



IBM



The background of the slide is a blue-tinted collage. On the left, a person is seen from behind, working at a computer with two monitors. The screens display a human figure and data. To the right, there are several 3D molecular models, including a DNA double helix with colored base pairs (A, T, C, G) and various protein structures represented as ribbons and spheres. The overall theme is the intersection of biology and computing.

# Computational Biology: Trends and Challenges in Computing

---

***IWLS 2004***

***Temecula, CA***

***June 3, 2004***

***Ajay K. Royyuru***

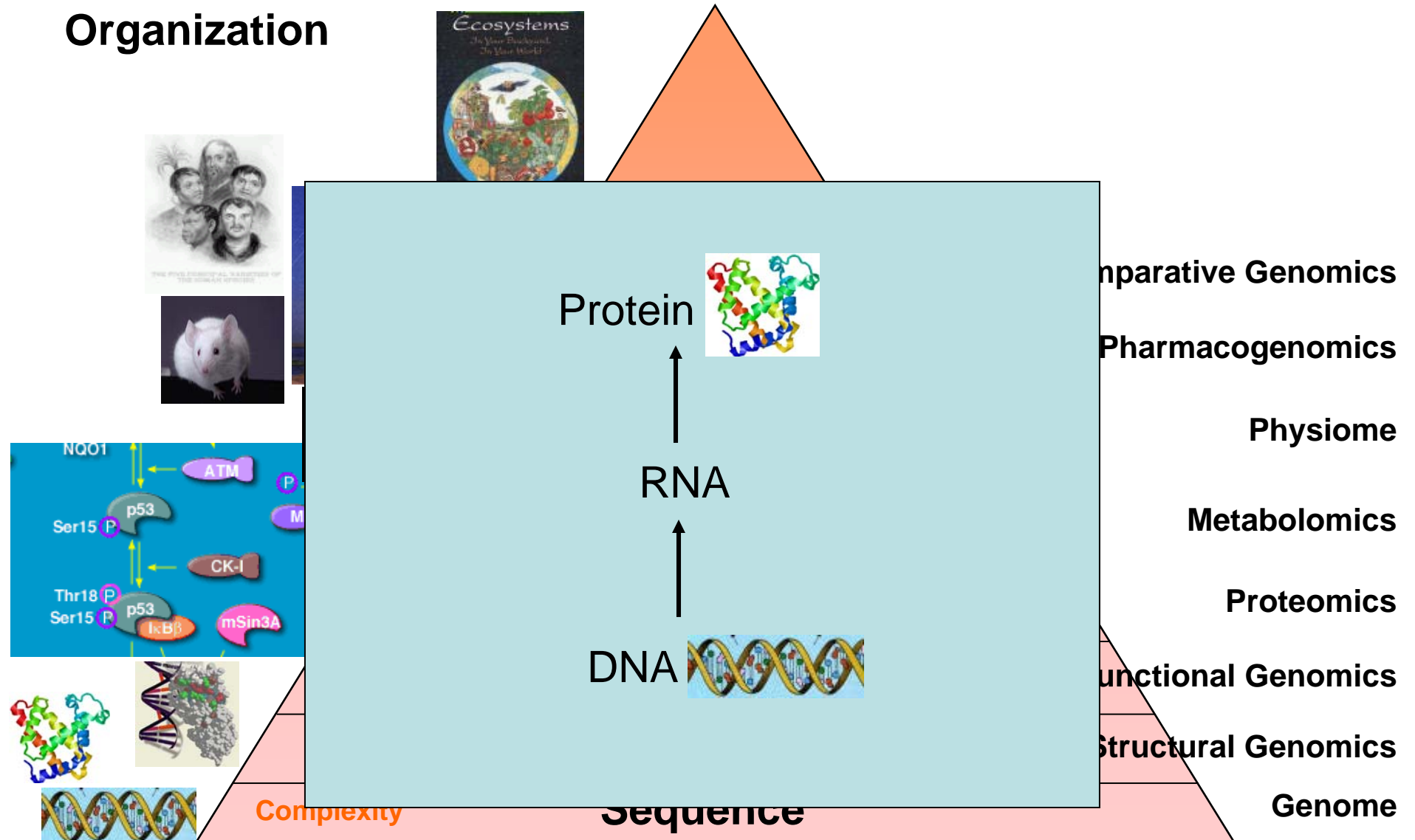
***Computational Biology Center***

***IBM Thomas J. Watson Research Center***

***ajayr@us.ibm.com***

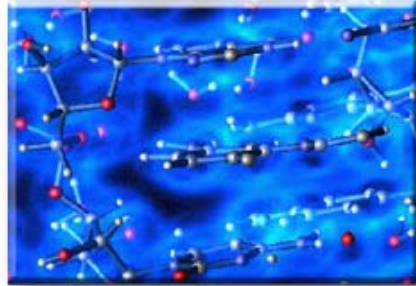
# Information in Biology

## Organization

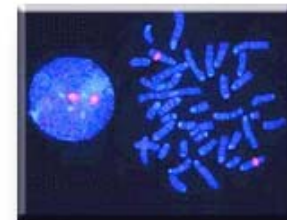
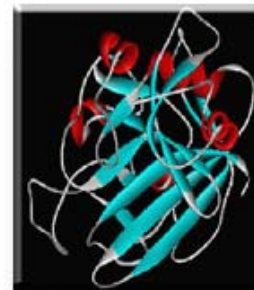
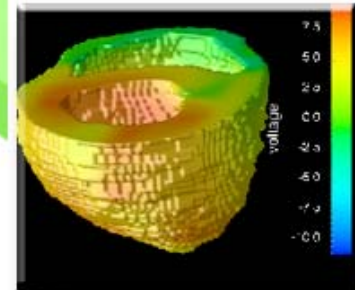
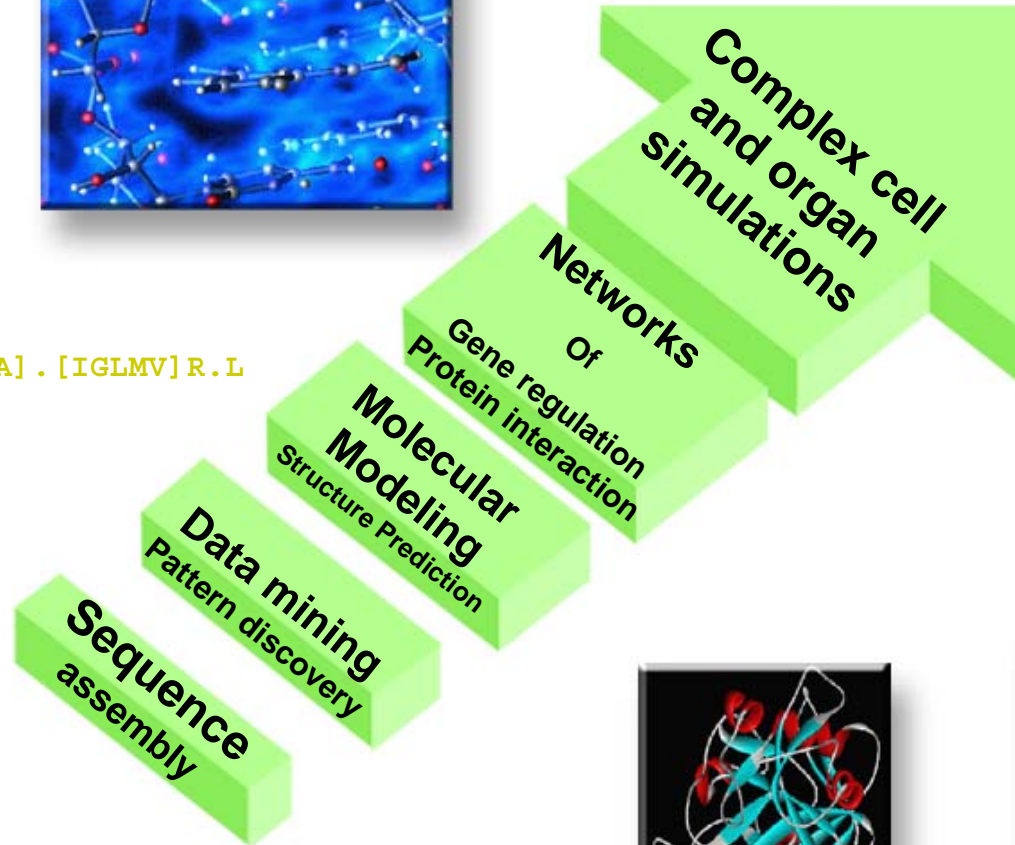


# Data Drives Computing

## Transforming Data Into Knowledge



G..G.GK [STG] TL  
H.....HRD.K..N  
SGG [QEMRY] ..R [VLIA] . [IGLMV] R.L  
V.I.G.G..G...A  
G.GLGL.I



# Comparative Genomics

*Nature* 428: 617 - 624 (08 April 2004)

Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*

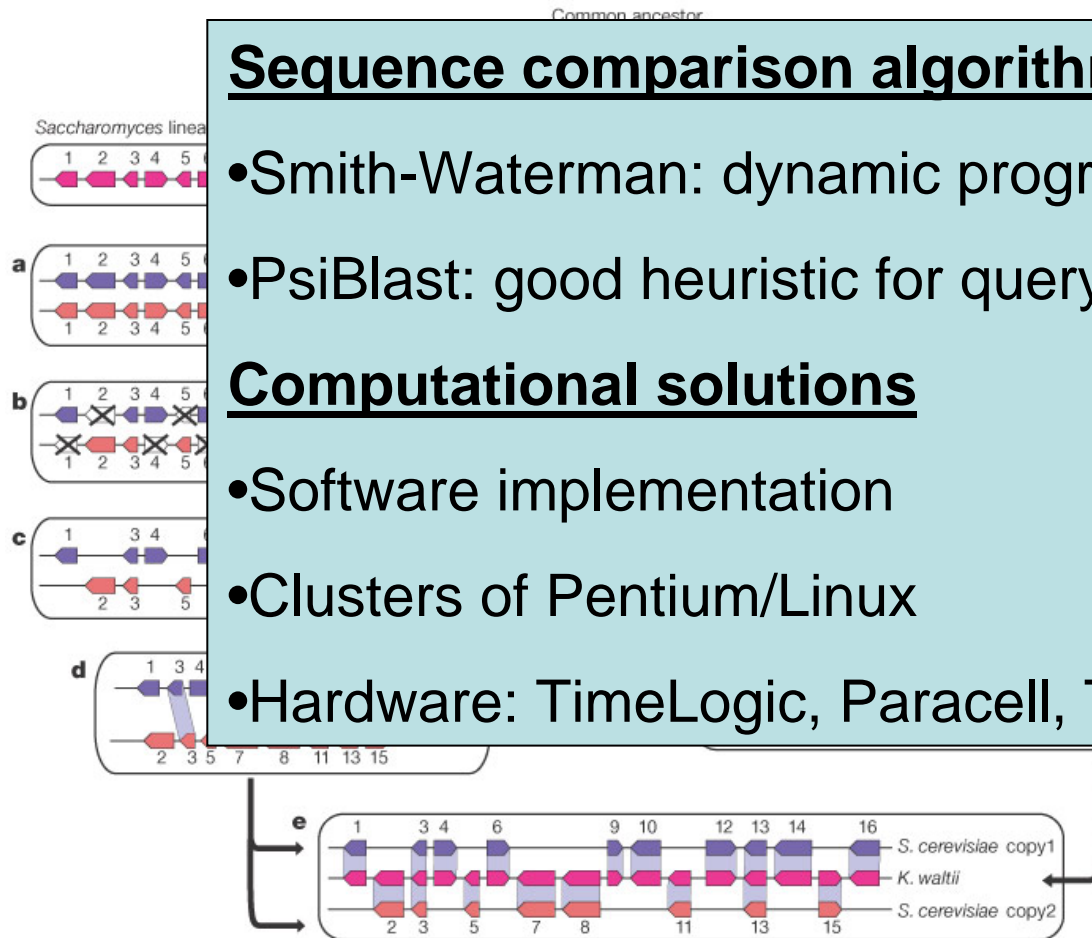
Manolis Kellis, Bruce W. Birren & Eric S. Lander

## Sequence comparison algorithms

- Smith-Waterman: dynamic programming for pairwise
- PsiBlast: good heuristic for query-vs-database

## Computational solutions

- Software implementation
- Clusters of Pentium/Linux
- Hardware: TimeLogic, Paracell, TurboWorx



... massive gene  
... ster regions. **a**,  
*Saccharomyces*  
... cation event,  
... and  
... of duplicated  
... e loss. **c**, Sister  
... s of the original  
... ly a small  
... were retained  
... *S. cerevisiae*, the  
... served order of  
... duplicated genes (numbered 3 and 13) across  
... different chromosomal segments; the intervening  
... genes are unrelated. **e**, Comparison with *K. waltii*  
... reveals the duplicated nature of the *S. cerevisiae*  
... genome, interleaving genes from sister segments  
... on the basis of the ancestral gene order.

# Common Problems in Bioinformatics

- **automated determination of PROSITE-like motifs**
- **multiple sequence alignment**
- **full-genome comparisons**
- **unsupervised clustering of SwissProt/NRDB**
- **tandem repeats in DNA**
- **homology searching using patterns**
- **3D-structure from 1D-sequence**
- **function from 1D-sequence**

# Teiresias – Pattern Discovery Algorithm

## Unsupervised pattern discovery on biological sequences

Algorithm allows the discovery of:

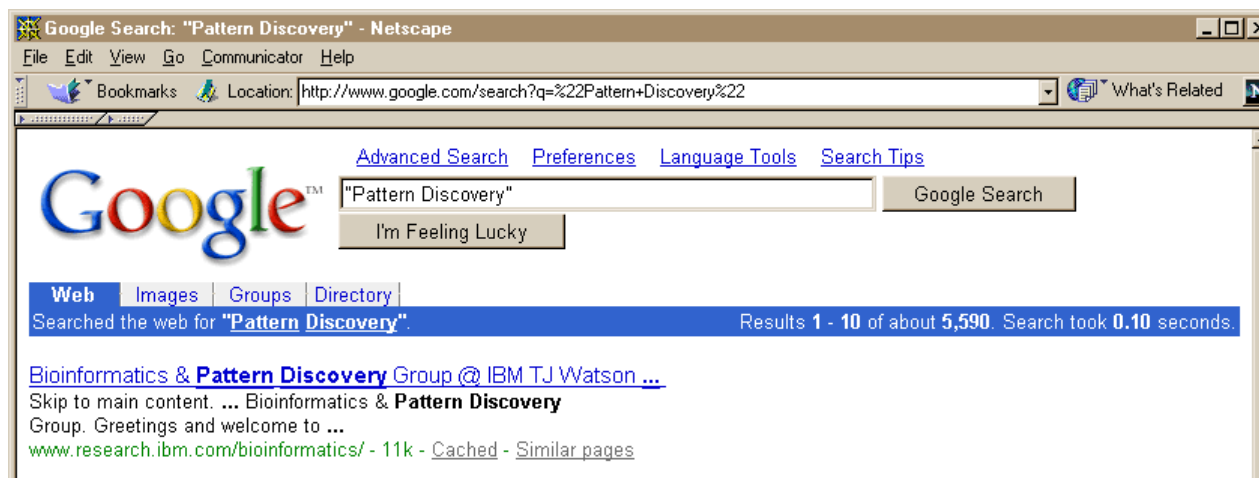
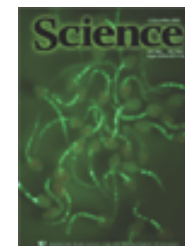
- *all* patterns appearing at least  $k$  times in the input
- patterns that in their most general form consist entirely of bracketed expressions, e.g.  
[NS][LIMYT][FYDN].[DNT][IMVY].[STGDN][DN]..[SGAP]
- patterns that are *maximal* in *composition*
- patterns that are *maximal* in *length*

# BioDictionary

#	"Word"	"Meaning"
1	G..G.GK[STG]TL	ABC transporters
2	H.....HRD.K..N	Ser/Thr-protein kinases
3	SGG[QEMRY]..R[VLIA].[IGLMV]R.L	ABC transporters
4	V.I.G.G..G...A	NAD/FAD-binding, Flavoproteins
6	G.GLGL.I	Sensory transd. His-prot. kinases
...		
10	GA.DY[LIV].KP	2-component sensory transduction
11	HR.GR..R....G	DEAD-box helicases
12	GDG[IVAMTD]ND[AILV][PEAS][AMV][LMIF]..A	Cation-transporting ATPases
13	D.FK.[IYVFL]N[DE].[YLFWR]GH..GD.[CLVF]L	Bacterial-type regulators*
14	DKT[GV]TLT	Cation-transporting ATPases
...		
16	KMSKS[LKDIR][GNDFQ]N	Amino-acyl-tRNA synthetases I
17	PTREL..Q	DEAD-box helicases
18	Q..GRAGR	DEAD-box helicases
19	F.[ASDN].[MIVTLA][SAT]HE[LIF]RTP	Sensory transd. His-prot. kinases
...		
27	DL[IVL][LIMVF]LD[ILVW].[ML]P..[DNST]G	2-component sensory transduction
...		
30	LD.GCG.G	Various methyltransferases
...		
34	T.[IVL][FLYMI]VTHD[QLIVP].[ELV]A	ABC transporters
...		
..		

# Bioinformatics Algorithms

- [www.research.ibm.com/compsci/compbio](http://www.research.ibm.com/compsci/compbio)
- No charge limited license for non-profit use
- Algorithms used to annotate >150 microbial genomes
- Site of the Month - Nov 2001, by Science
- Licensed to several major pharmas; mirrored at major univs
- Genome, annotation of *Ciona intestinalis*, Science, 298:2157 (2002)



# Protein Folds

Classification of protein folds

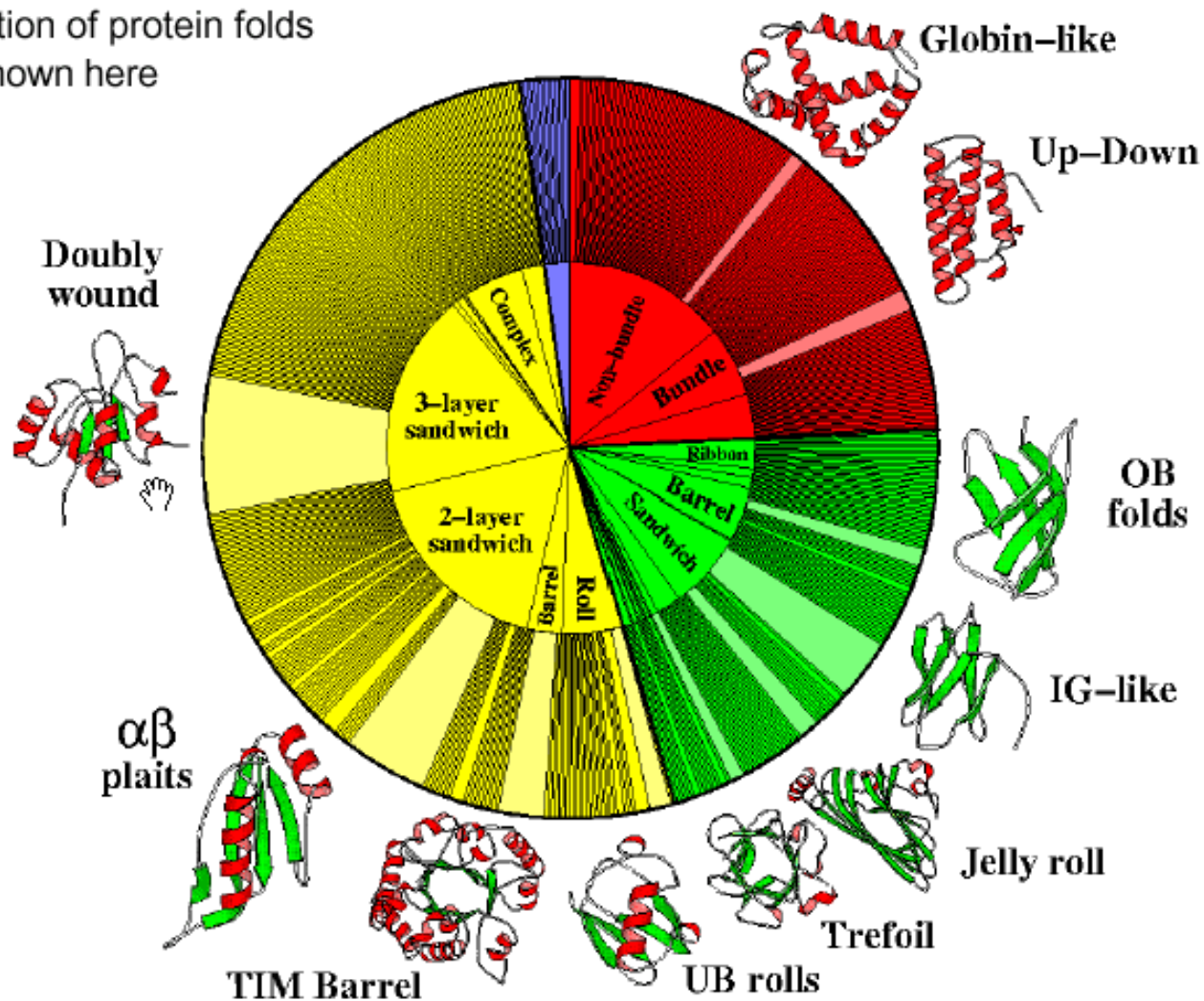
CATH - shown here

SCOP

FSSP

VAST

CE



# Protein Structure Prediction

## Comparative Modeling

### Psi-Blast

```

Psi-Blast to pdbs95
Query:  1 MTNPKVFFDMAIAGNPAGRI 20
        M  KVFFD+ I G  +GRI
Sbjct:  1 MVRSKVFFDITIGGKASGRI 20
    
```

## Fold Recognition

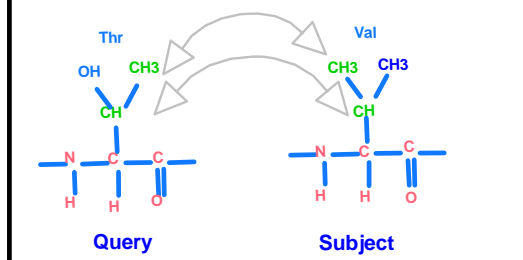
### CAFASP3

```

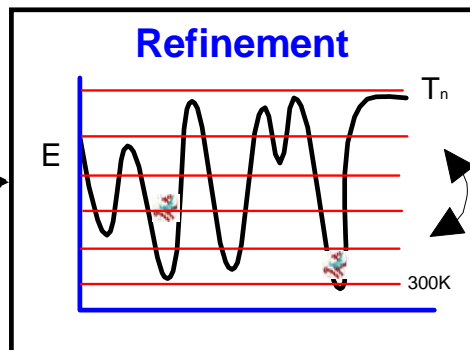
3D Jury
Query  : NRKWGLCIVGMGRLGSALAD
MGTH3  : DRRFGYAIVGLGKYALNILP
ORFS   : DRRFGYAIVGLGKYANQILP
PSSM   : DANIRVAIAGAGRMGRQLIQ
FUG    : ..MVNVAVNGYGTIGKRVD
    
```

Alignment

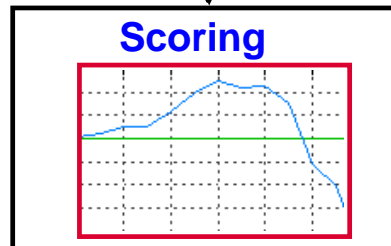
### All Atom Model



### Refinement

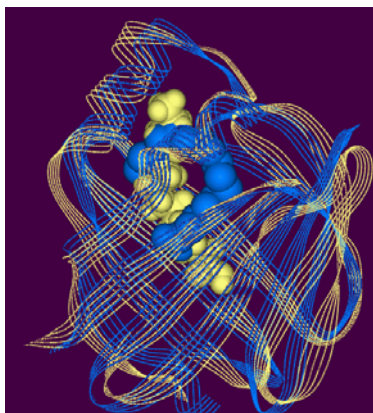


### Scoring

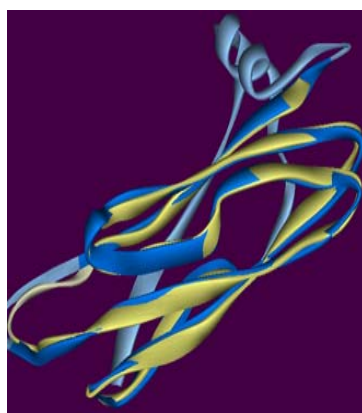


# CASP5 Examples

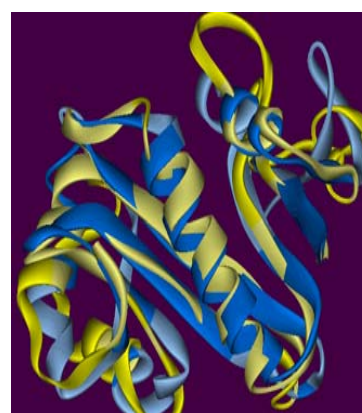
## Comparative Modeling:



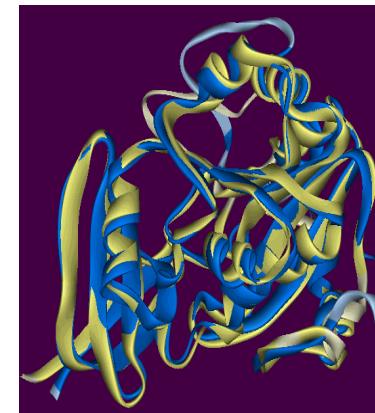
T137 Model



T160 Model

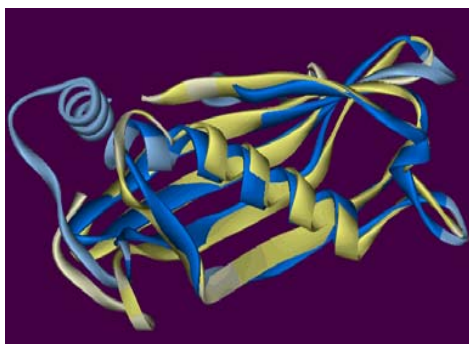


T169 Model

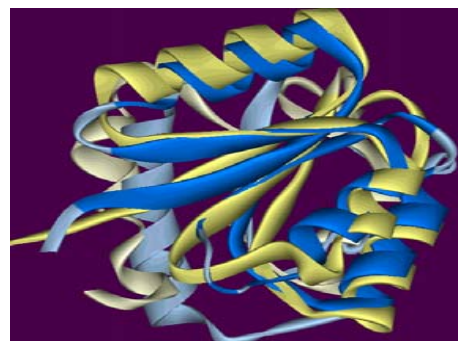


T179 Model

## Fold Recognition:



T132 Model



T138(1m2e\_A) Model

### Color legend

Target structure

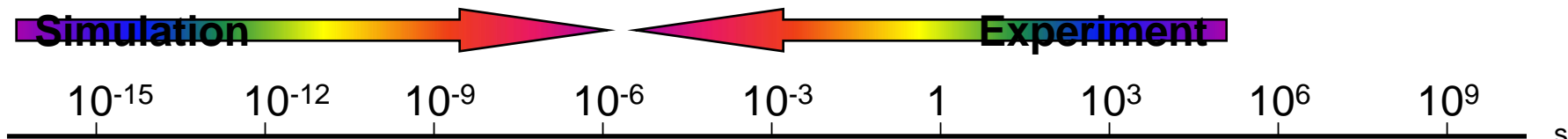
experimentally determined

revealed after CASP5 closes

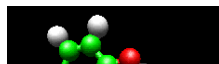
Model

Prediction submitted to CASP5

# Time scales in molecular motions



• Bond Vibration



## Molecular Dynamics

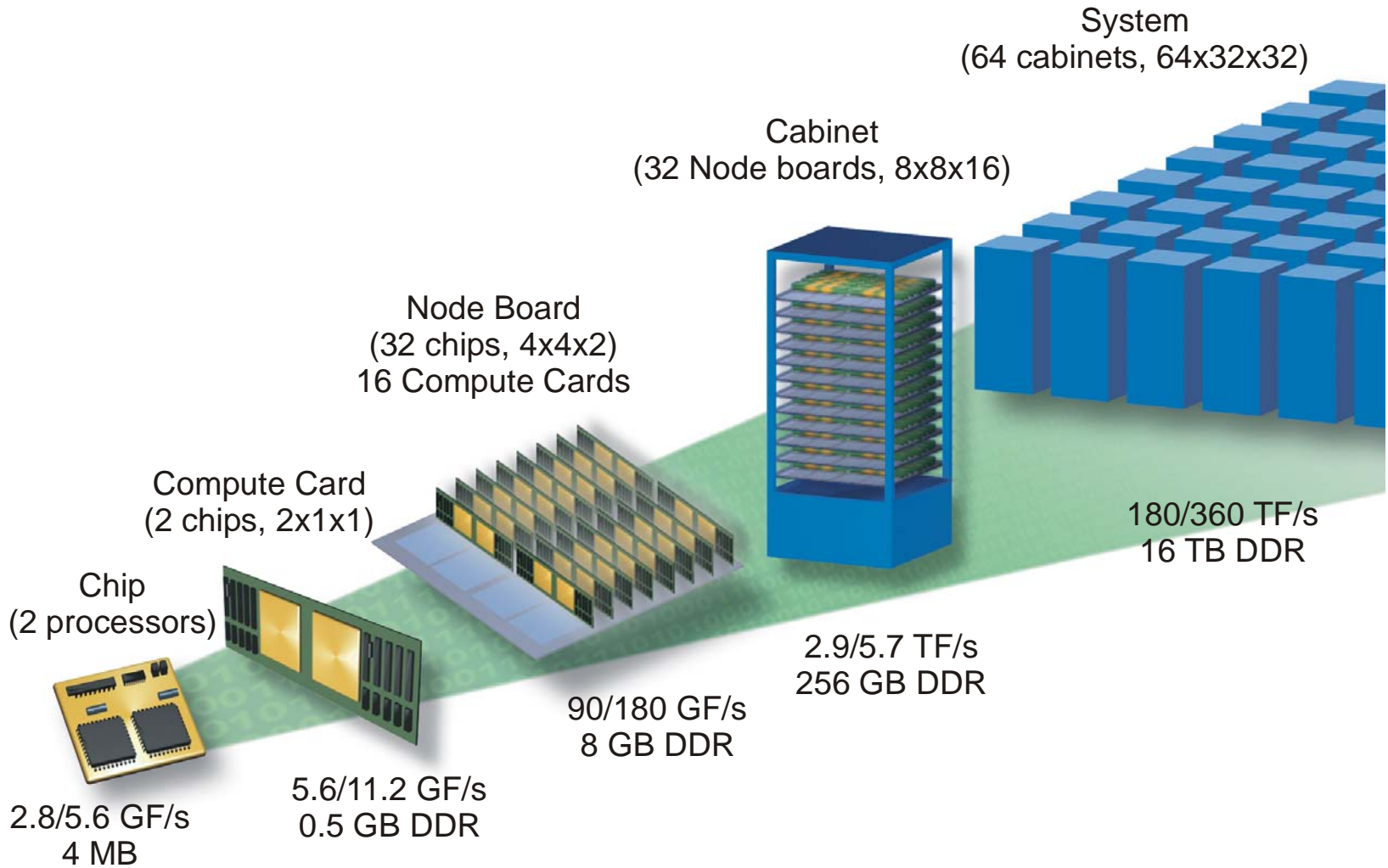
- Atoms represented as points, with physical characteristics – mass, charge, van der Waals, ...
- Forces computed for all-to-all atomic interactions, with bonded and non-bonded interactions
- Newton's equation of motion integrated with 1-2 fs time step
- Often used in conjunction with other conformational sampling techniques like Monte Carlo, Replica Exchange, ...

• Electron Transfer

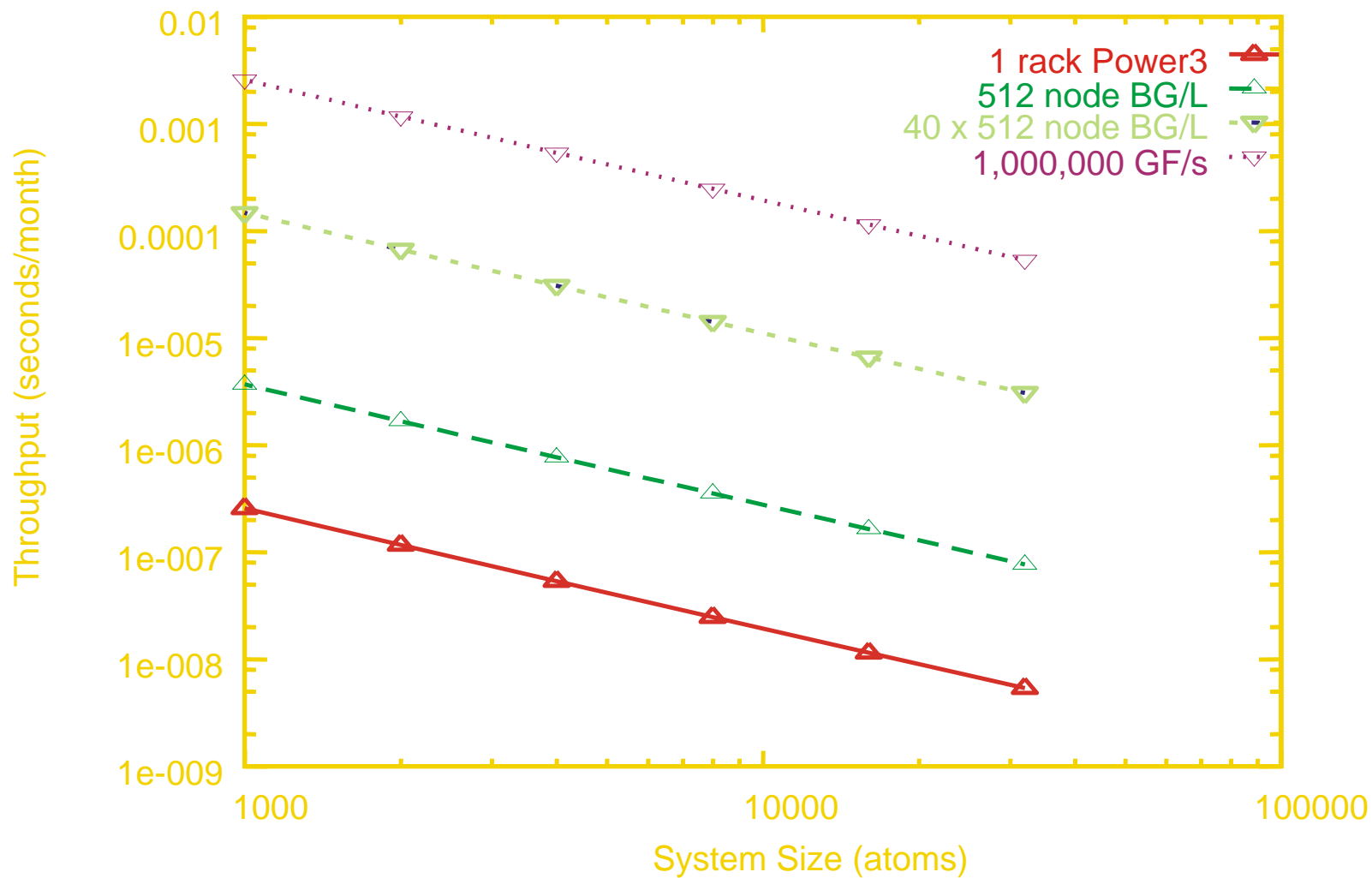
Torsional correlation in lipid headgroups



# Blue Gene/L



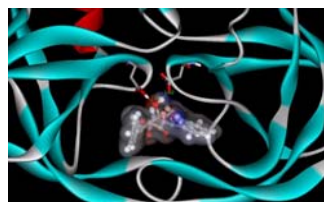
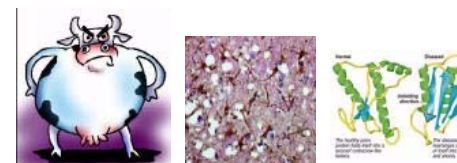
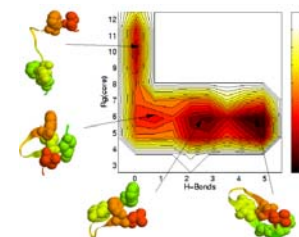
# Computational Throughput (Capacity)



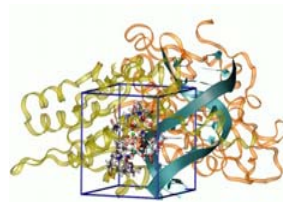
# Protein science on Blue Gene

## Science Goals

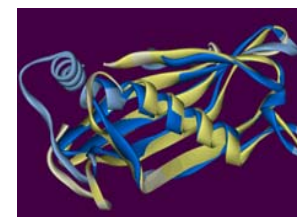
- Large scale studies of protein folding **thermodynamics** and long time scale **kinetics**
- Identifying **disease related processes** whose critical stage pathway can be studied using large scale simulation
- Connect with **experimental data**
- Advances in our understanding of biomolecular simulation can be **applied** to a variety of related problems, including:



- ▶ Drug protein interactions (docking)



- ▶ Enzyme catalysis with hybrid quantum methods

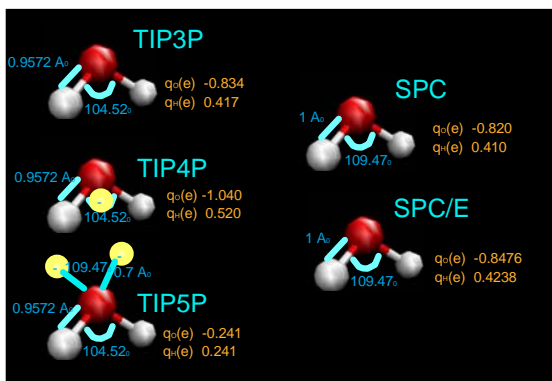
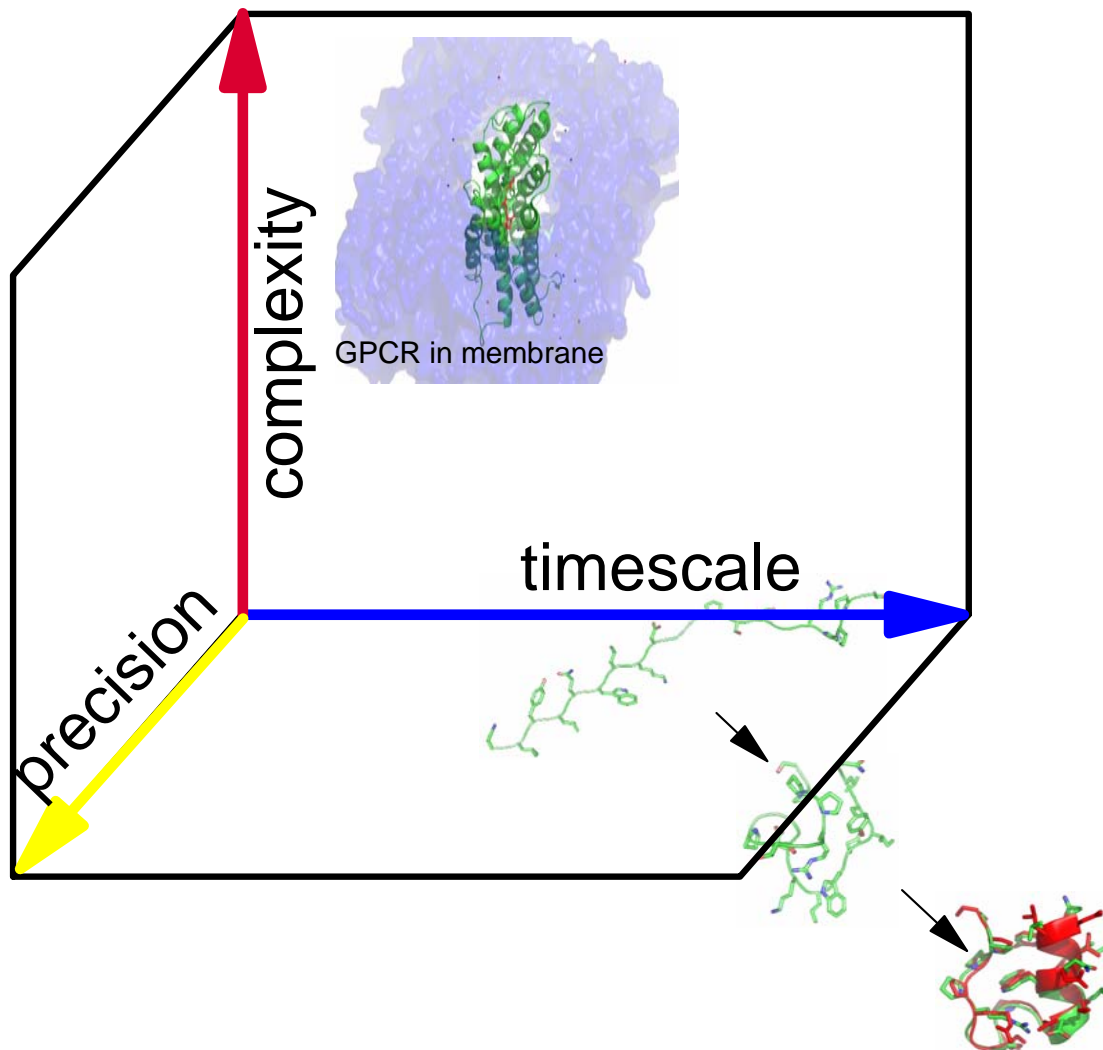


CASP5 T132 Model

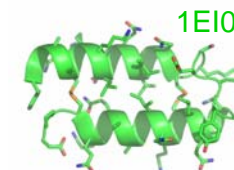
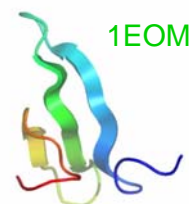
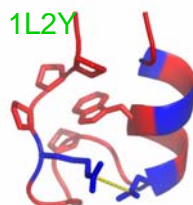
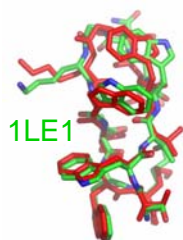
- ▶ Structure refinement and scoring for protein structure prediction

# Biomolecular Simulations

## Scaling directions



# Simulations – A Spectrum of Projects

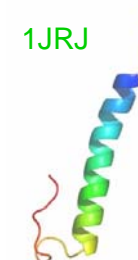


- systematically cover a range of system sizes, topological complexity

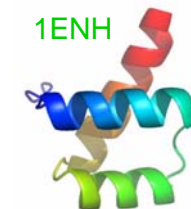
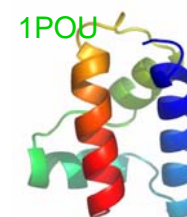
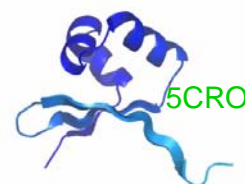
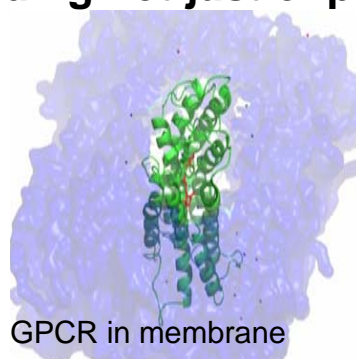
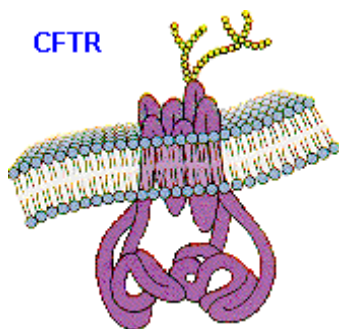
- ▶ discovering the "rules" of folding
- ▶ applying those rules to have impact on disease

- cover a broad range of scientific questions and impact areas:

- ▶ thermodynamics
- ▶ folding kinetics
- ▶ folding-related disease (CF, Alzheimer's, GPCR's)

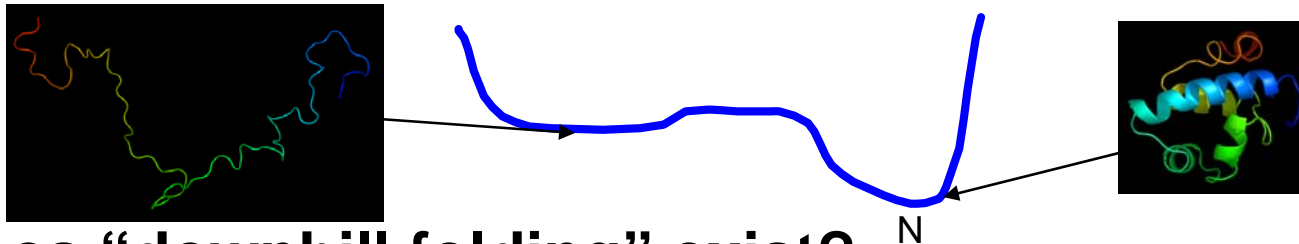


- improve our understanding not just of protein **folding** but protein **function**



# Protein folding: Outlook

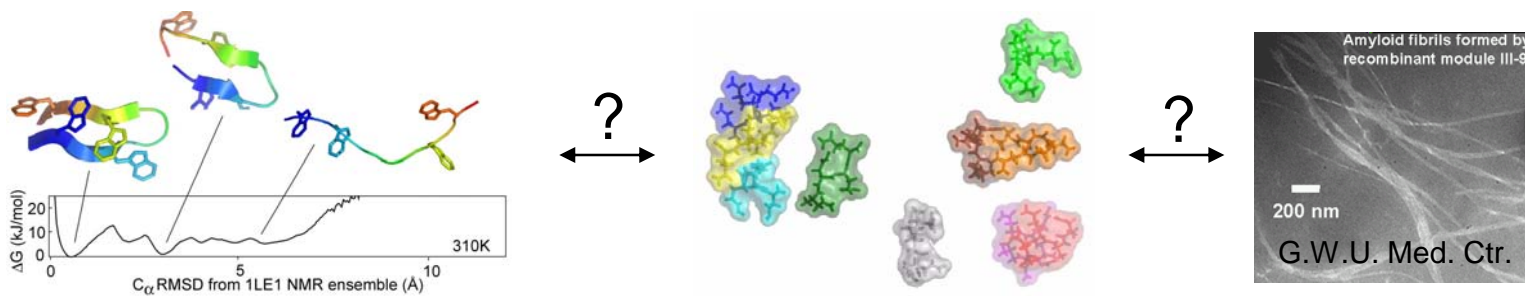
- Why is there a speed limit to protein folding?



- Does “downhill folding” exist?

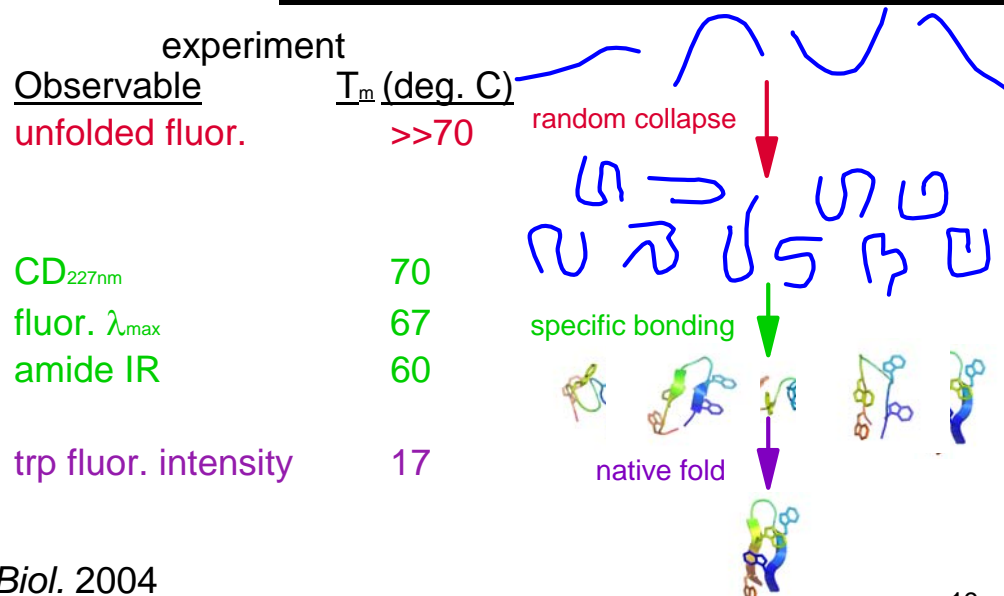
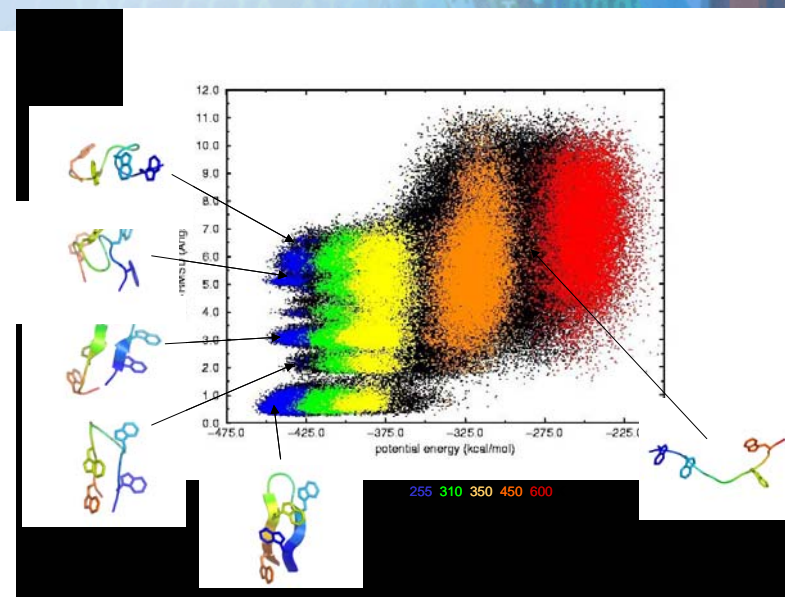


- How does roughness of the energy landscape affect folding & aggregation?



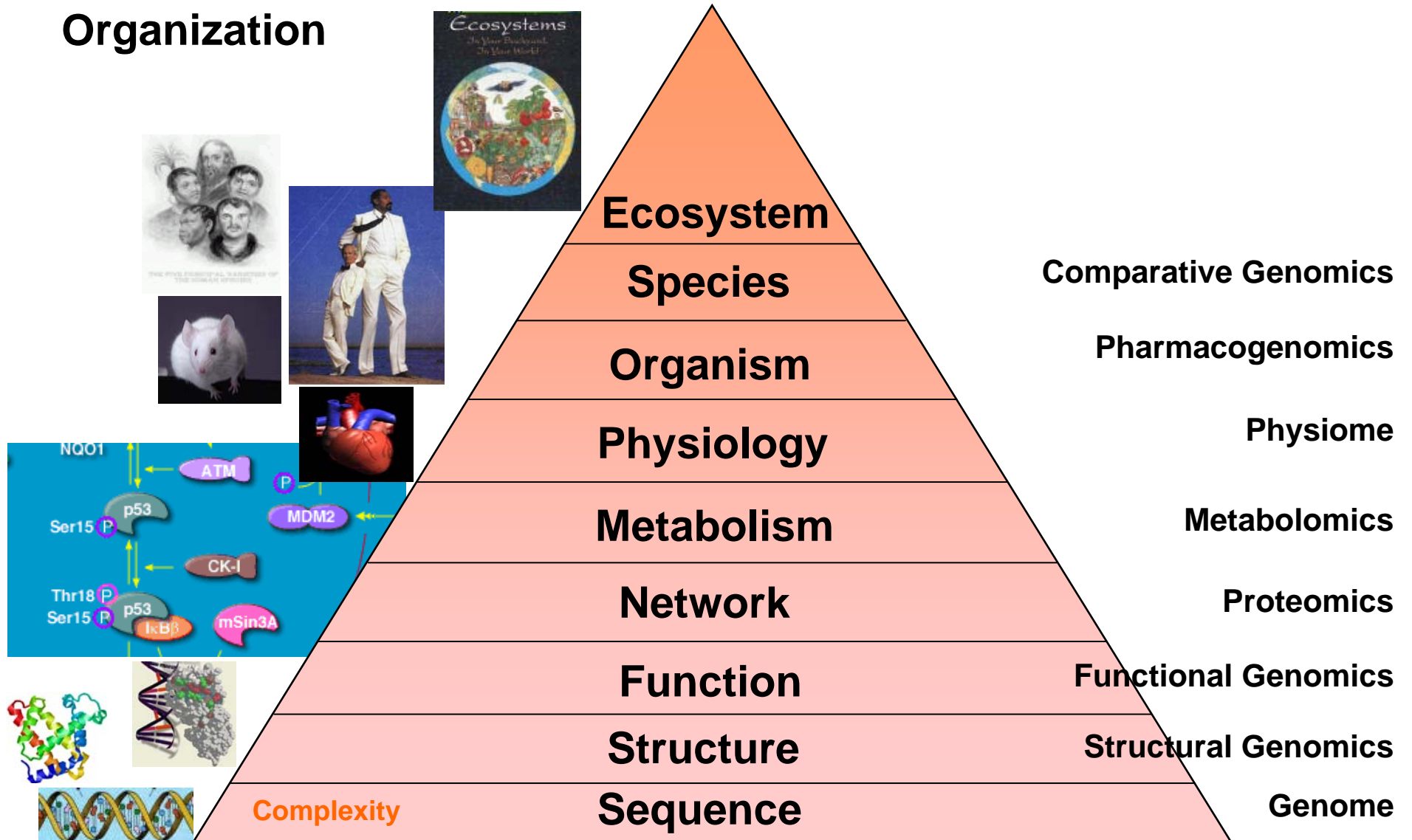
# trpzip2 folding

- Collaboration with Martin Gruebele and Wei Yang, UIUC
- UIUC group approached IBM, unable to explain their experiments
  - Different spectroscopic techniques yielded different melting temperatures ( $DT_m > 50$  K)
- Combined simulation + experiment to explain the observations



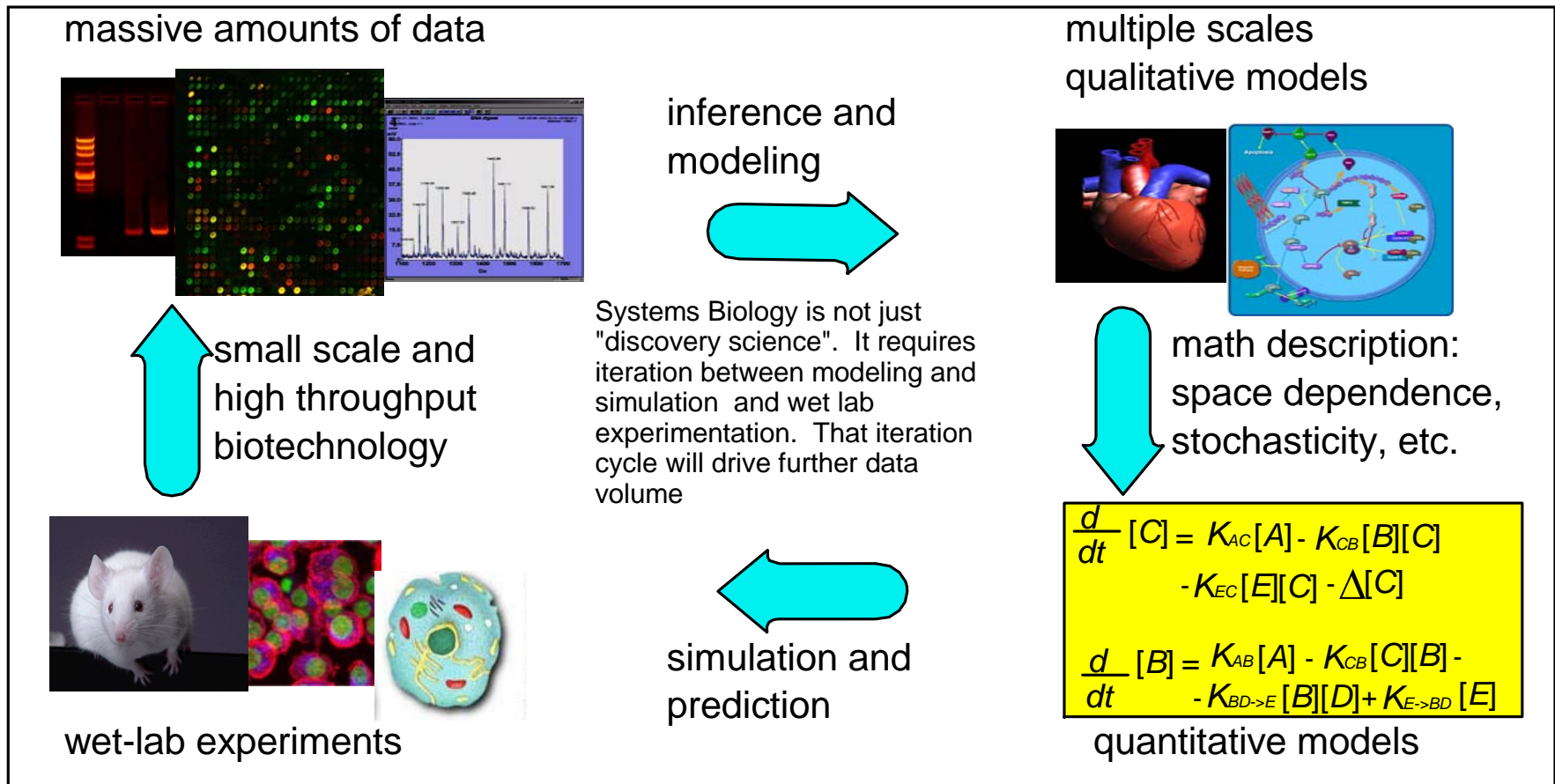
# Information in Biology

## Organization



# Systems Biology

Predictive simulations of complex biological problems





**IBM Thomas J. Watson Research Center, Yorktown Heights, NY**

**IBM Computational Biology Center**  
**[www.research.ibm.com/compsci/compbio](http://www.research.ibm.com/compsci/compbio)**

**IBM Healthcare and Life Sciences**  
**[www.ibm.com/solutions/lifesciences](http://www.ibm.com/solutions/lifesciences)**