From Genomics to Functional Omics: *In Silico* Approach

Nagiza F. Samatova

Computational Biology Institute

Oak Ridge National Laboratory

samatovan@ornl.gov

SIAM PPSC
February 25-27, 2004

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).

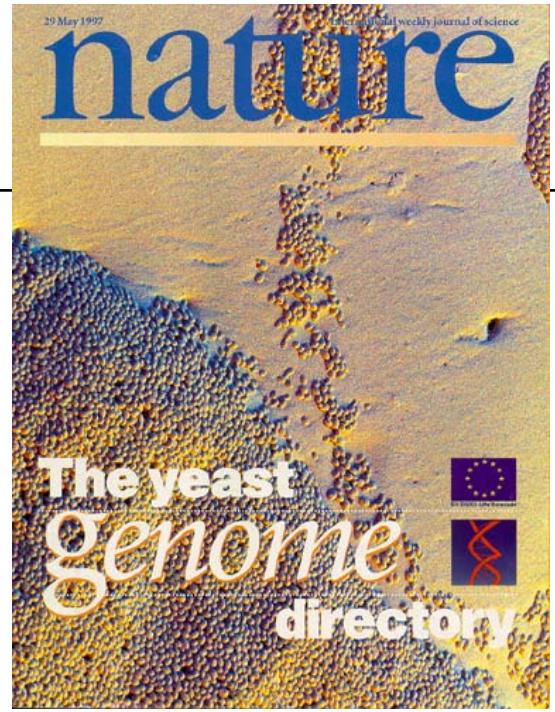
Highlights of the Genome Projects

1995

Bacteria,
1.6 Mb,
~1600 genes
[*Science* **269**: 496]

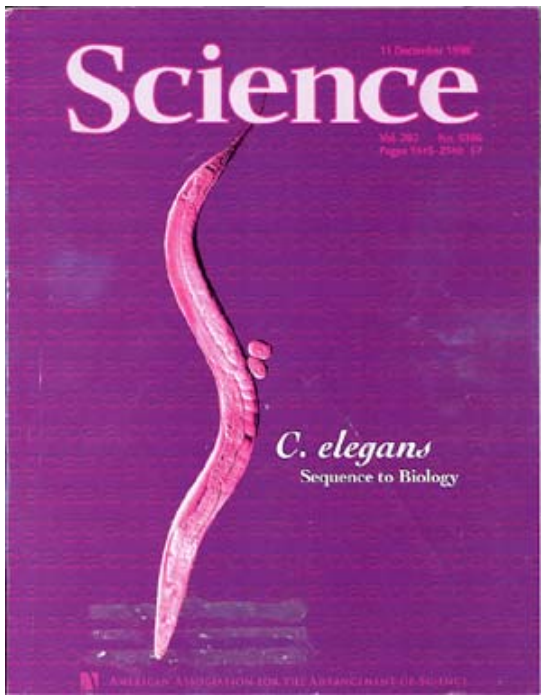
1997

Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]



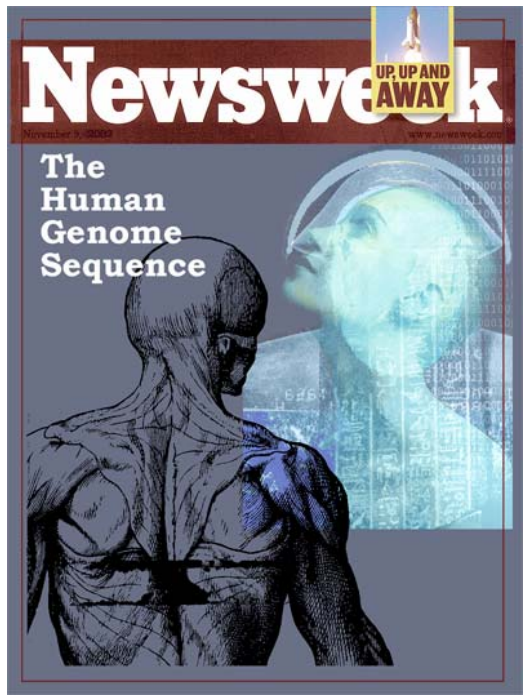
1998

Animal,
~100 Mb,
~20K genes
[*Science* **282**: 1945]



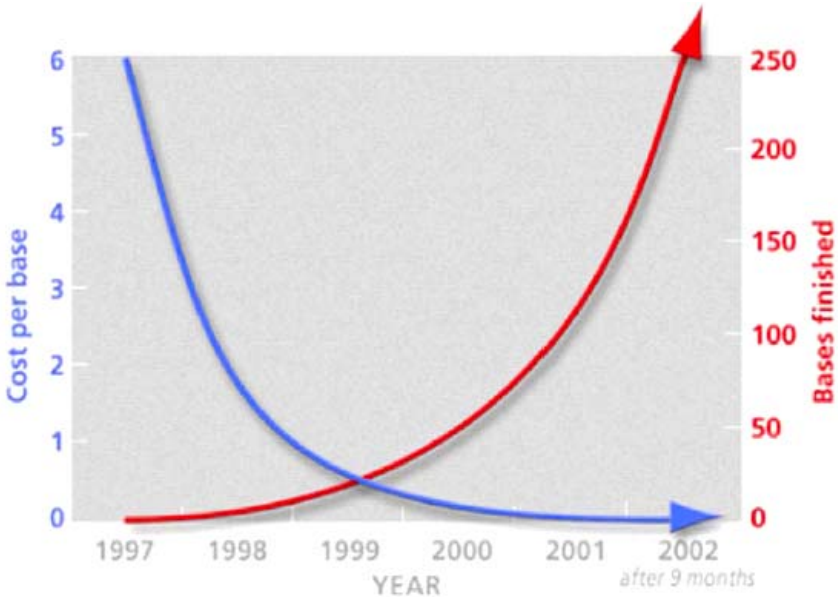
2000

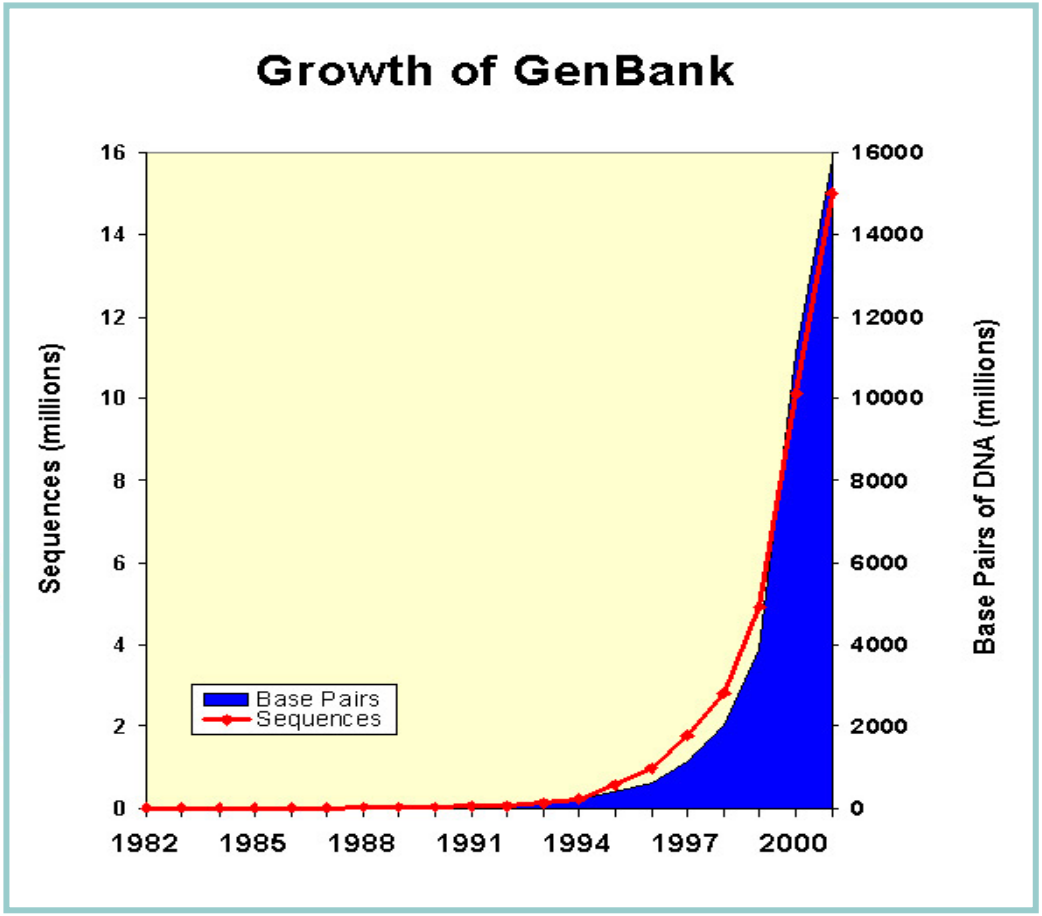
Human,
~3 Gb,
~30K genes



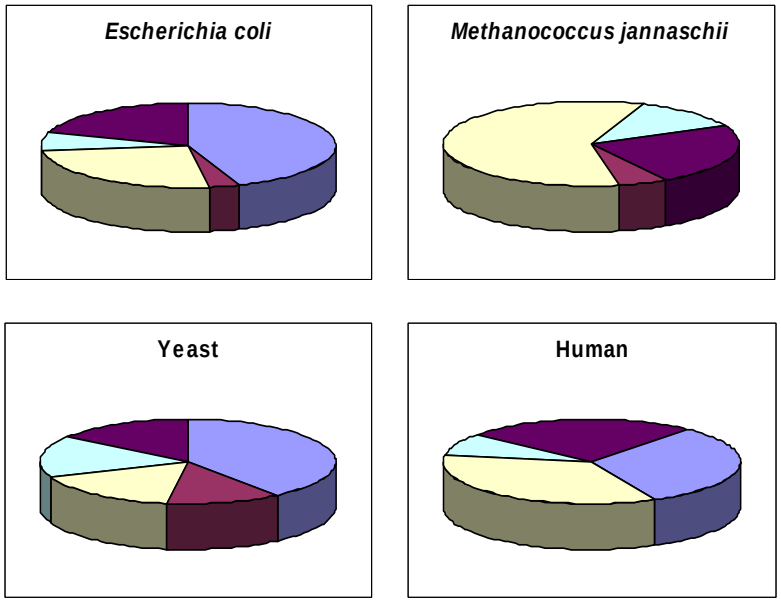
Status of Sequencing Data

High-performance computing has been essential to the high-throughput experimental approach to biology that has emerged in the last 10 years.





What About Annotations?



Although sequencing Mb regions of DNA has become quite routine, it remains a big challenge to characterize all the segments of DNA sequence with various biological roles.

Results from the completed prokaryotic genome sequences show that almost 50% of the predicted coding regions identified are of unknown biological function.

	E. coli	*M. jannaschii*	*S. cerevisiae*	*H. sapiens*
Characterized experimentally	2046	97	3307	10189
Characterized by similarity	1083	1025	1055	10901
Unknown, conserved	285	211	1007	2723
Unknown, no similarity	874	411	966	7965

from Koonin and Galperin, 2003, with modifications

From Genes to Protein Functions

Function assigned based on sequence similarity to another sequence with a

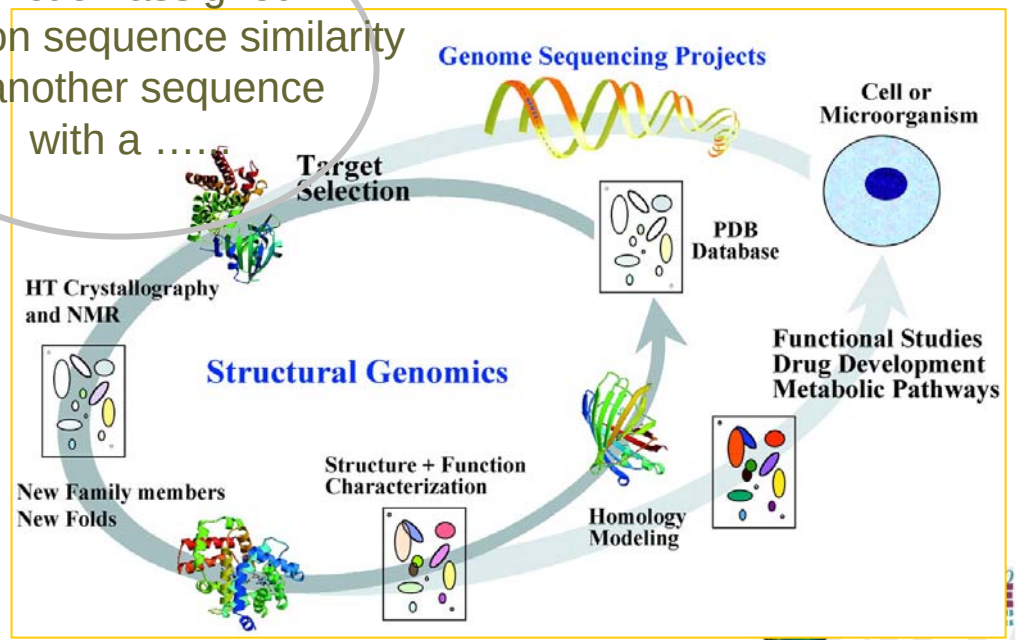
function assigned based on sequence similarity to another sequence with a

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

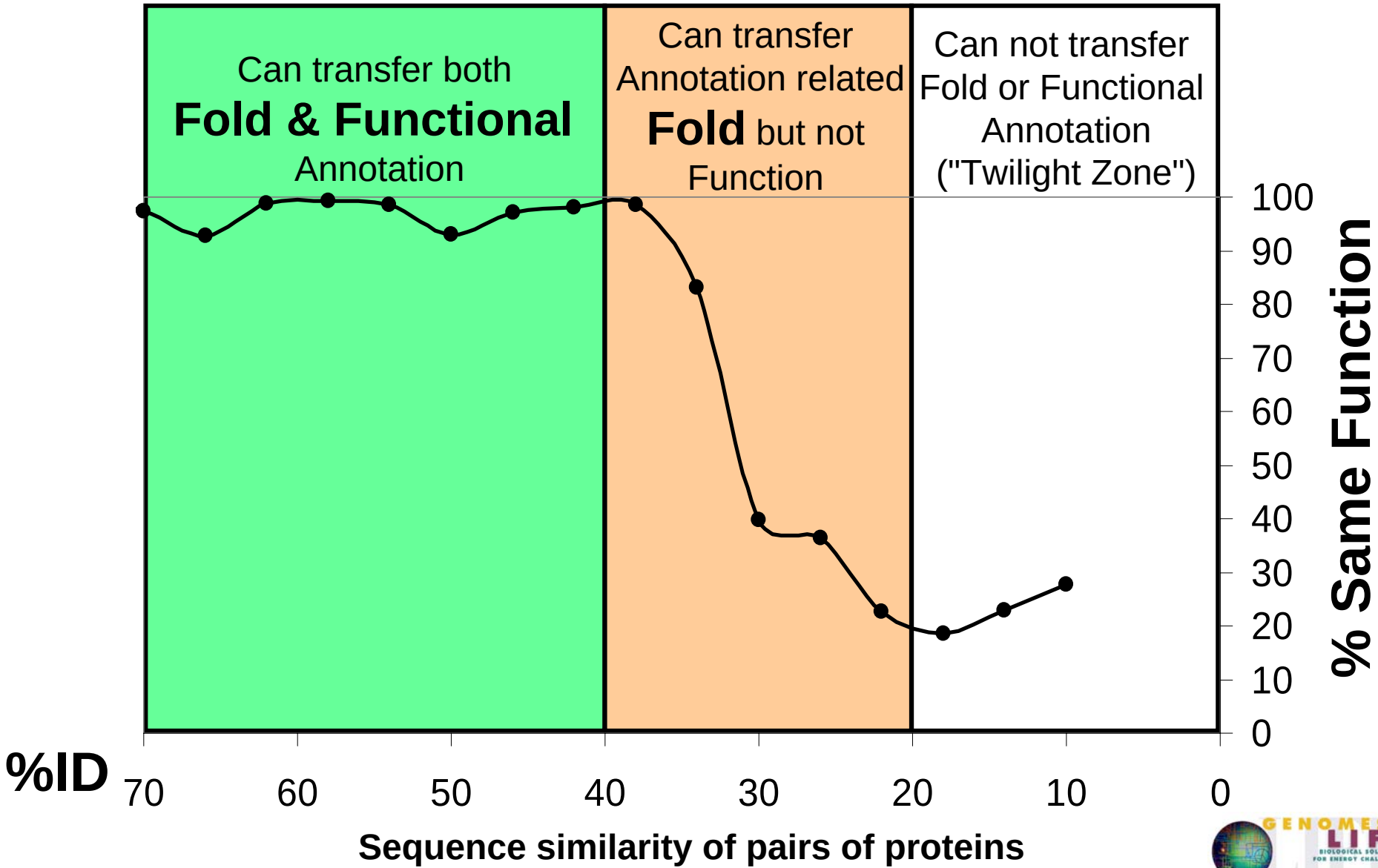
Weakness of sequence annotations

“guilt-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

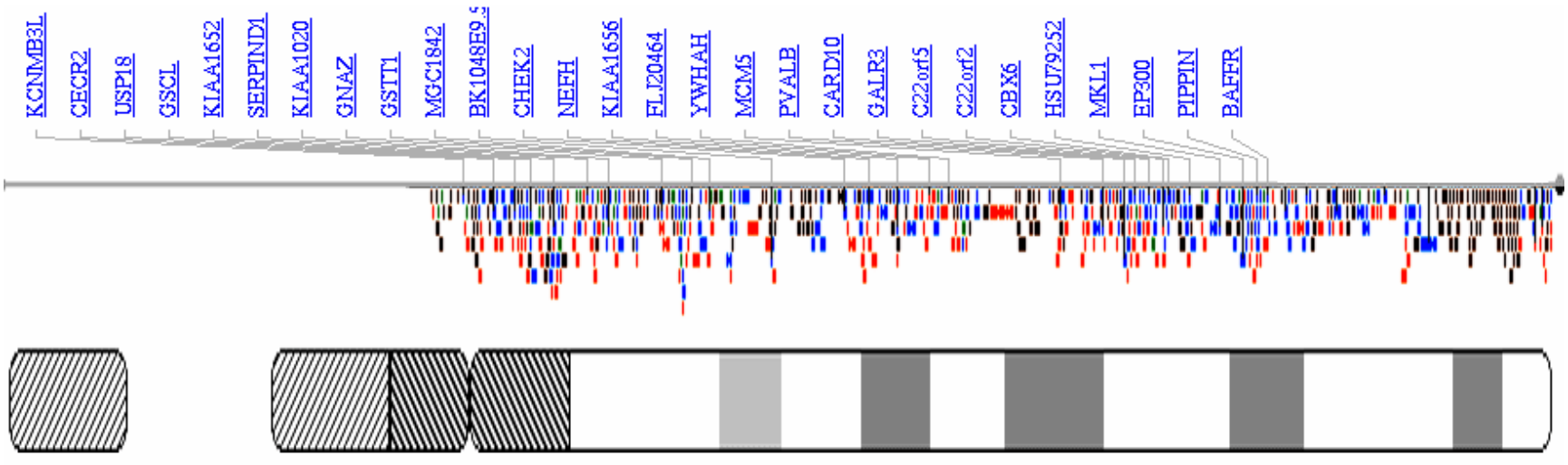


Relationship of Similarity in Sequence to that in Function



Why Function Annotation is Difficult?

- **Function exists at many hierarchical levels of description**
- It has temporal and spatial connections that are difficult to manage
- Functional descriptions do not correspond to well-defined physical models like biological structures defined by Cartesian coordinates for atoms.

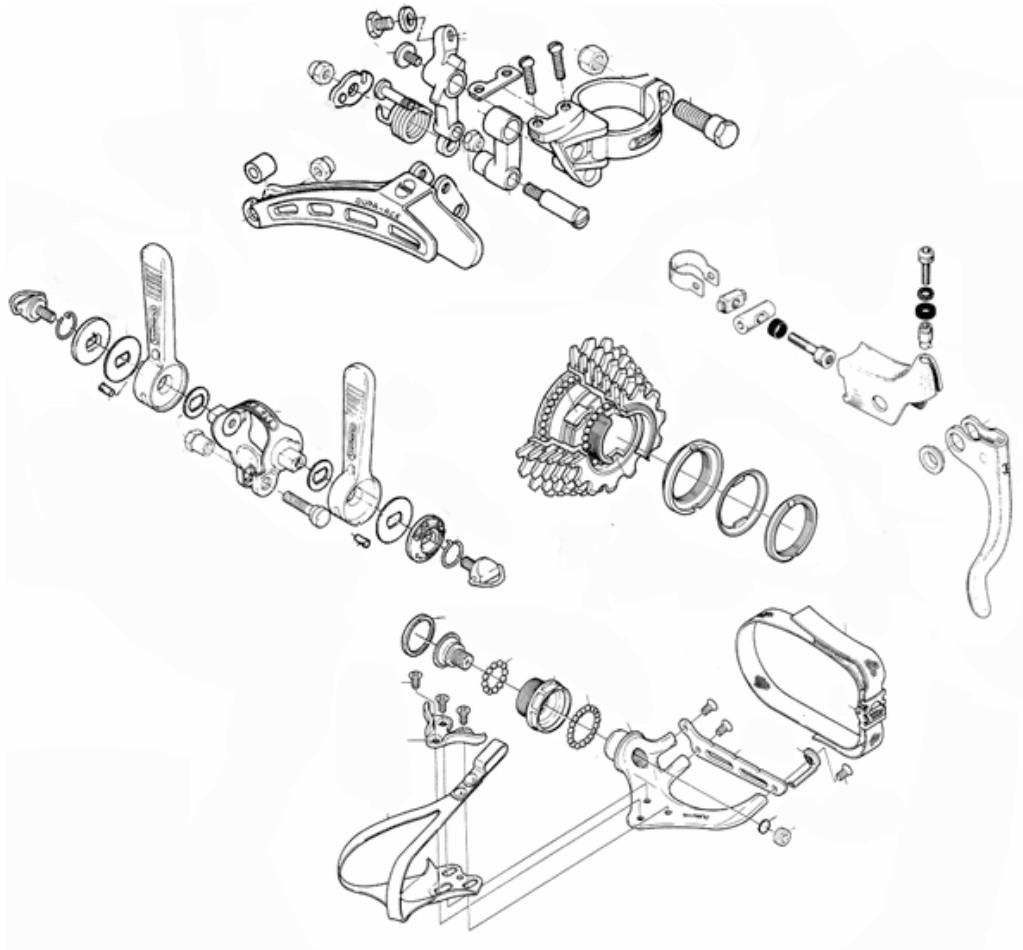


What is After the Genome?

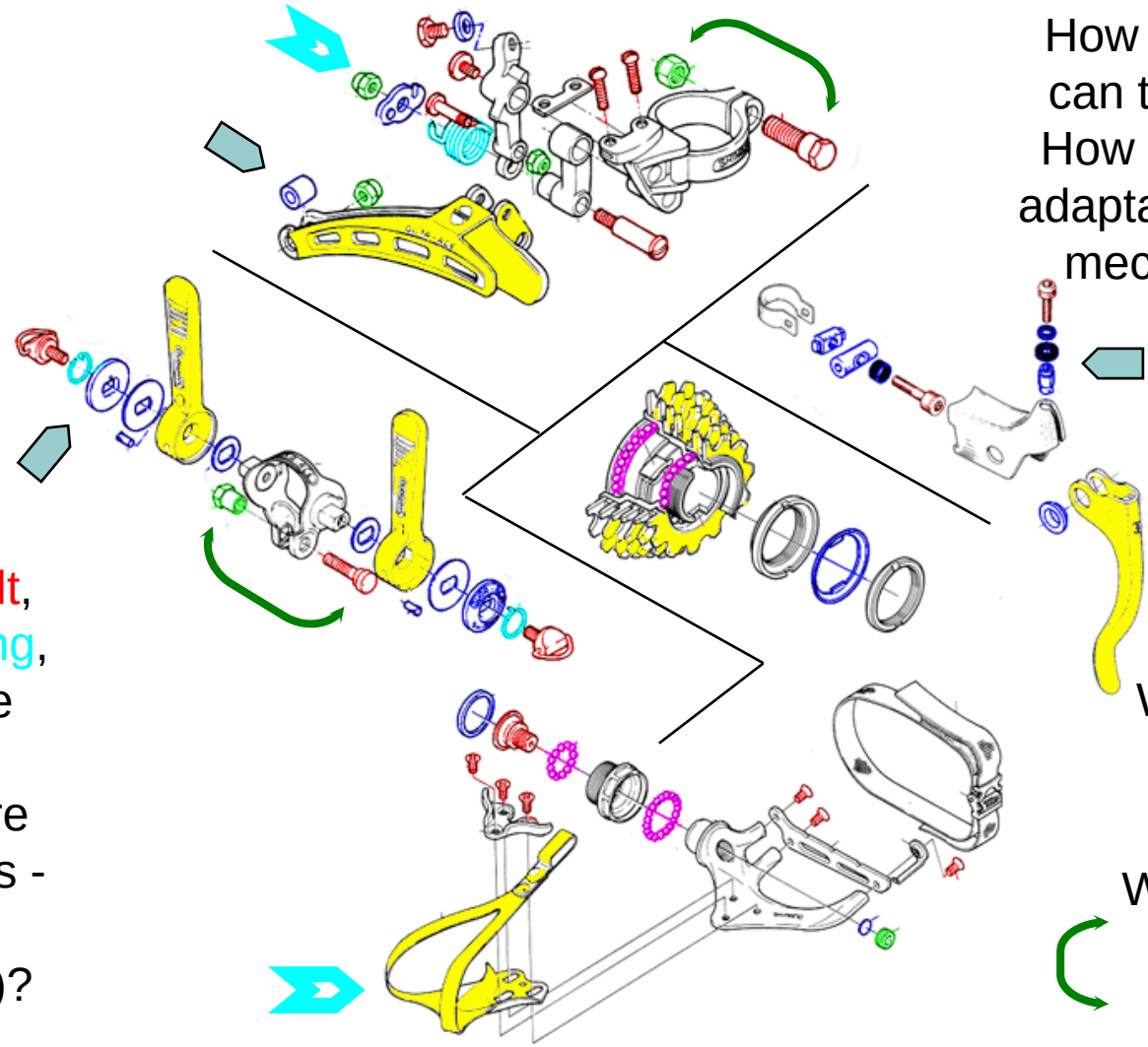
- Genome
- **Proteome**
- Transcriptome
- Physiome
- Metabolome
- Phenome
- Morphome
- Interactome
- Glycome
- Secretome
- Ribonome
- Orfeome
- Regulome
- Cellome
- Operome
- Transportome
- Functome

- Genomics projects have led to a fundamental way in which science is done. Genomics has opened the door to "discovery driven" studies as contrasted with traditional "hypothesis driven" research.
- A whole series of words have been coined to describe additional information people would like to have
  - Transcriptome – what are all the transcribed genes in an organism
    - *Gene expression microarrays*
  - Proteome – what are all the expressed proteins
    - *Two dimensional gel / Mass spectrometry*
  - Interactome – all the protein-protein interactions
    - *Two hybrid or tagging experiments*
  - Structurome – all three dimensional protein structures
  - Cardiome – complete electrophysiological model of the heart
  - Physiome – same for physiology
- Overall goal is to use the genome as a "parts list" and to build higher order models beginning from this complete list

A Collection of Parts



Assembly & Maintenance



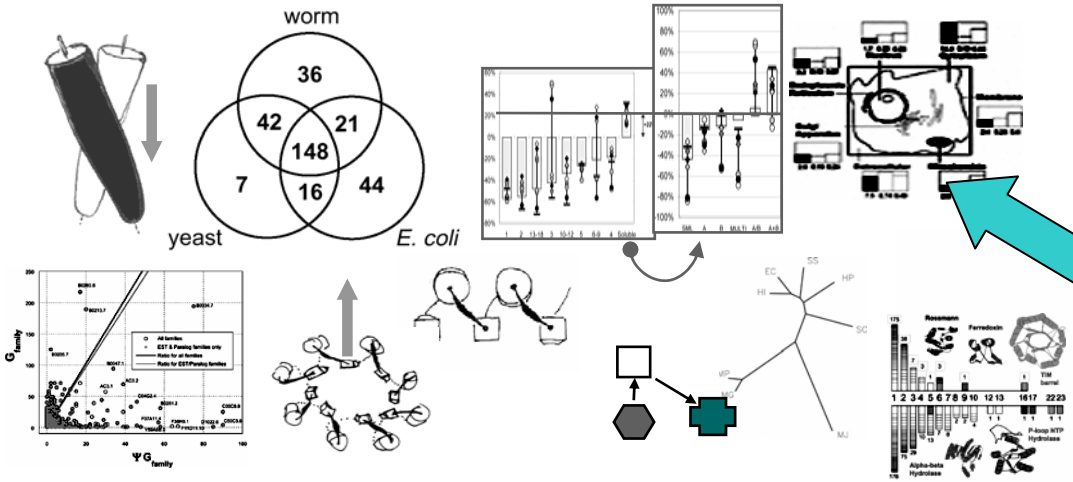
How many roles can these play? How flexible and adaptable are they mechanically?

What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

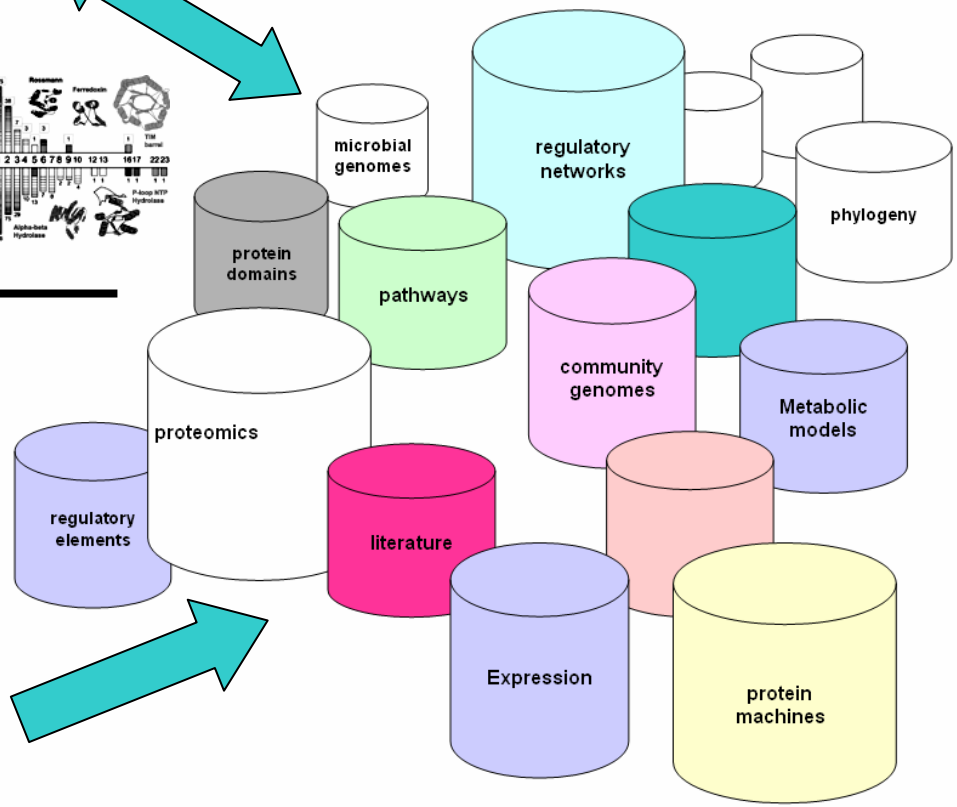
Where are the parts located? Which parts interact?

In Silico Integrative Omics

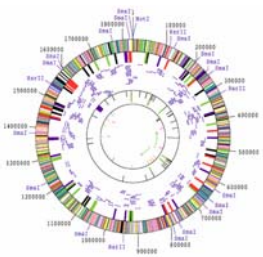
A Precursor to Functional Understanding



Data needs to be integrated and analysed in order to provide a rational basis for decisions

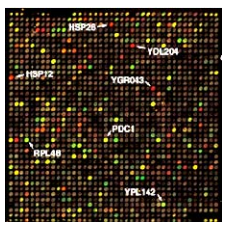


Smart Integration of Information



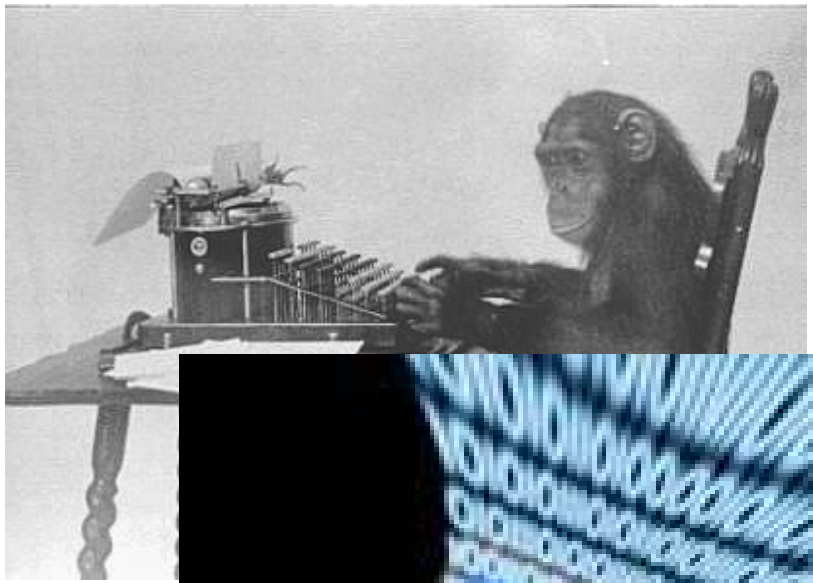
Sequences

Structures



Microarrays

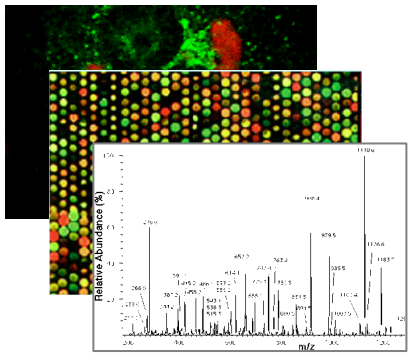
Challenge of Data Massiveness
Drinking from the firehose



- Genes for >30,000 proteins
- Up-to 10 variants each protein
- Up-to 5 genetic variants each protein
- Over 256 tissues in human body
- Each variant can interact with 10 proteins

billions of datapoints

Challenge of Breaking the Algorithmic Complexity Bottleneck



MS Data Rates:
100'sGB➔10'sTB/day(2004)➔1.0'sPB/day(2008)

Algorithmic Complexity:

Calculate means O(n)

Calculate FFT O(n log(n))

Calculate SVD O(r • c)

Clustering algorithms O(n^2)

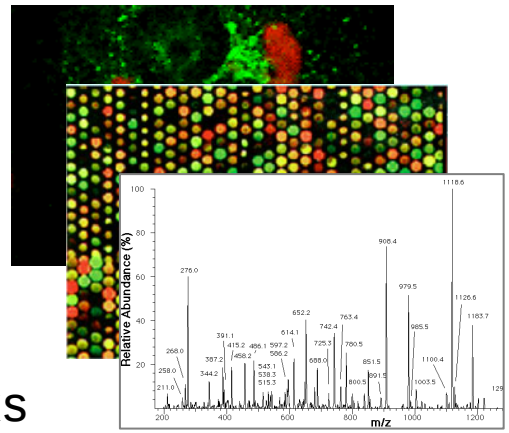
Data size, n	Algorithm Complexity		
	n	n*log*(n)	n^2
100B	10^{-10}sec.	10^{-10} sec.	10^{-8} sec.
10KB	10^{-8} sec.	10^{-8} sec.	10^{-4}sec.
1MB	10^{-6} sec.	10^{-5} sec.	1 sec.
100MB	10^{-4} sec.	10^{-3} sec.	3 hrs
10GB	10^{-2} sec.	0.1 sec.	3 yrs.

For illustration chart assumes 10^{-12} sec. calculation time per data point

Challenges of Data Heterogeneity

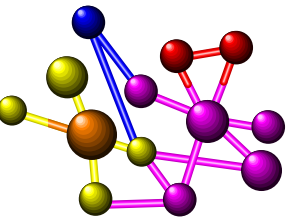
Data Complexity:

- Mass spectrometry
- Gene expression
- Genome sequences
- Protein structures
- Protein-protein interactions
- Metabolic and regulatory networks



Challenge of Data Complexity

Multi-scale, heterogeneous, distributed, dynamic,…



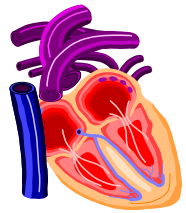
millions

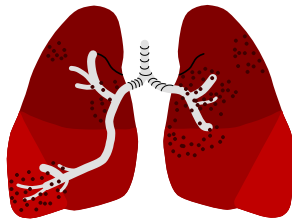
Protein-Protein Interactions
- metabolism
- pathways
- receptor-ligand
- 4º structure

billions

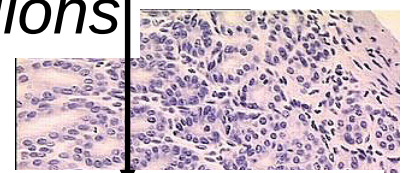
Polymorphism and Variants
- genetic variants
- individual patients
- epidemiology

Physiology
- Cellular biology
- Biochemistry
- Neurobiology
- Endocrinology
- etc.





millions



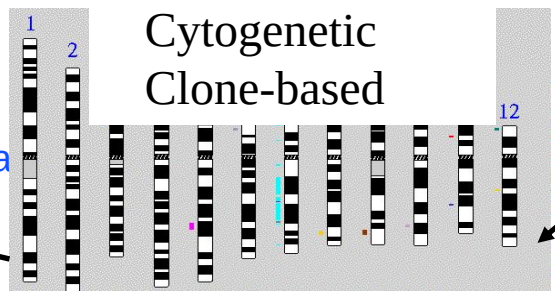
ESTs
Expression patterns
Large-scale screens

Hundred thousands

Proteins
- sequence
- 2º structure
- 3º structure

MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYT..

Genetics and Maps
- Linkage
- Cytogenetic
- Clone-based

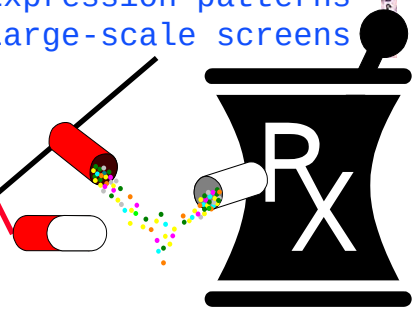


billions

DNA sequences
- alignments

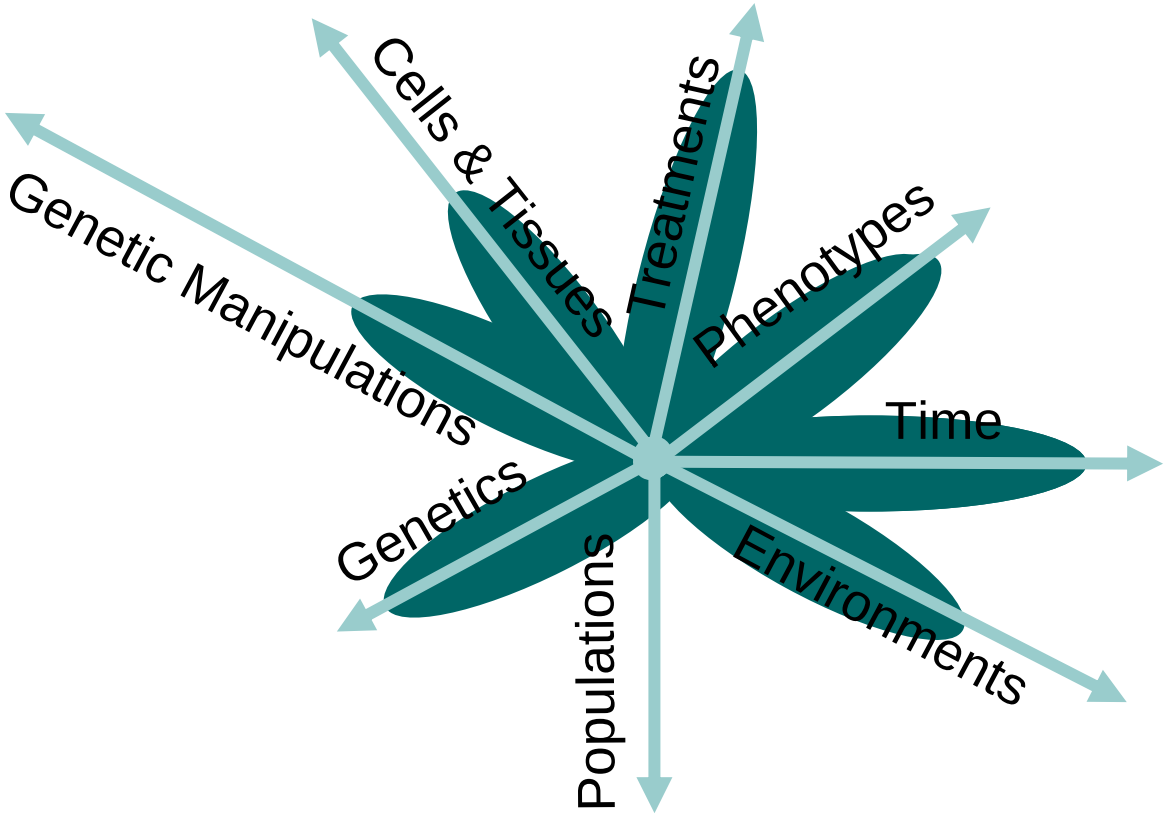
...atcgaattccaggcgtcacattctcaattcca

millions



GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

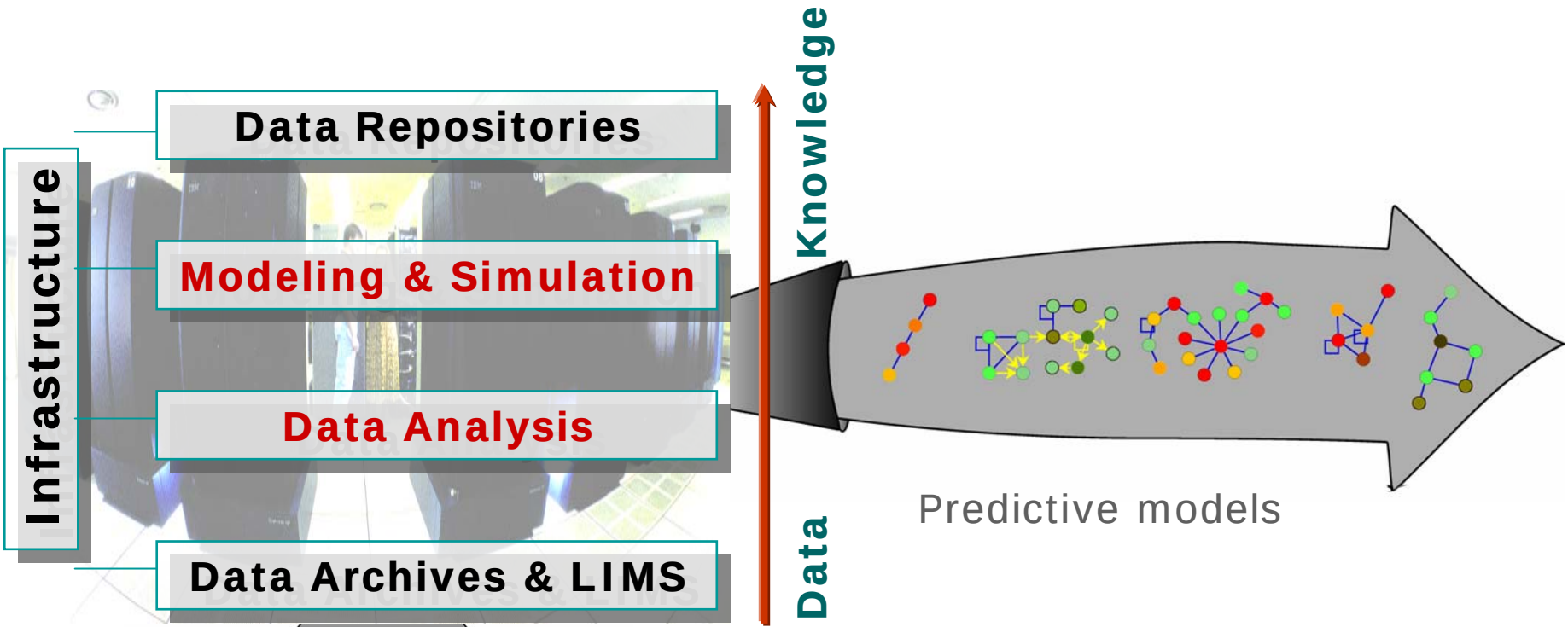
Challenge of High Dimensionity

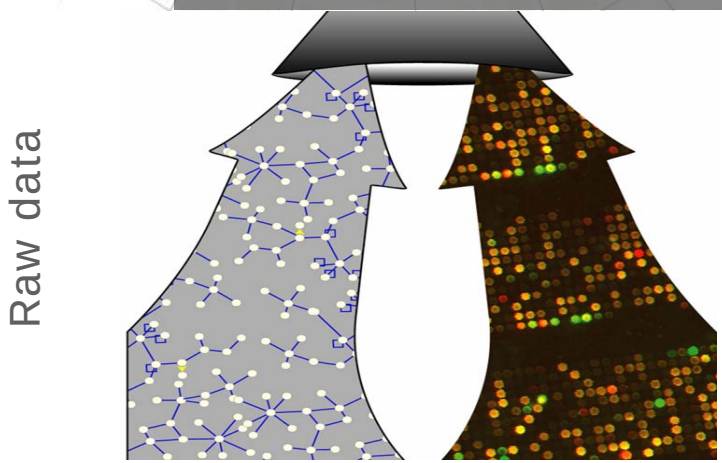


The experiment paradigm is changing to statistically capture the complexity. We will get maximum value when we explore three or more dimensions in a single experiment.

(From G. Michaels, PNNL)

The Challenges are Begging for "Smart" Data Analysis and Modeling Algorithms



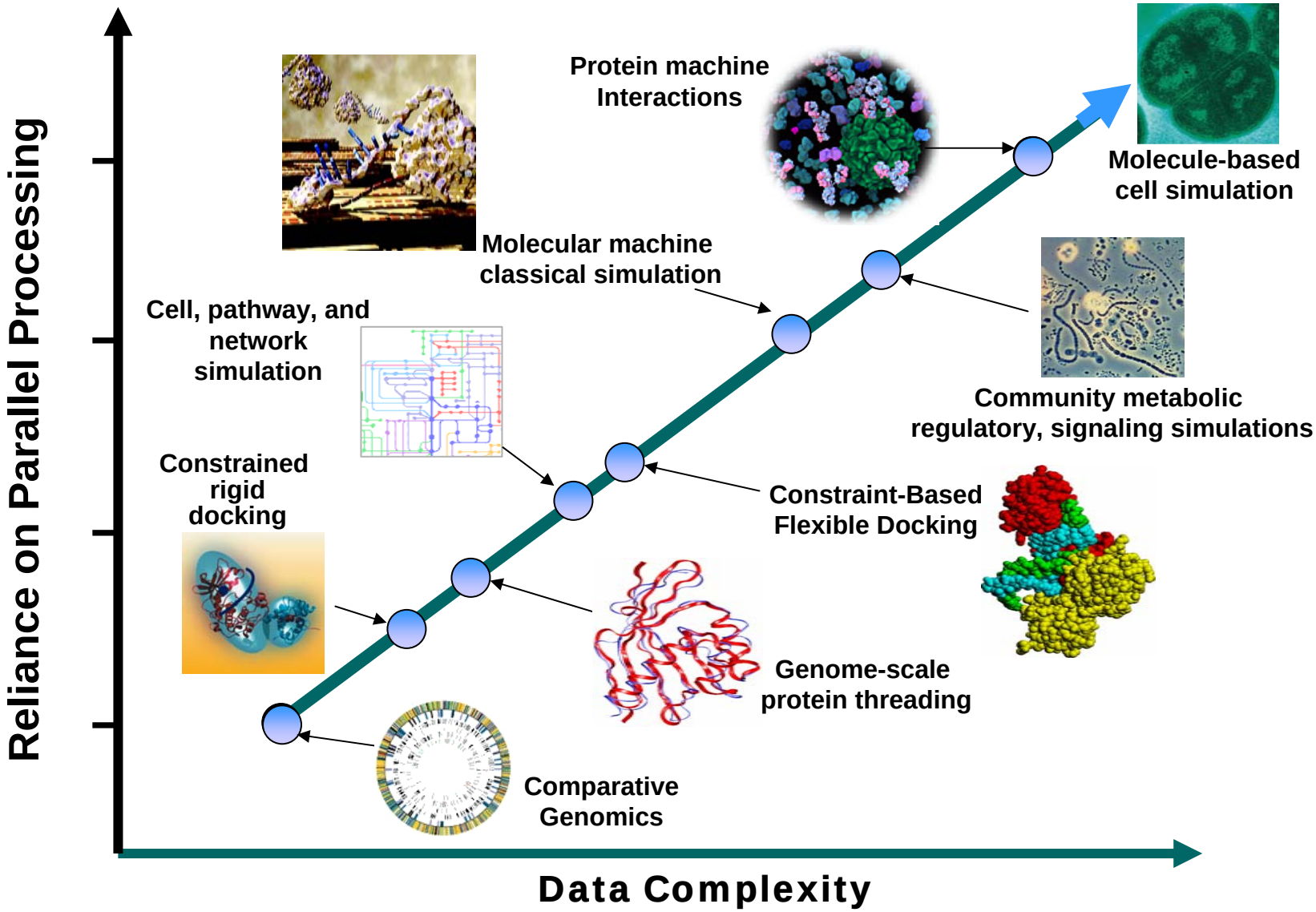


That will be able to "*learn the rules*" governing the multi-scale functional mechanisms of molecular machines

In order to provide predictive functional models

Multi-Scale Functional Characterization

Growing Reliance on Parallel Processing

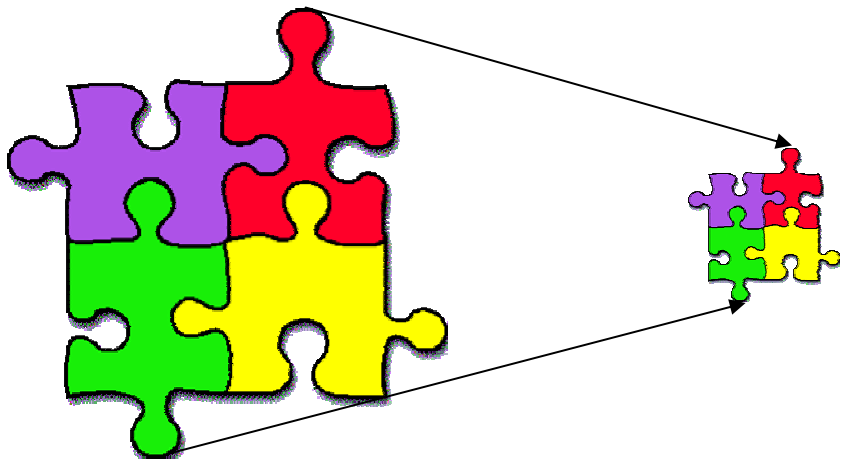


Paradigm Shift in Parallel Processing:

Increasing Not Only Speed But Prediction Accuracy

- **<u>From</u>: Transform a large problem into a set of small sub-problems with cumulative computational cost (time, memory, communication) much less than the aggregate problem**

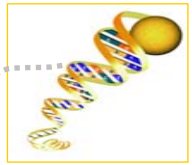
- **<u>To</u>: Transform a large problem into a set of small sub-problems with cumulative prediction accuracy much higher than the aggregate problem**



Example I. From Genotype to Phenotype

Through Genome-Scale Modeling of Biochemical Pathways

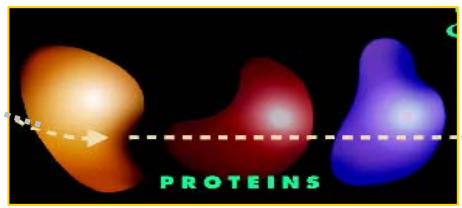
Genomes



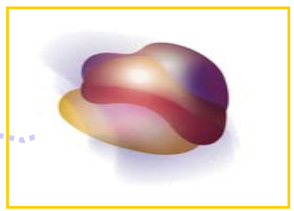
Challenge: Analyze, interpret, and predict metabolic function and control of a living cell

Metabolic maps:
provide a framework for studying the relationships between Metabolic Genotype & Phenotype

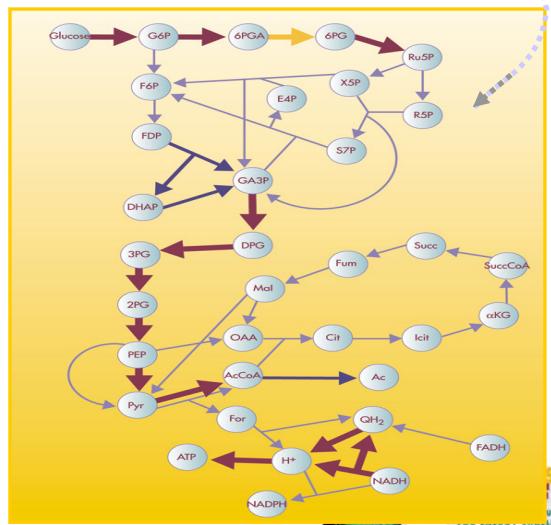
Gene Products



Structure & Function



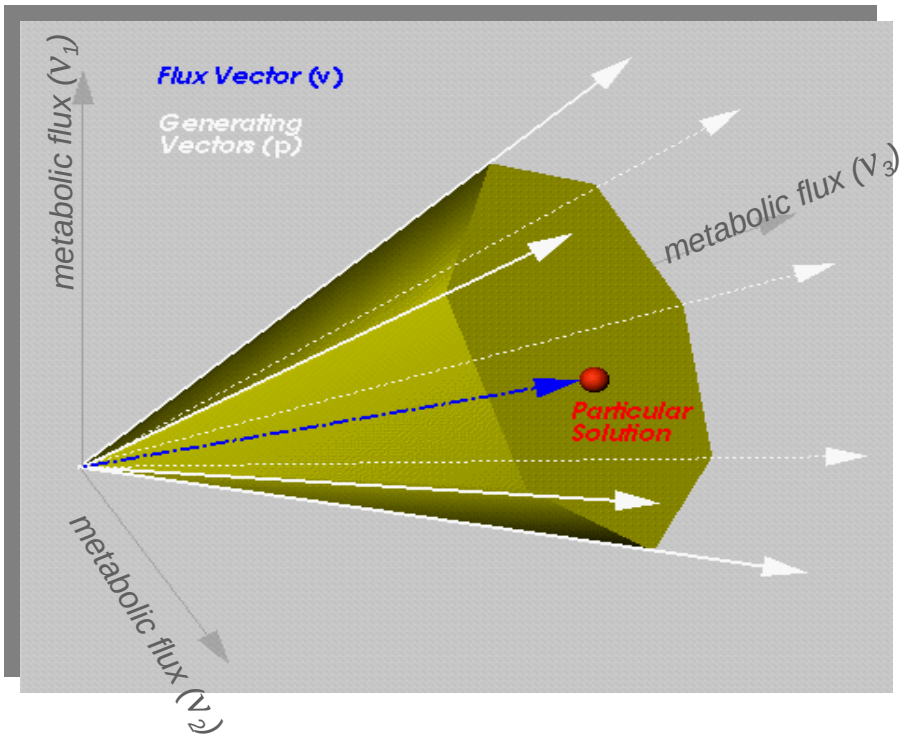
Pathways & Physiology



Optimal minimum solutions to the many possible metabolic pathways are called "Extreme Pathways" and are important to identify classes of cell Phenotypes

Extreme Pathways (EP)

Defined within the Context of Convex Analysis



$$C = \left\{ \mathbf{v} \in R^n \;\middle|\; \mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{p}_i, \quad \alpha_i \geq 0 \right\}$$

Convex Analysis	Cellular Biology
Convex Hull	Capabilities of a Metabolic Genotype
Unique Generating Vectors	Independent Extreme Pathways
Particular Solution	Metabolic Phenotype
Flux Vector	Positive Combination of Extreme Pathways

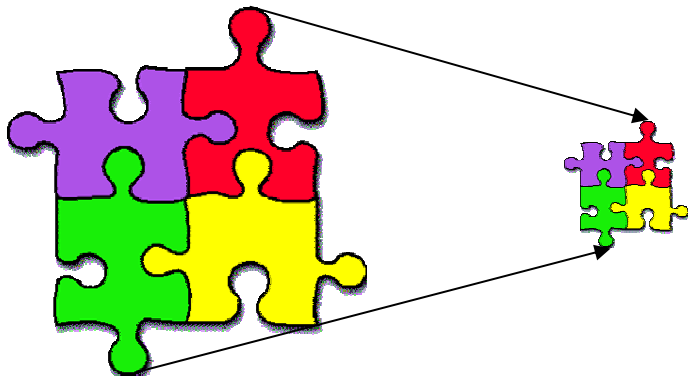
Allows the study of Metabolic Genotype-Phenotype relationships

Set of Extreme pathways is a basis for a defined metabolic genotype. The regulation of these pathways defines all possible phenotypes of the metabolic system.

Parallel Out-of-Core Extreme Pathways Algorithm

(with Bernhard Pallson, UCSD)

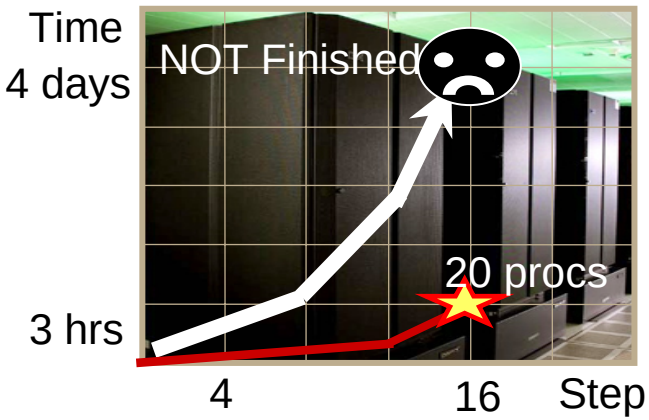
- **Strategy**
  - To transform a large problem into a set of small sub-problems with cumulative computational cost much less than the aggregate problem
  - To perform these sub-problems almost concurrently with reasonable data transfers, latency, and synchronization



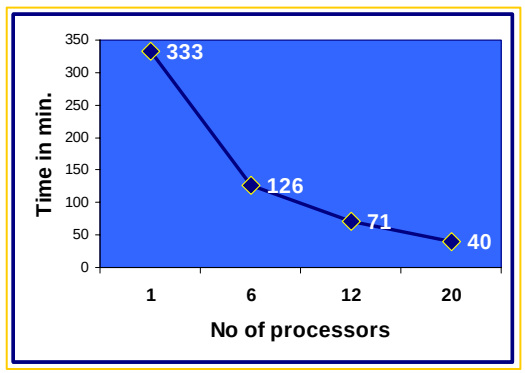
- **Key idea**
  - **Reduce memory requirements:**
    - Deploying out-of-core strategy
    - Using bitmap-based data representation scheme with a high compression rate
  - **Reduce computational time:**
    - Performing efficient bitwise logical operations without decompression overheads
    - Minimizing search space via maintaining descriptive statistics with cheap updates
    - RAM-based storage of all the info required for decision making
    - One-time synchronization/global initialization per long-running iteration
  - **Reduce network overheads**:
    - Partitioning data that minimizes data transfer needs

Performance of EP Algorithm

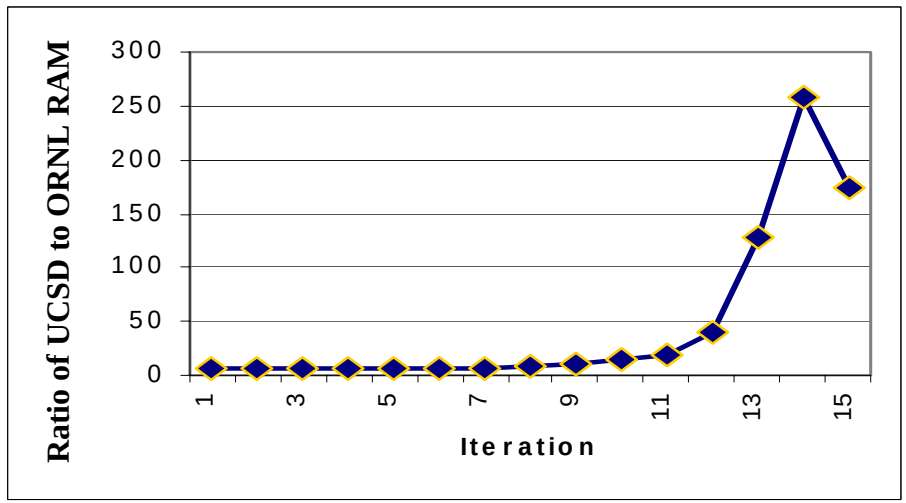
Previous algorithms could calculate Extreme Pathways only for small metabolic systems



- An order of magnitude larger systems
- Memory requirements: reduced by over 90%
- Computational time: reduced from days to hours



Reduction 300X Memory Utilization
(Fluxes=118 Metabolites=66)



Example II: Functional Discrimination of Homologous Sequences

Results of InterPro Analysis of four ADH Function Groups

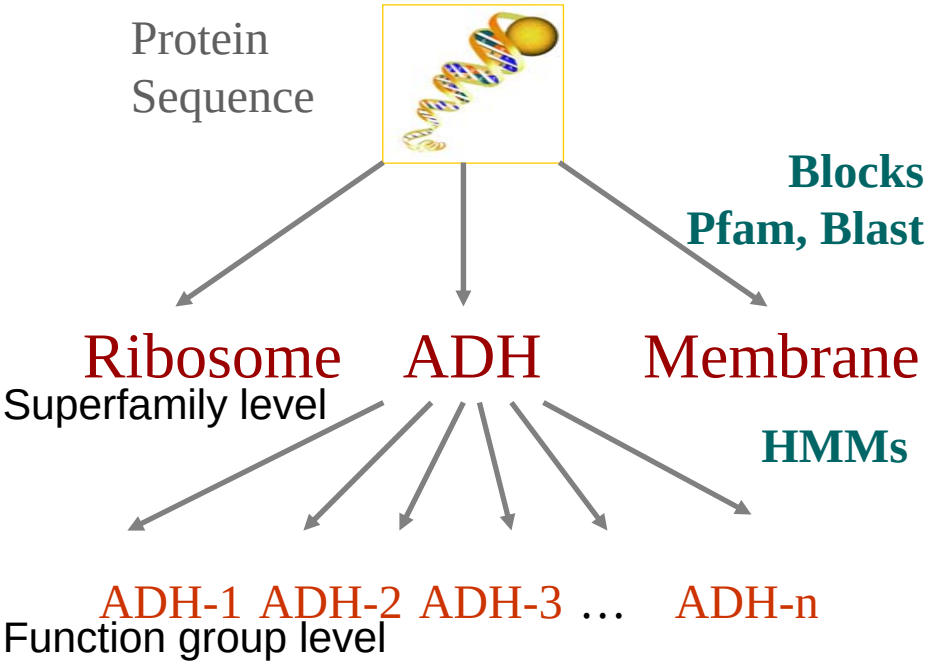
Function	Gene_id	Pfam	PROSITE	Tigr-Fam
EC1.1.1.1	ADH1_TRIRP	PF00107 adh_zinc	PS00059 ADH_ZINC	tdh: L-threonine 3-dehydrogenase
	ADH3_SOLTU	same	same	same
EC1.2.1.1	FADH_CANM	same	same	same
	FADH_PICPA	same	same	same
EC1.1.1.195	CADH_EUCGL	same	same	same
	CAD1_ARACO	same	same	same
EC1.1.1.255	MTD1_ARATH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

A tool for functional differentiation of highly homologous protein function groups



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

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

(In collaboration with Natalia Maltsev, ANL)

Performance Comparison

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

Example III. Prediction of Protein-Protein Interactions

- **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

Genomic context-based methods

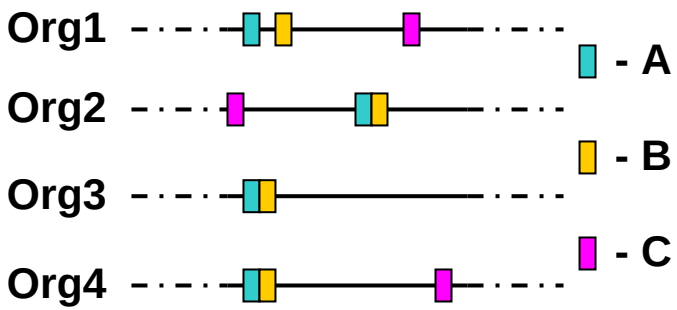
Phylogenetic Profiles

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

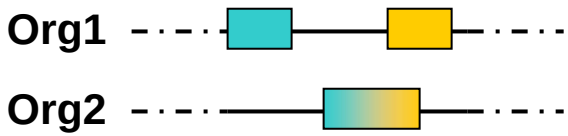
Conservation of gene-order



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.

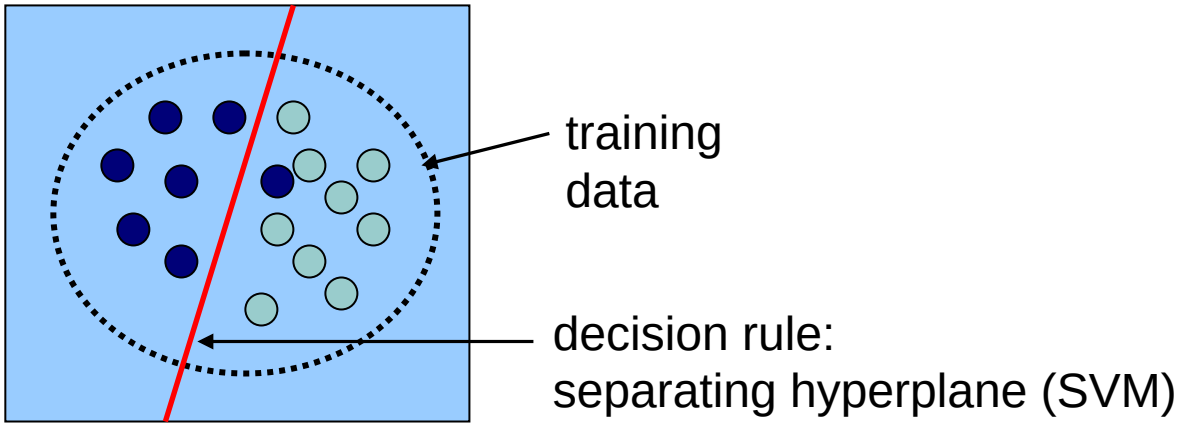
Advantages: Allow to identify functionally associated genes

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

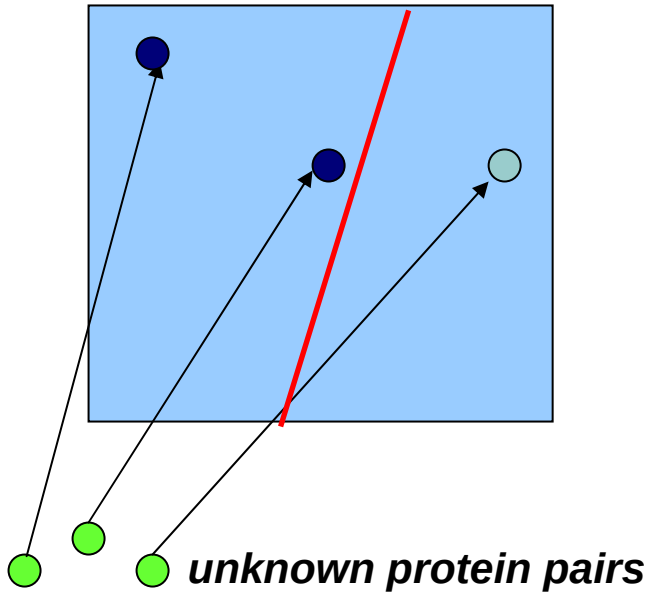
Coverage Statistics: 37%: Gene order; 6%: Gene fusion; 11%: Phylogenetic profiles (*M.genitalium*, Huynen et.al., 2000)

Classification-based methods

Training Stage



Prediction Stage



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

Limitations:

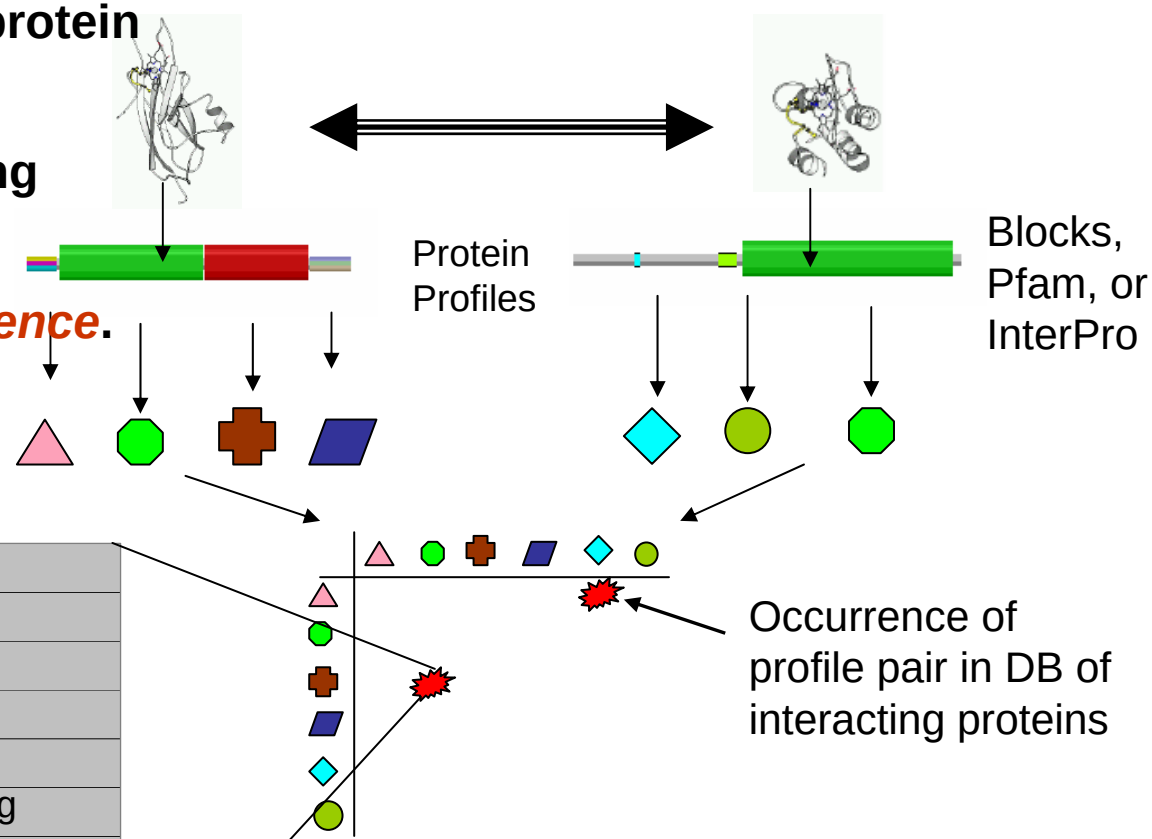
- Only positive data is available for protein interactions
- Ill-defined negative set: Shuffled protein may not belong to protein space
- Asymmetry: AB or BA protein pairs correspond to different points in space

Our Approach – PICCUP

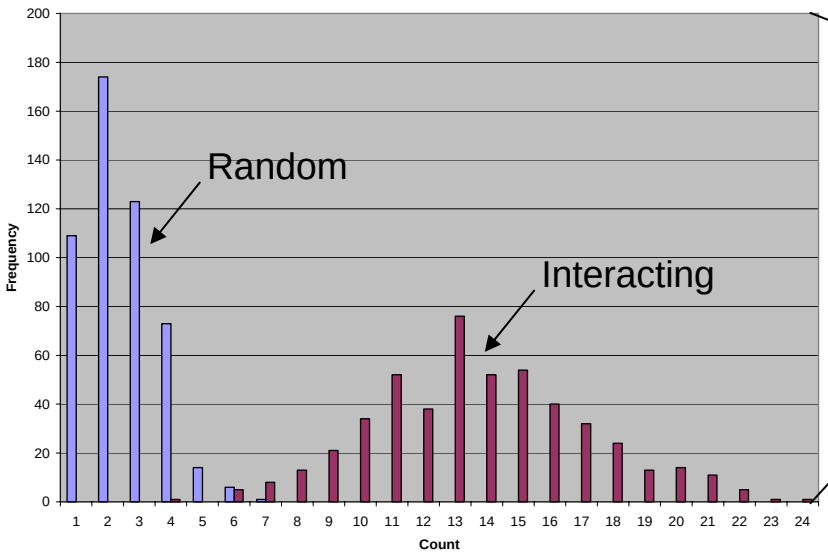
- **Given a set of positive interactions only, *statistical approach* is a natural candidate.**
- **PICCUP:**
  - Seeks to find statistically ***unusual*** protein profile pairs defined by Blocks, InterPro, Pfam, or Prosite.
  - Given protein interaction data, it estimates unusualness of profile pairs 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***.

Learning by Statistical Simulation

- **Find statistically *unusual* protein profile pairs.**
- **Estimate unusualness using statistical simulation.**
- **Compute *statistical confidence*.**

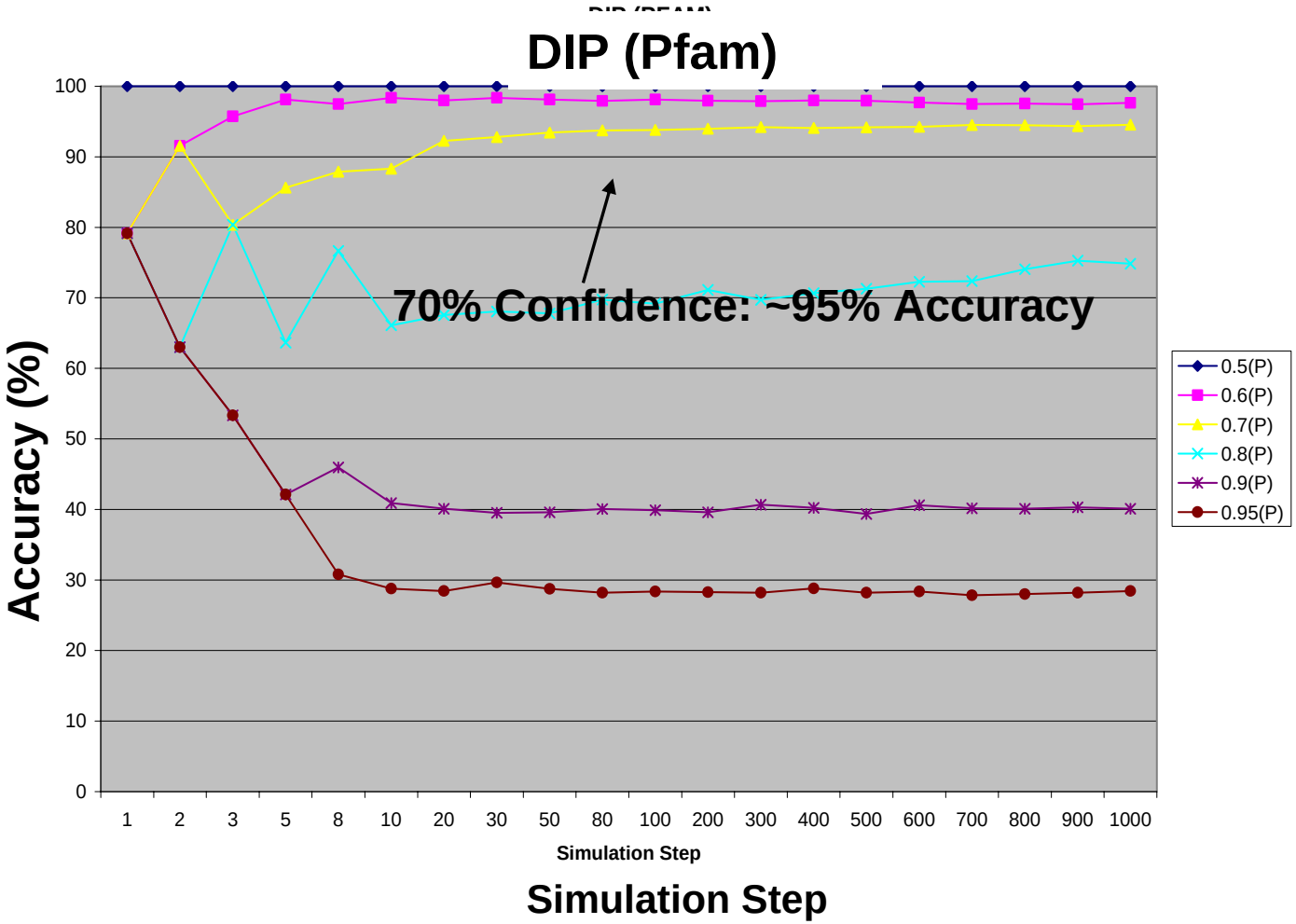


Pfam pair: PF0400:PF01423

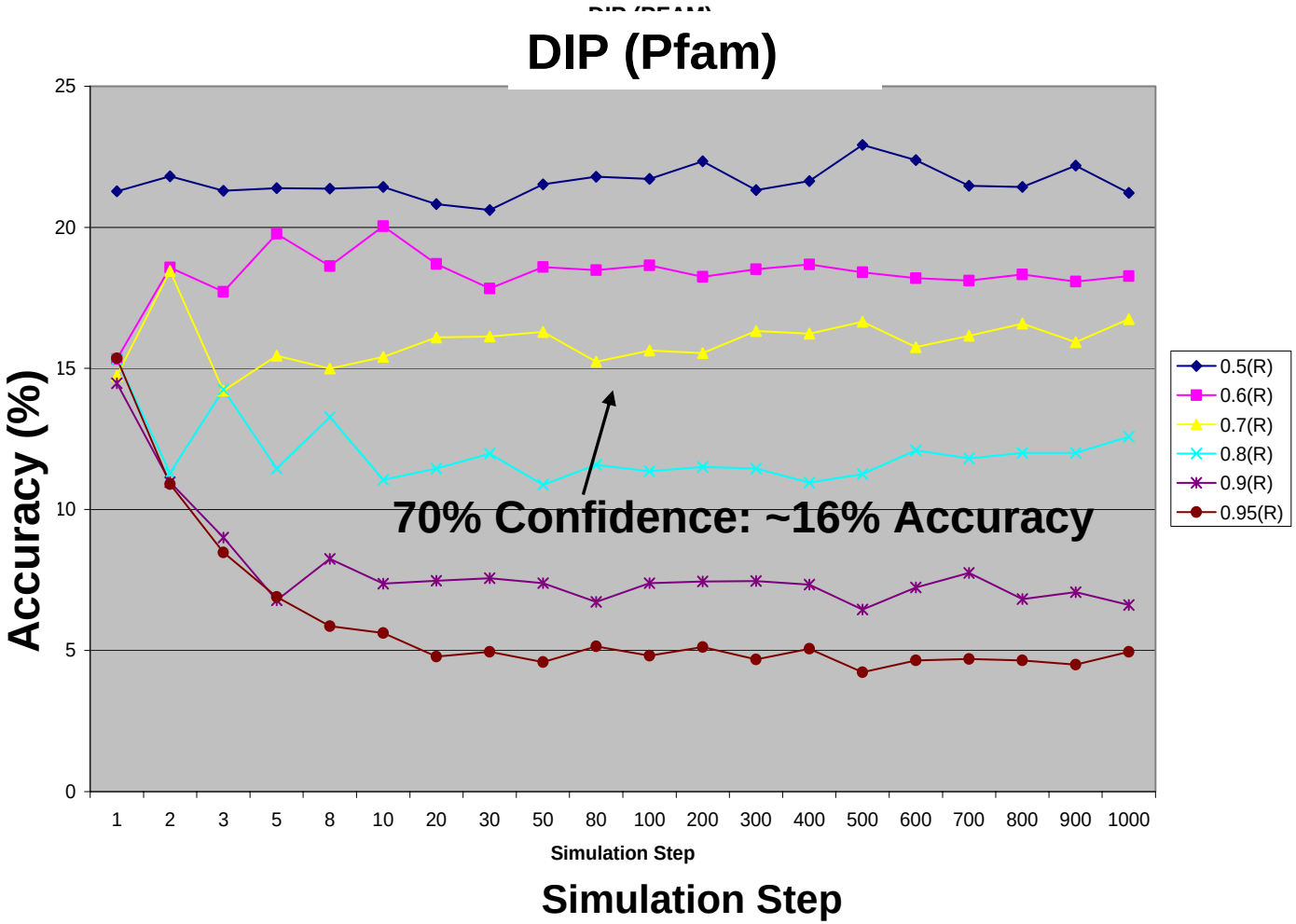


Profile Pairs Correlated for Protein Interaction

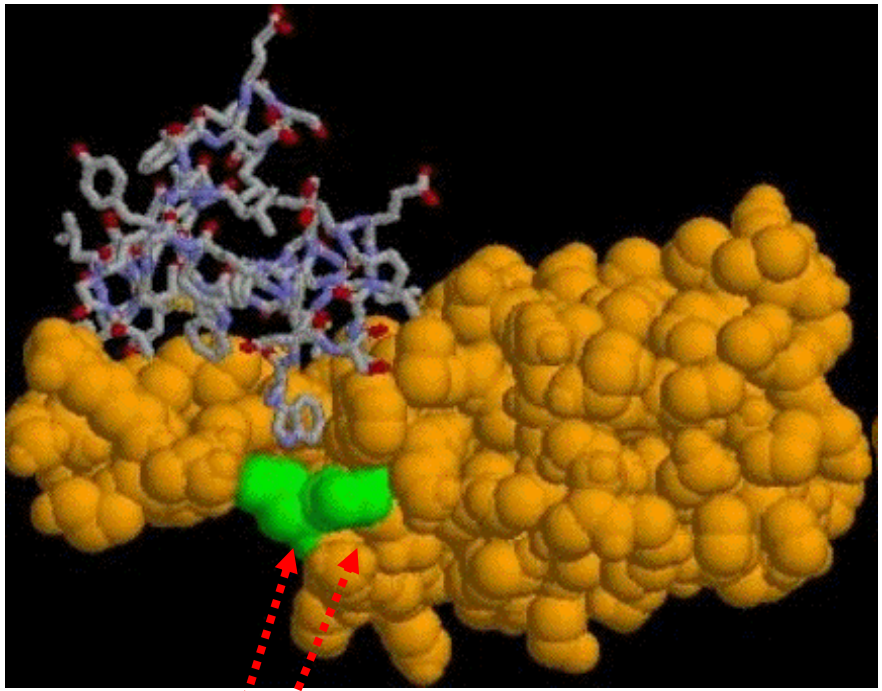
Sensitivity To Positive Pairs (Pfam)



Sensitivity To Random Pairs (Pfam)

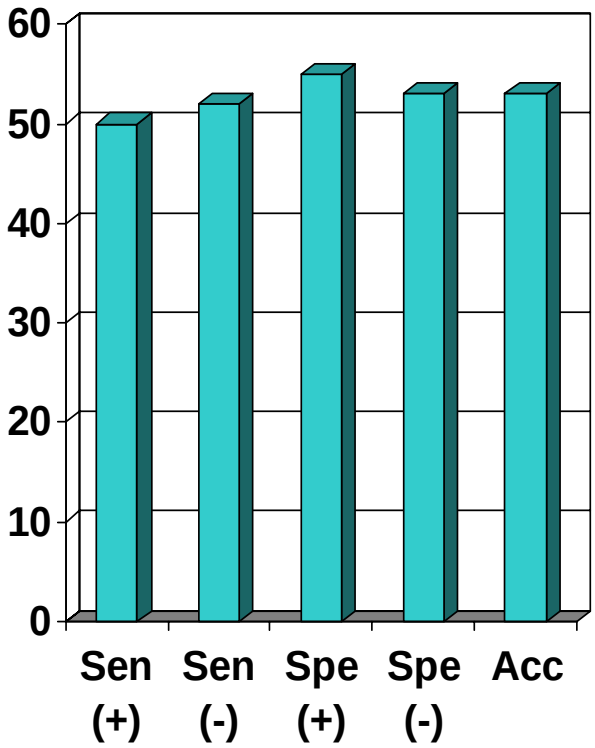


Example IV: Prediction of Protein Docking Interfaces



Interacting Residues

MGYPPPTRRL ... GDKKRYHYSN



A classifier from the data performs slightly better than a coin tossing.

Is there something in the Data then?

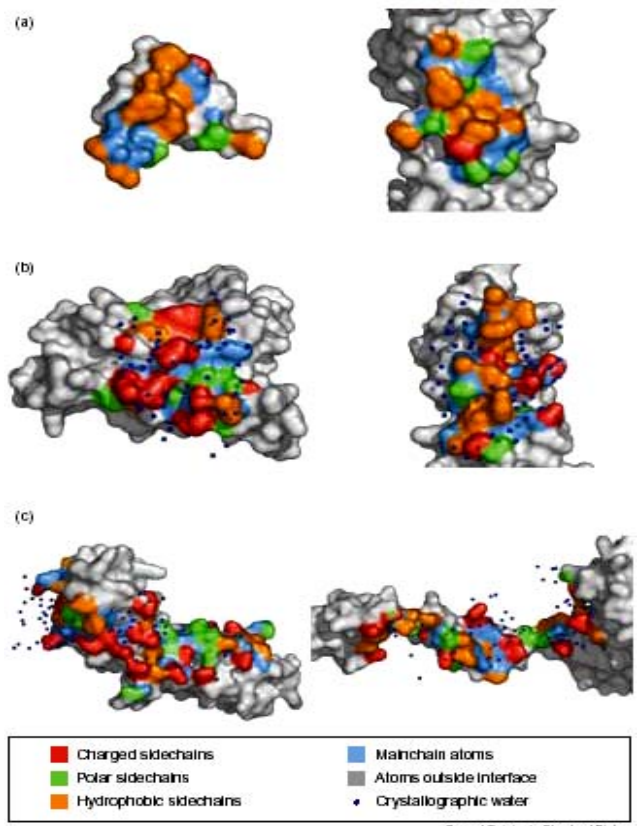
The complexity of interaction sites makes it extremely challenging to predict them computationally.

Interaction sites are diverse in:
- hydrophobicity
- electrostatics
- accessible surface area
- propensity
- degrees of planarity.

The diversity gives rise to the specificity

However, none of these was found to dominate.

In other words, no single reliable prediction rule can be inferred using these factors.

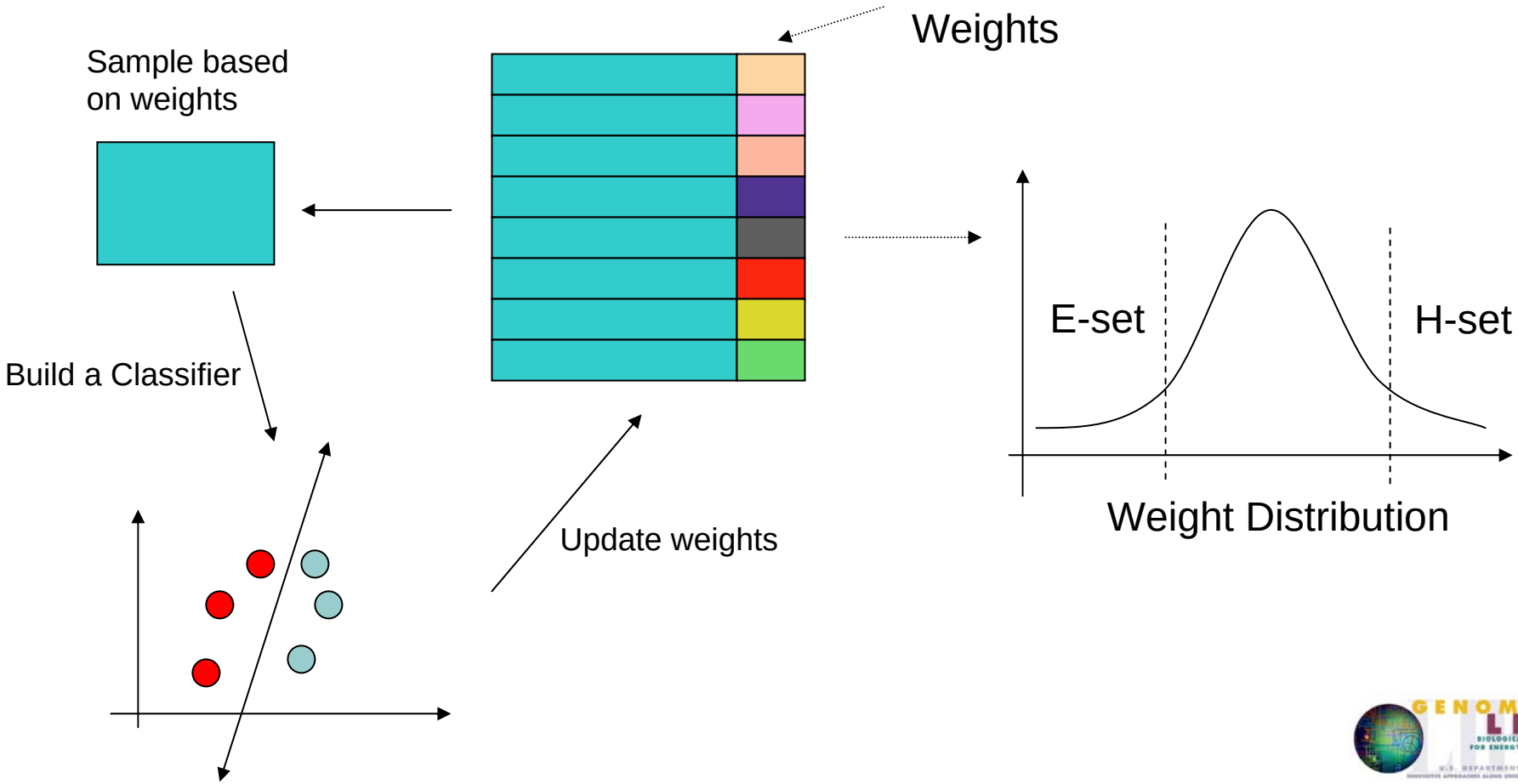


Current Opinion in Structural Biology

Data Separation with AdaBoost

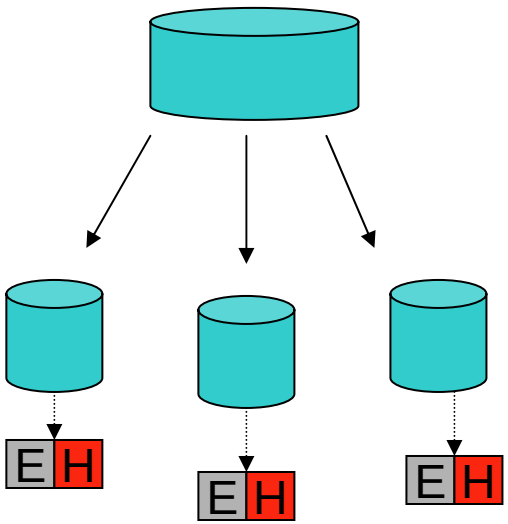
- **Noise Filtering (here "noise" can include "too complex pattern to capture")**

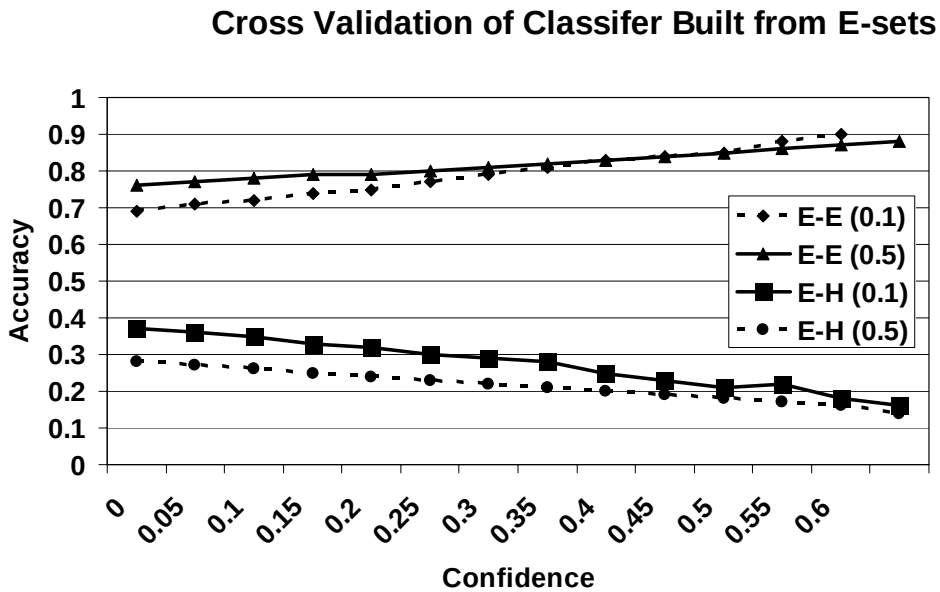
  Apply Boosting framework to identify problematic feature vectors.
- **Construct a classifier that predicts with confidences.**

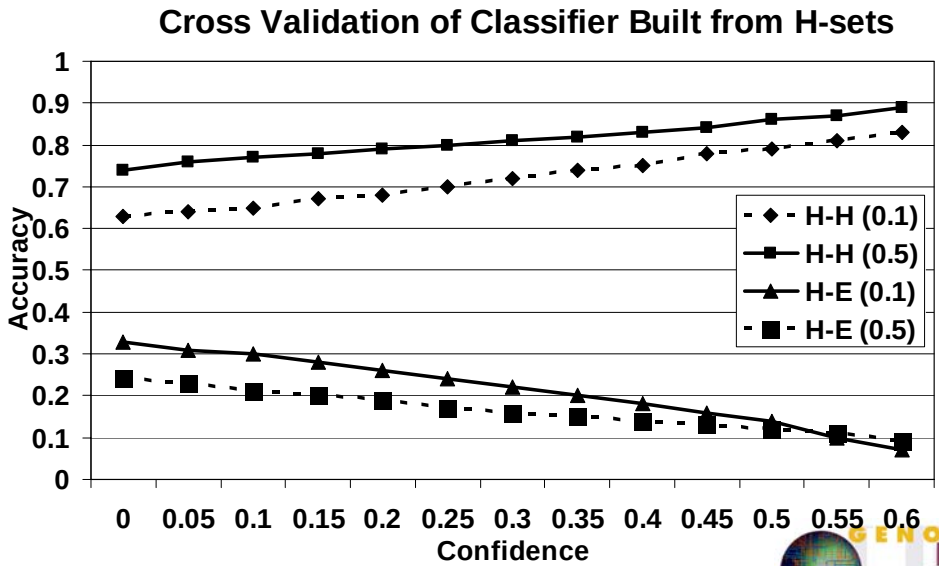


Performance Results

3-fold Cross Validation results show that E-set and H-set are really in conflict.







The future -*In Silico* omics

- **Convergence of omic data/informatics**
- **Critical mass offers opportunities for data mining and smart algorithms**
- **This leads to *in silico* prediction of protein functionality and interactions**
- ***In silico* functional omics increasingly depends on parallel processing**
- **Paradigm change in parallel processing: increasing not only speed but prediction accuracy**

Acknowledgements

- Geist, Al, ORNL
- Park, Hoony, ORNL
- Yu, Gong-Xin, ORNL
- Bernhard Palsson, UCSD
- Natalia Maltsev, ANL
- Gorin, Andrey, ORNL
- Ostrouchov, George, ORNL
- Chandramohan, Praveen, ORNL
- Choi, Huyk, ORNL
- Munavalli, Rajesh, ORNL

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.