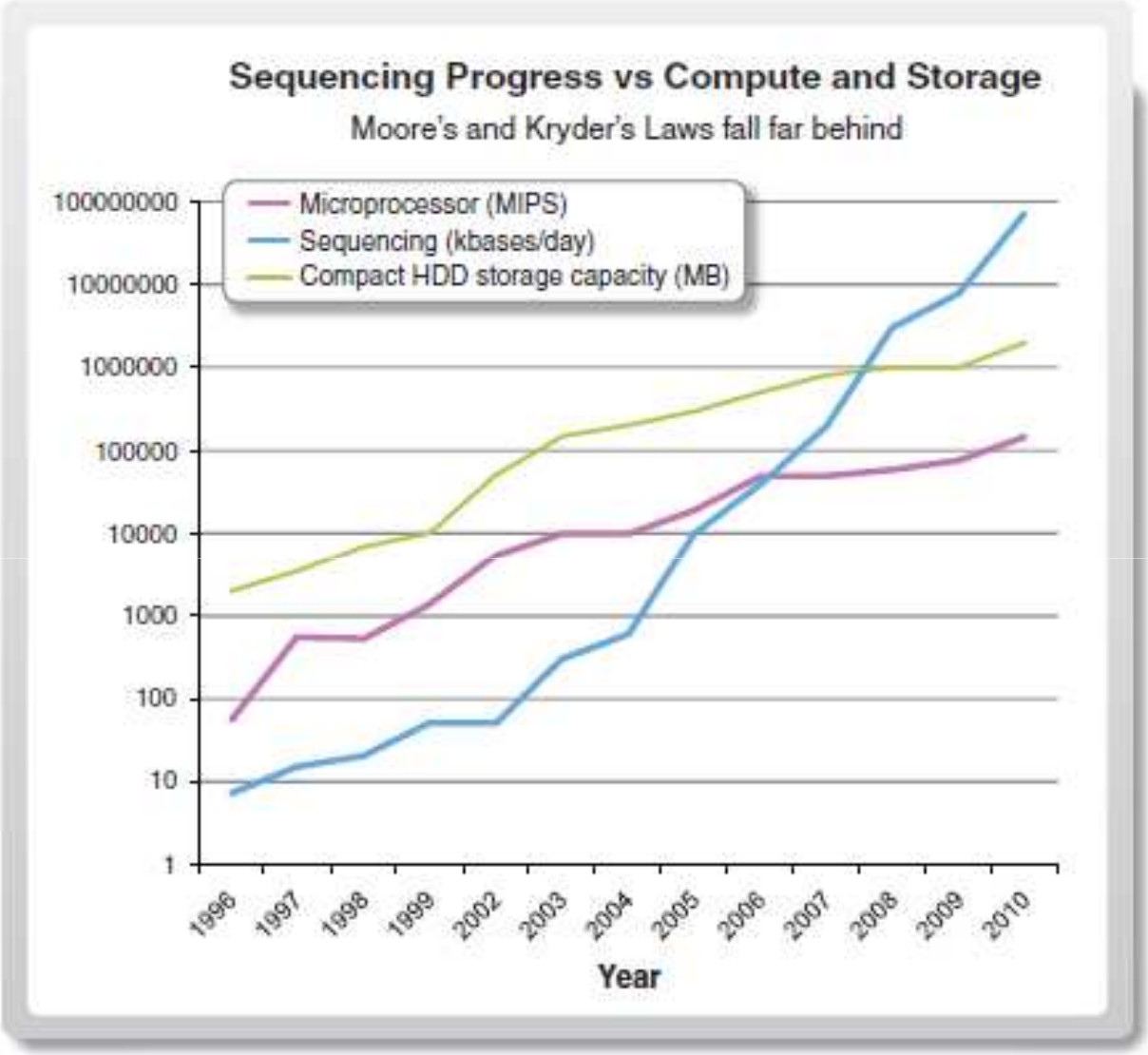


Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Genomic data and the problem of protein validation

Data production→Data analysis

DNA sequencing →gene recognition → protein translation

Experiments to validate protein structure and function produce data in a time << than that required to deposit putative protein sequences into data bases

Release 2011_08 of 27-Jul-11 of UniProtKB/Swiss-Prot contains 531,473 sequence entries

Protein existence (PE):	entries	
1: Evidence at protein level	73667	13.9%
2: Evidence at transcript level	68884	13%
3: Inferred from homology	372619	70.1%
4: Predicted	14435	2.7%
5: Uncertain	1868	0.4%

Release 2011_08 of 27-Jul-2011 of UniProtKB/TrEMBL contains 16,504,022 sequence entries

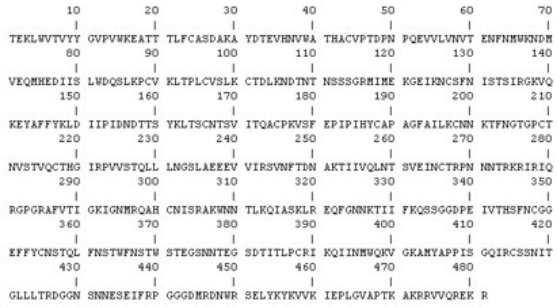
Protein existence (PE):	entries	
1: Evidence at protein level	14122	0.09%
2: Evidence at transcript level	530839	3.22%
3: Inferred from homology	3729175	22.60%
4: Predicted	12229886	74.10%
5: Uncertain	0	0.00%

Only 4.1 % sequences has evidence at the protein and trascript level and only 0.4 % proteins have structures in the Protein Data Bank.

A "BIG" problem of the "omic era" after genome sequencing:

code for

Genomes (~3000)



Protein sequences (~17 millions)

that are endowed with

Protein structures and functions

Protein sequence Annotation:

to endow with structural and functional features protein sequences after gene translation

Transfer of annotation *in silico* by homology search

```
ADH1_SULSO  ----------MRAVRLVEIGKP--LSLQEIGVPKPKGPQVLIKVEAAGVCHSDVHMRQGRFGNLRIVE
ADH_CLOBE   ----------MKGFAMLGINKLG---WIEKERPVAGSYDAIVRPLAVSPCTSDIHTVFEGA-------
ADH_THEBR   ----------MKGFAMLSIGKVG---WIEKEKPAPGPFDAIVRPLAVAPCTSDIHTVFEGA-------
ADH1_SOLTU  MSTTVGQVIRCKAAVAWEAGKP--LVMEEVDVAPPQKMEVRLKILYTSLCHTDVYFWEAKG-------
ADH2_LYCES  MSTTVGQVIRCKAAVAWEAGKP--LVMEEVDVAPPQKMEVRLKILYTSLCHTDVYFWEAKG-------
ADH1_ASPFL  ----MSIPEMQWAQVAEQKGGP--LIYKQIPVPKPGPDEILVKVRYSGVCHTDLHALKGDW-------
```

Sequence comparison is performed with alignment programs

Sequence identity ≥ 30 % ⇨ 3D ?; Similar function ??

Methods for similarity searches:

BLAST, Psi-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/)

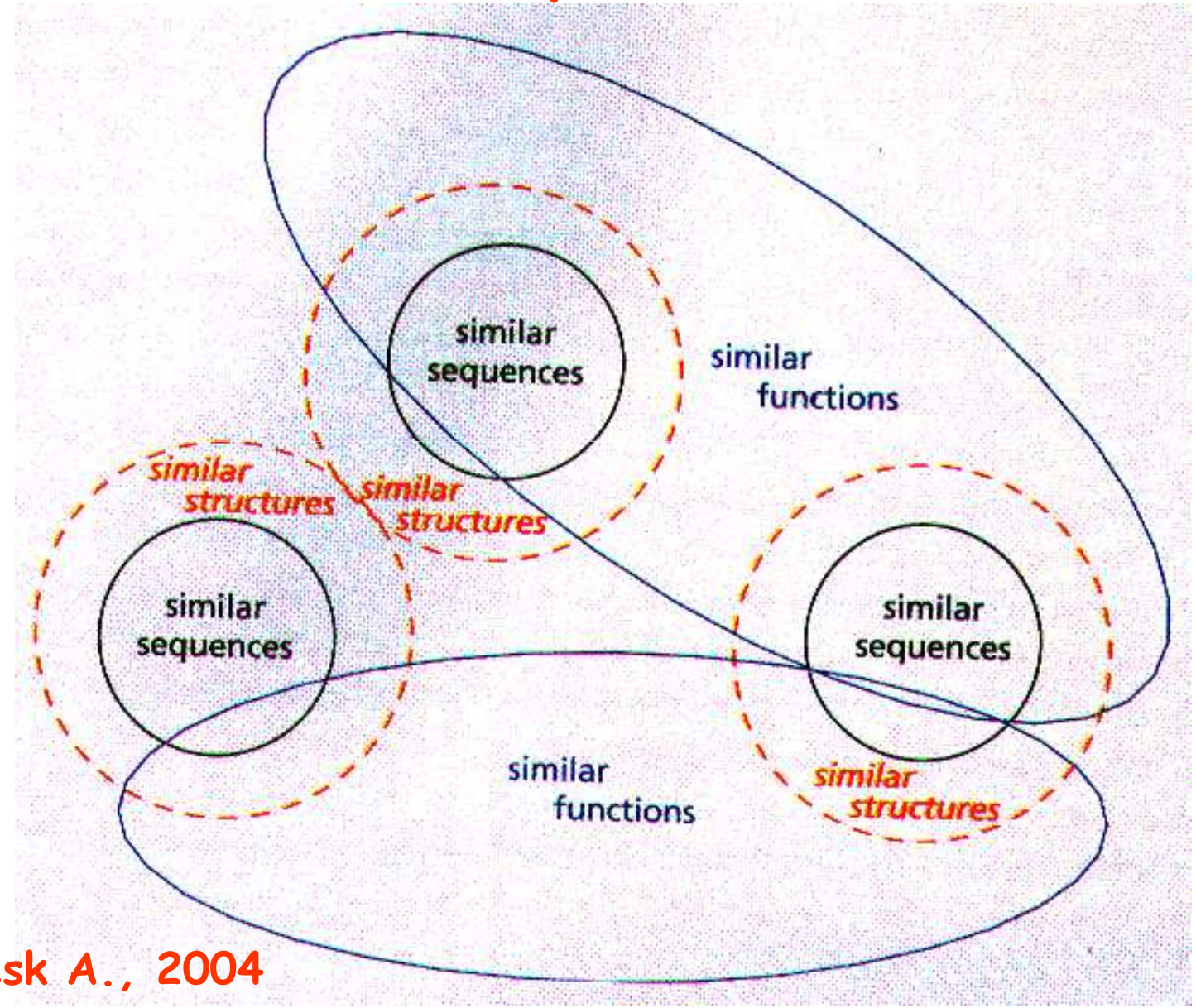
Altschul et al., (1990) J Mol Biol 215:403-410

Altschul et al., (1998) Nucleic Acids Res. 25:3389-3402

Pfam (http://pfam.wustl.edu/hmmsearch.shtml)

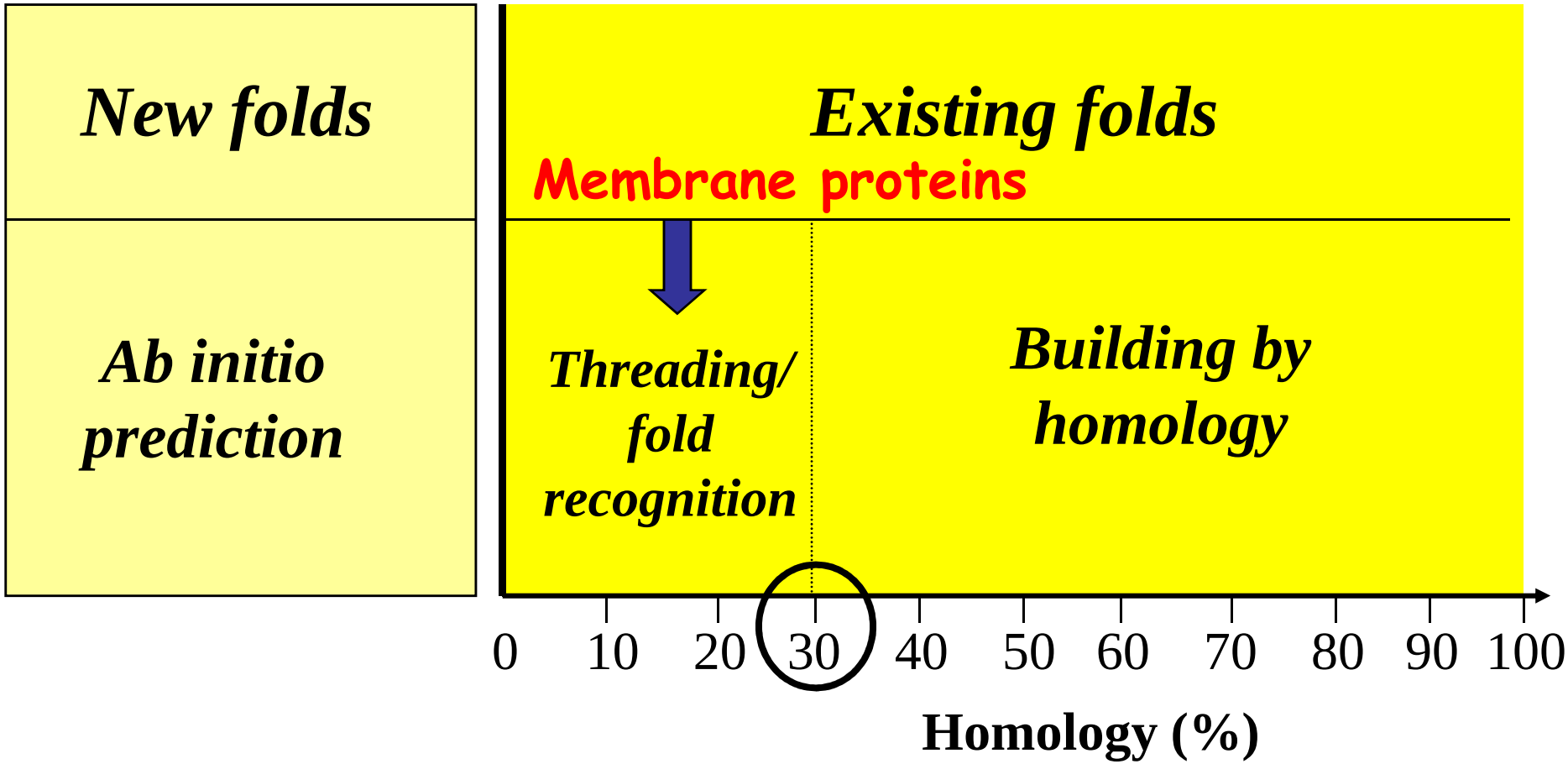
Bateman et al., (2000) Nucleic Acids Research 28:263-266

How can we infer function and structure from sequence?



Lesk A., 2004

3D structure prediction of proteins



the Gene Ontology Protein function

Open menus

GO function vocabulary:
http://www.geneontology.org/

The Ontologies

- Cellular component
- Biological process
- Molecular function

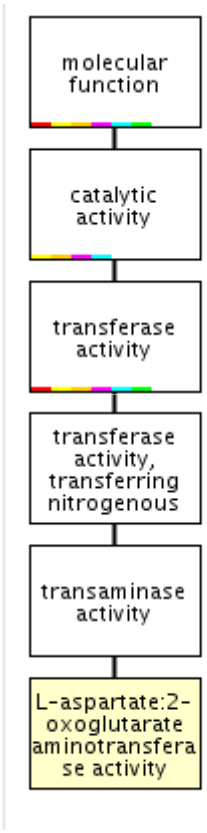
Ontology Structure

The Gene Ontology is a **controlled vocabulary**, a set of standard terms—words and phrases—used for indexing and retrieving information. In addition to defining terms, GO also defines the **relationships** between the terms, making it a **structured** vocabulary.

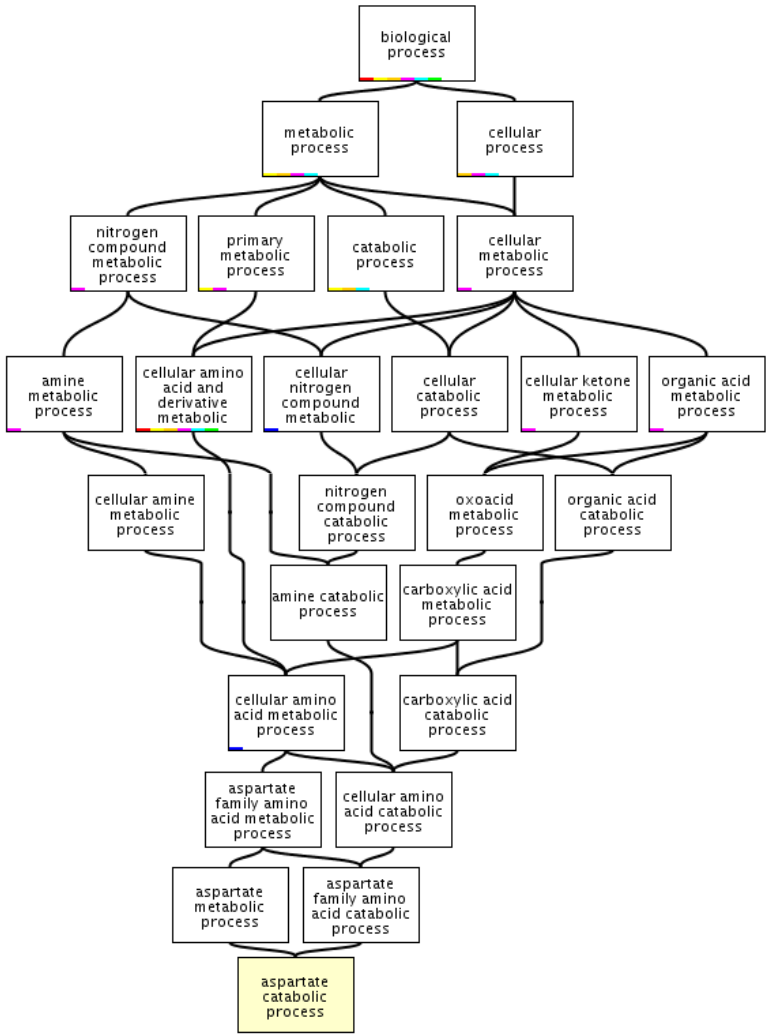
GO as a Graph

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes. The relationships used in GO are **directed**—for example, a mitochondrion *is an* organelle, but an organelle is not a mitochondrion—and the graph is **acyclic**, meaning that cycles are not allowed in the graph. The ontologies resemble a hierarchy, as child terms are more specialized and parent terms are less specialized, but unlike a hierarchy, a term may have more than one parent term. For example, the biological process term hexose biosynthetic process has two parents, hexose metabolic process and monosaccharide biosynthetic process. This is because biosynthetic process is a type of metabolic process and a hexose is a type of monosaccharide.

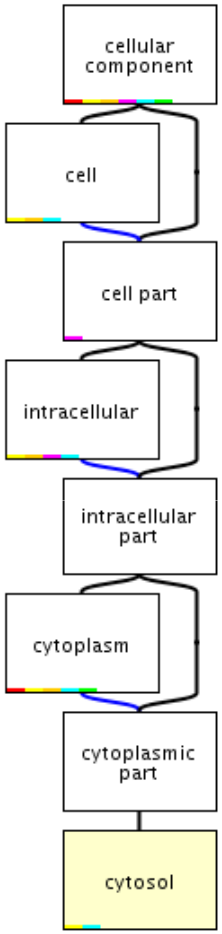
Gene Ontology classification:
E.G: The human cytoplasmic aspartate transaminase



GO:0004069



GO:0006533



GO:0005829

Validation of transfer of annotation by homology

How to infer structure and function from sequence…..

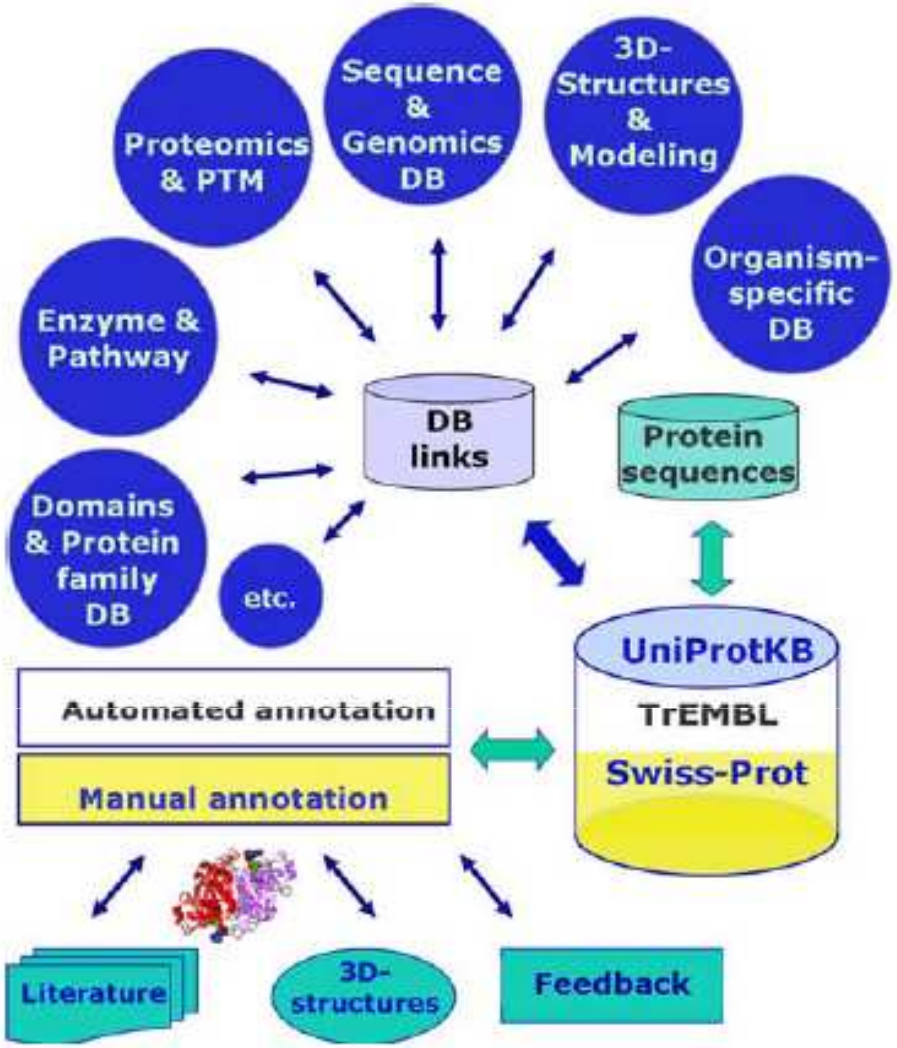
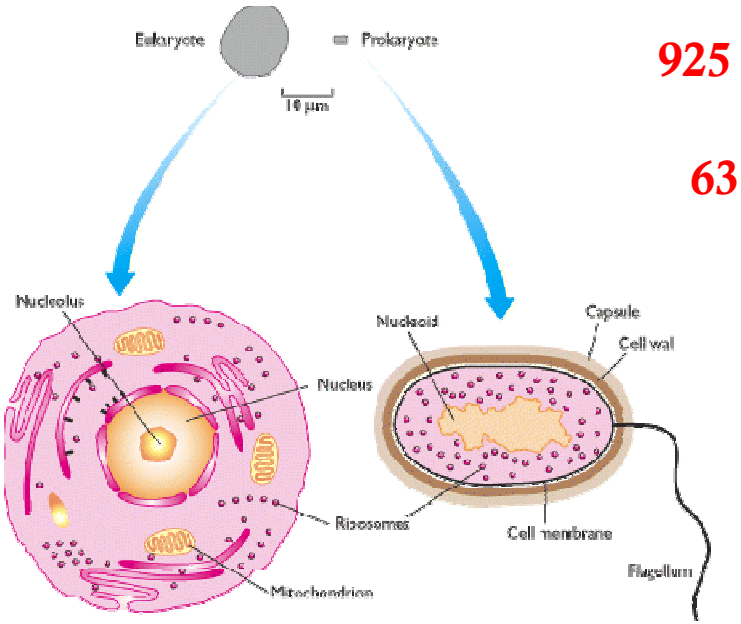


Fig. 1 UniProtKB serves as a knowledge repository and as a central hub that provides links to numerous other databases. New protein sequences are integrated in UniProtKB/TrEMBL and annotated by an automated procedure. UniProtKB/Swiss-Prot entries are manually annotated, combining carefully checked protein sequences with information from the scientific literature, protein 3D-structures, and specialised databases, together with feedback from the scientific community

UniProt KB: The largest annotation resource

Ursula Hinz • *The UniProt Consortium*
Cell. Mol. Life Sci. (2010) 67:1049–1064

988 genomes and the entire UniProt



925 Prokaryotes

63 Eukaryotes

\+

UniProt

30 %

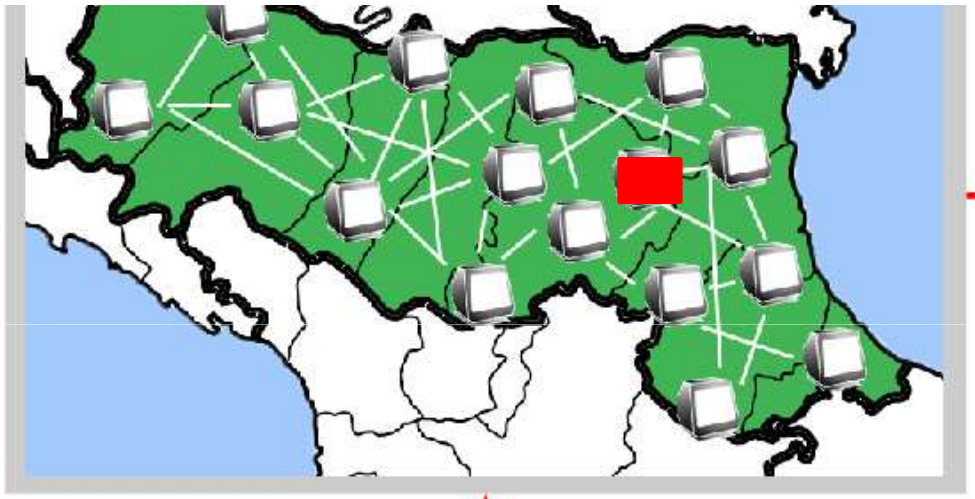
70 %

13,495,736 Protein sequences

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Large scale cross genome comparison: All-against-all BLAST alignments within GRID/COMPUT-ER

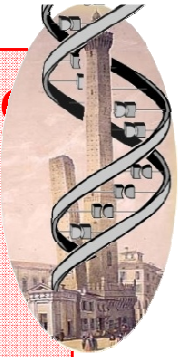
http://www.comput-er.it



13.5 M sequences

1 cpu	6,7 years
100 cpu	24 days
500 cpu	5 days

Biocomputing Group



Data warehouse

Filtering, clustering, mapping

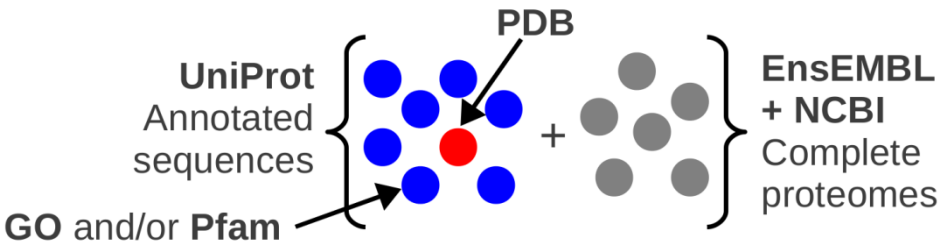
BAR+

BAR+

Bologna Annotation Resource PLUS

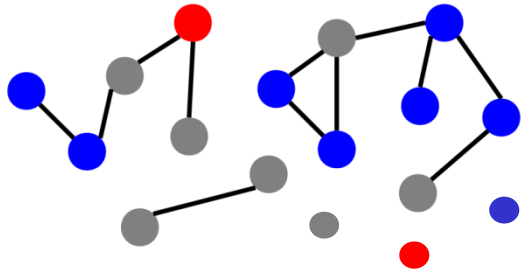
A non hierarchical clustering method for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis

Protein sequence universe



Sequence similarity network

Two sequences are linked when **SI ≥ 40% AND COV ≥ 90%** upon an all-against-all **BLAST** search

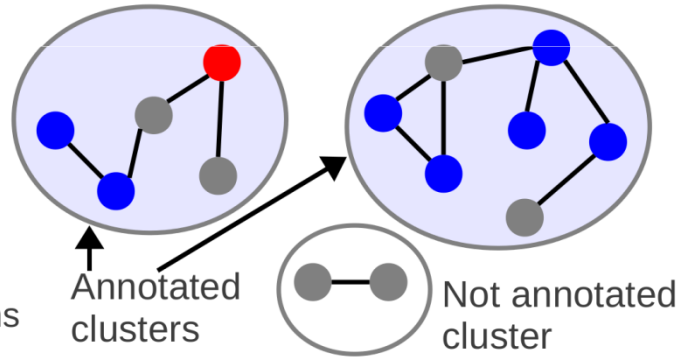


Cluster definition

Splitting of the connected components

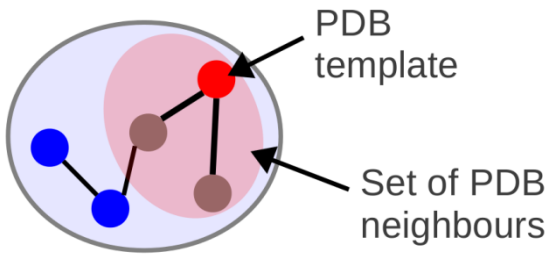
Cluster annotation

Clusters inherit UniProt annotations after a statistical validation of GO and Pfam terms



Cluster HMM

HMM profiles of the PDB templates are built from the alignment of neighbouring sequences

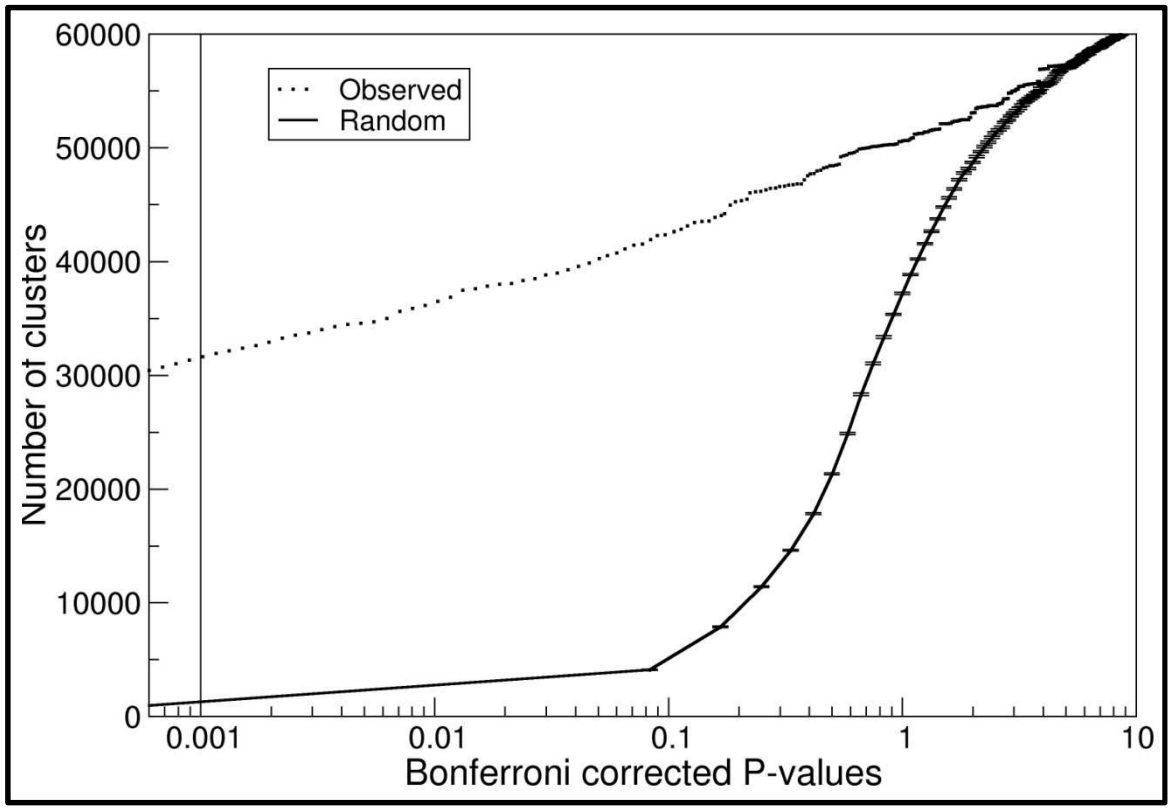


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Statistical validation of GO terms and Pfam domains

To assess whether a function is specific for a cluster:
P-value computation and bootstrapping procedure

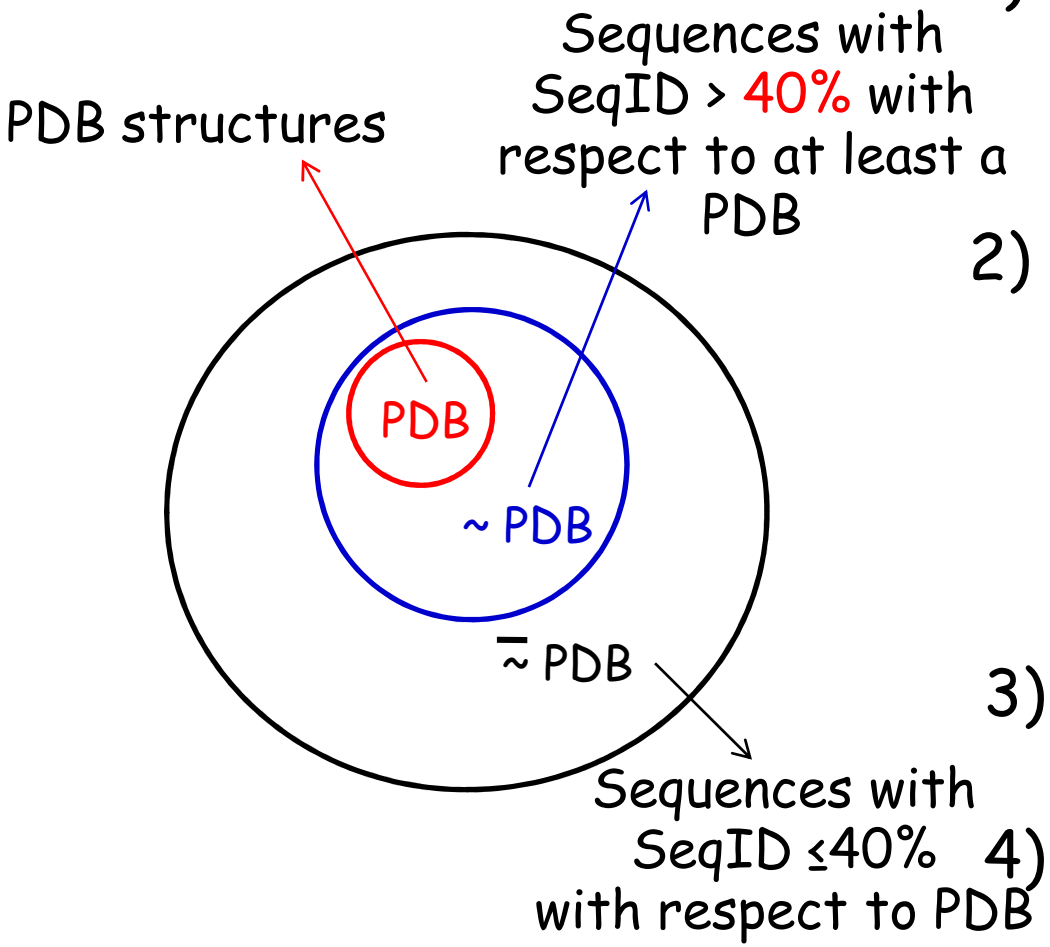
For each function within cluster: $$\text{P-value} = \sum_{i=N}^{\min(K,P)} \left(\frac{\binom{K}{i}\binom{D-K}{P-i}}{\binom{D}{P}} \right)$$



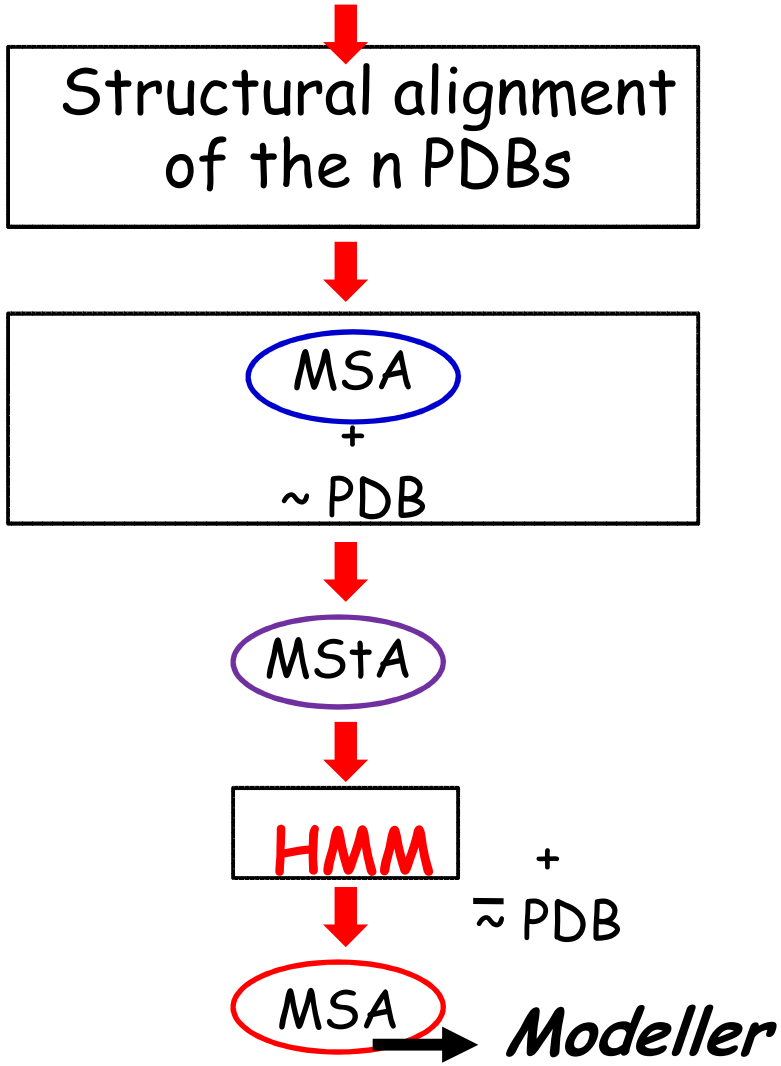
3D Modelling of the sequences in clusters with PDB structures with profile HMMs

1 PDB structure

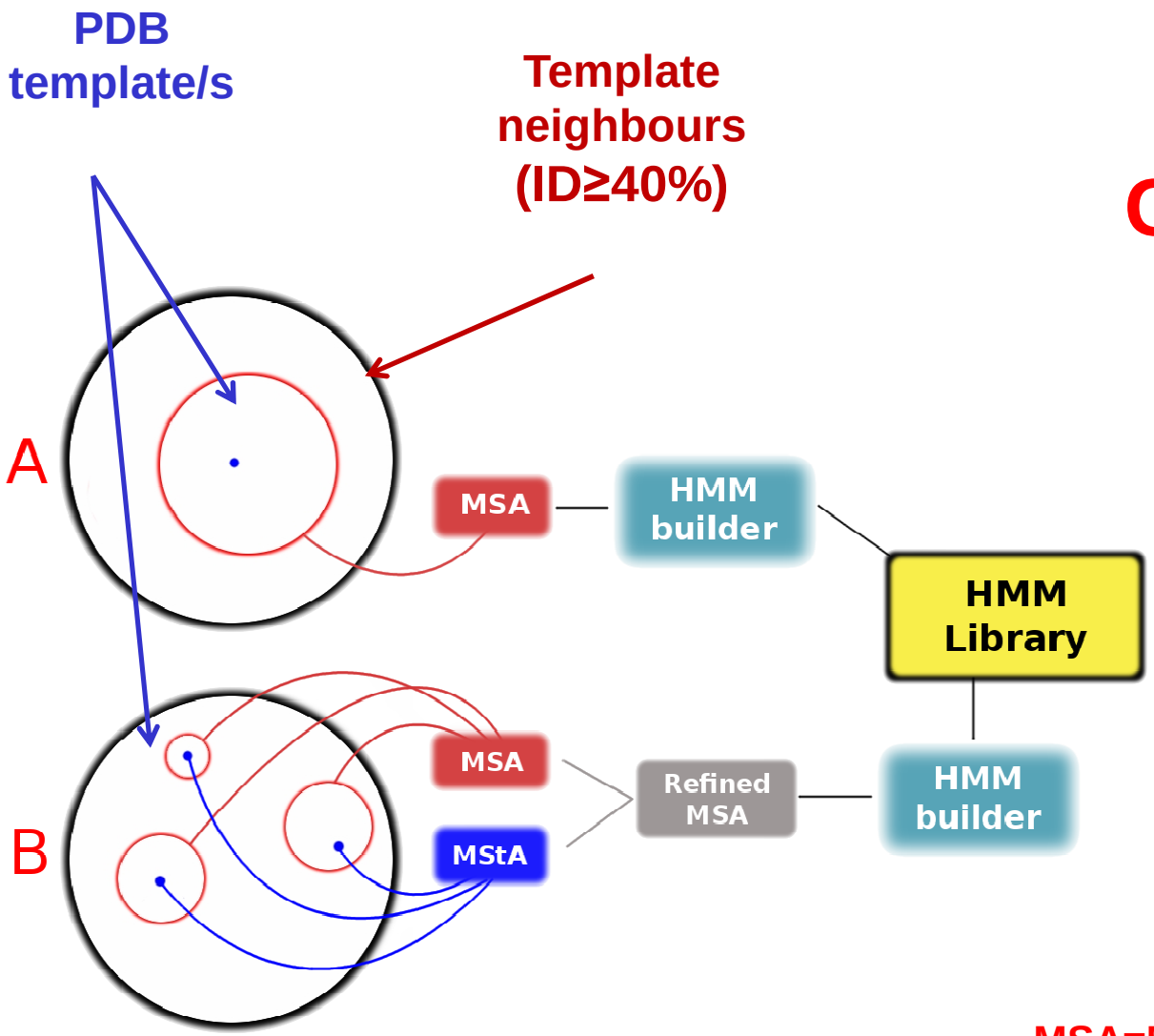
Start at point 2)



n PDB structures



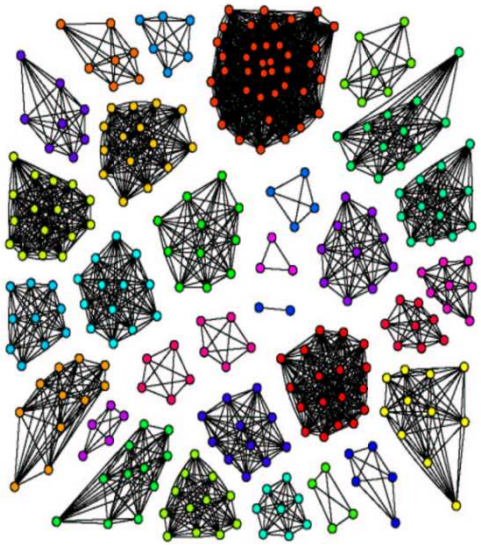
BAR+ Cluster-HMM library



Cluster HMMs: 10,858

MSA=Multiple Sequence Alignment
MStA=Multiple Structural Alignment

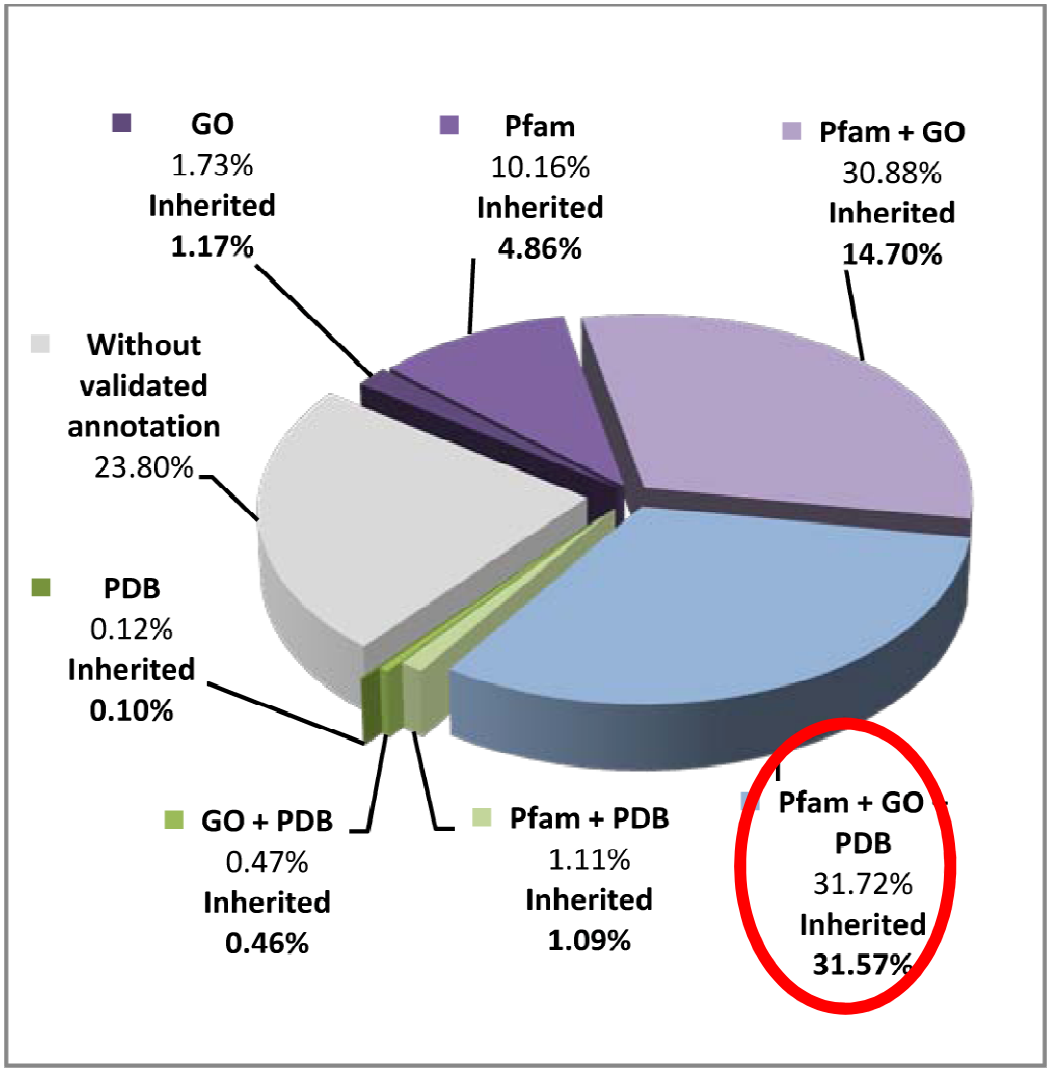
BAR+ anatomy



total sequences	13,495,736	
clusters	913,762	
sequences in clusters	9,401,223	(70%)
sequences in singletons	4,094,513	(30%)
clusters with validated GO	100,791	(11%)
sequences in clusters with a validated GO	6,092,668	(45%)
clusters with PDB	10,969	(1.2%)
sequences in clusters with PDB	3,167,555	(23%)
sequences per cluster	2/87,893	
clusters containing sequences whose standard deviation of the protein length is ≤ 5 residues		(87%)

BAR+annotation results within clusters

- 9,401,223 (70% of total) sequences are in 913,762 clusters
- 55% of the clusters include 84% of the cluster- included sequences
- 0.11% of the sequences are sufficient to annotate with the most complete annotation another 21.99% (gain factor:200)
- 11% of the sequences remains without annotation



Inherited validated annotation (P<0.01) → + 54%

http://bar.biocomp.unibo.it/bar2.0

BAR+ Bologna Annotation Resource

Protein clusters for sequence annotation

Home

Help

Tutorial

Statistics

Genome list

About us

References

BAR+ is a server for the annotation of protein sequences relying on a comparative large-scale genome analysis across 988 species and the entire UniProt. With BAR+ and a sequence/set of sequences (maximum number per run=50) you can annotate when possible: **function** (Gene Ontology), **structure** and **ligands** (Protein Data Bank), **structural classification** (SCOP), **protein domains** (Pfam). Also if your sequence falls into a cluster with a structural/some structural template/s we provide an alignment towards the template/templates based on the **Cluster-HMM** (HMM profile) that allows you to directly compute your 3D model. Cluster HMMs (over 10.000) are available for downloading.

Enter UniProt accessions

e.g. Q8DN23 P0CH25 Q28654 (*)

OR

Paste your sequences (FASTA format)

OR

Upload a file (FASTA format)

Sfoglia...

Submit Clear

E-mail

*Receive an E-mail when your job finished**

Job-ID

*Insert the job-ID to resume your work**

**Your jobs will be kept for 1 week!

(*)
Sequence already clustered in BAR+
Sequence that falls in clusters with Sequence Identity &ge 40% and Coverage &ge 90% (validated annotation)
Sequence that falls in clusters with Sequence Identity < 40% and Coverege < 90% (not-validated annotation)

http://bar.biocomp.unibo.it/bar2.0

Cluster-ID: 175

No. sequences: 2613

Average length: 346.1 (3.4% SD)

- Eukaryota 534
- Prokaryota 2079

Gene Ontology function (Total: 105)

Molecular Function: (Validated:9/19)

- zinc ion binding
- alcohol dehydrogenase (NAD) activity
- mannitol dehydrogenase activity
- cinnamyl-alcohol dehydrogenase activity
- oxidoreductase activity, acting on CH-OH group of donors
- oxidoreductase activity
- ion binding
- catalytic activity
- molecular_function

Biological Process: (Validated:11/72)

- NADH metabolic process
- ethanol biosynthetic process involved in glucose fermentation to ethanol
- glucose catabolic process to ethanol
- lignin biosynthetic process
- lignin metabolic process
- cellular amino acid derivative biosynthetic process
- monohydric alcohol biosynthetic process
- cellular amino acid derivative metabolic process
- phenylpropanoid metabolic process
- secondary metabolic process
- metabolic process

Structural annotation (No. PDB: 7)

- 1rjwA 1piwA 2hcyA 1lluA 1uufA 2cf5A 1yqdA
- Ligands: ZN 8ID NAD EDO NAP ETF DTT
- SCOP: b.35.1.2 # c.2.1.1

Pfam annotation (Total: 1)

- Alcohol dehydrogenase GroES-like domain(ADH_N)

Cluster HMM

Download HMM

Download **cluster annotation data**

Query

P0CH37

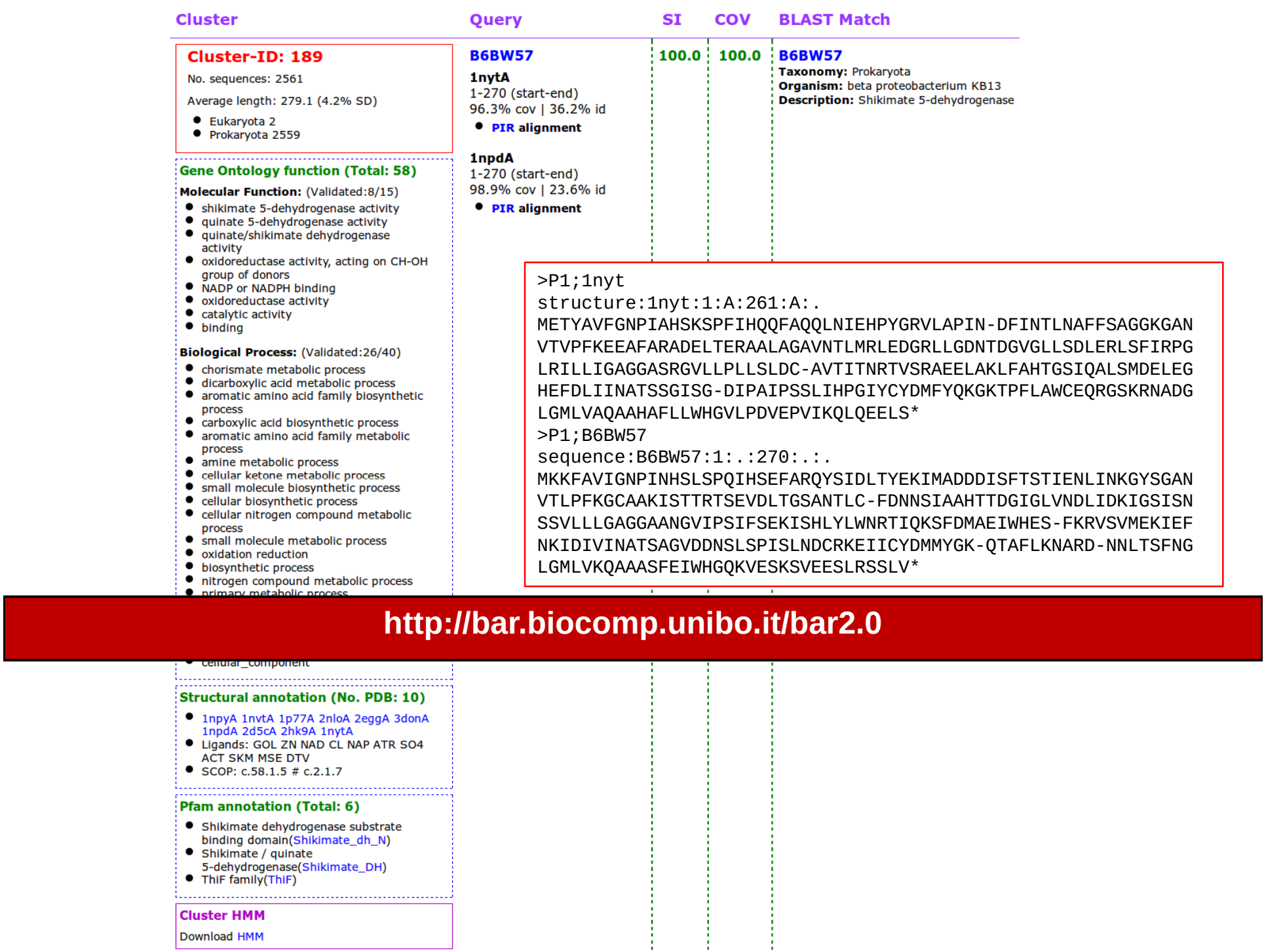
1uufA
3-346 (start-end)
96.8% cov | 60.0% id
- **PIR alignment**

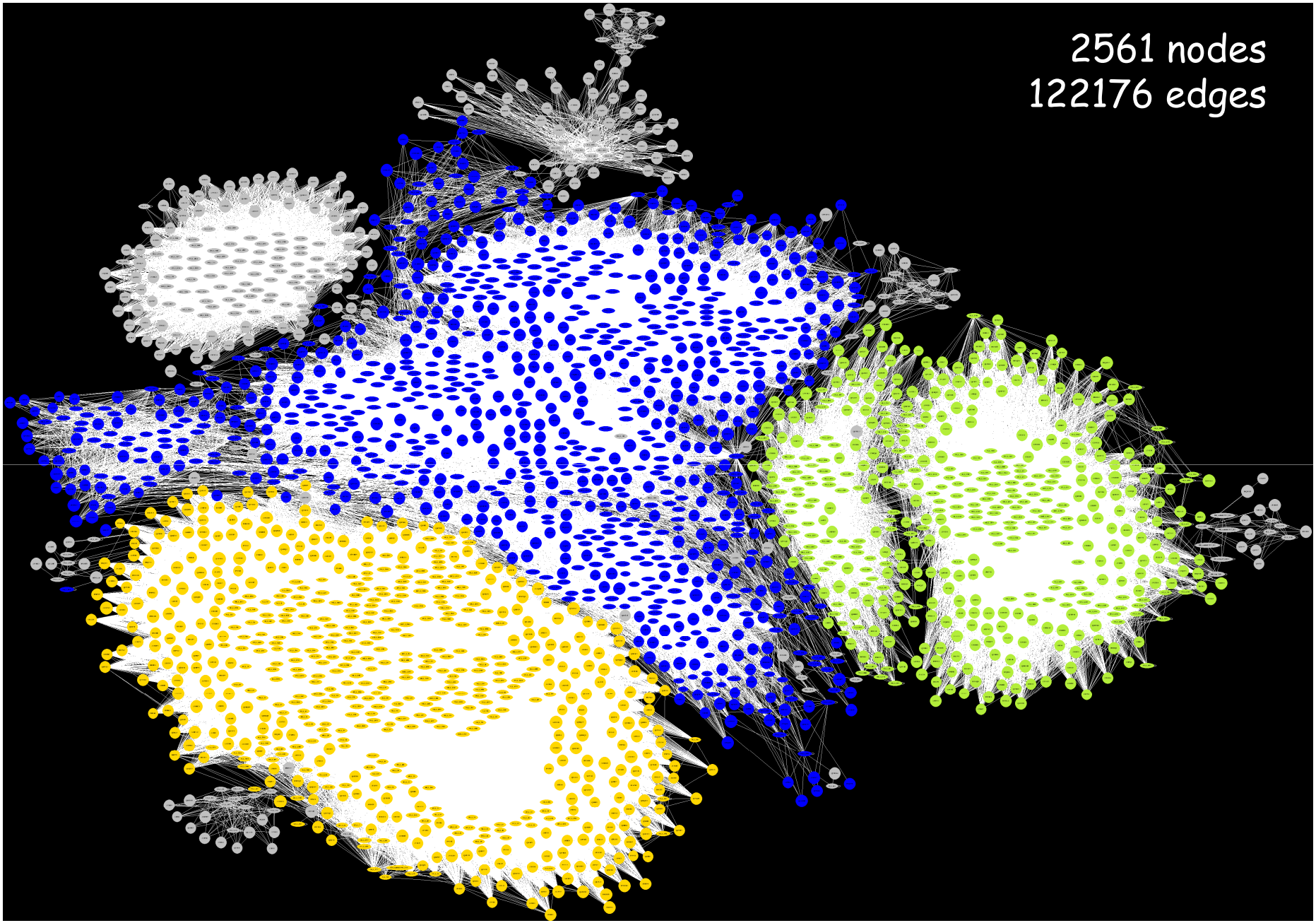
2cf5A
1-349 (start-end)
99.4% cov | 50.0% id
- **PIR alignment**

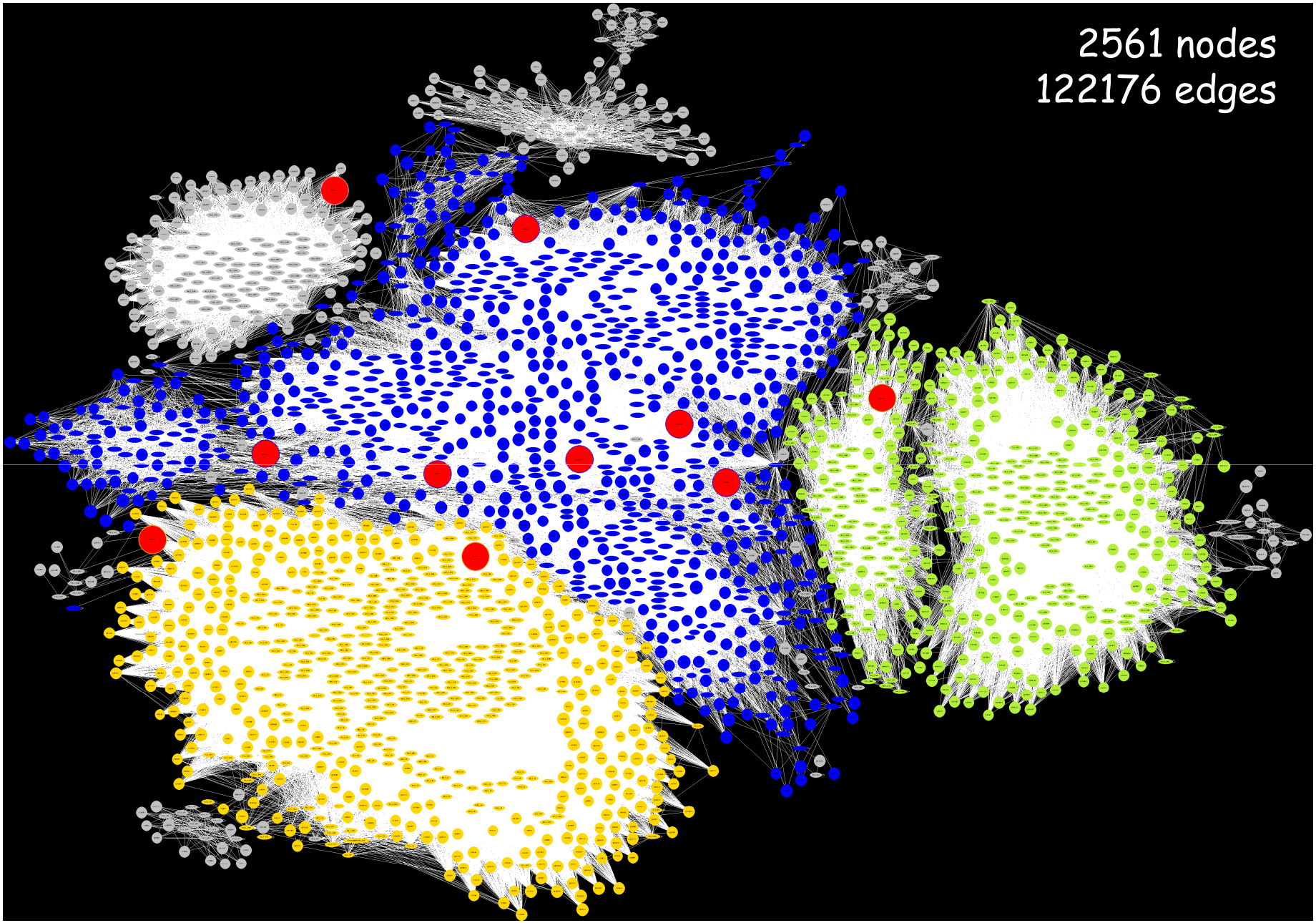
SI	COV	BLAST Match
100.0	92.8	**282_1_2241** **Taxonomy:** Prokaryota **Organism:** Mycobacterium smegmatis **Description:** gi|118468910|ref|YP_886664.1| alcohol dehydrogenase, zinc-containing [Mycobacterium smegmatis str. MC2 155]
100.0	92.8	**282_1_1012** **Taxonomy:** Prokaryota **Organism:** Mycobacterium smegmatis **Description:** gi|118468049|ref|YP_885435.1| alcohol dehydrogenase, zinc-containing [Mycobacterium smegmatis str. MC2 155]
100.0	92.8	**A0QR97** **Taxonomy:** Prokaryota **Organism:** Mycobacterium smegmatis (strain ATCC 700084 / mc(2)155) **Description:** Alcohol dehydrogenase, zinc-containing
99.1	92.8	**Q9AE96** **Taxonomy:** Prokaryota
83.9	92.0	**281_1_1823** **Taxonomy:** Prokaryota **Organism:** Mycobacterium (MCS) **Description:** gi|108798804|ref|YP_639001.1| Alcohol dehydrogenase GroES-like protein [Mycobacterium sp. MCS]
83.9	92.0	**Q1BAY9** **Taxonomy:** Prokaryota **Organism:** Mycobacterium sp. (strain MCS) **Description:** Alcohol dehydrogenase GroES-like protein
83.9	92.0	**A1UE25** **Taxonomy:** Prokaryota **Organism:** Mycobacterium sp. (strain KMS) **Description:** Alcohol dehydrogenase GroES domain protein
83.9	92.0	**278_1_1792** **Taxonomy:** Prokaryota **Organism:** Mycobacterium (JLS) **Description:** gi|126434405|ref|YP_001070096.1| Alcohol

Sequence Identity ≥ 40%

Coverage ≥ 90%

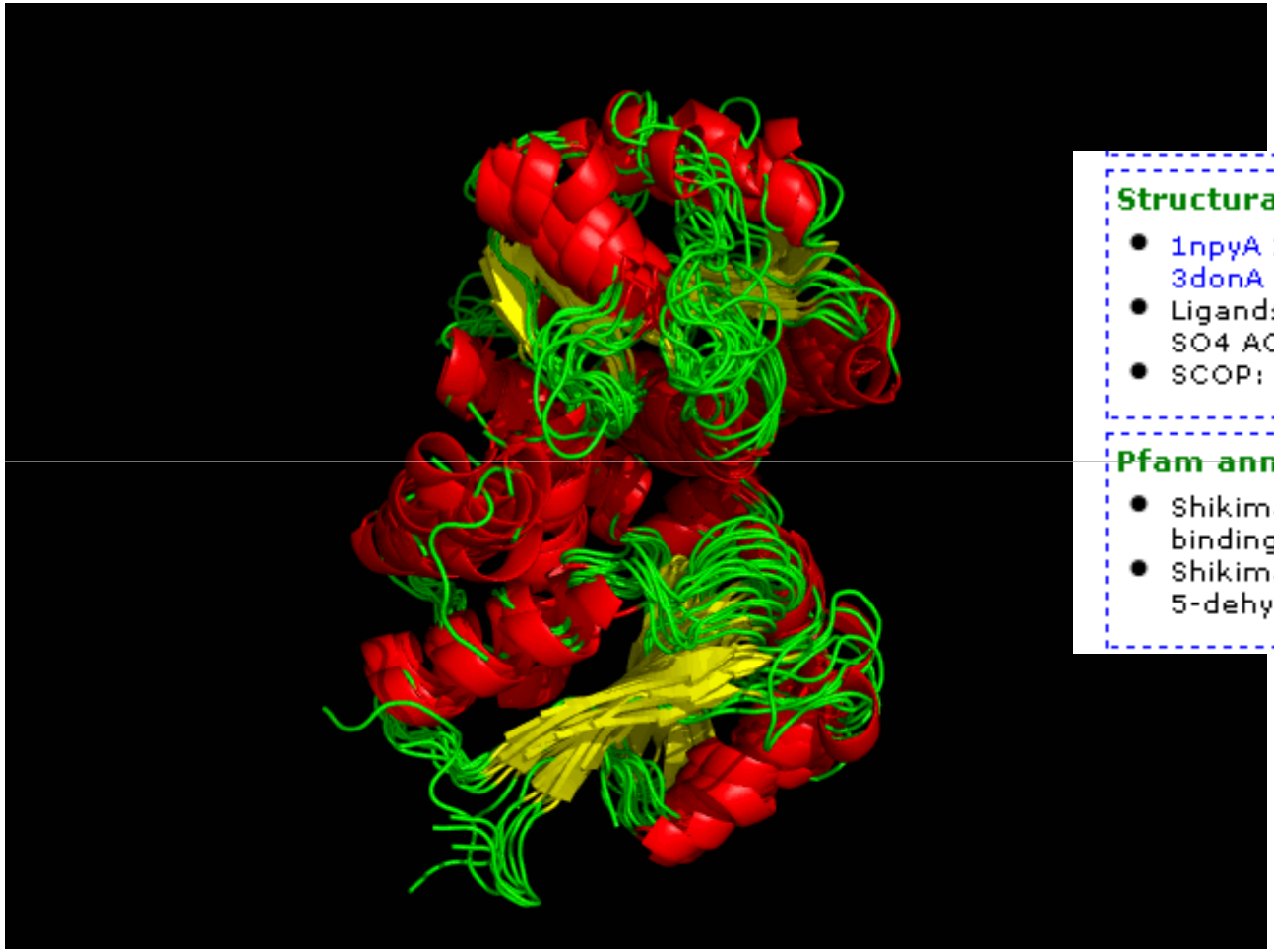

Cluster
Query
SI
COV
BLAST Match
Cluster-ID: 189
No. sequences: 2561
Average length: 279.1 (4.2% SD)
Eukaryota 2
Prokaryota 2559
B6BW57
1nytA
1-270 (start-end)
96.3% cov | 36.2% id
PIR alignment
1npdA
1-270 (start-end)
98.9% cov | 23.6% id
PIR alignment
100.0
100.0
B6BW57
Taxonomy: Prokaryota
Organism: beta proteobacterium KB13
Description: Shikimate 5-dehydrogenase
Gene Ontology function (Total: 58)
Molecular Function: (Validated:8/15)
shikimate 5-dehydrogenase activity
quinate 5-dehydrogenase activity
quinate/shikimate dehydrogenase activity
oxidoreductase activity, acting on CH-OH group of donors
NADP or NADPH binding
oxidoreductase activity
catalytic activity
binding
Biological Process: (Validated:26/40)
chorismate metabolic process
dicarboxylic acid metabolic process
aromatic amino acid family biosynthetic process
carboxylic acid biosynthetic process
aromatic amino acid family metabolic process
amine metabolic process
cellular ketone metabolic process
small molecule biosynthetic process
cellular biosynthetic process
cellular nitrogen compound metabolic process
small molecule metabolic process
oxidation reduction
biosynthetic process
nitrogen compound metabolic process
primary metabolic process
cellular_component
Structural annotation (No. PDB: 10)
1npyA 1nvtA 1p77A 2nloA 2eggA 3donA 1npdA 2d5cA 2hk9A 1nytA
Ligands: GOL ZN NAD CL NAP ATR SO4 ACT SKM MSE DTV
SCOP: c.58.1.5 # c.2.1.7
Pfam annotation (Total: 6)
Shikimate dehydrogenase substrate binding domain(Shikimate_dh_N)
Shikimate / quinate 5-dehydrogenase(Shikimate_DH)
ThiF family(ThiF)
Cluster HMM
Download HMM
>P1;1nyt
structure:1nyt:1:A:261:A:.
METYAVFGNPIAHSKSPFIHQQFAQQLNIEHPYGRVLAPIN-DFINTLNAFFSAGGKGAN
VTVPFKEEAFARADELTERAALAGAVNTLMRLEDGRLLGDNTDGVGLLSDLERLSFIRPG
LRILLIGAGGASRGVLLPLLSLDC-AVTITNRTVSRAEELAKLFAHTGSIQALSMDELEG
HEFDLIINATSSGISG-DIPAIPSSLIHPGIYCYDMFYQKGKTPFLAWCEQRGSKRNADG
LGMLVAQAAHAFLLWHGVLPDVEPVIKQLQEELS*
>P1;B6BW57
sequence:B6BW57:1:.:270:.:.
MKKFAVIGNPINHSLSPQIHSEFARQYSIDLTYEKIMADDDISFTSTIENLINKGYSGAN
VTLPFKGCAAKISTTRTSEVDLTGSANTLC-FDNNSIAAHTTDGIGLVNDLIDKIGSISN
SSVLLLGAGGAANGVIPSIFSEKISHLYLWNRTIQKSFDMAEIWHES-FKRVSVMEKIEF
NKIDIVINATSAGVDDNSLSPISLNDCRKEIICYDMMYGK-QTAFLKNARD-NNLTSFNG
LGMLVKQAAASFEIWHGQKVESKSVEESLRSSLV*
http://bar.biocomp.unibo.it/bar2.0


2561 nodes
122176 edges


2561 nodes
122176 edges

Cluster 189

Average RMSD 2.4 Å (0.4 Å SD)
Average sequence identity 31.9% (5.9% SD)



Structural annotation (No. PDB: 10)

- 1npyA 1nvtA 1p77A 2nloA 2eggA 3donA 1npdA 2d5cA 2hk9A 1nytA
- Ligands: GOL ZN NAD CL NAP ATR SO4 ACT SKM MSE DTV
- SCOP: c.58.1.5 # c.2.1.7

Pfam annotation (Total: 6)

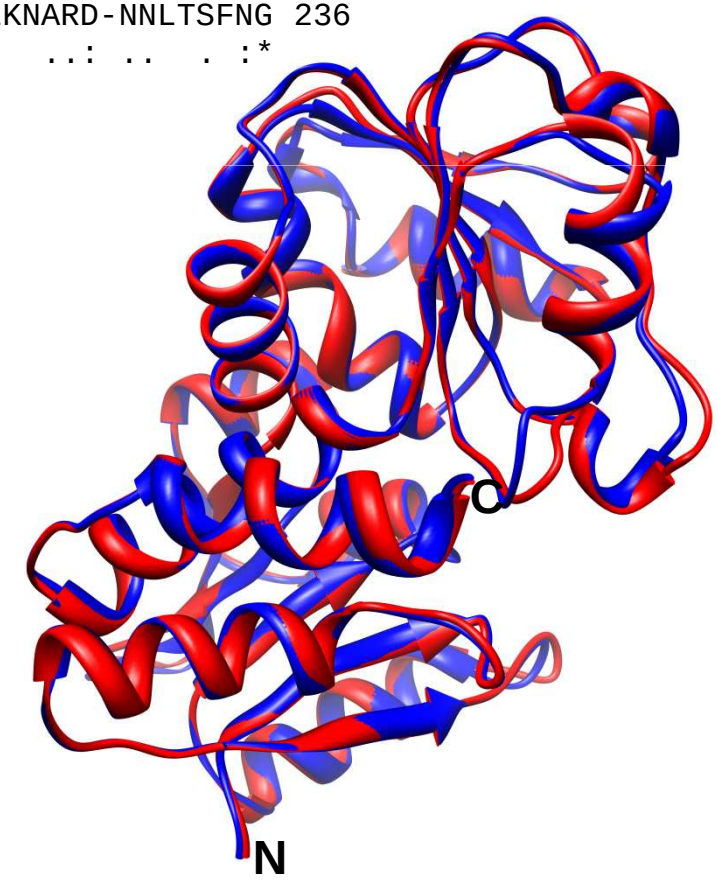
- Shikimate dehydrogenase substrate binding domain(Shikimate_dh_N)
- Shikimate / quinate 5-dehydrogenase(Shikimate_DH)

With MUSTANG v3.2 (Konagurthu et al., Proteins 64(3):559-574, 2006)

```
1NYT_A       METYAVFGNPIAHSKSPFIHQQFAQQLNIEHPYGRVLAPIN-DFINTLNAFFSAGGKGAN 59
B6BW57_9PROT MKKFAVIGNPINHSLSPQIHSEFARQYSIDLTYEKIMADDDISFTSTIENLINKGYSGAN 60
             *:.:**:**** ** ** **.:**:* .*: .* :::*  : .* .*:: ::. * .***

1NYT_A       VTVPFKEEAFARADELTERAALAGAVNTLMRLEDGRLLGDNTDGVGLLSDLERLSFIRPG 119
B6BW57_9PROT VTLPFKGCAAKISTTRTSEVDLTGSANTLC-FDNNSIAAHTTDGIGLVNDLIDKIGSISN 119
             **:***  *   :   *... *:*:.***  :::. : ...***:**:.**       ..

1NYT_A       LRILLIGAGGASRGVLLPLLSLDC-AVTITNRTVSRAEELAKLFAHTGSIQALSMDELEG 178
B6BW57_9PROT SSVLLLGAGGAANGVIPSIFSEKISHLYLWNRTIQKSFDMAEIWHES-FKRVSVMEKIEF 178
             :**:*****:.**: .::* .   : : ***:.:: ::*::: .:   :.  *:::*

1NYT_A       HEFDLIINATSSGISG-DIPAIPSSLIHPGIYCYDMFYQKGKTPFLAWCEQRGSKRNADG 237
B6BW57_9PROT NKIDIVINATSAGVDDNSLSPISLNDCRKEIICYDMMYGK-QTAFLKNARD-NNLTSFNG 236
             :::*::*****:*:.. .:..*. .  :  * ****:* * :*.**  ..: ..  . :*

1NYT_A       LGMLVAQAAHAFLLWHGVLPDVEPVIKQLQEELS 271
B6BW57_9PROT LGMLVKQAAASFEIWHGQKVESKSVEESLRSSLV 270
             ***** *** :* :***   : :.* :.*:..*
```

Alignment with cluster-specific profile-HMM

Sequence id= 32%
RMSD= 0.47 Å



PDB 1NYT vs B6BW57_9PROT

Nucleic Acids Research Advance Access published May 26, 2011

Nucleic Acids Research, 2011, 1–6
doi:10.1093/nar/gkr292

BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences

Damiano Piovesan[1], Pier Luigi Martelli[1], Piero Fariselli[2], Andrea Zauli[3], Ivan Rossi[3] and Rita Casadio[1,*]

[1]Department of Biology, Bologna Biocomputing Group, Bologna Computational Biology Network, [2]Department of Computer Science, University of Bologna, Bologna and [3]BioDec srl, Bologna, Italy

Received February 4, 2011; Revised April 4, 2011; Accepted April 13, 2011

---

technical notes Journal of proteome research

The Bologna Annotation Resource: a Non Hierarchical Method for the Functional and Structural Annotation of Protein Sequences Relying on a Comparative Large-Scale Genome Analysis

Lisa Bartoli,[†] Ludovica Montanucci,[†] Raffaele Fronza,[†] Pier Luigi Martelli,[†] Piero Fariselli,[†] Luciana Carota,[‡] Giacinto Donvito,[§] Giorgio P. Maggi,[§,||] and Rita Casadio*[,†]

Biocomputing Group, CIRB/Dept of Biology, University of Bologna, Italy, CNAF-INFN, National Institute of Nuclear Physics, Bologna, Italy, BA-INFN, National Institute of Nuclear Physics, Bari, Italy, and Physics Department, Politecnico di Bari, Italy

Received March 4, 2009

The Challenge: The CAFA EXPERIMENT and the protein function prediction (validated transfer of GO terms within clusters)

Our solutions: validated transfer of GO terms within clusters

- From protein sequence to function and structure with BAR+

Damiano Piovesan[1,*], Pier Luigi Martelli[1], Piero Fariselli[1], Andrea Zauli[2], Ivan Rossi[2] and Rita Casadio[1].
[1] Bologna Biocomputing Group, Bologna Computational Biology Network, University of Bologna, Italy;
[2] BioDec srl, Bologna, Italy
*To whom correspondence should be addressed: {piovesan, casadio}@biocomp.unibo.it

-

Table 1. Predicting annotations with BAR+

CAFA sets	Cov	MFO *OR* BPO	MFO	BPO	CCO	ALL-O	Pfam	PDB
Eukaryotes [35465]	*90%*	20,516	17,374	17,116	16,428	22,717	24,020	8,044
	70%	1,447						
Prokaryotes [11532]	*90%*	8,514	7,832	7,463	4,694	8,755	9,615	5,475
	70%	192						
"Add" set [1301]	*90%*	1,198	1,130	786	41	1,140	1,225	463
	70%	38						
Total [48298]		**30,228**	**26,336**	**25,365**	**21,163**	**32,612**	**34,860**	**13,982**
		62.6%	54.5%	52.5%	43.8%	67.5%	72.2%	28.9%

Table 1. Cov: Coverage, the ratio of the length of the intersection of the aligned regions on the two sequences and the overall length of the alignment (namely the sum of the lengths of the two sequences minus the intersection length). For both Cov values Sequence Identity is ≥40%. CCO: Cellular Component. ALL-O: number of sequences with predicted MFO *OR* BPO *OR* CCO. Pfam and PDB are inherited from the cluster with BAR+ (2).

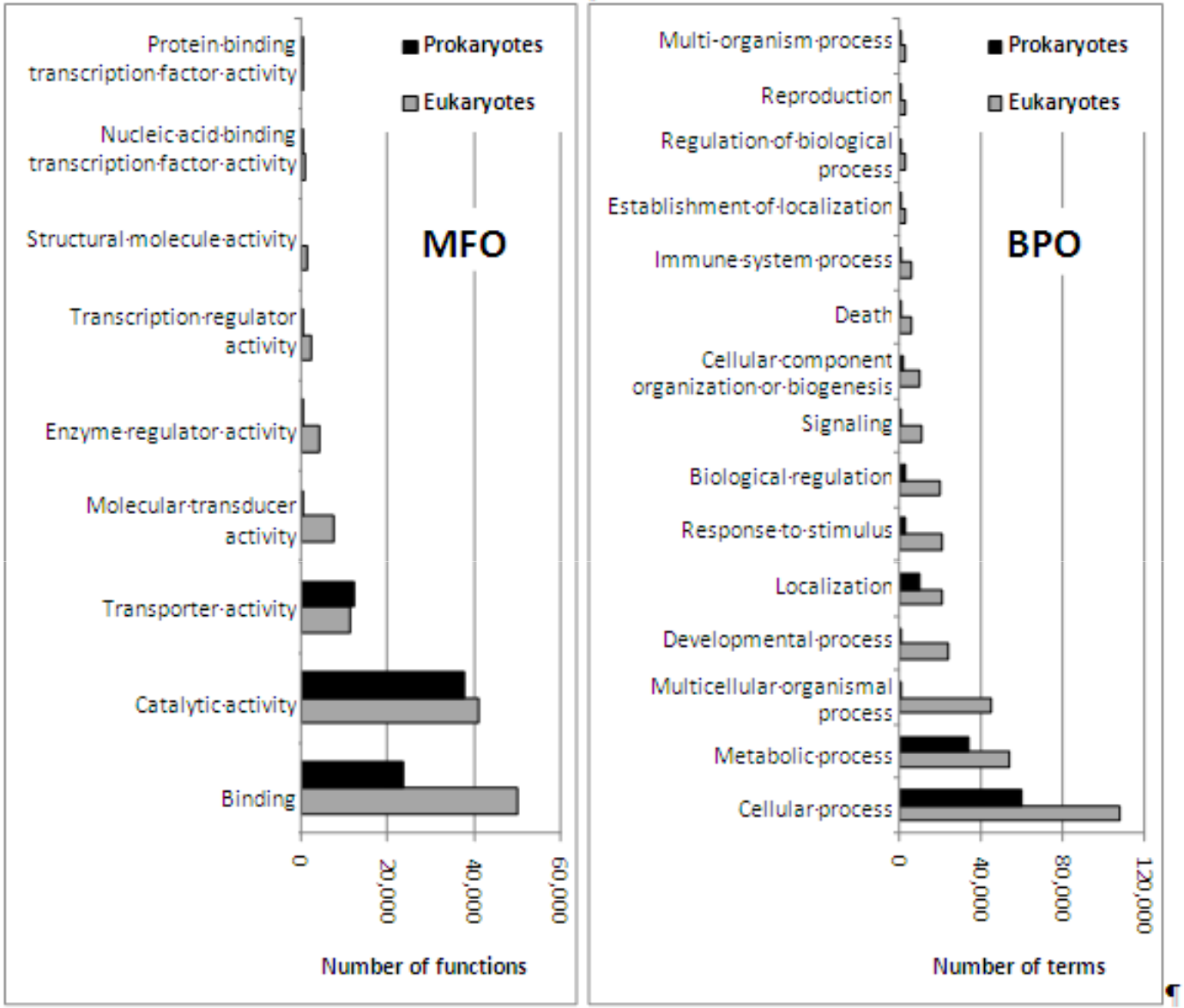


Figure 1. Histogram of the main statistically validated GO molecular functions (MFO) and biological processes (BPO) predicted by BAR+. Terms are included in main categories.

Two main results

1) Analyze the most represented genes among genomes

- Find genes that are expressed in all the kingdoms
- Infer common traits of the living machinery of all prokaryotic and eukaryotic organisms

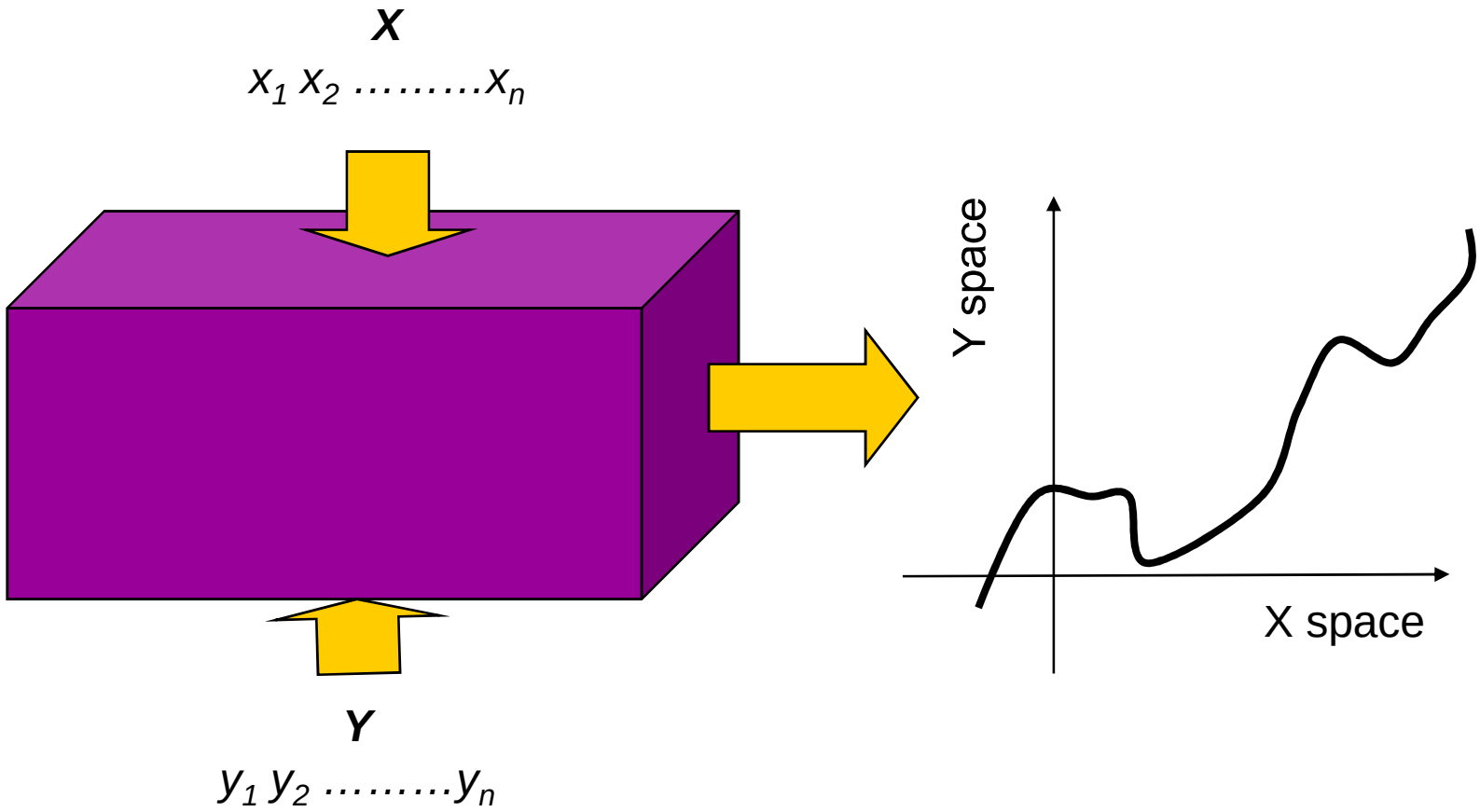
2) Sequence Annotation

- Assign functions and structures to sequences that do not share a high level of similarity with the annotated ones.

Machine learning-based methods for poorly and non-annotated protein sequences
(*putative, hypothetical*)

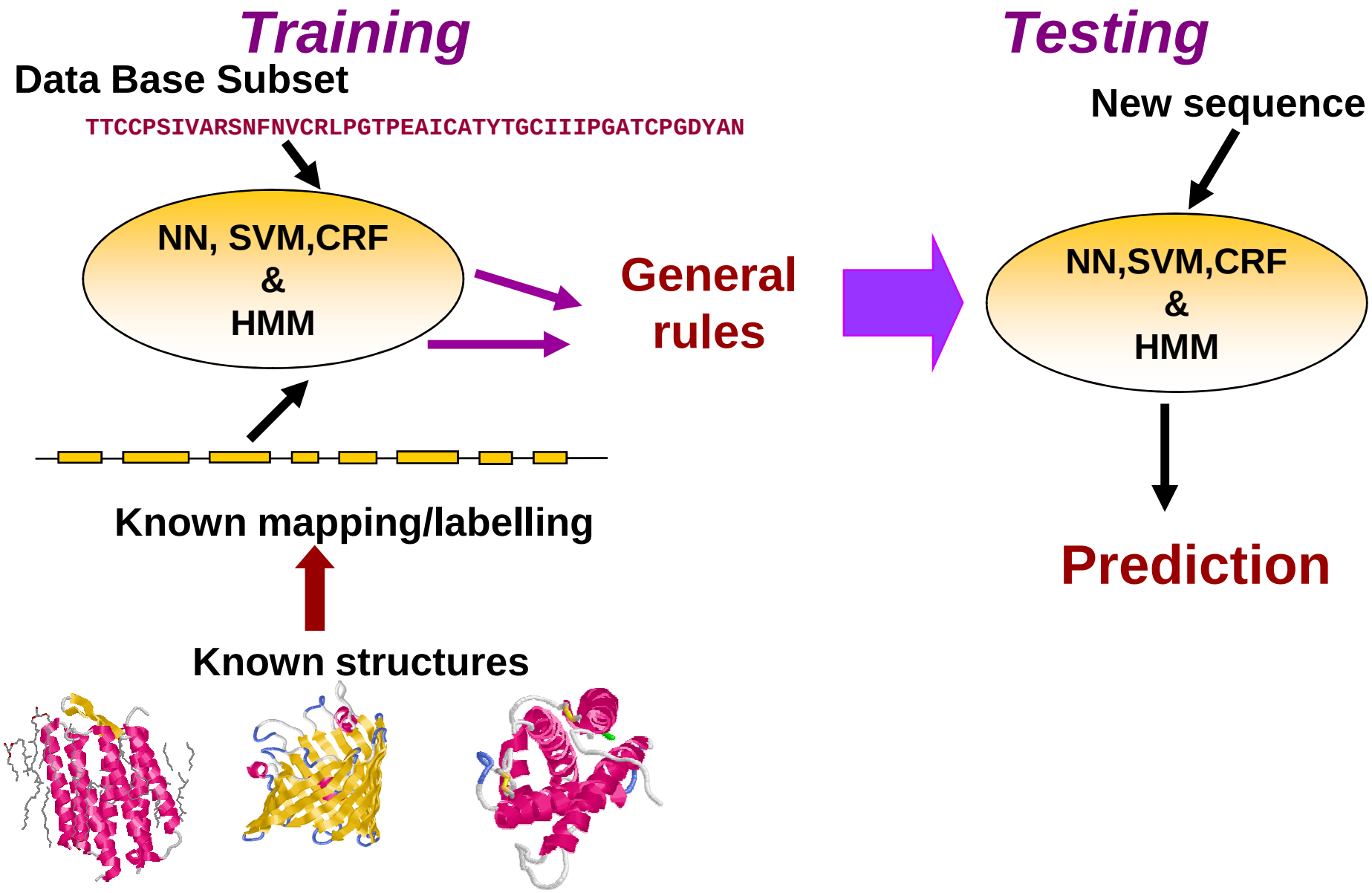
(after BAR filtering some 50% of the total)

General non-linear functional mapping



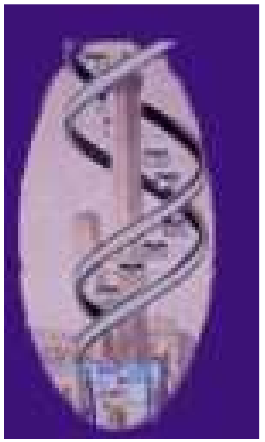
Machine learning approaches help data mining and, if existing, extraction of general rules relating inputs to outputs

Tools out of machine learning approaches



Key factors for a good performance

1) Data base selection for generating the input coding
2) Input coding based on evolution information/sequence profile
3) Choise of the best architecture/model for the problem at hand
4) Rigorous statistical validation (based on jacknife or cross validation procedures)



PREDICTORS (NN, SVM, HMM, GRHCRF)

BaCelLo - Balanced subCellular Localization predictor
BAR+ - Bologna Annotation Resource
BetAware - Detection of Prokaryotic outer-membrane betabarrel proteins
CCHMM - Predictor of Coiled-Coils Regions in Proteins
CCHMMPROF - Predictor of Coiled-Coils Regions in Proteins exploiting evolutionary information
CORNET - Predictor of Residue Contacts in Proteins
DCON - Predictor of Disulfide Connectivity in Proteins
DisLocate - Find Disulfide bonds in Eukaryotes with predicted subcellular Localization
FT-COMAR - Fault Tolerance Reconstruction of 3D Structure from Protein Contact Maps
HIPPIE - Protease Inhibitor engine
I-MUTANT - Neural Network based Predictor of Protein stability Changes upon Single Point Mutation
I-MUTANT 2.0 - Support Vector Machines based Predictor of Protein stability Changes upon Single Point Mutation from Protein Sequence and Structure
I-MUTANT Suite - Support Vector Machines based Predictor of Protein stability Changes of protein variants and of human SNPs
ISPRED - Predictor of Protein Interaction Sites
K-Fold - Predictor of the Protein Folding Mechanism and Rate
PhD-SNP - Support Vector Machines based Predictor of human Deleterious Single Nucleotide Polymorphysms
PredGPI - Predictor of GPI-Anchored Proteins
SNPs&GO - Predictor of Human Disease-related Mutations in Proteins with Functional Annotations
SPEPLip - Predictor of Signal Peptide and Lipoprotein Cleavage Sites in Proteins
YAP - Yet Another Alignment Program (Pairwise Sequence Alignment Using Secondary Structures)

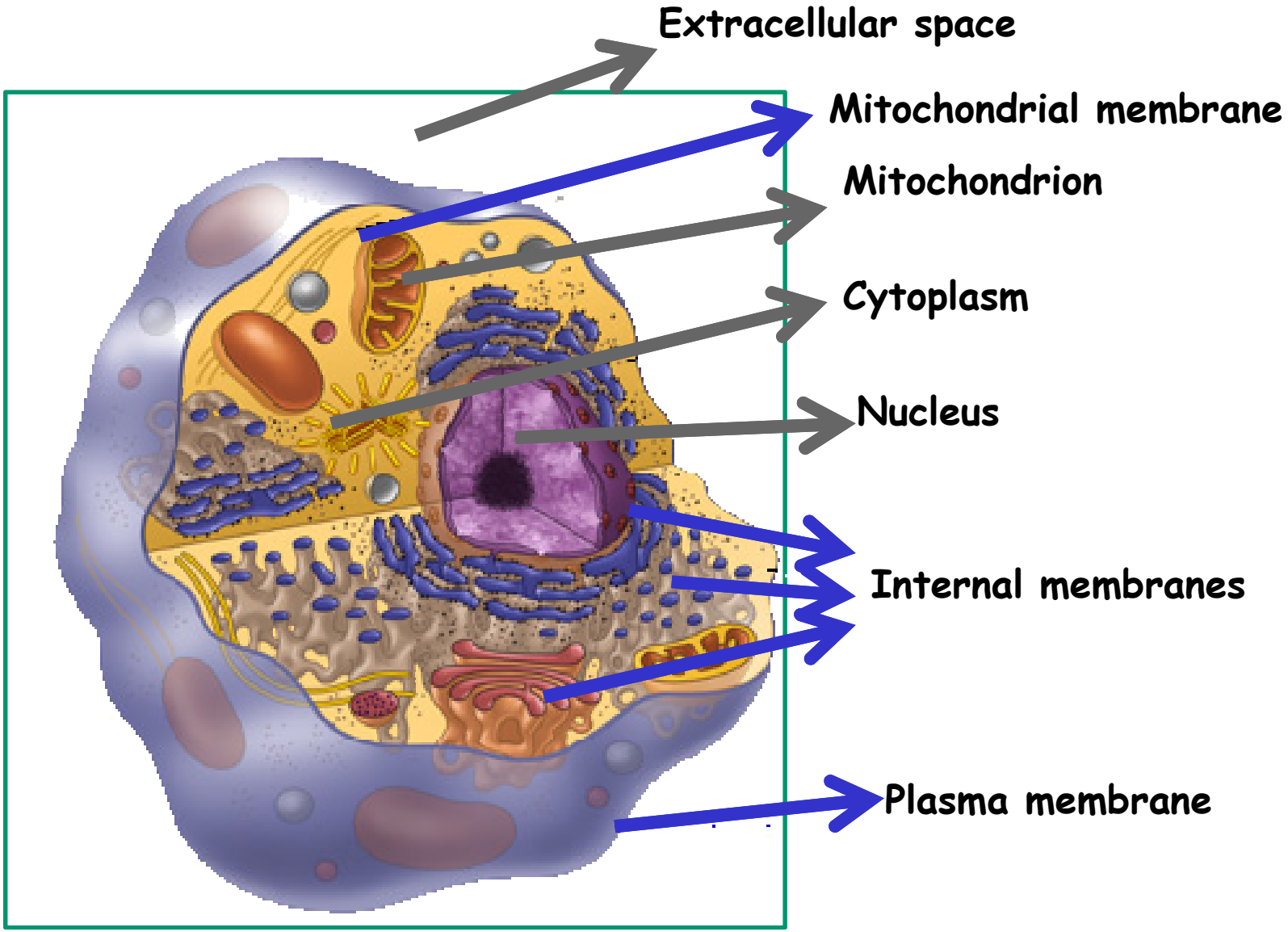
APPLICATION SERVERS

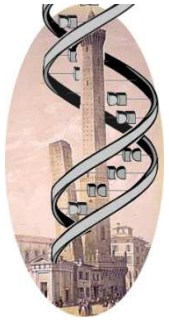
TRAMPLE: the transmembrane protein labelling environment
PONGO: a web server for multiple predictions of all-alpha transmembrane proteins

DATA BASES

eSLDB - eukaryotic Subcellular Localization DataBase
ZenPatches - Database of predicted protein interaction sites
DBMFHS - Data Base of Minimally-Frustrated Helical Segments

1) The role of Compartments in Eukaryotes





Subcellular localization prediction in Eukaryotes

BaCelLo (Balanced subCellular Localization predictor) is:

- **based on SVMs organised in a decision tree**
- **trained on non redundant data set**
- **kingdom specific (Animals, Plants, Fungi)**
- **trained with a balancing procedure that corrects biases in the data set**

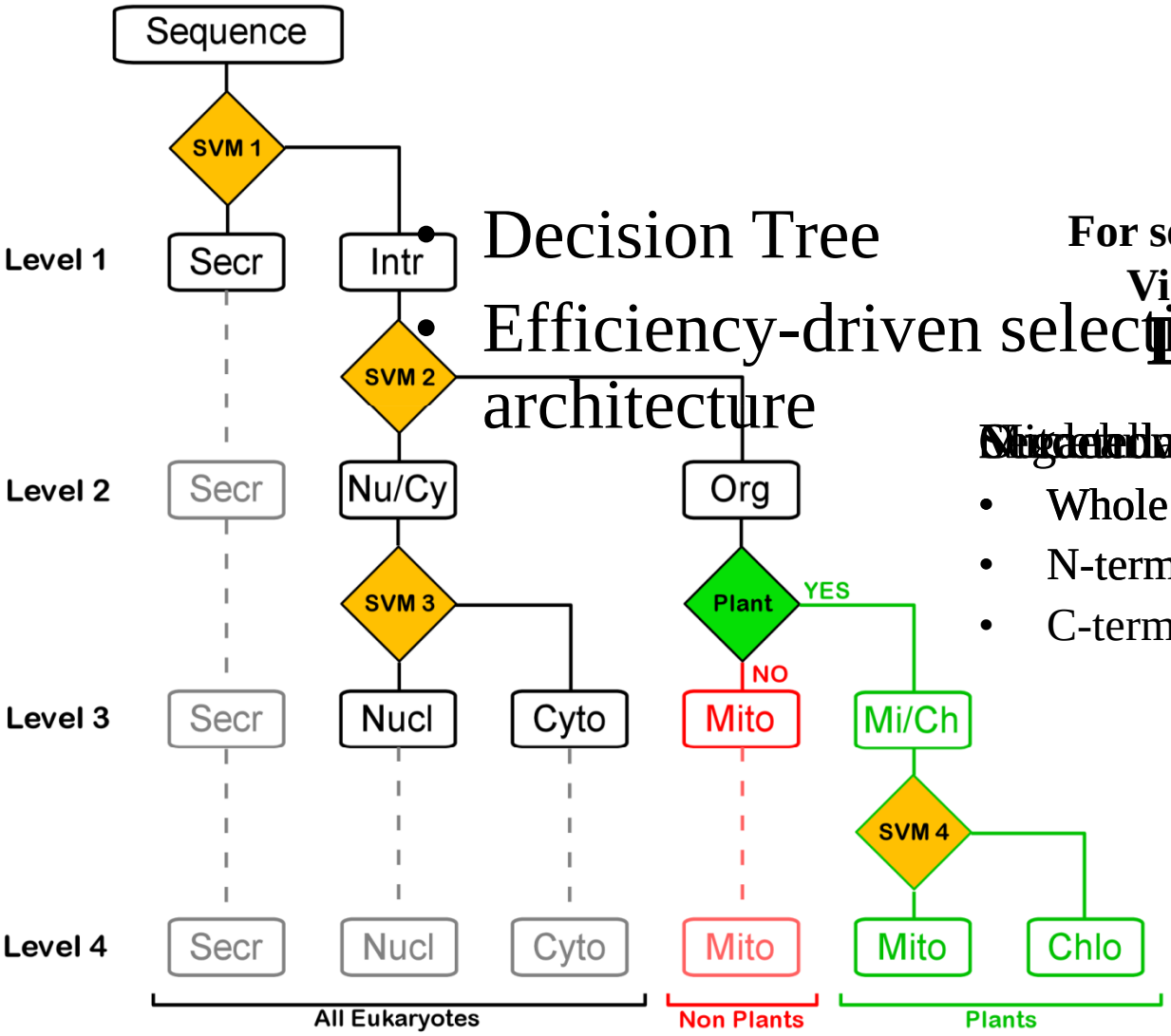
Pierleoni A, Martelli PL, Fariselli P, Casadio R -BaCelLo: a balanced subcellular localization predictor- *Bioinformatics* 22:e408-e416 (2006)

nature protocols
Recipes for Researchers

BaCelLo is also described in Nature Protocols

http://www.nature.com/nprot/
DOI: 10.1038/nprot.2007.165

Architecture of BaCelLo

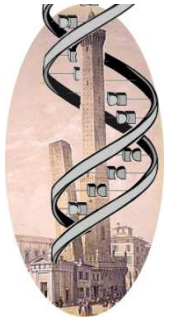


- Decision Tree
- Efficiency-driven selection of the architecture

For sequences from Viridiplantae

Level 2

- Whole sequence
- N-terminal regions
- C-terminal regions



BaCelLo achieves better results with respect to other predictors

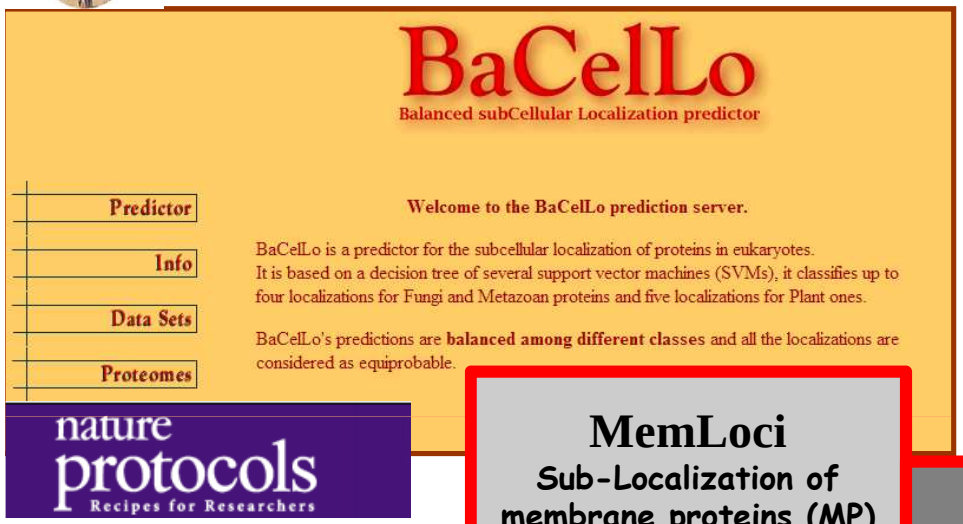
	Fungi				Animals			
	3 classes (level 2)		4 classes (level 3)		3 classes (level 2)		4 classes (level 3)	
Predictor	**nQ**	**GC**	**nQ**	**GC**	**nQ**	**GC**	**nQ**	**GC**
BaCelLo	**94.2**	**0.79**	**79.2**	**0.68**	**82.4**	**0.71**	**68.5**	**0.60**
LocTree	77.5	0.57	64.3	0.50	77.9	0.64	61.3	0.53
Psort II	69.3	0.53	56.9	0.47	67.8	0.48	55.5	0.43
SubLoc	66.8	0.45	51.9	0.39	60.1	0.38	46.5	0.32
ESLpred	64.1	0.58	50.7	0.49	55.7	0.36	45.5	0.31
LOCSVMpsi	80.8	0.77	65.7	0.65	69.0	0.57	55.3	0.48
SLP-Local	77.7	0.62			76.6	0.70		
Protein Prowler	77.3	0.70			79.0	0.75		
TARGETp	80.6	0.62			83.0	0.70		
PredoTar	87.6	0.70			79.4	0.67		

nQ: Accuracy
GC: Correlation index

3 classes: Extracellular/NucleoCytoplasmic/Organellar
4 classes: Extracellular/Nuclear/Cytoplasmic/Organellar

Subcellular localization prediction in Eukaryotes

http://www.biocomp.unibo.it/predictors.html



BaCelLo (Balanced subCellular Localization predictor) is:

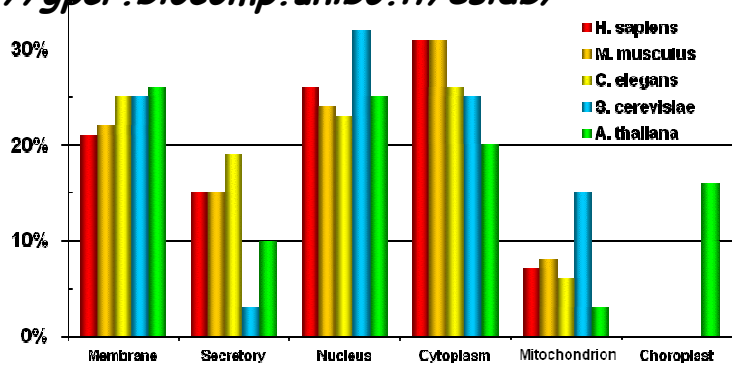
- based on SVMs organised in a decision tree
- trained on non redundant data set
- kingdom specific (Animals, Plants, Fungi)
- trained with a balancing procedure that corrects biases in the data set

MemLoci
Sub-Localization of membrane proteins (MP) plasma, internal and organelle membrane

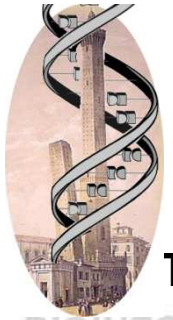
MemPype
Loc and topology of MP

eSLDB

Prediction of BaCelLo on whole proteomes are available on eSLDB (eukaryotic Subcellular Localization Data Base) http://gpcr.biocomp.unibo.it/esldb/



Pierleoni et al, Bioinformatics 22:e408-e416 (2006)
Pierleoni et al, Nucleic Acids Res 35:D208-12 (2007)
http://www.nature.com/nprot/DOI:10.1038/nprot.2007.165
Pierleoni et al, Bioinformatics 27:1224-1230 (2011)
Pierleoni et al, Nucleic Acids Res 39:W375-W380 (2011)



2)The effect of variations in proteins: •Decrease/increase in protein stability

The I-Mutant set of predictors

BIOINFORMATICS
Vol. 20 Suppl. 1 2004, pages i63–i68
DOI: 10.1093/bioinformatics/bth928

A neural-network-based method for predicting protein stability changes upon single point mutations

Emidio Capriotti, Piero Fariselli and Rita Casadio*

I-mutant
Based on NN
2 states: destabilizing/stabilizing mutations

BIOINFORMATICS
Vol. 21 Suppl. 2 2005, pages ii54–ii58
doi:10.1093/bioinformatics/bti1109

Protein Structure and Function

Predicting protein stability changes from sequences using support vector machines

Emidio Capriotti, Piero Fariselli, Remo Calabrese and Rita Casadio*
Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy

W306–W310 Nucleic Acids Research, 2005, Vol. 33, Web Server issue
doi:10.1093/nar/gki375

I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure

Emidio Capriotti, Piero Fariselli and Rita Casadio*

I-mutant 2
Based on SVMs
2 states: destabilizing/stabilizing mutations

BMC Bioinformatics
BMC Bioinformatics 2008, 9(Suppl 2):S6 doi:10.1186/1471-2105-9-S2-S6
BioMed Central
Research
Open Access

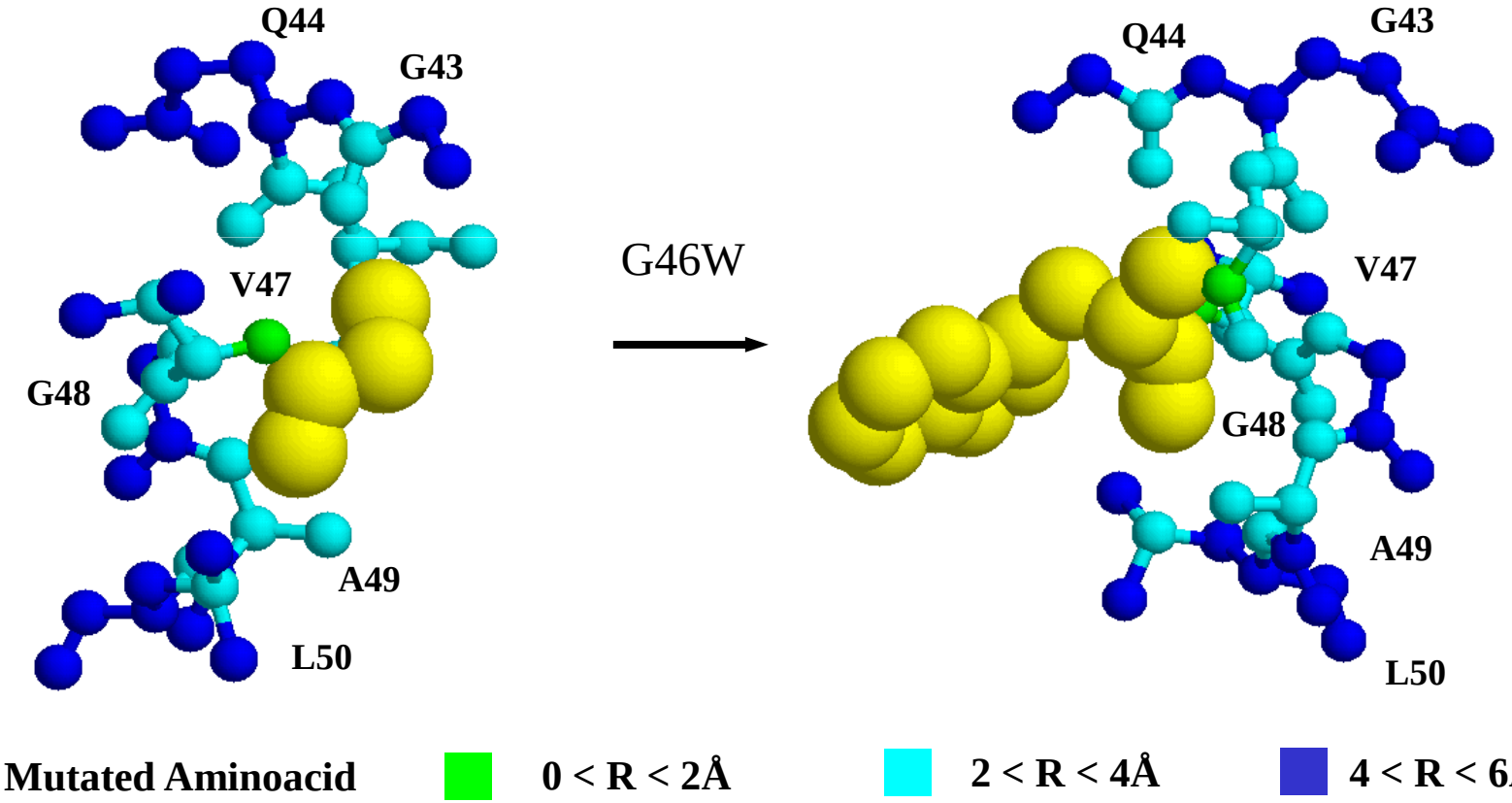
A three-state prediction of single point mutations on protein stability changes

Emidio Capriotti[1], Piero Fariselli[2], Ivan Rossi[2,3] and Rita Casadio*[2]

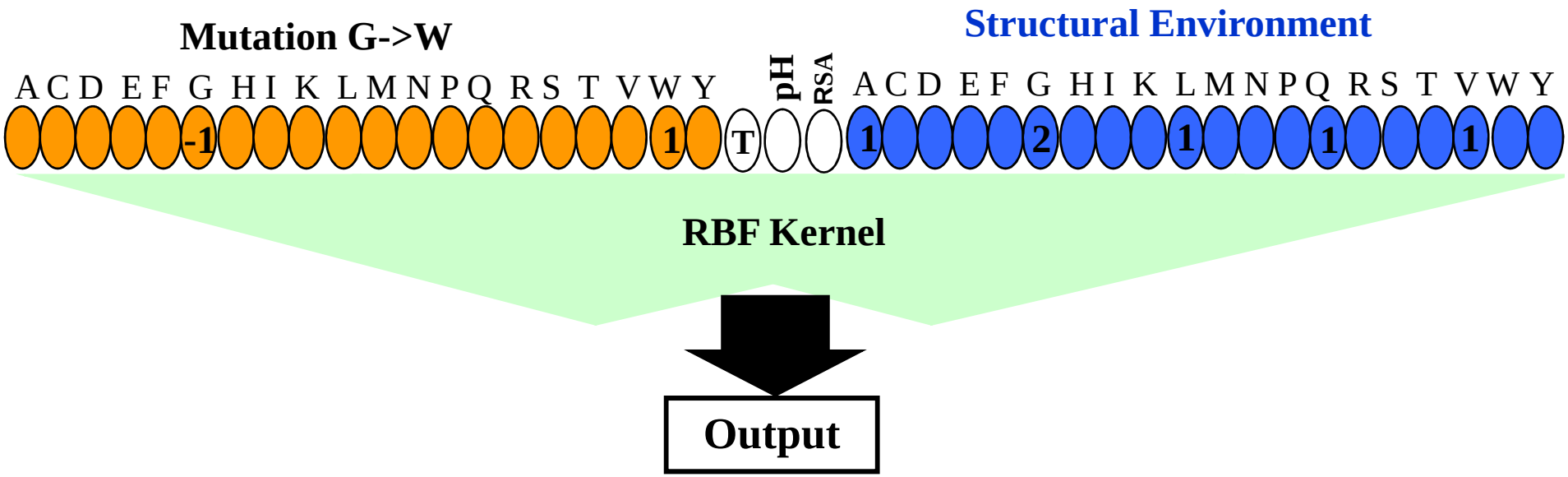
I-mutant 3
Based on SVMs
3 states: destabilizing/neutral/ stabilizing mutations

MUTATION STRUCTURAL ENVIRONMENT

We encode in a 20-element vector the contacts that the mutated residue has with the other side chains that fall inside a sphere of radius R (centered in the C-α carbon of the mutated residue).

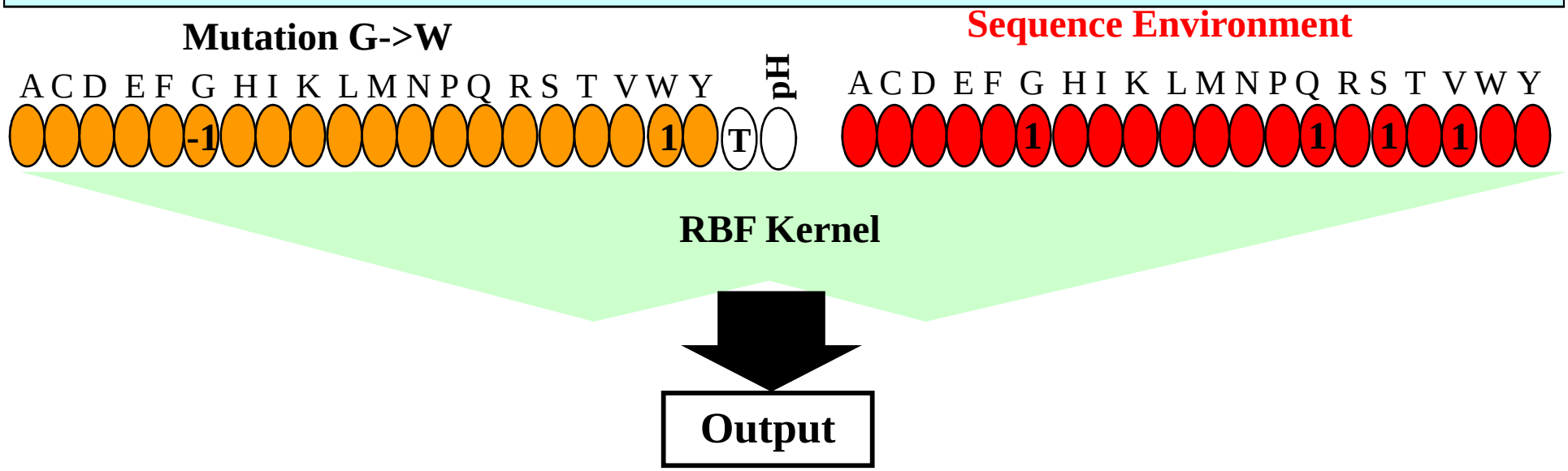


STRUCTURE-BASED PREDICTOR



SVM-3D12: **20 element vector that describes the aminoacid mutation,**
pH and T (experimental conditions)
relative accessibility surface (RSA)
20 more input neurons (43 in total) encoding the three-dimensional residue environment

SEQUENCE-BASED PREDICTOR: I-MUTANT



SVM-WIN15: 20 element vector that describes the residue mutation,
pH and T (experimental conditions)
20 more input neurons (42 in total) encoding the sequence residue environment

I-MUTANT 3.0

Decrease/increase of protein stability upon mutation

I-Mutant Home

Biocomputing Unit

Contact us

Last Update 25/12/06

Sequence File: fileseq.txt

```
Protein Sequence:  From 1 to 215
|---*----|----*----|----*----|----*----|----*----|
MLFNLRILLWMAAFRNGHNFMVRNFRCGQPLQNKVQLKGRDLLTLKNFTG
|---*----|----*----|----*----|----*----|----*----|
EEIKYMLWLSACLKFRIKQKGEYLPLLQGKSLGMIFEKRSTPTRLSTETG
|---*----|----*----|----*----|----*----|----*----|
FALLGGHPCFLTTQDIHLGVNESLTDTARVLSSMADAVLARVYKQSDLDT
|---*----|----*----|----*----|----*----|----*----|
LAKEASIPIINGLSDLYHPIQILADYLTLQEHYSSLKGLTLSWIGDGNNI
|---*----|----*
LHSIMMSAAKFGMML
```

Position	WT	NEW	pH	T
45	L	P	7.0	25

SVM2 Prediction Effect: Decrease RI: 6

SVM3 Prediction Effect: Large Decrease RI: 6

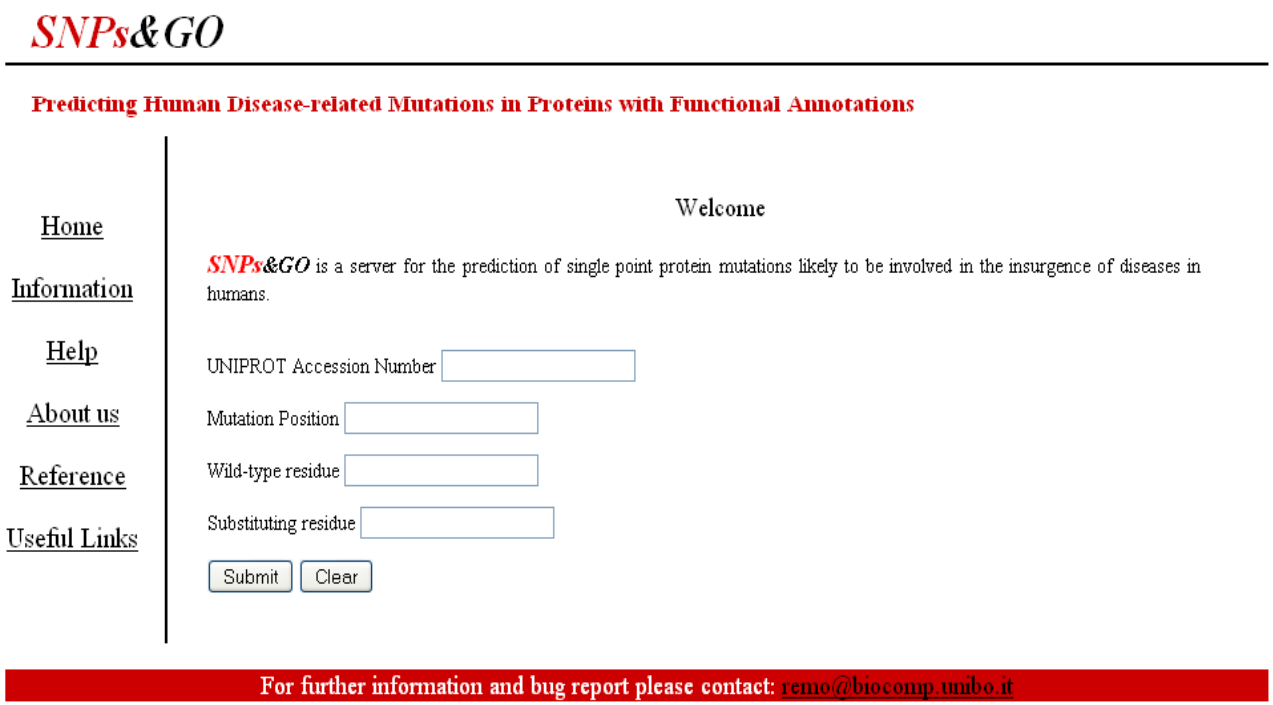
DDG Value Prediction: -1.51 Kcal/mol

Stability Predictors:
- DDG Value:
  - DG(NewProtein)-DG(WildType) in Kcal/mole
- Binary Classification (SVM2):
  - DDG<0: Decrease Stability
  - DDG>0: Increase Stability
- Ternary Classification (SVM3):
  - DDG<-0.5: Large Decrease of Stability
  - DDG>0.5: Large Increase of Stability
  - -0.5<=DDG<=0.5: Neutral Stability

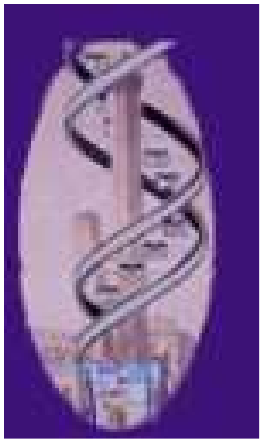
Capriotti et al,
BMC Bioinformatics 2008

3)The effect of variations in proteins:

•Disease related mutations in humans



http://snps-and-go.biocomp.unibo.it/snps-and-go/



METHOD

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Functional Annotations Improve the Predictive Score of Human Disease-Related Mutations in Proteins

Remo Calabrese, Emidio Capriotti, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio*

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy

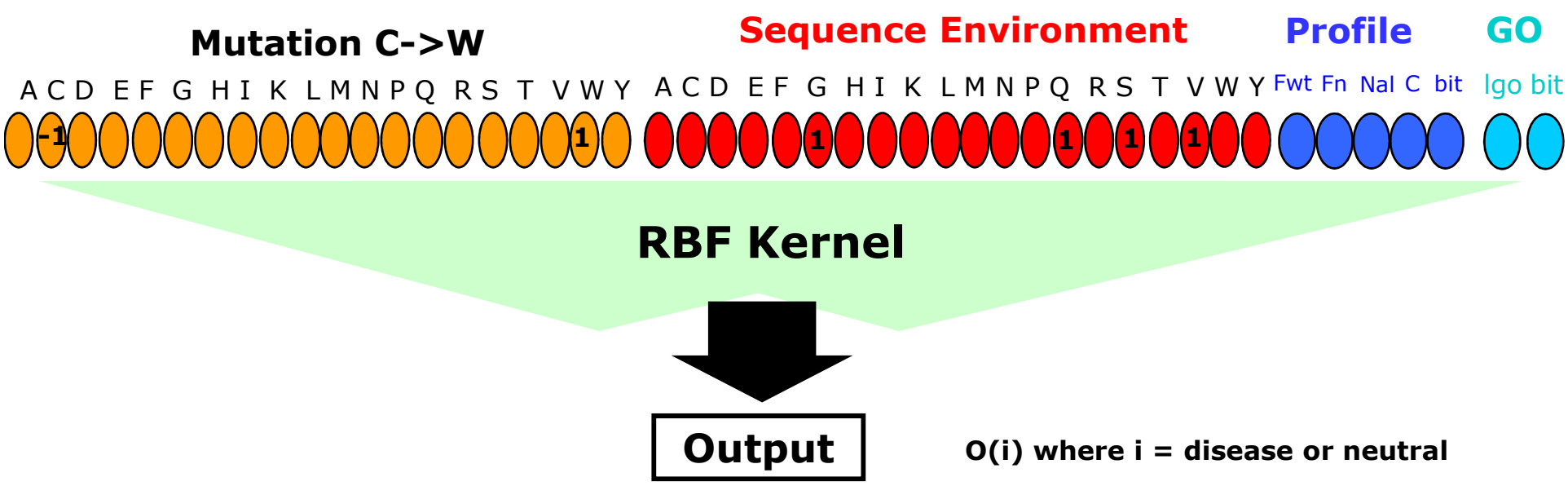
Communicated by Mauno Vihinen

Received 16 December 2008; accepted revised manuscript 1 May 2009.
Published online 12 May 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21047

The role of function/SNPs&GO:

an accurate method that, starting from a protein sequence, can predict whether a mutation is disease related or not by exploiting the protein functional annotation

SNPs&GO: the SEQUENCE-PROFILE AND GO-BASED PREDICTOR



Svm-Seq-Prof-GO: Mutation+ Sequence Environment + Profile + GO

Fwt, Fn= occurrence of wild-type and mutated residue in sequence Profile
Nal= Number of sequences aligned at the mutated position
C= Conservation index based on variance at the mutated position
bit= flag for the presence/absence of the above information
LGO= log-odd ratio of the frequency with which any GO term is found in Disease and Neutral sets
bit= flag for the presence/absence of the above information:

$$LGO = \Sigma \log_2[(GTO(GO)_D + 1)/(GTO(GO)_N + 1)]$$

http://snps-and-go.biocomp.unibo.it/snps-and-go/

Predicting Human Disease-related Mutations in Proteins with Functional Annotations

Home

Information

Help

About us

Reference

SNPs&GO Prediction

****

RESULTS

SEQ File: P13569

Position	WT	NEW	Effect	RI
67	P	L	Disease	10

BP: GO:0001764 BP: GO:0001937 BP: GO:0006811 BP: GO:0007268 BP: GO:0007420 BP: GO:0007585 BP: GO:0007595
MF: GO:0000166 MF: GO:0004713 MF: GO:0005216 MF: GO:0005224 MF: GO:0005260 MF: GO:0005515 MF: GO:0005524
CC: GO:0005624 CC: GO:0005625 CC: GO:0005739 CC: GO:0005829 CC: GO:0005886 CC: GO:0005925 CC: GO:0009925

WT: Residue in Wild-Type Protein
NEW: Mutated Residue
RI: Reliability Index
Effect:
Neutral: Neutral Polymorphism
Disease: Disease-related Polymorphism
BP: Biological Process GO term
MF: Molecular Function GO term
CC: Cellular Component GO term

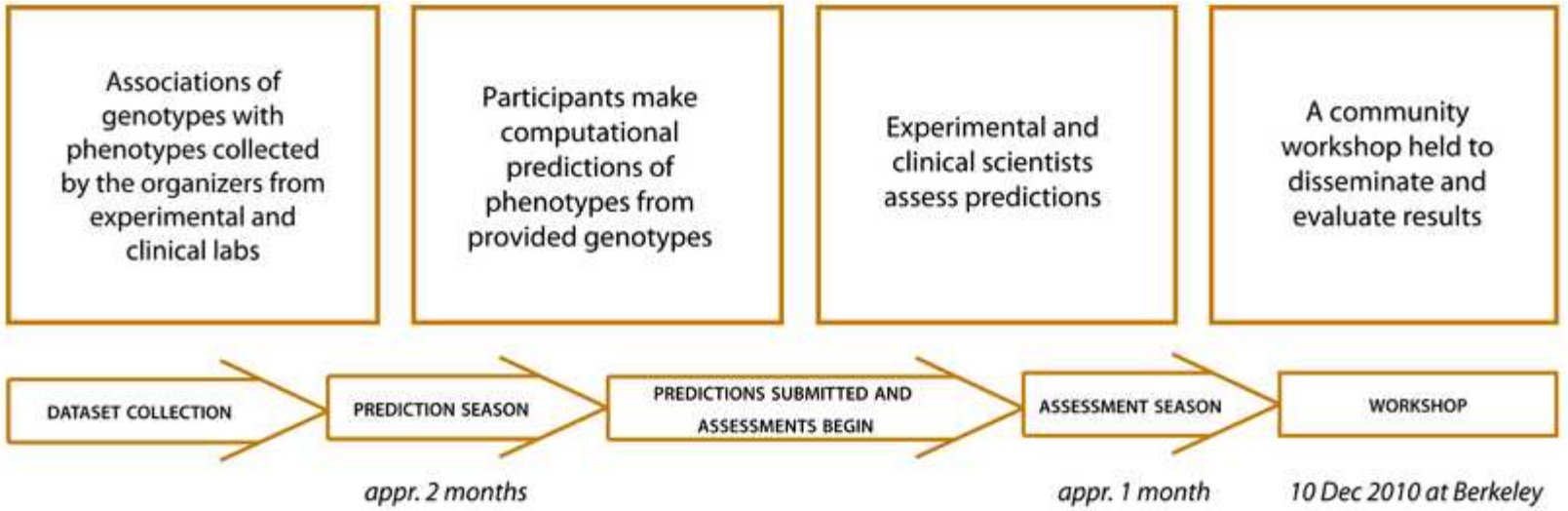
Table 3. Performance of Prediction Methods

	MutPred	nsSNPAnalyzer	Panther	PhD-SNP	PolyPhen 1a	PolyPhen 1b	PolyPhen 2a	PolyPhen 2b	SIFT	SNAP	SNPs&GO
Performance of prediction methods (full data)											
tp[a]	13829	4360	9689	11900	10093	14285	13807	16206	10464	16000	13736
fn[a]	2507	2778	2859	6896	9185	4993	5102	2703	4856	2146	5487
tn[a]	15891	1319	8676	16788	17669	13671	13863	10199	12188	8190	17028
fp[a]	4557	943	2797	4377	3199	7197	6010	9674	7433	6387	1382
cases +[a]	16336	7138	12548	18796	19278	19278	18909	18909	15320	18146	19223
cases −[a]	20448	2262	11473	21165	20868	20868	19873	19873	19621	14577	18410
Accuracy[b]	0.81	0.60	0.76	0.71	0.69	0.70	0.71	0.69	0.65	0.72	0.82
Precision[b]	0.79	0.59	0.76	0.75	0.77	0.68	0.71	0.64	0.64	0.67	0.90
Specificity[b]	0.78	0.58	0.76	0.79	0.85	0.66	0.70	0.51	0.62	0.56	0.92
Sensitivity[b]	0.85	0.61	0.77	0.63	0.52	0.74	0.73	0.86	0.68	0.88	0.71
NPV[b]	0.84	0.60	0.77	0.68	0.64	0.72	0.72	0.78	0.66	0.83	0.76
MCC[b]	0.63	0.19	0.53	0.43	0.39	0.40	0.43	0.39	0.30	0.47	0.65
Performance of prediction methods (3D structure)											
tp[a]	5625	2857	3934	5041	4563	5980	5814	6726	4303	6751	5887
fn[a]	517	1603	1009	2411	3074	1657	1842	930	1329	714	1746
tn[a]	1101	569	735	1090	1361	1070	1163	843	904	700	1378
fp[a]	697	527	441	754	462	753	672	992	901	777	318
cases +[a]	6142	4460	4943	7452	7637	7637	7656	7656	5632	7465	7633
cases −[a]	1798	1096	1176	1844	1823	1823	1835	1835	1805	1477	1696
Accuracy[b]	0.76	0.58	0.71	0.63	0.67	0.68	0.70	0.67	0.63	0.69	0.79
Precision[b]	0.70	0.57	0.68	0.62	0.70	0.65	0.67	0.62	0.60	0.63	0.80
Specificity[b]	0.61	0.52	0.63	0.59	0.75	0.59	0.63	0.46	0.50	0.47	0.81
Sensitivity[b]	0.92	0.64	0.80	0.68	0.60	0.78	0.76	0.88	0.76	0.90	0.77
NPV[b]	0.88	0.59	0.75	0.65	0.65	0.73	0.72	0.79	0.68	0.83	0.78
MCC[b]	0.55	0.16	0.43	0.27	0.35	0.38	0.40	0.37	0.27	0.42	0.58
Performance of prediction methods (pathogenic dataset only from LSDBs, not in SwissProt)											
tp	2248	1175	1368	1436	1651	2410	2190	2764	2131	2615	2547
fn	899	862	1252	2158	1943	1184	1361	787	1145	917	952
tn	2655	212	1508	2842	3004	2333	2334	1705	2073	1382	2898
fp	804	165	501	752	534	1205	1028	1657	1268	1069	259
cases +[a]	3139	2037	2620	3594	3594	3594	3551	3551	3276	3532	3499
cases −[a]	3459	377	2009	3594	3538	3538	3362	3362	3341	2451	3157
Accuracy[b]	0.74	0.57	0.64	0.6	0.65	0.66	0.66	0.64	0.64	0.65	0.82
Precision[b]	0.75	0.57	0.68	0.66	0.75	0.66	0.67	0.61	0.63	0.63	0.90
Specificity[b]	0.77	0.56	0.75	0.79	0.85	0.66	0.69	0.51	0.62	0.56	0.92
Sensitivity[b]	0.71	0.58	0.52	0.4	0.46	0.67	0.62	0.78	0.65	0.74	0.73
NPV[b]	0.73	0.57	0.61	0.57	0.61	0.67	0.64	0.70	0.64	0.68	0.77
MCC[b]	0.48	0.14	0.28	0.21	0.33	0.33	0.31	0.30	0.27	0.31	0.66

[a]Total number of cases used by the given program (not normalized).
[b]Accuracy, precision, specificity, sensitivity, NPV, and MCC are calculated from normalised numbers.

Thusberg J, Olatubosun A, Vihinen M. 2011. Hum Mutat. 2:358-68

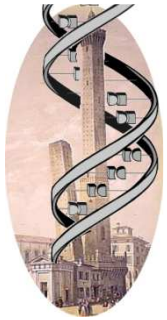
CAGI: Critical Assessment of Genome Interpretation

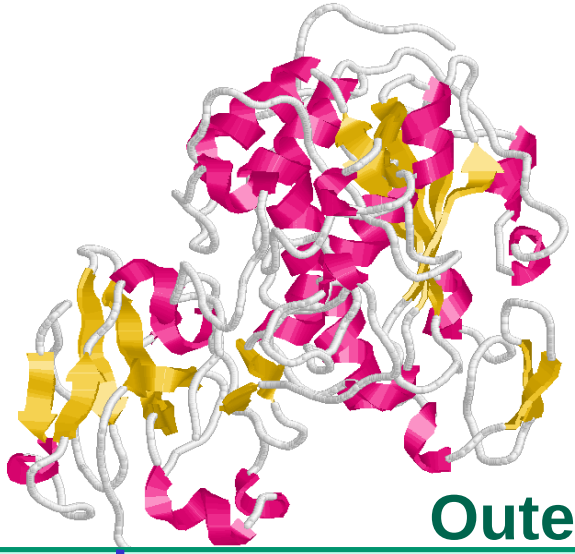


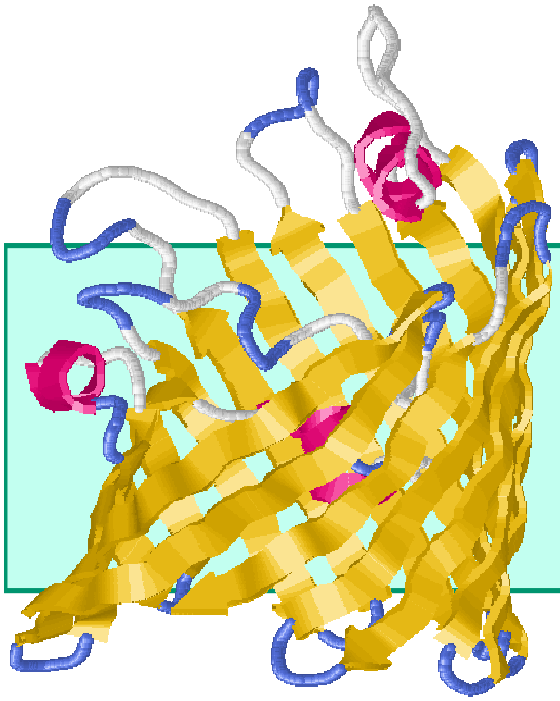
Berkeley, December 2010

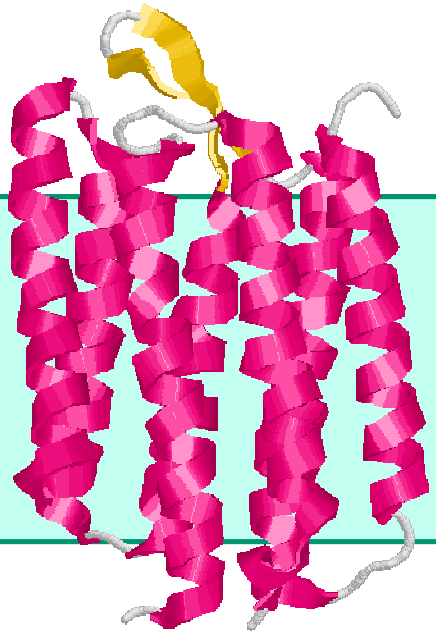
The CAGI-2010 dancing "award" of SNPs&GO

4)The topology, localization and post-translational modification of membrane proteins









Outer Space

Glycophosphatidylinositol (GPI) -anchor

BIOLOGICAL MEMBRANE

Inner Space

Porin
(*Rhodobacter capsulatus*)

Bacteriorhodopsin
(*Halobacterium salinarum*)

Leishmaniolysin
(*Leishmania major*)

β-barrel transmembrane proteins

All-α transmembrane proteins

GPI-anchored membrane proteins

•ANNOTATION OF INTEGRAL MEMBRANE PROTEIN TOPOLOGY

•**HMM-B2TMR**: *prediction of the topology of β-barrel TM proteins in Gram – Bacteria. Based on HMMs (Martelli et al., Bioinformatics, 2002) (SGI Best Paper Award at ISMB2002)(Evaluated as Recommended by Faculty of 1000, Biology reports)*

•***ENSEMBLE***: *prediction of the topology of all-α TM proteins. Based on HMMs and NNs (Martelli et al., Bioinformatics, 2003)*

•***CINTHIA***: *prediction of the topology of all-α TM proteins. Based on HMMs and GRHCRFs (in progress)*

•***TRAMPLE***: *the transmembrane protein labelling environment incorporating different predictors for all-α and β-barrel TM proteins . (Fariselli et al., NAR, 2005)*

•***PONGO***: *database/DAS server for multiple predictions of all-α TM proteins. proteins (Amico et al., NAR, 2006)*

•ANNOTATION OF POST-TRANSLATIONAL MODIFICATIONS

- *SPEPLIP: prediction of the N-terminal signal peptide. Based on NNs. (Fariselli et al, Bioinformatics, 2003)*

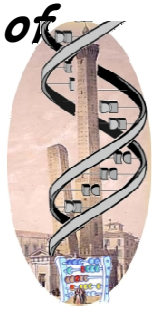
- *PREDGPI: discrimination of GPI-anchored proteins. Based on SVMs and HMMs (Pierleoni et al., BMC Bioinformatics, 2008)*

- *CYSPRED2 : prediction of the bonding state of cysteine residues in proteins. Based on HNN (Martelli et al., Proteomics,2004)*

•ANNOTATION OF MEMBRANE PROTEIN LOCALIZATION

- *MEMLOCI: prediction of the subcellular localization of integral and peripheral membrane proteins in Eukaryotes . Based on SVMs (Pierleoni et al, Bioinformatics 2011)*

- *MEMPYPE: A pipeline for predicting the topology and the localization of membrane proteins (Pierleoni et al. NAR, 2011)*

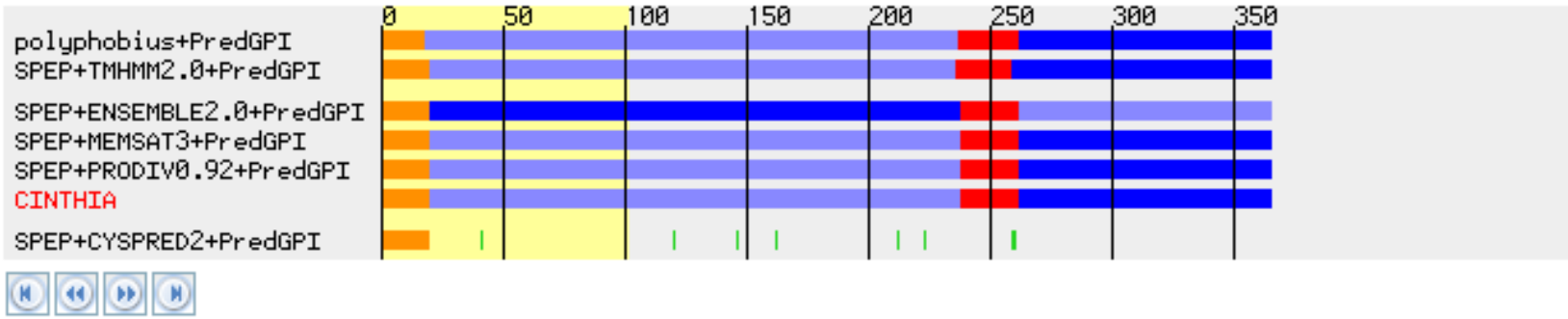


Prediction results

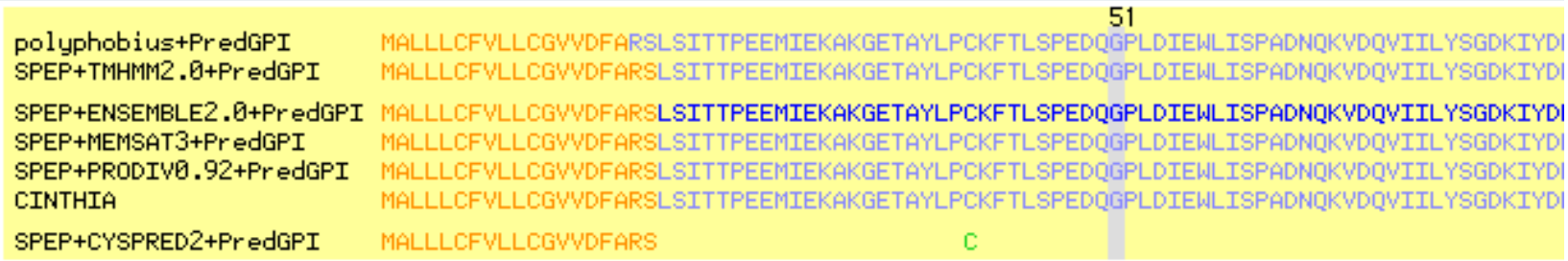
Uniprotcode: P78310
Uniprotcode: P78310-1
CRC64: AB01C6346CB7FE64
submitted by: biocomp.unibo.it
submitted on: 2008

http://pongo.biocomp.unibo.it/

predictions

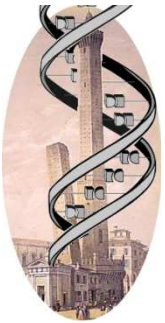


zoom



legend

transmembrane_region
cytoplasmic_region
non_cytoplasmic_region
extramembrane_region
signal_peptide
gpi_propeptide
ss_bond
non_ss_bond



As time goes by… ***the prediction of membrane protein topology from:***

Martelli PL, Fariselli P, Krogh A, Casadio R –
A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins- *Bioinformatics* 18: S46-S53 (**2002**) ***(SGI Best Paper Award at ISMB2002)*** ***(Evaluated as Recommended by Faculty of 1000, Biology reports)***

Martelli PL, Fariselli P, Casadio R –
An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins- *Bioinformatics* 19:I205-I211 (**2003**)

..to recent developments at the Biocomputing Group : GRHCRFs

Algorithms for Molecular Biology BioMed Central

Published: 22 October 2009

Research Open Access

Grammatical-Restrained Hidden Conditional Random Fields for Bioinformatics applications

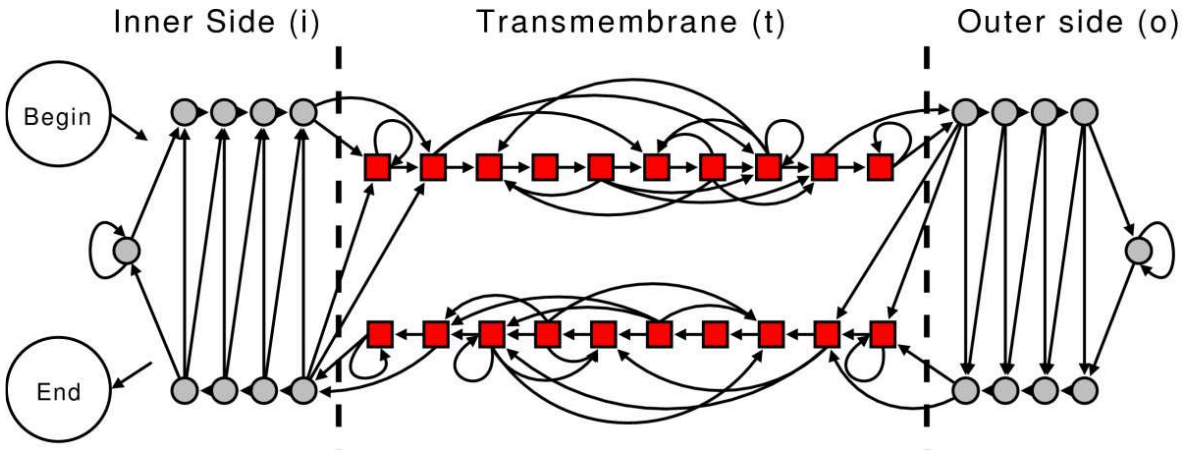
Piero Fariselli*, Castrense Savojardo, Pier Luigi Martelli and Rita Casadio

HMMs vs HCRFs/GRHCRFs

HMM	CRF/HCRF	GRHCRF
•Model joint probability p(x,y)	•Model conditional probability p(y|x)	•Model conditional probability p(y|x)
•Model parameters are transition and emission probabilities	•Model parameters are weights associated to feature functions	•Model parameters are weights associated to feature functions
•Per-state normalization	•Global normalization	•Global normalization
•Emission probabilities describe a single position in the sequence	•Emission parameters can describe several features of the sequence (e.g. local windows)	•Emission parameters can describe several features of the sequence (e.g. local windows)
•Decoupling between states and labels	• States and labels coincide	•Decoupling between states and labels

HMM=Hidden Markov Models (Rabiner ,1989)
CRF/HCRF= Conditional Random Fields /Hidden CRF (Lafferty et al., 2001; Quattoni et al., 2005)
GRHCRF= Grammatical Restrained Hidden Random Fields (Fariselli et al., 2009)

A GRHCRF MODEL FOR TMBB



Three labels [i (periplasmic side), t (trans-membrane segment) and o (extra-cellular side)]
40-state model
Problem constraints are included into the model

Table 1: Prediction of the topology of the Prokaryotic outer membrane proteins.

Method		POV	Q2	C(t)	Sn(t)	Sp(t)
CRF-1 (Vit)	3 different state/label associations	0.26 ± 0.05	0.72 ± 0.01	0.47 ± 0.02	0.59 ± 0.01	0.80 ± 0.01
CRF-1 (Pvit)		0.39 ± 0.05	0.77 ± 0.01	0.54 ± 0.02	0.71 ± 0.01	0.80 ± 0.01
CRF-2 (Vit)		0.34 ± 0.05	0.76 ± 0.01	0.52 ± 0.03	0.63 ± 0.02	0.82 ± 0.02
CRF-2 (Pvit)		0.47 ± 0.05	0.80 ± 0.01	0.60 ± 0.03	0.74 ± 0.02	0.82 ± 0.02
CRF-3 (Vit)		0.29 ± 0.04	0.72 ± 0.01	0.45 ± 0.02	0.60 ± 0.02	0.79 ± 0.01
CRF-3 (Pvit)		0.45 ± 0.04	0.76 ± 0.01	0.52 ± 0.02	0.70 ± 0.02	0.79 ± 0.01
GRHCRF		0.66 ± 0.04	0.85 ± 0.01	0.70 ± 0.03	0.83 ± 0.01	0.84 ± 0.01
HMM-B2TMR		0.58 ± 0.04	0.80 ± 0.01	0.62 ± 0.02	0.82 ± 0.02	0.83 ± 0.01

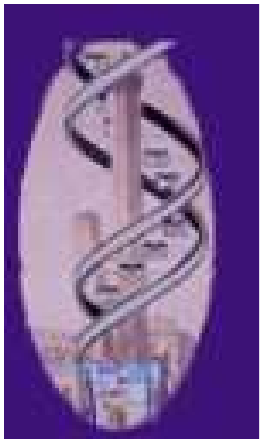
C(t), Sn(t) and Sp(t) are reported for the transmembrane segments (t).
Vit = Viterbi decoding, Pvit = posterior-Viterbi decoding.
For GRHCRF and HMM-B2TMR we used the posterior-Viterbi decoding.
Models are detailed in the text. Scoring indices are described in Measure of Accuracy section.

Fariselli et al, BMC ALMOB 2009

5) Prediction of subcellular localisation and disulfide bonds

DisLocate

Find Disulfide bonds in Eukaryotes with predicted subcellular Localization



Prediction

Method

Please, choose the organism's kingdom (this is required by the BaCelLo predictor used for subcellular localization prediction):

- (•) Animals
- ( ) Fungi
- ( ) Plants

Please, paste a protein sequence in FASTA format (only one sequence at a time is allowed):

Based on Bacello and GRHCRFs

BIOINFORMATICS **ORIGINAL PAPER** *Vol. 27 no. 16 2011, pages 2224–2230*
doi:10.1093/bioinformatics/btr387

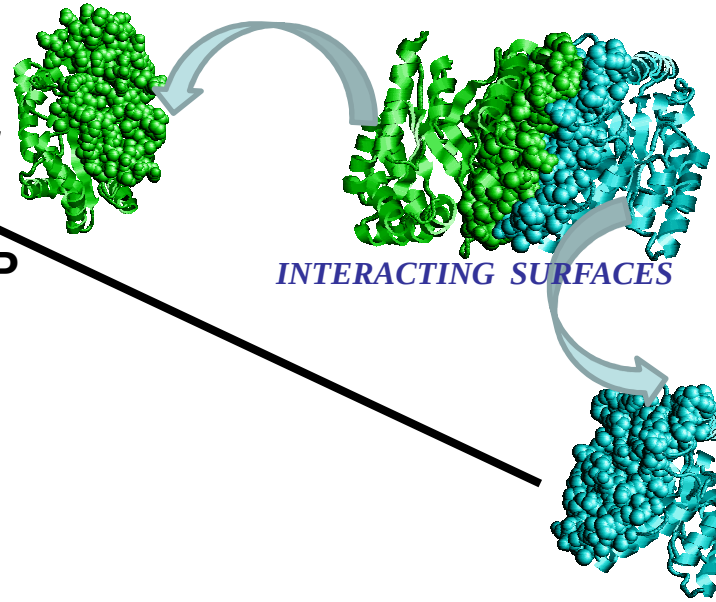
Structural bioinformatics Advance Access publication June 29, 2011

Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization

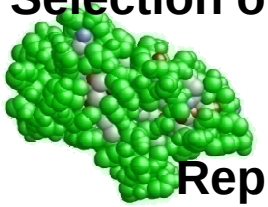
Castrense Savojardo[1,2], Piero Fariselli[1,2,*], Monther Alhamdoosh[1], Pier Luigi Martelli[1], Andrea Pierleoni[3] and Rita Casadio[1]

6)The protein-protein interaction sites

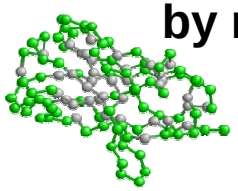
PDB-coordinates



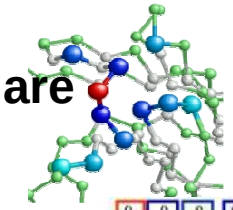
Selection of exposed residues with the DSSP



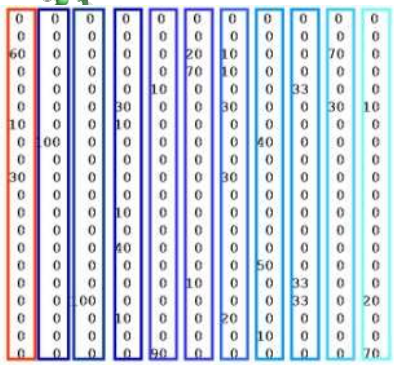
Representation of the protein by means of Ca atoms



For each exposed Ca the 10 closest exposed Ca's are selected within 1.2 nm



For each selected Ca the corresponding column of th sequence profile is extracted and relative solvent accessibility is computed



HMM-SVM predictor

Our new solution:
Hidden Markov Support Vector Machines for PPI

Altun et al, ICML 2003:

- A discriminative learning technique combining HMMs and SVMs

1. Dependencies between neighboring labels are handled by Viterbi decoding

2. The learning procedure is discriminative and based on a maximum/soft margin criterion

3. Learning is kernel-centric

Table 1. Cross-validation performance of the different methods

Method	Q2 (%)	Sp (%)	Sn(%)	MCC(%)	Encoding
NN	64	55.1	77	31	(Profile+RSA)
NN	66	62	84	35	(PSSM+RSA)
NN	69	65	82	39	(PSSM+dSA)
NN	69	65	85	40	<(PSSM+dSA)>*
HM-SVM	68	70	65	36	(Profile+RSA)
HM-SVM	70	72	66	40	(PSSM+RSA)
HM-SVM	71	73	67	42	(PSSM+dSA)
PredPPI	71	73	68	43	<(PSSM+dSA)>*

NN= neural network as described in Fariselli et al. 2002. HM-SVM=Hidden Markov Support Vector Machine. Q2=overall accuracy. Sp= interaction site specificity. Sn=interaction site sensitivity. MCC=Mattews' correlation coefficient. Profile=sequence profile. PSSM=position specific score matrix as computed by BLAST. *=prediction averaged over a window of 11 neighbors.

PredPPI at work

Table 2. Comparison with other recent methods

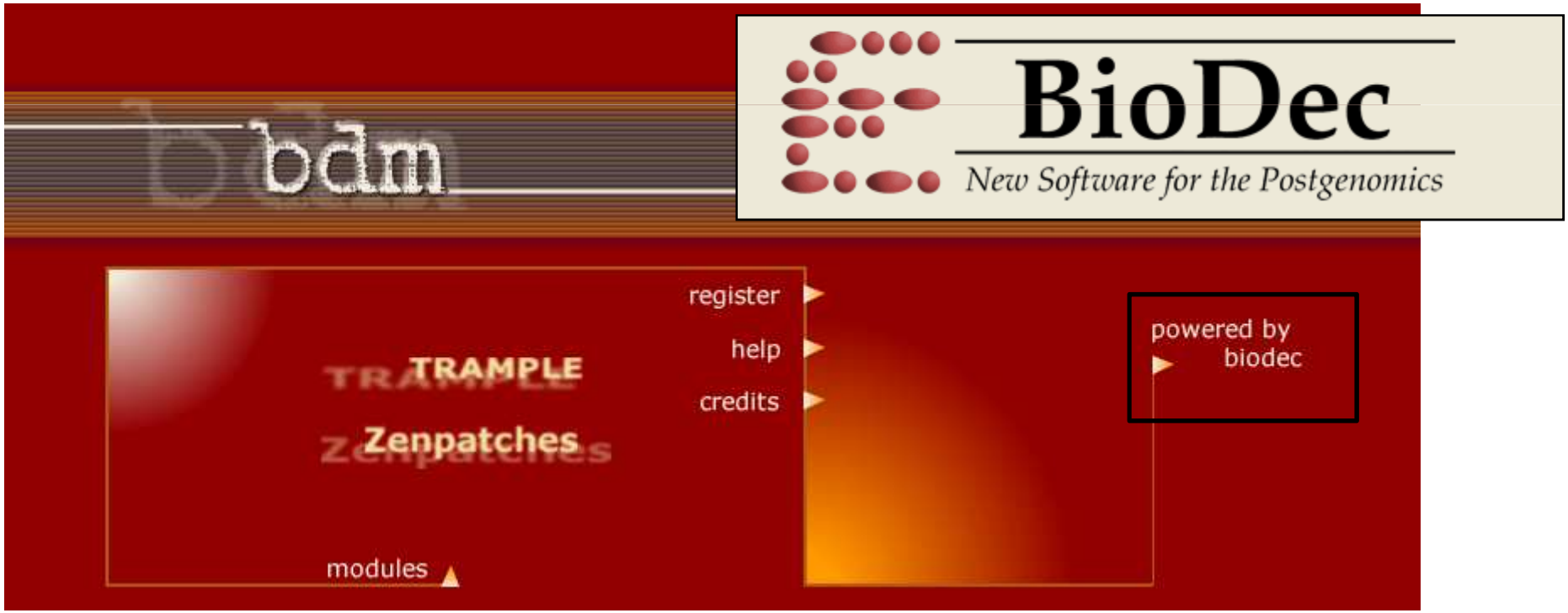
Method	Q2 (%)	Sp (%)	Sn(%)	MCC(%)	F1(%)
Wang et al. 2006	NA	65	69	28	67
Nguyen-Rajapakse 2006	NA	93	36	33	52
Deng et al. 2009	NA	63.	77	35	69
Liu et al. 2009 *	69	52	59	33	55
PredPPI *	71	73	68	43	71

*= On the same data set. The reported performances are taken from the corresponding papers (with the exception of PredPPI).
NA=not available.

$$F1 = 2\,(Sp(s) * Sn(s)) \,/\, (Sp(s) + Sn(s))$$

The ZenPatches database: A PDB Supplement Annotated for Putative Protein-Protein Interaction Patches

http://gpcr.biocomp.unibo.it/biodec/
www.biodec.com



Patches groups of predicted protein contacts

Protein Information 1a1r (chain: AB)

Title: HCV NS3 PROTEASE DOMAIN:NS4A PEPTIDE COMPLEX

Classification: VIRAL PROTEIN

Each Patch is obtained grouping the predicted contacts according to a C-αlpha distance threshold of 4.5 Angstrom.

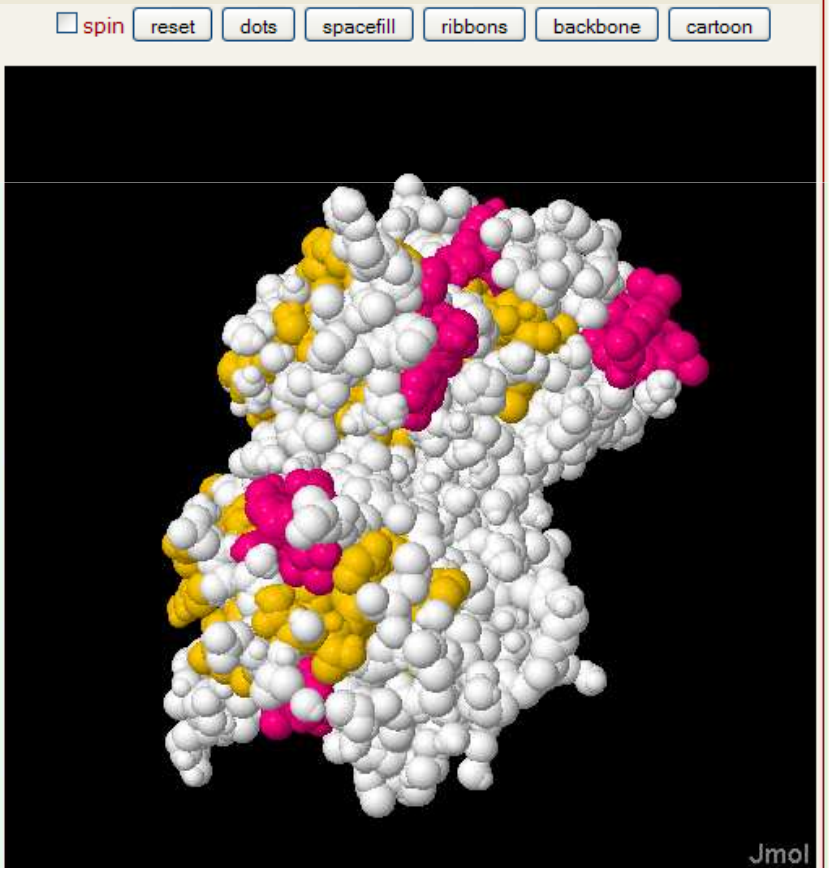
Source: MOL_ID: 1; ORGANISM_SCIENTIFIC: HEPATITIS C VIRUS; ORGANISM_COMMON: HCV; STRAIN: H; GENE: NS3; EXPRESSION_SYSTEM: ESCHERICHIA COLI; EXPRESSION_SYSTEM_STRAIN: JM109; EXPRESSION_SYSTEM_CELLULAR_LOCATION: CYTOPLASM; EXPRESSION_SYSTEM_PLASMID: PET-BS(+); MOL_ID: 2; SYNTHETIC: BIOLOGICAL SEQUENCE; ORGANISM_SCIENTIFIC: HEPATITIS C VIRUS; ORGANISM_COMMON: HCV; STRAIN: H; GENE: NS4A

PDB (file): 1a1r.pdb -- **PDB (@rcsb.org):** 1a1r

Patches

Patch 1 on off			
35 GLN B (1)	36 THR B (1)	37 ARG B (2)	**38 GLY B (3)**
39 LEU B (4)	**40 LEU B (3)**	**41 GLY B (3)**	**42 CYS B (3)**
43 ILE B (3)	44 ILE B (1)	46 SER B (0)	47 LEU B (0)
Patch 2 on off			
154 SER B (0)	155 PRO B (1)	156 ARG B (2)	157 PRO B (2)
185 CYS A (2)	186 THR A (2)	187 ARG A (2)	188 GLY A (2)
189 VAL A (2)	191 LYS A (1)		
Patch 4 on off			
121 THR B (0)	122 PRO B (1)	**123 CYS B (3)**	**124 THR B (3)**
125 CYS B (2)	126 GLY B (1)	172 PRO B (2)	**173 THR B (3)**
174 GLY B (4)	**175 HIS B (3)**		
Patch 5 on off			
100 MET A (1)	**101 TYR A (3)**	**102 THR A (3)**	103 ASN A (2)
104 VAL A (2)	105 ASP A (2)	106 GLN A (0)	
Patch 6 on off			
100 MET B (2)	101 TYR B (2)	102 THR B (2)	103 ASN B (1)
104 VAL B (1)	105 ASP B (1)	106 GLN B (0)	
Patch 7 on off			
133 VAL A (1)	137 ALA A (0)	138 ASP A (1)	139 VAL A (2)
Patch 8 on off			
131 TYR A (1)	141 PRO A (1)		

High-confidence residues (R=3-9) are highlighted in graphics.



BaCelLo
Balanced subCellular Localization predictor

SNPs&GO

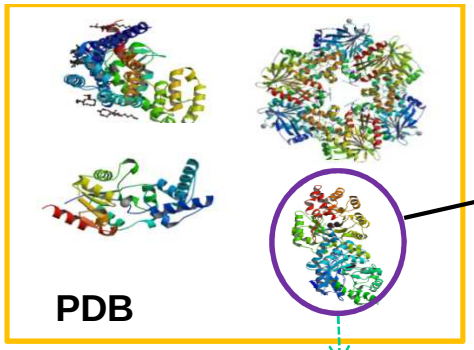
BAR

Alternative splicing variant annotation

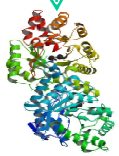
http://t.caspur.it/ASPicDB/
Martelli et al., 2011, NAR

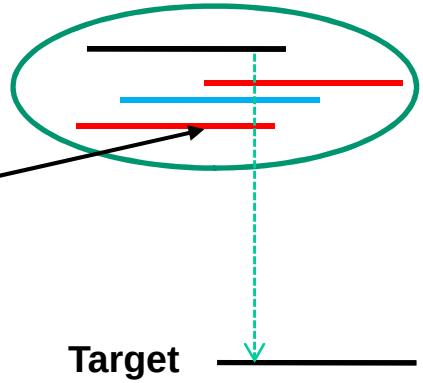
The Bologna functional annotation resource

www.biocomp.unibo.it



Template





NN,HMM, SVM, GRHCRF

Prediction of:
Secondary structure
Disulfide bonding state
Solvent accessibility
Active sites, Protein protein interaction sites, protein DNA/RNA interactions, GPI anchor
Functional domains

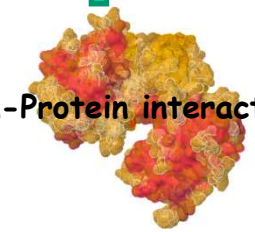
Feature-driven alignment

Modelling

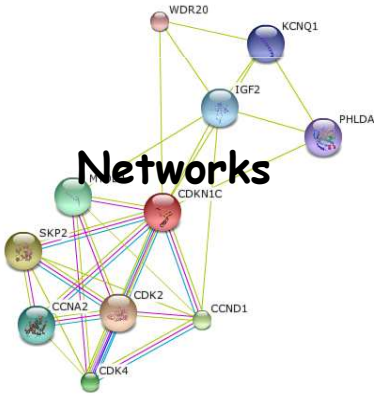
Check of model stereochemistry

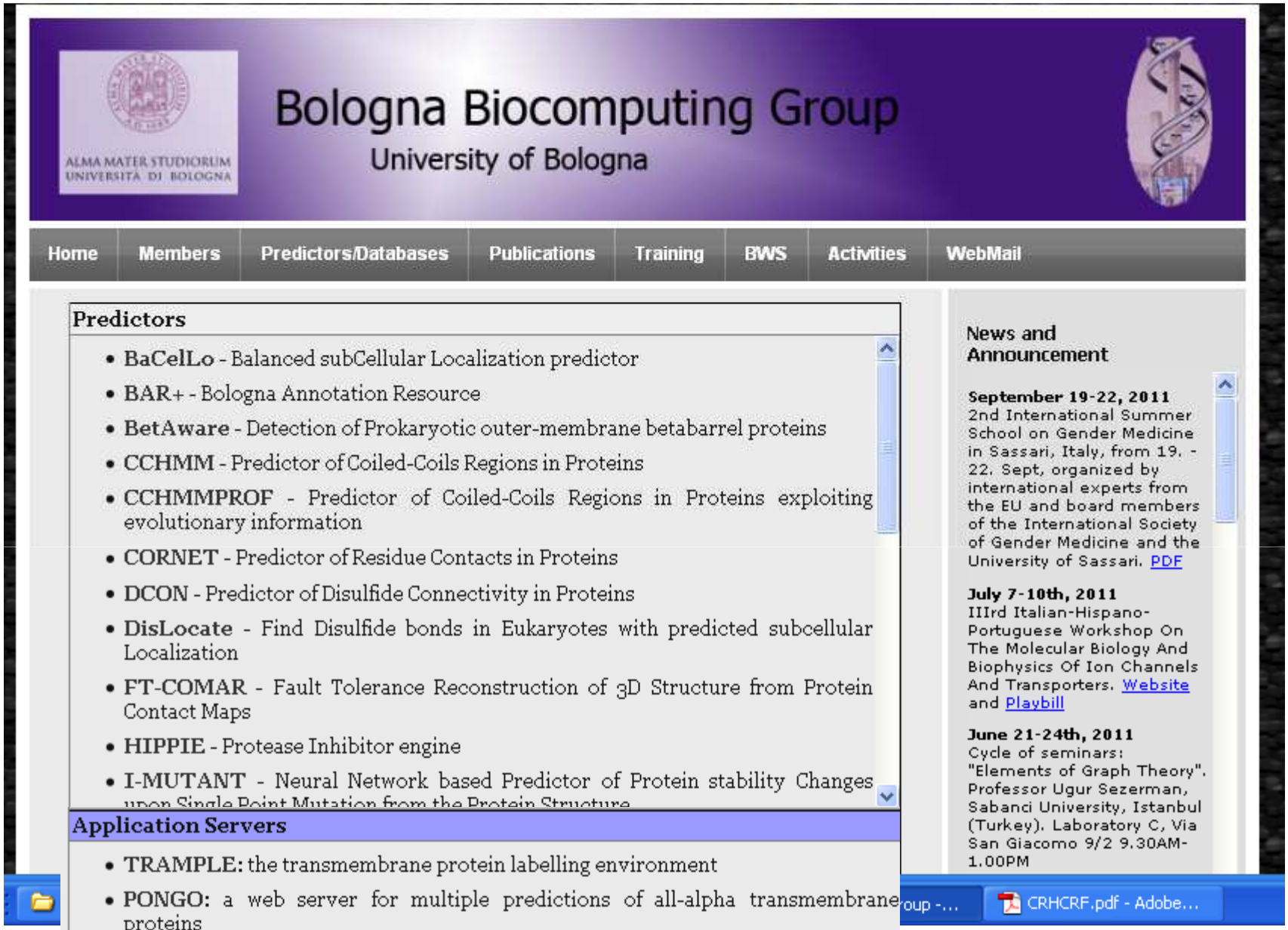
Model

Structural location of variation



Protein-Protein interaction




ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
Bologna Biocomputing Group
University of Bologna
Home
Members
Predictors/Databases
Publications
Training
BWS
Activities
WebMail
Predictors
BaCelLo - Balanced subCellular Localization predictor
BAR+ - Bologna Annotation Resource
BetAware - Detection of Prokaryotic outer-membrane betabarrel proteins
CCHMM - Predictor of Coiled-Coils Regions in Proteins
CCHMMPROF - Predictor of Coiled-Coils Regions in Proteins exploiting evolutionary information
CORNET - Predictor of Residue Contacts in Proteins
DCON - Predictor of Disulfide Connectivity in Proteins
DisLocate - Find Disulfide bonds in Eukaryotes with predicted subcellular Localization
FT-COMAR - Fault Tolerance Reconstruction of 3D Structure from Protein Contact Maps
HIPPIE - Protease Inhibitor engine
I-MUTANT - Neural Network based Predictor of Protein stability Changes
Application Servers
TRAMPLE: the transmembrane protein labelling environment
PONGO: a web server for multiple predictions of all-alpha transmembrane proteins
Databases
eSLDB - eukaryotic Subcellular Localization DataBase
ZenPatches - Database of predicted protein interaction sites
DBMFHS - Data Base of Minimally-Frustrated Helical Segments
News and Announcement
September 19-22, 2011
2nd International Summer School on Gender Medicine in Sassari, Italy, from 19. - 22. Sept, organized by international experts from the EU and board members of the International Society of Gender Medicine and the University of Sassari. PDF
July 7-10th, 2011
IIIrd Italian-Hispano-Portuguese Workshop On The Molecular Biology And Biophysics Of Ion Channels And Transporters. Website and Playbill
June 21-24th, 2011
Cycle of seminars: "Elements of Graph Theory". Professor Ugur Sezerman, Sabanci University, Istanbul (Turkey). Laboratory C, Via San Giacomo 9/2 9.30AM-1.00PM
CRHCRF.pdf - Adobe...

Bologna Biocomputing Group

University of Bologna

Group Leader

Rita Casadio

Senior Researchers

Piero Fariselli
Pier Luigi Martelli

Technical Staff

Gianluca Tasco

Post-Doc

Pietro Di Lena
Marco Vassura

External Collaborators

Alberto Eusebi (BioDec)
Andrea Pierleoni (Externautics)
Ivan Rossi (BioDec)

PhD Students

Raffaele Fronza
Damiano Piovesan
Deepak Rajan
Castrense Savojardo
Valentina Indio

Former Members

Lisa Bartoli
Remo Calabrese
Emidio Capriotti
Paola Marani
Ludovica Montanucci
Priyank Shukla
Shalinee Tiwari

News and Announcement

July 7-10th, 2011
IIIrd Italian-Hispano-Portuguese Workshop On The Molecular Biology And Biophysics Of Ion Channels And Transporters. Website and Playbill

14-18 Febbraio 2011
Seminario Scienza e Filosofia - Indagare la Mente: chi sono io?

February 7-11 2011
Bologna Winter School 2011: Deep Sequencing Data Analysis: a challenge for personalized medicine

10/12/2010 The **CAGI-2010** dancing award of SNPs&GO

22-23/09/2010 Seminar: Where does sequence information come from? Native state, function and evolution of proteins. Prof. Gustavo Parisi, Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes, Argentina. September 22 and 23 10:00-1p.m. Aula C - Via San

Acknowledgements..

BioDec
New Software for the Postgenomics

Web Predictors at http://www.biocomp.unibo.it

CORNET/ Predictor of Residue Contacts in Proteins

CYSPRED/ Predictor of Bonding State of Cysteines in Proteins

DCON/ Predictor of Disulfide Connectivity in Proteins

RCNPRED/ Predictor of Residue Coordination Numbers in Proteins

SECPRED/ Predictor of Secondary Structures of Proteins

ISPRED/ Predictor of Protein Interaction Sites

I-MUTANT/ Predictor of Protein Stability Changes upon Single Point Mutation

B2TMR/ Predictor of Transmembrane Topology of all-beta Membrane Proteins (Neural Network-based)

HMM-B2TMR/ Predictor of Transmembrane Topology of all-beta Membrane Proteins (Hidden Markov Model-based)

HTMR/ Predictor of Transmembrane Topology of all-alpha Membrane Proteins (NN based)

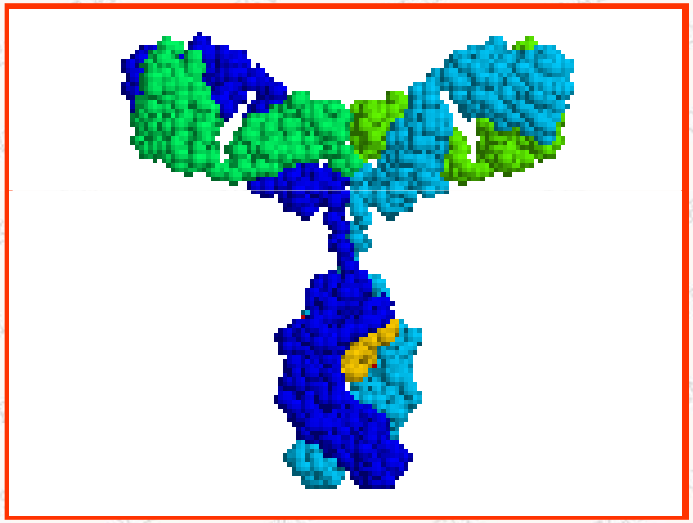
SPEPLip/ Predictor of Signal Peptide and Lipoprotein Cleavage Sites in Proteins

Curated Databases at http://www.biocomp.unibo.it

ZenPatches/ Predicted Protein-Protein Interacting Surface Patches in collaboration with BioDec (www.biodec.com)

TOPDB/ Proteins from thermophiles and their counterparts in mesophiles

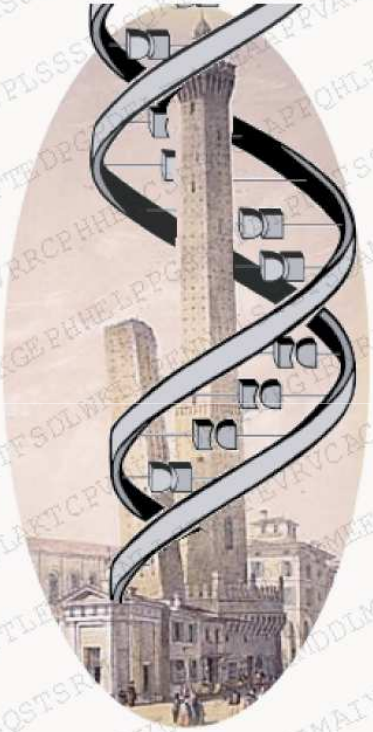
The research interests of **The Biocomputing Group** focus on different aspects of protein sequence analysis, mainly the implementation of predictive algorithms based on different machine learning approaches. The Biocomputing Group organises International Advanced Schools on Bioinformatics at the Bologna University since 1999.
The Biocomputing Group is presently also:
> Node of the BioSapiens Network of Excellence
> Italian node of the FIRB project "Bioinformatics for Genomics and Proteomics"
> Member of the European Computational Biology Community (ECCB)



The **Biocomputing Group** of the University of Bologna

Rita Casadio, Piero Fariselli, Pier Luigi Martelli, Gianluca Tasco, Andrea Pierleoni, Raffaele Fronza, Alberto Eusebi, Damiano Piovesan Ivan Rossi, Michele Finelli, Mauro Amico, Andrea Zauli.

The Biocomputing Group of the University of Bologna



> Protein Biosequence Analysis
> Protein Modelling
> Prediction of
  - Protein Secondary Structure
  - Contact Map
  - Protein Folding
  - Bonding State of Cysteines
  - Topology of Disulfide Bridges
  - Protein Stability upon Mutation

Protein and Peptide Design
Molecular Docking
Molecular Dynamics