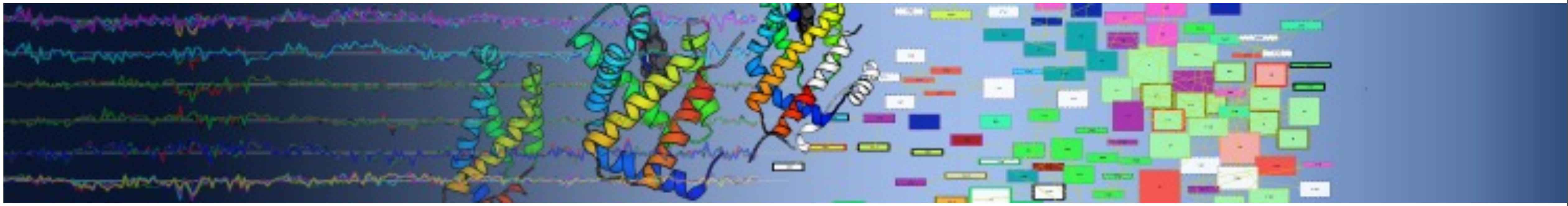
Reconstructing biological networks from data: cMonkey & Inferelator



Richard Bonneau

bonneau@nyu.edu

http://www.cs.nyu.edu/~bonneau/

New York University,

Dept. of Biology &

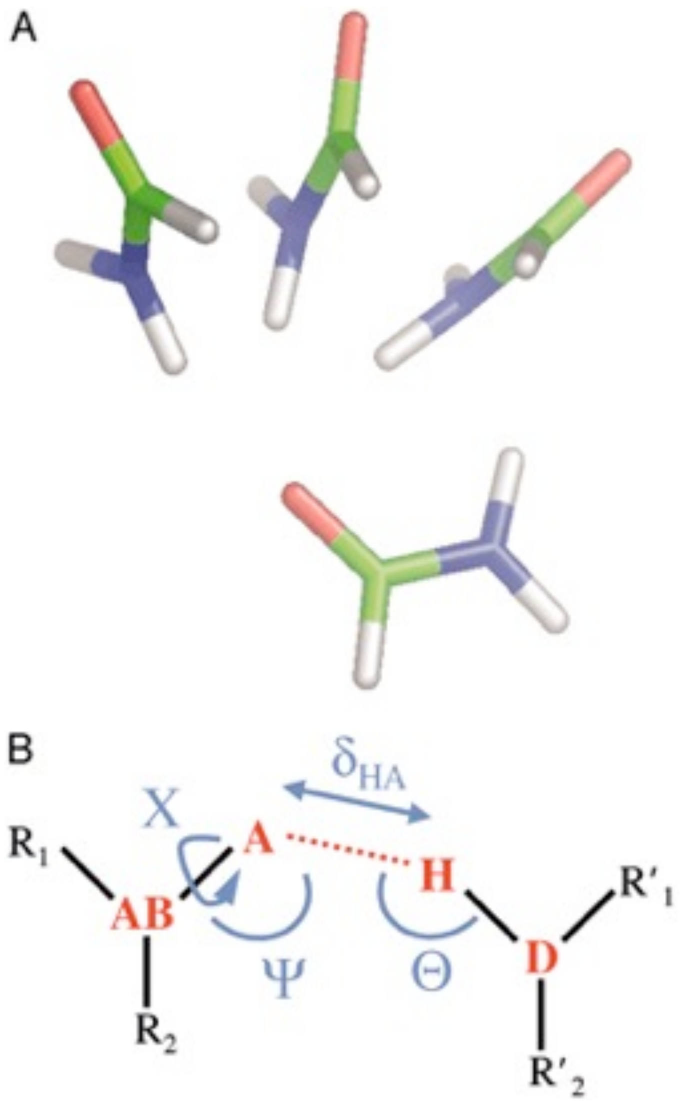
Computer Science Dept.

CENTER FOR GENOMICS
AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY

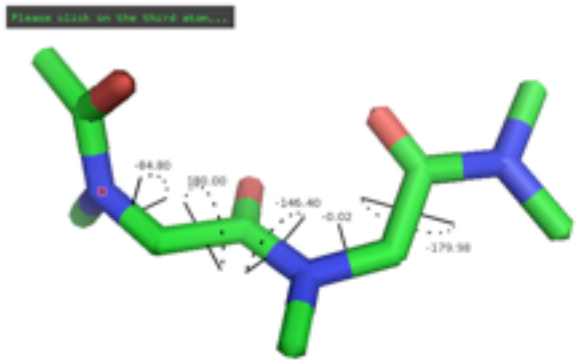
COURANT
INSTITUTE

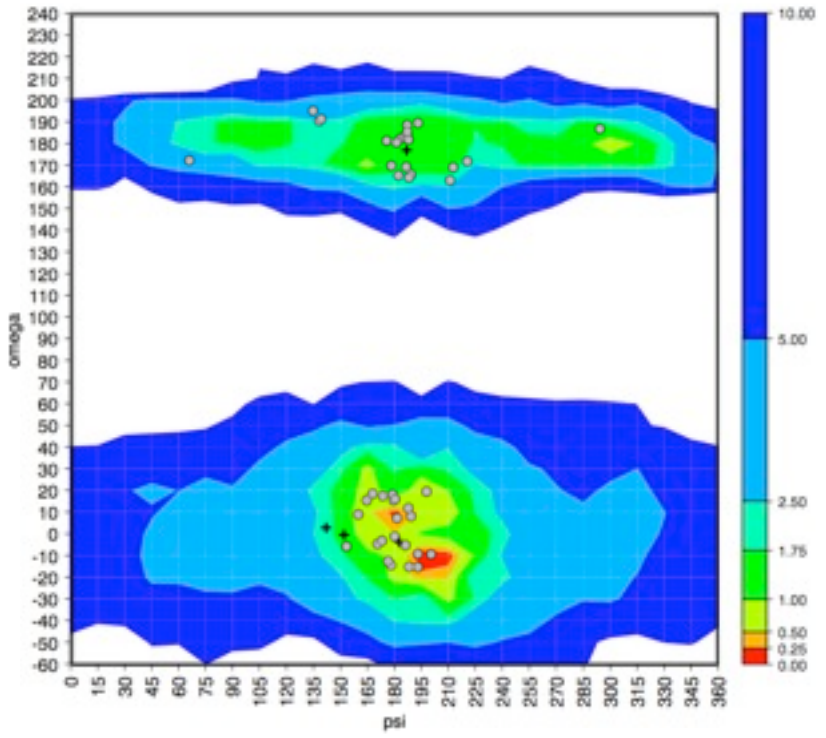
justification of functional-from-principles, parameters-from-data by example

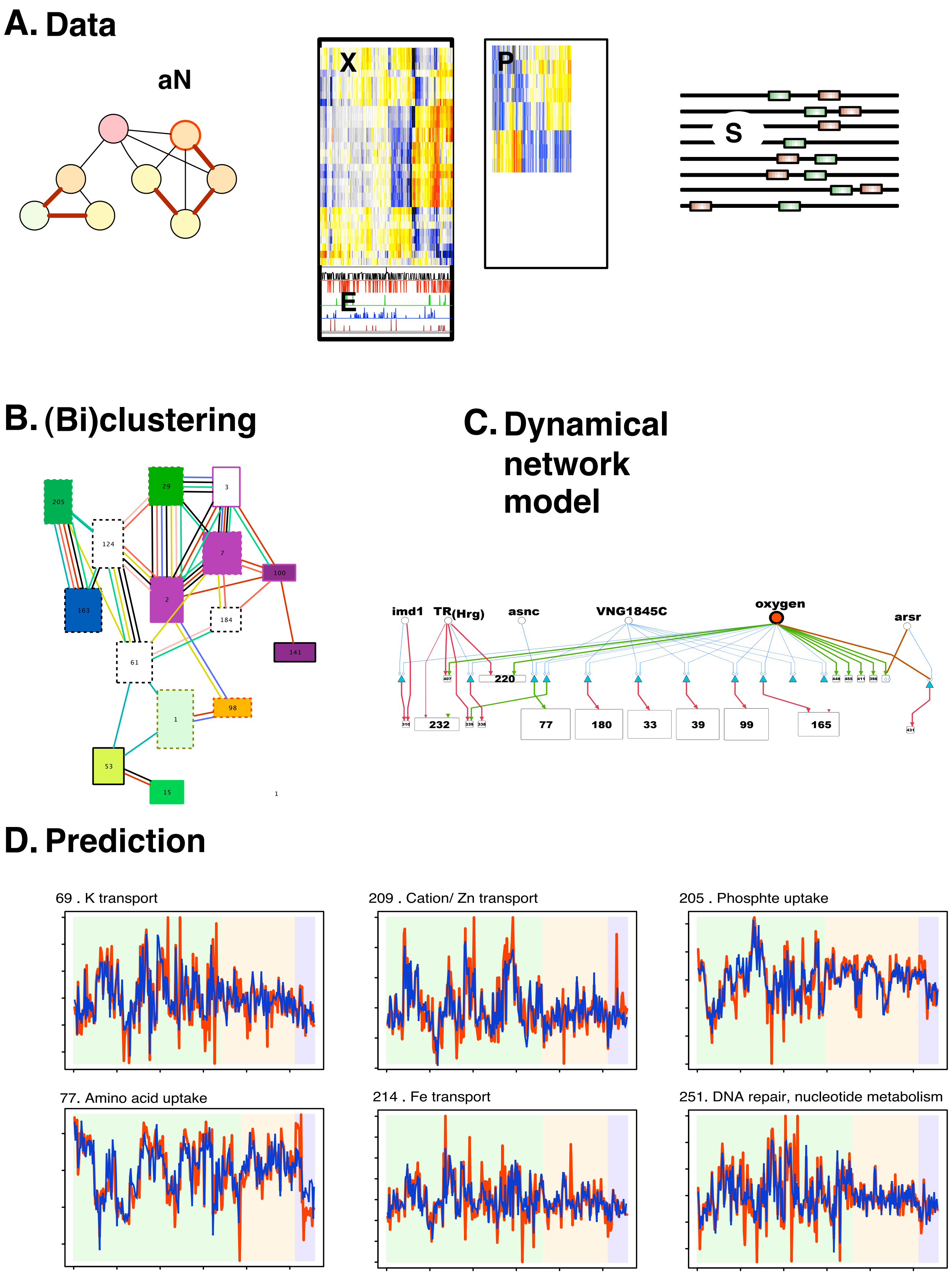
Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Morozov, Kortemme, Baker, PNAS, 2003



Oligo(N-aryl glycines): A New Twist on Structured Peptoids, Shah, Butterfoss, Bonneau, Kirshenbaum, 2008, JACS


Please click on the third atom...




A. Data

B. (Bi)clustering

C. Dynamical network model

D. Prediction

Overview

1. co-regulated modules (integrate data types).

2. Learn topology and Dynamics with greedy / local aprox. (inferelator 1.0, 1.1)

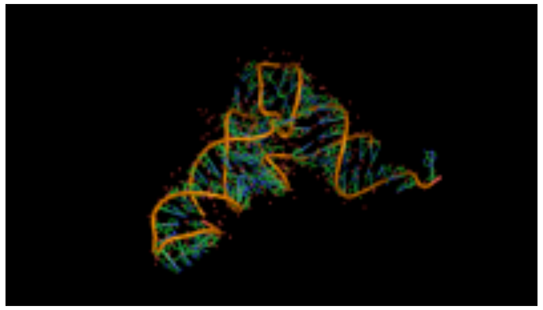
3. improving performance over multiple time-scales (Inferelator 2.x)

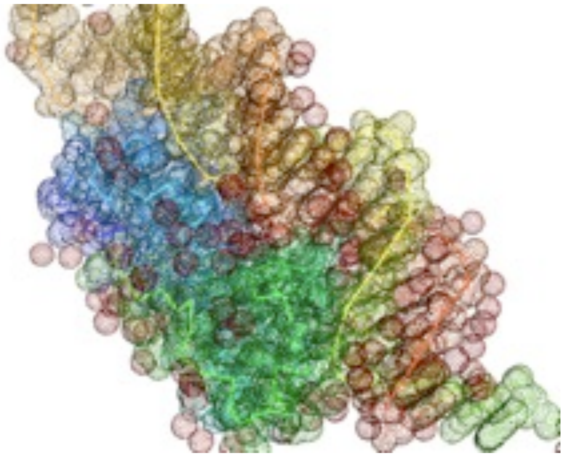
Main results:

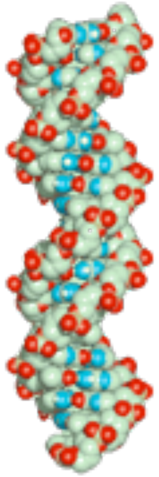
- Surprising predictive performance for prokaryotic networks, T-cell and macrophage differentiation EE Networks

- Longer time scale stability

- model flexibility

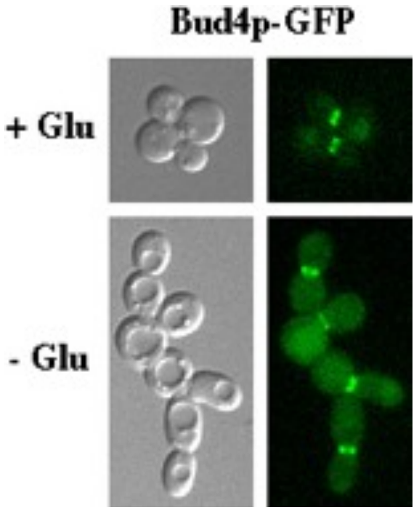






DNA → RNA → PROTEIN → PHENOTYPE

AUTOMATED MICROSCOPY, ETC. ★



http://www.molbio.uoregon.edu/images/research/sprague1.jpg

★ GENE SEQUENCING
WHOLE GENOME ASSEMBLY

★ CHIP-CHIP
TF-DNA BINDING
EXPERIMENTS

★ MICROARRAYS

cDNA
ESTs

LIBRARIES OF
FUNCTIONAL RNA

PROTEIN
SEQUENCE
DATABASES,

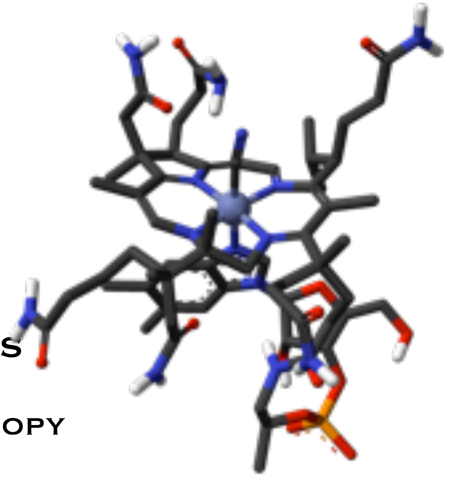
PROTEIN STRUCTURES,

★ PROTEOMICS

★ PROTEIN-PROTEIN INTERACTIONS

★ METABOLOMICS

MASS-SPECTROSCOPY
NMR
CHROMOTOGRAPHY



★ GENOTYPE & SEQUENCING

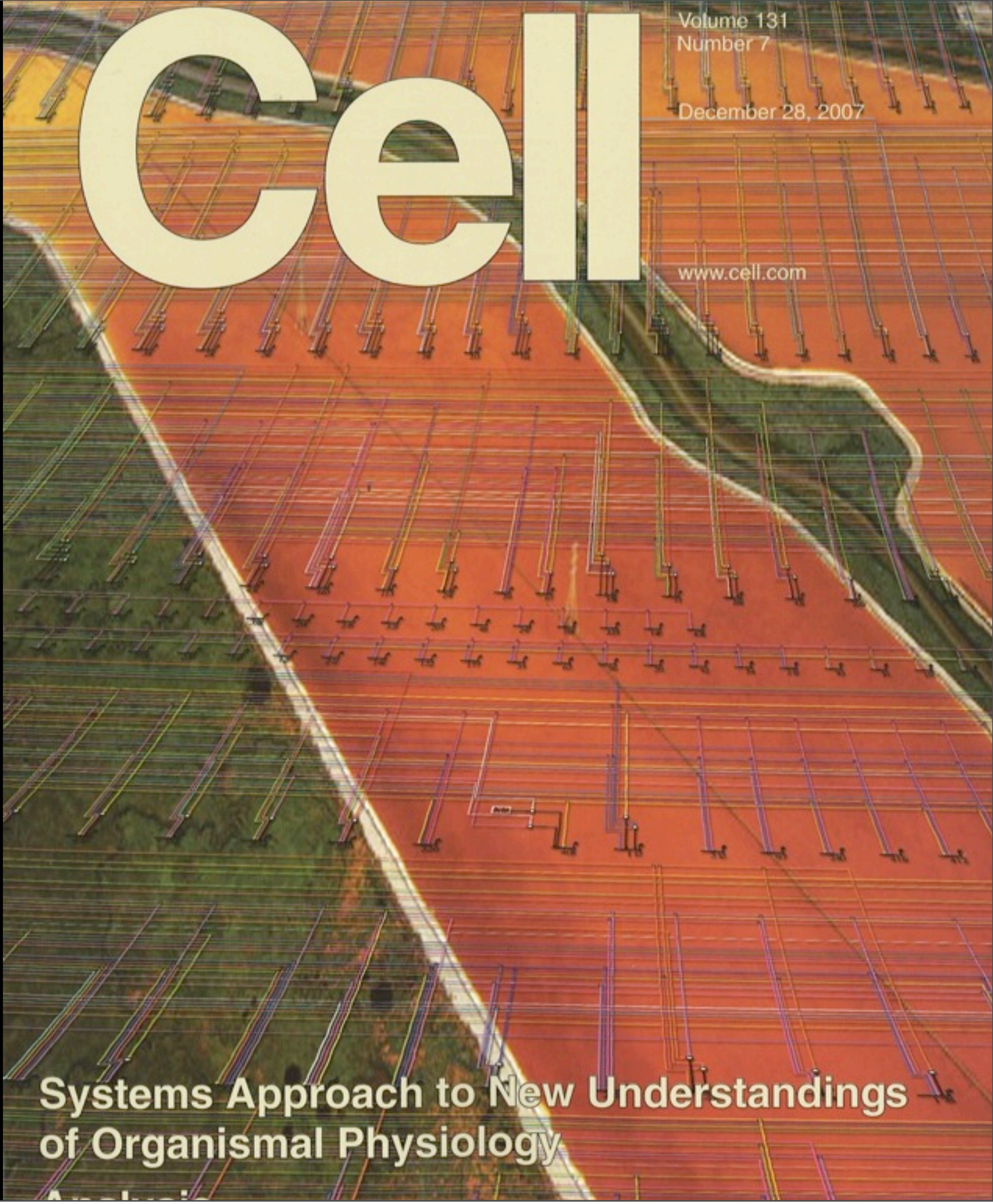
★ MEASURING AFFINITIES / BINDING

★ MEASURING LEVELS

★ ASSAYING FUNCTIONAL OUTCOME

algorithms:
David J. Reiss (cMonkey)
Vesteinn Thorsson (Inferelator)
Richard Bonneau

functional genomics:
Marc T. Facciotti
Amy Schmid,
Kenia Whitehead
Min Pan, Amardeep Kaur,
Leroy Hood
Nitin S. Baliga


Cell
Volume 131
Number 7
December 28, 2007
www.cell.com
Systems Approach to New Understandings of Organismal Physiology
Analysis

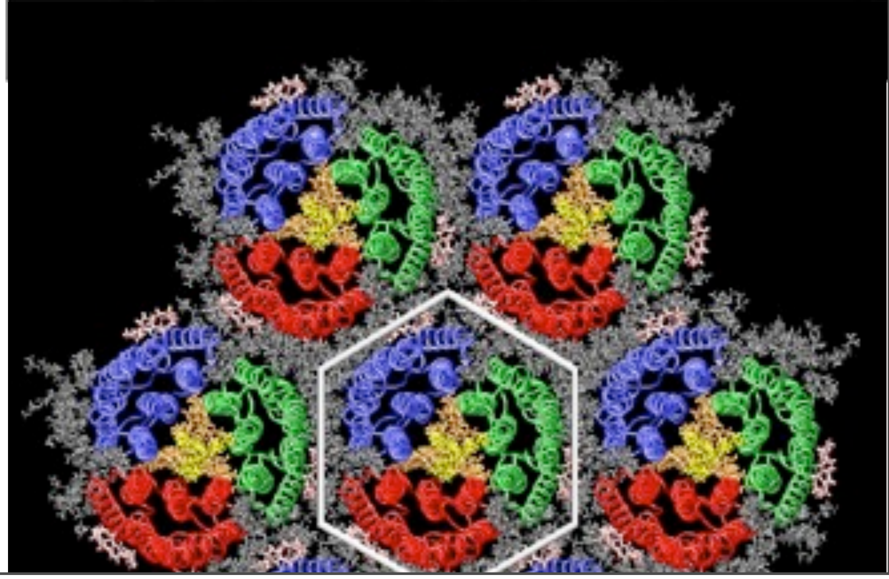
An example : Halobacterium

why halobacterium:

- if your friends are working on halo ... (Hood, Baliga)
- not a “model” system (originally)
- high IQ
- diverse environment
- small genome
- good genetics, cultivable, etc.
- a very tough extremophile, bioengineering

Data collection and modeling effort

- genome and genome annotation
- microarrays
- genetic and environmental perturbations
- proteomics
- ChIP-chip
- some protein-protein



Halobacterium dataset including

>800 microarrays

time series

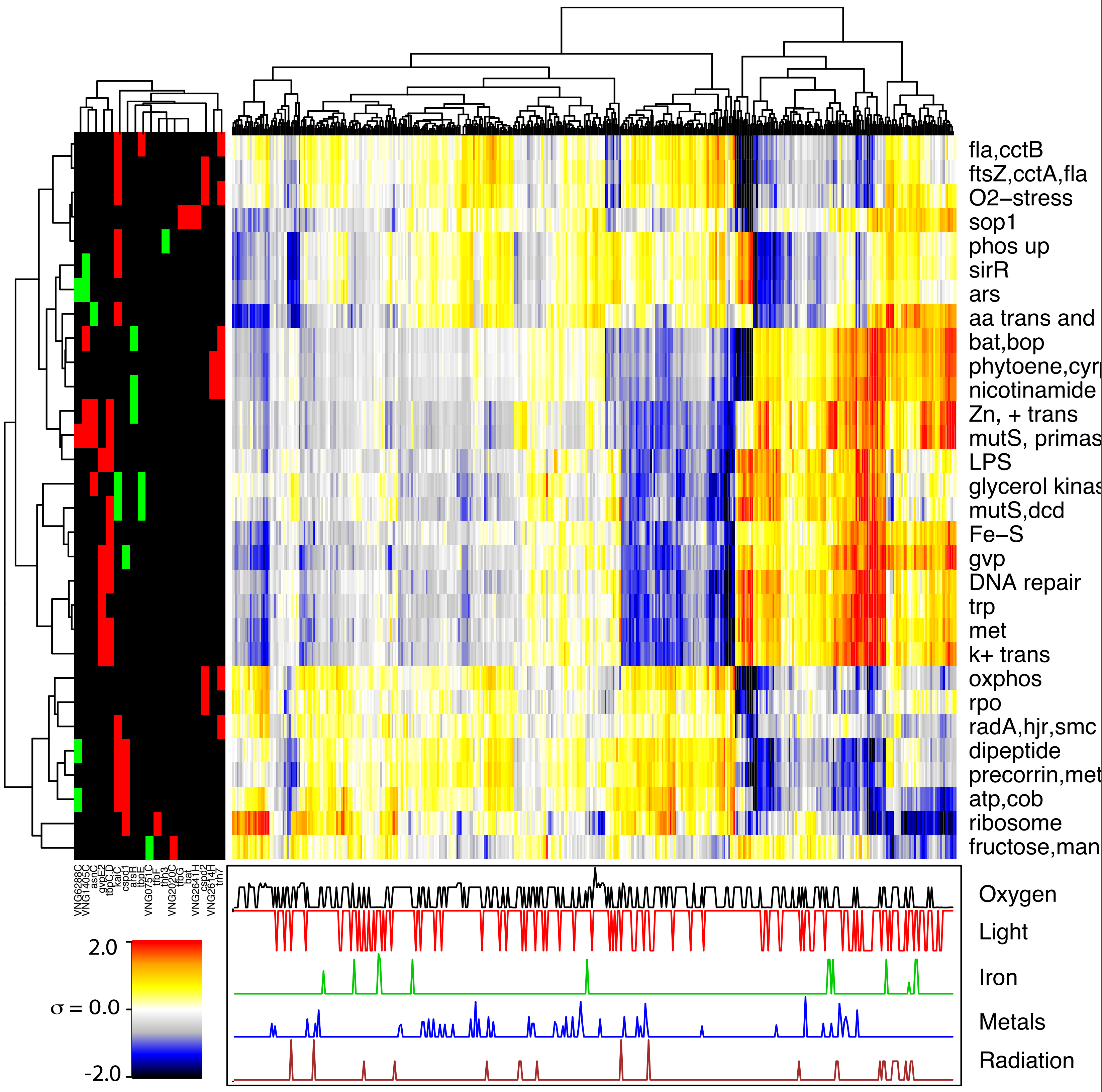
knock outs

ChIP-chip experiments

proteomics

phenotype

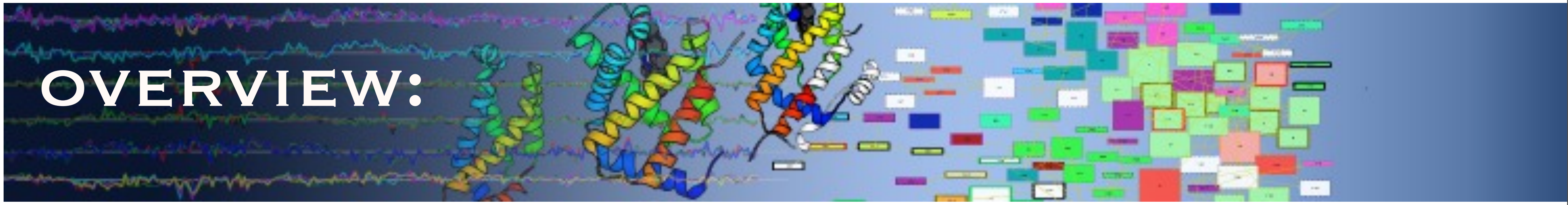
among the most complete prokaryotic datasets

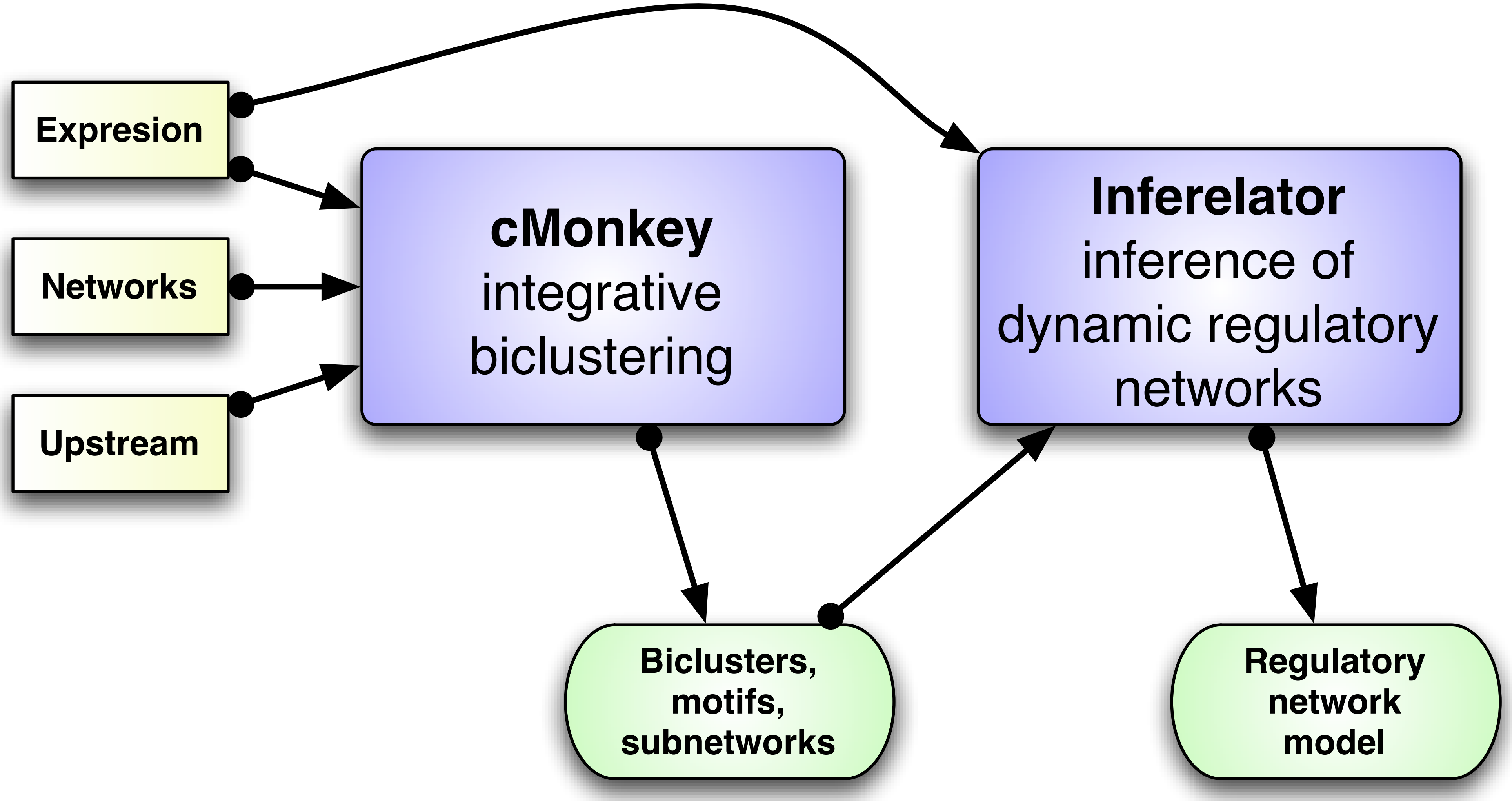


M. Facciotti, N. Baliga

min pan, Kenia Whitehead, Amy Schmid

OVERVIEW:





Expresion

Networks

Upstream

cMonkey
integrative biclustering

Inferelator
inference of dynamic regulatory networks

Biclusters, motifs, subnetworks

Regulatory network model

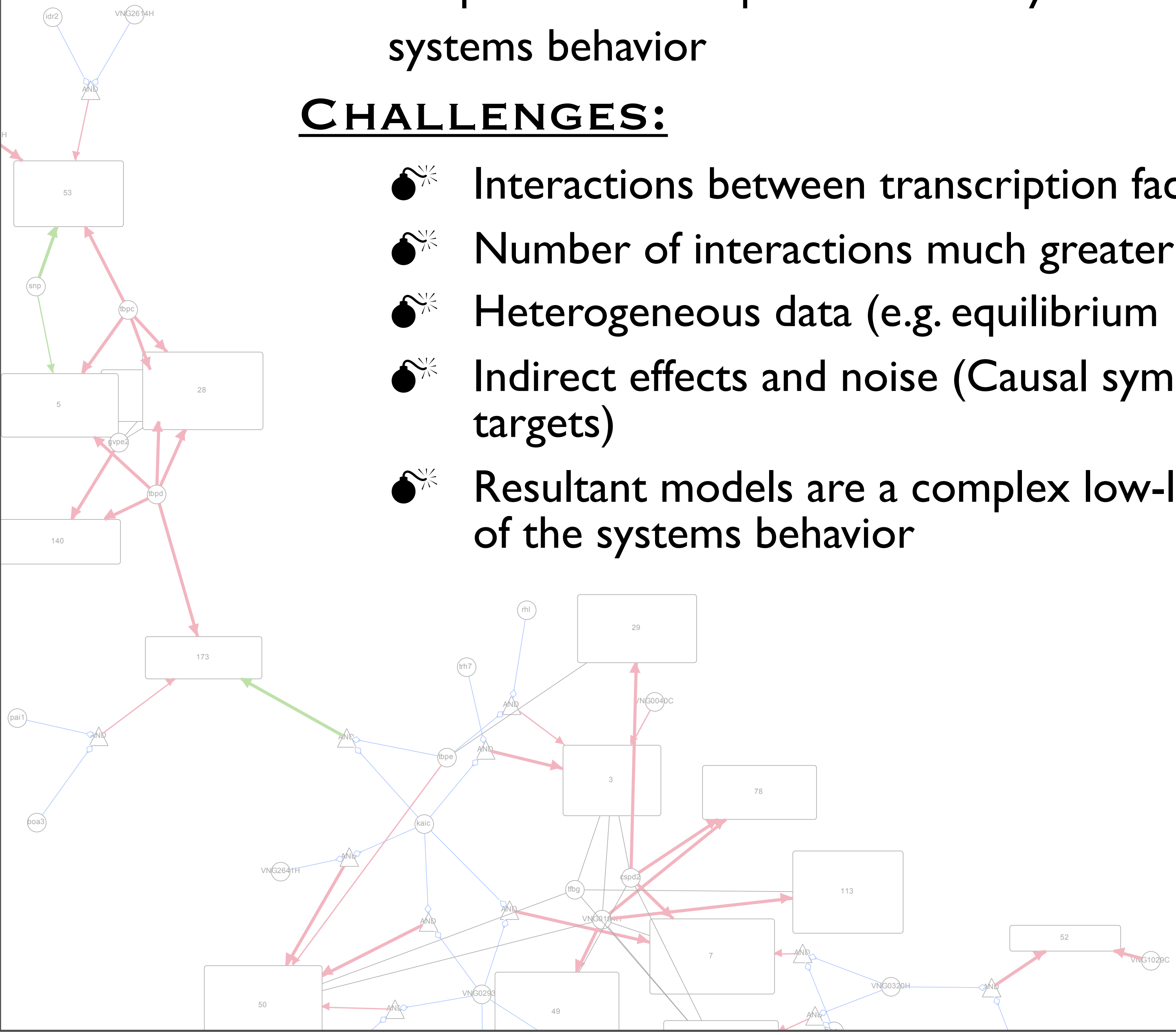
II. The Inferelator: regulatory network inference

Biological motivation:

Learn regulatory interactions from data that are predictive of equilibrium and dynamical systems behavior

Challenges:

- Interactions between transcription factors
- Number of interactions much greater than number of observations
- Heterogeneous data (e.g. equilibrium and kinetic measurements)
- Indirect effects and noise (Causal symmetry between activators and targets)
- Resultant models are a complex low-level abstraction of the systems behavior



EXPERIMENTAL DESIGN

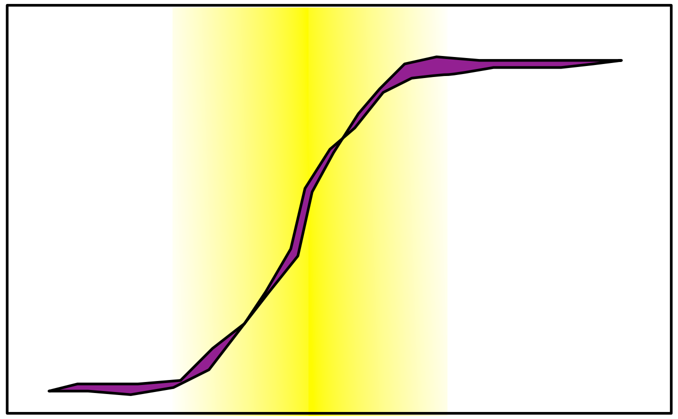
1. KNOW YOUR MODEL/ FRAMEWORK:

$$\tau \frac{dy}{dt} = -y + g(\beta \bullet Z) \qquad (1)$$

$$\beta \mathbf{Z} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 \min(x_1, x_2)$$

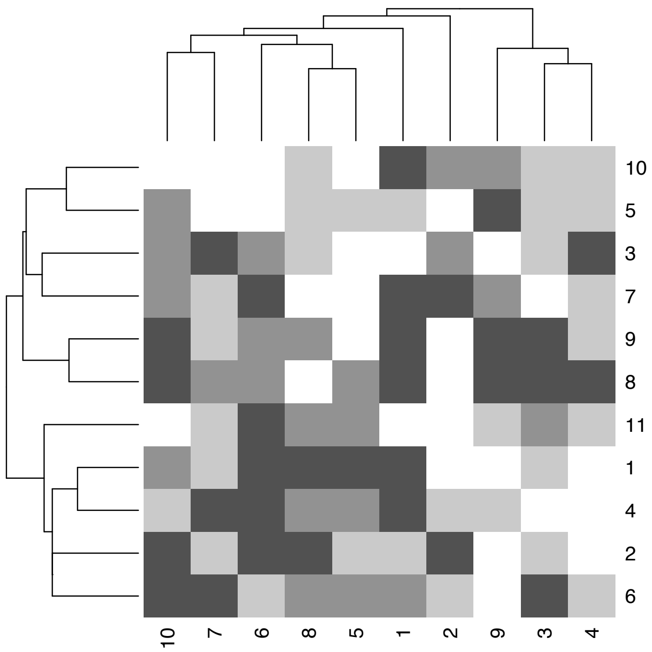
2. TIME SERIES: SAMPLING WITH CORRECT RATE(S)!

3. SAMPLING IN CORRECT REGIME.

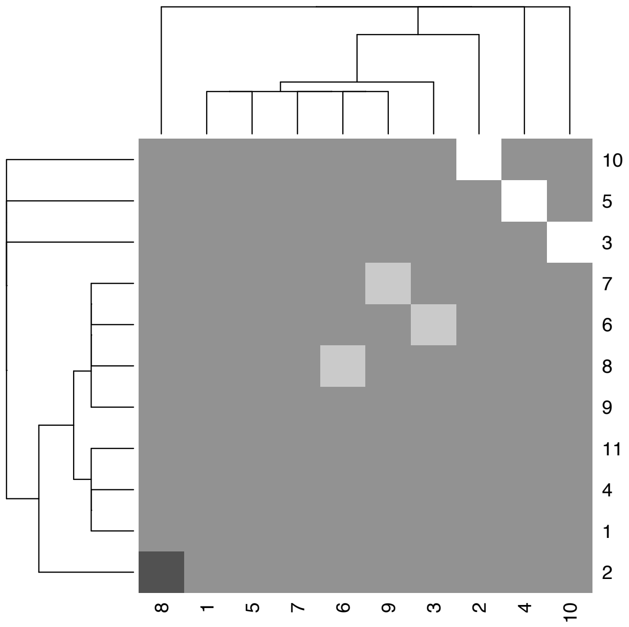


4. MULTIFACTORIAL ON A BUDGET:

OPTIMAL (?) :-)



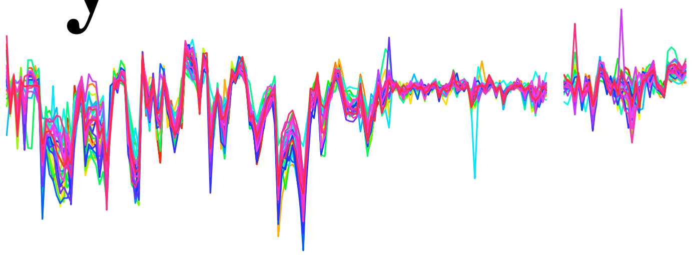
THE USUAL :-(

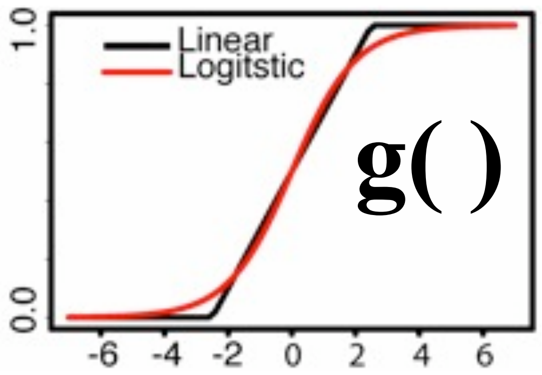


Inferelator v 1.0

$$\tau \frac{\partial y}{\partial t} = -y + g(\beta Z)$$

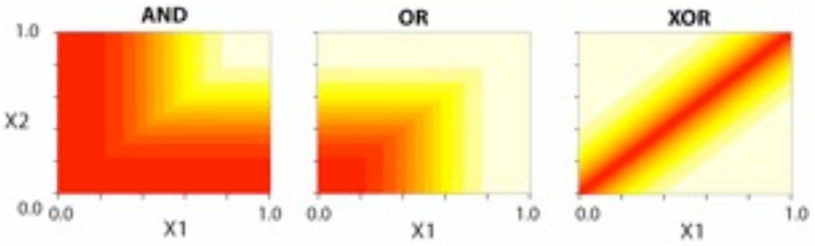
y

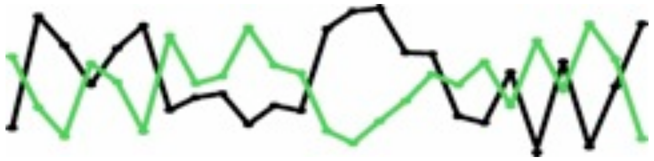




z = x1

z = min(x1,x2)





STEADY-STATE

$$y = g(\beta Z_{eq})$$

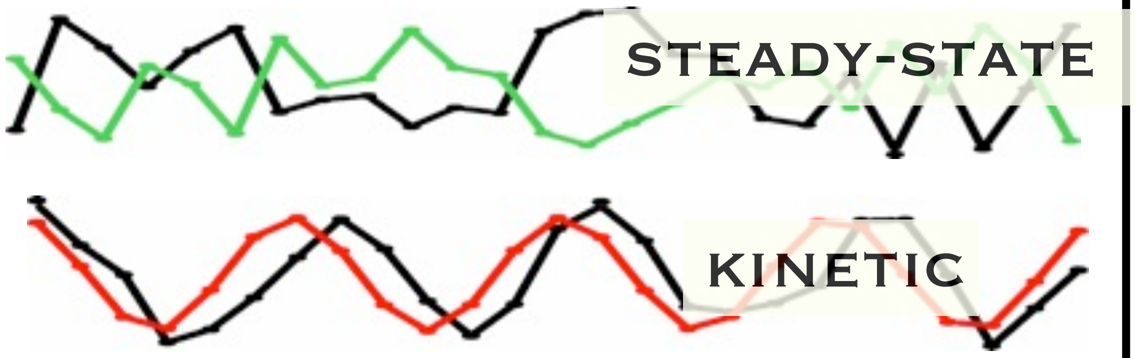
KINETIC

$$\frac{\tau}{\Delta t}(y_{t+1} - y_t) + y_t = g(\sum_{j=1}^{p} \beta_i z_{tj})$$

Bonneau, Reiss, Baliga, Thorsson, 2006

CORE ASSUMPTION

$$\tau \frac{\partial y}{\partial t} = -y + g(\beta Z)$$



STEADY STATE

$$y = g(\beta Z_{eq})$$

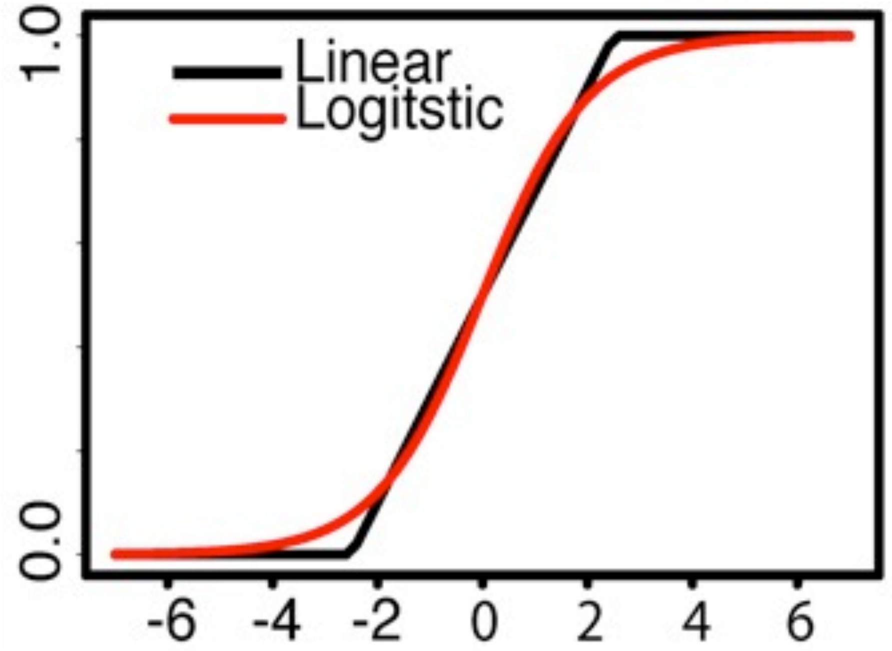
TIME SERIES/ KINETIC

$$\tau \frac{(y_{t+1} - y_t)}{\Delta t} = -y_t + g\left(\sum_{j=1}^{p} \beta_i z_{tj}\right)$$

$$\frac{\tau}{\Delta t}(y_{t+1} - y_t) + y_t = g\left(\sum_{j=1}^{p} \beta_i z_{tj}\right)$$

Bonneau, Baliga, Thorsson, 2006

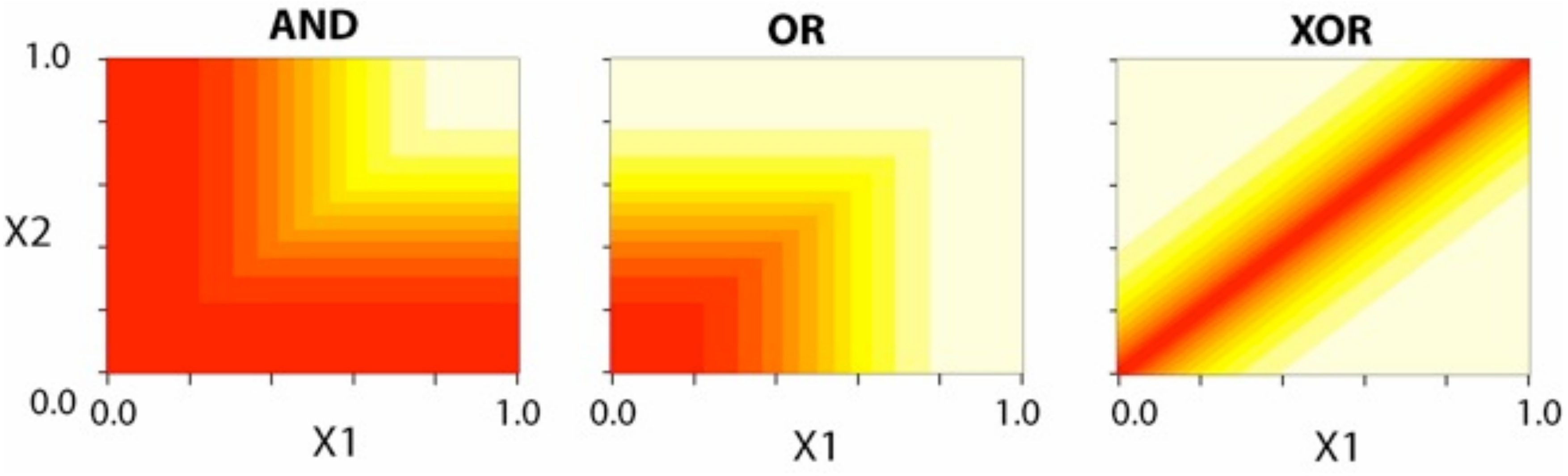
2 Squashing functions: promoter saturation



$$g(\beta Z) = \frac{1}{1+e^{\beta Z}}$$

$$g(\beta Z) = \begin{cases} \beta Z: \text{ if } \min(y) < \beta Z < \max(y) \\ \max(y): \quad \text{if } \beta Z > \max(y) \\ \min(y): \quad \text{if } \beta Z < \min(y) \end{cases}$$

Representing Interactions:



$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 \min(x_{1j}, x_{2j}) + 1$$

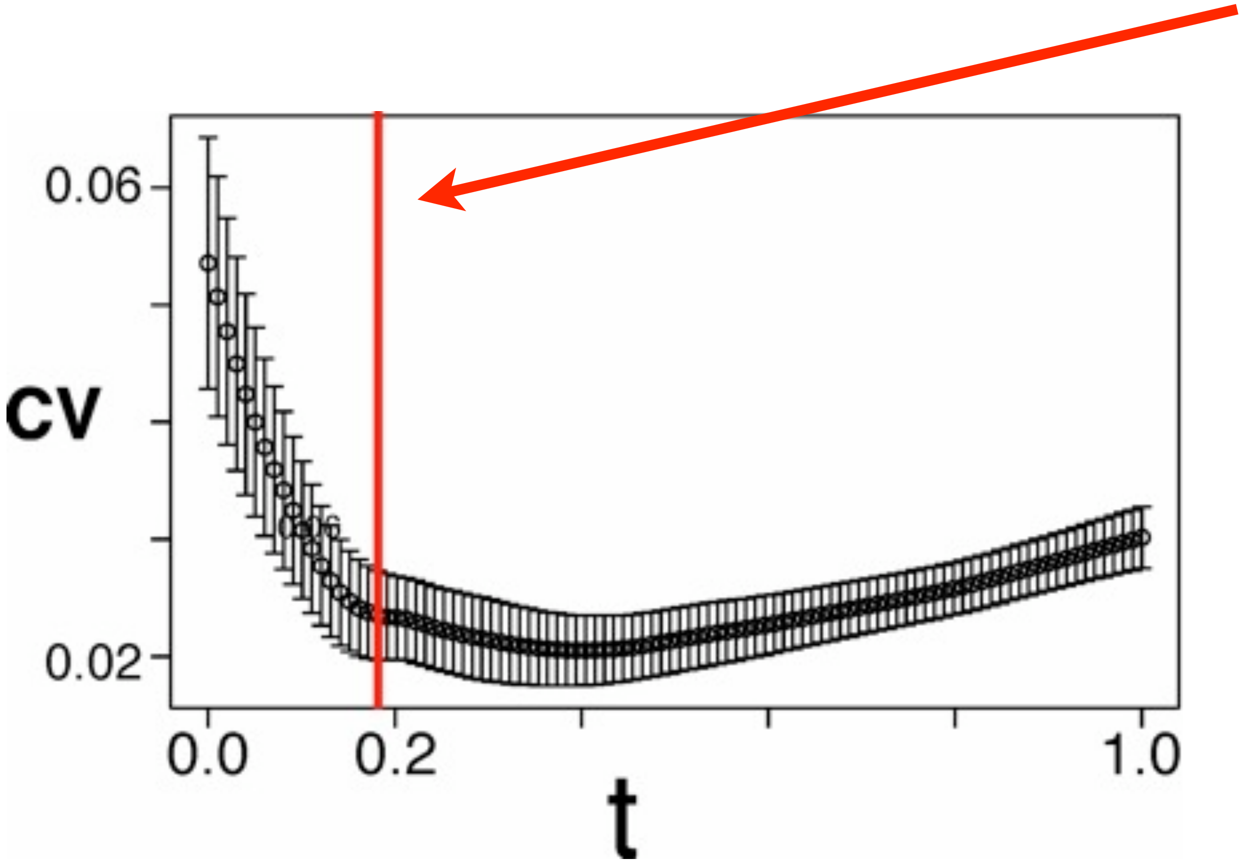
$$(\mathbb{R}, \odot, \oplus) \quad x \oplus y \coloneqq \min\{x, y\} \quad x \odot y \coloneqq x + y$$

$$y_j = \beta_1 x_1 \odot \beta_2 x_2 \odot \left(\beta_1 x_1 \oplus \beta_2 x_2\right)$$

Model selection using L1-shrinkage:
avoiding overfitting

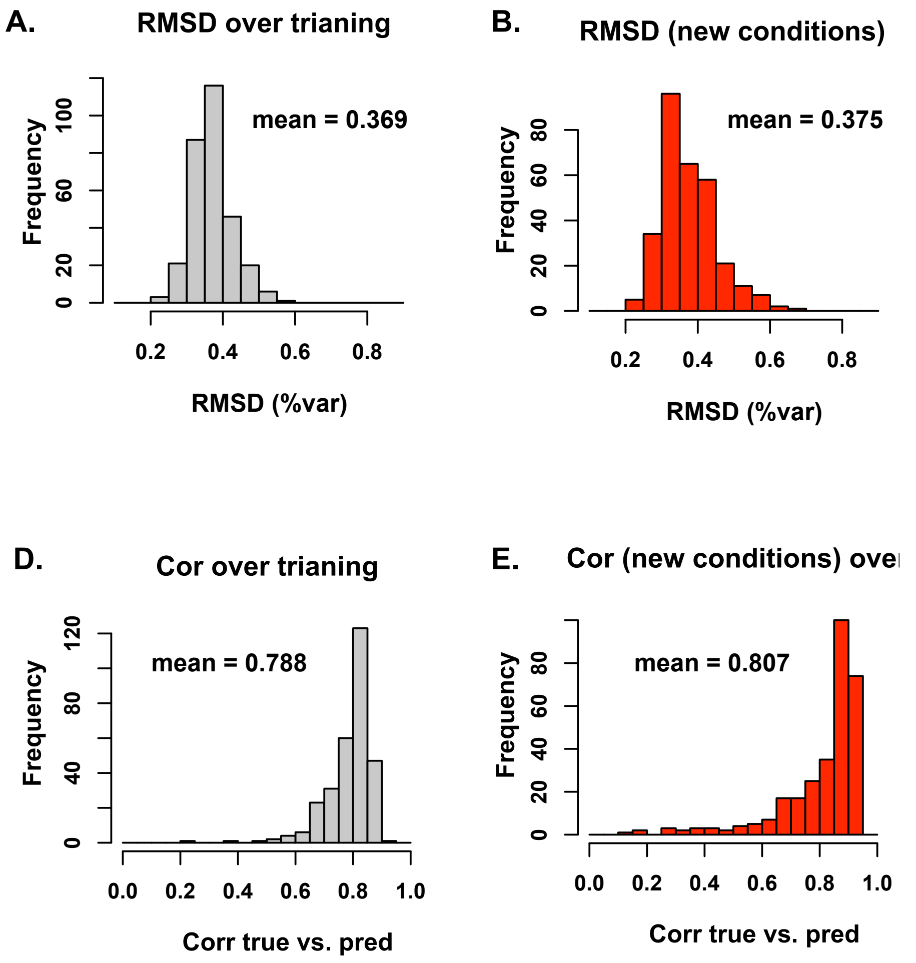
$$\sum_{j=1}^{p}\left|\hat{\beta}_j\right| \le t\sum_{j=1}^{p}\left|\beta_{ols_j}\right| \qquad (\hat{\alpha},\hat{\beta}) = \underset{(\hat{\alpha},\hat{\beta})}{\arg\min}\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_{j=1}^{p}\beta_j z_{ij}\right)^2\right\}$$

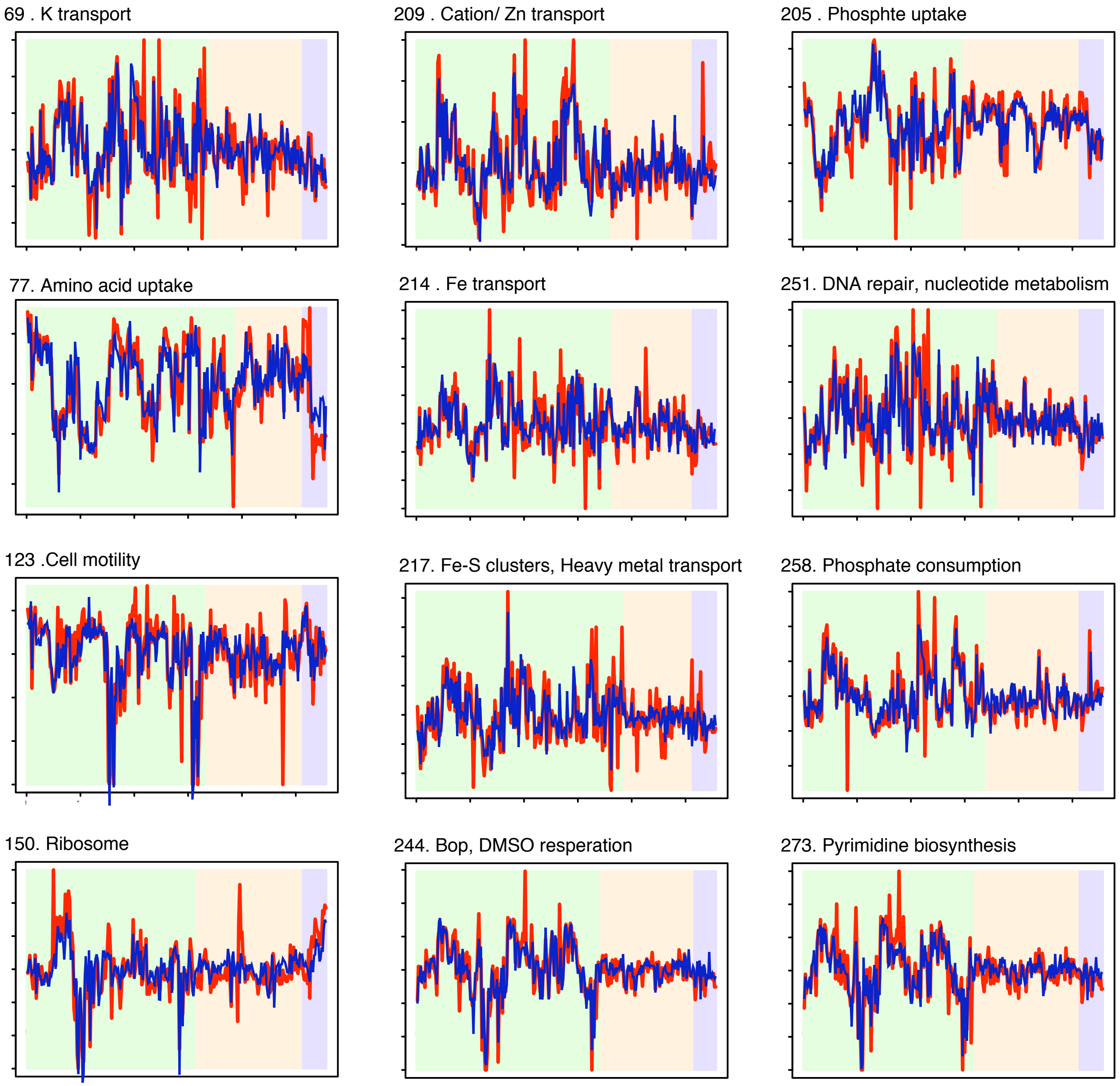
model size



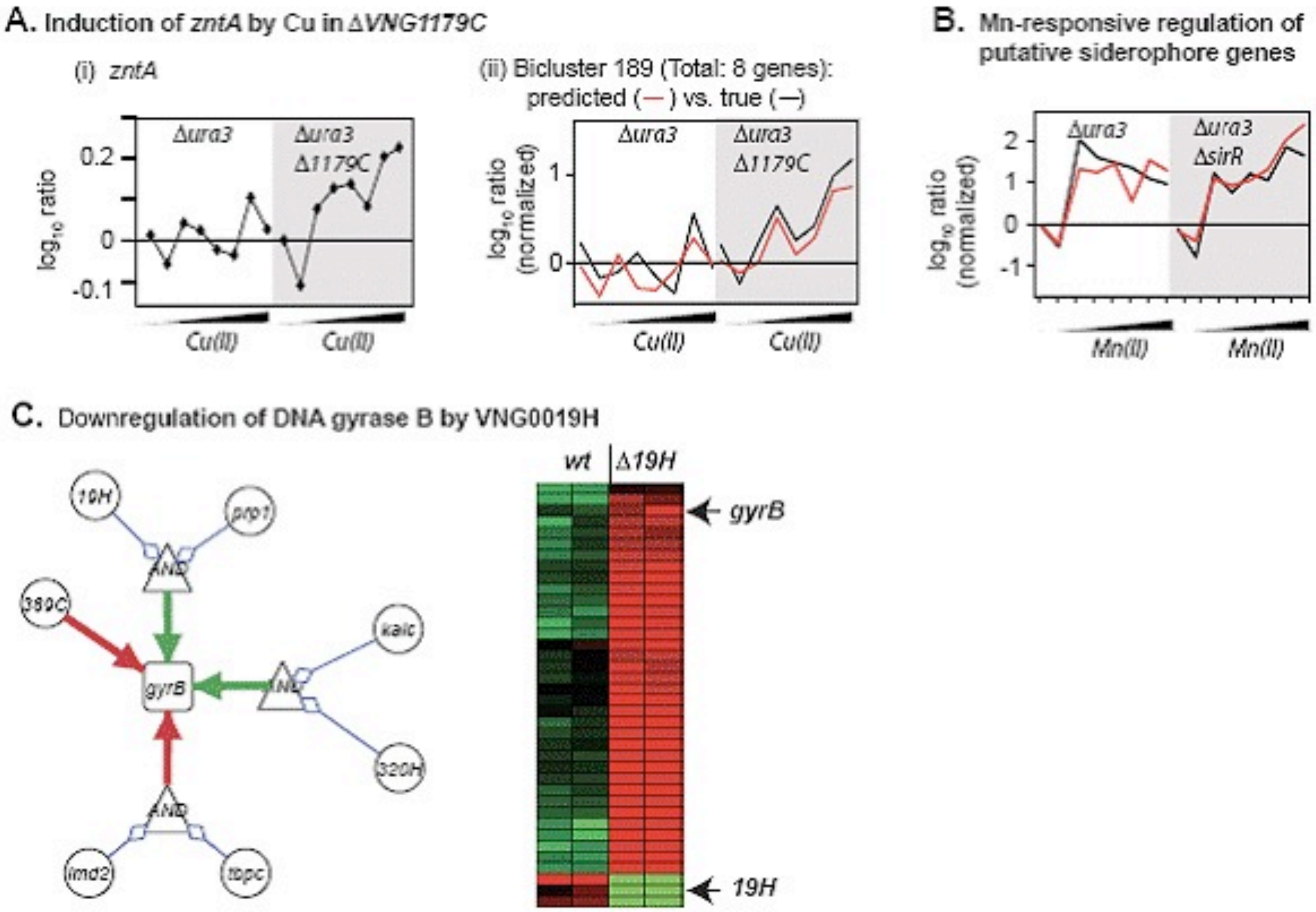
<u>why L1?</u>
beta -> 0
LARS

Predictive power over 130 new experiments





Prediction of outcome following genetic and Environmental perturbations



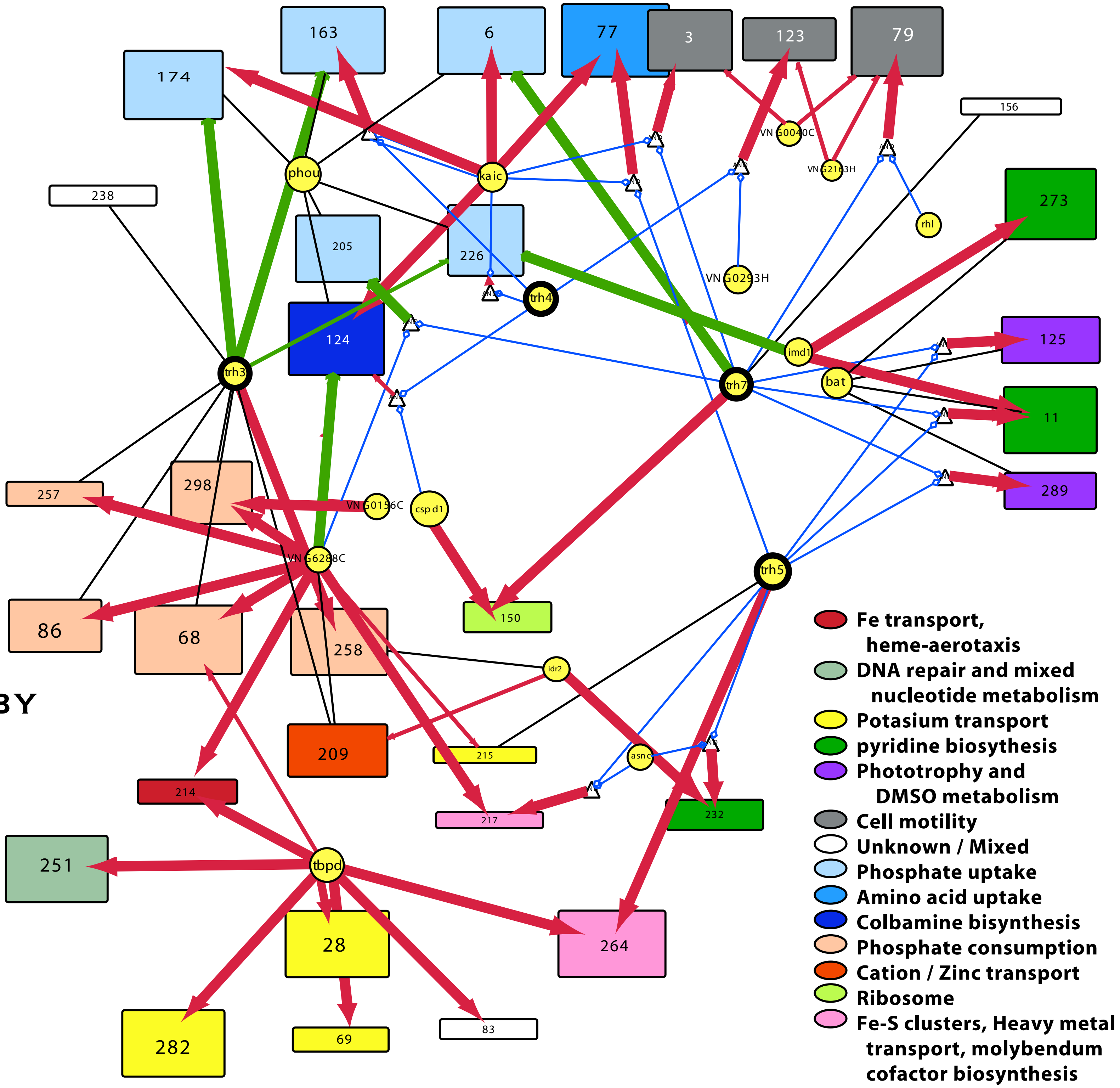
A. Induction of *zntA* by Cu in Δ*VNG1179C*

B. Mn-responsive regulation of putative siderophore genes

C. Downregulation of DNA gyrase B by VNG0019H

INFERRED TRH CONTROLLED SUBNETWORK:

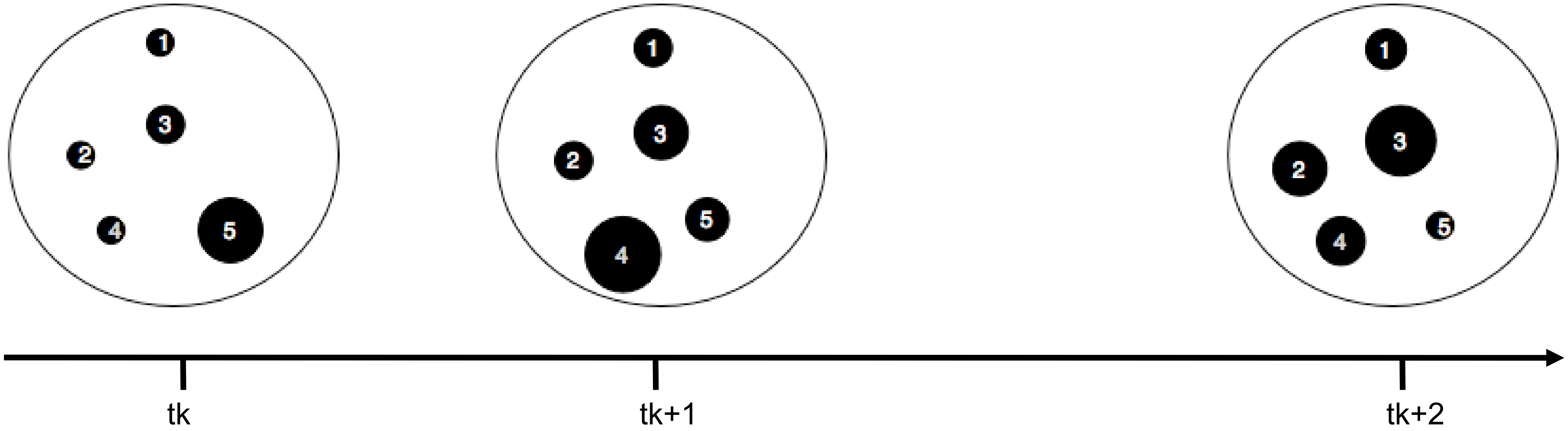
Homeostasis is an emergent property of the global network



Edges validated since by ChIP-chip
M. Facciotti, N. Baliga

Bonneau, et al, Genome Biology, 2006, Bonneau, et al. Cell, 2007

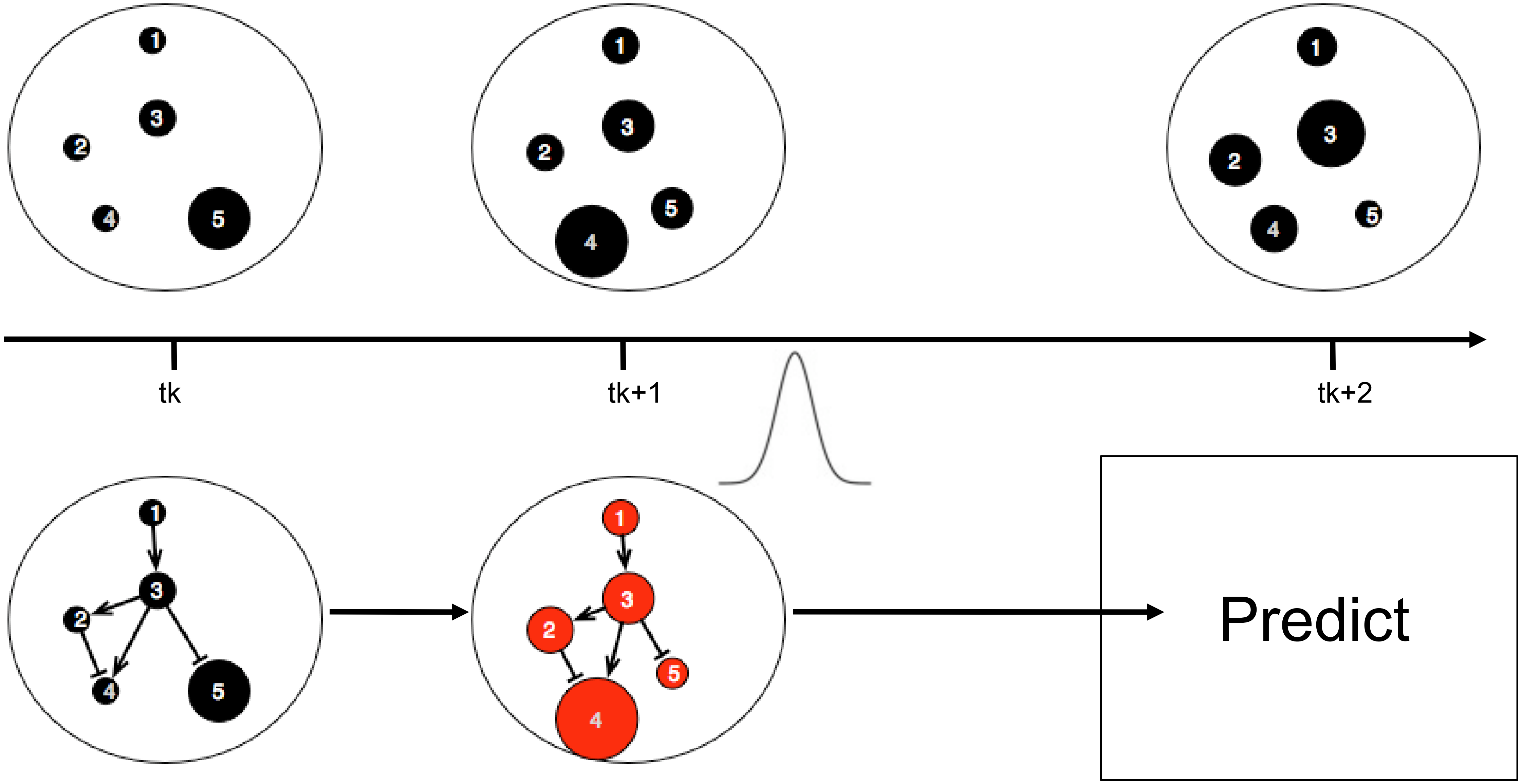
Inferelator 1--- Limitations



Error propagation

Error over long time intervals

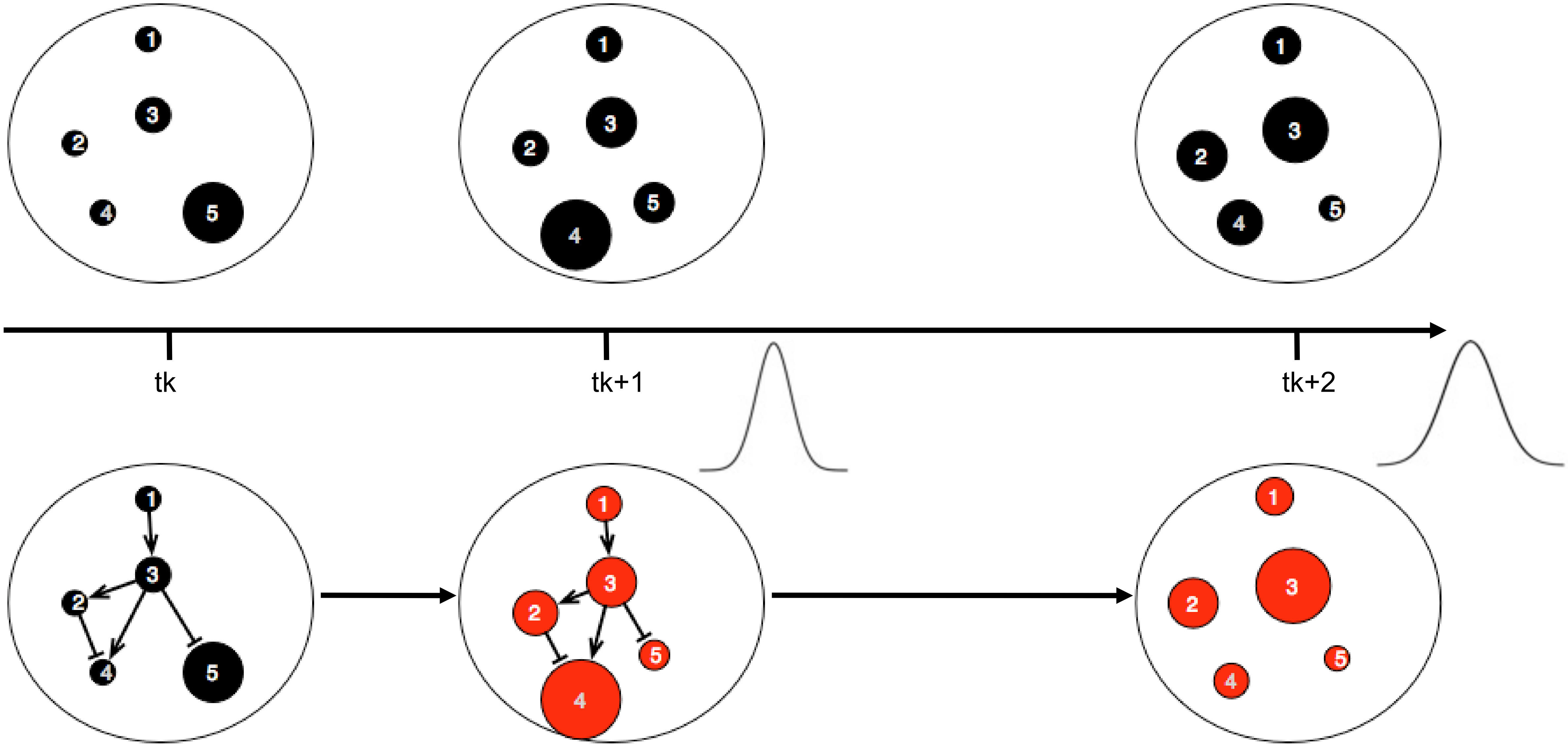
Inferelator 1--- Limitations



Predict

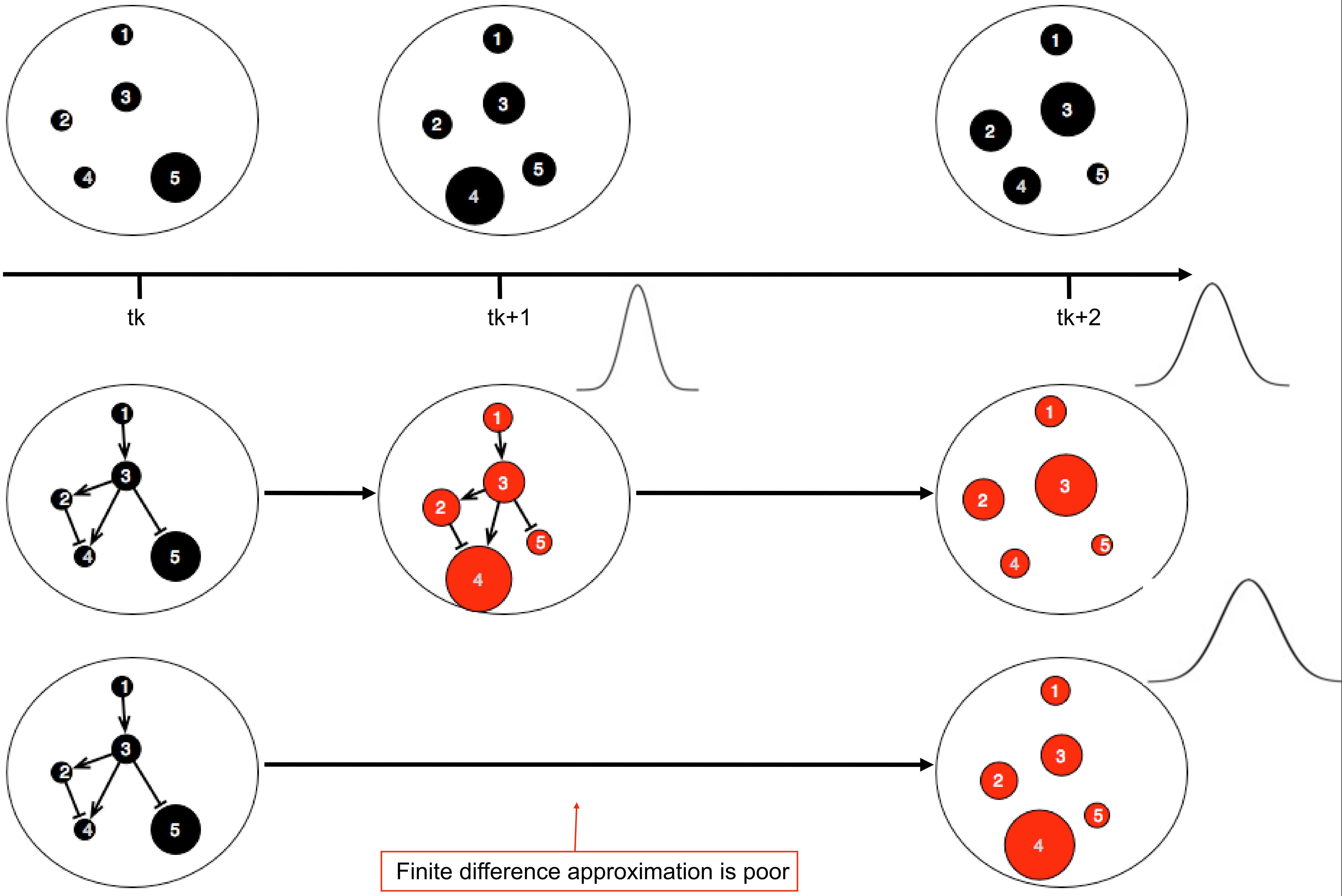
Error over long time intervals

Inferelator 1--- Limitations



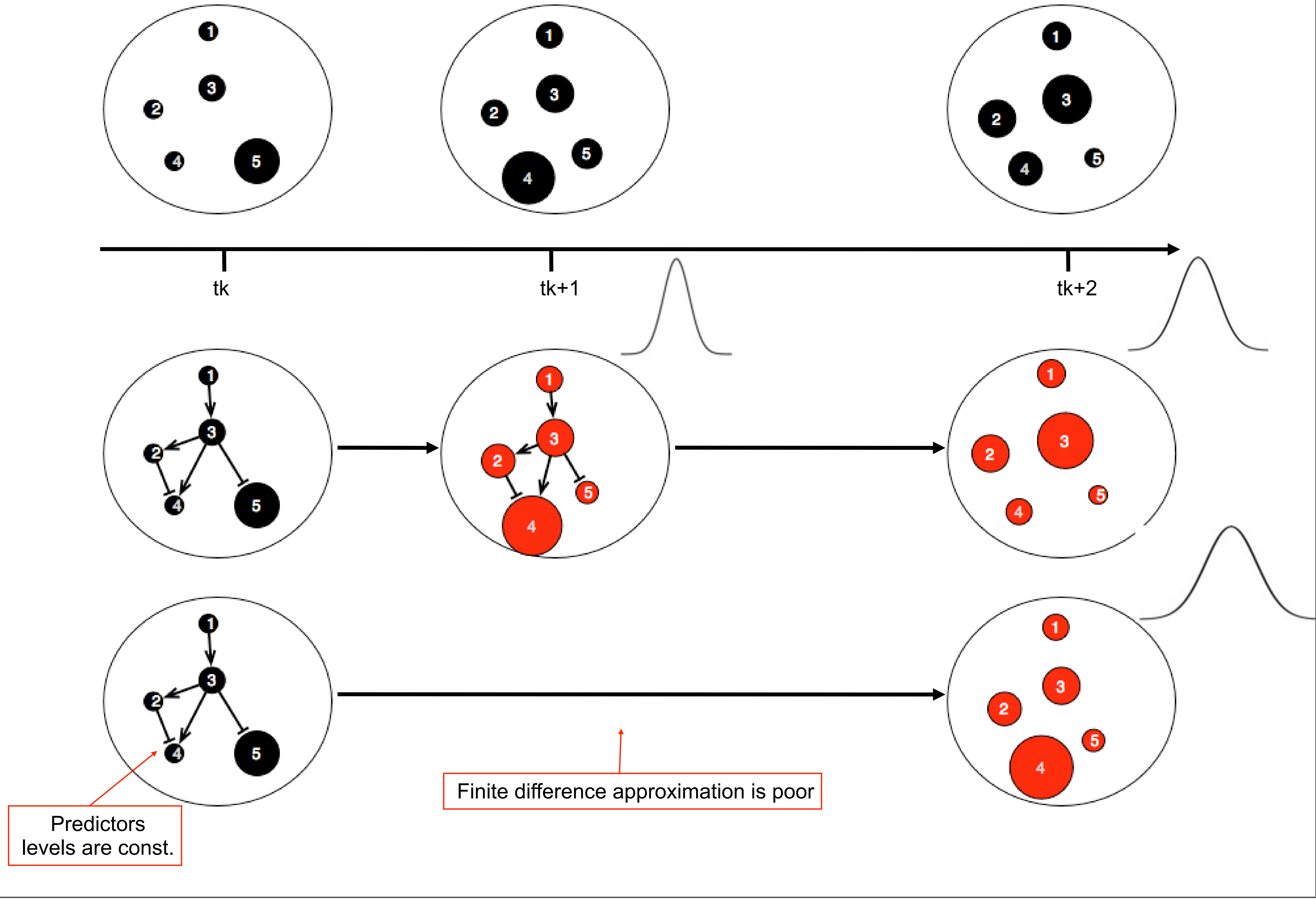
Error over long time intervals

Inferelator 1--- Limitations



Finite difference approximation is poor

Inferelator 1--- Limitations



Predictors
levels are const.

Finite difference approximation is poor

Explicit global solutions using Metropolis-Hastings:

(1) $$E(\beta)=\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{M}\left(x_i(t_k)-x_i^{obs}(t_k)\right)^{2}$$

Calculate gradient of the Energy with respect to

(2) $$\frac{\partial E(\beta)}{\partial \beta_j}=\sum_{k=1}^{K}\sum_{i=1}^{M}\left(x_i(t_k)-x_i^{obs}(t_k)\right)\frac{\partial x_i(t_k)}{\partial \beta_j}$$

Slope

Step

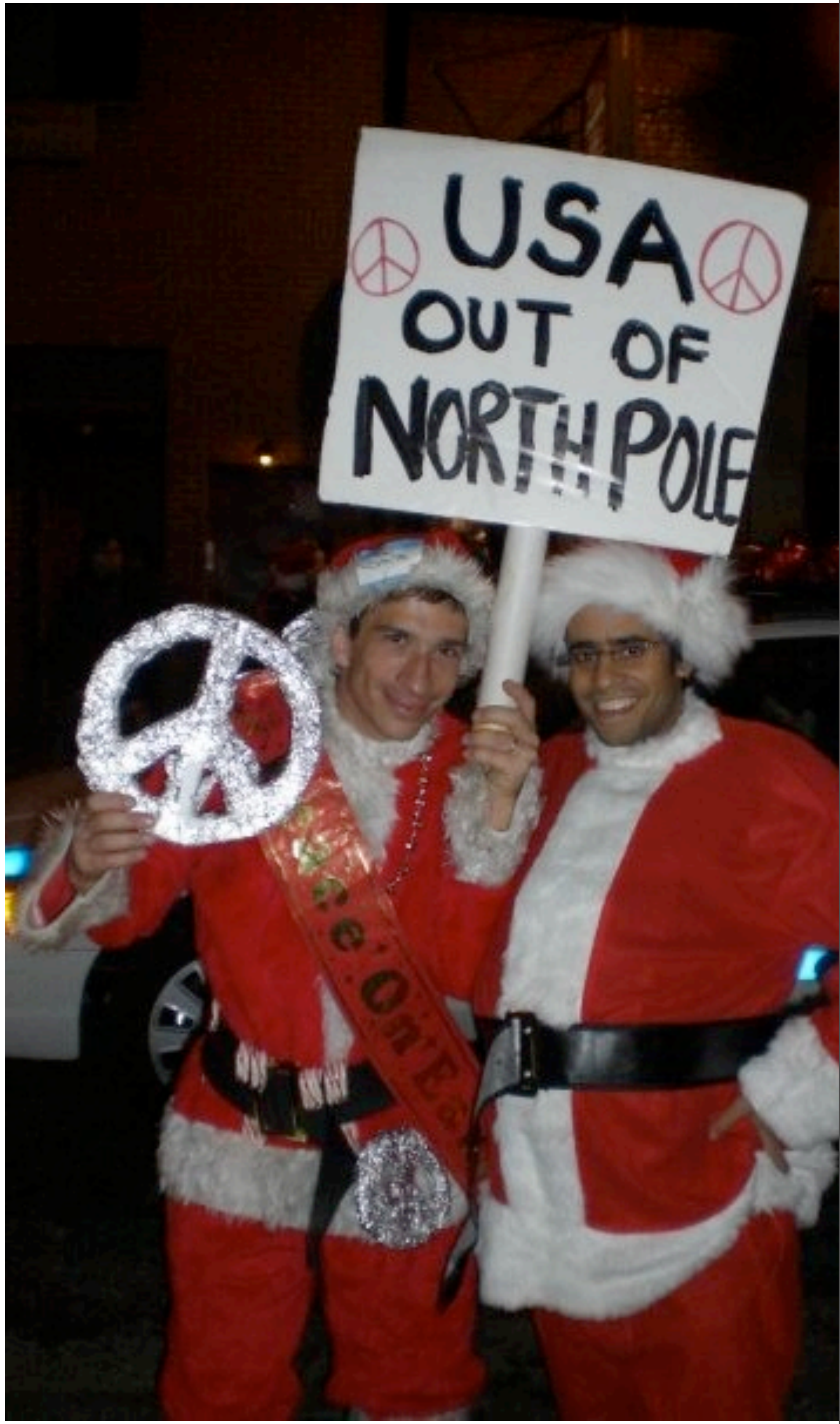
Slope

~N(0,1)

(3) $$\beta_j^{n+1}=\beta_j^{n}-h\frac{\partial E(\beta^{n})}{\partial \beta_j}+\sqrt{2\sigma h}*\zeta_j^{n}$$

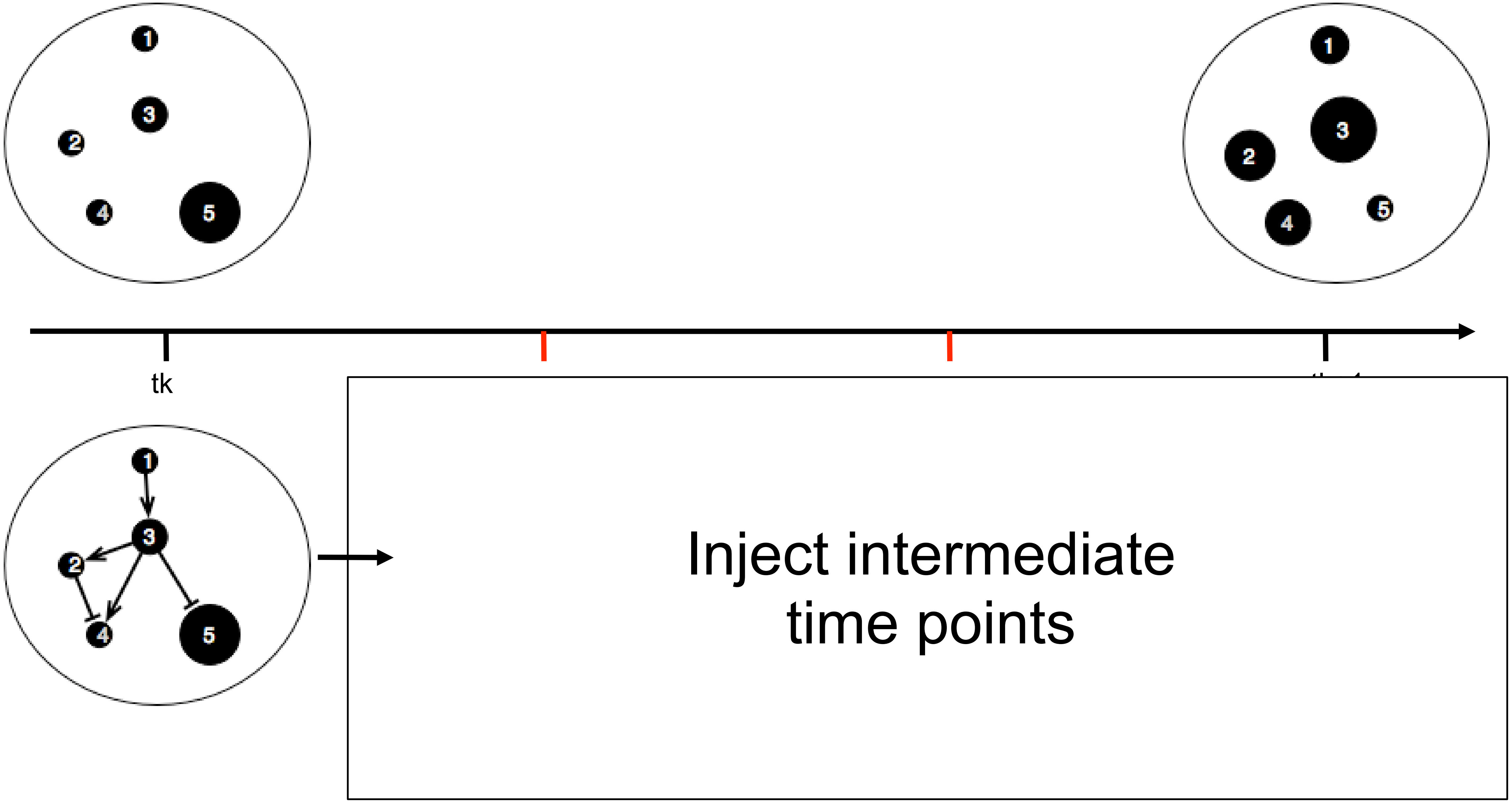
Temperature

Aviv Madar on Right



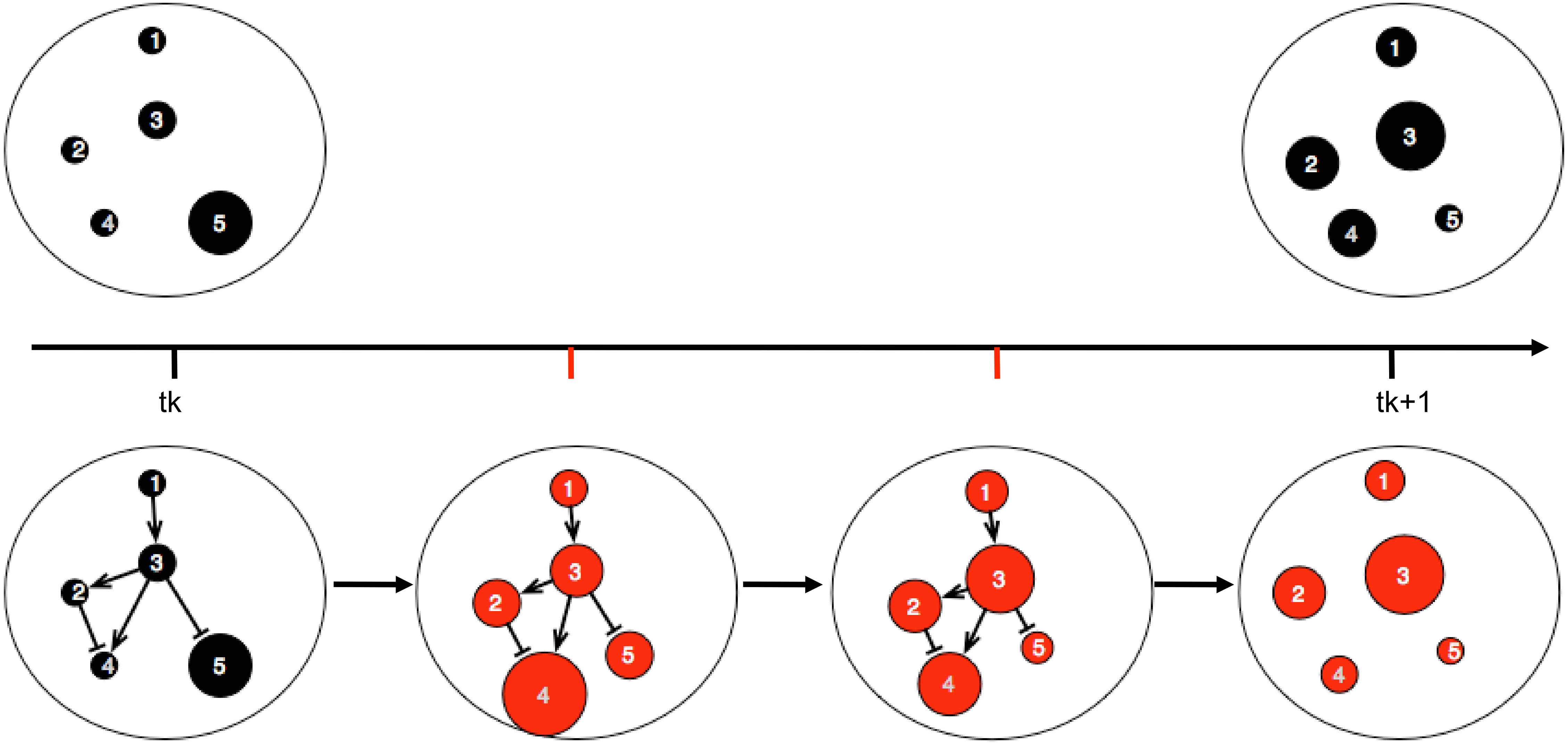
Work done with Eric Vanden-Eijnden, Courant Institute of Mathematical Sciences, NYU

Inferelator 2: Concepts

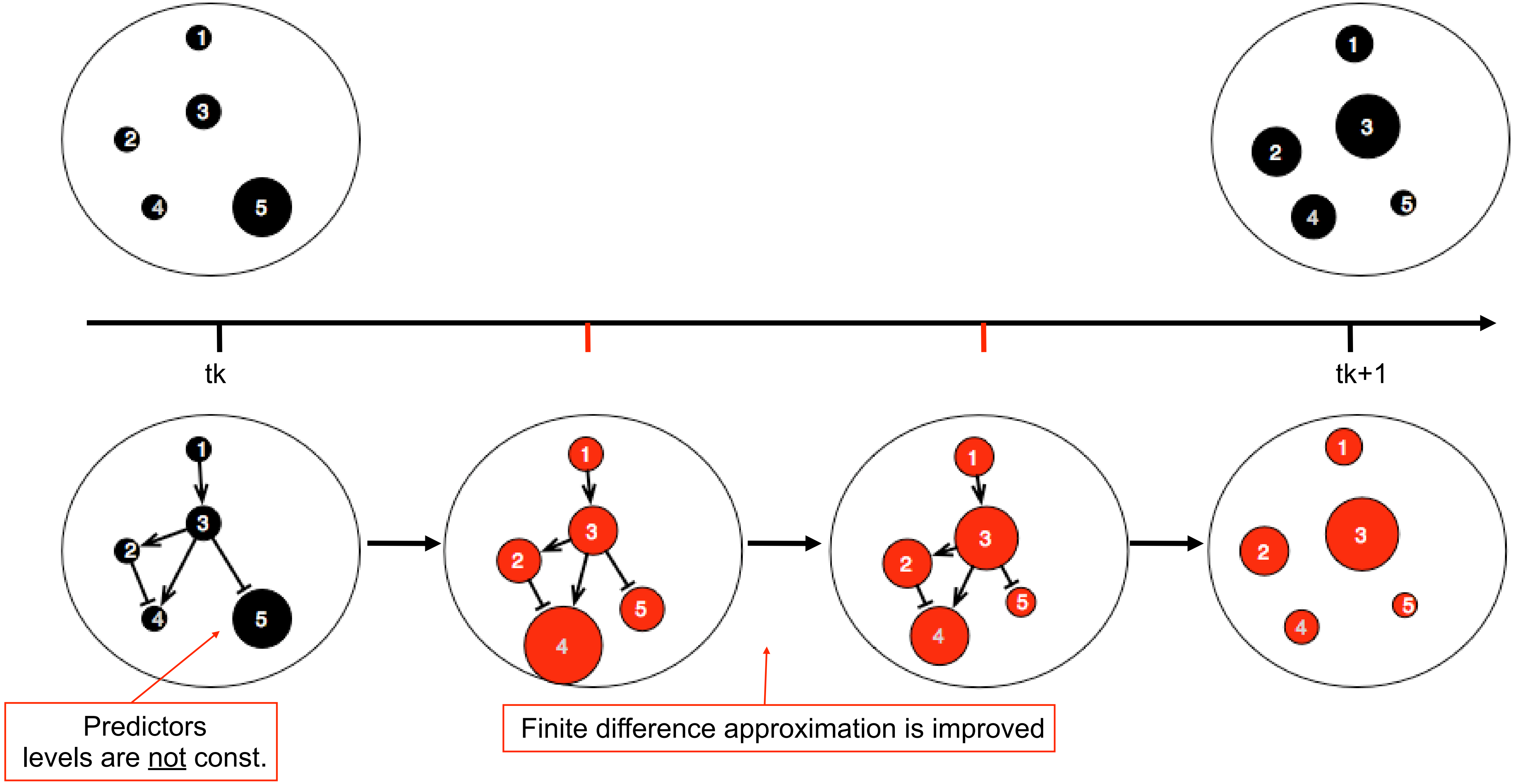


Inject intermediate time points

Inferelator 2: Concepts



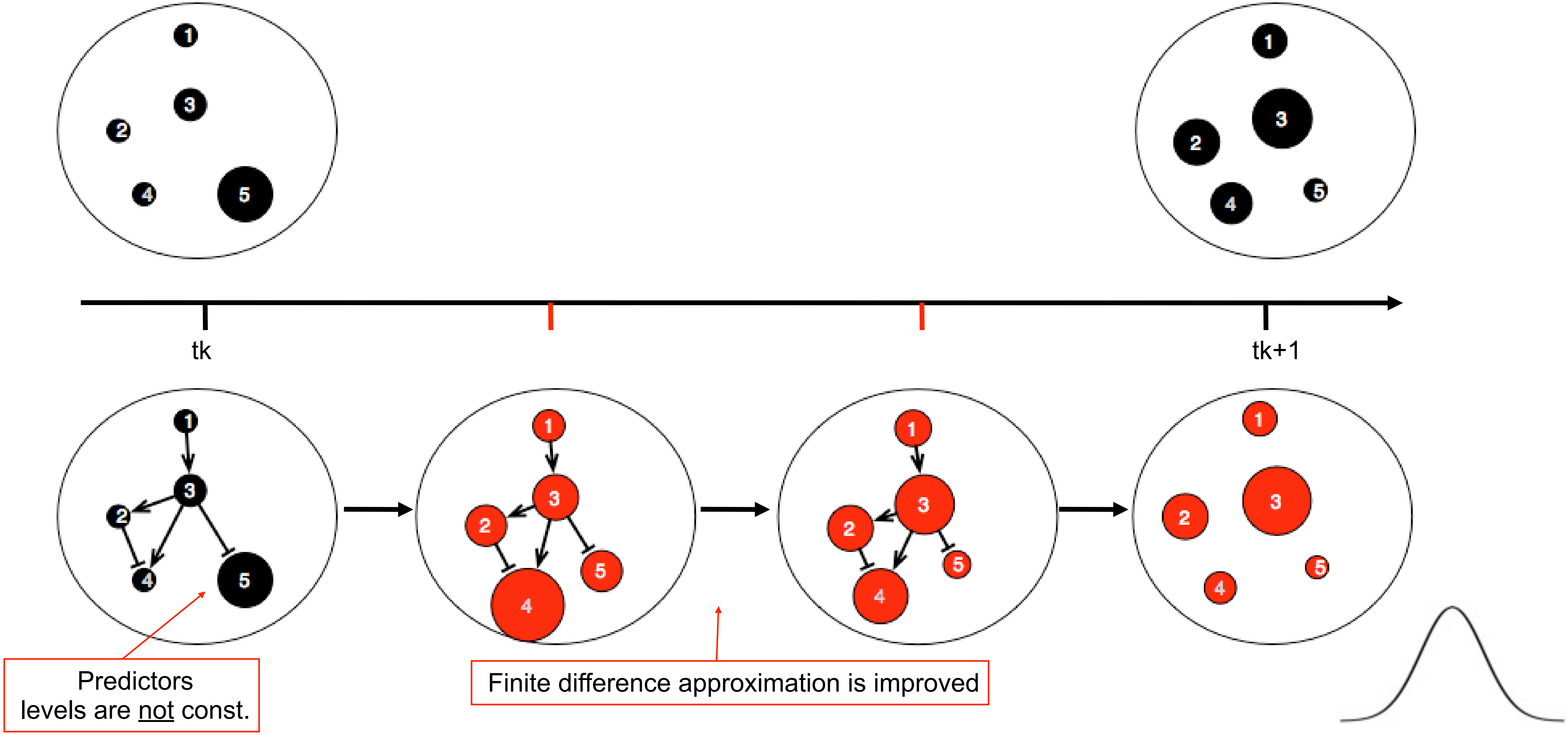
Inferelator 2: Concepts



Predictors levels are not const.

Finite difference approximation is improved

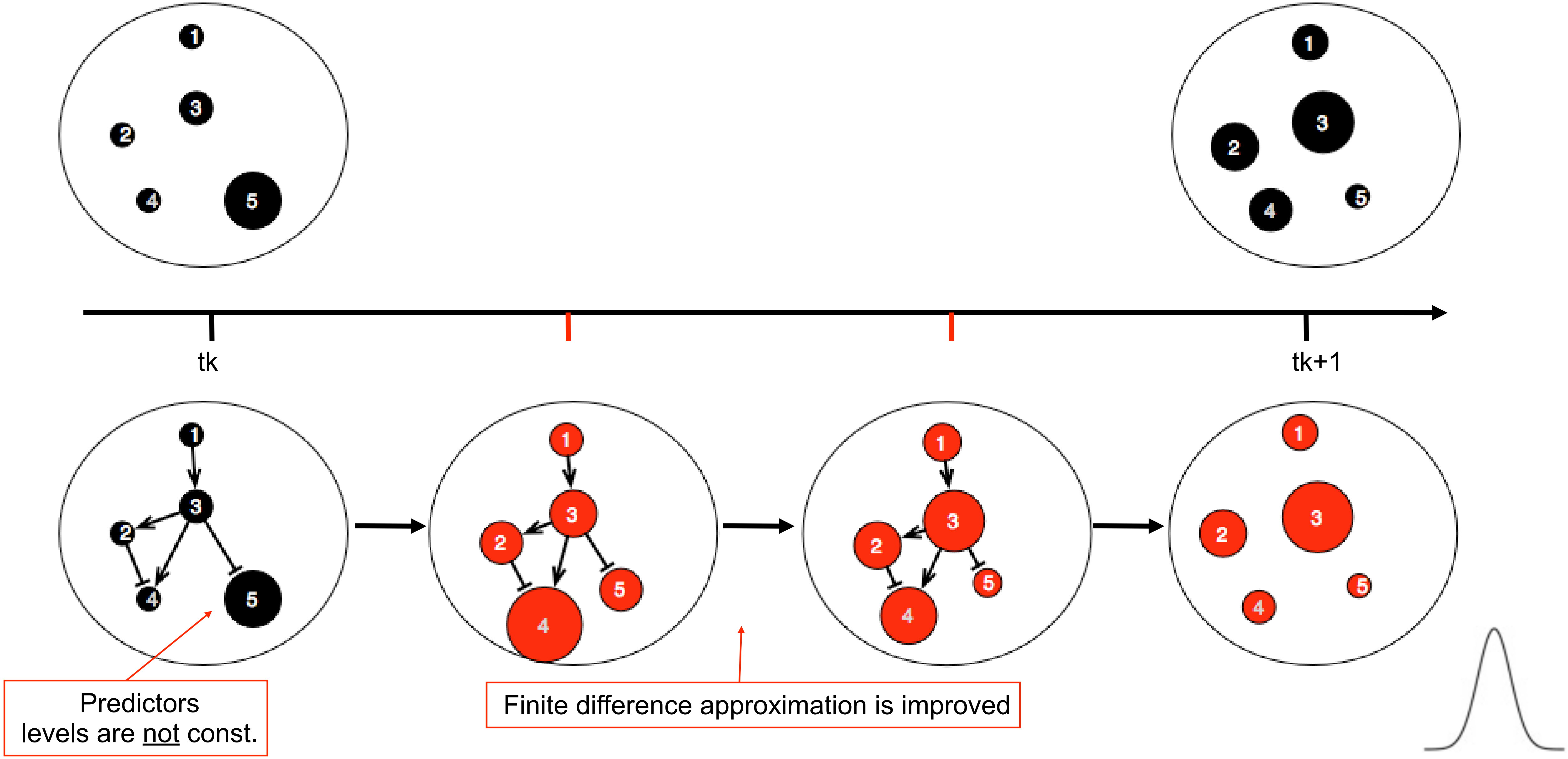
Inferelator 2: Concepts



Predictors levels are not const.

Finite difference approximation is improved

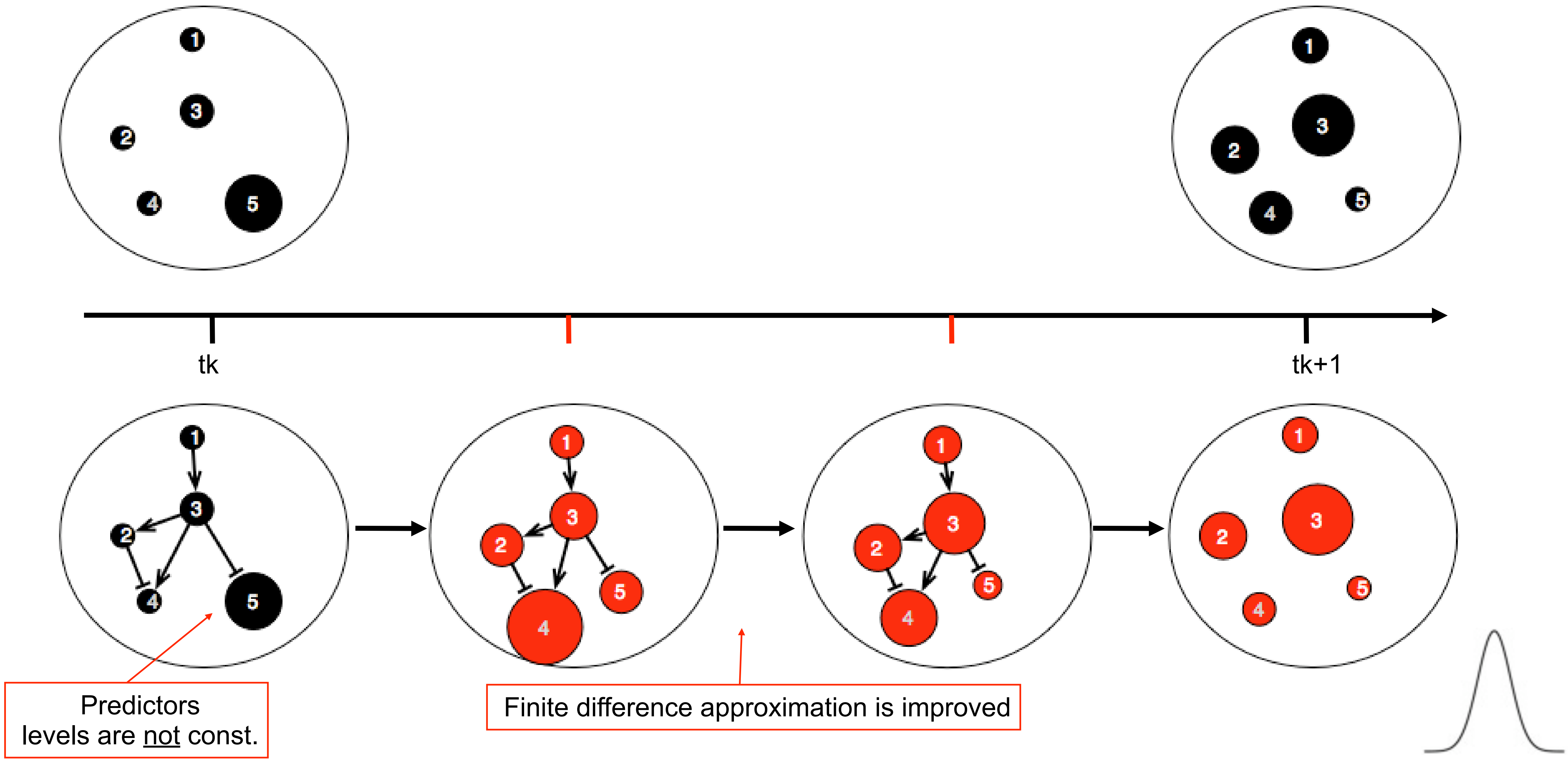
Inferelator 2: Concepts



Predictors
levels are <u>not</u> const.

Finite difference approximation is improved

Inferelator 2: Concepts



Predictors
levels are not const.

Finite difference approximation is improved

How do we estimate parameters?

Inferelator 2: Mathematical Overview

Minimize Energy (scoring/objective function)

Markov Chain Monte Carlo (MCMC) scheme
to sample parameters

Inferelator 2: Mathematical Overview

$$E(\boldsymbol{\beta}) = c_1 \sum_{k=1}^{K_{ts}} \sum_{i=1}^{N} |x_i^{\text{obs}}(t_k) - x_i^{\text{pred}}(t_k, \boldsymbol{\beta})|^2$$

Error over time series data

$$+ c_2 \sum_{k=1}^{K_{ss}} \sum_{i=1}^{N} | \sum_{j=1}^{P} \beta_{i,j} x_i^{\text{obs}}(t_k)|^2$$

Error over steady state data

$$+ c_3 \sum_{i=1}^{N} \sum_{j=1}^{P} |\beta_{i,j}|^2$$

L2 norm constraint/regularizer

Markov Chain Monte Carlo (MCMC) scheme
to sample parameters

Inferelator 2: Mathematical Overview

$$E(\boldsymbol{\beta}) = c_1 \sum_{k=1}^{K_{ts}} \sum_{i=1}^{N} |x_i^{\text{obs}}(t_k) - x_i^{\text{pred}}(t_k, \boldsymbol{\beta})|^2$$

Error over time series data

$$+ c_2 \sum_{k=1}^{K_{ss}} \sum_{i=1}^{N} |\sum_{j=1}^{P} \beta_{i,j} x_i^{\text{obs}}(t_k)|^2$$

Error over steady state data

$$+ c_3 \sum_{i=1}^{N} \sum_{j=1}^{P} |\beta_{i,j}|^2$$

L2 norm constraint/regularizer

$$\beta_{i,j}^{n+1} = \beta_{i,j}^{n} + \sqrt{\sigma h}\, \xi_{i,j}^{n} - h \frac{\partial E(\boldsymbol{\beta}^n)}{\partial \beta_{i,j}}$$

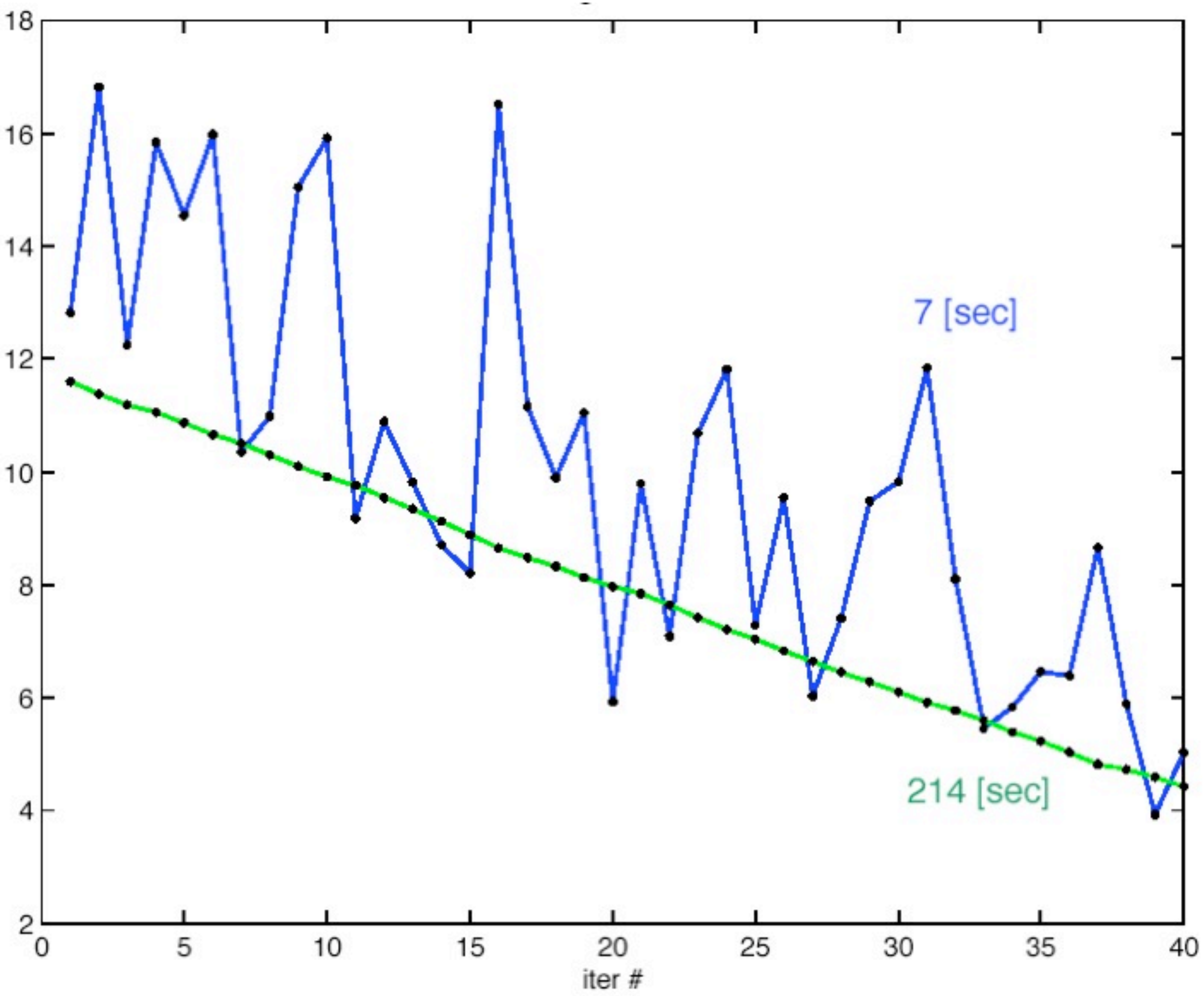
Markov chain

Gaussian Noise term

Importance sampling

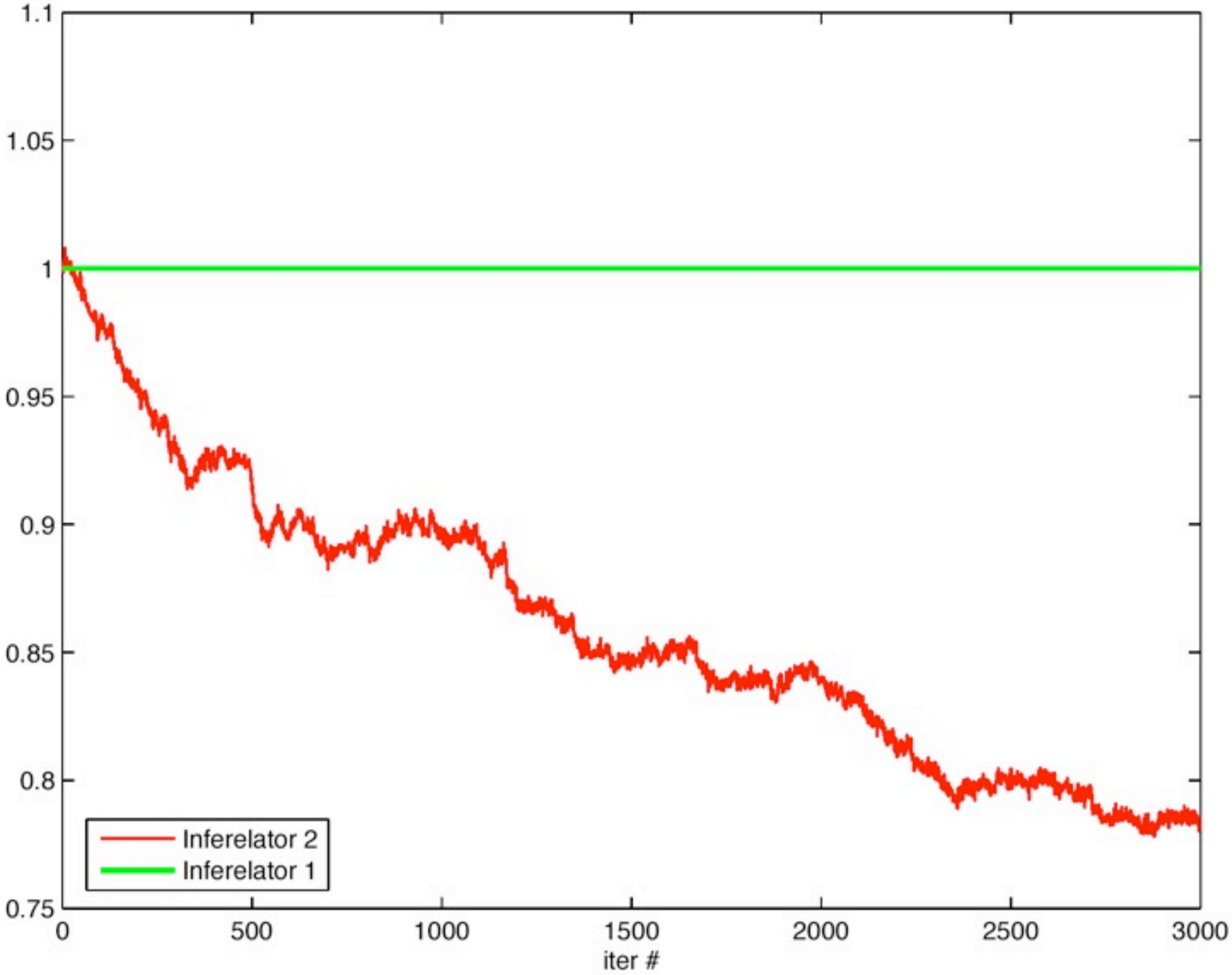
Inferelator 2: Gradient Approximation

$$\sum_{i,j} \left| \frac{\partial E(\boldsymbol{\beta}^n)}{\partial \beta_{i,j}} \right|$$



Inferelator 2: L1-norm of Parameters

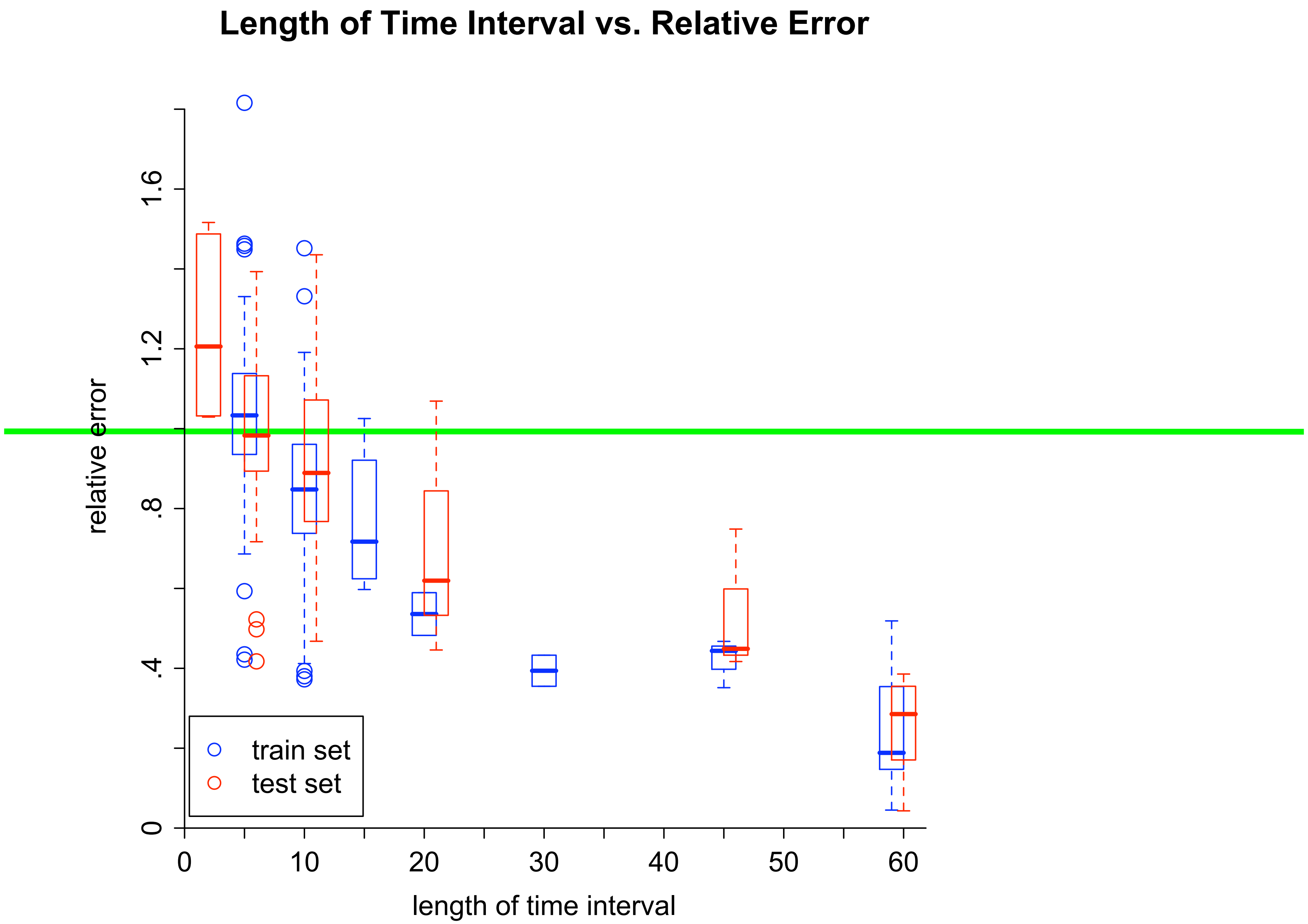
$$\frac{L1^{\text{Inf-2}}}{L1^{\text{Inf-1}}}$$



$$\frac{L1\_norm^{\text{Inf-2}}}{L1\_norm^{\text{Inf-1}}}$$

Inferelator 2: Performance 5

$$\frac{E(\boldsymbol{\beta}, t_k)^{\text{Inf-2}}}{E(\boldsymbol{\beta}, t_k)^{\text{Inf-1}}}$$



time interval, minutes ->

II. The Inferelator: Future Directions

mixed time scales / mixed data-types:

Learn regulatory interactions from sub-optimal datasets

Mixed signaling & regulatory nets

Adding metabolic effects

Inferelator 2, more explicit dynamics:

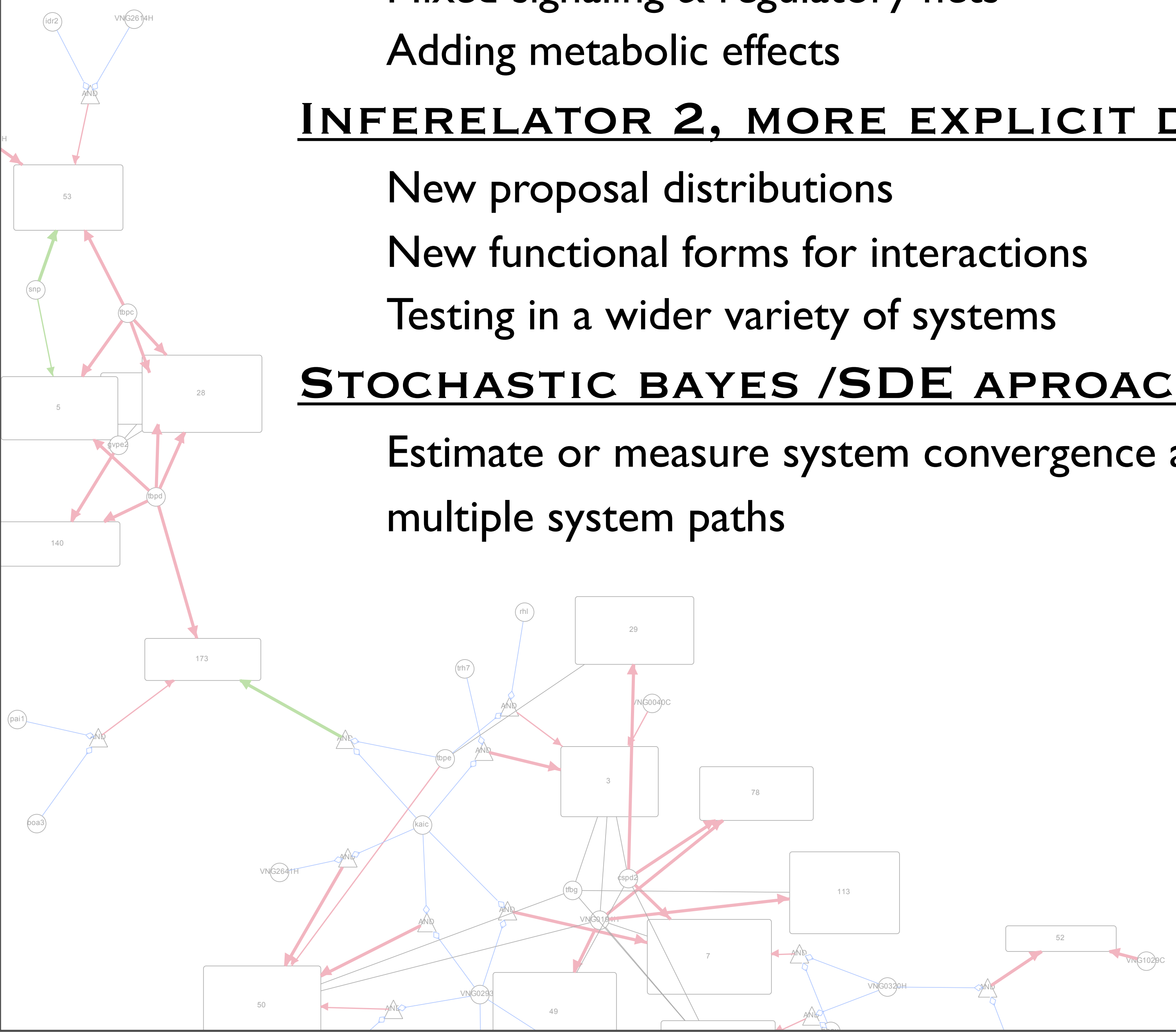
New proposal distributions

New functional forms for interactions

Testing in a wider variety of systems

Stochastic Bayes /SDE aproach:

Estimate or measure system convergence as well as mean, model error, multiple system paths



post-docs for protein design, prediction & network inference

ashington Square (February 1998)

Acknowledgments

Bonneau lab:

Glenn Butterfoss
Kevin Drew
Aviv Madar
Peter Waltman
Thadeous Kacmarczyk
Shailla Musharof
Devorah Kengmana
Chris Poultny (Shasha)
Irina Nudelman
Alex Pearlman (Ostrer)
Alex Pine

NYU:

Eric Vanden-Eijnden
Harry Ostrer
Mike Purugganan
Patrick Eichenberger
Dennis Shasha

Tacitus-
Howard Coale

- **IBM**
  - Robin Wilner
  - Bill Boverman
  - Viktors Berstis
  - Rick Alther
- ETH Zurich
  - Reudi Aebersold
  - Lars Malmstroem

Mike Boxem
Marc Vidal

Dave Goodlett

Jochen Supper (Zell Lab)

- ISB

- Nitin Baliga (&lab)
- Leroy Hood
- Marc Facciotti
- David Reiss
- Vesteinn Thorsson
- Paul Shannon
- Iliana Avila-Campillo (MERC)

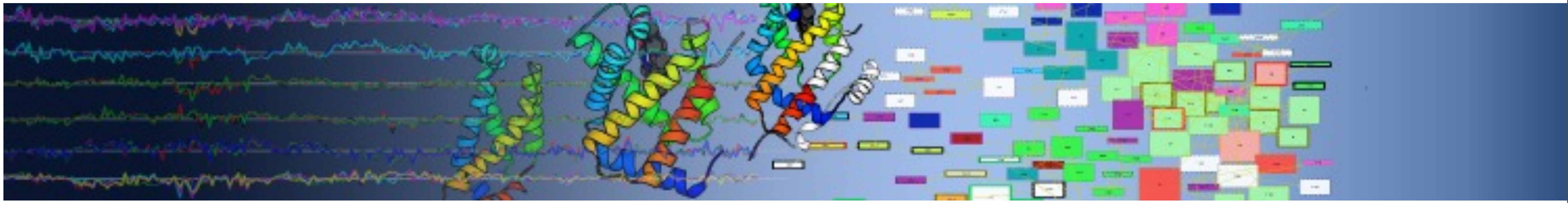
Alan Aderem

Rosetta Commons

Charlie Strauss (los alamos)
David Baker (UW seattle)

DOD-computing and society,
NSF ABI,
NSF Plant genome
NSF DBI,
DOE GTL

Rosetta de novo structure prediction: The Human Proteome folding Project



Kevin Drew,

Lars Malmstroem,

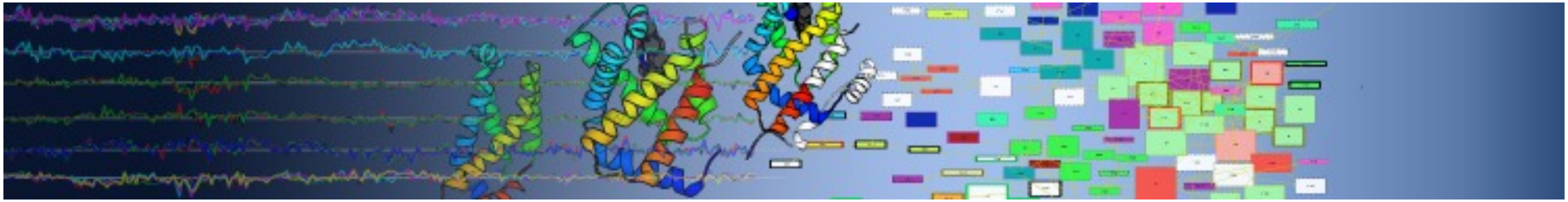
Glenn Butterfoss,

Richard Bonneau

Rosetta Commons

CENTER FOR GENOMICS
AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY

COURANT
INSTITUTE



Motivation: Genome Annotation

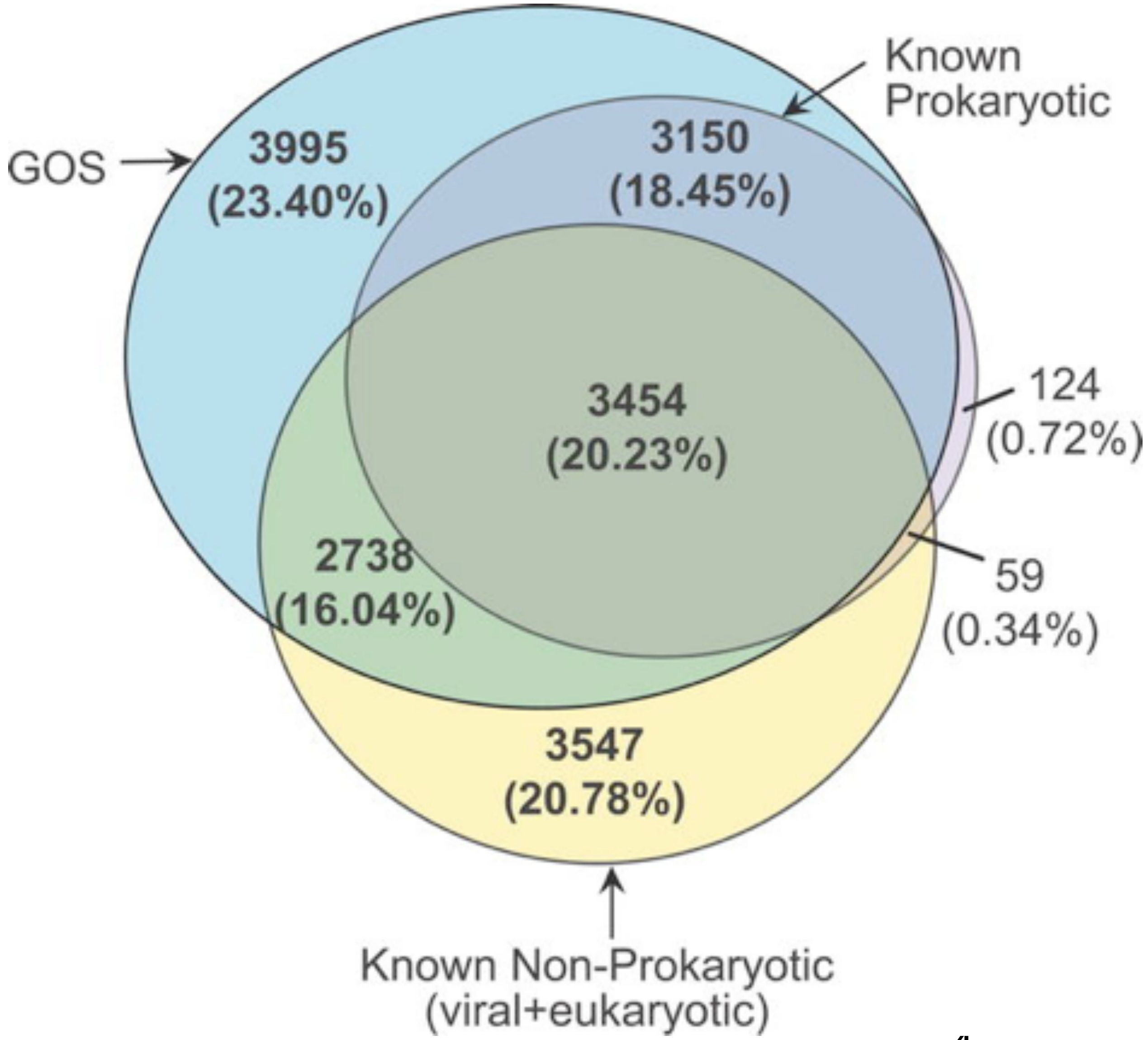
Cheaper sequencing technologies

↓

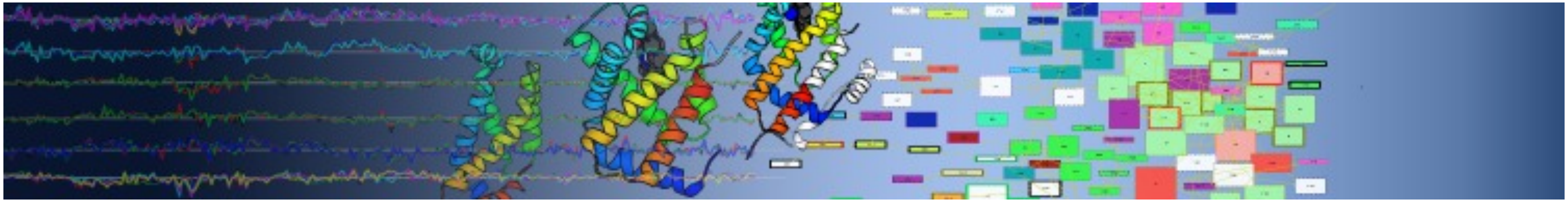
New protein sequences

↓

Proteins w/ unknown function

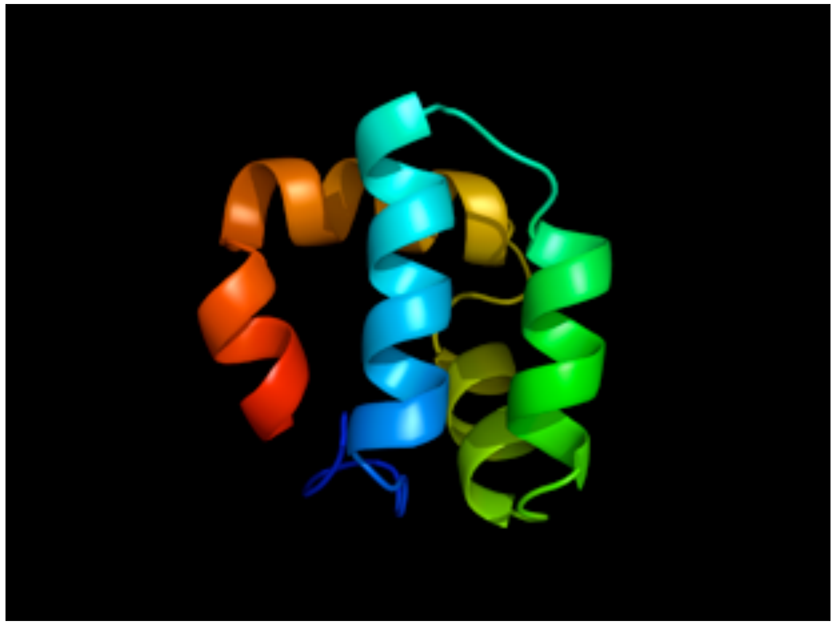
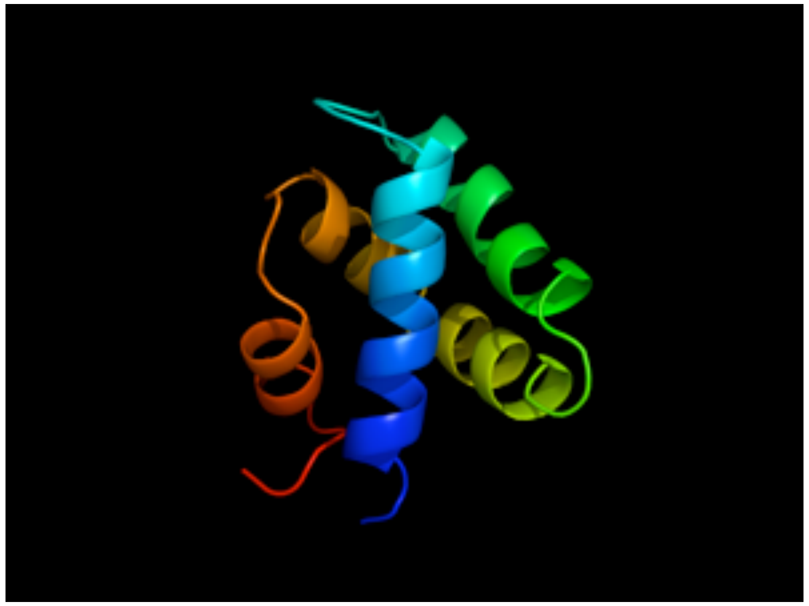


Shibu Yooseph et al. 2007 Plos Biology

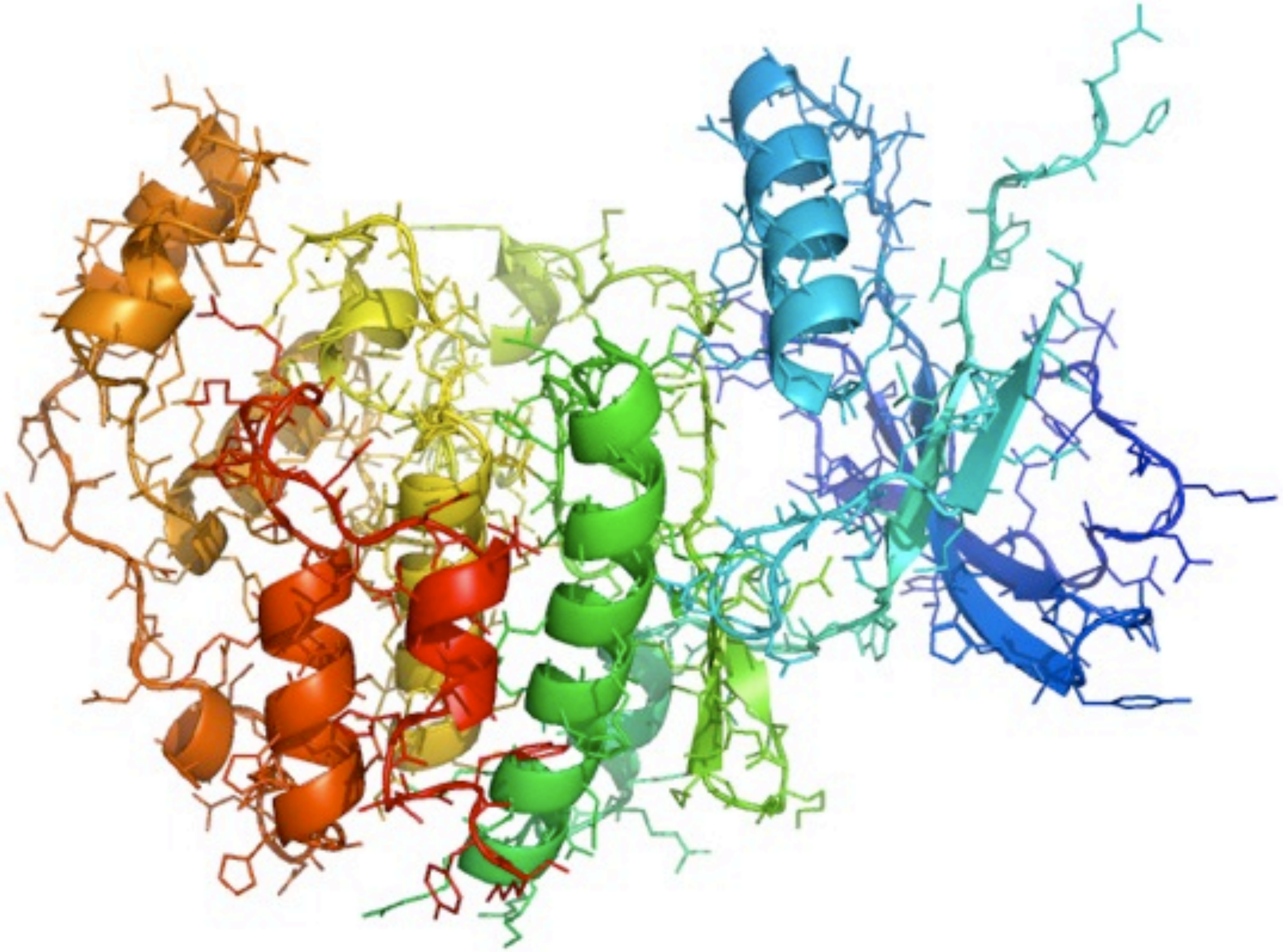


Background: Quick Example

Bacteriocin AS-48, Casp 4

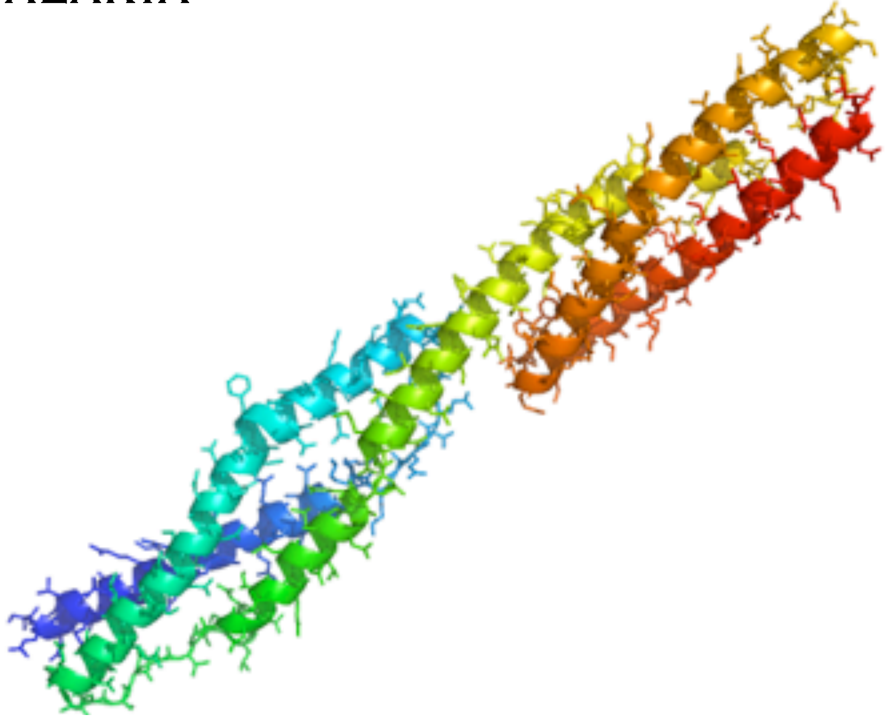
	1E68		1NKL
Sequence:	MAKEFGIPAAVAGTVLNVVEAGGW VTTIVSILTAVGSGGLSLLAAAGRES IKAYLKKEIKKKGKRAVIAW	4%=	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAAS QVCDKLKILRGLCKKIMRSFLRRISWDILTGKKP QAICVDIKICKE
Structure:		=	
Function:	Cyclic Bacterial Lysin	=	NK Lysin

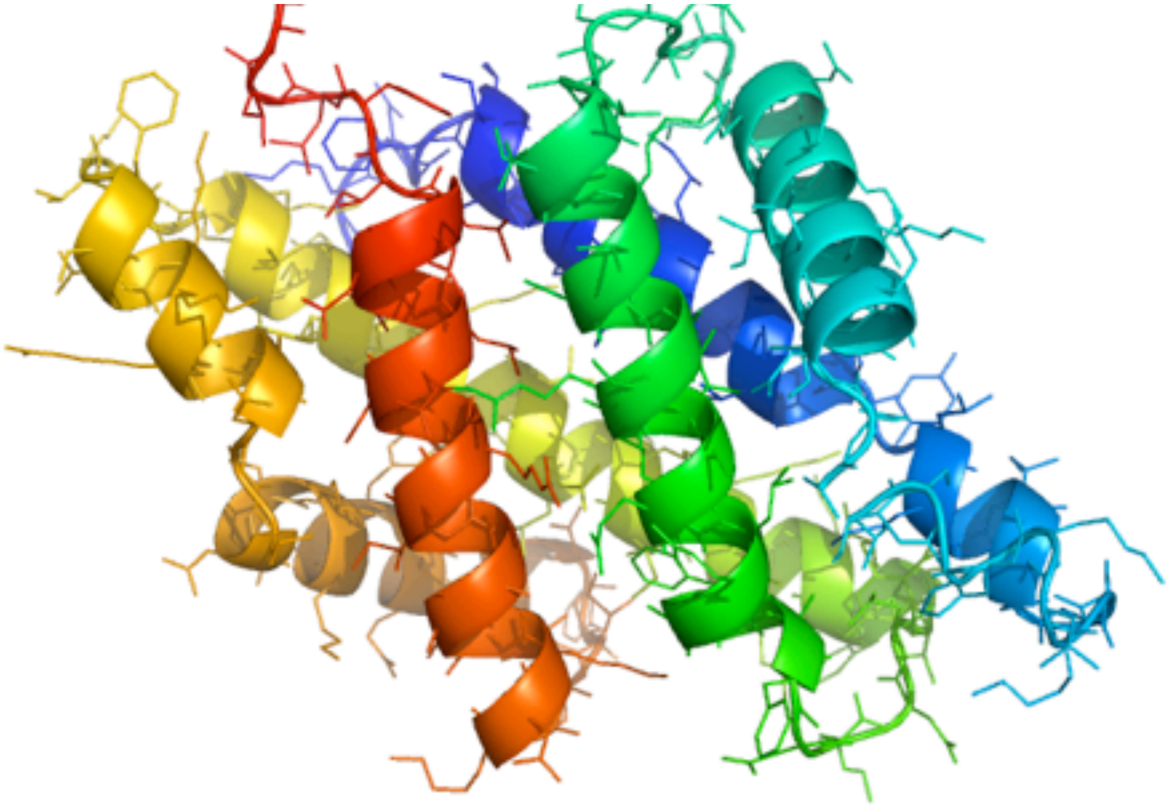
Bonneau, R., Tsai, J., Ruczinski, I., Baker, D. Functional Inferences from Blind ab Initio Protein Structure Predictions. J. Structural Biology. (2001)

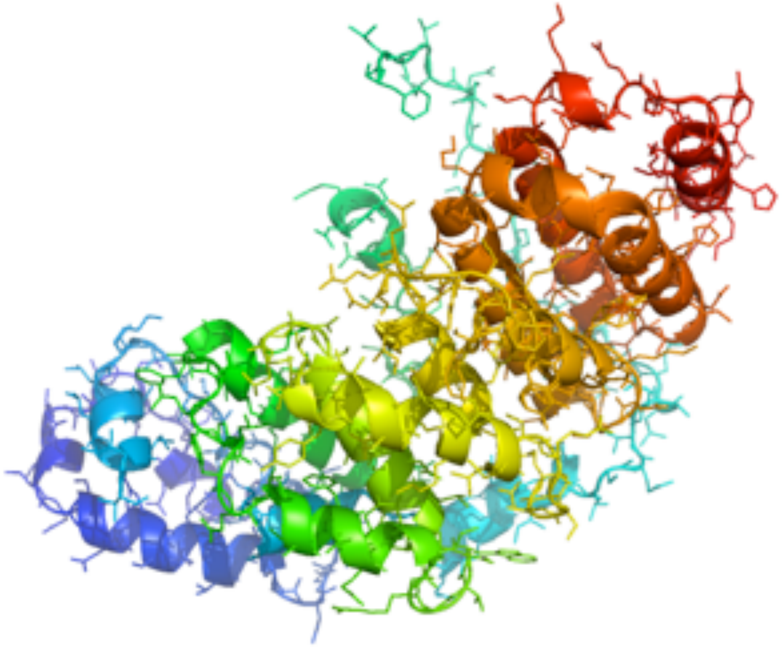


Plasmodium

SBRI top candidates for Vaccine for preventing pregnancy malaria





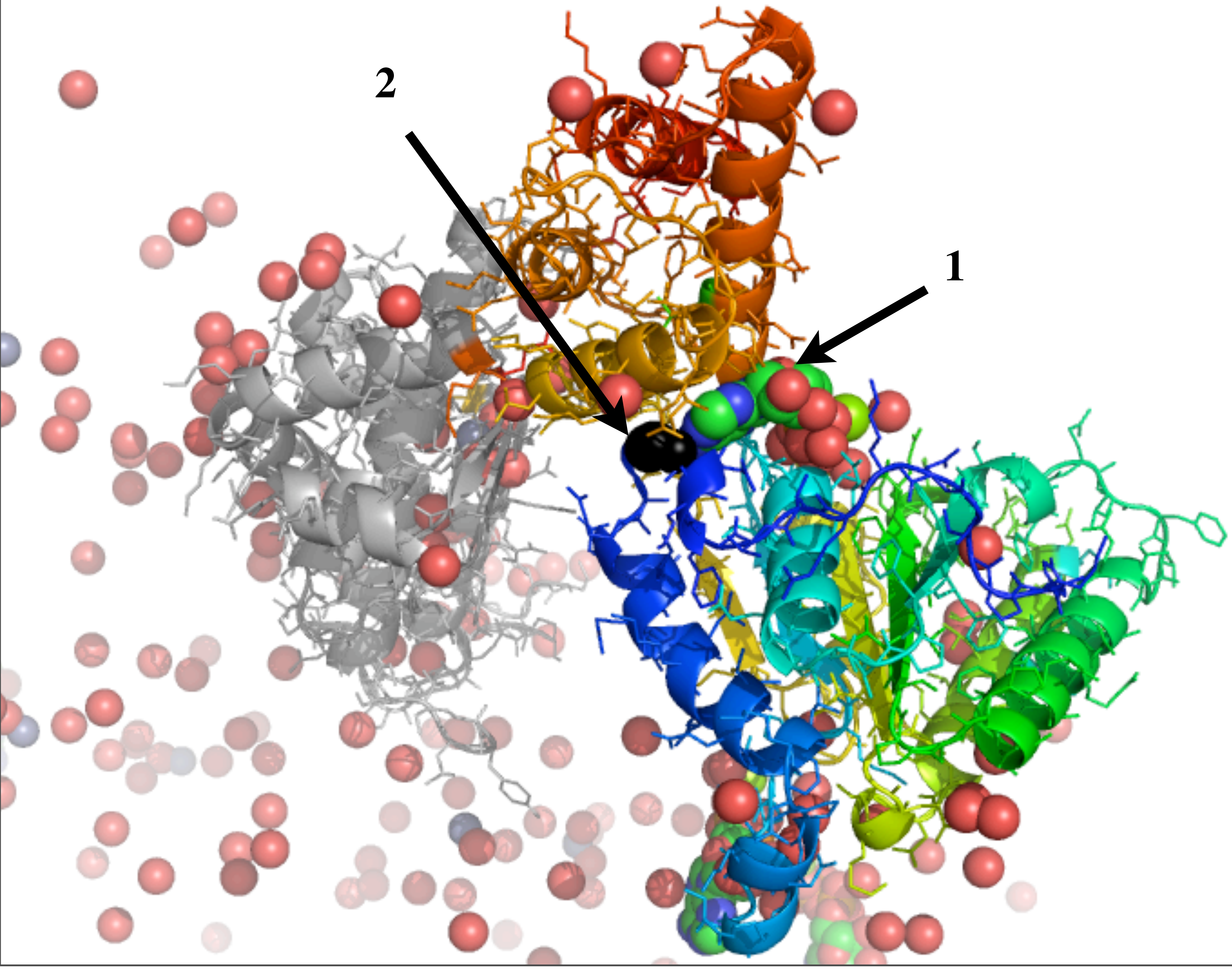


Arabidopsis example:

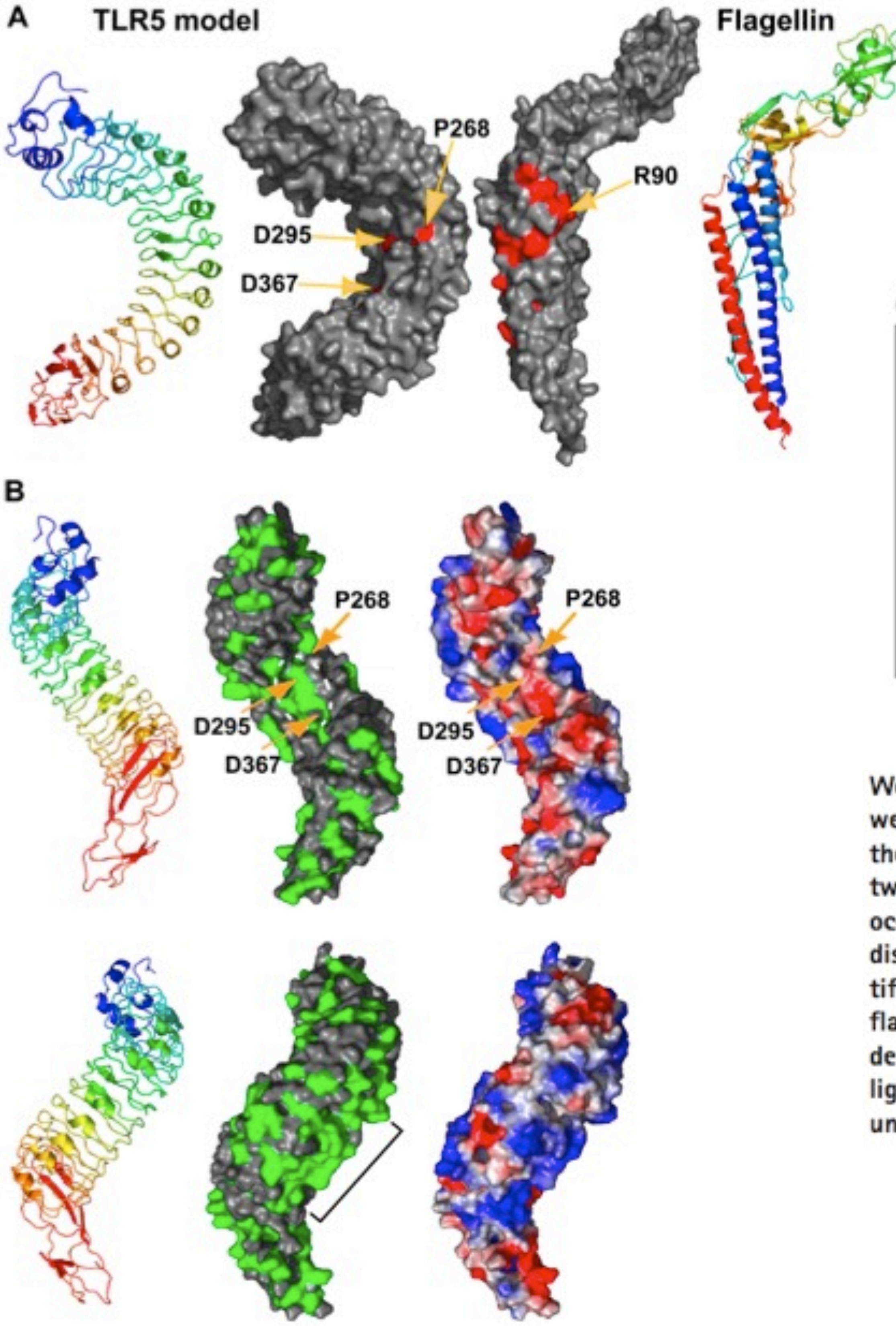
RPT3

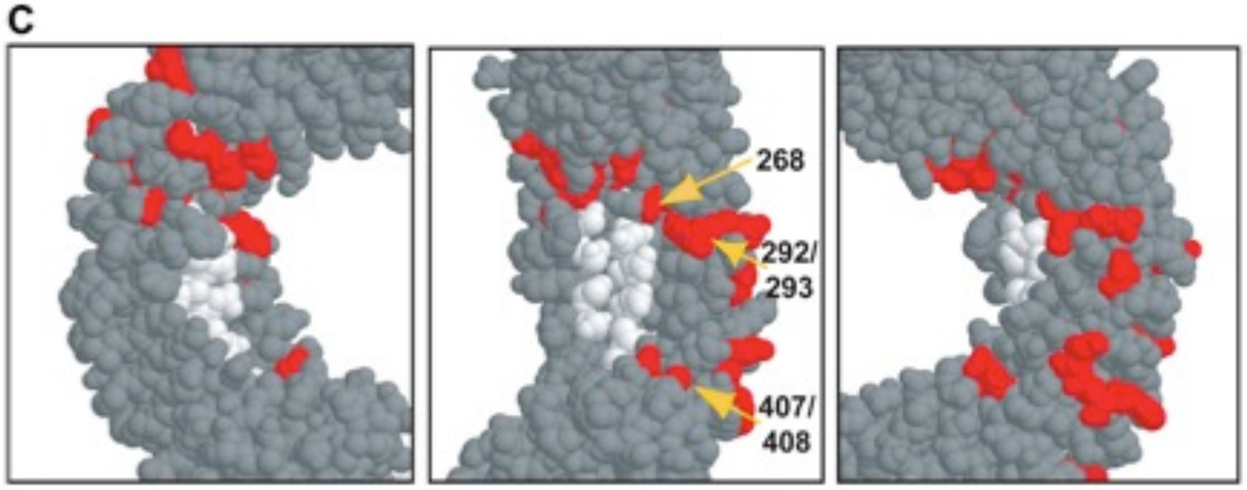
1: cofactor

2: point of mutation causing differential response to morphogen



Distant Multi-template fold recognition for Toll receptors





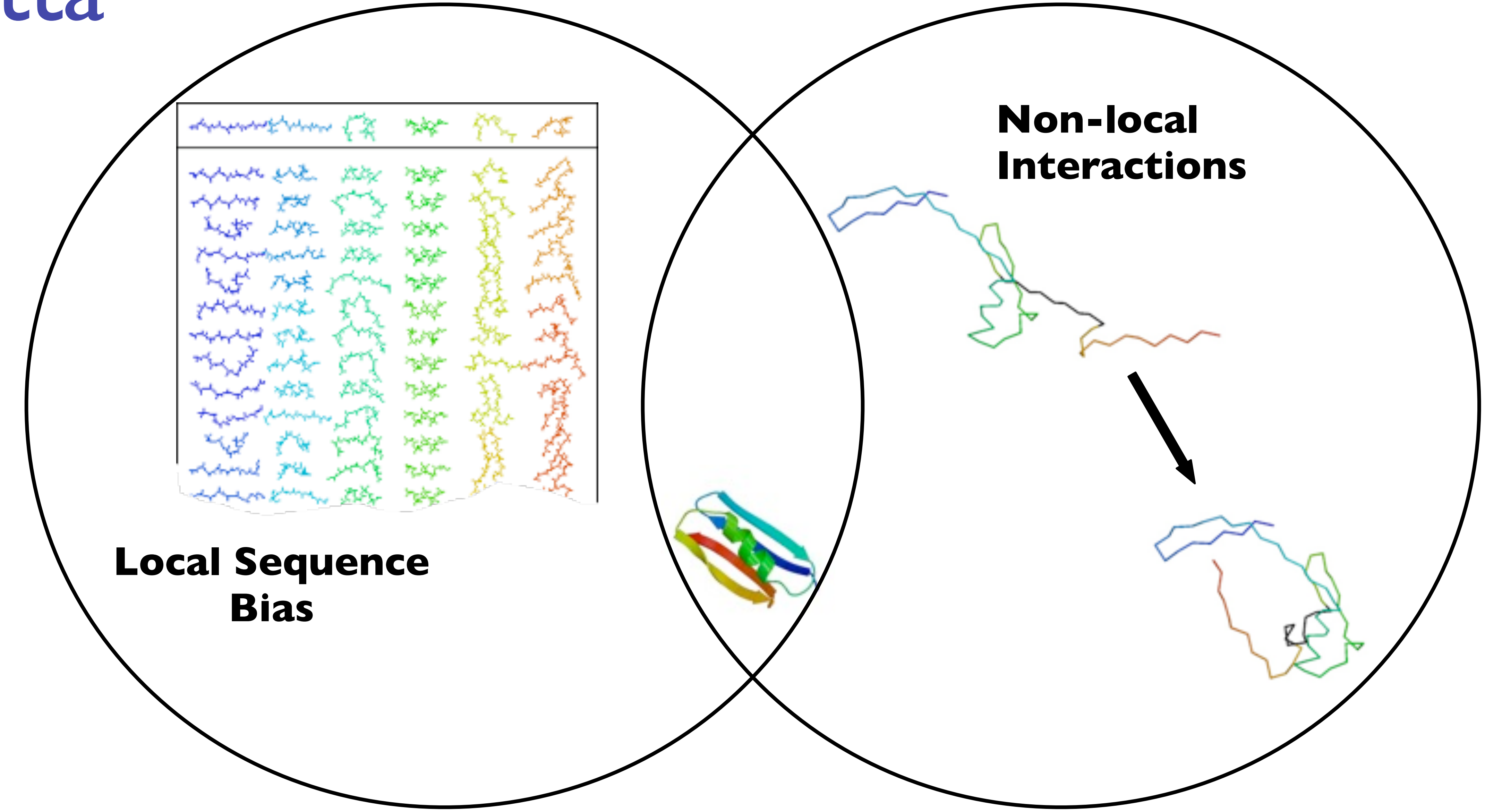
We demonstrate that mouse and human TLR5 discriminate between different flagellins, and we use this difference to map the flagellin recognition site on TLR5 to 228 amino acids of the extracellular domain. Through molecular modeling of the TLR5 ectodomain, we identify two conserved surface-exposed regions. Mutagenesis studies demonstrate that naturally occurring amino acid variation in TLR5 residue 268 is responsible for human and mouse discrimination between flagellin molecules. Mutations within one conserved surface identify residues D295 and D367 as important for flagellin recognition. These studies localize flagellin recognition to a conserved surface on the modeled TLR5 structure, providing detailed analysis of the interaction of a TLR with its ligand. These findings suggest that ligand binding at the β sheets results in TLR activation and provide a new framework for understanding TLR–agonist interactions.

E Andersen-Nissen, R Bonneau,
R Strong, A Aderem
Journal of Experimental Medicine, 2007

Lars Malmstroem

Kevin Drew

Rosetta

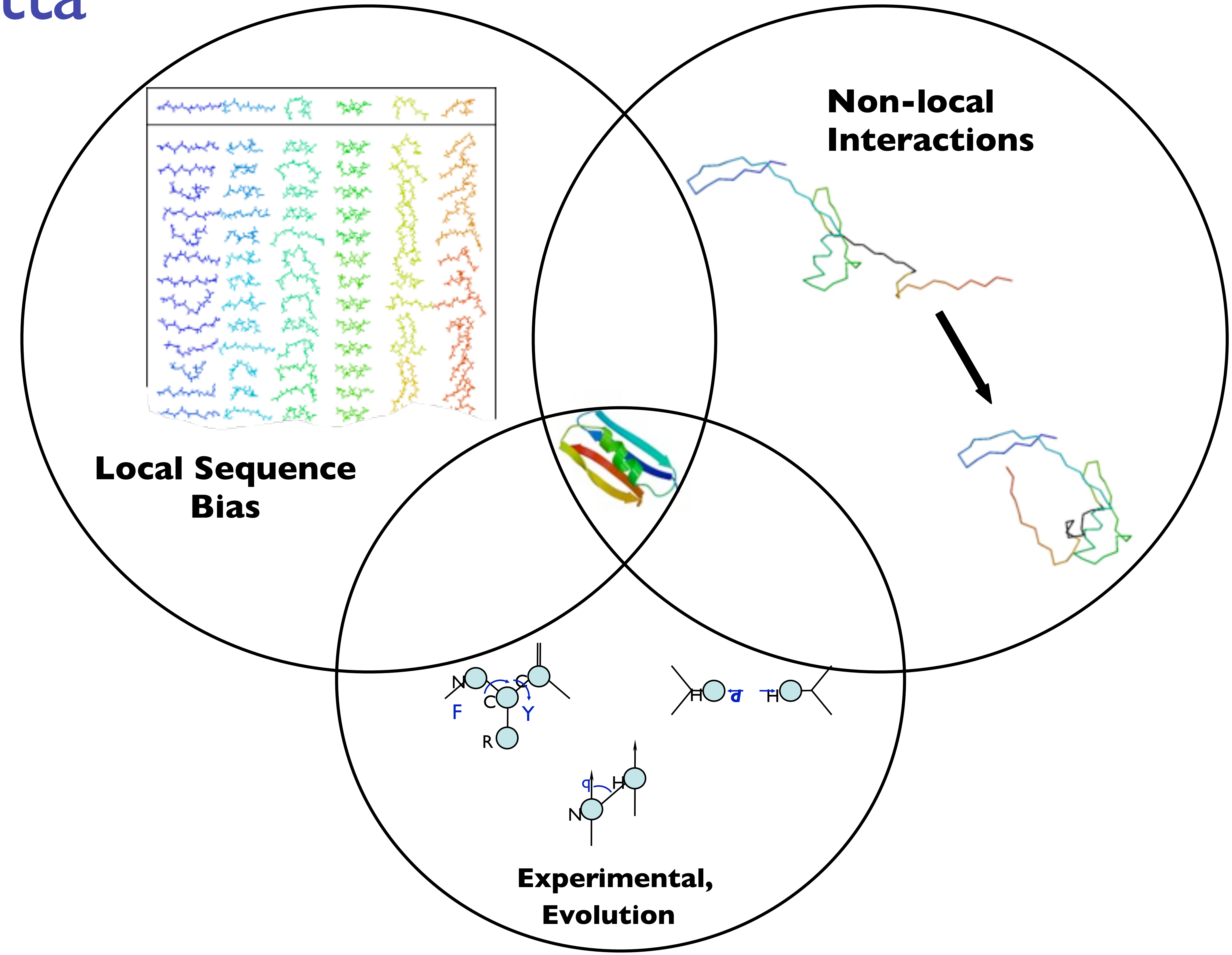


Local Sequence Bias

Non-local Interactions

Kevin Drew, Chivian, D., Bonneau, R. Ab initio structure prediction. (In) Bourne, P.E. (2007) Structural Bioinformatics (Methods of Biochemical Analysis, V. 44). New York: John Wiley & Sons; ISBN: 0471201995. Second Edition.

Rosetta



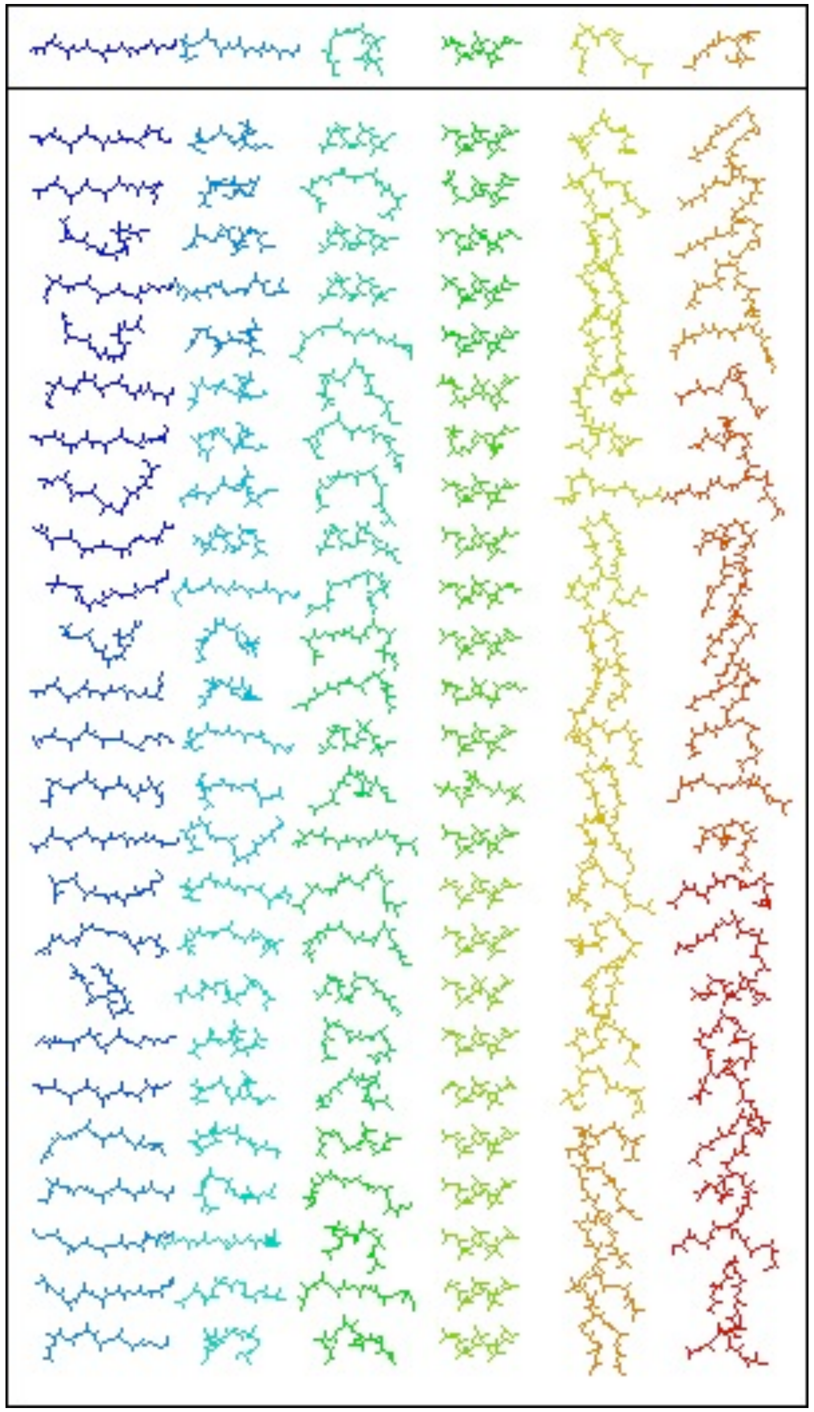
Non-local Interactions

Local Sequence Bias

Experimental, Evolution

Kevin Drew, Chivian, D., Bonneau, R. Ab initio structure prediction. (In) Bourne, P.E. (2007) Structural Bioinformatics (Methods of Biochemical Analysis, V. 44). New York: John Wiley & Sons; ISBN: 0471201995. Second Edition.

Rosetta Fragment Libraries



- 25-200 fragments for each/every 3 and 9 residue sequence window (overlapping)

- Selected from database of known structures
  > 2.5Å resolution
  < 50% sequence identity

- Ranked by sequence similarity and similarity of predicted and known secondary structure

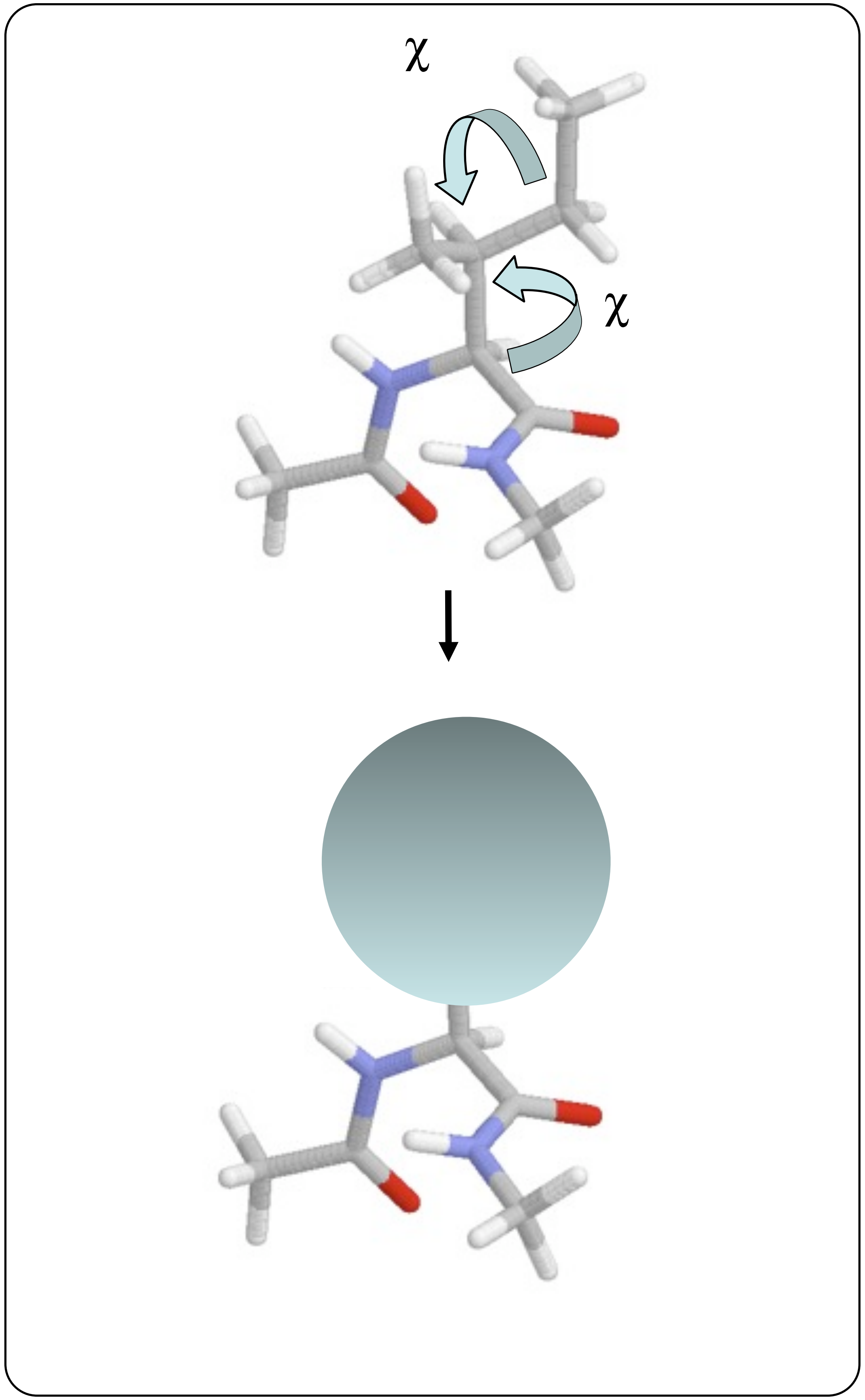
- Fragments restrict search to protein-like local conformations

Sequence similar or exact sequence **BUT** not long enough that the similarity is attributable to evolution (not homologous strictly speaking).

Low resolution:

Atom Model

centroid reduction of side chains



Energy function terms

van der Waals repulsion

“pair” terms (electrostatics)

residue environment (prob of burial)

2º structure pairing terms (H-bonds)

radius of gyration

packing density

Implicit terms

fragments (local interactions)

Low resolution:

Atom Model

centroid reduction of side chains

Energy function terms

van der Waals repulsion

“pair” terms (electrostatics)

residue environment (prob of burial)

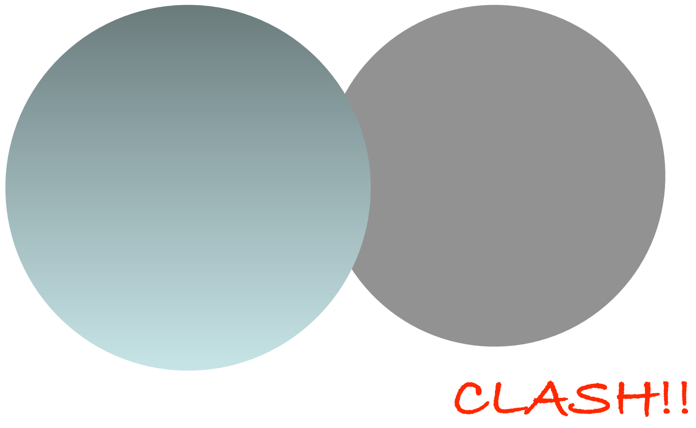
2º structure pairing terms (H-bonds)

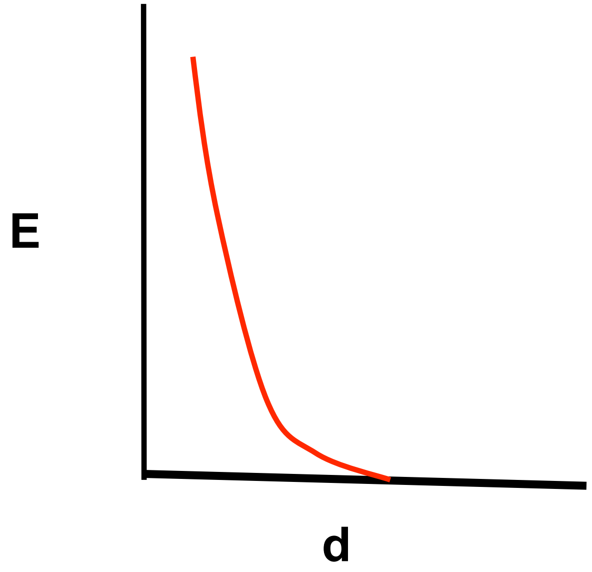
radius of gyration

packing density

Implicit terms

fragments (local interactions)





$$\sum_{i}\sum_{j<i}\frac{(r_{ij}^2-d_{ij}^2)^2}{r_{ij}};d_{ij}<r_{ij}$$

$$d = distance$$

$$r = \sum radii$$

Evaluate between Centoids and Backbone Atoms

Low resolution:

Atom Model

centroid reduction of side chains

Energy function terms

van der Waals repulsion

“pair” terms (electrostatics)

residue environment (prob of burial)

2º structure pairing terms (H-bonds)

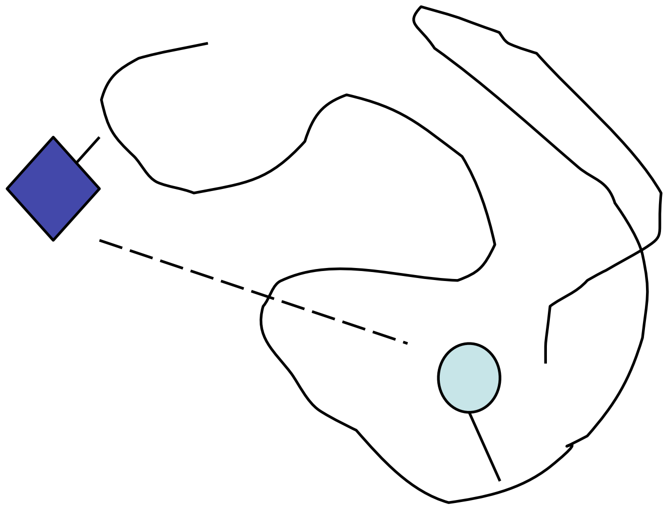
radius of gyration

packing density

Implicit terms

fragments (local interactions)

Pair-wise probability based on PDB statistics (electrostatics)



$$\sum_{i}\sum_{j>i}-\ln\left\lfloor\frac{P(aa_i,aa_j \mid s_{ij}d_{ij})}{P(aa_i \mid s_{ij}d_{ij})P(aa_i \mid s_{ij}d_{ij})}\right\rfloor$$

aa = residue type
d = centroid distance (binned, interpolated)
s = sequence seperation (must be > 8 res)

Low resolution:

Atom Model

- centroid reduction of side chains

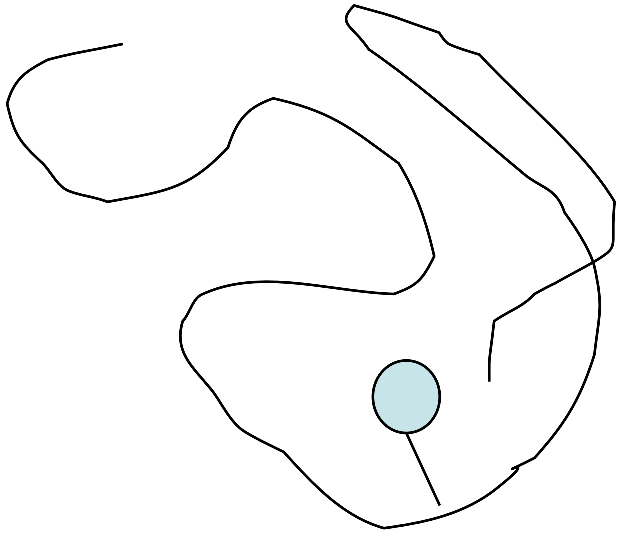
Energy function terms

- van der Waals repulsion
- “pair” terms (electrostatics)
- **residue environment (prob of burial)**
- 2º structure pairing terms (H-bonds)
- radius of gyration
- packing density

Implicit terms

- fragments (local interactions)

Probability of burial /exposure (solvation)



$$\sum_{i} -\ln\left[P(aa_i \mid neighbors_i)\right]$$

neighbors within 10 Å of Cβ

binned by : 0-3, 4,5, … , >30

also interpolated

Low resolution:

Atom Model

centroid reduction of side chains

Energy function terms

van der Waals repulsion

“pair” terms (electrostatics)

residue environment (prob of burial)

2º structure pairing terms (H-bonds)

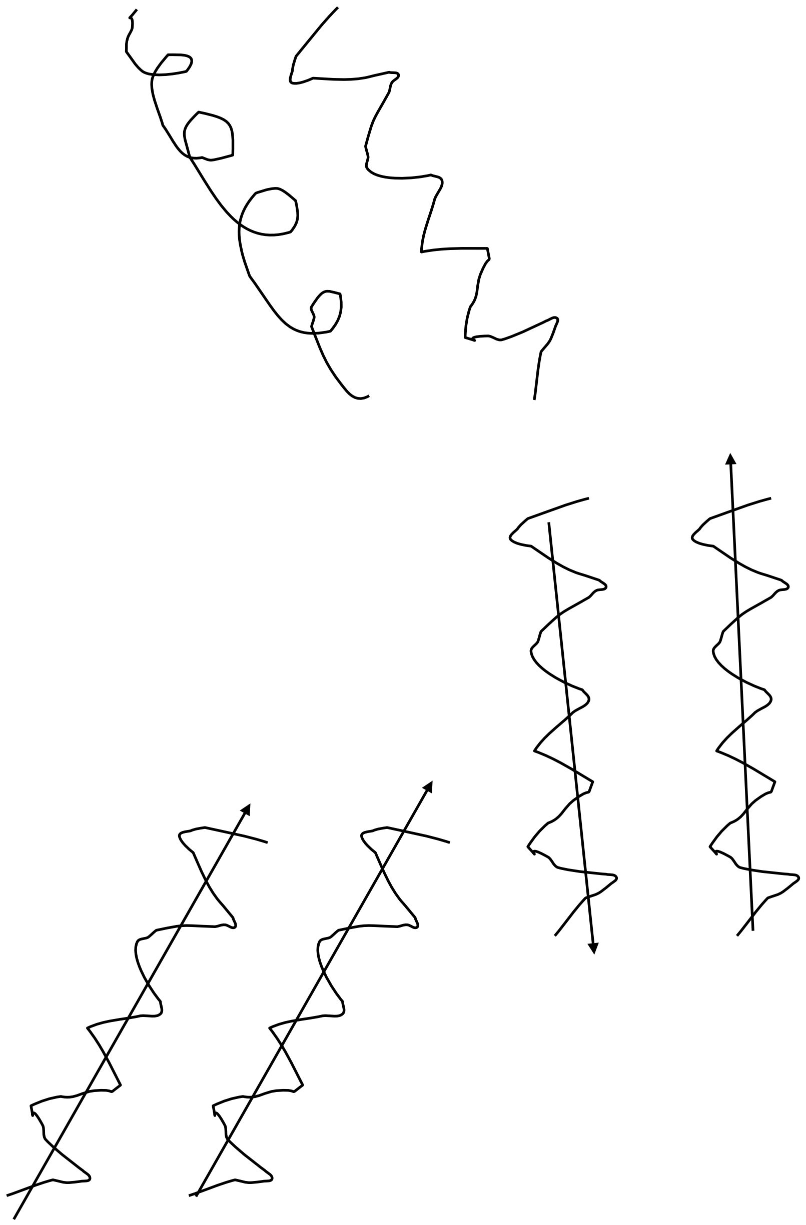
radius of gyration

packing density

Implicit terms

fragments (local interactions)

Optimize 2º orientation



Low resolution:

Atom Model

- centroid reduction of side chains

Energy function terms

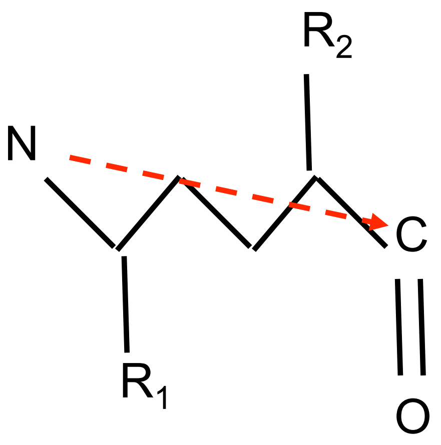
- van der Waals repulsion
- “pair” terms (electrostatics)
- residue environment (prob of burial)
- **2^{o} structure pairing terms (H-bonds)**
- radius of gyration
- packing density

Implicit terms

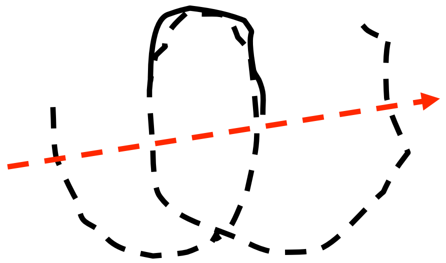
- fragments (local interactions)

Represent protein as vectors of 2 residue “strands”

sheet vector



helix vector



Low resolution:

Atom Model

centroid reduction of side chains

Energy function terms

van der Waals repulsion

“pair” terms (electrostatics)

residue environment (prob of burial)

2º structure pairing terms (H-bonds)

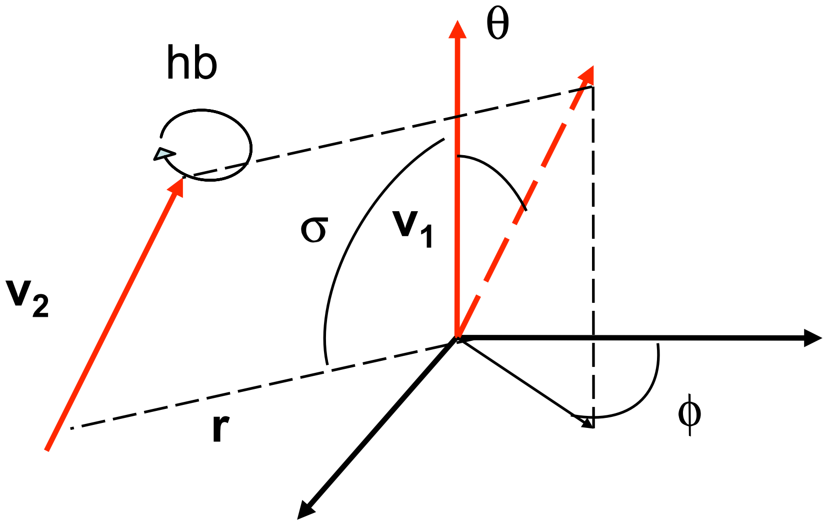
radius of gyration

packing density

Implicit terms

fragments (local interactions)

Coordinate system



Scores selected to discriminate “near native structures for “non native”:

Relative direction (φ,θ)

Relative H-bond orientation (hb)

Distance (r, rσ)

Number of sheets given number of strands

Helix-Strand Packing

Low resolution:

Atom Model

centroid reduction of side chains

Energy function terms

van der Waals repulsion

“pair” terms (electrostatics)

residue environment (prob of burial)

2º structure pairing terms (H-bonds)

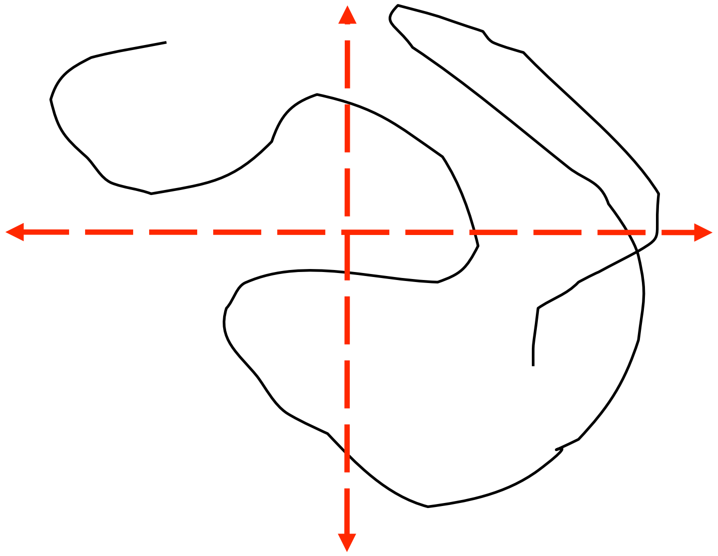
radius of gyration

packing density

Implicit terms

fragments (local interactions)

Promote a compact fold



$$RG = \sqrt{\langle d_{ij}^2 \rangle}$$

$$Density = \sum_{i} \sum_{sh} -\ln\left[\frac{P_{compact}(neighbors_{i,sh})}{P_{random}(neighbors_{i,sh})}\right]$$

Used in earlier stages and for filtering

High resolution:

Atom Model

full atom representation

Energy function terms

Rotamer (Dunbrack)

Ramachandran

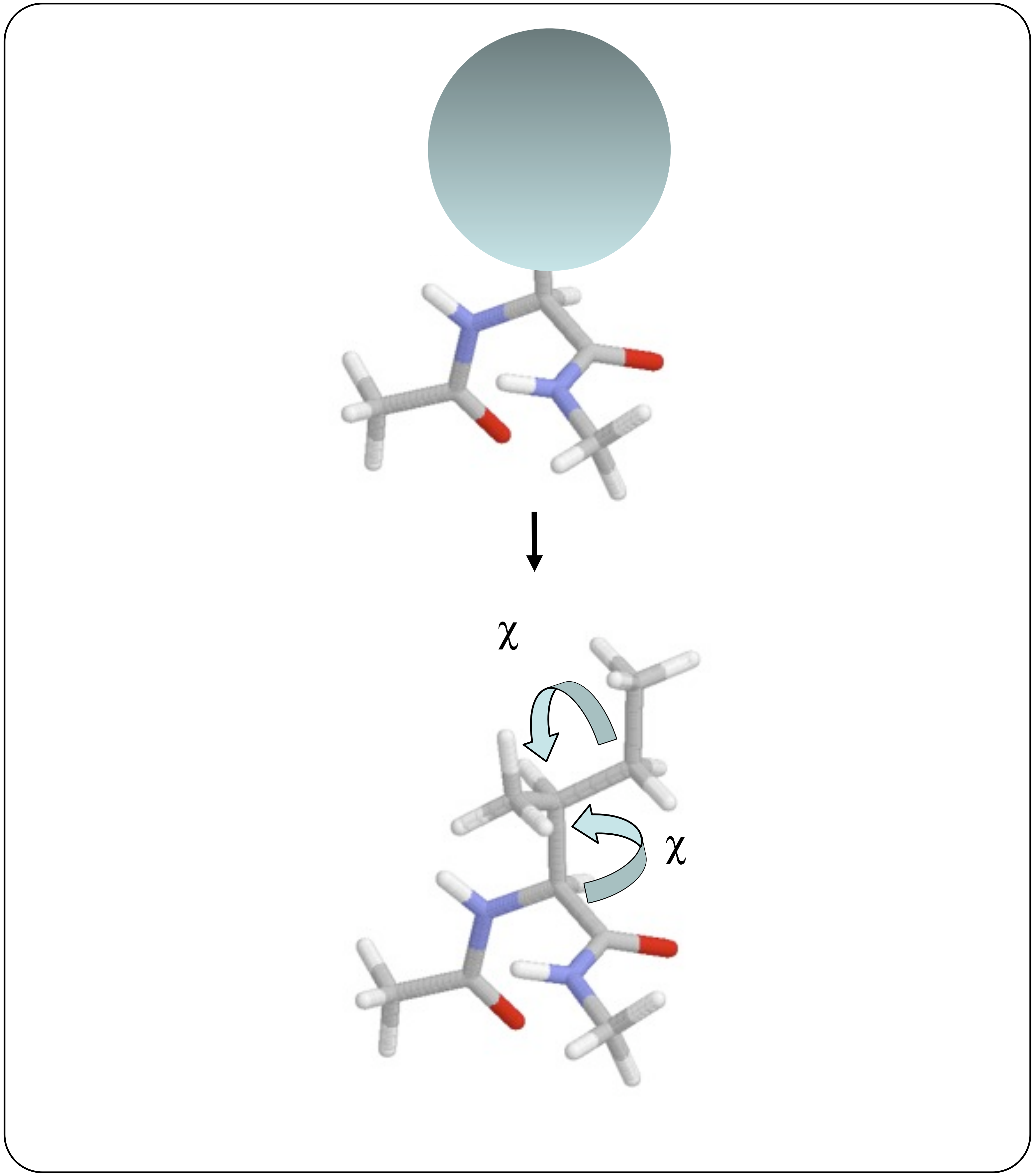
Solvation (Lazaridius Karplus)

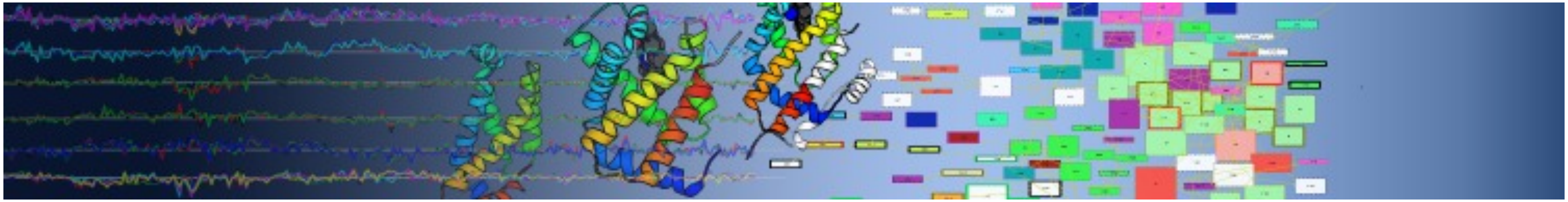
Hydrogen bonding

Lennard-Jones

Pair (electrostatic)

Reference energies





Process of Obtaining Structures

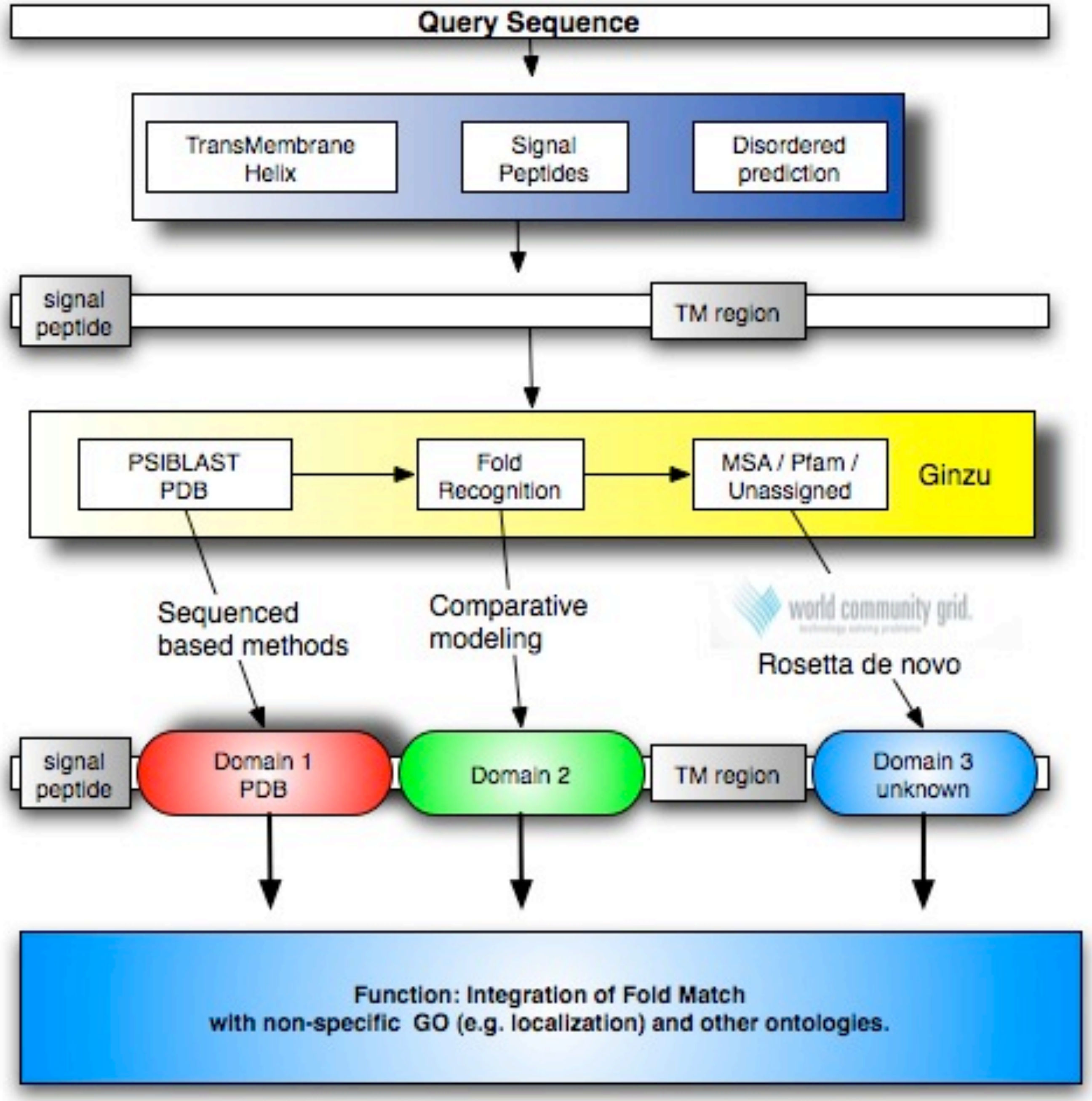
1. Split proteins into domains (ginzu, Chivian)
(chop, Rost)

2. Find domains we can annotate using Rosetta

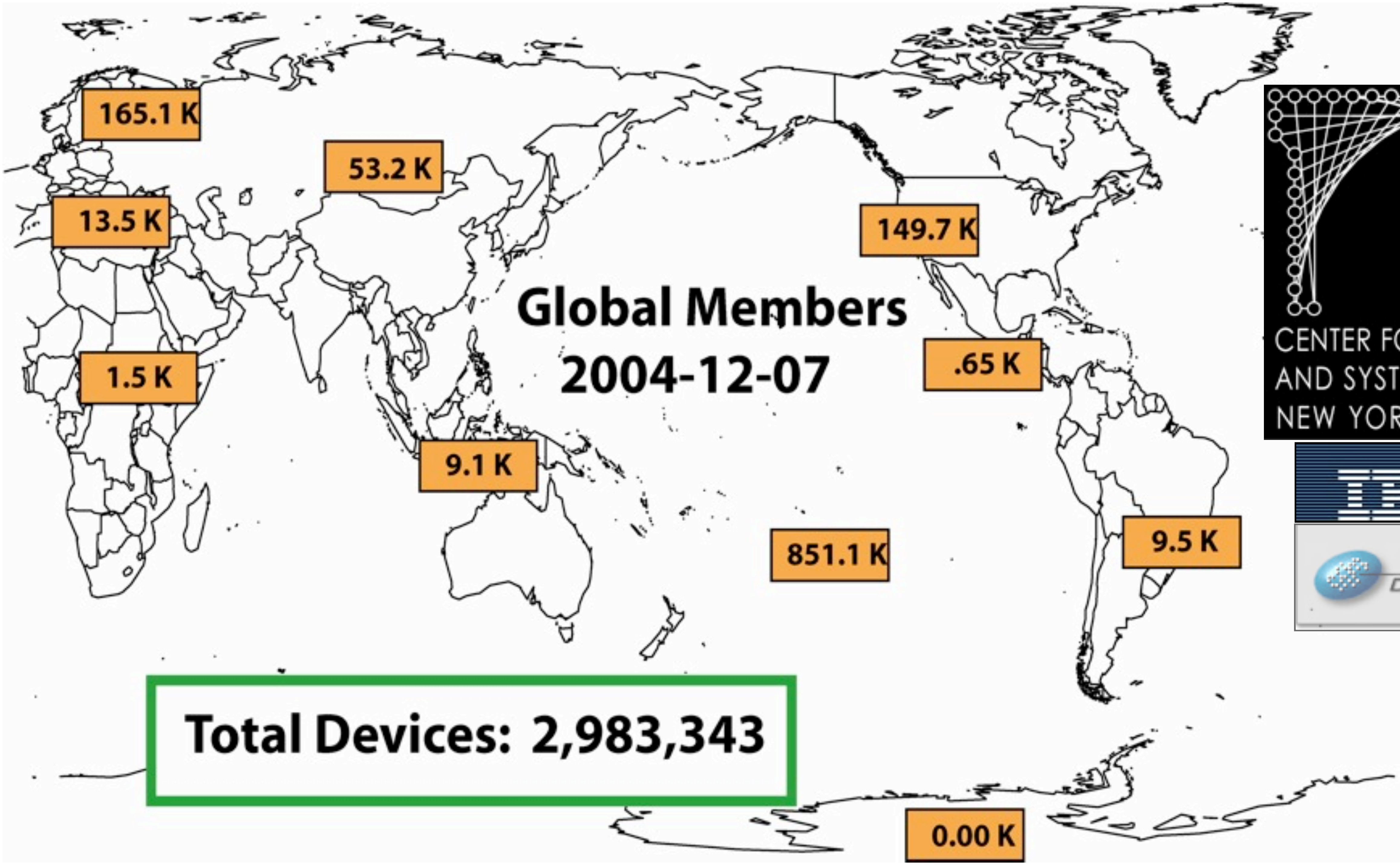
3. Fold remaining domains using Rosetta on IBM's World Community Grid

- 180,000 domains folded from 120 genomes

BIG caveat emptor:
all results from this point for domains < 170 aa



worldcommunitygrid.org & grid.org



CENTER FOR GENOMICS AND SYSTEMS BIOLOGY NEW YORK UNIVERSITY

IBM

UNITED DEVICES™

collaborators: Lars Malmstroem, Viktors Berstis, Mike Riffle, Leroy Hood, David Baker

Completed and ongoing projects

Bacterial and Archaea:

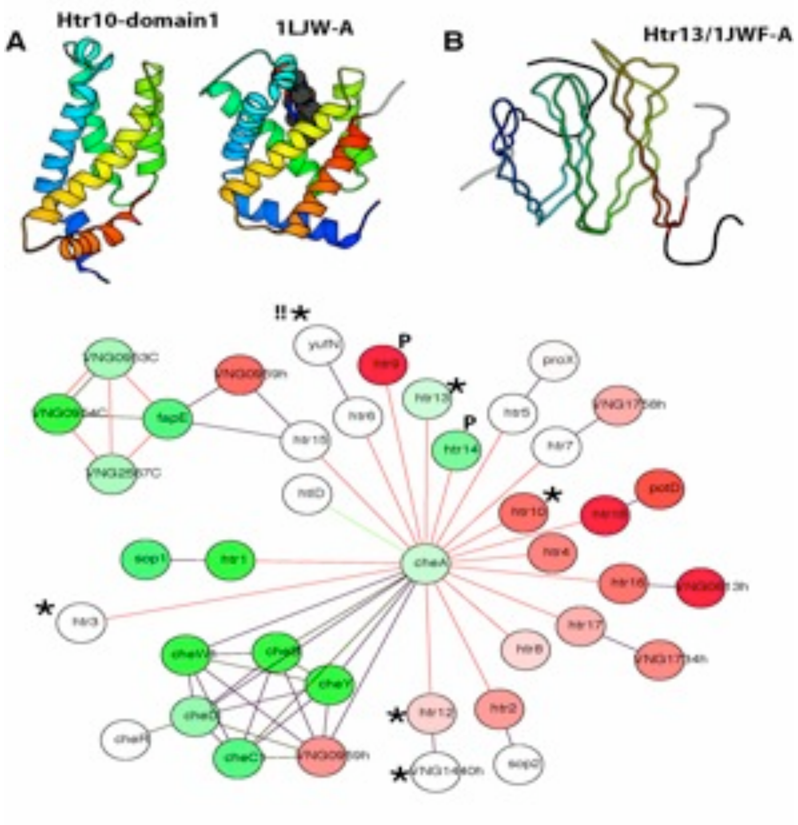
Bonneau, & Baliga. (2004)Genome Biology:
Annotaion of **Halobacterium NRC-1**
identification of transcription factors
role of chemotaxis sensing domains

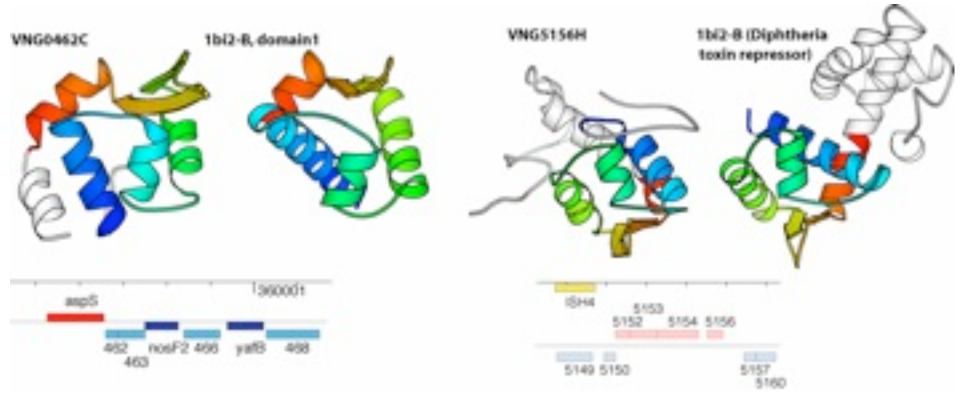
Yeast:

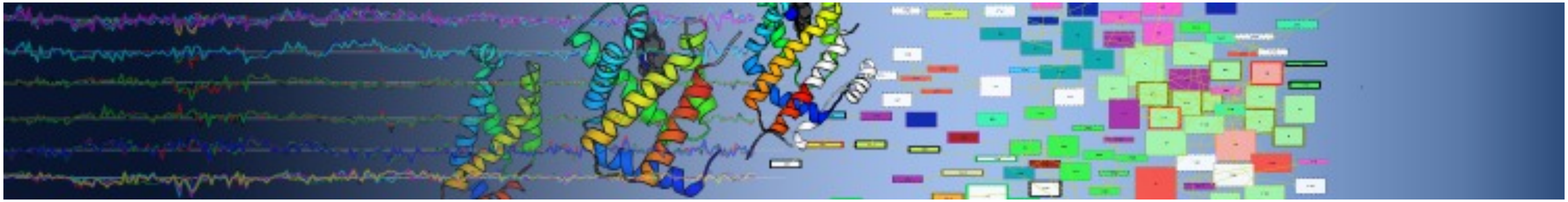
Malstroem, Baker, Bonneau (2006) Plos Biology

Human & others:

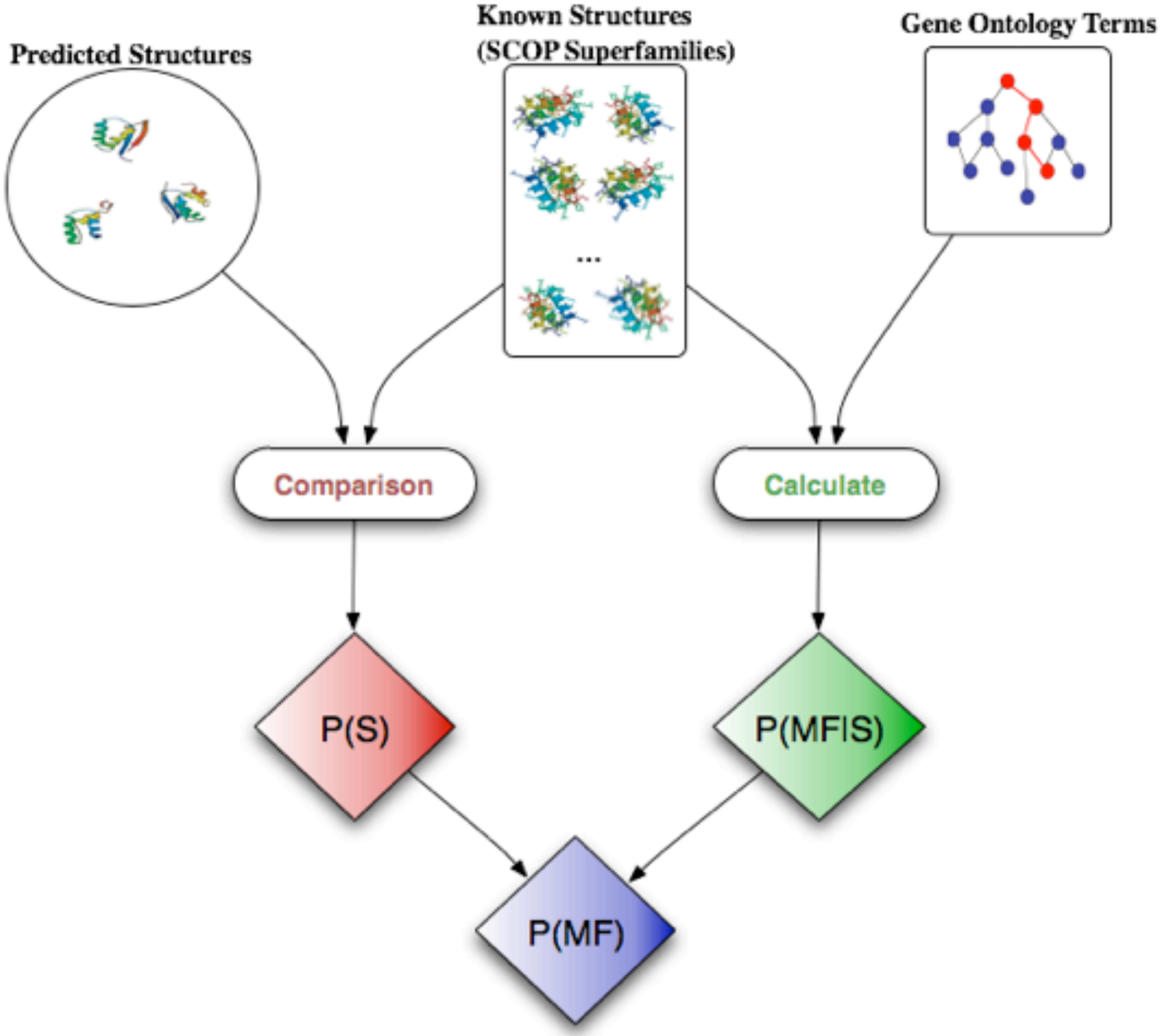
Bonneau, Malstroem, IBM:
Human and others (in Progress)

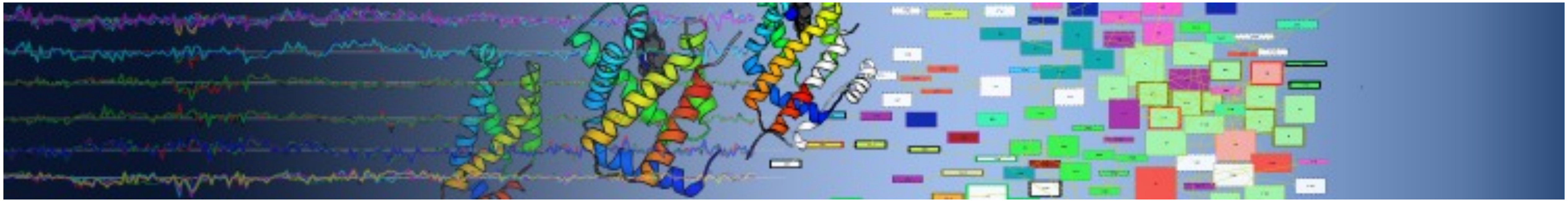






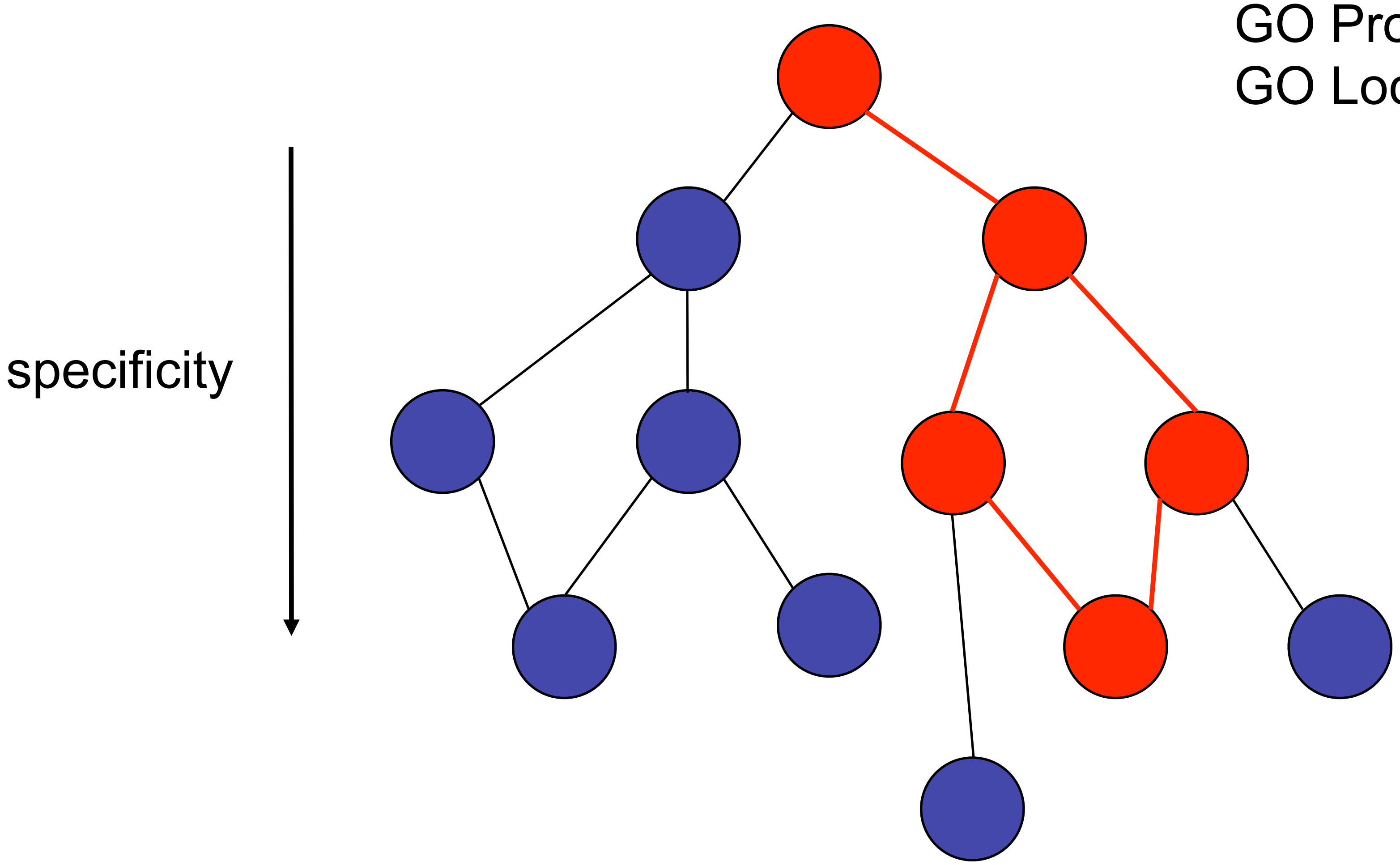
overview of approach



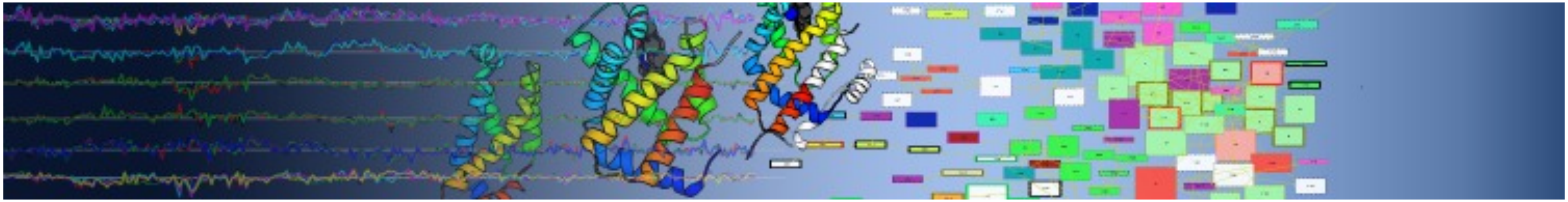


overview of approach

P(MF | Predicted Structure,
GO Process,
GO Localization, …)



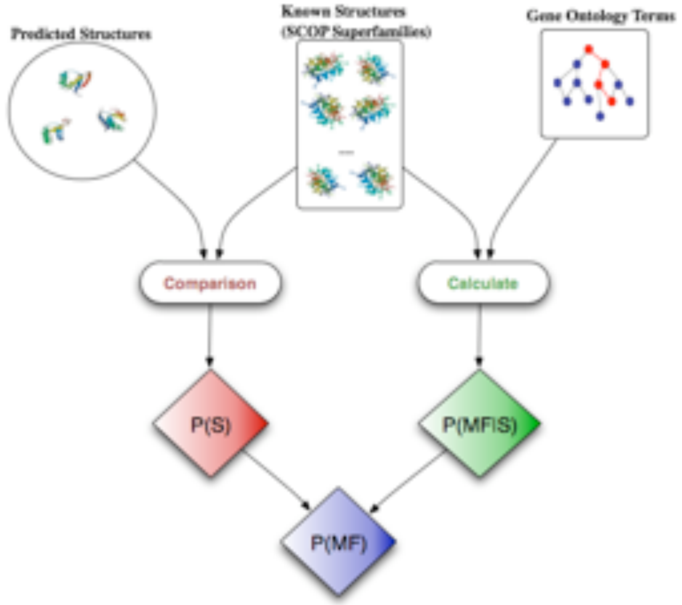
Molecular Function GO Tree



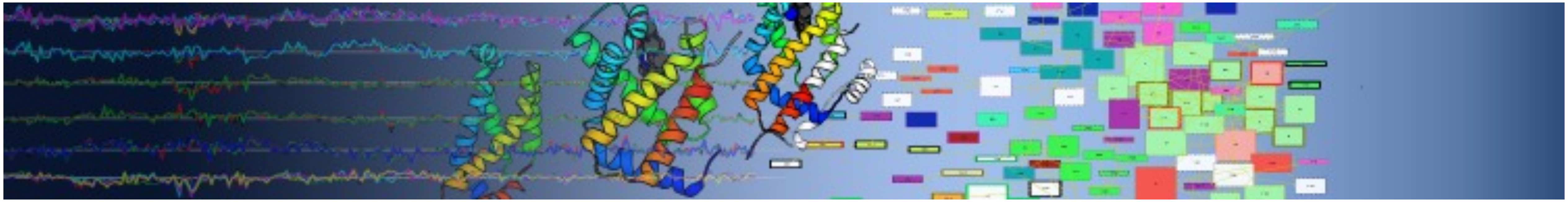
overview of approach

1. Structure (MCM) Score

2. Training Set (attaching structure to GO)

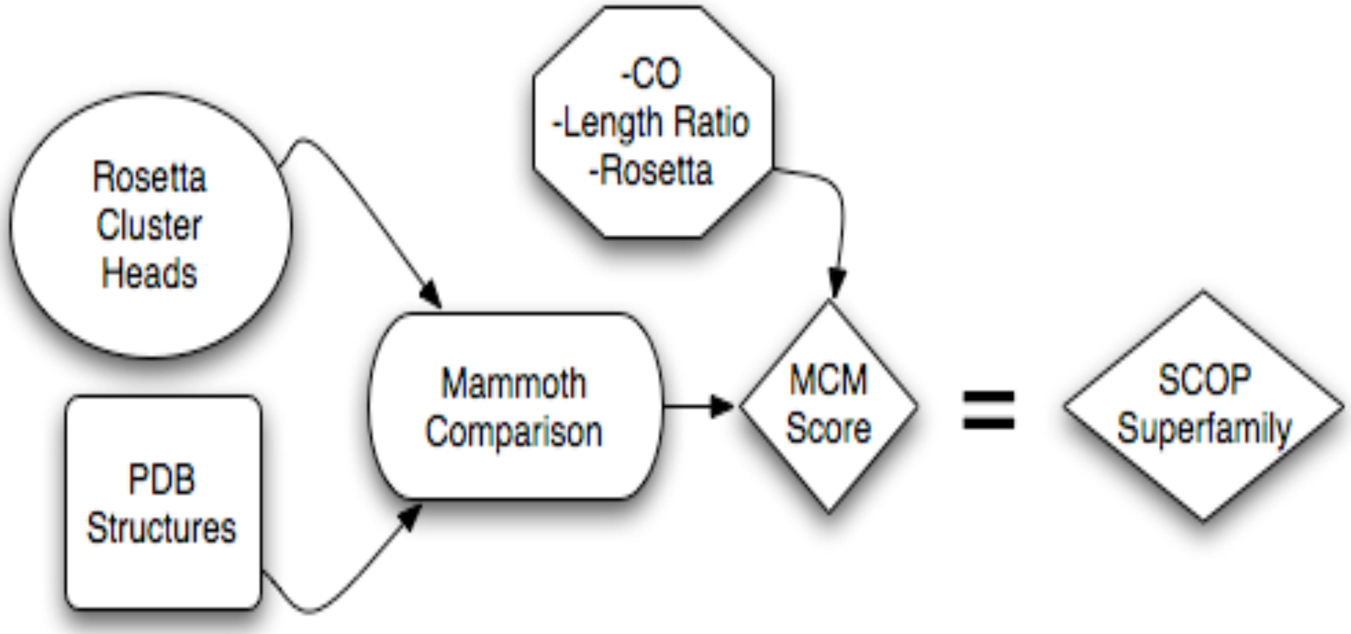


3. Naïve Bayes
   - Naïve Bayes with continuous SF prob
   - Naïve Bayes with GO terms



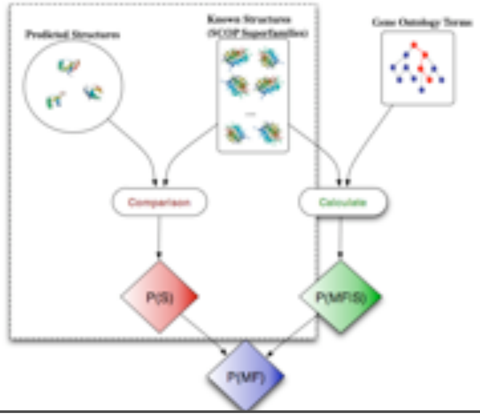
Mammoth Confidence Metric (MCM)

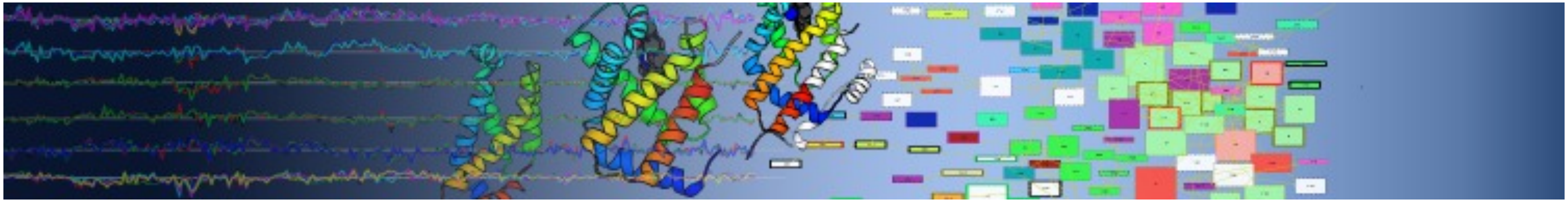
- Compare Cluster Representatives to PDB Structures
- MCM Score [0…1] probability
- based on:
  - Quality of match
  - Rosetta quality
  - Length ratio of PDB and cluster rep
  - Contact Order



$$\log\left(\frac{P_{MCM}}{1-P_{MCM}}\right) = a \cdot zscore + b \cdot CO + c \cdot converg + d \cdot \left|\log\left(\frac{L_{Astral}}{L_{predicted}}\right)\right| + C$$

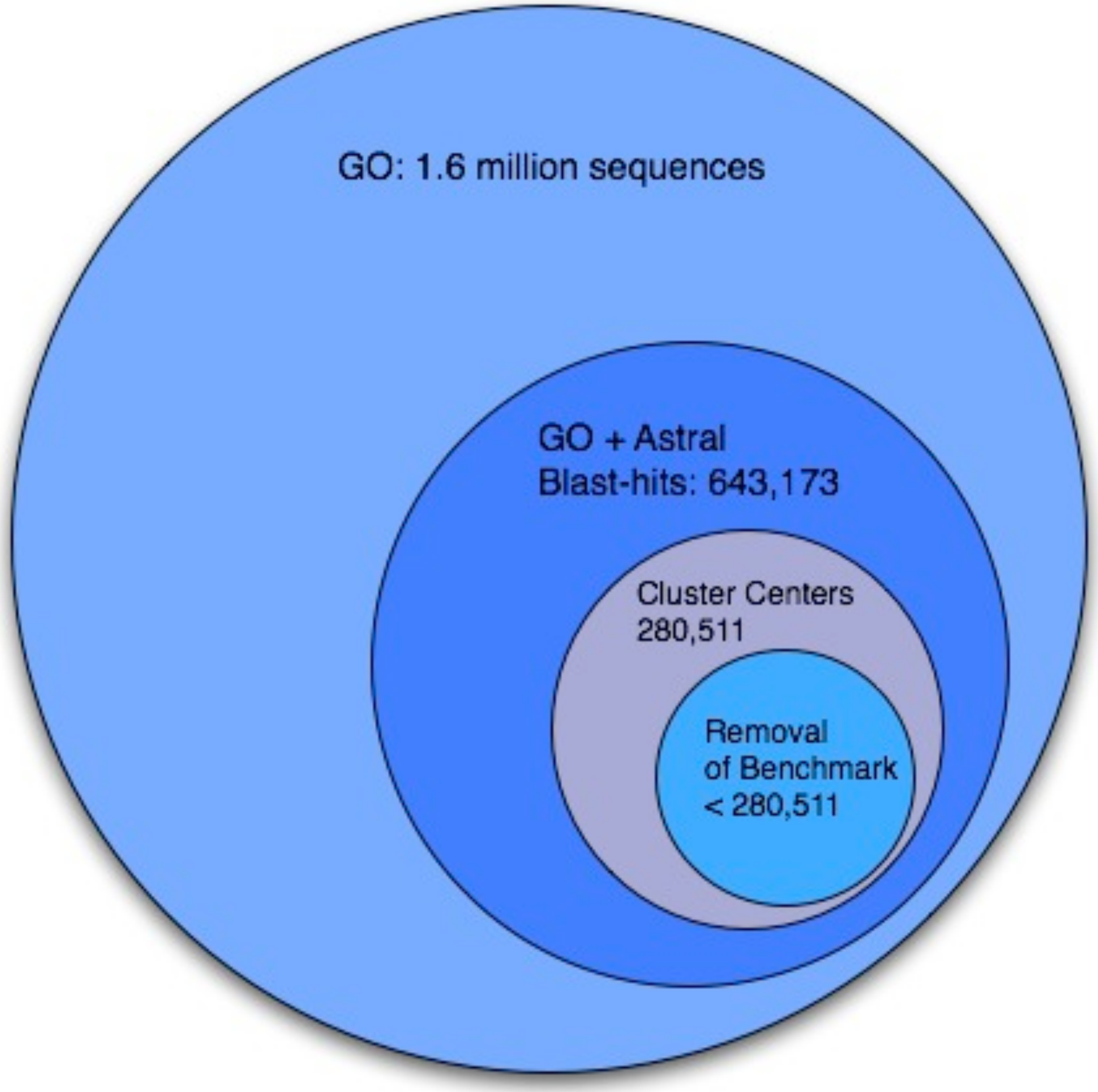
$$P(sf)_{mcm} = 1 - \prod_{k=1}^{n} (1 - p_k)$$

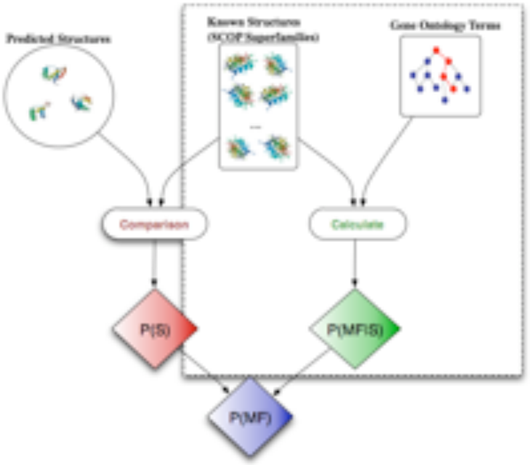


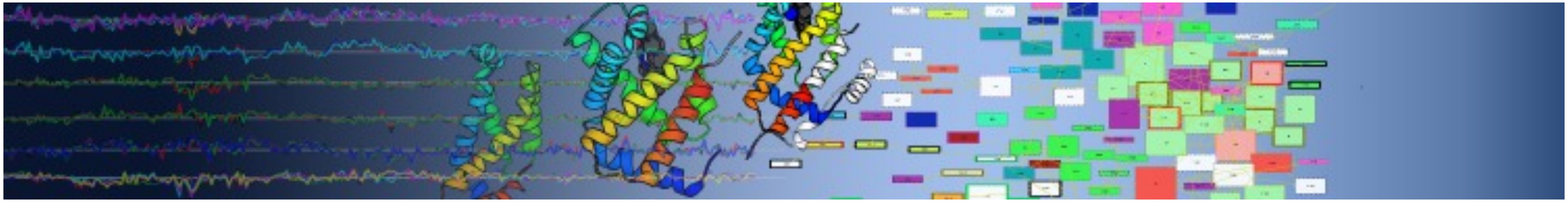


Gene Ontology (GO) & Training Data

- Function, Process, Localization terms
- 1.6 million sequences with annotations
- BLAST astral sequences to GO sequences (astral = pdb w/ SCOP SF)
- Cluster using CD-hit to reduce redundancy
- Cluster again using genome of benchmark sequences and remove matches



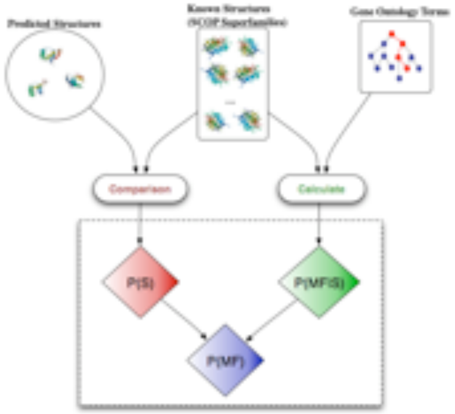


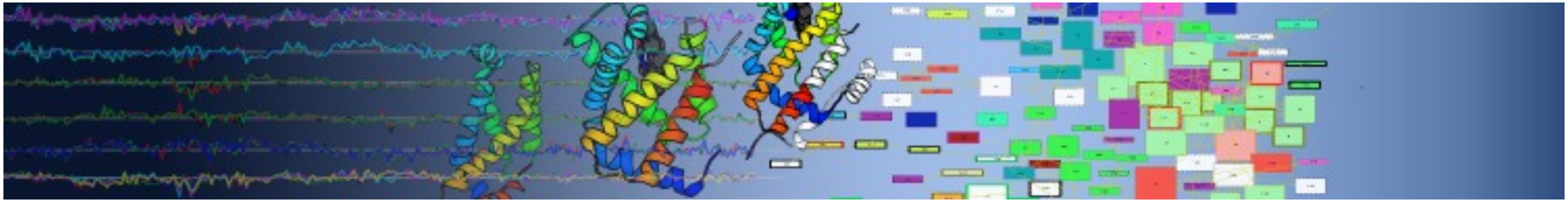


Naïve Bayes

y = molecular function and **x** = {sf, bp, cc}

$$LL_X = log\left(\frac{P(y = TRUE)}{P(y = FALSE)}\right) + \sum_{j=1}^{d} log\left(\frac{P(x_j|y = TRUE)}{P(x_j|y = FALSE)}\right)$$

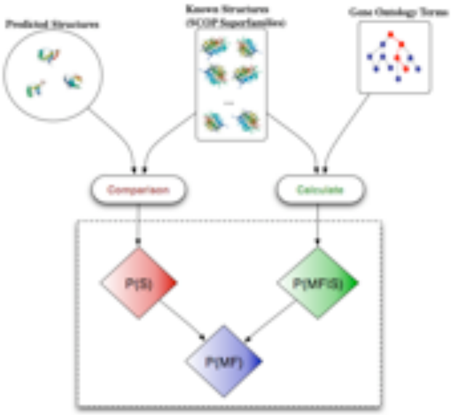


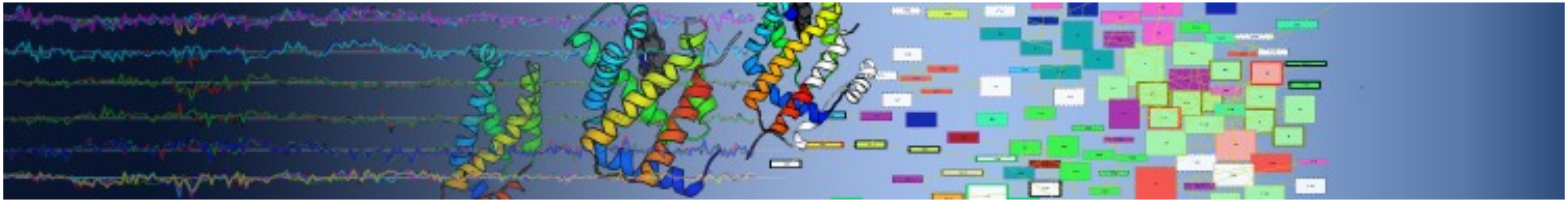


Naïve Bayes w/ Superfamilies

- How to take continuous probabilities of SF (by way of mcm scores)
  - we weight log-likelihood by the mcm scores:

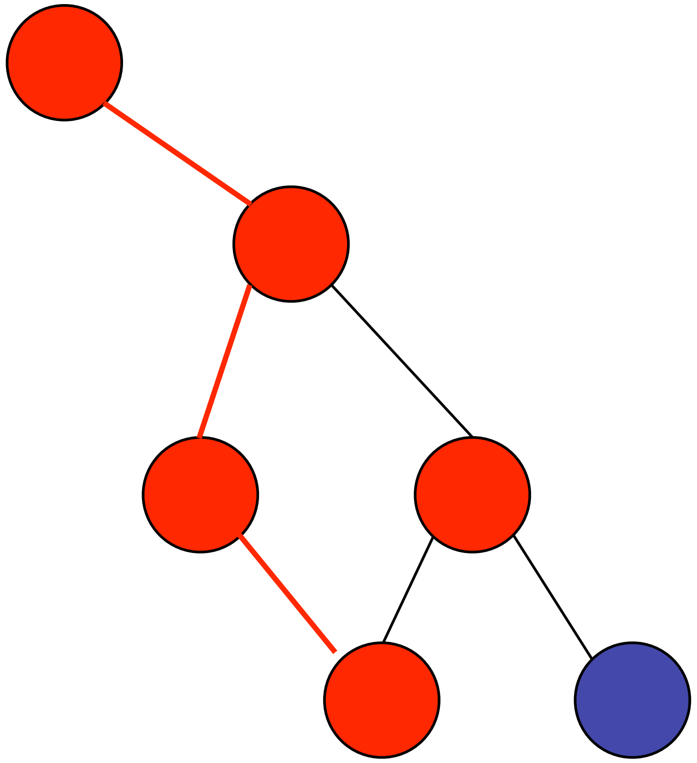
$$LL_{PLS} = log\left(\frac{P(MF)}{P(\bar{M}F)}\right) + \sum_{i=1}^{N}\left[P_{mcm}(sf_i) * log\left(\frac{P(sf_i|MF)}{P(sf_i|\bar{M}F)}\right)\right] + \sum_{j=P,L} log\left(\frac{P(x_j|MF)}{P(x_j|\bar{M}F)}\right)$$



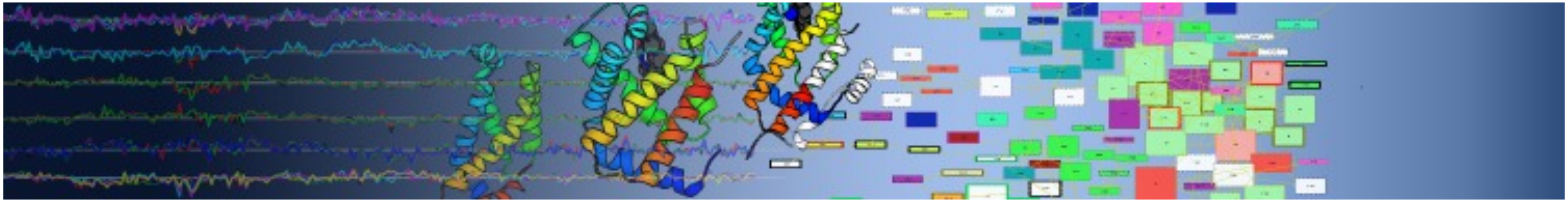


Naïve Bayes w/ GO terms

- **Problem**: Go terms are not independent
  - if we use all terms annotated to a sequence we end up double counting
- **Solution:** pick a term that will be predictive
  - Mutual information between term and MF

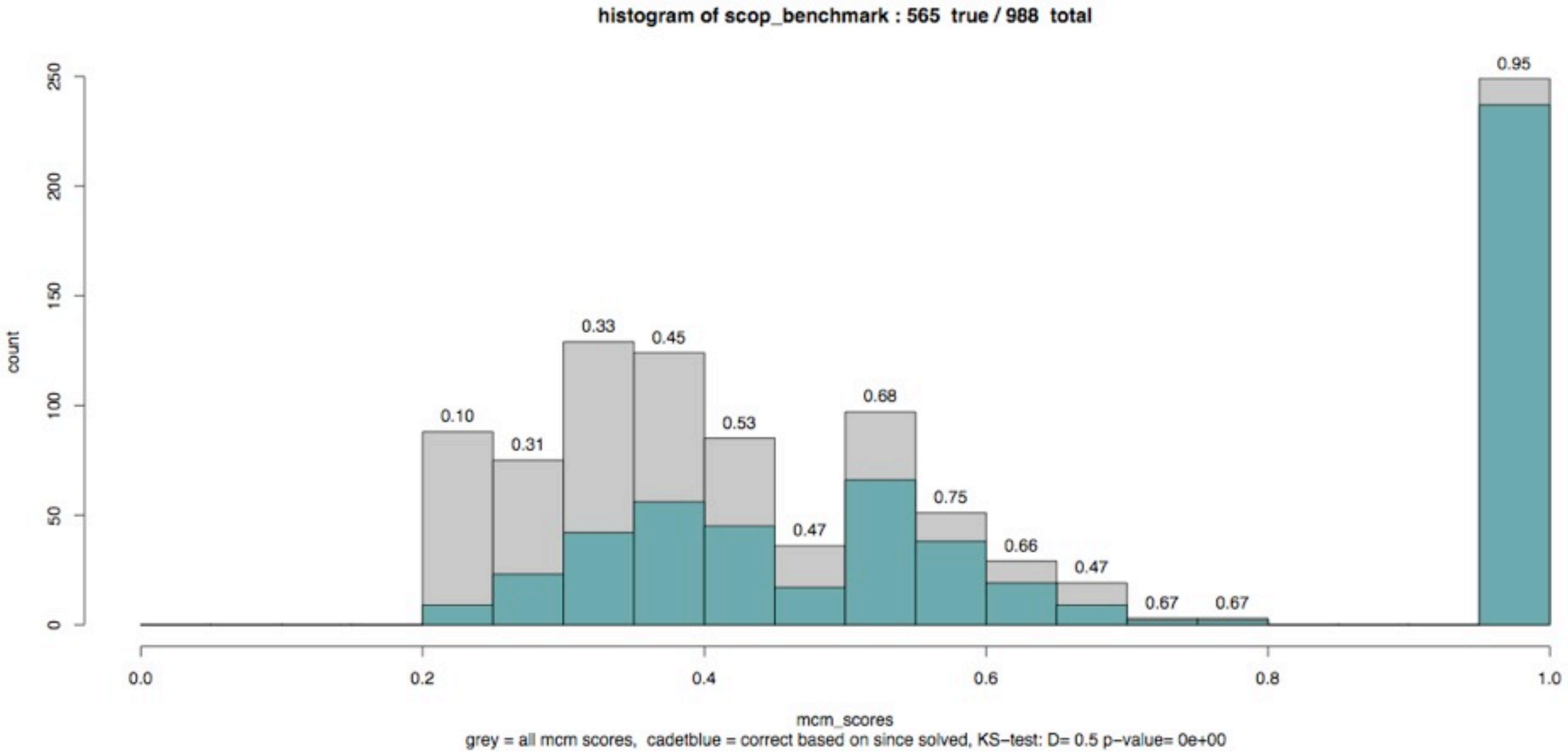


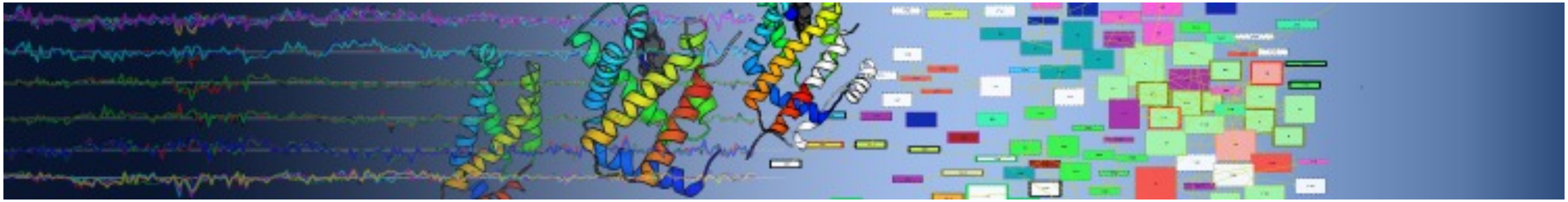
$$I(F;Y) = \sum_{f \in F} \sum_{y \in Y} P(F=f, Y=y) \log \frac{P(F=f, Y=y)}{P(F=f)P(Y=y)}.$$



Results: Solved Structures

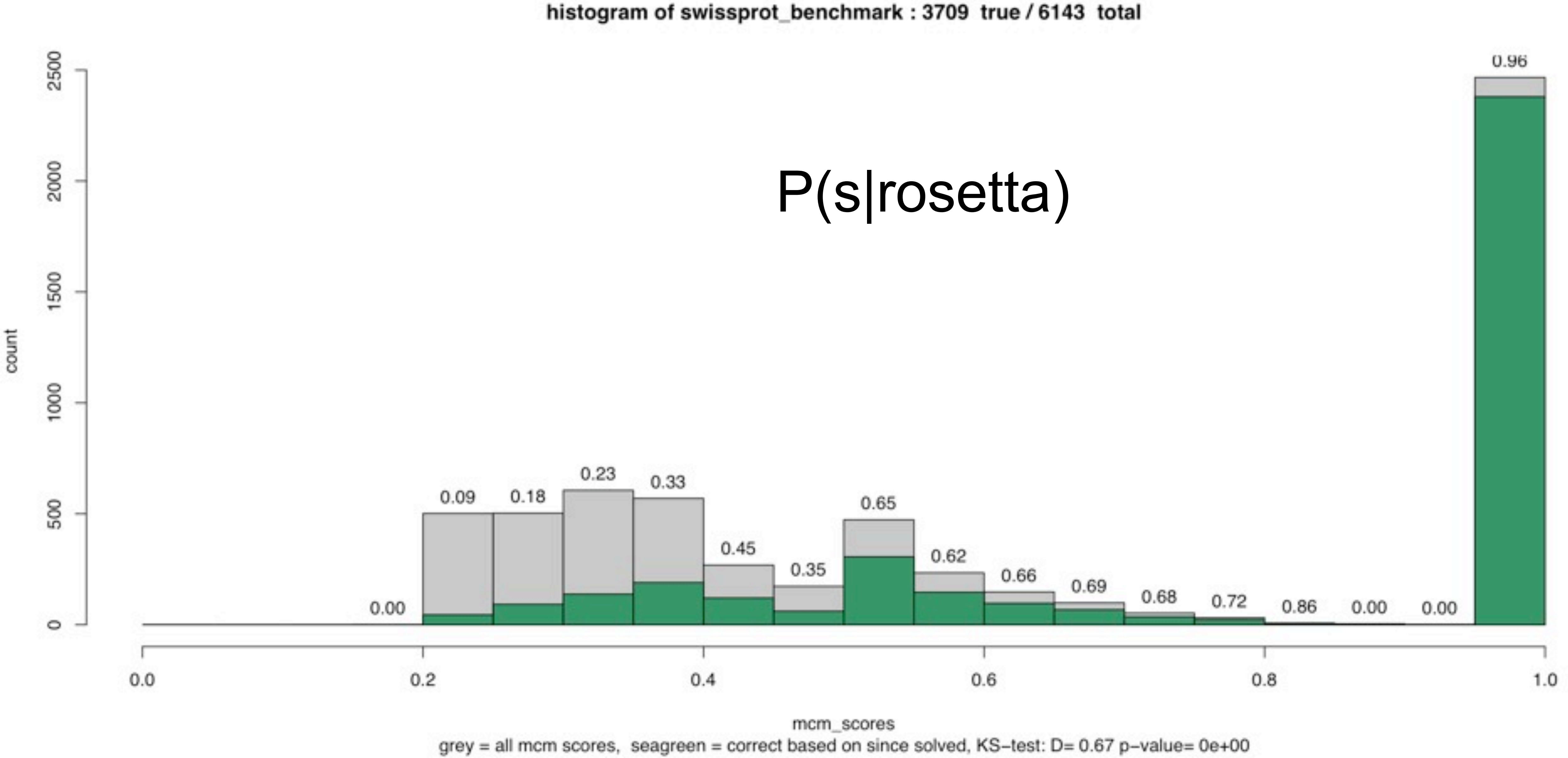
How accurate are we when we predict SCOP Superfamily for PDB Structures?



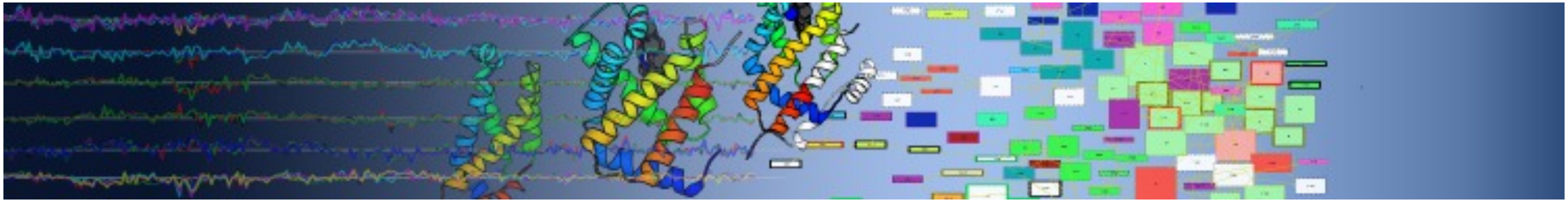


Results: Since Solved Structures (2005)

How accurate are we when we predict SCOP Superfamily for Swissprot Proteins?

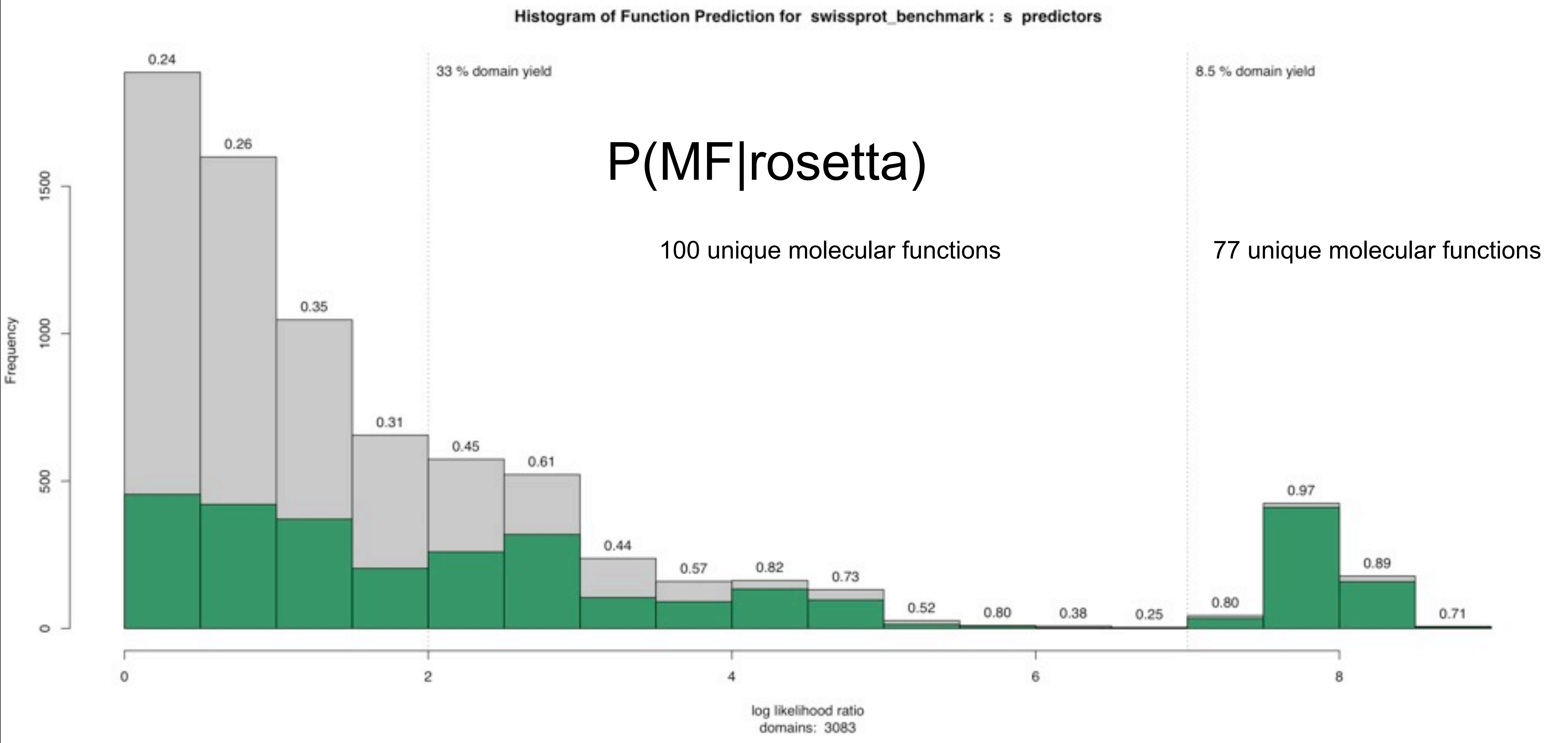


P(s|rosetta)



Results: Bayes Function Prediction (Swissprot Benchmark)

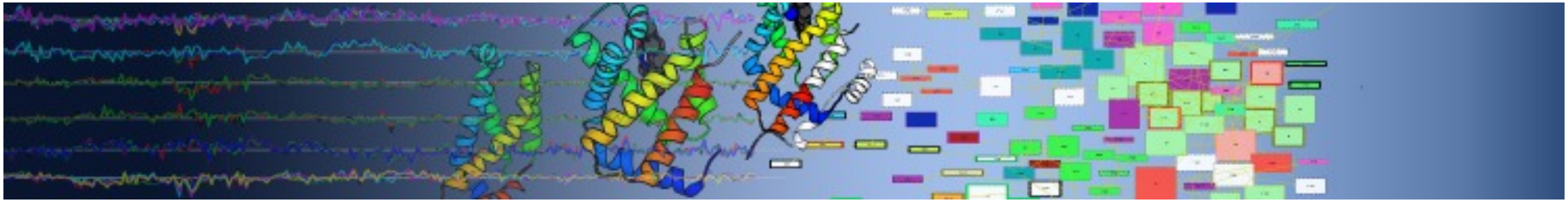
How accurate are our function predictions using structure only?



P(MF|rosetta)

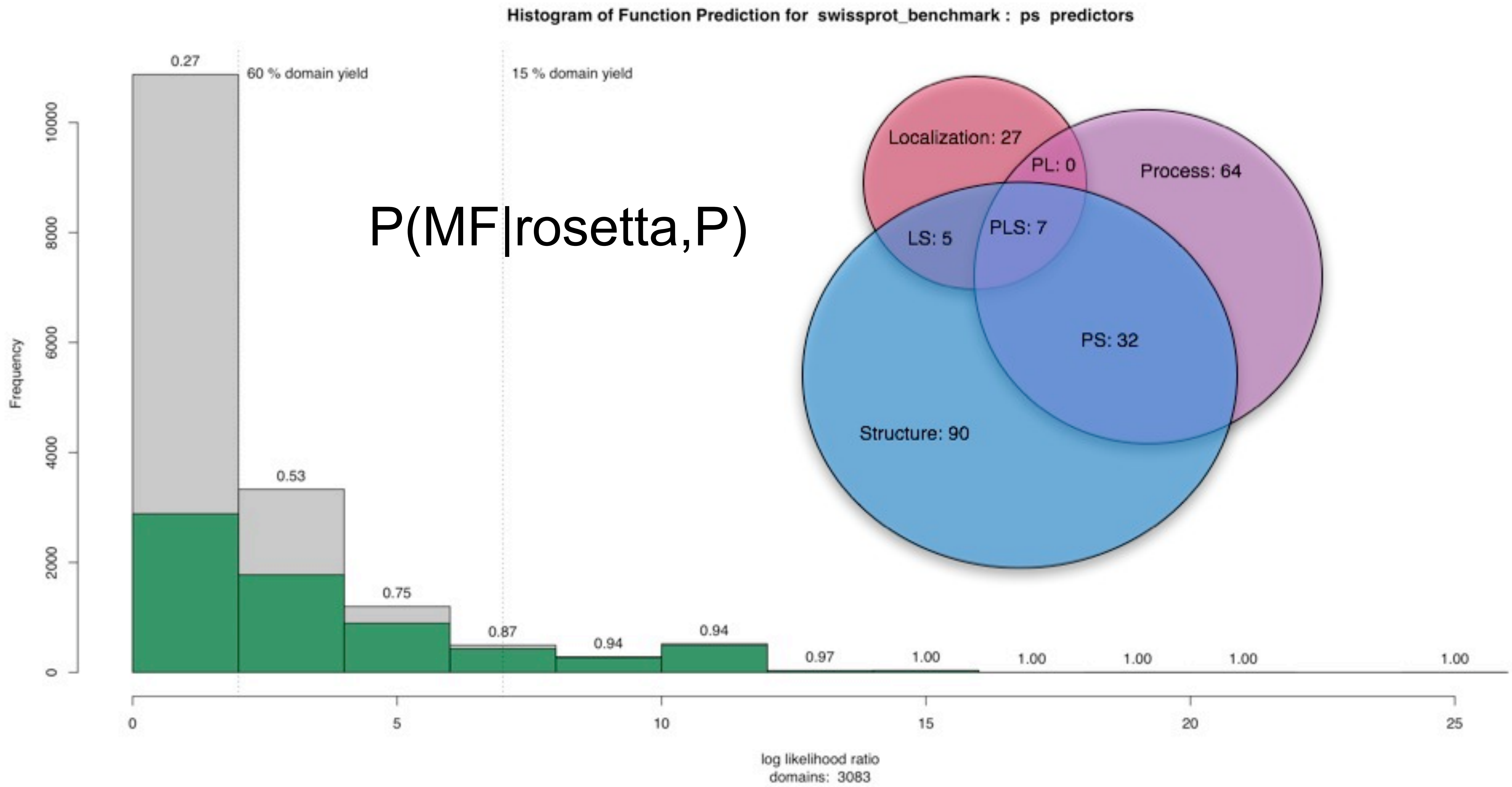
100 unique molecular functions

77 unique molecular functions



Results: Bayes Function Prediction (Swissprot Benchmark)

How accurate are our function predictions using GO process & structure?

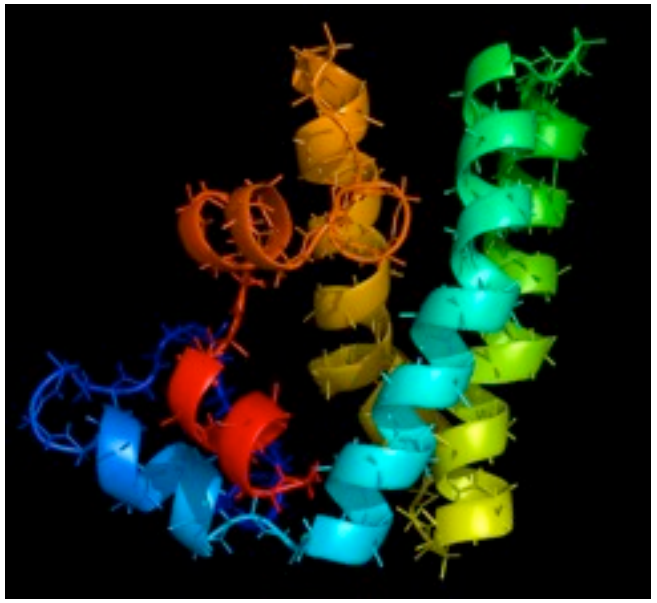


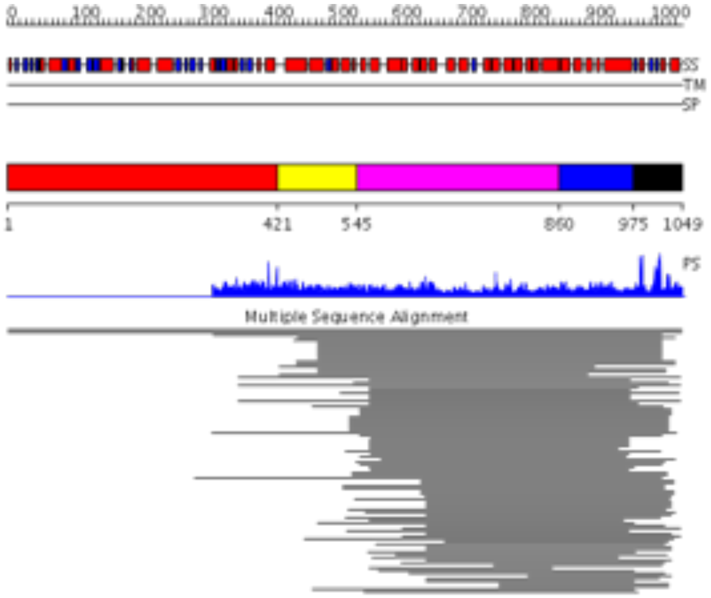
P(MF|rosetta,P)

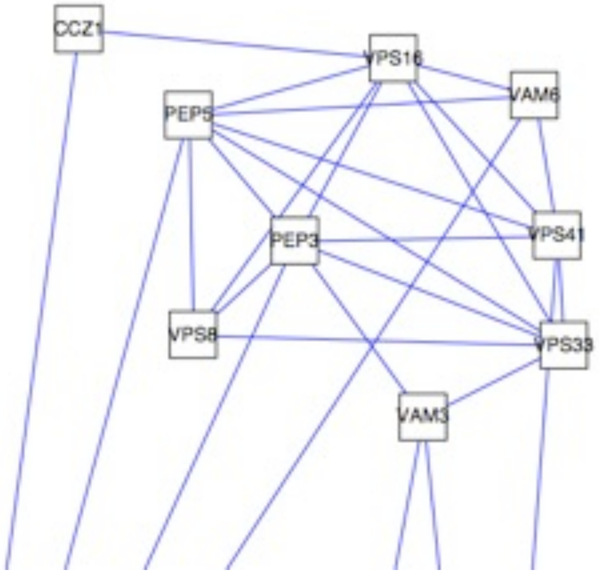
Results: HPF:

vesicle transport

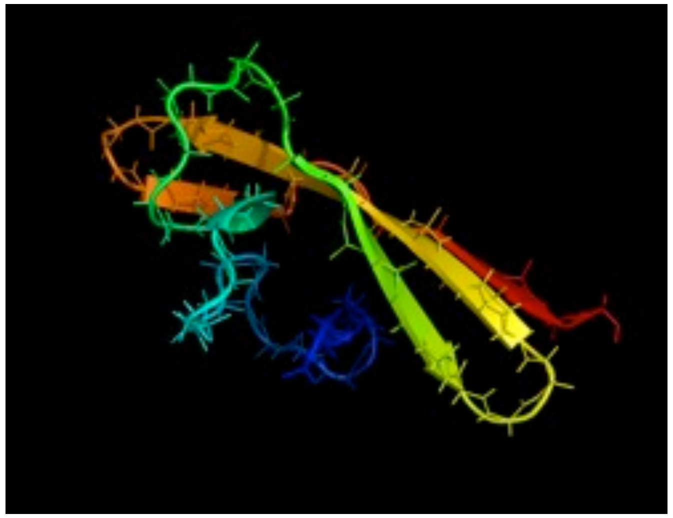
VAM6/ YDL077C: Vacuolar protein that plays a critical role in the tethering steps of vacuolar membrane fusion by facilitating guanine nucleotide exchange on small guanosine triphosphatase Ypt7p.We find the following: (domain1) unknown (domain 2) Rosetta hit to Polynucleotide phosphorylase/guanosine pentaphosphate synthase (PNPase/GPSI), domain 3 (domain 3) Clathrin proximal leg (domain 4) Rosetta de novo hit to Hemerythrin (domain 5) Rosetta hit to SAM/ Pointed domain.

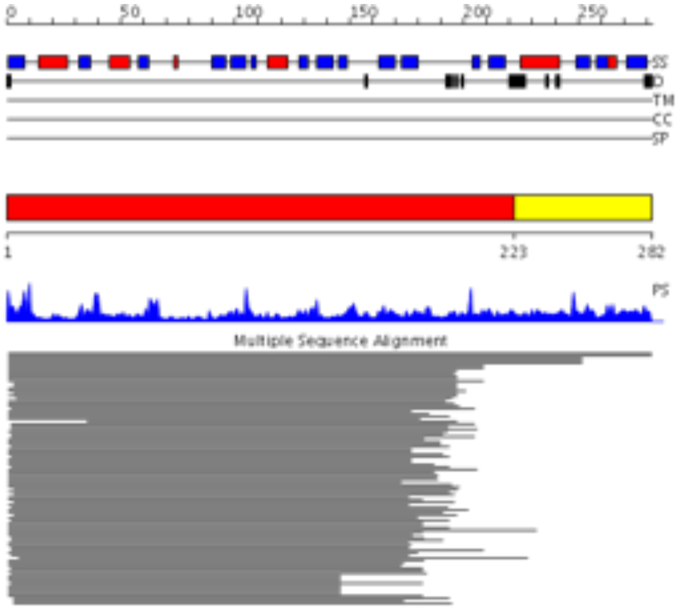






VPS29: Endosomal protein that is a subunit of the membrane-associated retromer complex essential for endosome-to-Golgi retrograde transport; forms a subcomplex with Vps35p and Vps26p that selects cargo proteins for endosome-to-Golgi retrieval. But, with this context so well defined, there is still no molecular function known, that is to say there is no precise mechanistic role known for this protein. We find a strong hit to Mre11 (a double stranded mismatch repair protein, metal dependent phosphotase for domain 1 and a strong Rosetta hit for domain 2 to the PUA-domain like fold (implicated in RNA binding OR ATP sulfurylase N-terminal domain). The fold predictions are as confident as we ever see (MCM = 0.95, psiblast evalue to domain 1 hit Z = 13.).




0
50
100
150
200
250
SS
D
TM
CC
SP
1
223
282
PS
Multiple Sequence Alignment