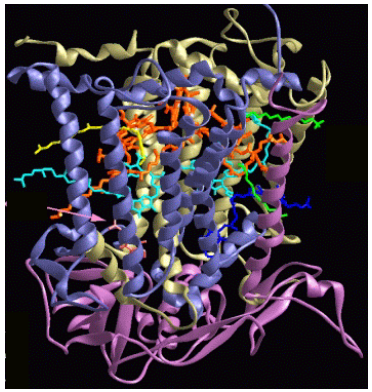
Bioinformatics Tools for Characterization of Protein-Protein Interactions



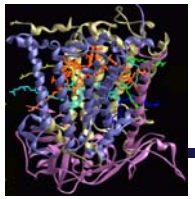
Nagiza F. Samatova

Computer Science & Mathematics Division

Oak Ridge National Laboratory

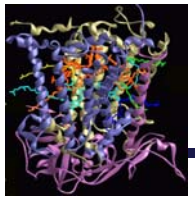
This work is funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Cray X1 Planning – May 9, 2003



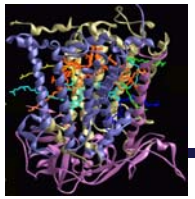
Our Team

- Geist, Al
- Gorin, Andrey
- Chandramohan, Praveen
- Munavalli, Rajesh
- Ostrouchov, George
- Park, Hoony
- Samatova, Nagiza – Project Lead
- Yu, Gong-Xin



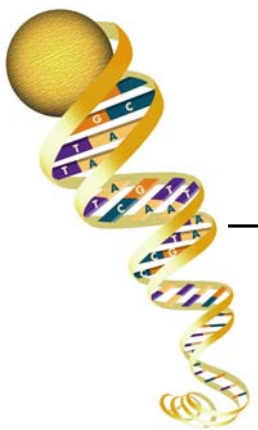
Importance of Protein Interactions

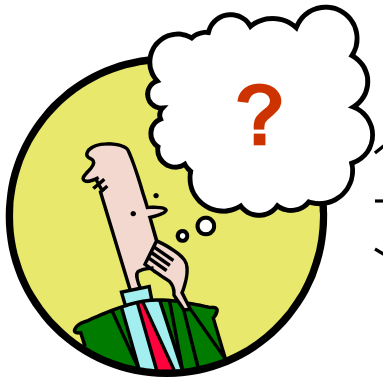
- Protein-Protein interaction is essential in most cellular processes.
- They are basis of phenomena like:
  - DNA replication and transcription
  - Metabolism, signal transduction, etc.
- Understanding role of proteins is essentially about discovering its biological context.
- Discovering proteins that interact in a cell is key to understand its functional networks.
- Used to predict a function of a protein

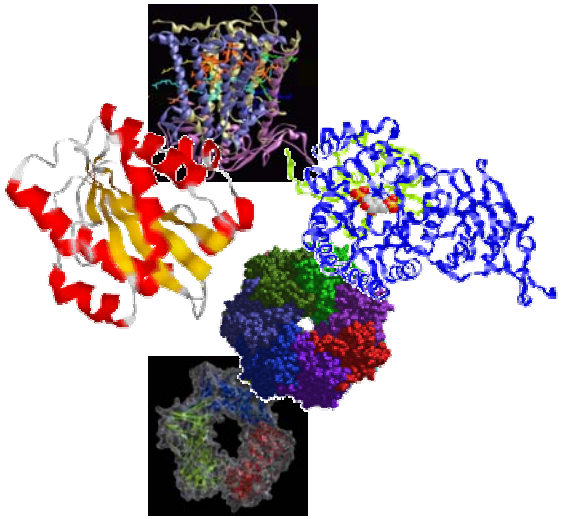


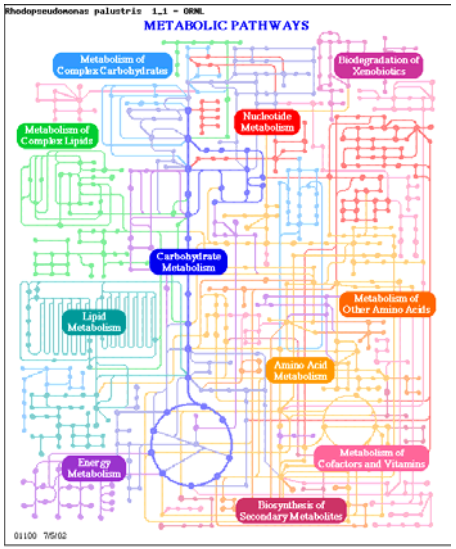
Characterization of P-P Interactions

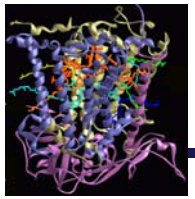
- Who? – Identify interacting proteins
- Where? – Identify interface residues
- How? – Characterize interaction dynamics
- Why? – Characterize a function





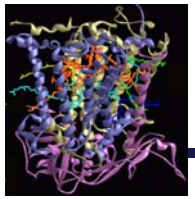






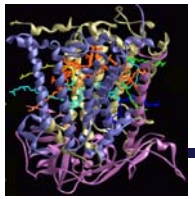
W^3 Goals of the Project

- Develop tools and algorithms that will increase precision and reliability of high-throughput characterization of protein complexes:
  - Prediction of interacting proteins (Who)
  - Identification of interface residues (Where)
  - Automated assignment of functions to genes/complexes utilizing various information sources (sequence, structure, biochemical properties) (Why)



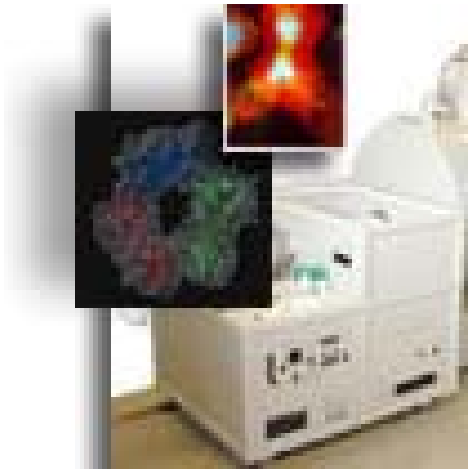
The Outline

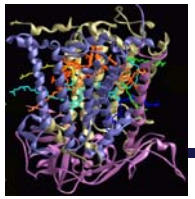
- **PICCUP** for protein-protein interaction prediction (Who Problem)
- **SVMMER** for high-resolution functional characterization of proteins (Why Problem)
- **KeyGeneMiner** for identification of photosynthesis-specific genes (Why Problem)



High-Throughput Experimental Methods

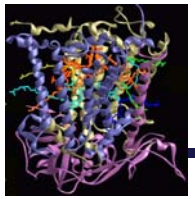
- Y2H: Yeast two-hybrid system (Fields et al. 1989)
- Protein arrays (Finley et al. 1994; Zhu et al. 2001)
- Phage display (Rodi et al. 1999)
- Mass spectrometry (Gavin et al. 2002; Ho et al. 2002).
- Using these techniques, the protein interaction networks for a few simple cellular systems are obtained.
  - Yeast (Ito et al. 2000; Uetz et al. 2000)
  - Helicobacter pylori (Rain et al. 2001)





Computational Methods

- Genomic context-based methods:
  - Gene fusion events (Marcotte et.al., 1999; Enright et.al., 1999)
  - Conservation of gene-order or co-occurrence of genes in potential operons (Tamames et.al.,1997; Dandekar et.al.,1998; Overbeek et.al., 1999)
  - Presence/absence of genes in different species (phylogenetic profiles) (Pelligrini et.al., 1999)
  - Similarity of phylogenetic trees (Pazos and Valencia, 2001)
- Classification-based methods:
  - SVM-based (Bock and Gough, 2001)
- Statistical methods:
  - PICUPP: **P**rotein **I**nteraction **C**lassification by **U**nlikely **P**rofile **P**air



Genomic context-based methods

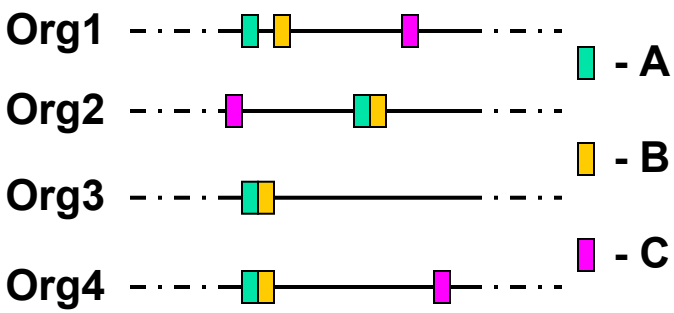
Phylogenetic Profiles

Protein	A	B	C	D
Org1	1	1	1	1
Org2	0	1	0	1
Org3	1	0	1	0
Org4	1	0	1	1

A ↔ C

Idea: Pairs of proteins w/ similar phylogenetic profiles

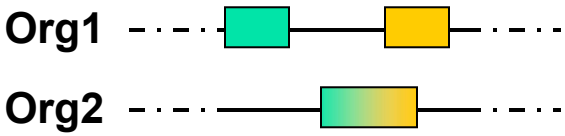
Conservation of gene-order



A ↔ B

Idea: Proteins whose genes are co-located in multiple genomes

Gene fusion



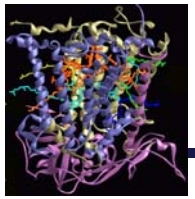
A ↔ B

Idea:If two proteins form a part of a single protein in other orgs.

Advantages: Allow to identify functionally associated genes

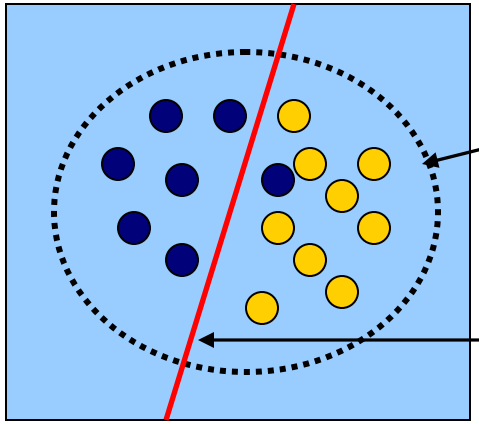
Disadvantages: Provide a small coverage of direct physical interactions (~30%, Huynen et.al., 2000)

Coverage Statistics: 37%: Gene order; 6%: Gene fusion; 11%: Phylogenetic profiles (*M.genitalium*, Huynen et.al., 2000)



Classification-based methods

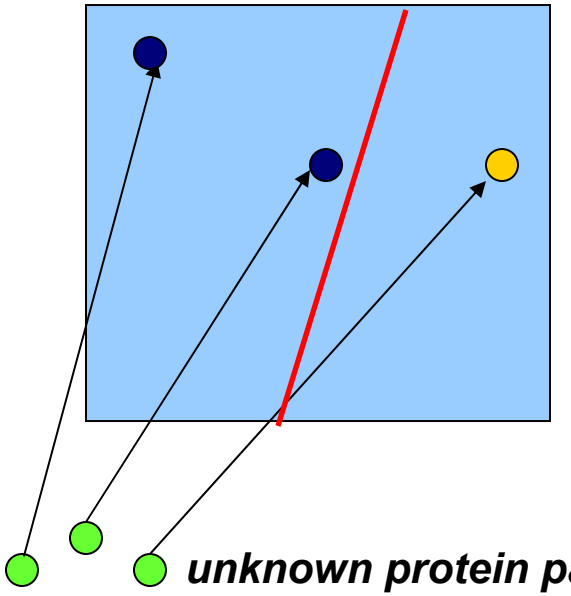
Training Stage



training data

decision rule:
separating hyperplane (SVM)

Prediction Stage

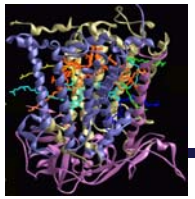


unknown protein pairs

- positive (interacting protein pairs)
- negative ("non-interacting" protein pairs) – generated by shuffling AA residues

Limitations:

- Only positive data is available for protein interactions
- Ill-defined negative set: Shuffled protein may not belong to protein space
- Asymmetry: AB or BA protein pairs correspond to different points in space

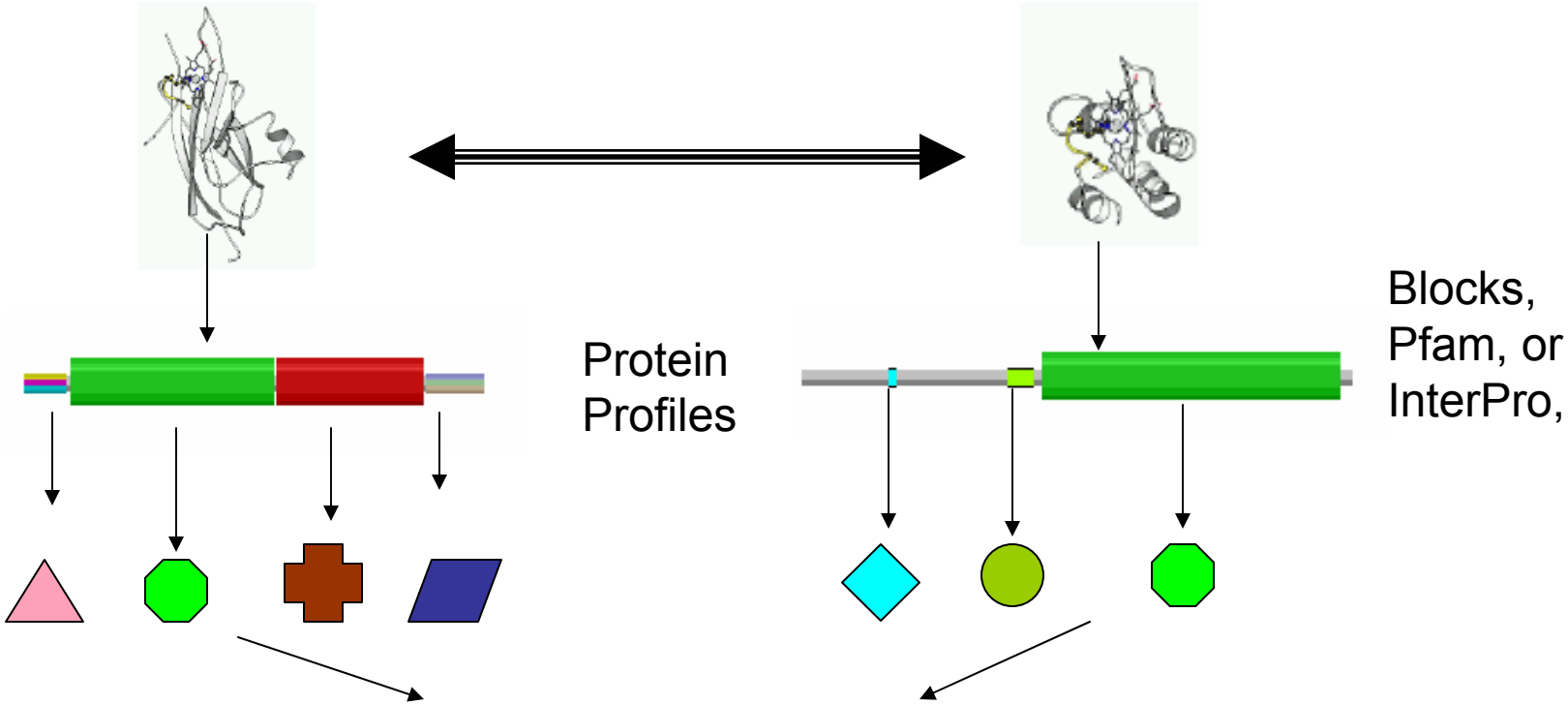


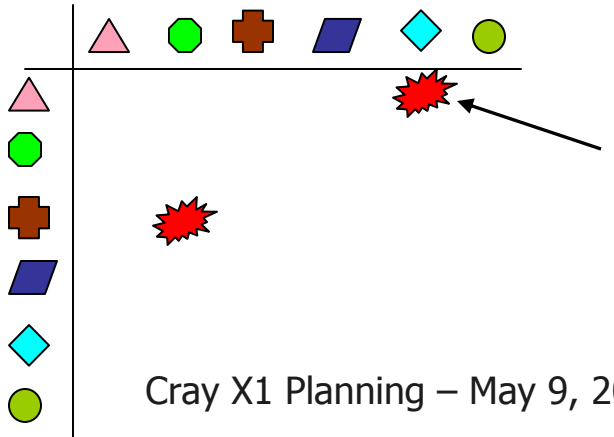
Our Approach

Learning by Statistical Simulation

- Given a set of positive interactions only, *statistical approach* is a natural candidate.
- **PICCUP**:
  - Seeks to find statistically ***unusual*** protein profile pairs defined by Blocks, InterPro, Pfam, or Prosite.
  - Given protein interaction data, it estimates unusualness of profile pairs using statistical simulation (via bootstrapping).
  - It computes how apart a profile pair is from what is expected at random. The degree of separation is essentially a ***statistical confidence***.

Profile Pair Correlated for Protein Interaction

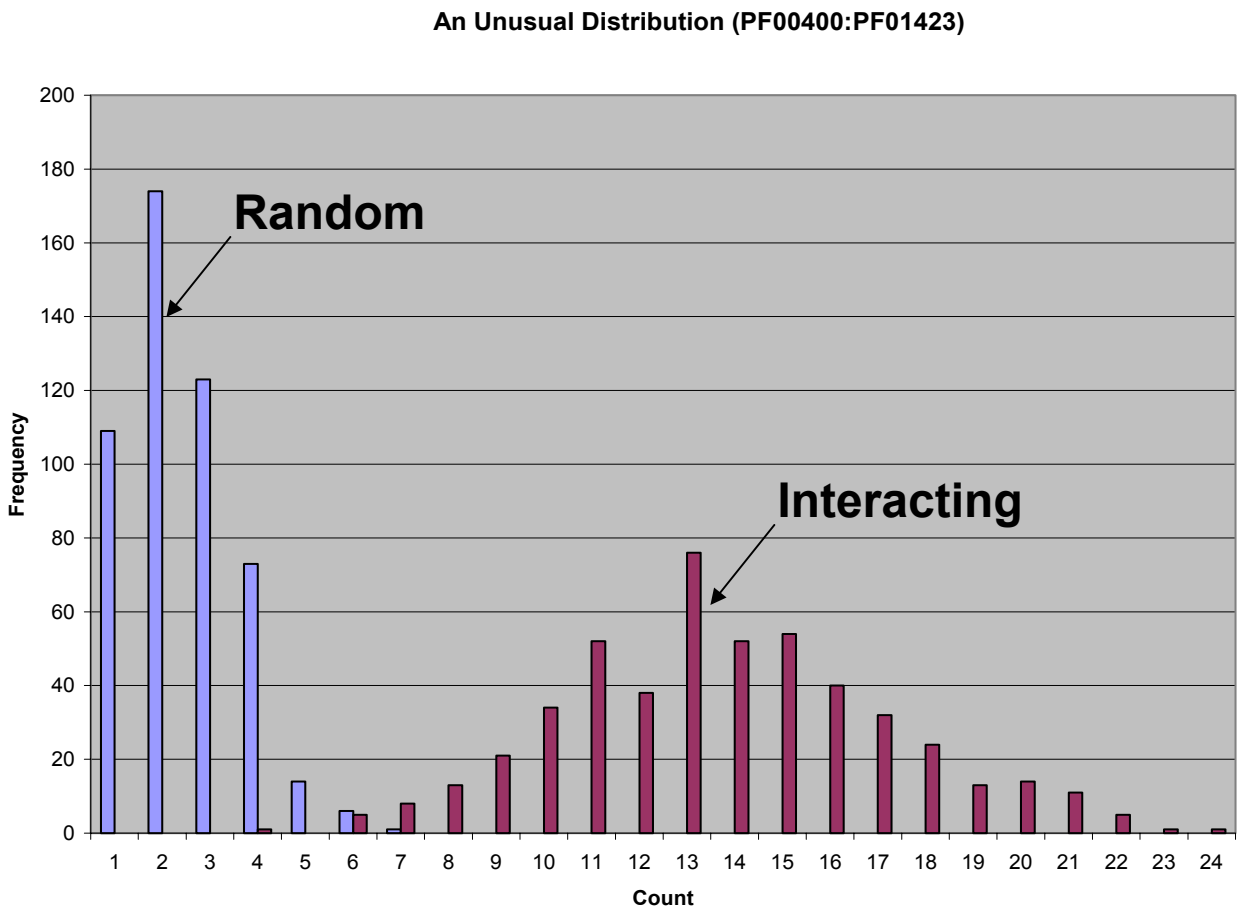




Occurrence of profile pair in DB of interacting proteins

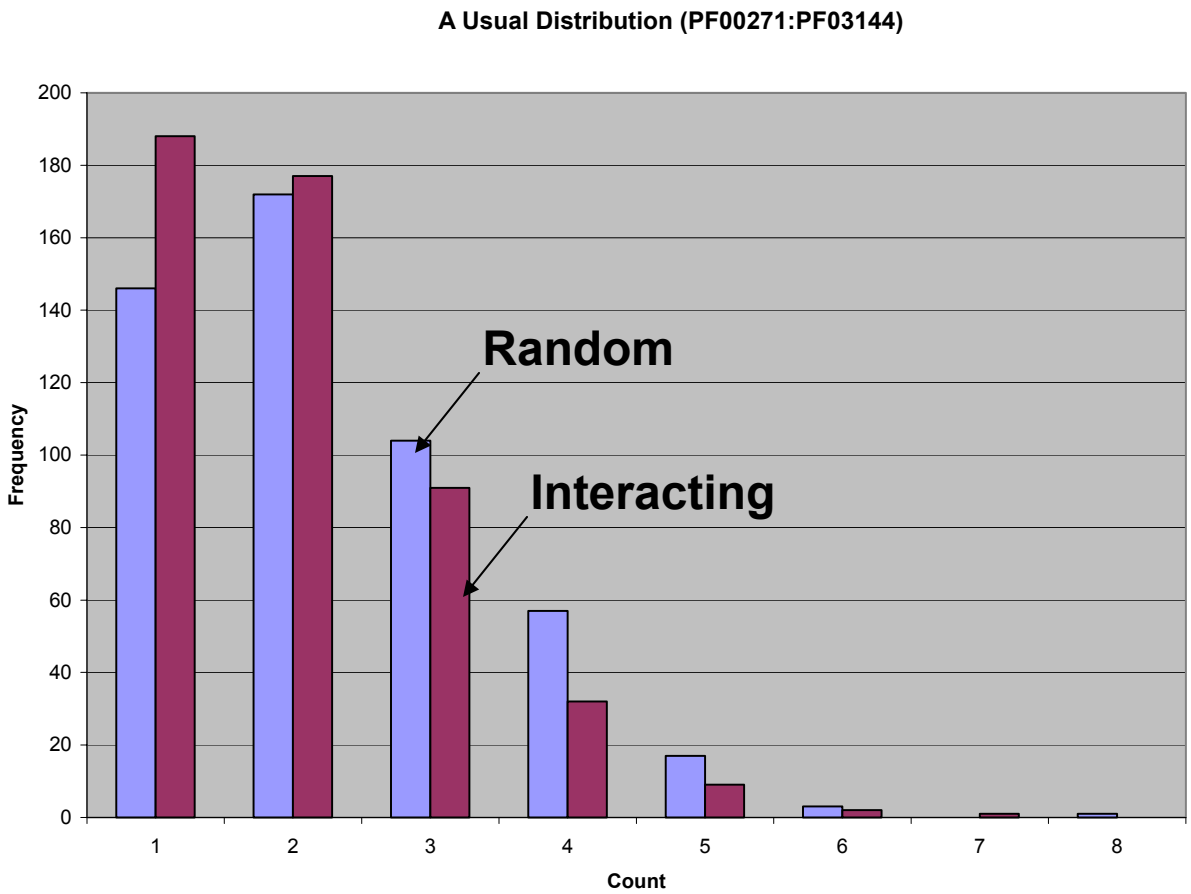
Unusual Profile Pair

Pfam pair: PF0400:PF01423

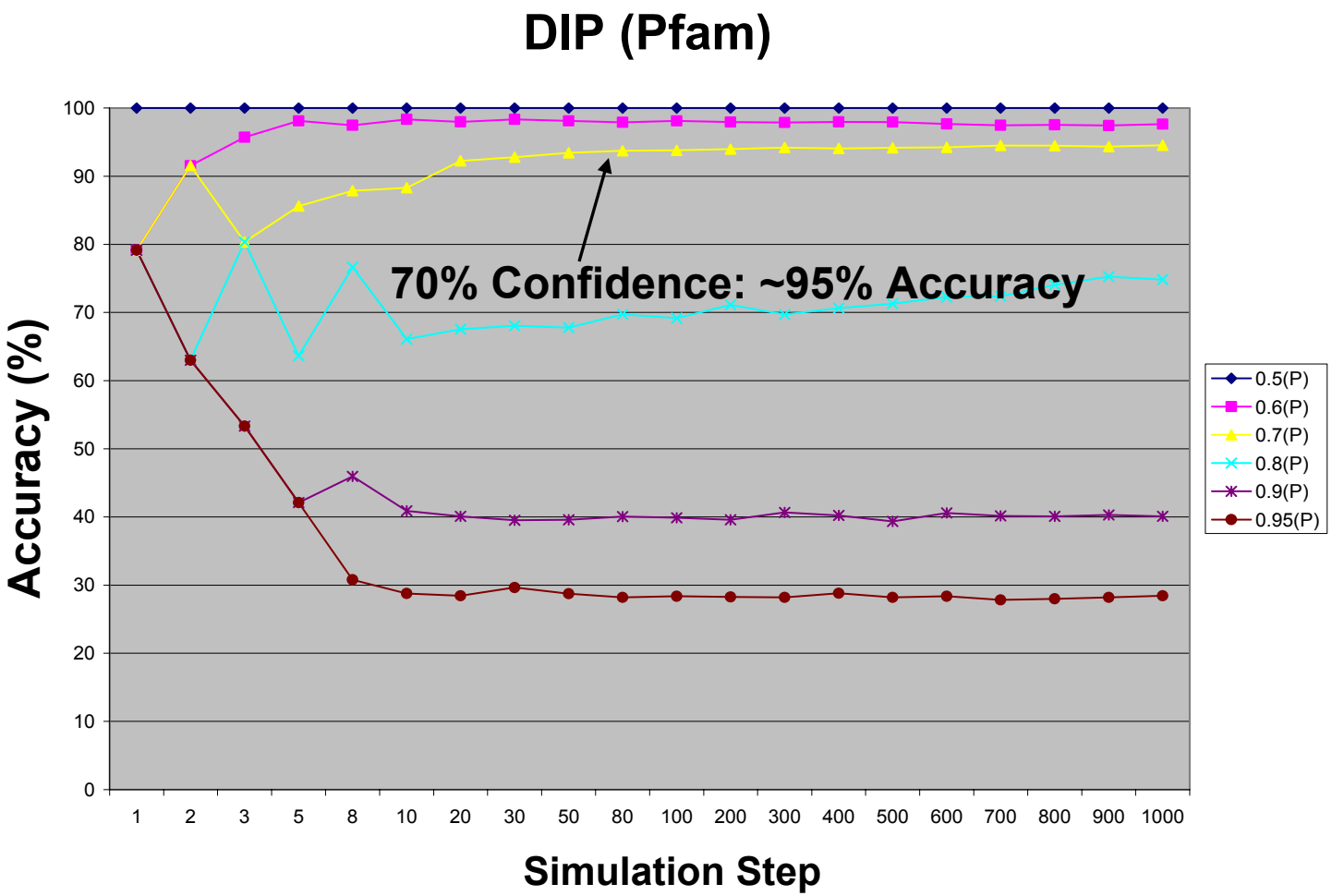


Usual Profile Pair

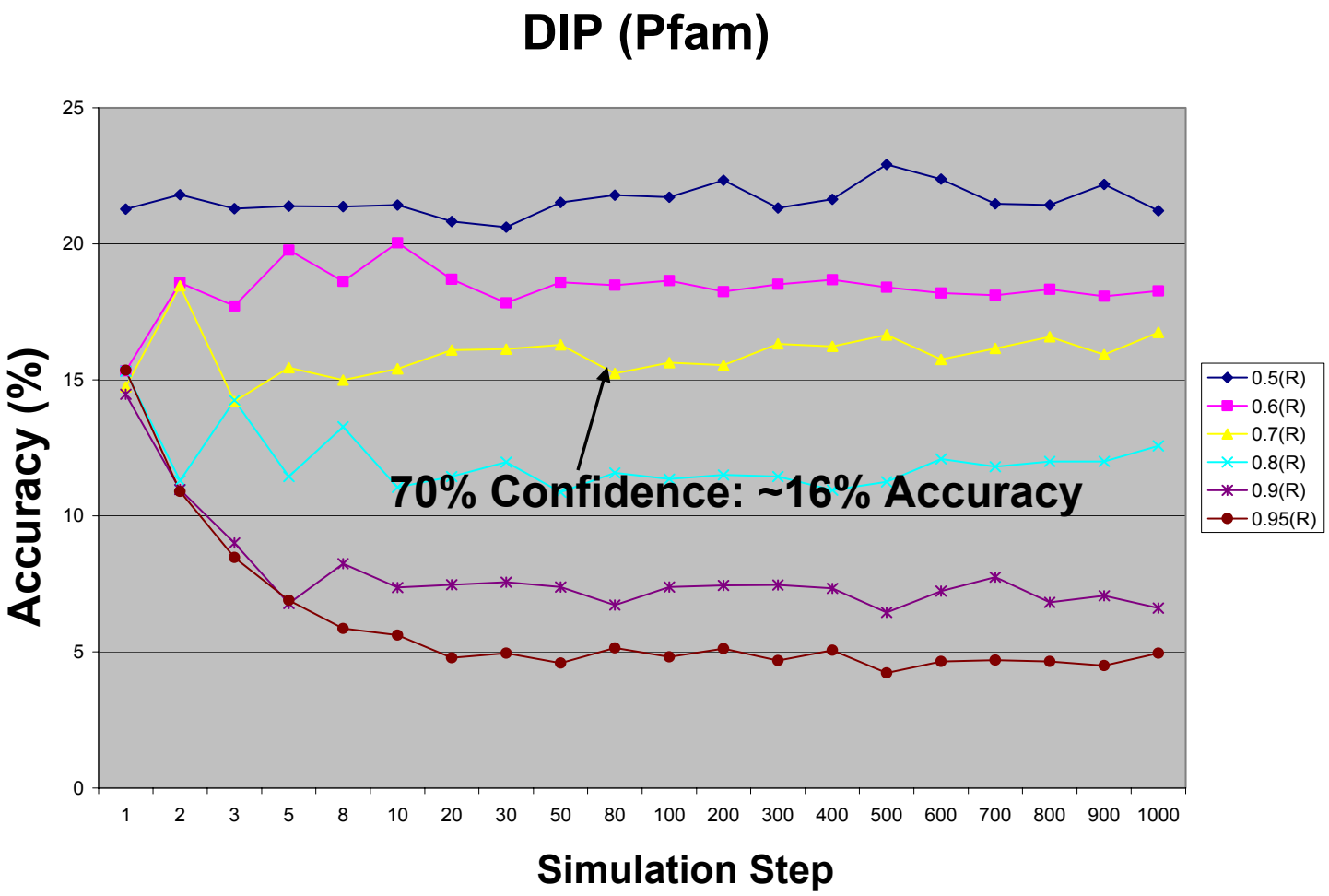
Pfam pair: PF00271:PF03144

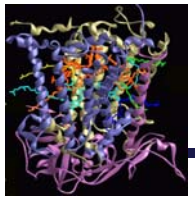


Sensitivity To Positive Pairs (Pfam)



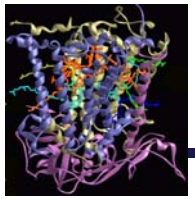
Sensitivity To Random Pairs (Pfam)





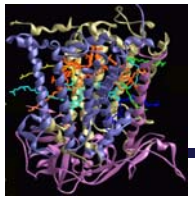
How Unusual the Unusual Domains Are

- **~108** pairs of Pfam domains are identified at **98%** confidence
- A case-by-case investigation discovered many well-established Pfam-Pfam associations among these interactions, including:
  - G-protein beta WD-40 repeat ↔ Small nuclear ribonucleoprotein (Achsel et al. 1998)
  - Actin ↔ Cofilin/tropomyosin-type actin-binding protein (Nishida et al. 1984)
  - Protein kinase domain ↔ Fibroblast growth factor (Taniguchi et al. 2003)
  - EF hand ↔ Myosin head (motor domain) (Messer and Kendrick-Jones 1991)



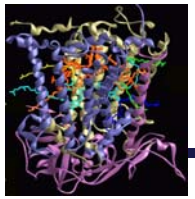
Application to *Synechococcus Sp*. WH8102

- **1479** interacting pairs predicted at **98%** confidence
- A case-by-case study of these interactions coupled with other methods is under way



Summary: PICUPP

- PICUPP is statistical approach to sift out profile pairs that account for protein interactions.
- Empirical studies show positive results.
- Selecting optimal similarity threshold is very important.
- Too sparsely represented pairs in DB are still problematic. Cluster similar profiles is worth to consider.



From Genes to Protein Functions

The first and most crucial step in systems biology

Function assigned based on sequence similarity to another sequence with a

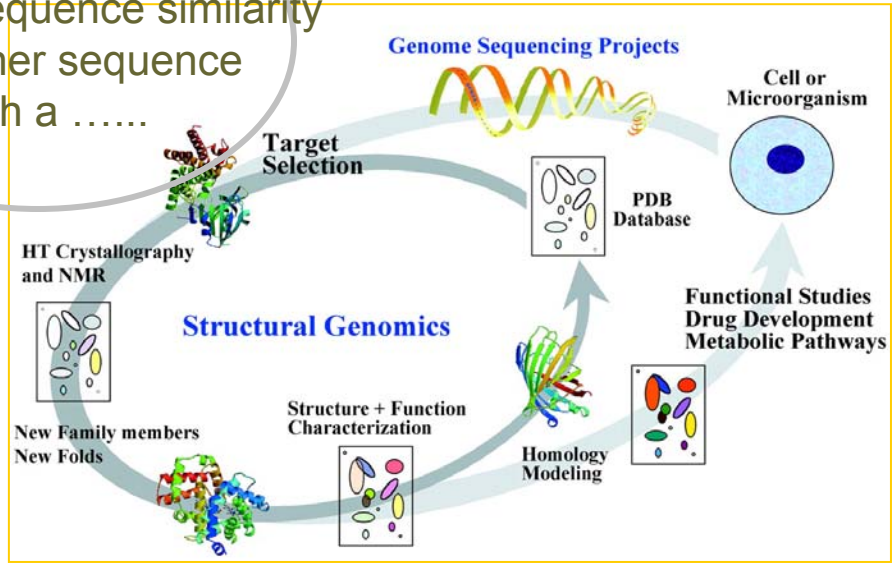
function assigned based on sequence similarity to another sequence with a

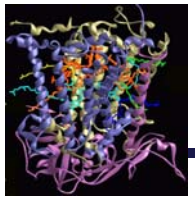
Function assigned based on sequence similarity to another sequence with a ……

Weakness of sequence annotations

"Guilty-by-association" goes global

- Among some orthologous genes in phylogenetically distantly related organisms, sequence similarities are no longer recognizable
- No sequence similarity existed at all for genes that have resulted from convergent evolution





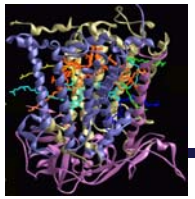
Resolution of Bioinformatics Tools is Insufficient to Discriminate Highly-Homologous Function Groups

Results of InterPro Analysis of four ADH Function Groups

Function	Gene_id	Pfam	PROSITE	Tigr-Fam
EC1.1.1.1	ADH1_TRIRP	PF00107 adh_zinc	PS00059 ADH_ZINC	tdh: L-threonine 3-dehydrogenase
	ADH3_SOLTU	same	same	same
EC1.2.1.1	FADH_CANM	same	same	same
	FADH_PICPA	same	same	same
EC1.1.1.195	CADH_EUCGL	same	same	same
	CAD1_ARACO	same	same	same
EC1.1.1.255	MTD1_ARATH	same	same	same
	MTDH_ARATH	same	same	same

1196 ORFs (out of 2522 ORFs) in Syn. 81902 that are conserved hypothetical or hypothetical (Brian Palenik)

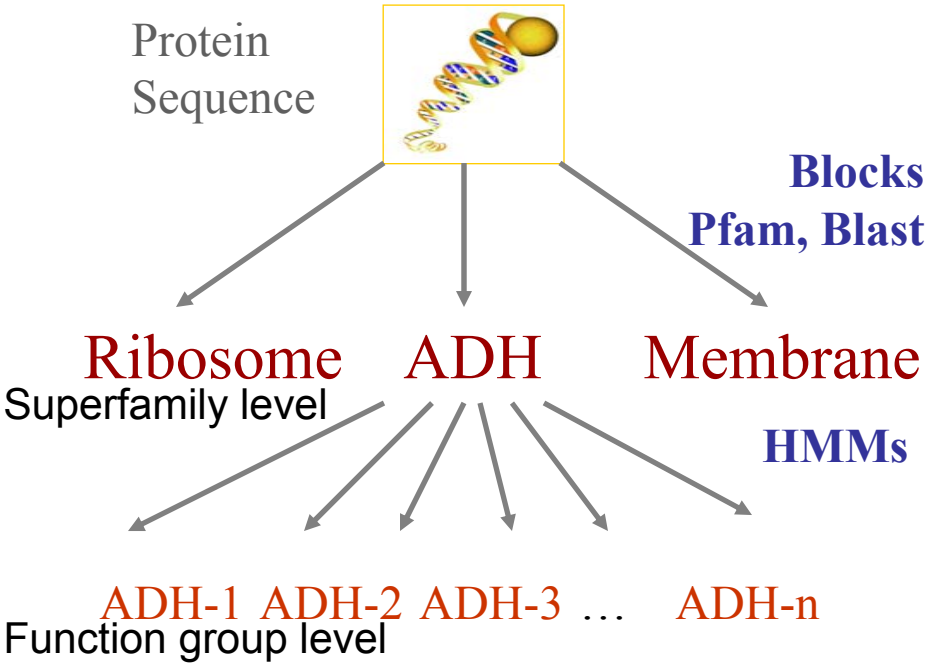
The current tools (e.g., BLOCKS, Pfam, PROSITE), while containing a wealth of information for characterization and identification of proteins, are unable to discriminate and classify closely related homologous sequences



Our Approach – SVMMER

Hierarchical High-Resolution Feature Extraction

(In collaboration with Natalia Maltsev, ANL)



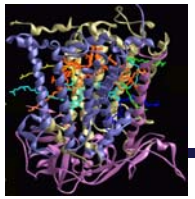
Approach:

- Using *a priori* knowledge, partition proteins into a hierarchy of classes.
- Extract features specific to each layer of the hierarchy
- Train a classifier on multi-layered hierarchy of classes

Technical Challenges:

- Classification algorithms (CA) require feature vectors of fixed length. Protein sequences are of variable length.
- Most CAs (SVM, NN) are binary classifiers. We deal with multi-class classification problem.
- CAs are designed for one layer of classes. We deal with multi-layered hierarchy of classes.

There is no "universal hammer" when dealing with such a complex concept as protein function



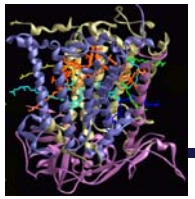
Performance Comparison for Different Classification Algorithms

Superfamily	Linear Kernel SVM			C4.5			RIPPER		
	Sp	*Sn*	*CC*	*Sp*	*Sn*	*CC*	*Sp*	*Sn*	*CC*
Aldehyde dehydrogenase	0.93	0.92	0.88	0.64	0.63	0.34	0.67	0.67	0.44
Phycobilisome	0.93	0.93	0.91	0.92	0.92	0.89	0.84	0.77	0.64
Cys/Met metabolism	0.94	0.94	0.87	0.61	0.56	0.16	0.55	0.56	0.03
Alpha-amylase signature	0.94	0.94	0.81	0.96	0.96	0.88	0.95	0.95	0.84
GMP synthase C terminal domain	1.0	1.0	1.0	0.95	0.95	0.89	0.95	0.95	0.89
AMP-dependent synthetase	0.88	0.84	0.84	0.89	0.87	0.85	0.79	0.73	0.72
NAD dependent epimerase	0.97	0.97	0.95	0.93	0.92	0.88	0.71	0.77	0.58
Polyprenyl synthetase	0.94	0.93	0.91	0.85	0.79	0.73	0.66	0.66	0.52
Zinc-containing ADH	0.94	0.94	0.93	0.83	0.87	0.79	0.77	0.73	0.66
Aminolevulinic acid synthase	1.0	1.0	1.0	0.96	0.96	0.94	0.86	0.83	0.74
Average	0.95	0.94	0.91	0.85	0.64	0.74	0.78	0.76	0.61

Classification Algorithm

***Sp*: Specificity; *Sn*: Sensitivity; *CC*: Correlation coefficient**

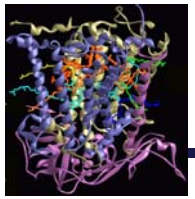
SVM performance is superior compared to the other classification methods based on results from leave-one-out cross validation



Multiple functional assignments for genes in *Synechococcus sp*

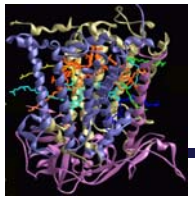
Gene id	Function Description	EC Number	Blocks Family
Synechococcus sp			
Contig48_gene556	Cystathionine gamma-lyase (EC 4.4.1.1) (Gamma-cystathionase)	4.4.1.1	IPB000277
	Probable cystathionine gamma-synthase (EC 4.2.99.9) (CGS) (O-succinylhomoserine (Thiol)-lyase)	4.2.99.9	IPB000277
	Methionine gamma-lyase (EC 4.4.1.11) (L-methioninase)	4.4.1.11	IPB000277
Contig51_gene26	Heptaprenyl diphosphate synthase component II (EC 2.5.1.30) (HEPPP synthase subunit 2)	2.5.1.30	IPB000092
	Bifunctional short chain isoprenyl diphosphate synthase. Includes: Farnesyl pyrophosphate synthetase (EC 2.5.1.1) (FPP synthetase) (Dimethylallyltransferase); Geranyltranstransferase (EC 2.5.1.10)	2.5.1.1 2.5.1.10	IPB000092

Predicted



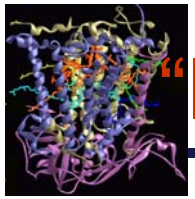
Summary: SVMMER

- SVMMER:
  - a tool for functional differentiation of highly homologous protein function groups
- Has unique features:
  - Incorporates *a priory* knowledge about protein functions when building a hierarchy of protein function groups.
  - Extracts protein features that are specific to each node of the hierarchy and trains a machine learning method to discriminate groups of proteins at such node.
  - Provides a possibility to expand the hierarchy to capture more and more functional specificities among homologous proteins.
- As a result:
  - Provides a significant step forward in the category of sequence similarity based methods by increasing the resolution level at which highly homologous but functionally different proteins can be discriminated



Identification of photosynthesis-specific genes

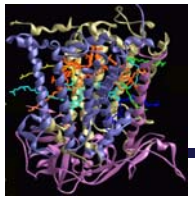
- Importance:
  - Understanding the genetic basis of photosynthesis is critical: it transforms solar energy into other forms of energy available to all living organisms.
- Problem:
  - Identification of “key” genes that are responsible for or contributing to photosynthesis is an essential step toward this understanding
- Challenge:
  - The problem presents a tremendous challenge due to the complexity of gene networks involved into this process



"Key" Genome Features Identification Problem

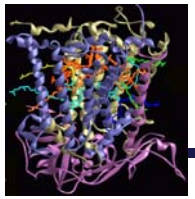
- Given:
  - A target biochemical process of interest, *P* (e.g., photosynthesis, pathogenesis)
  - A set X of *n* genomes: $X = X_P^+ \cup X_P^-$

    (X_P^+: *P* is present; X_P^- : *P* is absent)
  - A set F of *m* genome features (e.g., function groups, domains)
- Find:
  - A subset $S_P \subseteq 2^F$ of genome features or their combinations that contribute to or responsible for the process *P* in any of the genomes from X_P^+.



Approach to “Key” Features Identification

- Assumption:
  - Given a biochemical process ***P***, a set of genomes ***X*** , and a set of genome features ***F***, individual genome features can be identified as “key” features if the discrimination between two classes of genomes with respect to ***P*** becomes almost impossible once these features are not considered.
- Approach:
  - Discrimination procedure utilizes Support Vector Machines (SVMs), a supervised classification technique
  - Changes in classification performance in response to features deletion/addition measure the contribution of these features to the process ***P***



KeyGeneMiner for "Key" Features Identification

Genome Matrix:

PF: F_1 F_2 … F_j

Function group	f_1	f_2	f_3	…	f_m	Class Label
Genome$_1$	0	1	0		1	+
Genome$_2$	0	1	1		1	-
Genome$_3$	1	0	1		0	-
…	…	…	…	**Vik**	…	…
Genome$_n$	1	0	0		1	+

- **P** is present (**+**); **P** is absent (**-**)
- **n = 52** genomes; **m = 21,656** function groups organized into protein families: $F_1,\ldots, F_j$
- **Vik = 0** if no genes in **Genome$_i$** belong to **f$_k$**

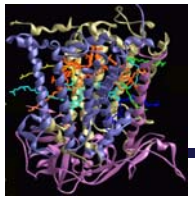
Steps:

A. Select target protein families

1. Take one protein family (PF) at a time
2. Compute leave-one SVM classification error
3. If error is big, then Go to Step 1; else identify PF as a target protein family

B. Select "key" function groups

1. Take one target protein family at a time
2. Apply backward-feature-propagation (remove function group w/ replacement)
3. Apply forward-feature-propagation (add function group)
4. If leave-one-out SVM classification error is small, then identify the function group as a key function group



Performance on Random vs. Genome Data

Kernel	SV (%)	ROC	Fp	Fn	Tp	Tn	SP	SN	CC
I. Genome data									
1	94.20	1	0	0	7	45	1	1	1
2	82.70	1	0	0	7	45	1	1	1
3	80.80	1	0	0	7	45	1	1	1
4	80.80	1	0	0	7	45	1	1	1
5	78.80	0.99	1	0	7	44	0.88	1	0.92
6	48.10	1	0	1	6	45	1	0.86	0.92
7	38.50	0.717	4	6	1	41	0.2	0.14	0.06
8	34.60	0.746	15	4	3	30	0.17	0.43	0.07
Radial	100%	1	0	0	7	45	1	1	1
II. Random data									
1	100	0.43	0	7	0	45	0	0	0
2	100	0.46	0	7	0	45	0	0	0
3	100	0.38	0	7	0	45	0	0	0
4	100	0.39	0	7	0	45	0	0	0
5	98.10	0.41	0	7	0	45	0	0	0
6	88.50	0.63	4	6	1	41	0.2	0.14	0.06
7	67.30	0.37	2	6	1	43	0.33	0.14	0.14
8	69.20	0.62	10	5	2	35	0.17	0.29	0.05
Radial	100	0.18	0	7	0	45	0	0	0

SVMs fail to identify structure in random data generated by shuffling class labels but succeed with genome data

SV: Support vectors,

Fp: False positives,

Fn: False negatives,

Tp: True positives,

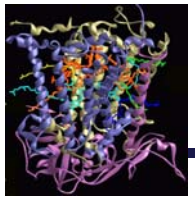
Tn: True negatives,

SP: Specificity,

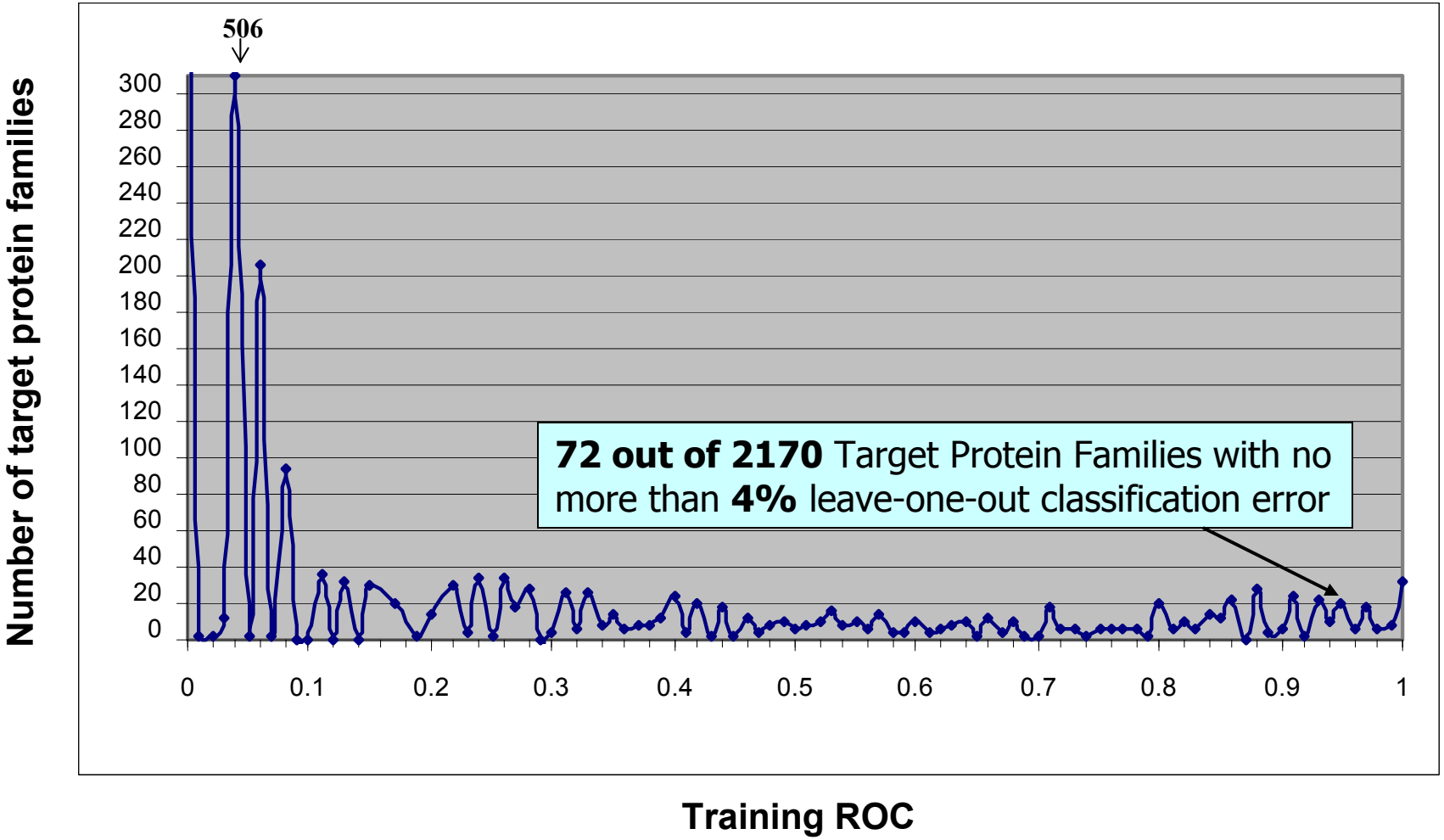
SN: Sensitivity,

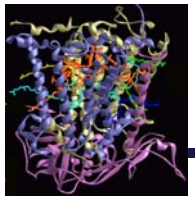
CC: Correlation co-efficiency

ROC: A score of the normalized area under a curve that plots true positives as a function of false positives for varying decision thresholds.



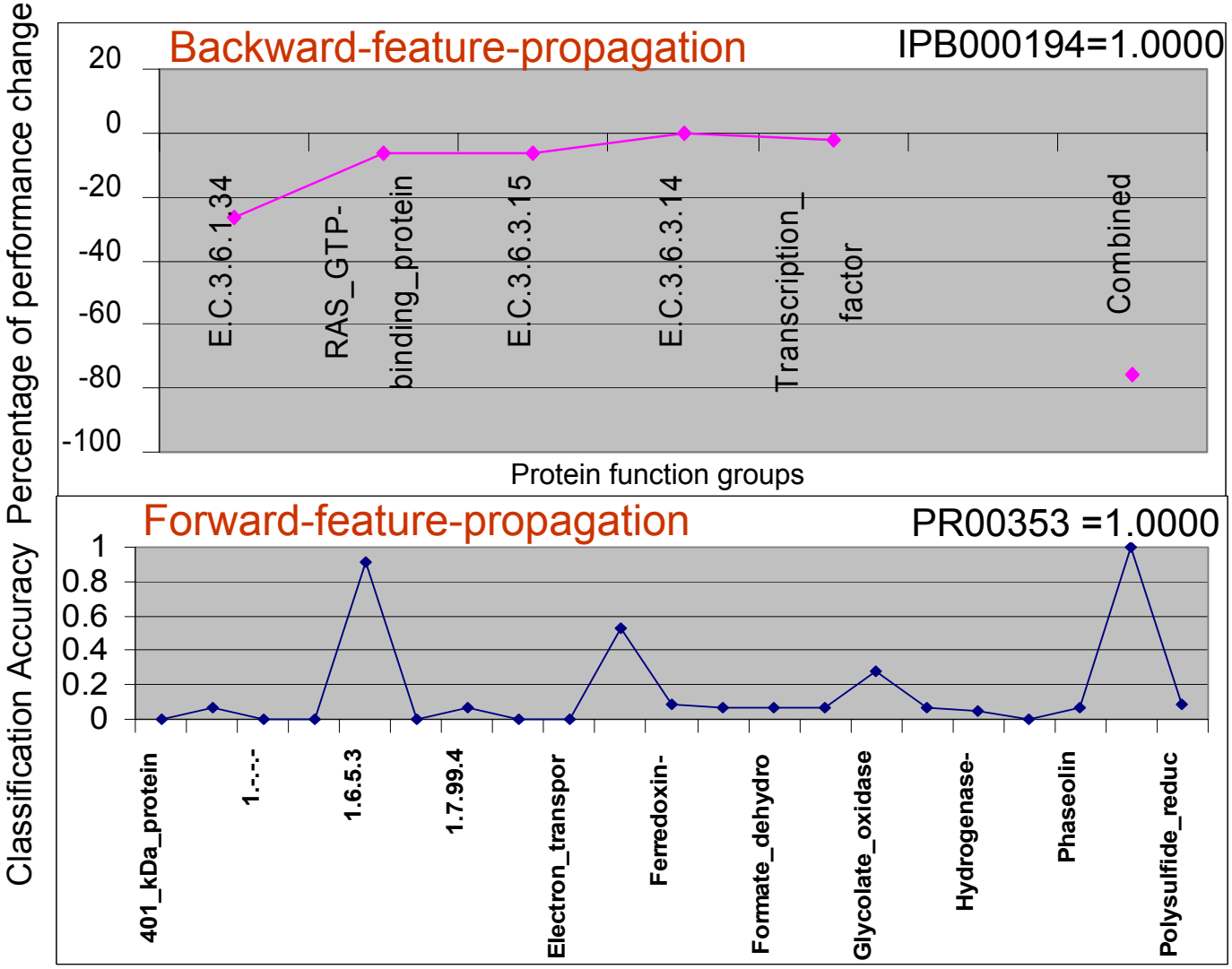
Selection of Target Protein Families

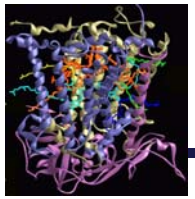




Selection of "Key" Function Groups

- **76** key function groups (by backward-feature-propagation)
- **50** additional key function groups (by forward-feature-propagation)





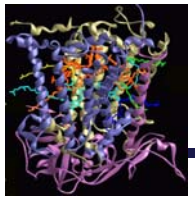
Validation of "Key" Function Groups

Gene clusters of "Key" function groups

5 clusters were established

Synechococcus sp. WH8102			
*or2996	2123	Light-dependent protochlorophyllide reductase, EC 1.3.1.33 IPB002529	I
or2997	2124	hypothetical protein	I
*or2998	2125	Carbon dioxide concentrating mechanism protein IPB000249 B	I
*or2999	2126	Ribulose bisphosphate carboxylase large chain 2, EC 4.1.1.39 IPB000685	I
*or3000	2127	Ribulose bisphosphate carboxylase small chain, EC 4.1.1.39 PR00152	I
or3001	2128	carboxysome shell polypeptide	I
or3002	2129	carboxysome shell polypeptide	I
*or3003	2130	Ethanolamine utilization protein eutN. Ethanolamine utilization protein PR00910	I
or3005	2131	carboxysome polypeptide	I
*or3006	2132	Major carboxysome shell protein 1A. Major carboxysome shell protein IPB000249	I
*or3007	2133	NADH-plastoquinone oxidoreductase chain 5, chloroplast, EC 1.6.5.3 IPB002128	I
*or3008	2134	NADH dehydrogenase subunit 4, EC 1.6.5.3 PR01437	I
*orl3145	2225	Iron(III)-transport ATP-binding protein IPB001140	II
*orl3146	2226	ATP phosphoribosyltransferase, EC2.4.2.17 IPB001348	II
*orl3147	2227	Hemolysin IPB001140	II
*orl3148	2228	Ycf52 protein IPB000182	II

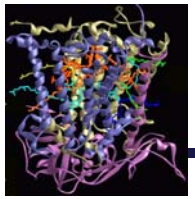
* Predicted



Identified Gene Clusters are Conserved across Multiple Cyanobacterial Genomes

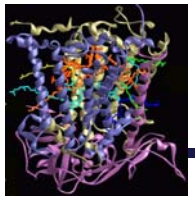
***Prochlorococcus marinus* MIT9313**		
*mit230	230	Light-dependent protochlorophyllide reductase, EC 1.3.1.33 IPB002529
*mit231	231	Putative sulfate transporter ybaR. Sulfate transporter IPB001902
mit232	232	hypothetical protein
mit233	233	hypothetical protein
mit234	234	hypothetical protein
mit235	235	hypothetical protein
*mit236	236	Carbon dioxide concentrating mechanism protein IPB000249
*mit237	237	Ribulose bisphosphate carboxylase large chain 2, EC 4.1.1.39 IPB000685
*mit238	238	Ribulose bisphosphate carboxylase small chain, EC 4.1.1.39 PR00152
mit239	239	carboxysome shell polypeptide
mit240	240	carboxysome shell polypeptide
mit241	241	carboxysome polypeptide
mit242	242	carboxysome polypeptide
*mit243	243	Major carboxysome shell protein 1A. Major carboxysome shell protein IPB000249
*mit244	244	Pterin 4 alpha carbinolamine dehydratase EC4.2.1.96 IPB001533
mit245	245	CfxQ protein . CFXQ protein PR00819 B chl
*mit246	246	ATP phosphoribosyltransferase, EC 2.4.2.17 IPB001348
*mit247	247	Leukotoxin secretion ATP-binding protein. Hemolysin IPB001140
*mit248	248	Ycf52-like protein Ycf52 protein IPB000182

***Prochlorococcus marinus* MED4**		
*med944	944	Light-dependent protochlorophyllide reductase, EC 1.3.1.33
med945	945	hypothetic protein
*med946	946	Carbon dioxide concentrating mechanism protein IPB000249
*med947	947	Ribulose bisphosphate carboxylase large chain 2, EC 4.1.1.39 IPB000685
*med948	948	Ribulose bisphosphate carboxylase small chain, EC 4.1.1.39 PR00152
med949	949	carboxysome shell polypeptide
med950	950	carboxysome shell polypeptide
med951	951	Ethanolamine utilization protein eutN. Ethanolamine utilization protein
med952	952	carboxysome polypeptide
*med953	953	Pterin 4 alpha carbinolamine dehydratase
*med954	954	ATP phosphoribosyltransferase (EC 2.4.2.17). 2.4.2.17 IPB001348
*med955	955	RTX-III toxin IPB001140
*med956	956	Ycf52-like protein. Ycf52 protein IPB000182



Summary: KeyGeneMiner

- Developed a method for identification of genes critical to a biochemical process of interest
- Applied the method to oxygenic photosynthetic process
- Identified 126 highly-confident “key” genome features:
  - Cover not only dominant features (that always occur in oxygenic photosynthetic genomes but not in the other genomes) but also weak yet complementary features (their combinations make unique dominant key genome features)
  - Many of these features are readily-recognized gene components in the oxygenic photosynthetic process
  - Many of the corresponding key genes are co-located on a genome, and clusters are conserved across multiple cyanobacterial genomes
  - Some clusters include hypothetical proteins => provide clues about their functional role



Summary of Accomplishments

- PICUPP: a tool for protein-protein interactions prediction
- SVMMER: a tool for high-resolution functional characterization
- KeyGeneMiner: a tool for identification of “key” genes responsible for a target biochemical process

Thank You!