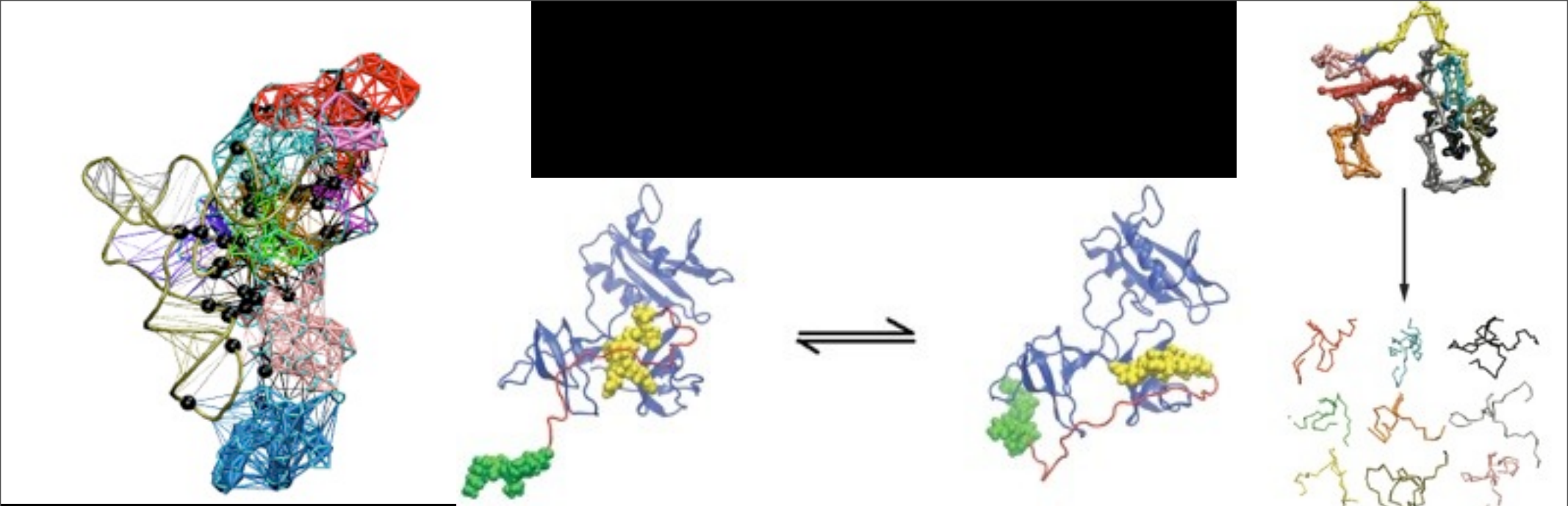
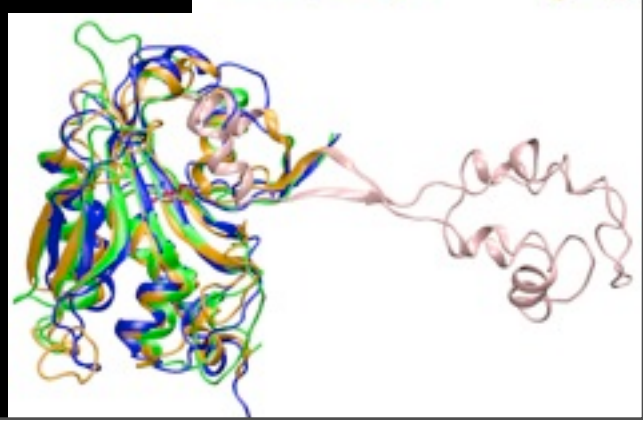


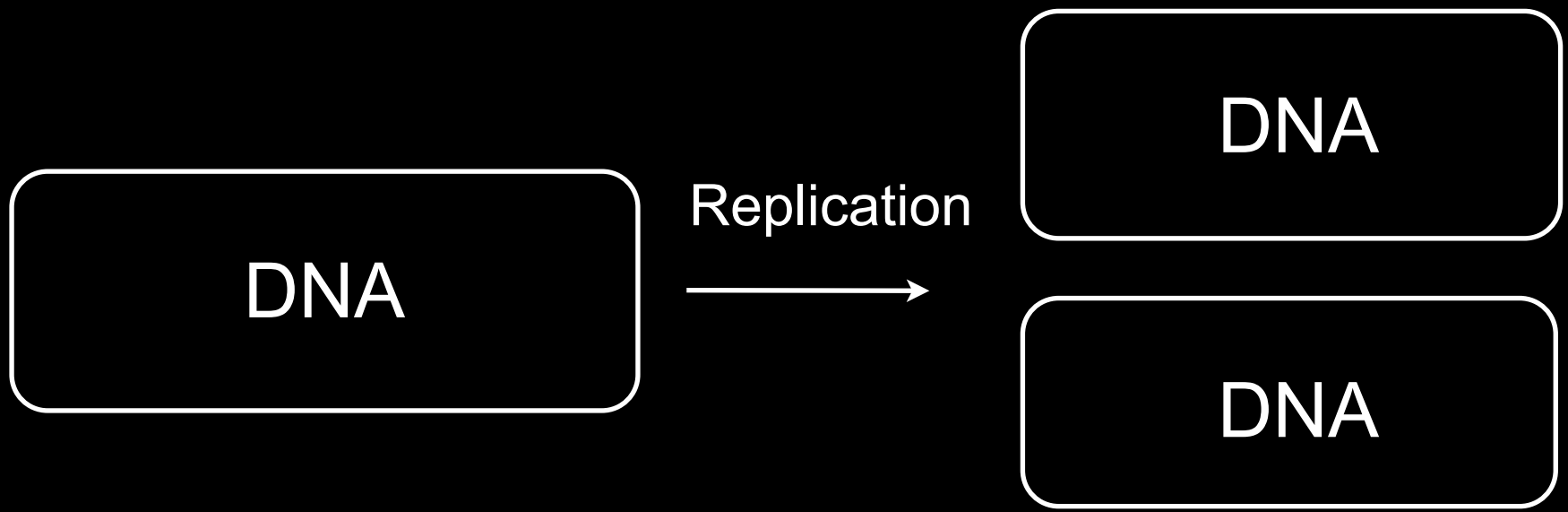
Modeling Biological Information Processing Pathways

Anurag Sethi
LANL
Yale University
June 2013

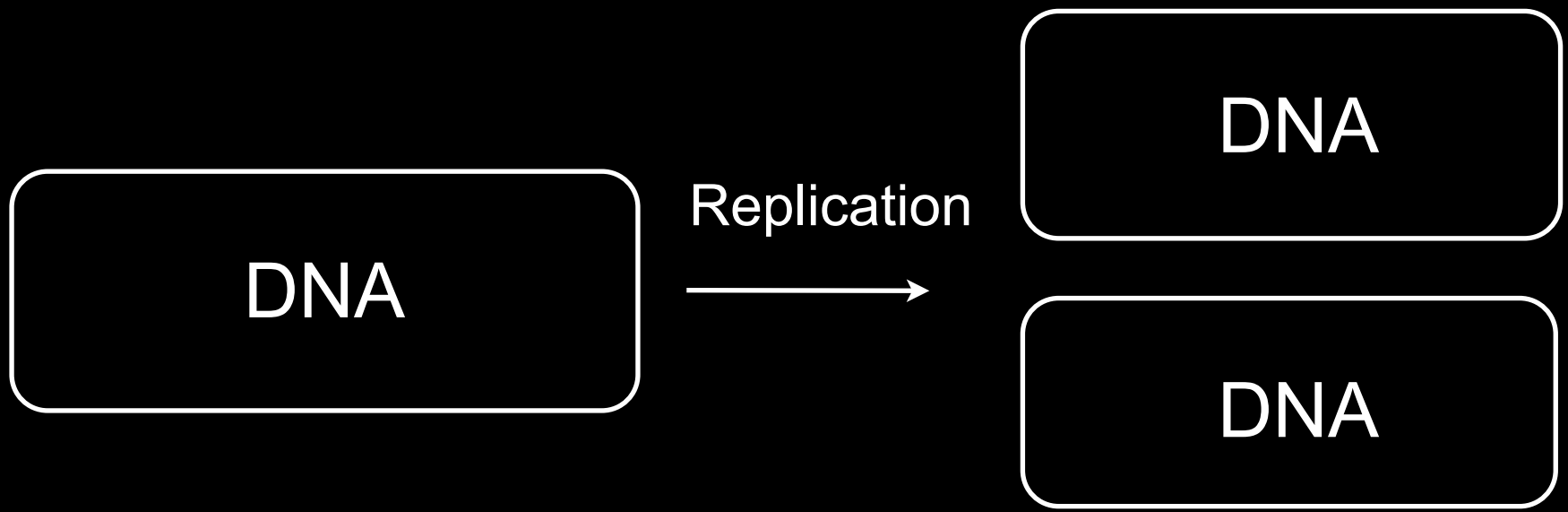


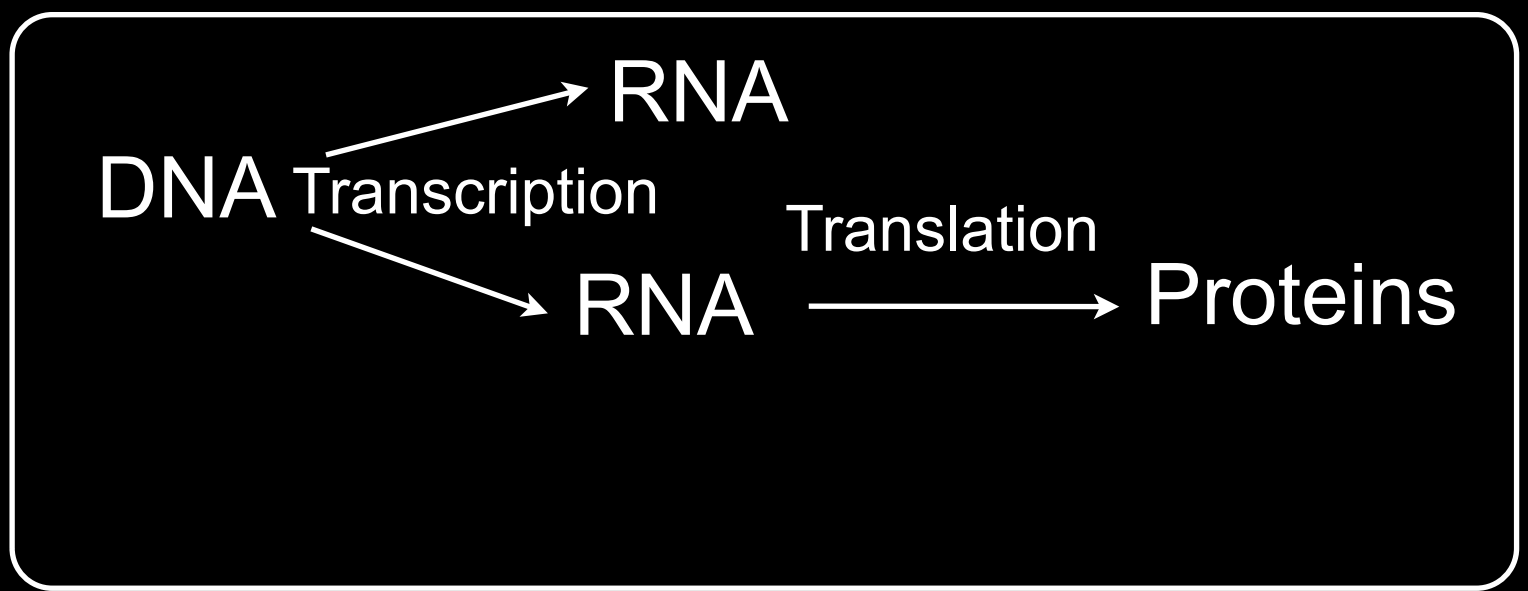
Life is Information Processing

Life is Information Processing



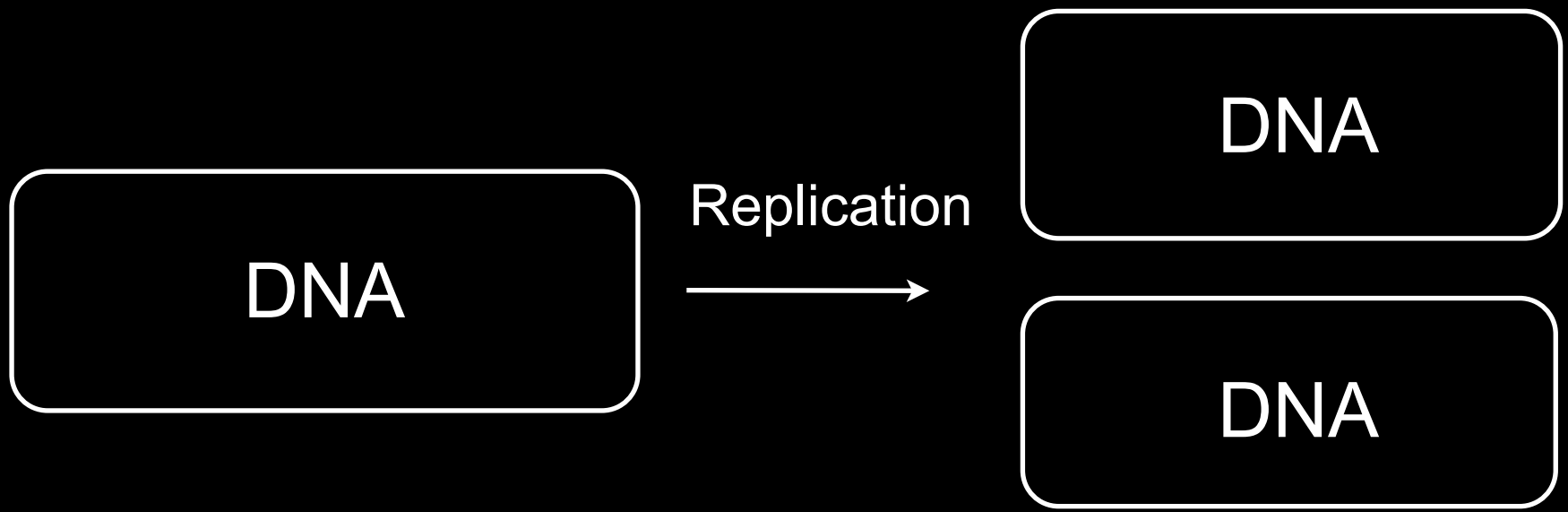
Life is Information Processing

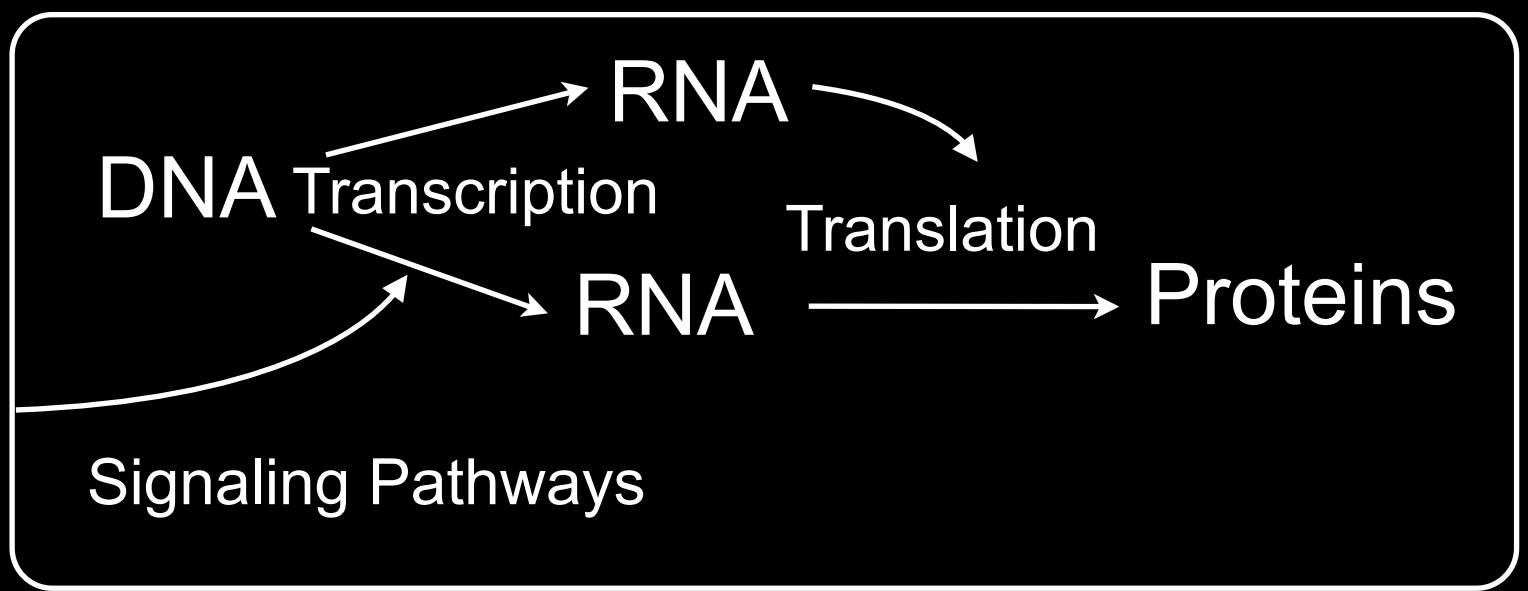




Genes are translated into proteins and proteins perform functions subject to cues from the environment.

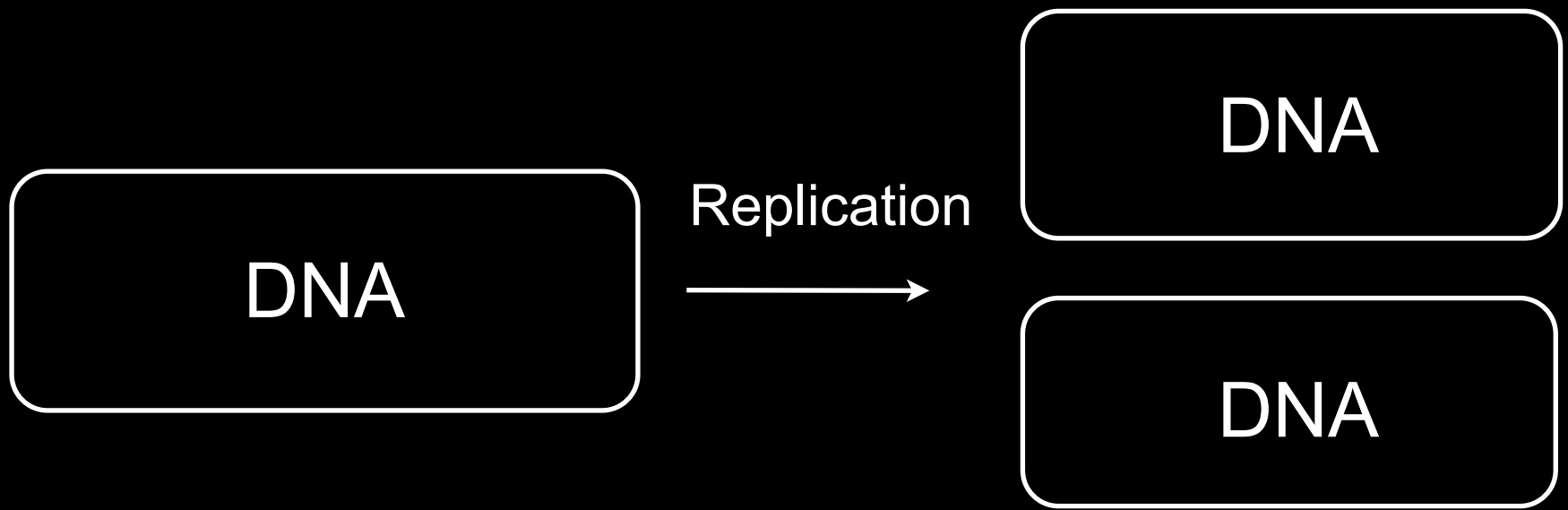
Life is Information Processing

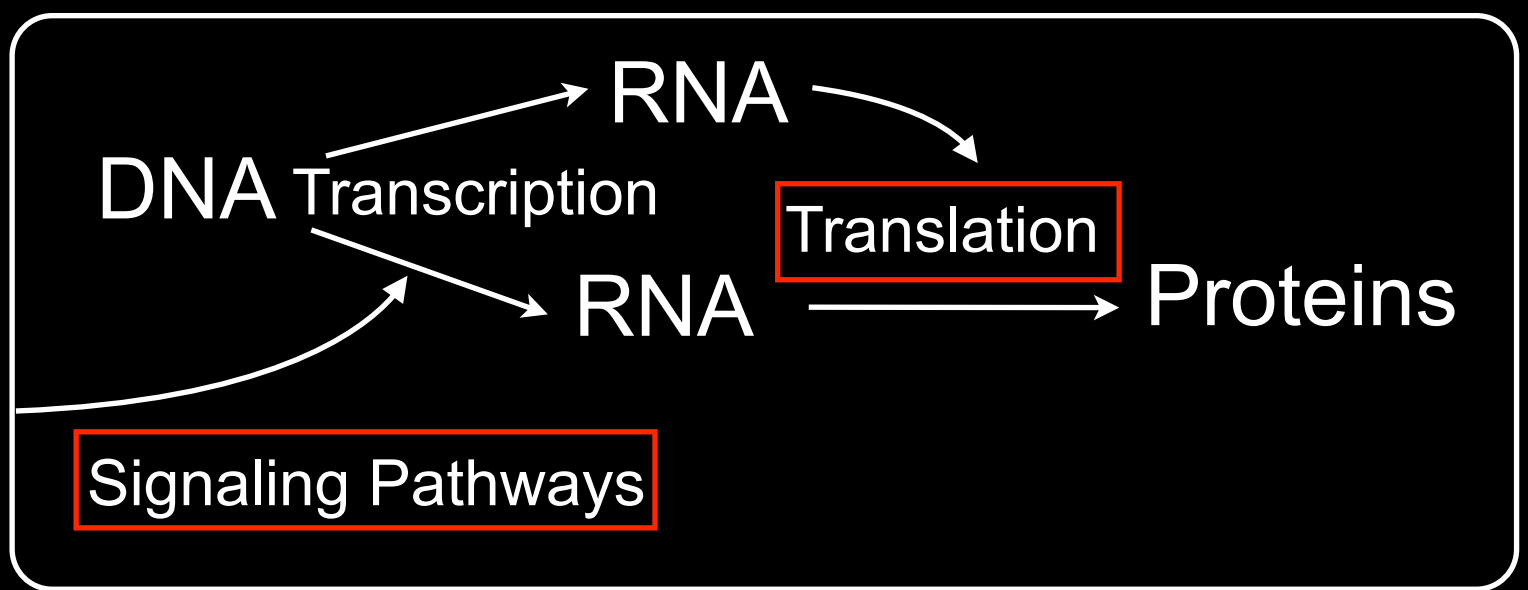




Genes are translated into proteins and proteins perform functions subject to cues from the environment.

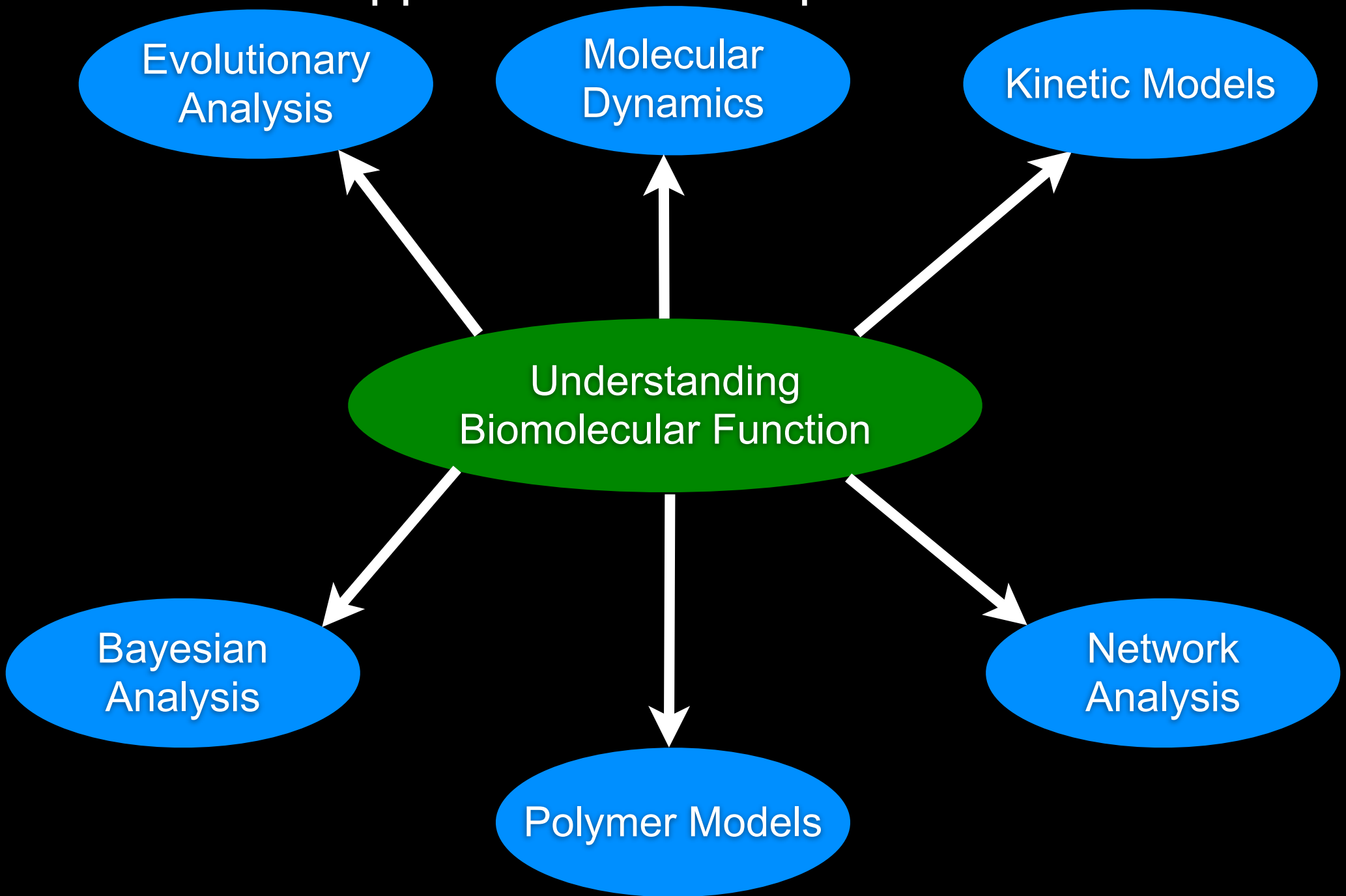
Life is Information Processing





Genes are translated into proteins and proteins perform functions subject to cues from the environment.

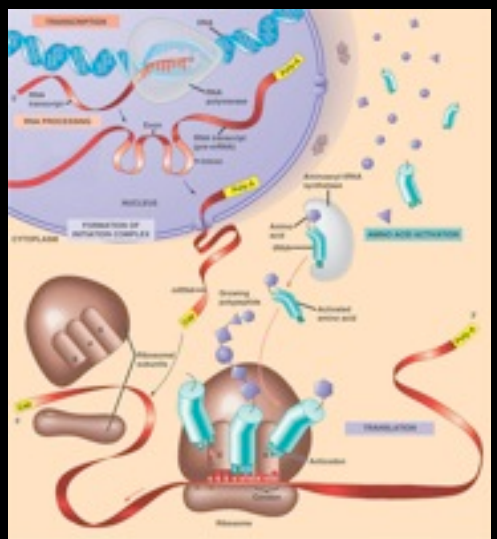
Investigation of complex systems requires the application of multiple tools



Organization of Talk

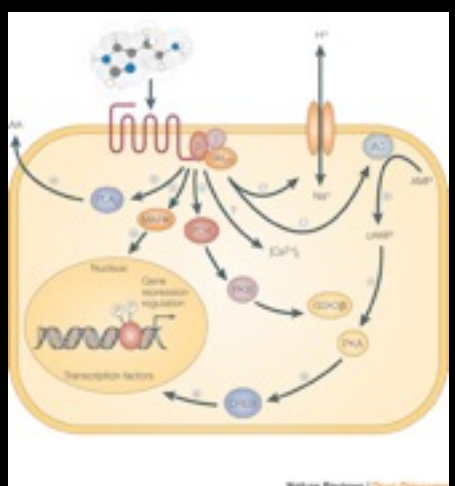
Translating Information in RNA

- Evolutionary analysis of biomolecules
- Allosteric signaling pathways



Disordered Regions in Signaling Cascades

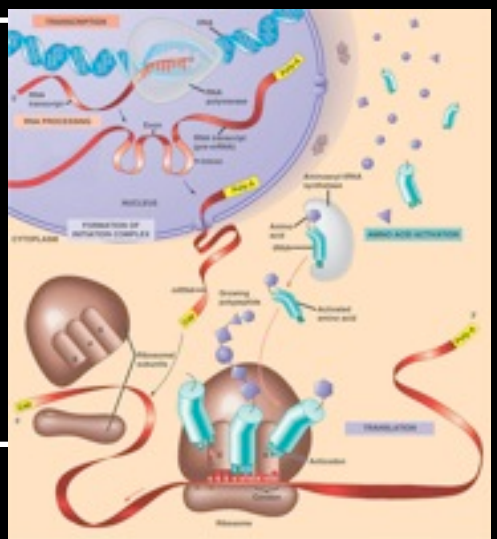
- Multivalent Proteins
- Modeling Disordered Proteins



Organization of Talk

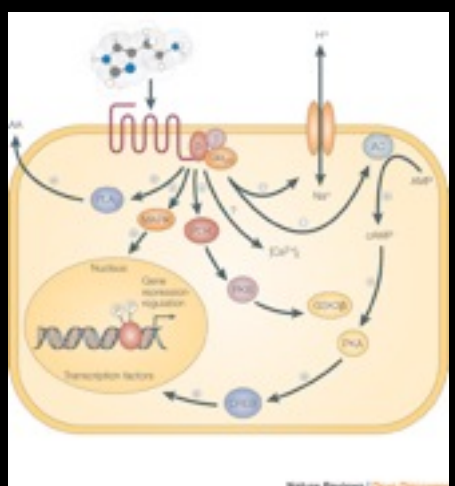
Translating Information in RNA

- Evolutionary analysis of biomolecules
- Allosteric signaling pathways



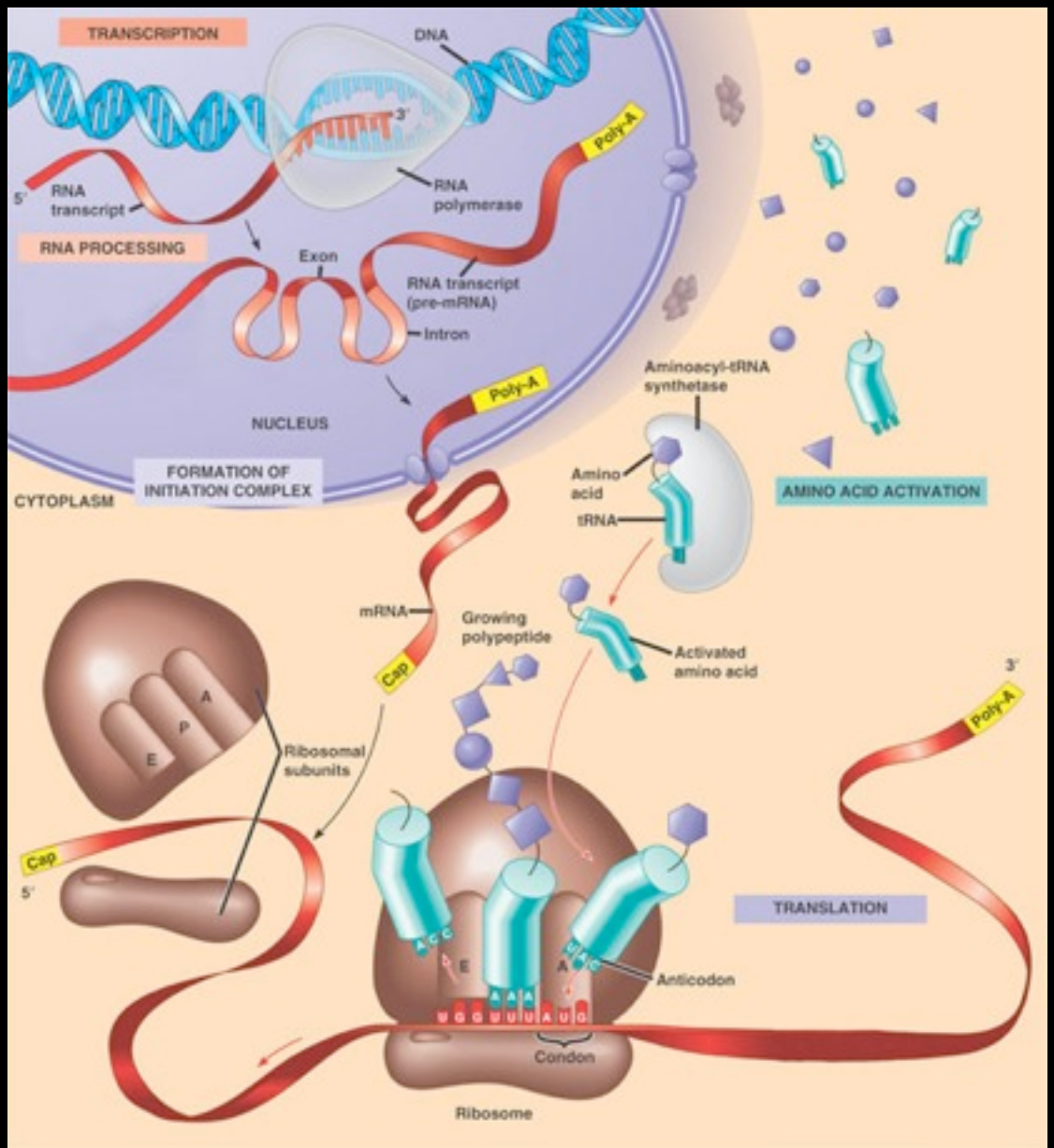
Disordered Regions in Signaling Cascades

- Multivalent Proteins
- Modeling Disordered Proteins



Setting the Genetic Code: Evolutionary Analysis

The central dogma of molecular biology describes how DNA is translated into proteins

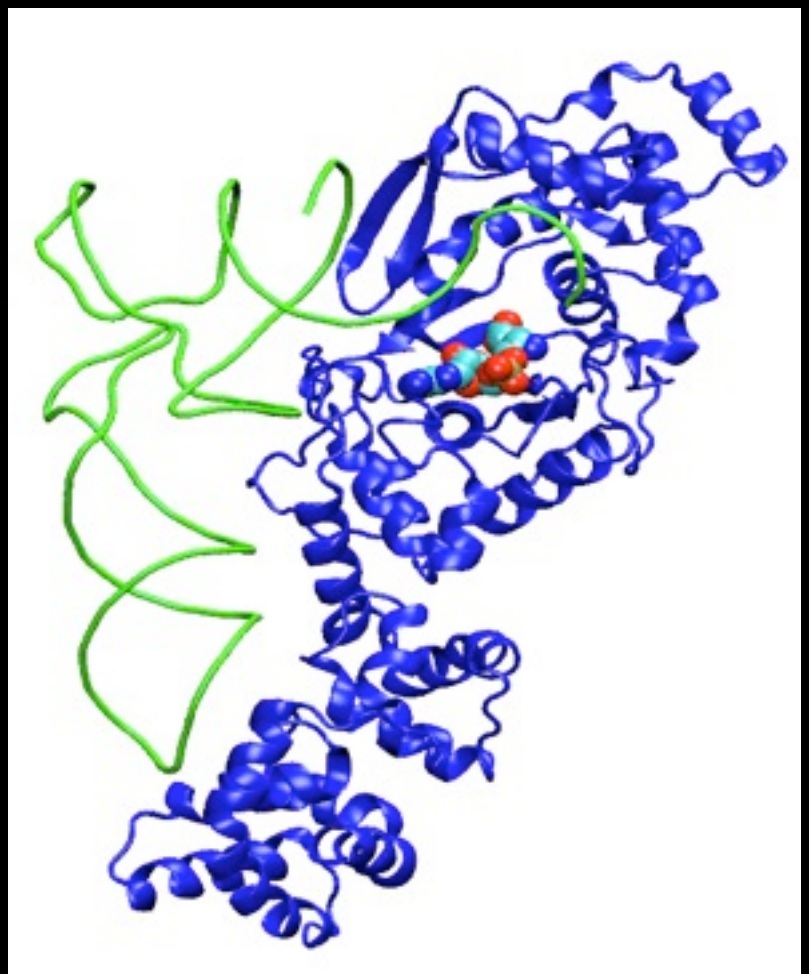


The universal genetic code is used to translate the information in the DNA into functional proteins.

http://compbio.pbworks.com/w/page/16252897/Introduction%20and%20Basic%20Molecular%20Biology

Setting the Genetic Code: Evolutionary Analysis

The genetic code is set by enzymes called the aminoacyl-tRNA synthetases



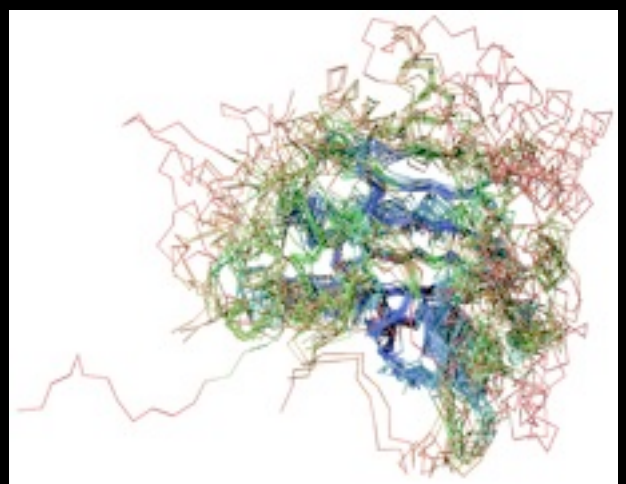
First Letter	Second Letter: T	Second Letter: C	Second Letter: A	Second Letter: G	Third Letter
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA Stop	TGA Stop	A
	TTG Leu	TCG Ser	TAG Stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAC His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

How do complex biomolecules perform their function?

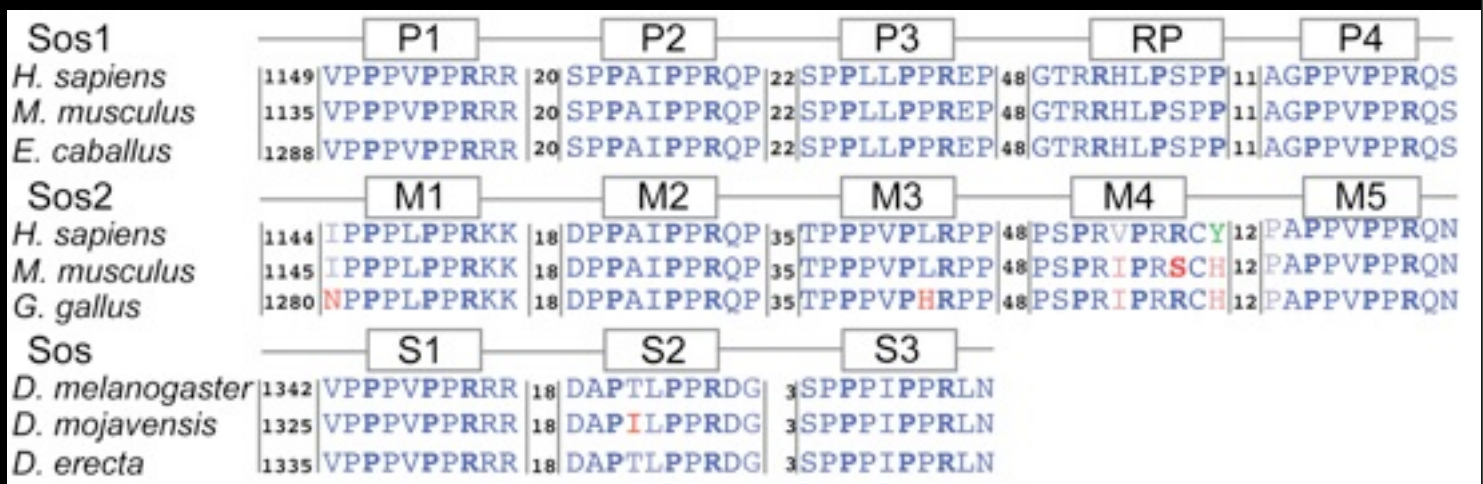
http://plato.stanford.edu/entries/information-biological/

Goal of Evolutionary Analysis

Nothing in biology makes sense except in the light of evolution. - Theodosius Dobzhansky



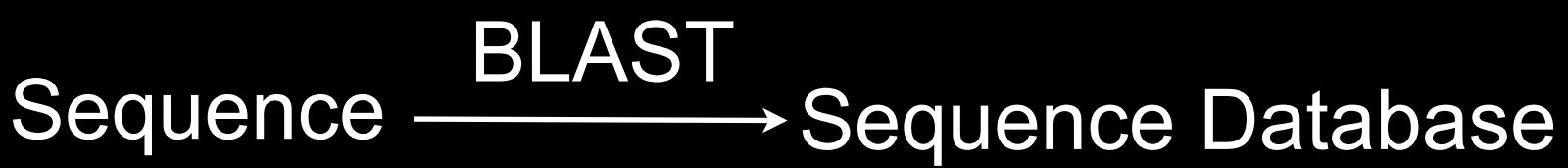
Structural alignment



Sequence alignment

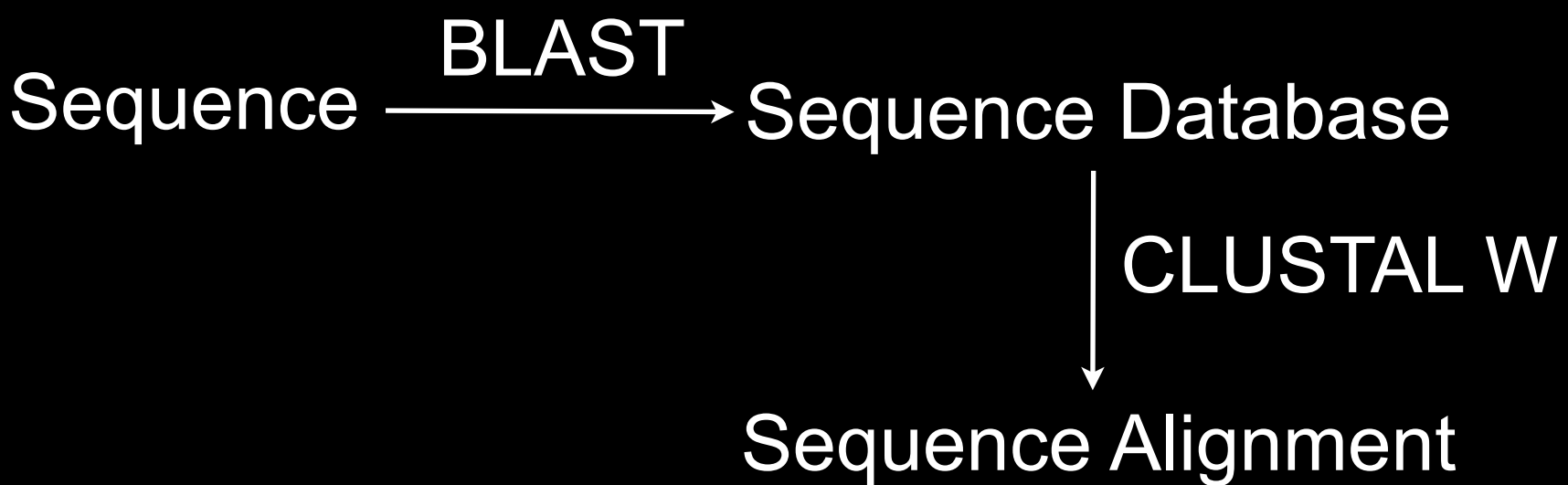
Conservation analysis provides information on the constraints in the evolution of biomolecular families.

Evolutionary Analysis



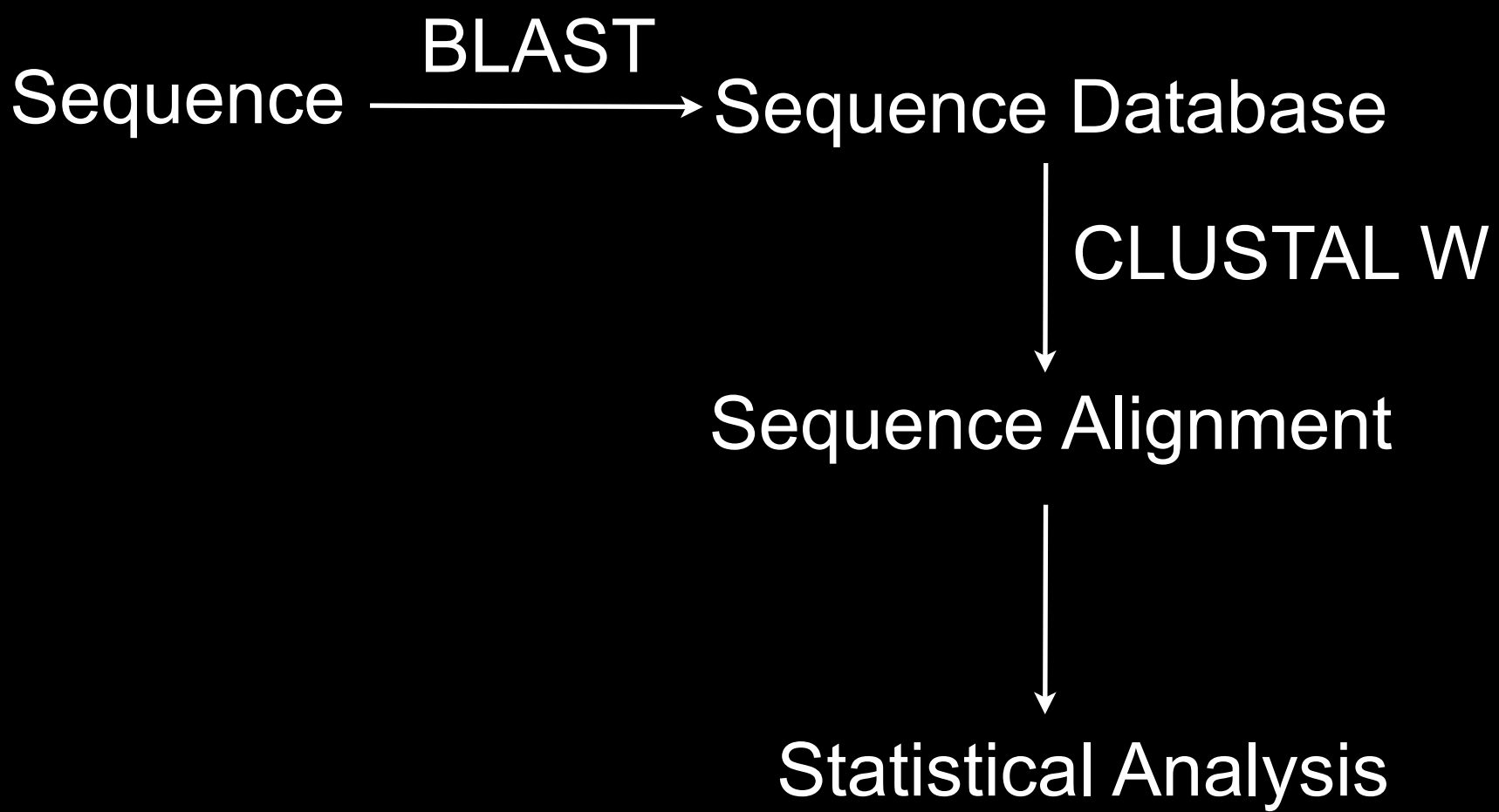
The sequences used to represent the evolutionary history of a biomolecule are the sequences found in a database.

Evolutionary Analysis



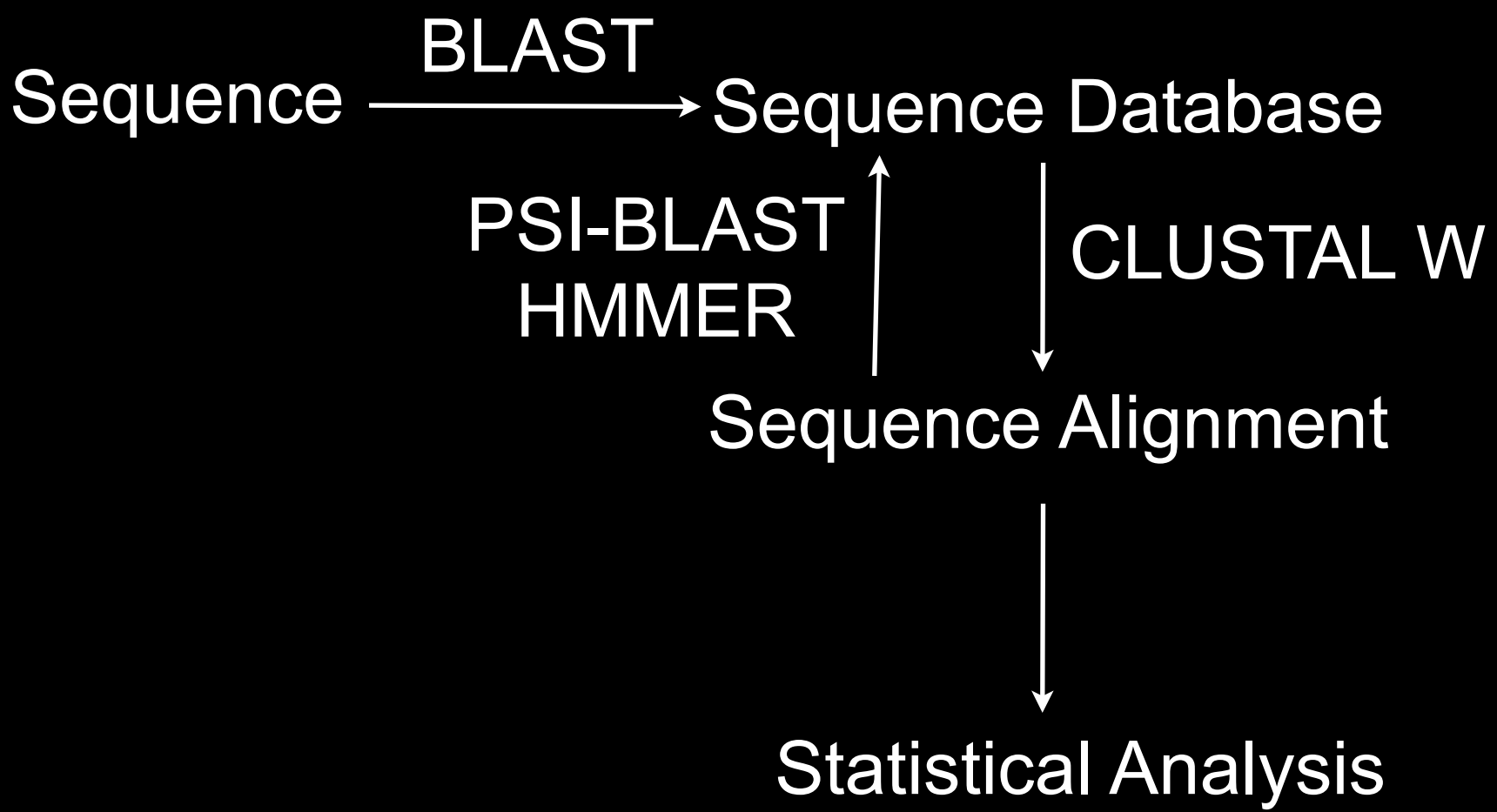
The sequences used to represent the evolutionary history of a biomolecule are the sequences found in a database.

Evolutionary Analysis



The sequences used to represent the evolutionary history of a biomolecule are the sequences found in a database.

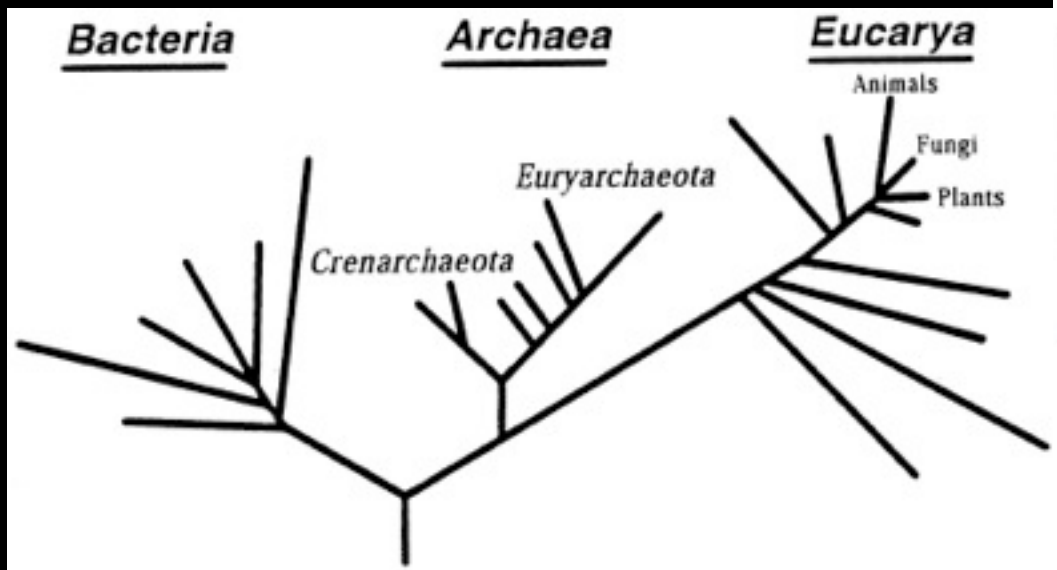
Evolutionary Analysis



The sequences used to represent the evolutionary history of a biomolecule are the sequences found in a database.

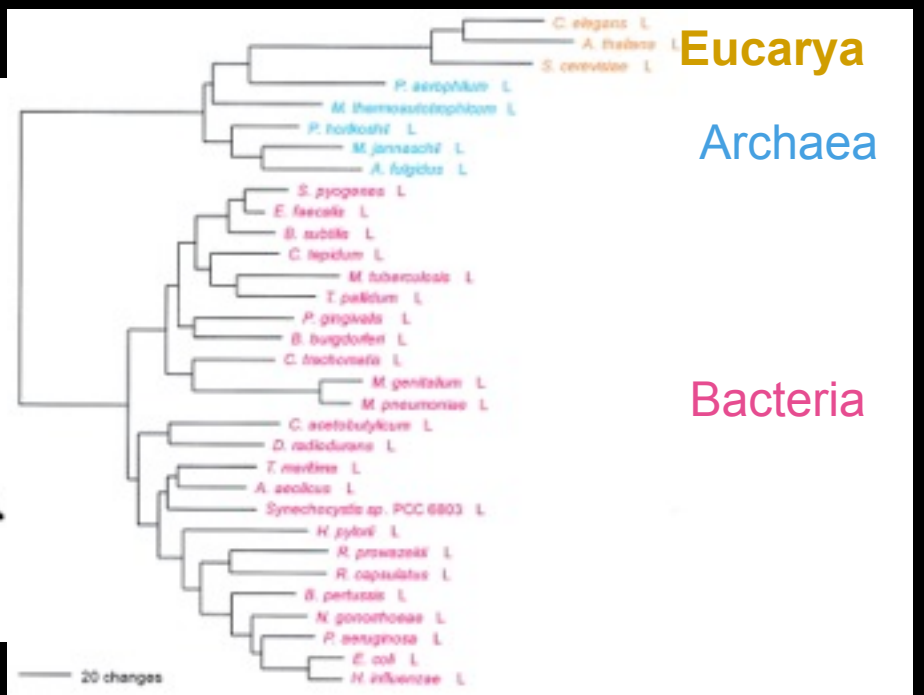
Setting the Genetic Code: Evolutionary Analysis

The sequence and structure databases are biased towards bacterial domain of life



Universal Phylogenetic Tree
Three domains of life

Based on rRNA



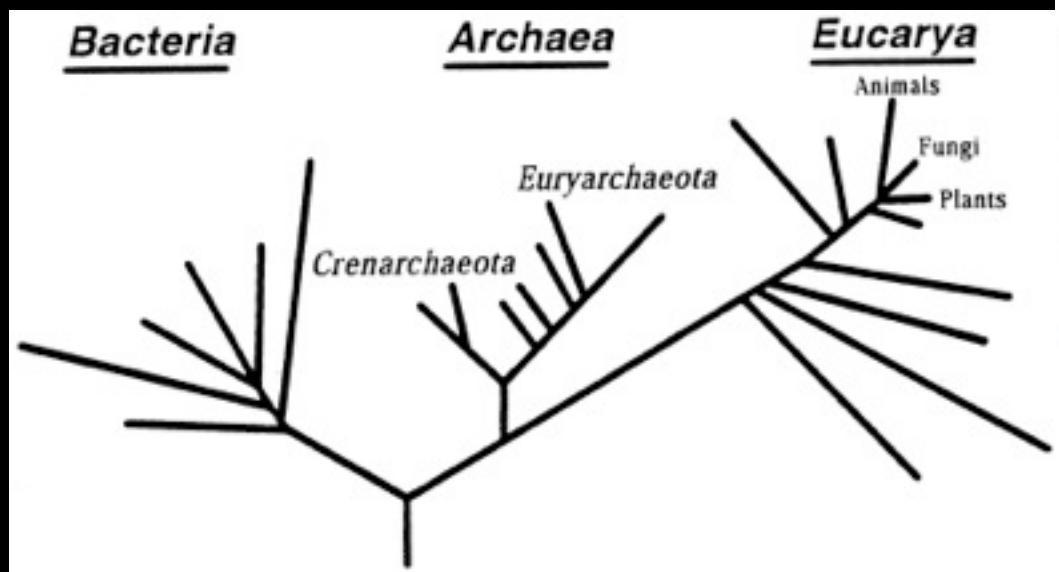
Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

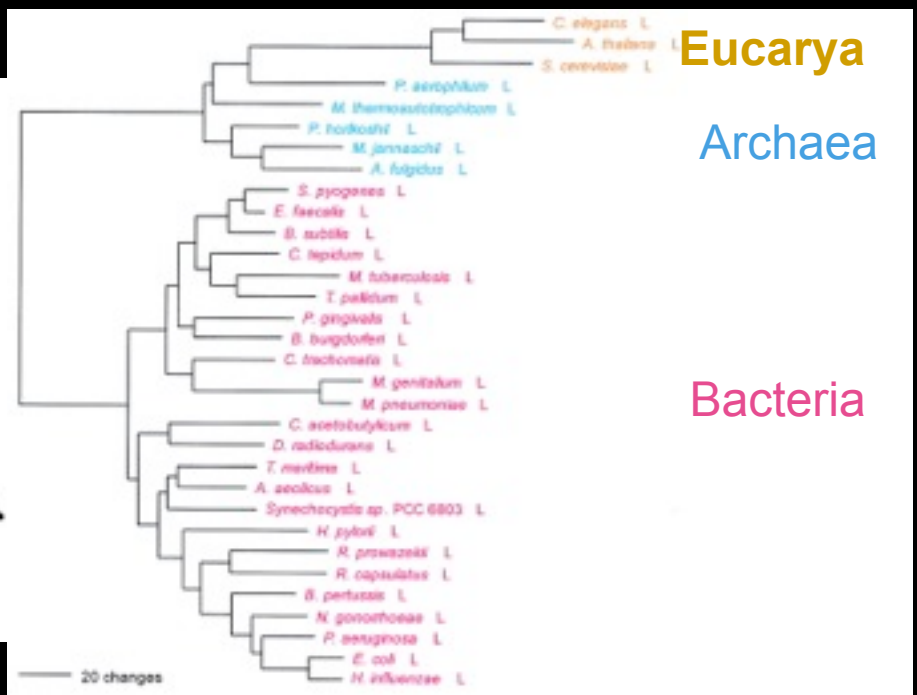
for review see Woese *PNAS* 2000

Woese, et al., *MMBR* 2000.

Setting the Genetic Code: Evolutionary Analysis

The sequence and structure databases are biased towards bacterial domain of life





Universal Phylogenetic Tree
Three domains of life

Based on rRNA

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

The databases are biased and statistical analysis of sequence and structure profiles implement ad-hoc sequence weighting methods.

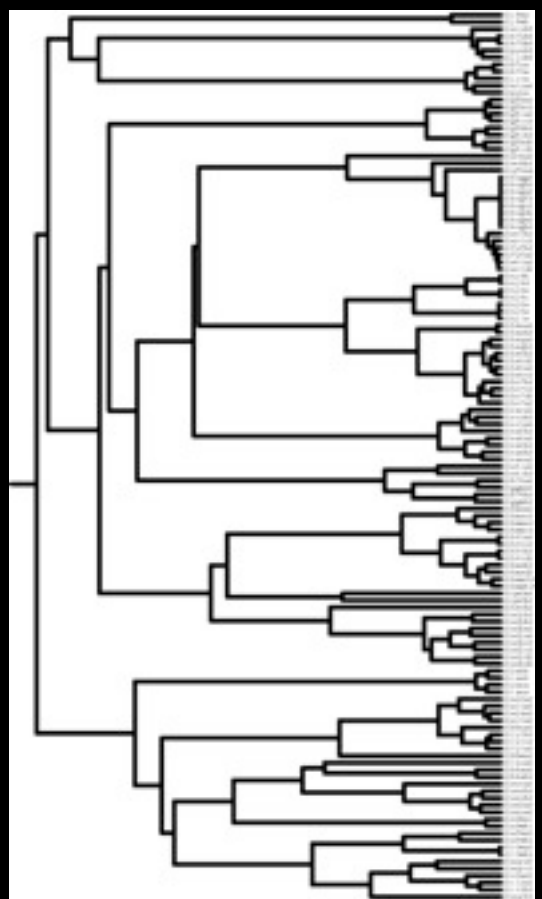
for review see Woese *PNAS* 2000

Woese, et al., *MMBR* 2000.

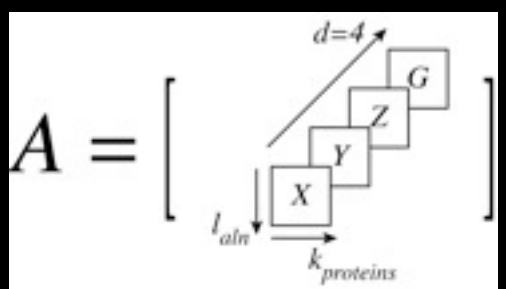
Setting the Genetic Code: Evolutionary Analysis

Non-redundant sets of sequences and structures can be used for statistical analysis.

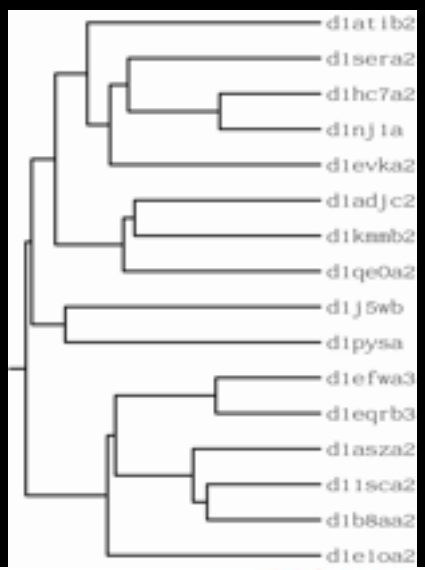
Too much information
129 Structures

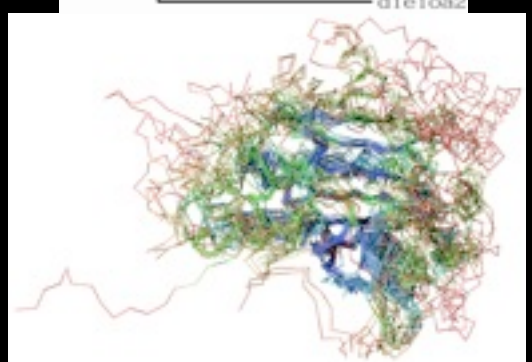


Multidimensional QR factorization of alignment matrix, ***A***.



Economy of information
16 representatives



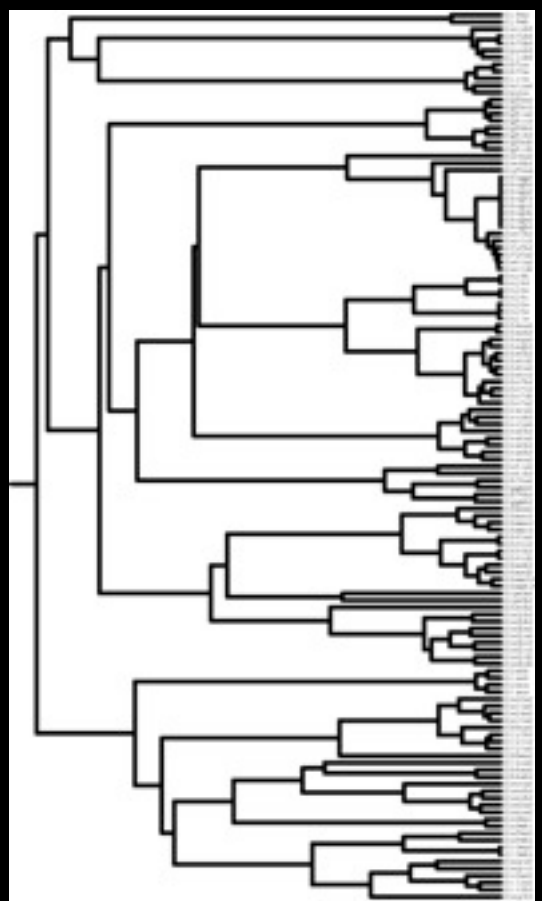


P. O'Donoghue, et al. JMB, 2005; A. Sethi, et al., PNAS, 2005

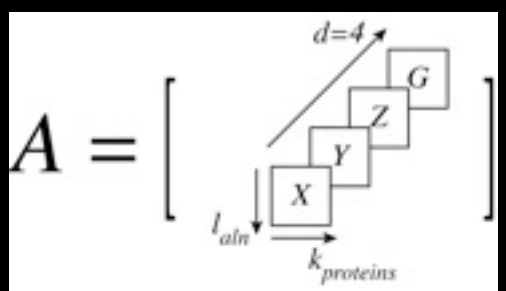
Setting the Genetic Code: Evolutionary Analysis

Non-redundant sets of sequences and structures can be used for statistical analysis.

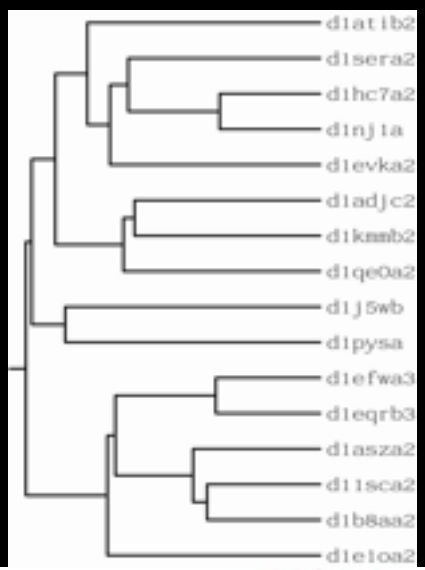
Too much information
129 Structures

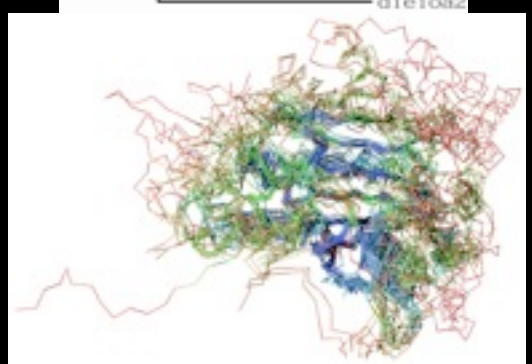


Multidimensional QR factorization of alignment matrix, ***A***.



Economy of information
16 representatives





QR computes a set of maximal linearly independent sequences/structures.

P. O'Donoghue, et al. JMB, 2005; A. Sethi, et al., PNAS, 2005

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS
GlnRS
LysRS
CysRS

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS, GlnRS } Indirect Mechanism
LysRS
CysRS

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS, GlnRS } Indirect Mechanism

LysRS Class I aaRS

CysRS

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS, GlnRS } Indirect Mechanism

LysRS Class I aaRS

CysRS

Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186:**8-14. Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS, GlnRS } Indirect Mechanism

LysRS Class I aaRS

CysRS

Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186:**8-14. Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

M. jannaschii
genome database
search using EP of
class II aaRS with
HMMER

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS, GlnRS } Indirect Mechanism

LysRS Class I aaRS

CysRS

Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186:**8-14. Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

M. jannaschii genome database search using EP of class II aaRS with HMMER

Protein	E-value
HisRS	1.1e-10
AspRS	1.9e-10
PheRS α-chain	9.5e-10
ThrRS	6.6e-04
ProRS	9.1e-03
SerRS	9.2e-03
putative CysRS	1.6e-02
AlaRS	5.1e-02
GlyRS	0.12
PheRS β-chain	0.15
DNA repair protein	7.5

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

The story of the missing archaeal CysRS

M.jannaschii genome was completely sequenced in 1996.
Genome had four missing aaRSs:

AsnRS, GlnRS } Indirect Mechanism

LysRS Class I aaRS

CysRS

Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186:**8-14. Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

M. jannaschii genome database search using EP of class II aaRS with HMMER

Protein	E-value
HisRS	1.1e-10
AspRS	1.9e-10
PheRS α-chain	9.5e-10
ThrRS	6.6e-04
ProRS	9.1e-03
SerRS	9.2e-03
putative CysRS	1.6e-02
AlaRS	5.1e-02
GlyRS	0.12
PheRS β-chain	0.15
DNA repair protein	7.5

← MJ1660

A Sethi, et al., PNAS, 2005.

Setting the Genetic Code: Evolutionary Analysis

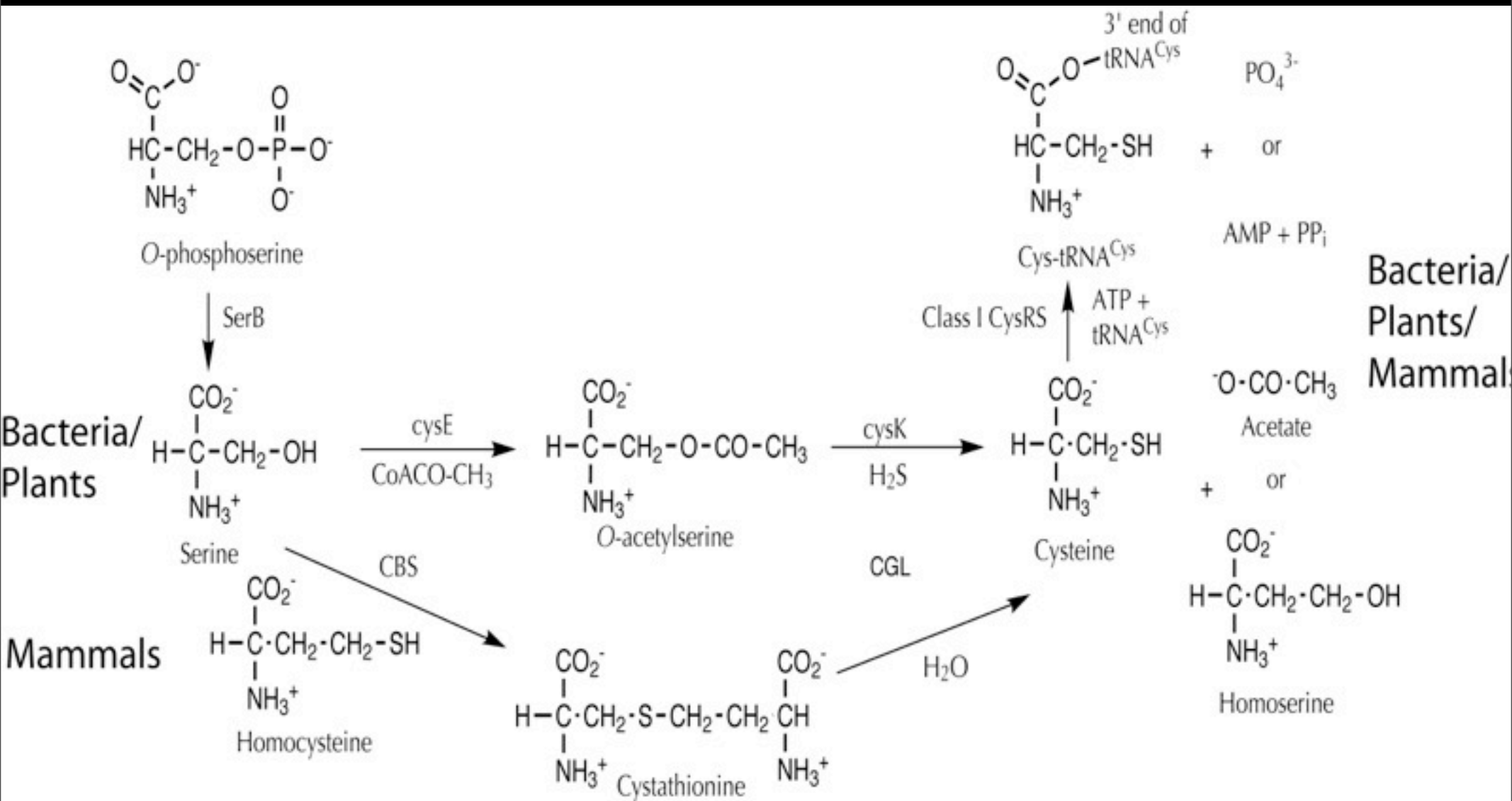
The Complete Story

Sauerwald, et al., Science, 2005.

Setting the Genetic Code: Evolutionary Analysis

The Complete Story

Direct pathway for cysteine aminoacylation



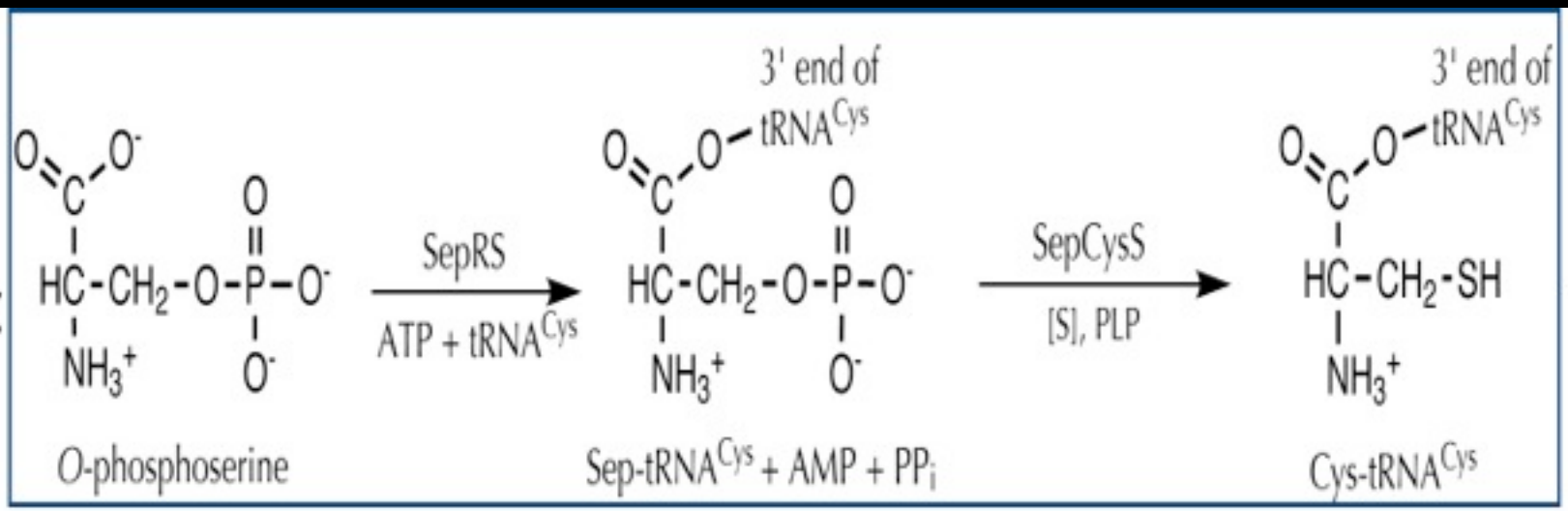
Sauerwald, et al., Science, 2005.

Setting the Genetic Code: Evolutionary Analysis

The Complete Story

Indirect pathway for cysteine aminoacylation

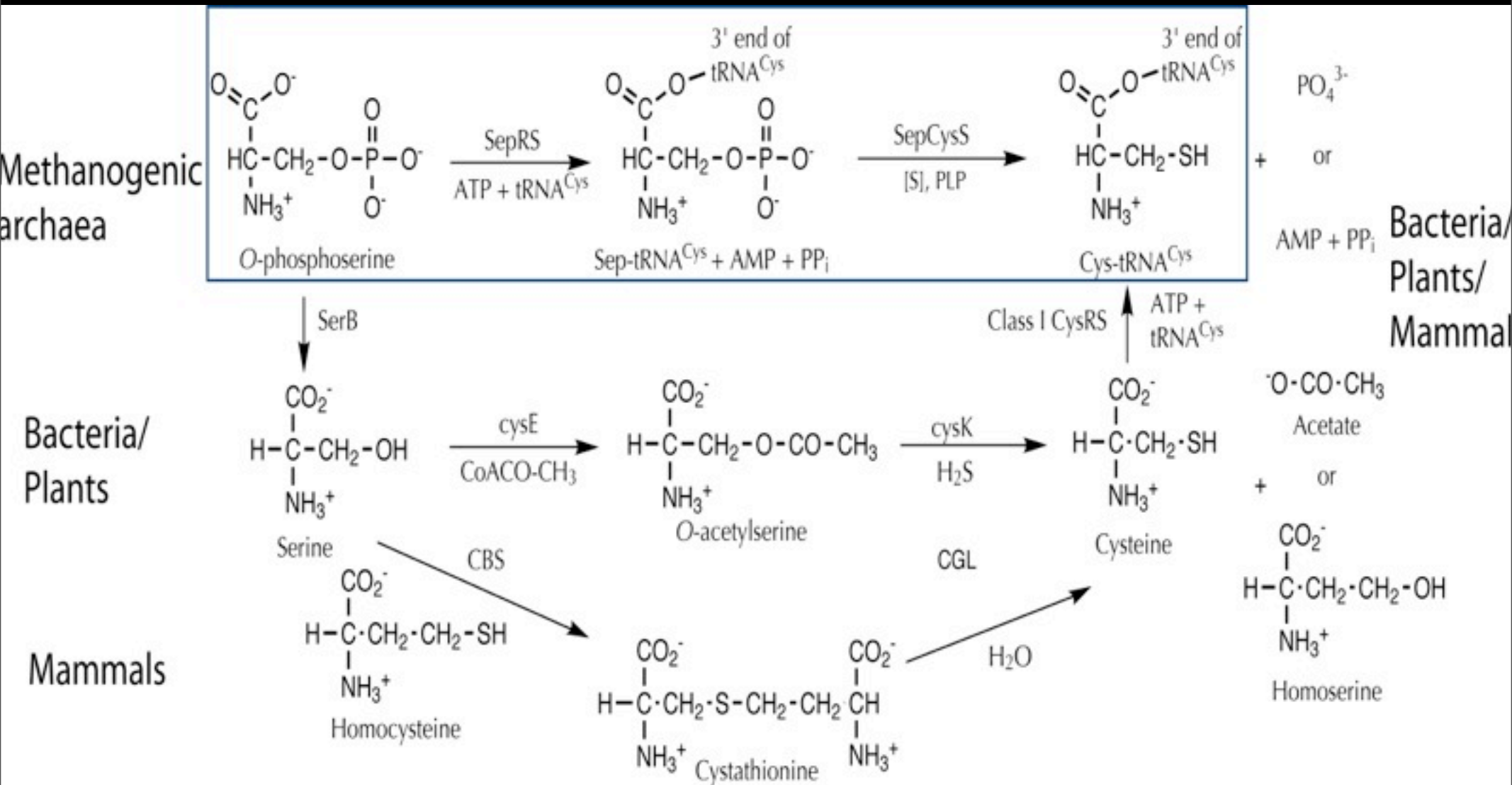
Methanogenic archaea



Sauerwald, et al., Science, 2005.

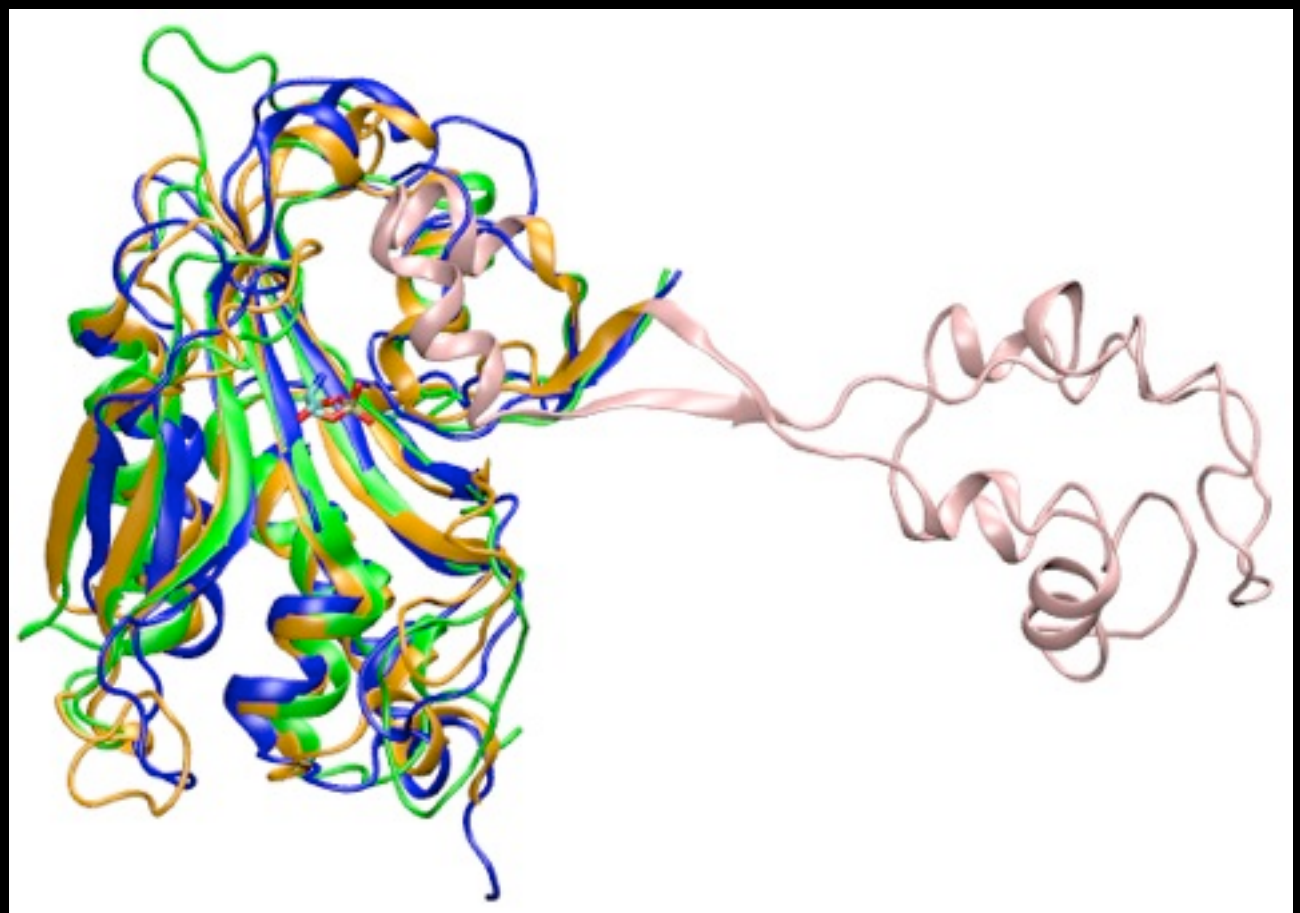
Setting the Genetic Code: Evolutionary Analysis

The Complete Story



Sauerwald, et al., Science, 2005.

Success from Structure Prediction



RMSD = 2.72 Å

Orientation of O-phosphoserine in SepRS is different from that of all other amino acid substrates in the catalytic site of class II aaRS.

Fukunaga and Yokoyama, Nature Str. Mol. Biol., 2007.

Sethi, et al., PNAS, 2005.

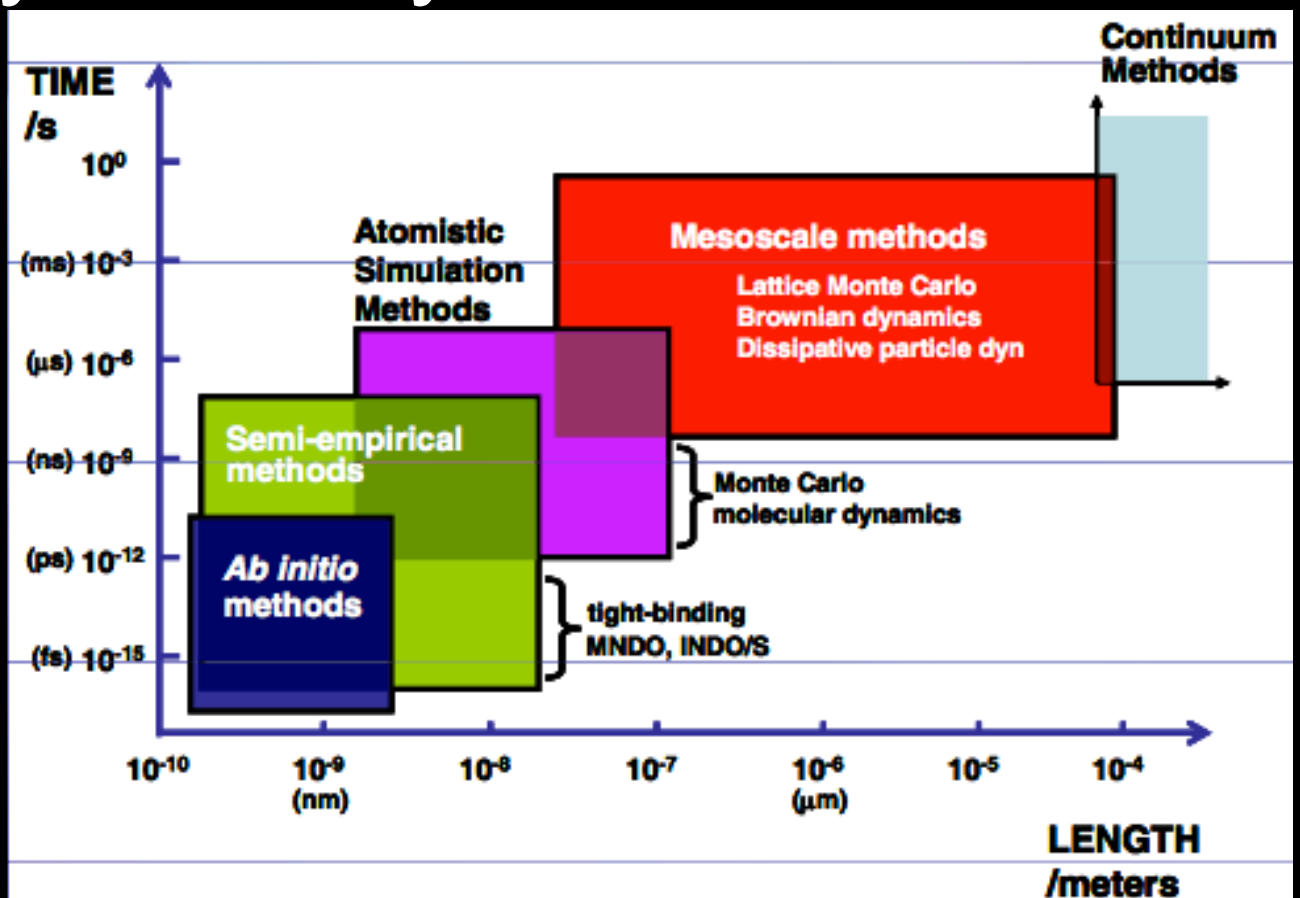
SeqQR was used to study all the steps of translation.

Evolution gives valuable clues to understand how complex systems function.

However, conservation can be due to a variety of reasons and teasing out the details requires physical models.

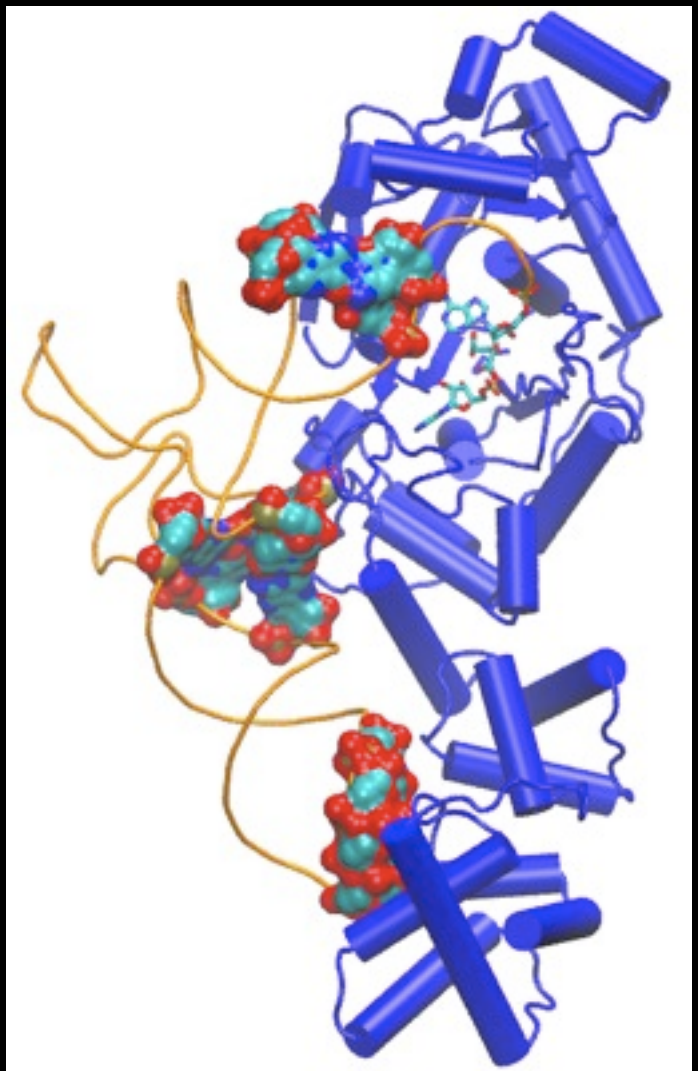
Setting the Genetic Code: Signaling Within Biomolecules

Molecular dynamics simulations can be used to analyze the dynamics within biomolecules



$$U\left(\vec{R}\right) = \sum_{\text{bonds}} K_b \left(b - b_0\right)^2 + \sum_{\text{UB}} K_{UB} \left(S - S_0\right)^2 + \sum_{\text{angle}} K_\theta \left(\theta - \theta_0\right)^2$$
$$+ \sum_{\text{dihedrals}} K_\chi \left(1 + \cos\left(n\chi - \delta\right)\right) + \sum_{\text{impropers}} K_{\text{imp}} \left(\phi - \phi_0\right)^2$$
$$+ \sum_{\text{nonbond}} \epsilon \left[\left(\frac{R_{min_{ij}}}{r}\right)^{12} - \left(\frac{R_{min_{ij}}}{r}\right)^{6}\right] + \frac{q_i q_j}{\epsilon_1 r_{ij}}$$

Long range communication is necessary for setting the genetic code



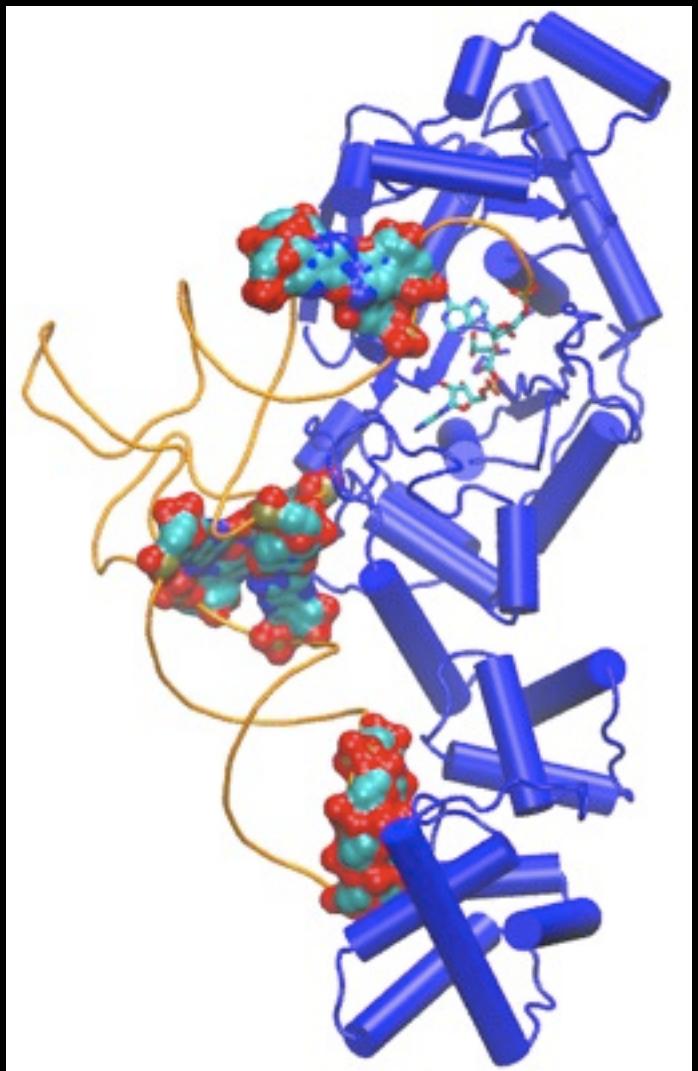
PDB ID: 1N78

$$aa + ATP \xrightarrow{aaRS} aa\text{-}AMP + PP_i$$

$$aa\text{-}AMP + tRNA^{aa} \xrightarrow{aaRS} aa\text{-}tRNA^{aa} + AMP$$

Sekine, et. al., JMB, 1996.Sekine, et. al., Eur. J. Biochem, 1999.

Long range communication is necessary for setting the genetic code



PDB ID: 1N78

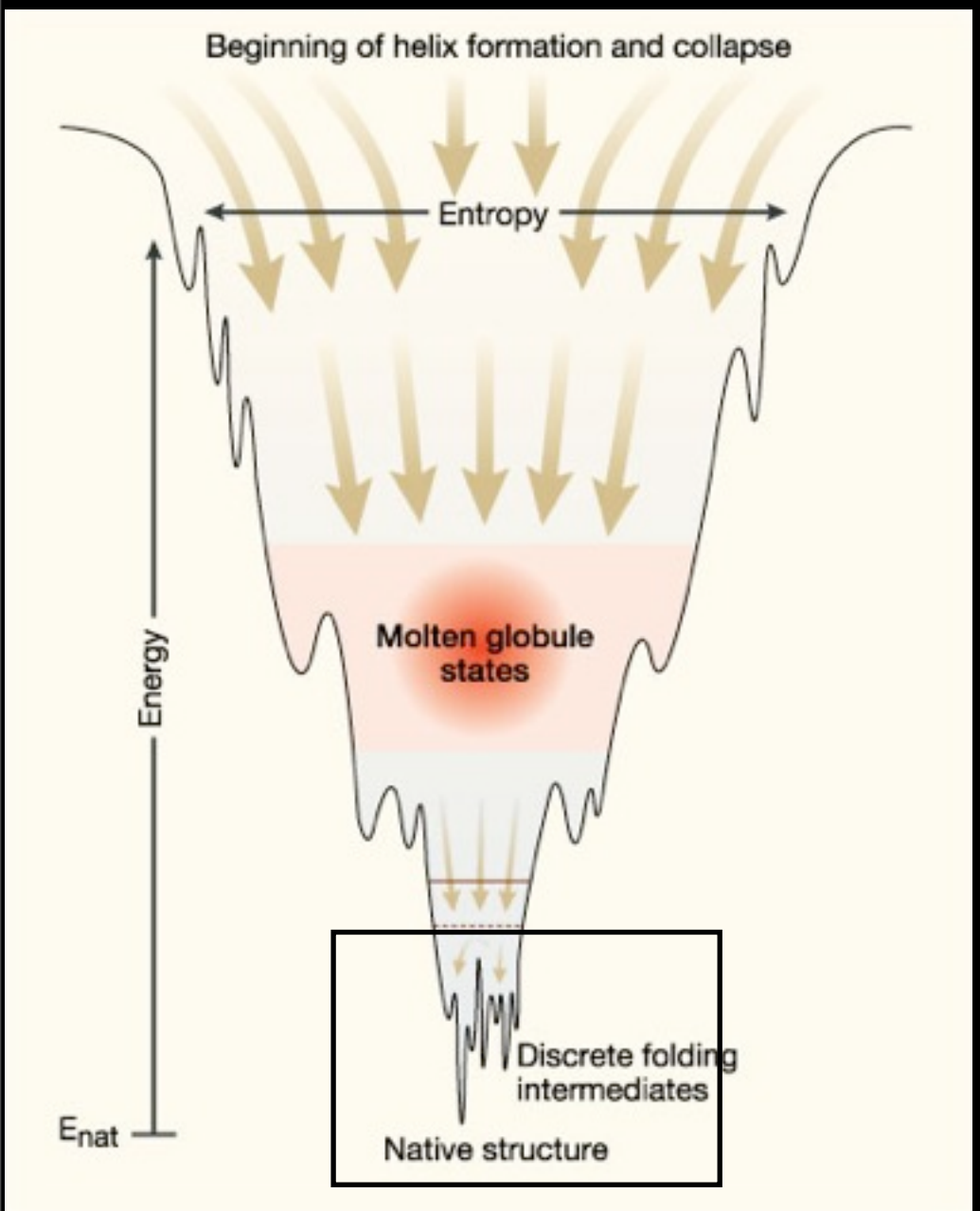
Nucleotides that affect the efficiency of the reaction (catalytic rate) can occur up to 50-70 Angstoms away from the catalytic site.

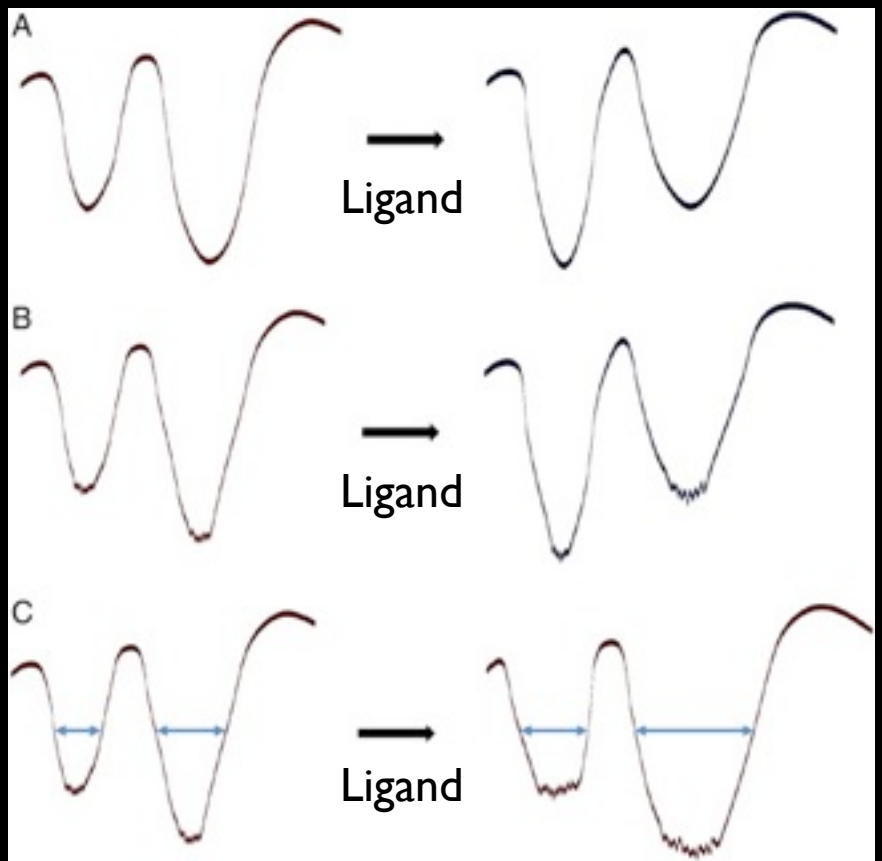
$$aa + ATP \xrightarrow{aaRS} aa\text{-}AMP + PP_i$$

$$aa\text{-}AMP + tRNA^{aa} \xrightarrow{aaRS} aa\text{-}tRNA^{aa} + AMP$$

Sekine, et. al., JMB, 1996.Sekine, et. al., Eur. J. Biochem, 1999.

Energy landscape theory explains how macromolecules fold and function



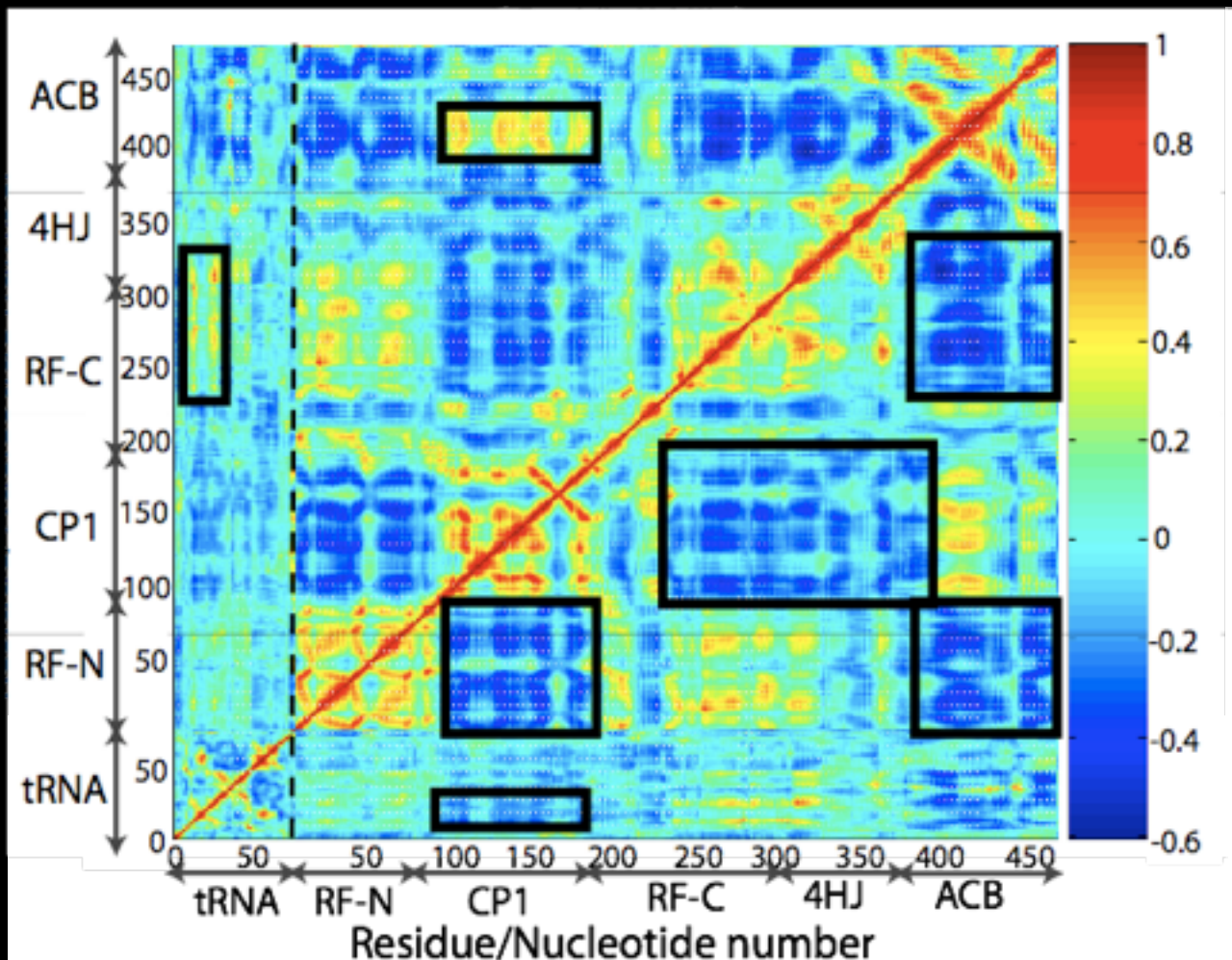


Allostery involves a change in conformation or dynamics.

Structural changes might occur through a network of local changes.

Onuchic, et al., Ann Rev Phys Chem, 1997
Lila Gierarsch, Curr Opin Str Biol, 2006

Observation of Allostery in MD Simulations

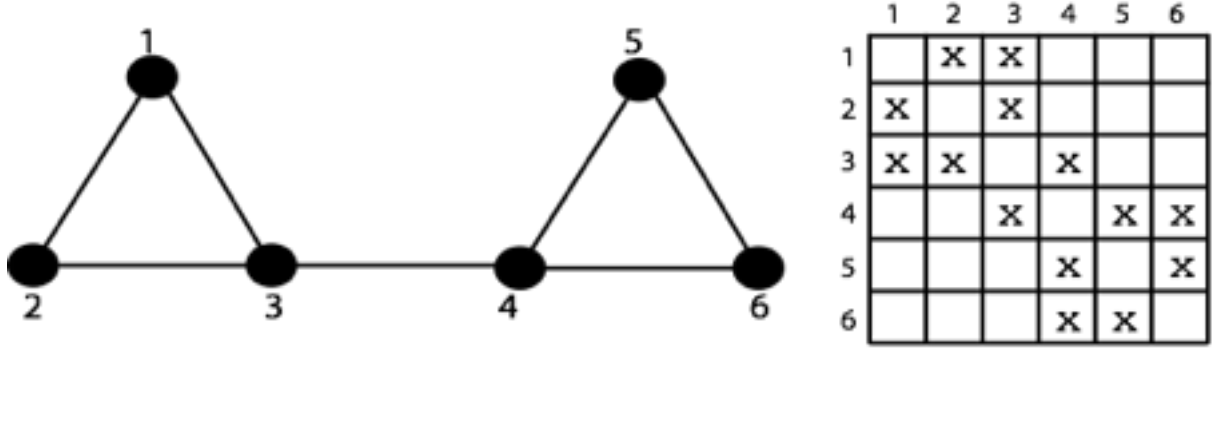


$$C(i,j) = \frac{< \Delta r_i(t) \cdot \Delta r_j(t) >}{(< \Delta r_i(t)^2 >< \Delta r_j(t)^2 >)^{0.5}}$$

Correlation in motion between residues and nucleotides in protein:RNA complex.

A Sethi, et al., PNAS, 2009.

Ideas borrowed from network theory



	1	2	3	4	5	6
1		x	x			
2	x		x			
3	x	x		x		
4			x		x	x
5				x		x
6				x	x	

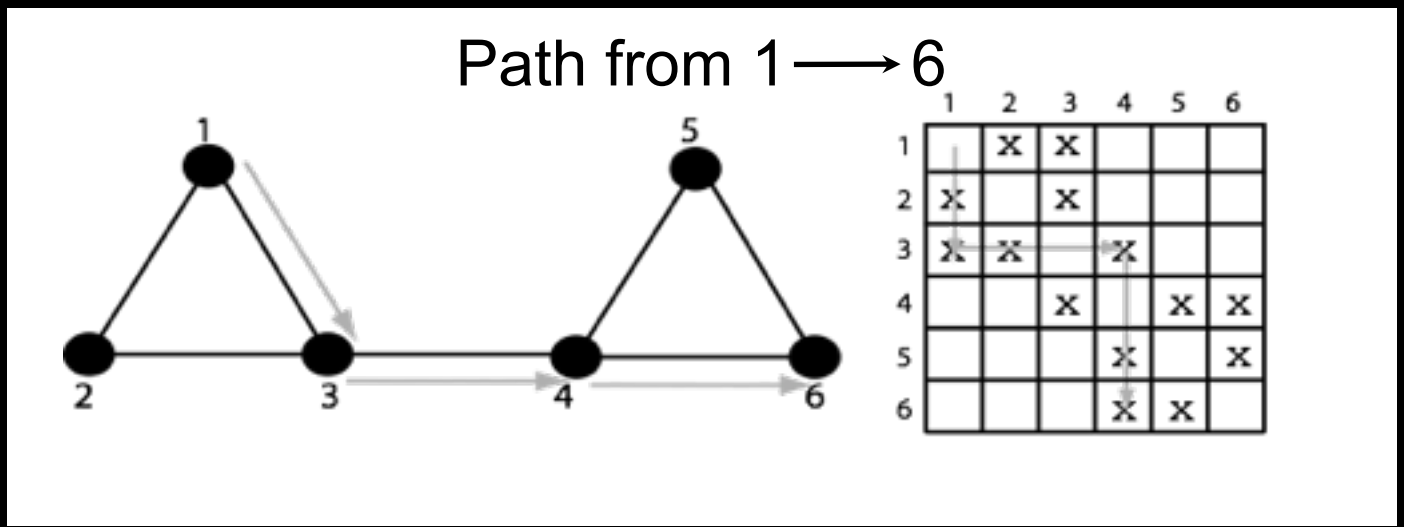
Nodes represent the residues/nucleotides in the complex.

Edges represent contact between monomers in the complex.

The edges can either be unweighted or weighted. The edges are weighted by correlation (C_{ij}) between contacts in the simulation:

$$w_{ij} = -\ln(|C_{ij}|)$$

Ideas borrowed from network theory



Nodes represent the residues/nucleotides in the complex.

Edges represent contact between monomers in the complex.

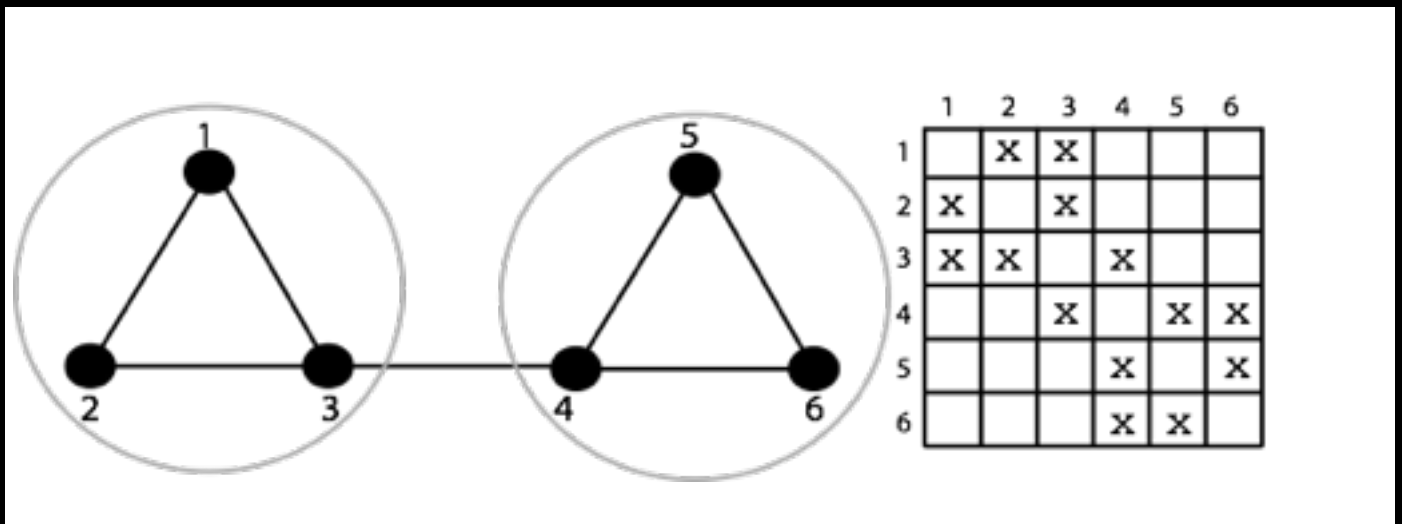
The edges can either be unweighted or weighted. The edges are weighted by correlation (C_{ij}) between contacts in the simulation:

$$w_{ij} = -\ln(|C_{ij}|)$$

The path distance is the sum of the weights or distance of each edge:

$$D_{ij} = \sum_{k,l} w_{kl}$$

Ideas borrowed from network theory



Nodes represent the residues/nucleotides in the complex.

Edges represent contact between monomers in the complex.

The edges can either be unweighted or weighted. The edges are weighted by correlation (C_{ij}) between contacts in the simulation:

$$w_{ij} = -\ln(|C_{ij}|)$$

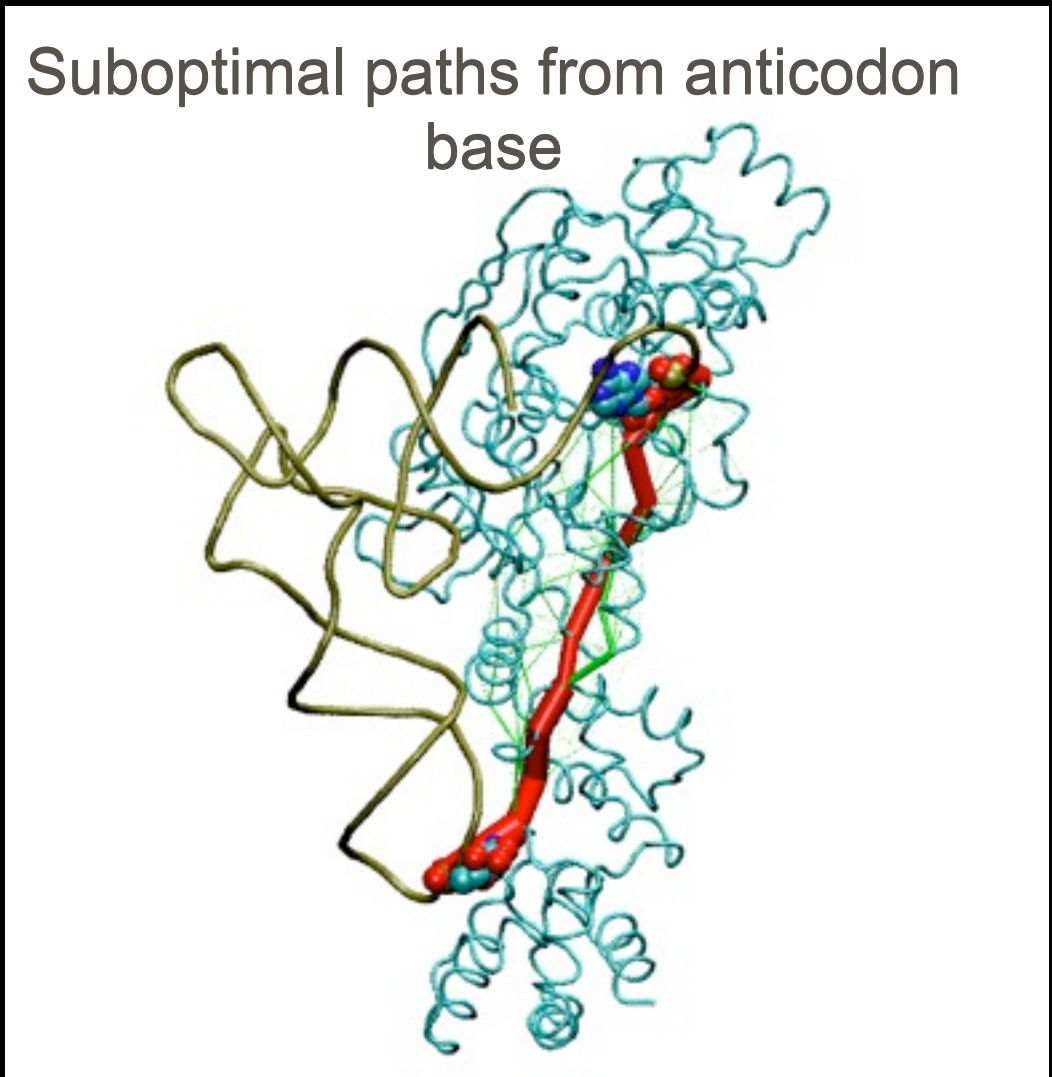
Modules have fewer connections between them

The path distance is the sum of the weights or distance of each edge:

$$D_{ij} = \sum_{k,l} w_{kl}$$

Girvan and Newman, PNAS, 2002

The Importance of Suboptimal Paths



There are a number of suboptimal paths for communication between identity elements and active site.

A Sethi, et al., PNAS, 2009.

The Importance of Suboptimal Paths

Suboptimal paths from anticodon base

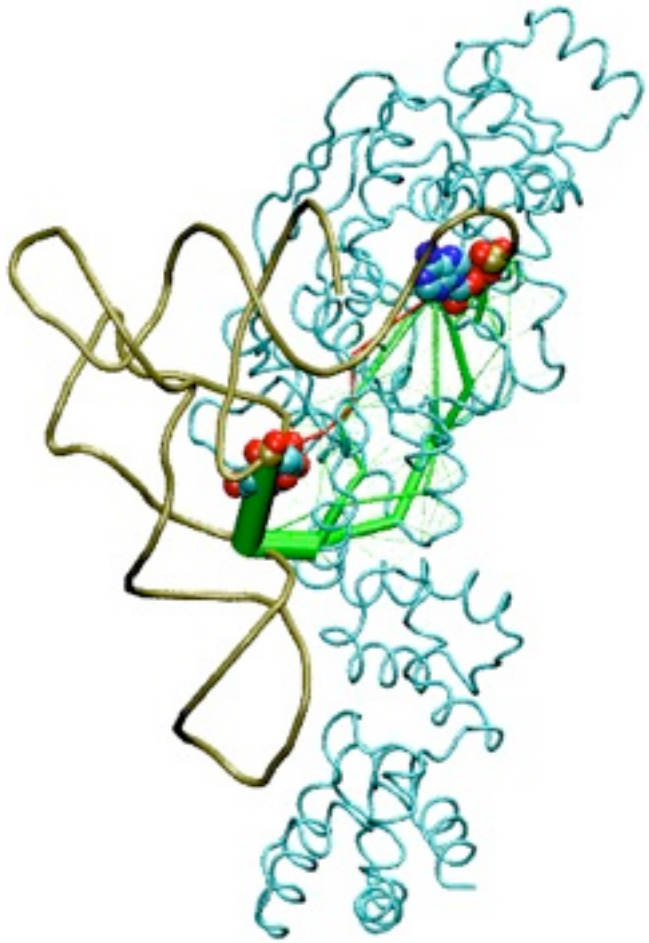
There are a number of suboptimal paths for communication between identity elements and active site.

A Sethi, et al., PNAS, 2009.

The Importance of Suboptimal Paths

Suboptimal paths from anticodon base

Suboptimal paths from Ura11



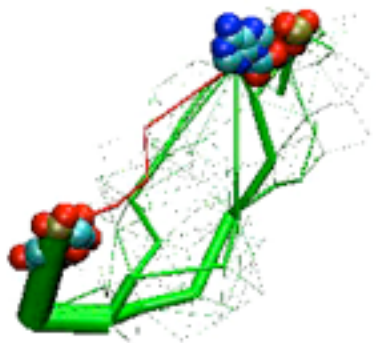
There are a number of suboptimal paths for communication between identity elements and active site.

A Sethi, et al., PNAS, 2009.

The Importance of Suboptimal Paths

Suboptimal paths from anticodon base

Suboptimal paths from Ura11



There are a number of suboptimal paths for communication between identity elements and active site.

A Sethi, et al., PNAS, 2009.

Regions connecting modules form hotspots for communication in the network

Communities are modules within the network that move in a correlated fashion during the MD simulation.

Residues connecting modules are critical for communication in the biomolecular network

They are conserved in evolution

They affect network properties

They occur in majority of suboptimal pathways

A Sethi, et al., PNAS, 2009.

Regions connecting modules form hotspots for communication in the network

Communities are modules within the network that move in a correlated fashion during the MD simulation.

Residues connecting modules are critical for communication in the biomolecular network

They are conserved in evolution

They affect network properties

They occur in majority of suboptimal pathways

A Sethi, et al., PNAS, 2009.

The modules in the protein are more highly conserved

CAP210 YU2 HXBC2

The communities in the network are highly conserved between different sequences of gp120.

The intermodular contacts are under high immune pressure to evolve.

A Sethi, et al., PLoS Comp Biol, 2013

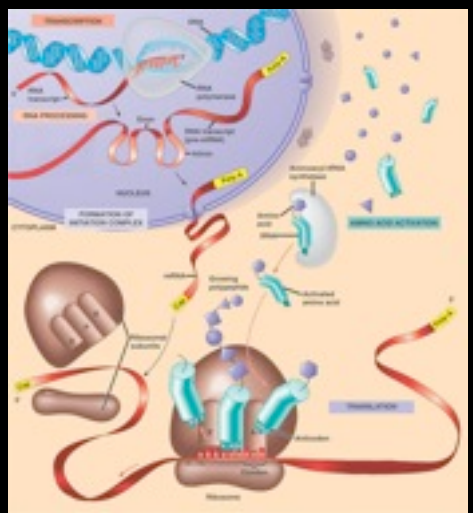
Conclusions

- Balanced evolutionary profiles provide an economy of information that can be used for gene annotation.
- Evolutionary profiles were successful at identifying the protein responsible for cysteine aminoacylation in methanogens.
- The suboptimal paths should be considered while studying communication between distant sites in biomolecular complexes.
- The residues involved in communication between modules in the dynamical network are highly conserved and form hot spots for communication in biomolecular complexes.

Organization of Talk

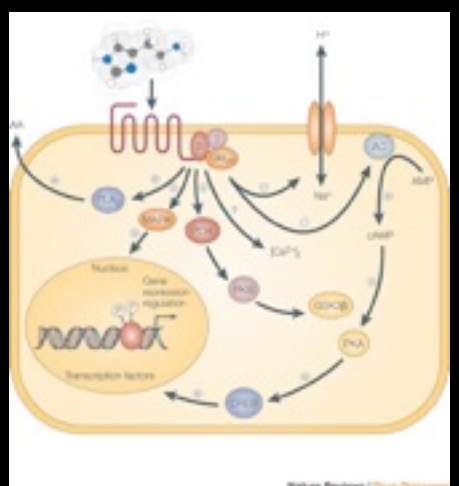
Translating Information in RNA

- Evolutionary analysis of biomolecules
- Allosteric signaling pathways

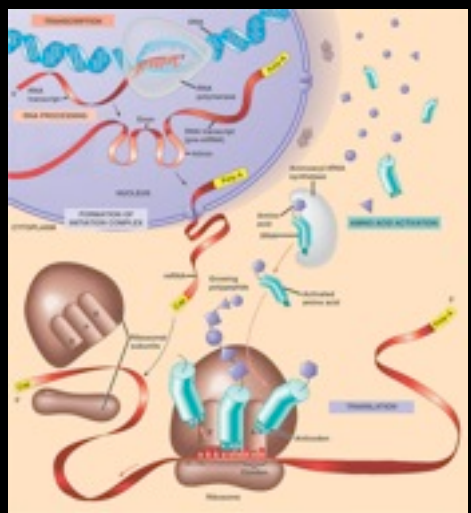


Disordered regions in Signaling Proteins

- Multivalent Proteins
- Modeling Disordered Proteins

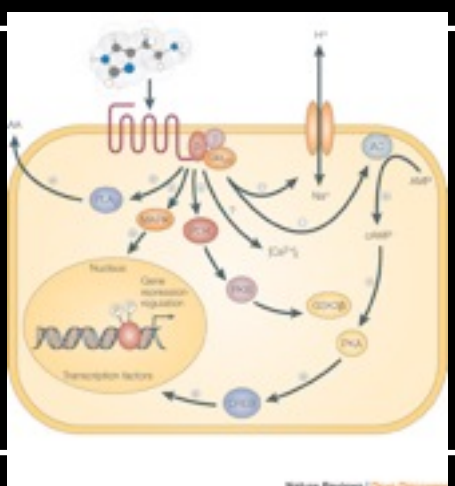


Organization of Talk



Disordered regions in Signaling Proteins

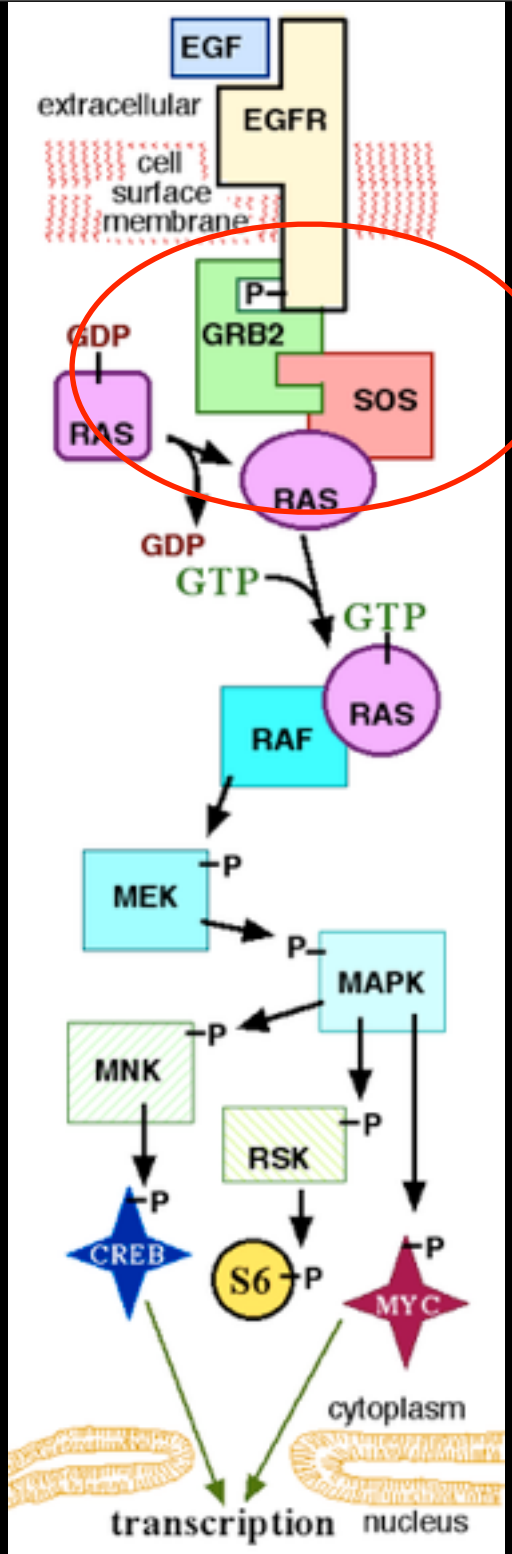
- Multivalent Proteins
- Modeling Disordered Proteins



Part 2:

Structurally Modeling Signaling Cascades

Signaling cascades regulate information transfer inside the cell.

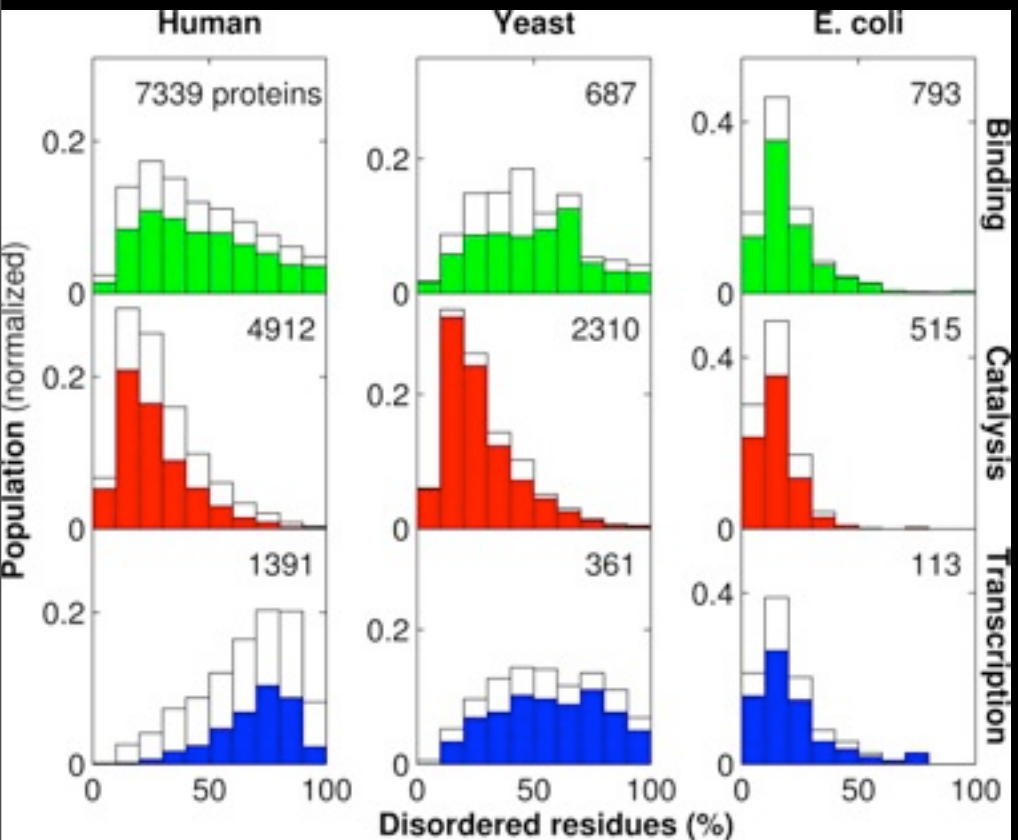


http://en.wikipedia.org/wiki/Cell_signaling

Eukaryotes have a significant proportion of their proteins that are disordered.

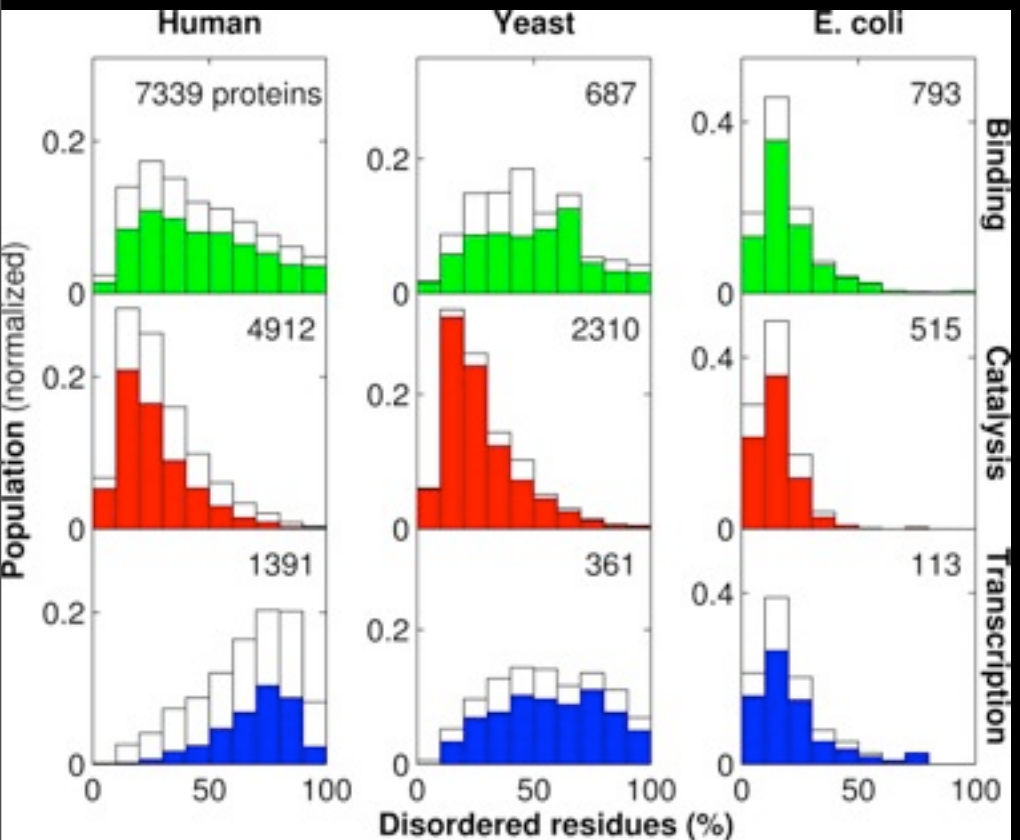
Classical interpretation:

Protein's ordered structure is related to its function.



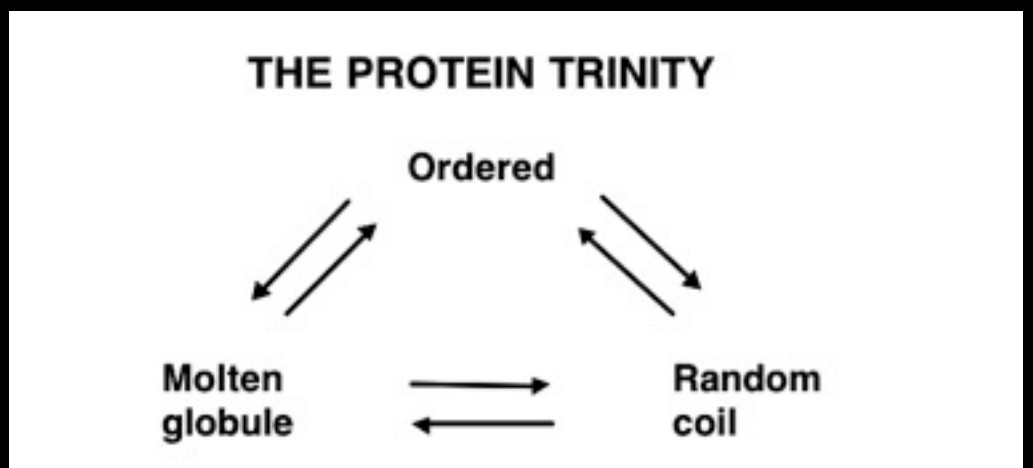
Liu, et al. PNAS, 2009

Eukaryotes have a significant proportion of their proteins that are disordered.



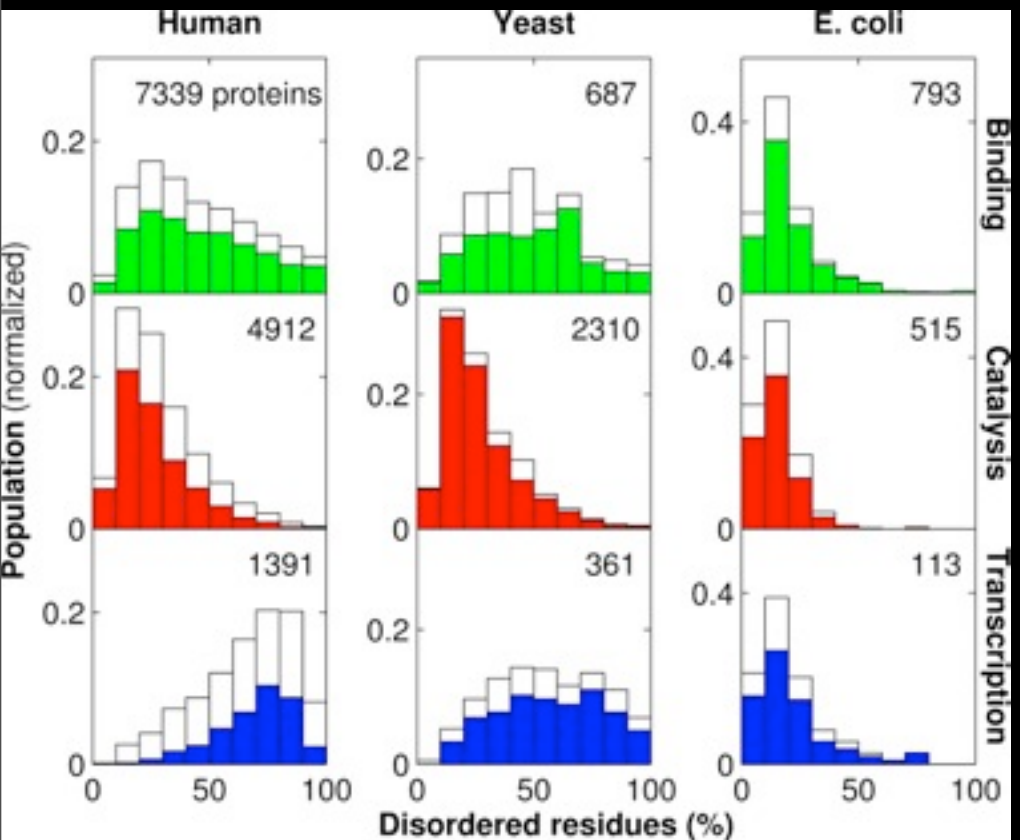
Classical interpretation:

Protein's ordered structure is related to its function.



Liu, et al. PNAS, 2009

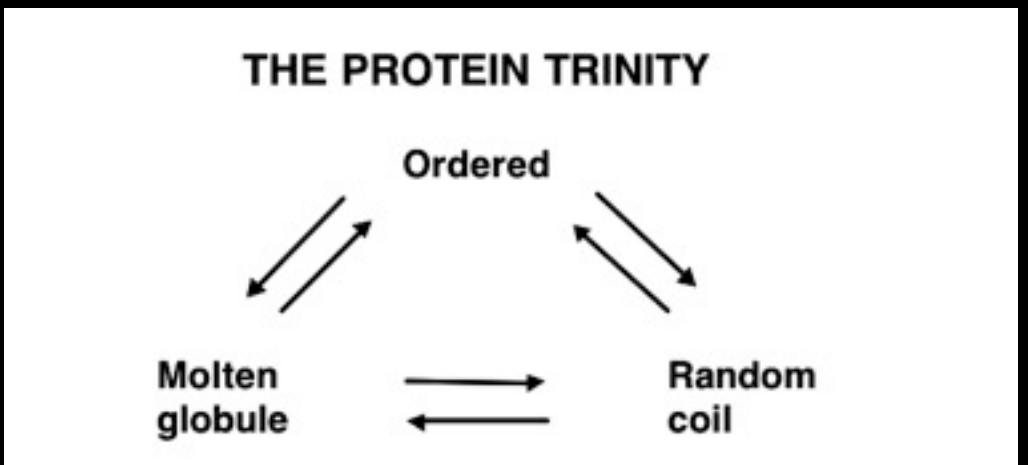
Eukaryotes have a significant proportion of their proteins that are disordered.



Liu, et al. PNAS, 2009

Classical interpretation:

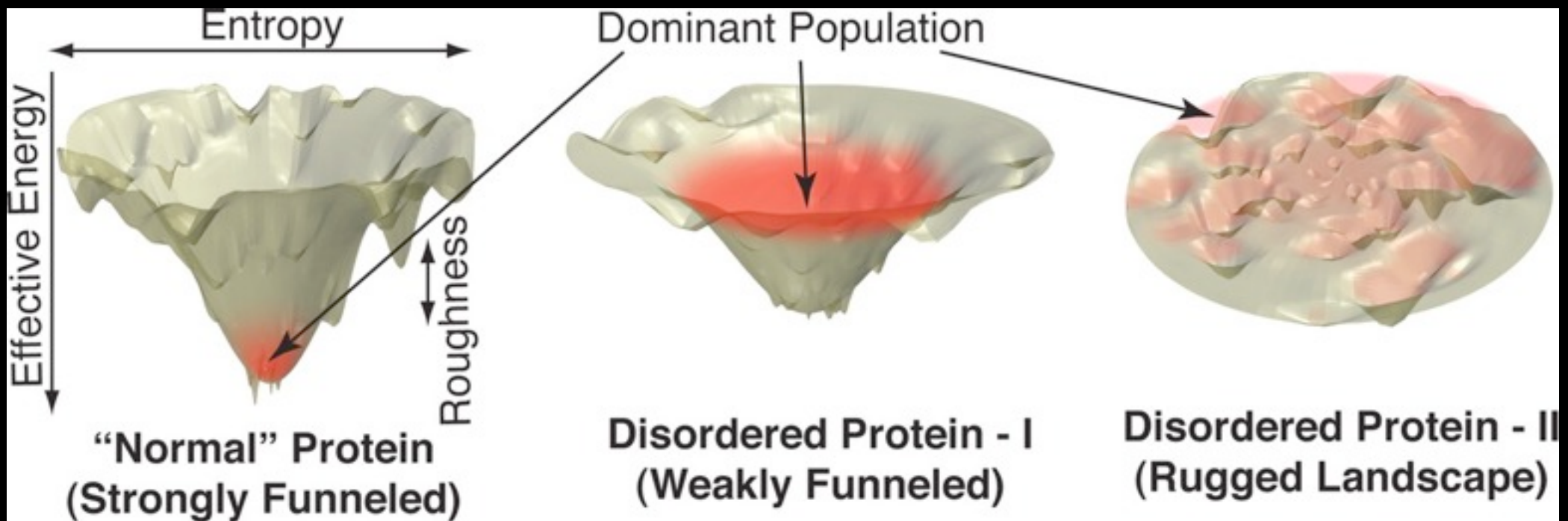
Protein's ordered structure is related to its function.



Folding upon binding
Fuzzy complexes
Entropic chains

Signaling Cascades: Disordered Regions

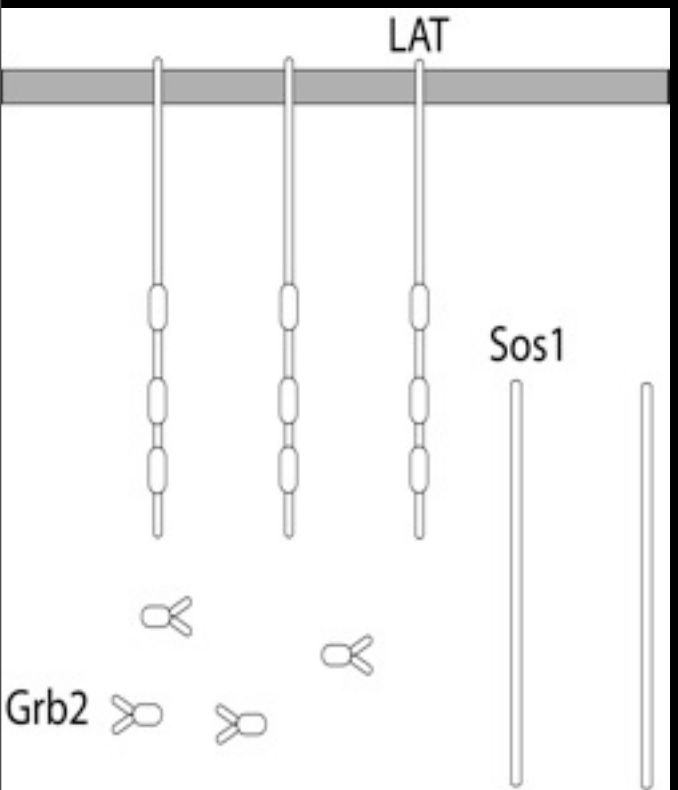
Disordered proteins are considered to either have a weakly funneled or a rugged energy landscape



Onuchic, et al., Ann Rev Phys Chem, 1997

Papoian, PNAS, 2008

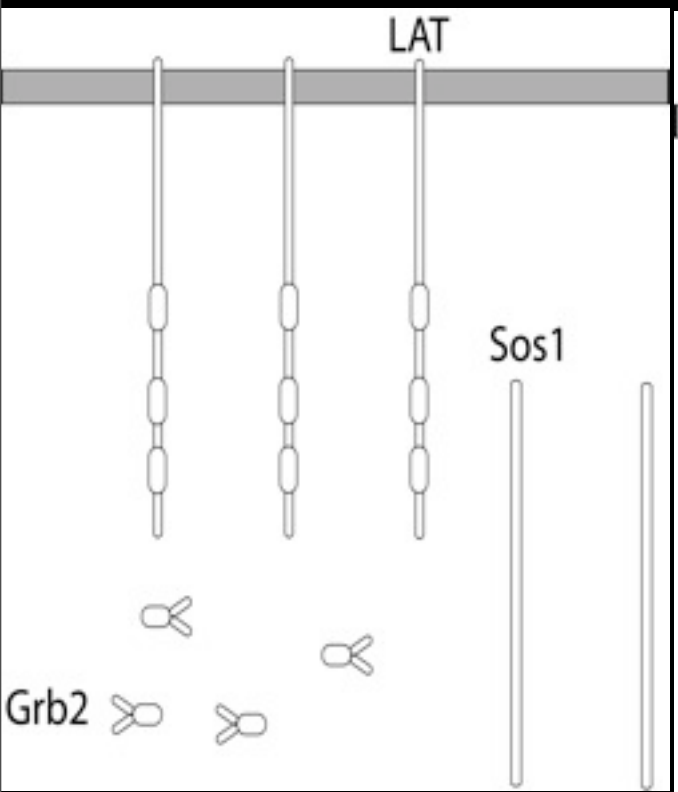
Signaling proteins utilize multivalent interactions

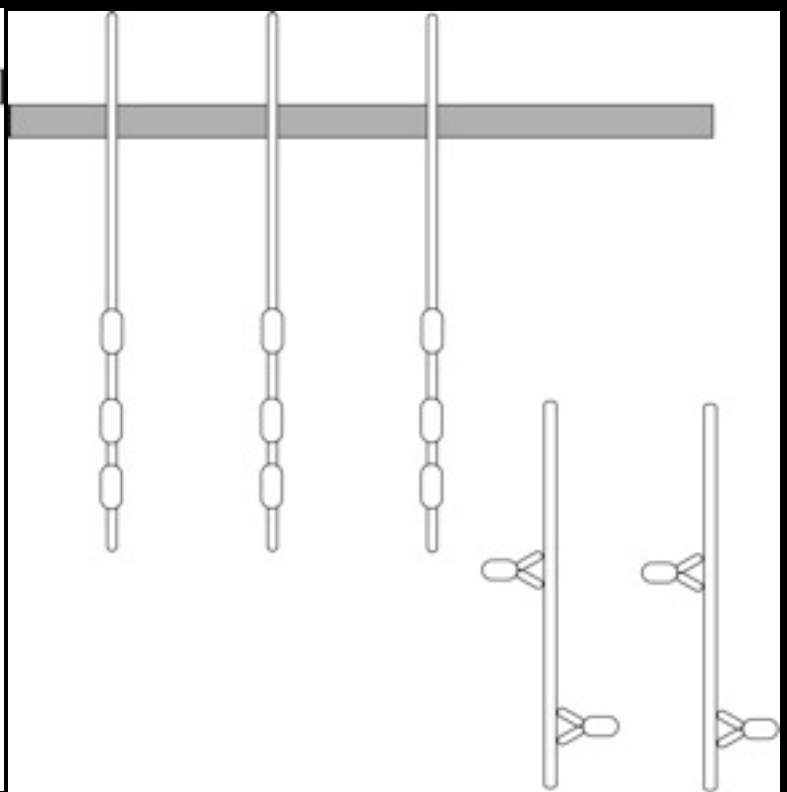


The nucleotide exchange factor Sos1 has to be localized near the plasma membrane.

Houtman, et al., NSB, 2006.
Nag, et al., Biophys J., 2009.

Signaling proteins utilize multivalent interactions



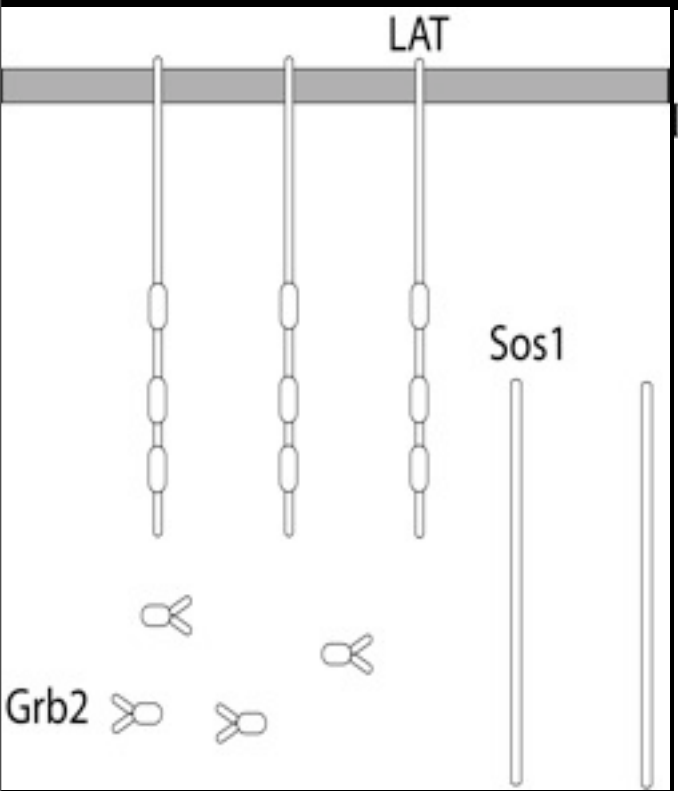


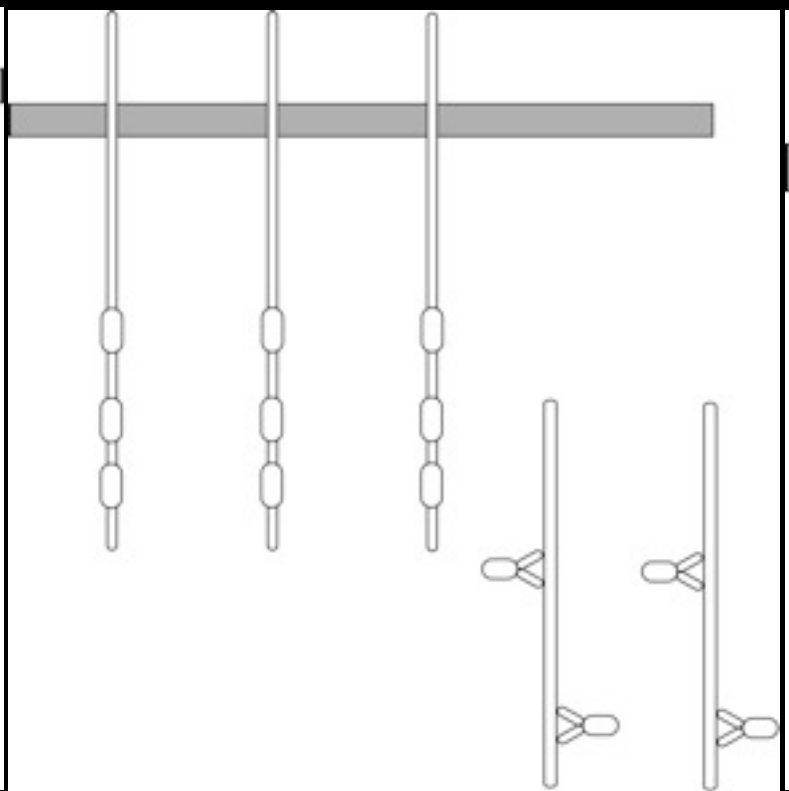
The nucleotide exchange factor Sos1 has to be localized near the plasma membrane.

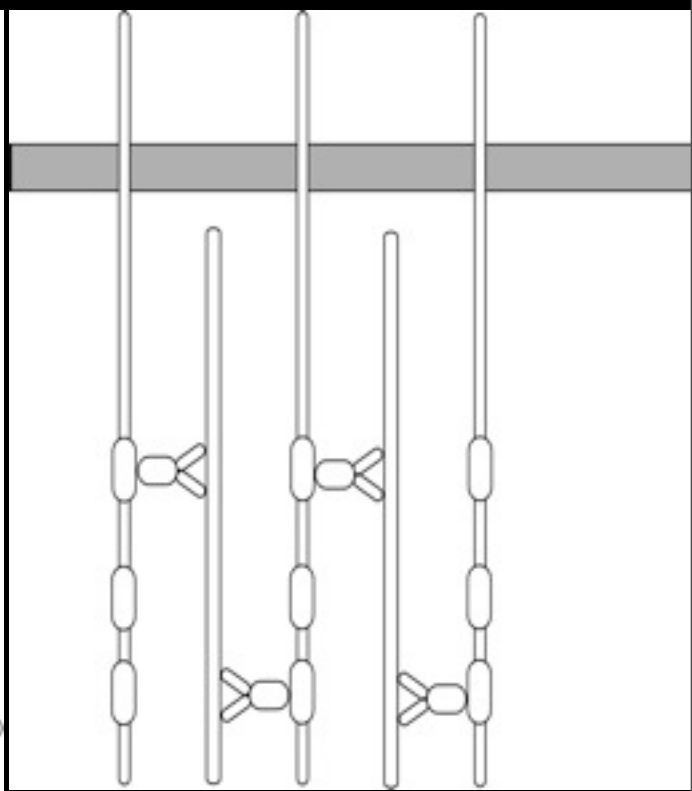
Grb2 consists of two SH3 domains that interact with Sos1.

Houtman, et al., NSB, 2006.
Nag, et al., Biophys J., 2009.

Signaling proteins utilize multivalent interactions







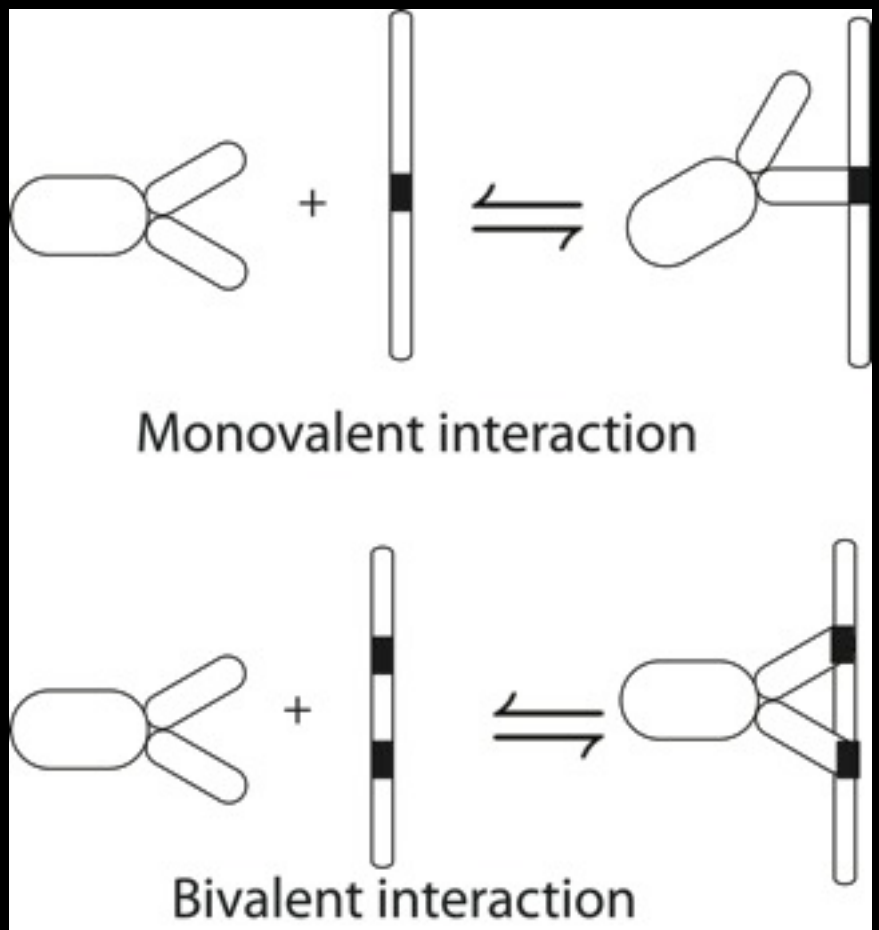
The nucleotide exchange factor Sos1 has to be localized near the plasma membrane.

Grb2 consists of two SH3 domains that interact with Sos1.

Grb2 consists of a SH2 domain that interacts with LAT.

Houtman, et al., NSB, 2006.
Nag, et al., Biophys J., 2009.

Multivalent interactions are ubiquitous in biology.

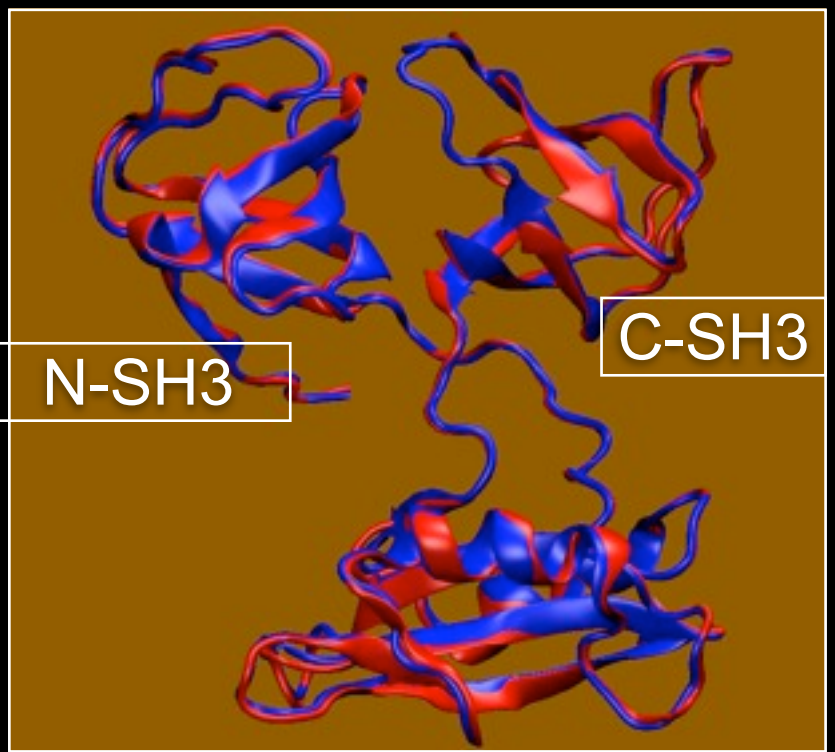


Signaling proteins often use multiple binding sites to increase the overall strength and specificity of the complexes formed.

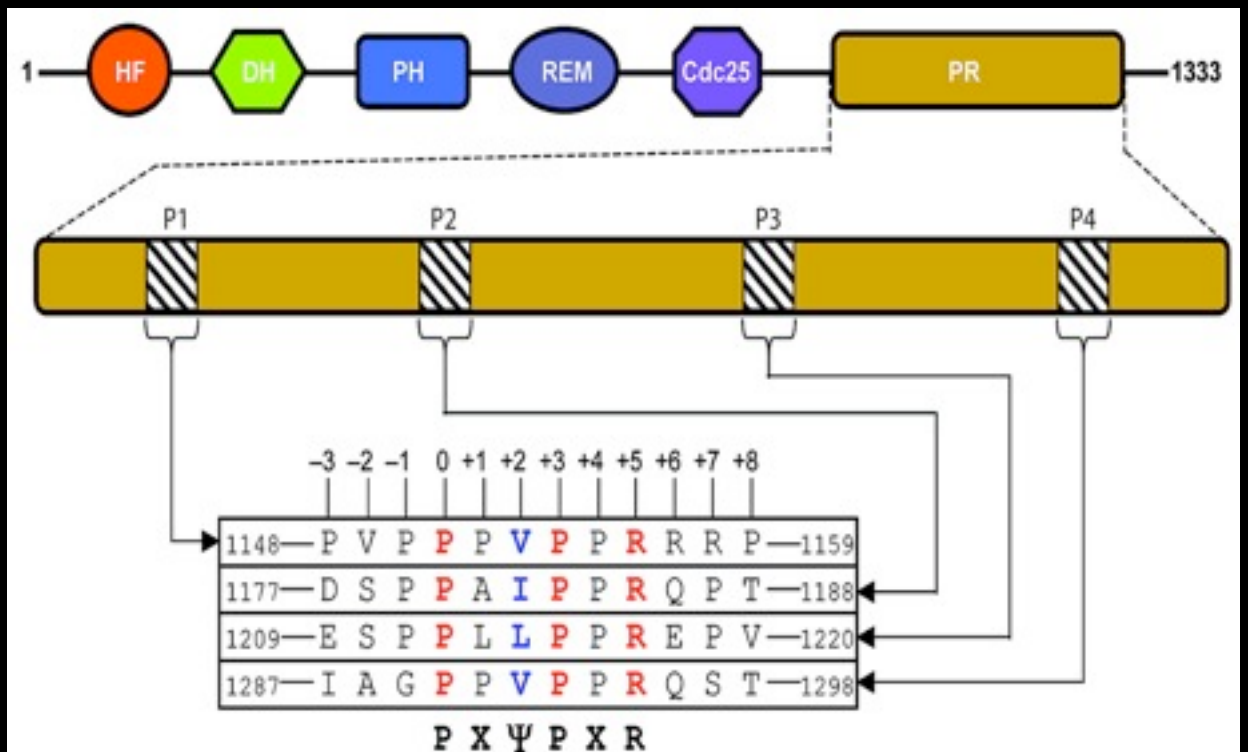
Mammen, et al., Angewandte Chemie, 1998

Signaling Cascades: Quantifying Multivalent Binding

Controversy on stoichiometry of complexes formed under physiological conditions.



Grb2



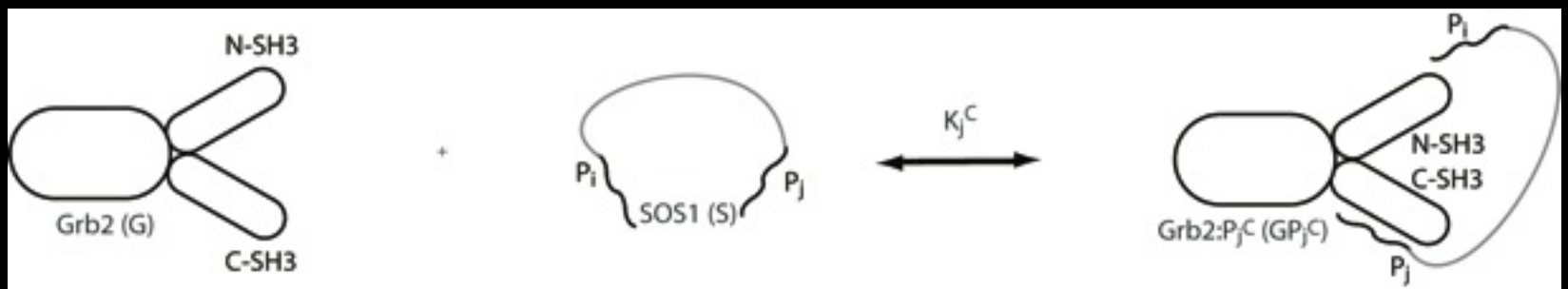
Sos1

Maignan, et al., Science, 1995.

McDonald, et al., Biochemistry, 2009

Signaling Cascades: Quantifying Multivalent Binding

Simultaneous binding of both SH3 domains to two motifs in Sos1

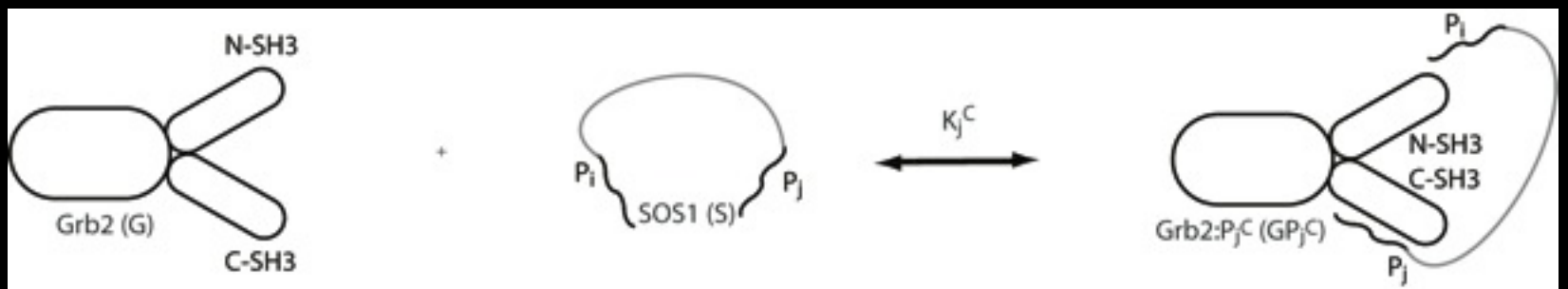


$$[GP_j^C] = K_j^C\,[G]\,[P_j]$$

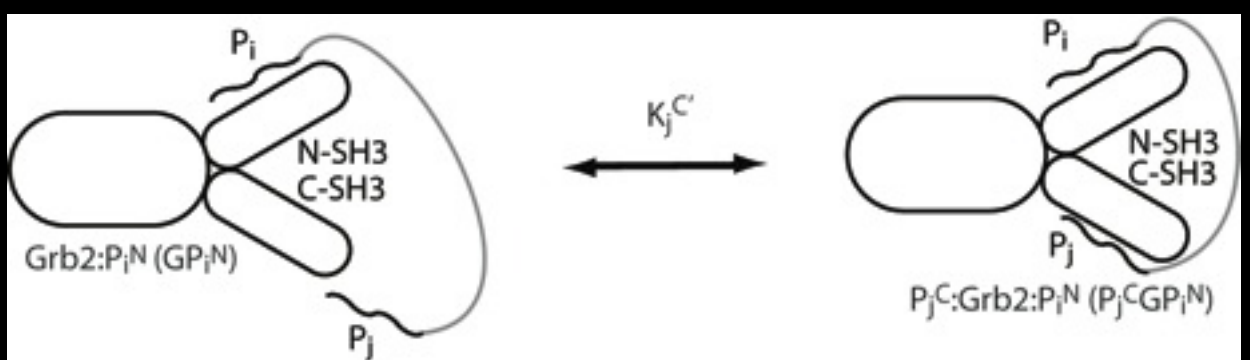
A Sethi, et al., PLoS Comp. Biol., 2011

Signaling Cascades: Quantifying Multivalent Binding

Simultaneous binding of both SH3 domains to two motifs in Sos1



$$[GP_j^C] = K_j^C\,[G]\,[P_j]$$

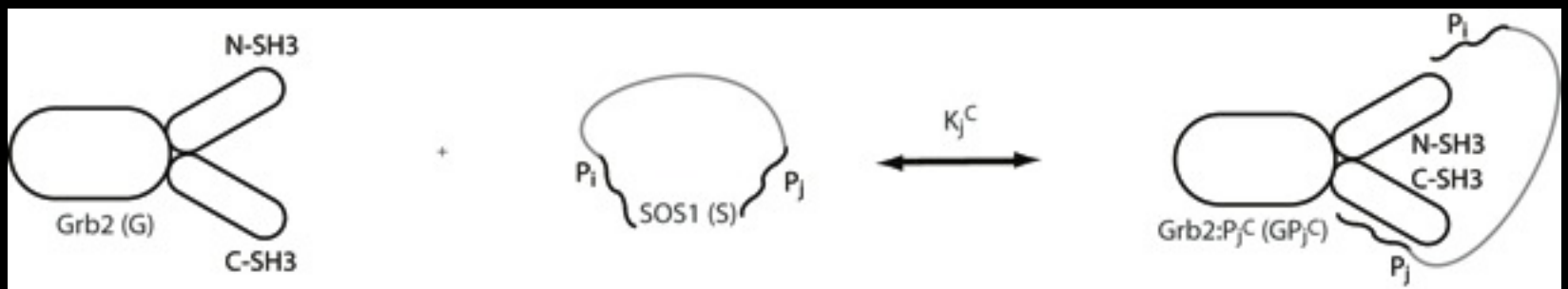


$$[P_j^C GP_i^N] = K_j^{C'}\,[GP_i^N]$$

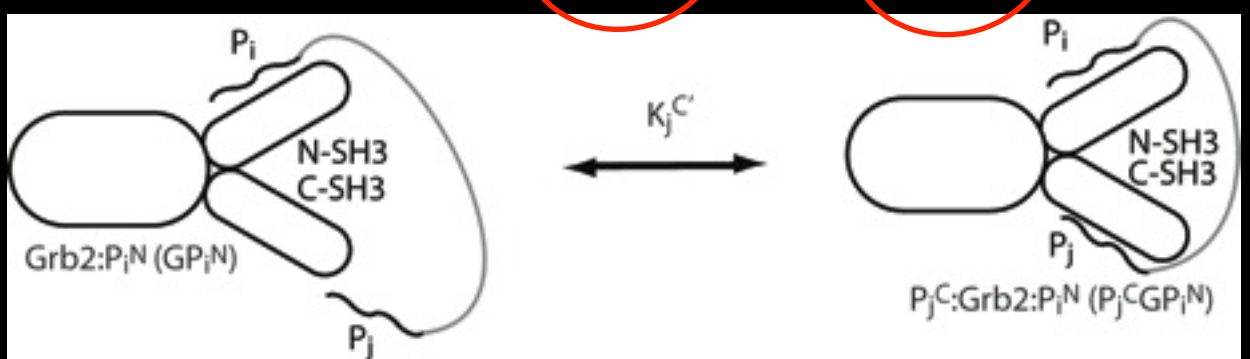
A Sethi, et al., PLoS Comp. Biol., 2011

Signaling Cascades: Quantifying Multivalent Binding

Simultaneous binding of both SH3 domains to two motifs in Sos1



$$[GP_j^C] = K_j^C [G][P_j]$$



$$[P_j^C GP_i^N] = K_j^{C'} [GP_i^N]$$

$$K_j^{C'} = K_j^C \times C_{eff}(P_i^N, P_j^C)$$

A Sethi, et al., PLoS Comp. Biol., 2011

Modeling multivalent interactions

Motifs of Sos1 that bind to Grb2

- Evolutionary analysis
- Binding Energy Calculations

A Sethi, et al., PLoS Comp. Biol., 2011

Modeling multivalent interactions

Motifs of Sos1 that bind to Grb2
- Evolutionary analysis
- Binding Energy Calculations

Simultaneous binding of two motifs in Sos1 to Grb2
- Hybrid model from polymer theory and MD Simulations.

A Sethi, et al., PLoS Comp. Biol., 2011

Modeling multivalent interactions

Motifs of Sos1 that bind to Grb2
- Evolutionary analysis
- Binding Energy Calculations

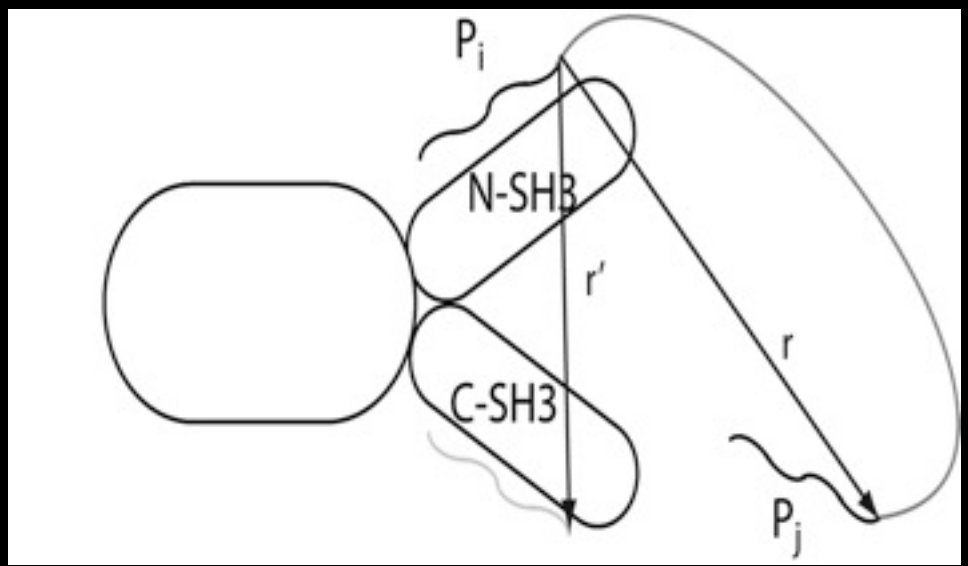
Simultaneous binding of two motifs in Sos1 to Grb2
- Hybrid model from polymer theory and MD Simulations.

Binding of Grb2 to Sos1
- Thermodynamic modeling of Grb2-Sos1 complexes.

A Sethi, et al., PLoS Comp. Biol., 2011

Signaling Cascades: Quantifying Multivalent Binding

The flexibility of both binding partners determine C_{eff}.

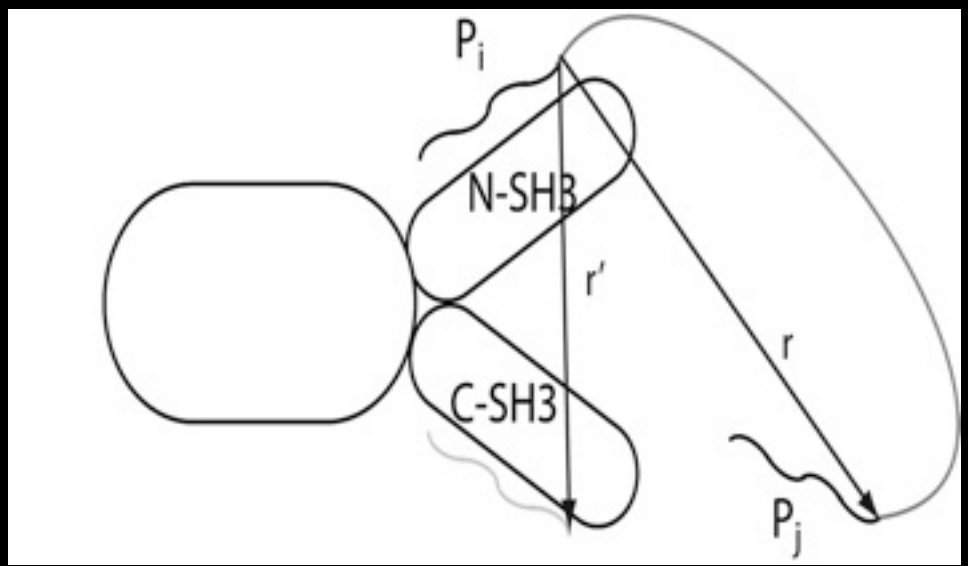


$$C_{eff}\left(P_i^N, P_j{}^C\right) = \int_{r=0}^{\infty}\int_{r'=0}^{\infty} p_{bs}(\vec{r}')p_{pep}\left(P_i^N, P_j^C, \vec{r}\right)\delta(\vec{r}-\vec{r}')d^3r d^3r'$$

$$= \int_{r=0}^{\infty} p_{bs}(\vec{r})p_{pep}\left(P_i^N, P_j^C, \vec{r}\right)\, d^3r$$

Signaling Cascades: Quantifying Multivalent Binding

The flexibility of both binding partners determine C_{eff}.



$$C_{eff}\left(P_i^N, P_j{}^C\right) = \int_{r=0}^{\infty}\int_{r'=0}^{\infty} p_{bs}(\vec{r}')\, p_{pep}\left(P_i^N, P_j^C, \vec{r}\right)\delta(\vec{r}-\vec{r}')\, d^3r\, d^3r'$$

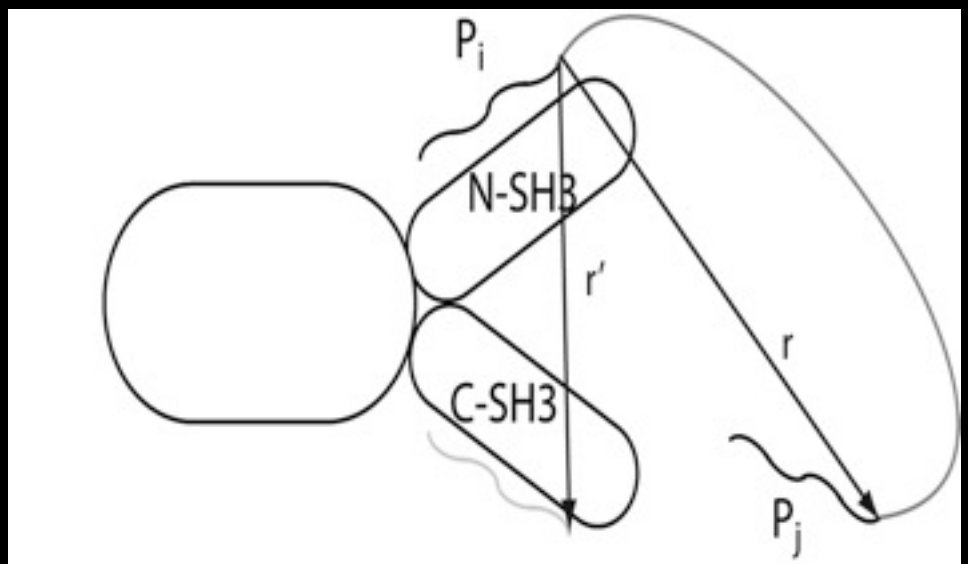
$$= \int_{r=0}^{\infty} p_{bs}(\vec{r})\, p_{pep}\left(P_i^N, P_j^C, \vec{r}\right)\, d^3r$$

Worm-like chain model for linker

$$p_{pep}\left(P_i^N, P_j^C, r\right) = \left(\frac{3}{4\pi l_p l_c}\right)^{3/2} \times \exp\left(\frac{-3r^2}{4 l_p l_c}\right)$$

Signaling Cascades: Quantifying Multivalent Binding

The flexibility of both binding partners determine C_{eff}.



$$C_{eff}\left(P_i^N, P_j{}^C\right) = \int_{r=0}^{\infty}\int_{r'=0}^{\infty} p_{bs}(\vec{r}')p_{pep}\left(P_i^N, P_j^C, \vec{r}\right)\delta(\vec{r}-\vec{r}')d^3r d^3r'$$

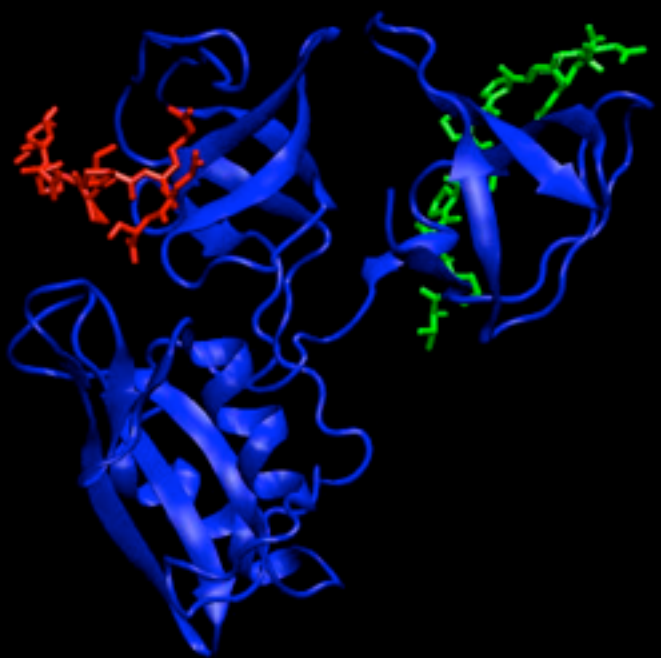
$$= \int_{r=0}^{\infty} p_{bs}(\vec{r})p_{pep}\left(P_i^N, P_j^C, \vec{r}\right)\, d^3r$$

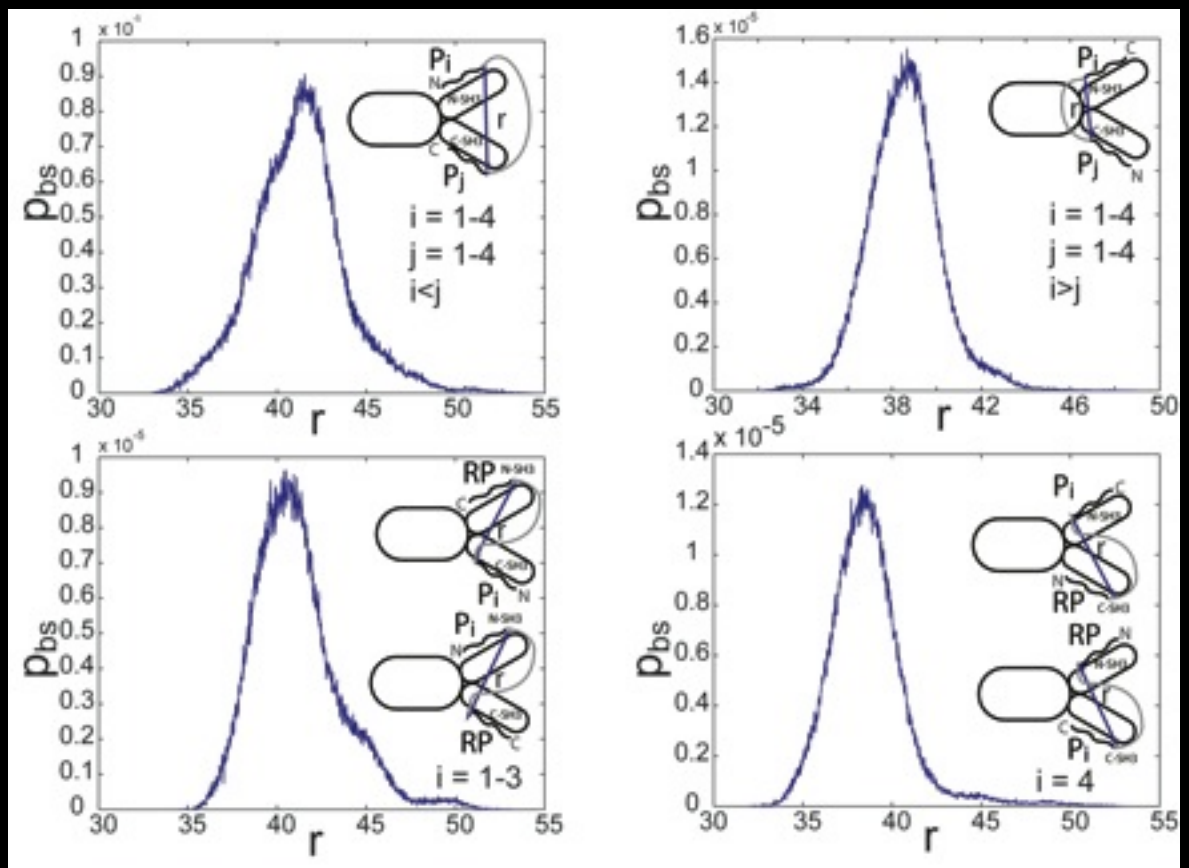
Worm-like chain model for linker

$$p_{pep}\left(P_i^N, P_j^C, r\right) = \left(\frac{3}{4\pi l_p l_c}\right)^{3/2} \times \exp\left(\frac{-3r^2}{4l_p l_c}\right)$$

Signaling Cascades: Quantifying Multivalent Binding

The flexibility of the linker and the motion of the two domains with respect to each other influence C_{eff}





A Sethi, et al., PLoS Comp. Biol., 2011

Signaling Cascades: Quantifying Multivalent Binding

The local concentration of the other motifs near the second binding site of Grb2 is in mM range

Motif bound to N-SH3 domain

Motif bound to C-SH3 domain	P1	P2	P3	RP	P4
P1	-	0.6	2.1	1.6	1.6
P2	0.3	-	0.7	1.7	1.9
P3	1.6	0.4	-	1.5	2.1
RP	1.6	1.7	1.5	-	0.07
P4	1.4	1.6	1.7	0.07	-

C_{eff} (mM)

A Sethi, et al., PLoS Comp. Biol., 2011

Signaling Cascades: Quantifying Multivalent Binding

The local concentration of the other motifs near the second binding site of Grb2 is in mM range

Motif bound to N-SH3 domain

Motif bound to C-SH3 domain	P1	P2	P3	RP	P4
P1	-	0.6	2.1	1.6	1.6
P2	0.3	-	0.7	1.7	1.9
P3	1.6	0.4	-	1.5	2.1
RP	1.6	1.7	1.5	-	0.07
P4	1.4	1.6	1.7	0.07	-

C_{eff} (mM)

Motif bound to N-SH3 domain

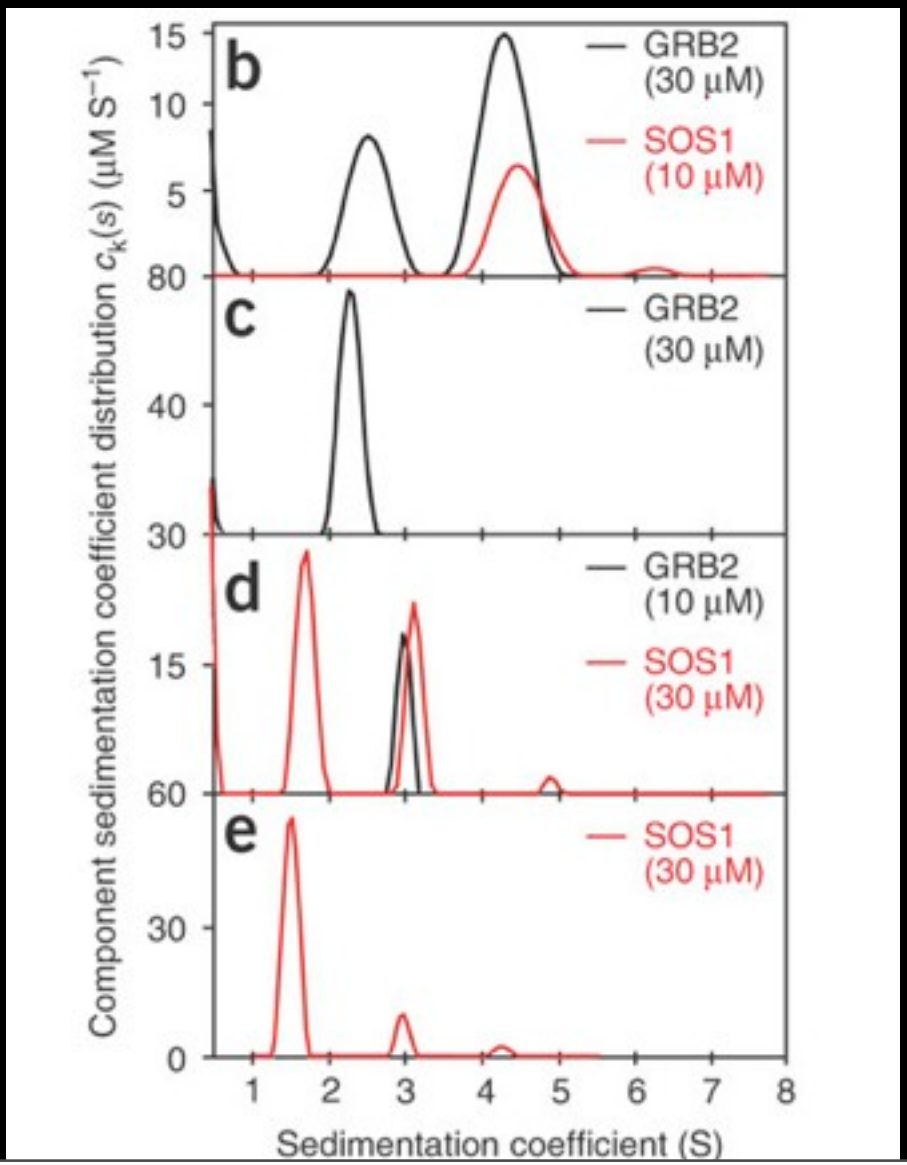
Motif bound to C-SH3 domain	P1	P2	P3	RP	P4
P1	-	12	7	9	7
P2	164	-	220	94	62
P3	42	220	-	133	67
RP	43	56	130	-	2100
P4	37	46	90	2300	-

$1/\bar{K}_{ij}^{NC}(\mu M)$

A Sethi, et al., PLoS Comp. Biol., 2011

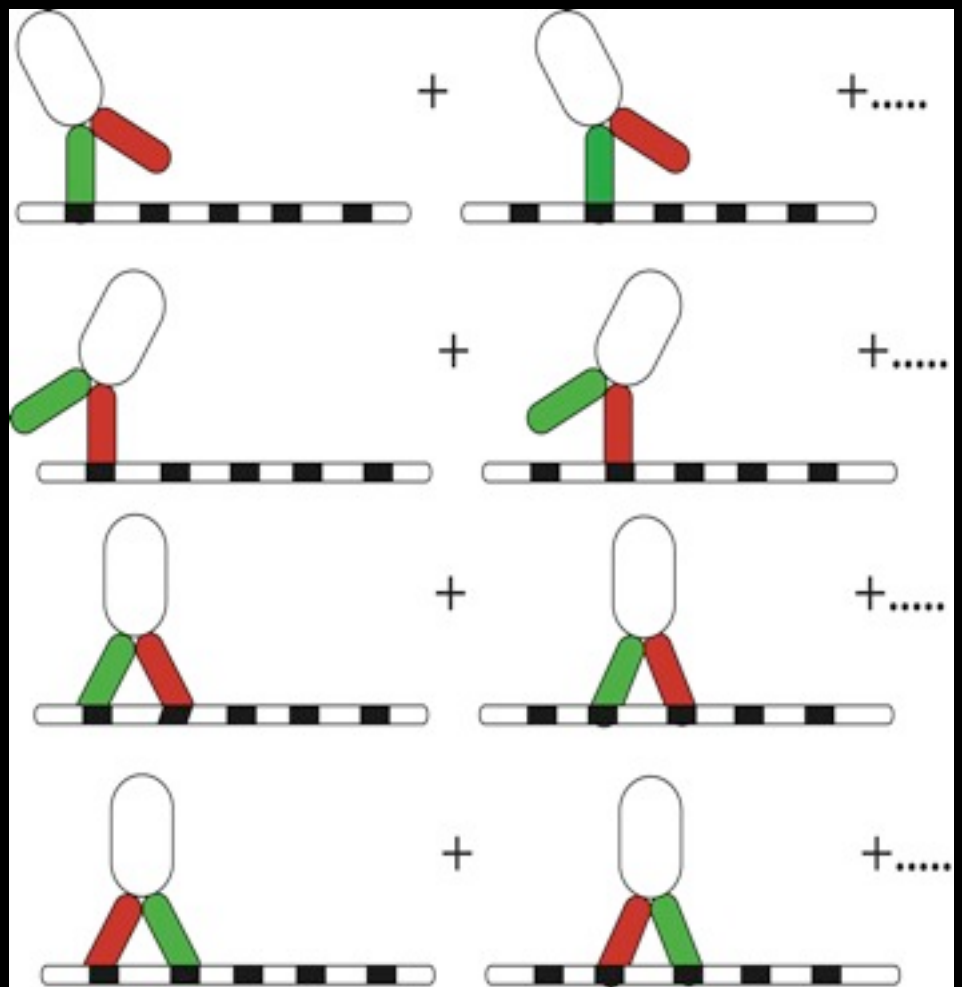
Signaling Cascades: Quantifying Multivalent Binding

It is very difficult to distinguish different complexes formed between Sos1 and Grb2 in experiments



Signaling Cascades: Quantifying Multivalent Binding

The 1:1 complex is actually a large combination of complexes that should be taken into account.

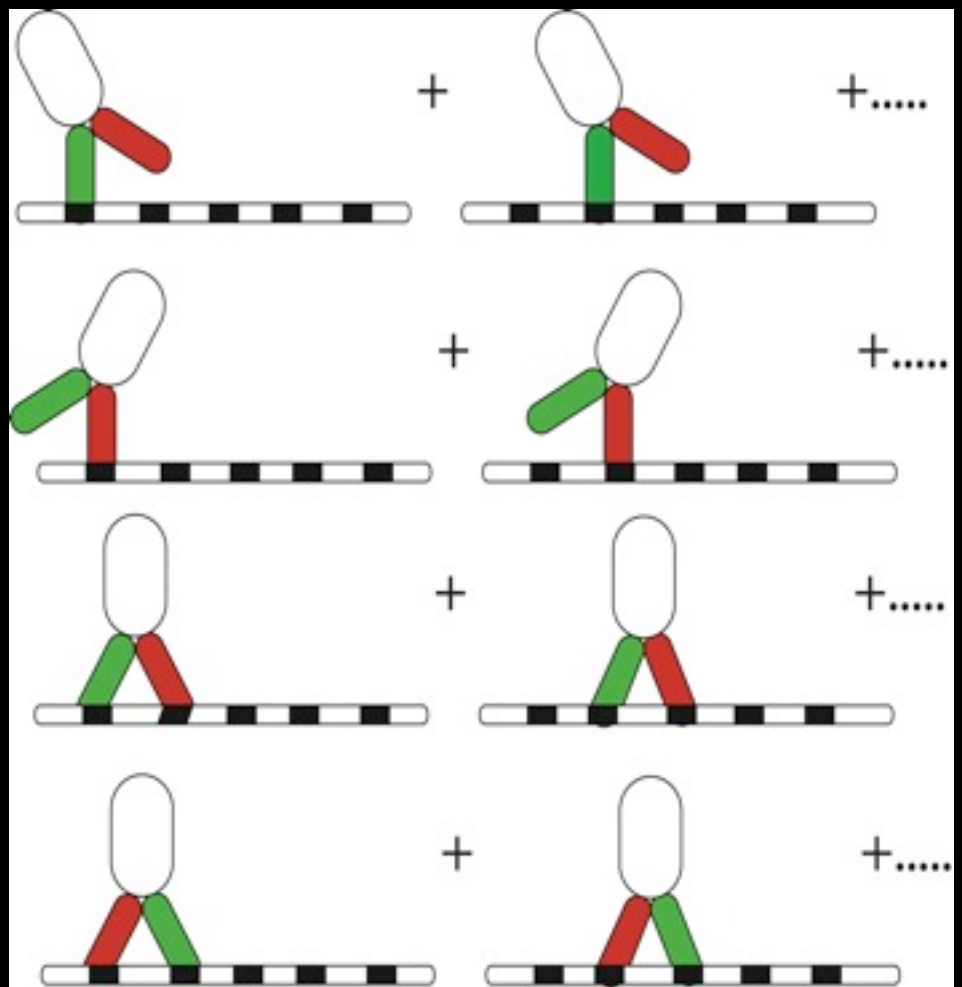


$$\bar{K}_1 = \sum_{X=N,C} \sum_{i=1}^{5} K_i^X + \sum_{i=1}^{5} \sum_{j=1, j\neq i}^{5} \bar{K}_{ij}^{NC}$$

A Sethi, et al., PLoS Comp. Biol., 2011

Signaling Cascades: Quantifying Multivalent Binding

The 1:1 complex is actually a large combination of complexes that should be taken into account.

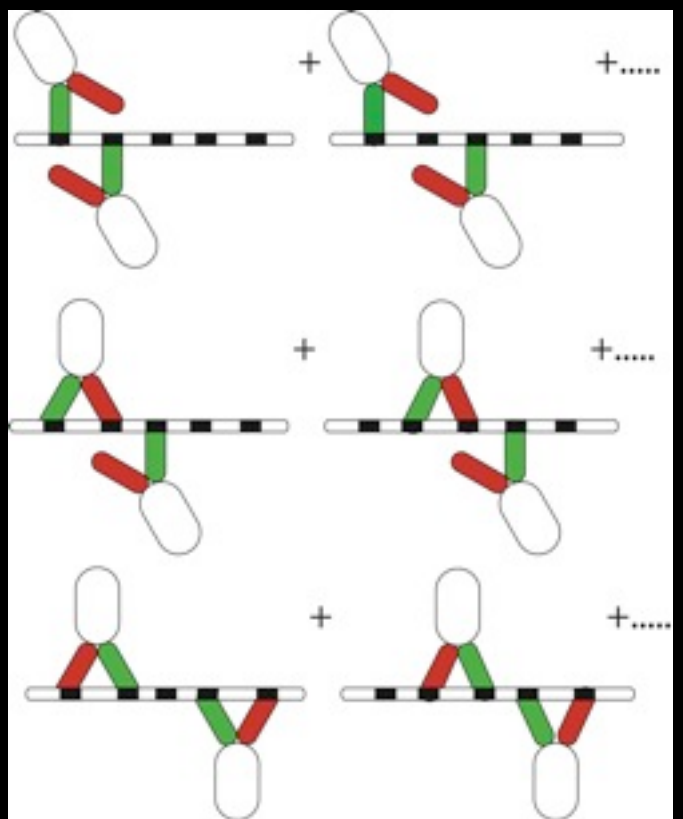


$$\bar{K}_1 = \sum_{X=N,C} \sum_{i=1}^{5} K_i^X + \sum_{i=1}^{5} \sum_{j=1, j\neq i}^{5} \bar{K}_{ij}^{NC}$$

Predicted value = 1.2 μM.
Experimental value = 0.3 μM.

A Sethi, et al., PLoS Comp. Biol., 2011

The 2:1 complex contains an even larger combination of complexes.



Predicted value = 14 μM.
Experimental value = 1 μM.

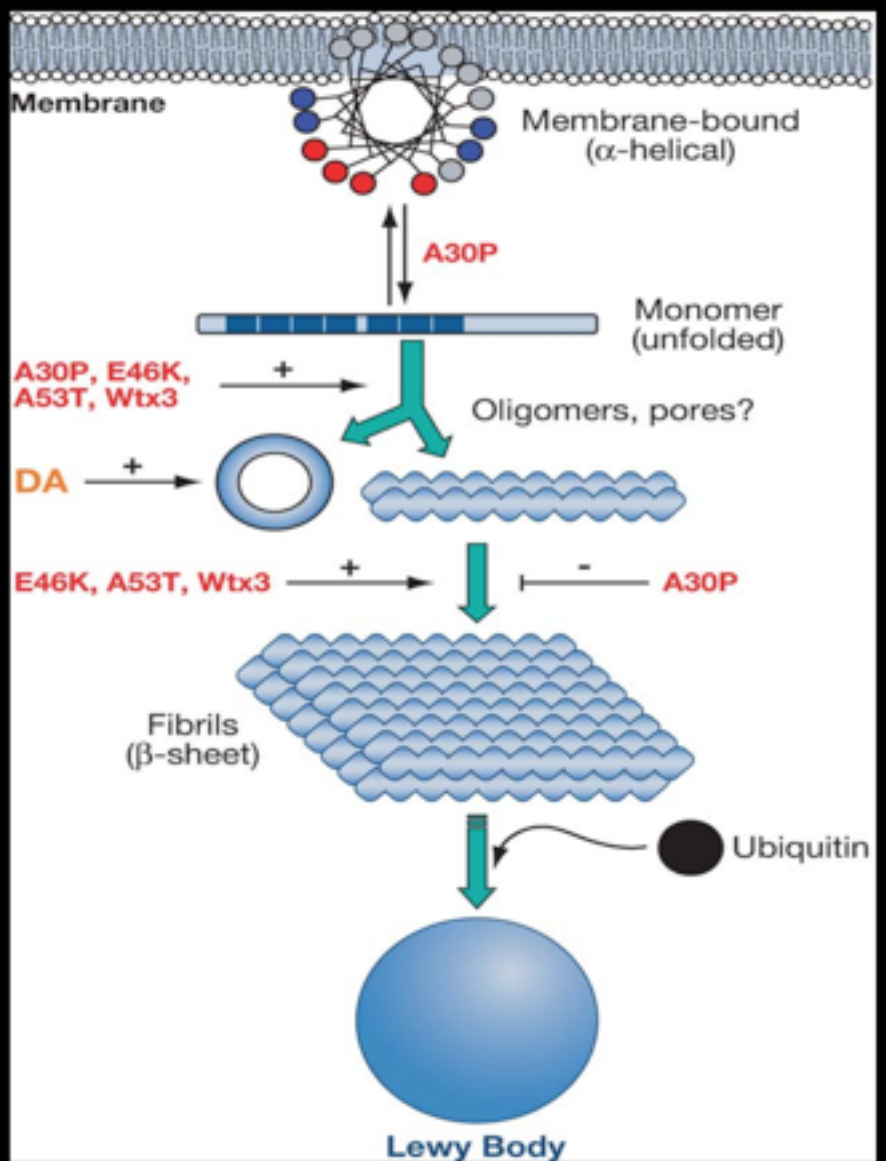
$$\bar{K}_2 = \frac{1}{2\bar{K}_1}\Bigg(\sum_{X,Y=N,C} \sum_{i=1}^{5} \sum_{j=1,j\neq i}^{5} K_i^X K_j^Y + \sum_{X=N,C} \sum_{i=1}^{5} \sum_{j=1}^{5} \sum_{k=1,k\neq j\neq i}^{5} K_i^X \bar{K}_{jk}^{NC} + \sum_{i=1}^{5} \sum_{j=1}^{5} \sum_{k=1}^{5} \sum_{l=1,l\neq k\neq j\neq i}^{5} \bar{K}_{ij}^{NC} \bar{K}_{kl}^{NC} \Bigg)$$

A Sethi, et al., PLoS Comp. Biol., 2011

Is a WLC model a good model to get the probability of the distance between the two ends of a linker?
Can we use MD-based methods to figure this probability density?

Signaling Cascades: Disordered Regions

Fibrils or oligomers of disordered proteins are often implicated in neurological diseases



Cookson, Ann. Rev. Biochem., 2005

Signaling Cascades: Disordered Regions

Fibrils or oligomers of disordered proteins are often implicated in neurological diseases

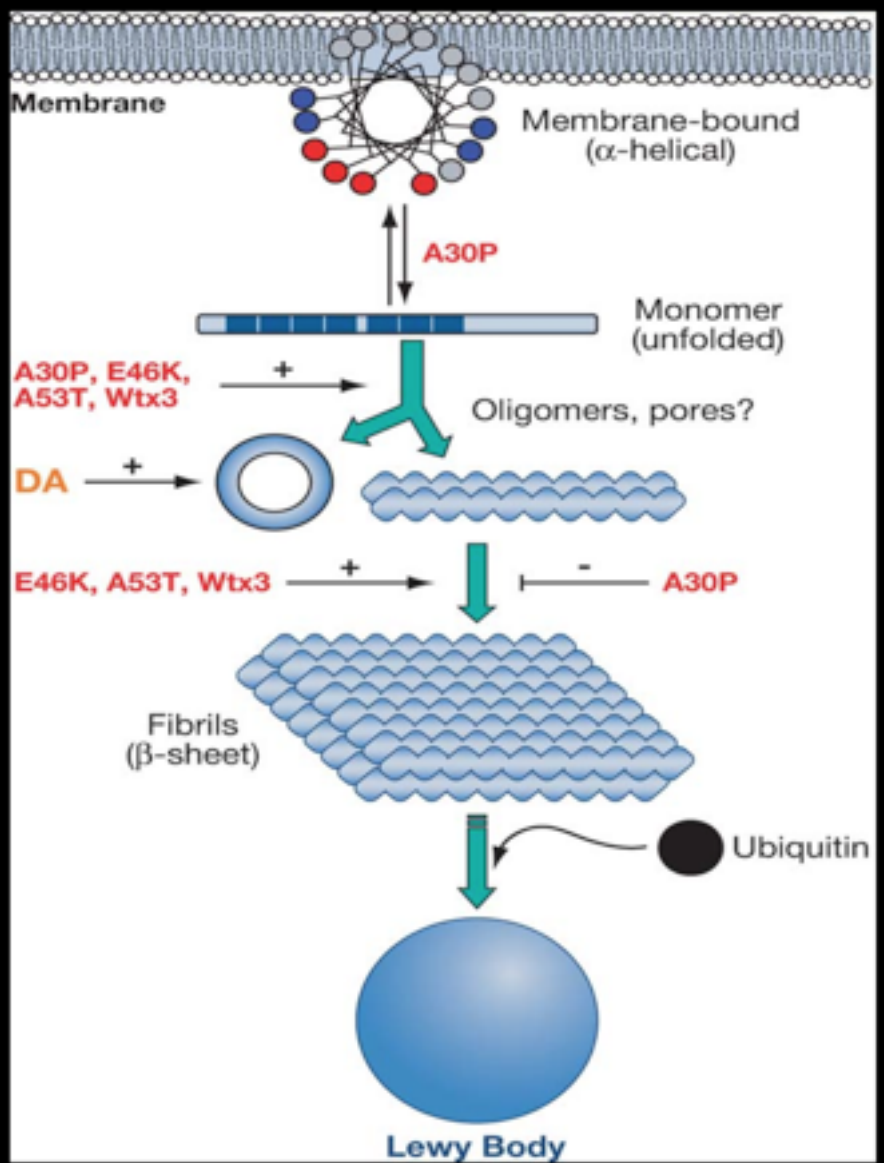
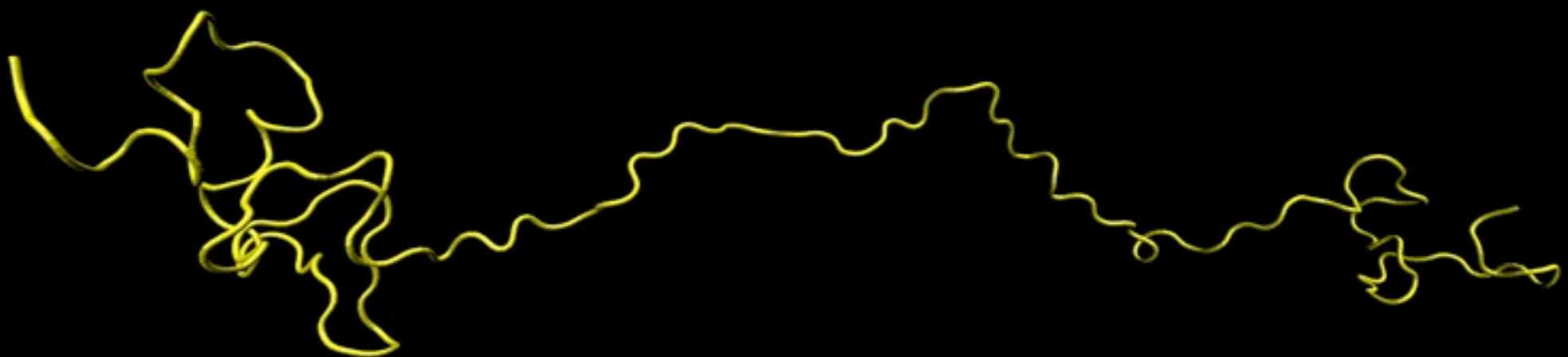


Image Source:
http://www.genome.gov/pressDisplay.cfm?photoID=10004

Cookson, Ann. Rev. Biochem., 2005

Signaling Cascades: Disordered Regions

Is divide and conquer technique possible for IDPs?

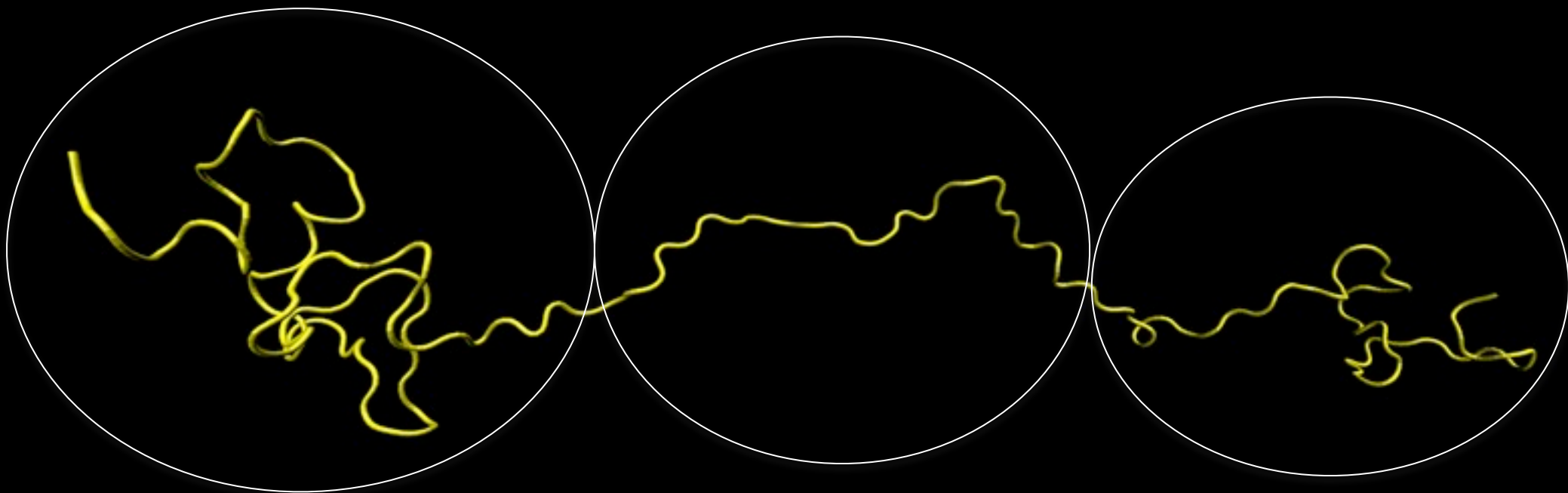


Can we break an IDP into smaller, more manageable, pieces in order to calculate it's conformational network (divide and conquer approach)?

A Sethi, et al., Biophys. J., 2012

Signaling Cascades: Disordered Regions

Is divide and conquer technique possible for IDPs?

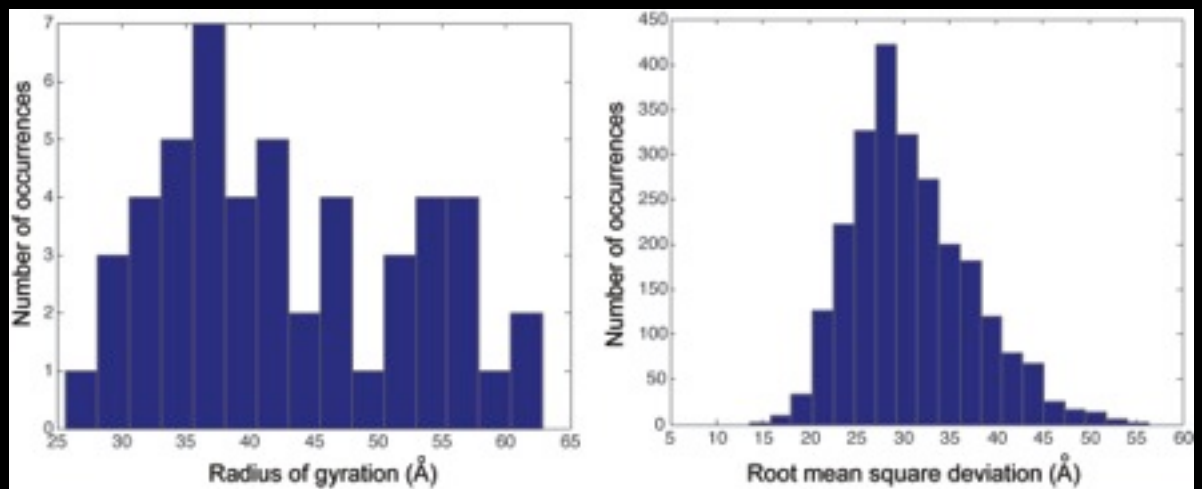


Can we break an IDP into smaller, more manageable, pieces in order to calculate it's conformational network (divide and conquer approach)?

A Sethi, et al., Biophys. J., 2012

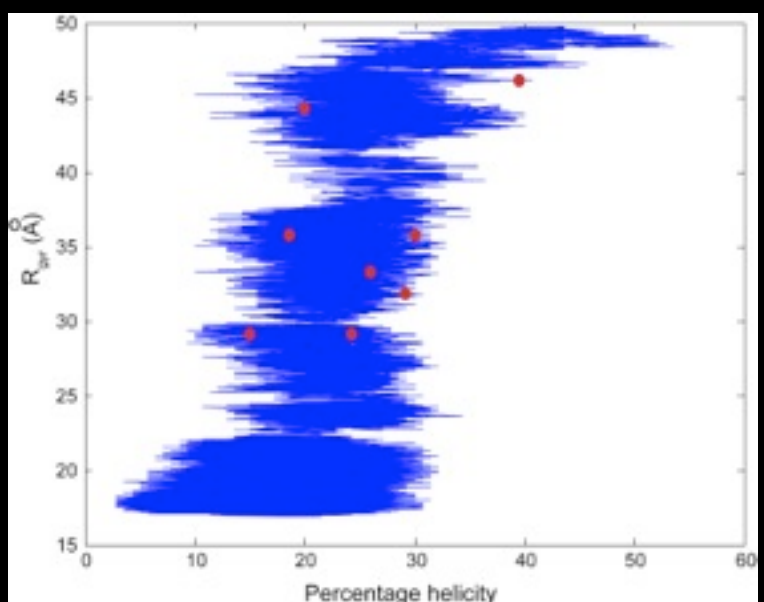
Signaling Cascades: Disordered Regions

We performed 100 simulations to sample the entire phase space of α-synuclein



50 conformations generated using random coil generator (conformations of random coils in pdb database)

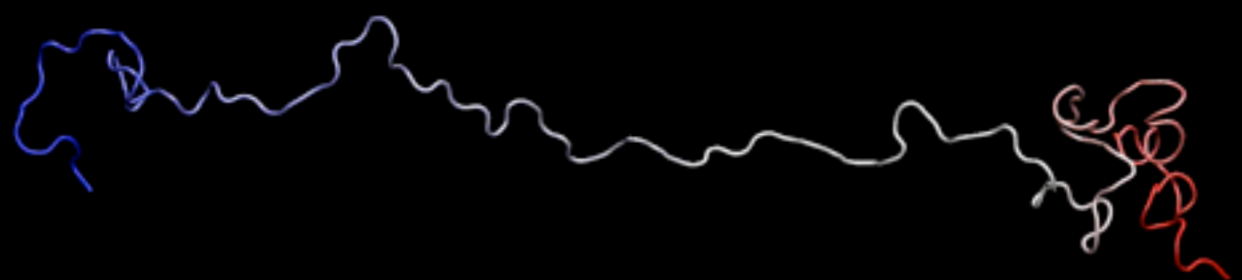
8 partially helical conformations chosen from one simulation of the helix turn helix conformation + 2 native conformations of α-synuclein = 50 partially helical simulations (five simulations starting from 10 different conformations).



A Sethi, et al., Biophys. J., 2012

Signaling Cascades: Disordered Regions

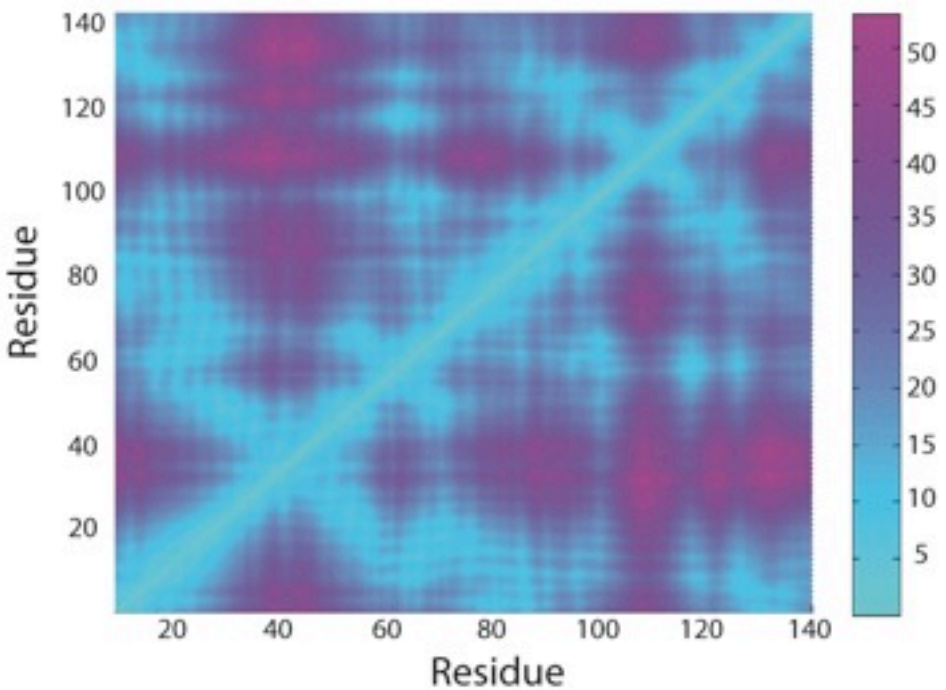
The protein collapses and remains collapsed during the timescale of these simulations.



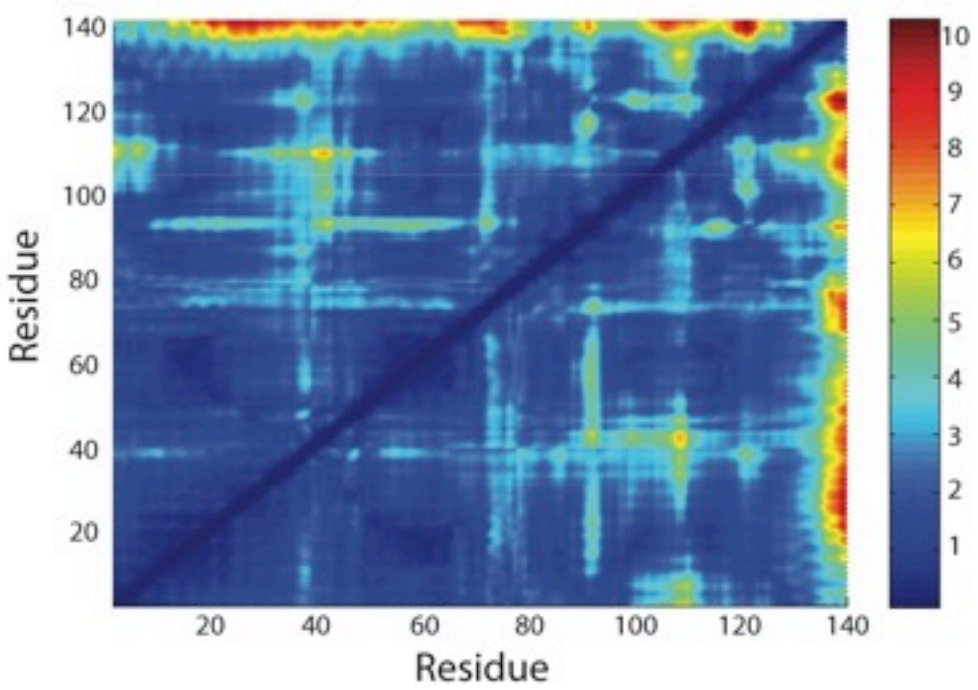
A Sethi, et al., Biophys. J., 2012

Signaling Cascades: Disordered Regions

The protein stays trapped in a minima upto hundreds of nanoseconds.



<r(i,j)> (in Angstroms) during 300ns of 298K (NMR)

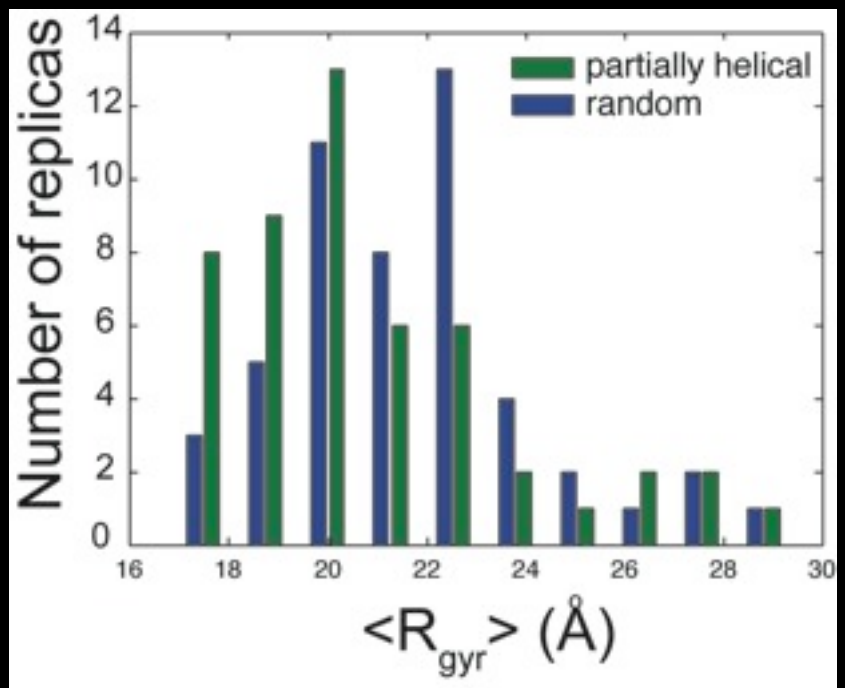


Standard deviation in r(i,j) (in Angstroms) during 300ns of the 298K (NMR) simulation.

A Sethi, et al., Biophys. J., 2012

Signaling Cascades: Disordered Regions

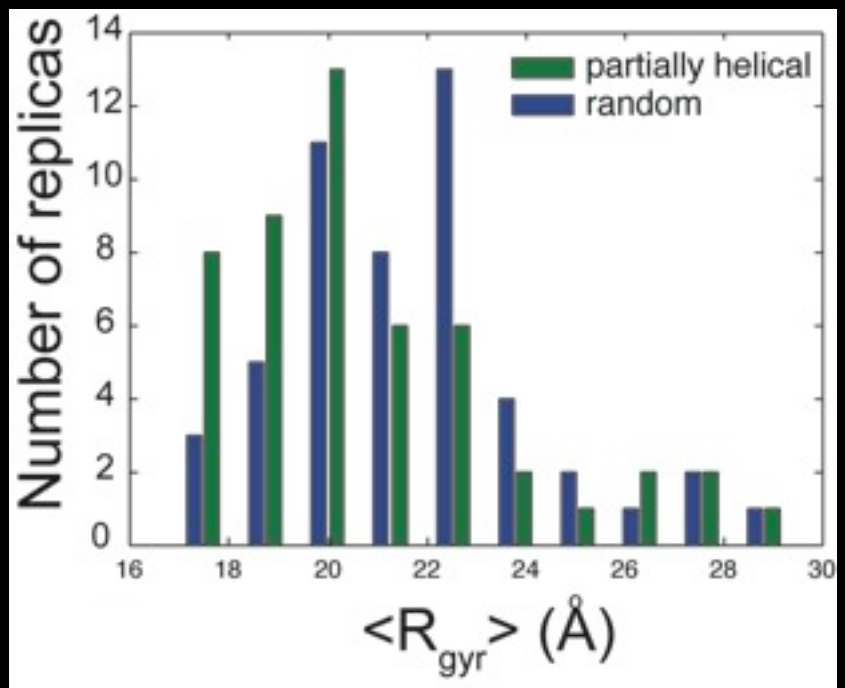
The protein forms a heterogeneous collapsed state.

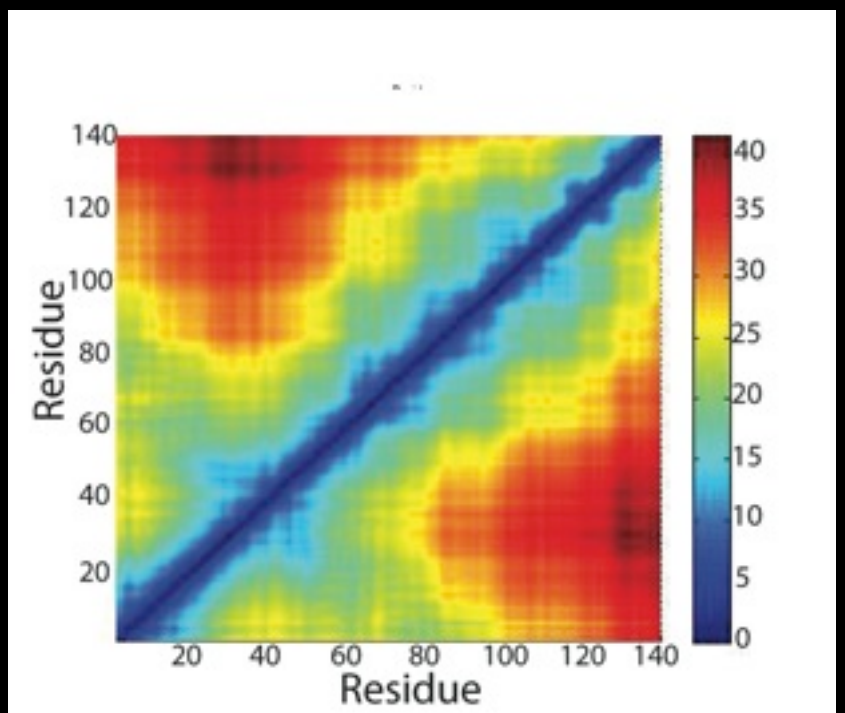


A Sethi, et al., Biophys. J., 2012

Signaling Cascades: Disordered Regions

The protein forms a heterogeneous collapsed state.



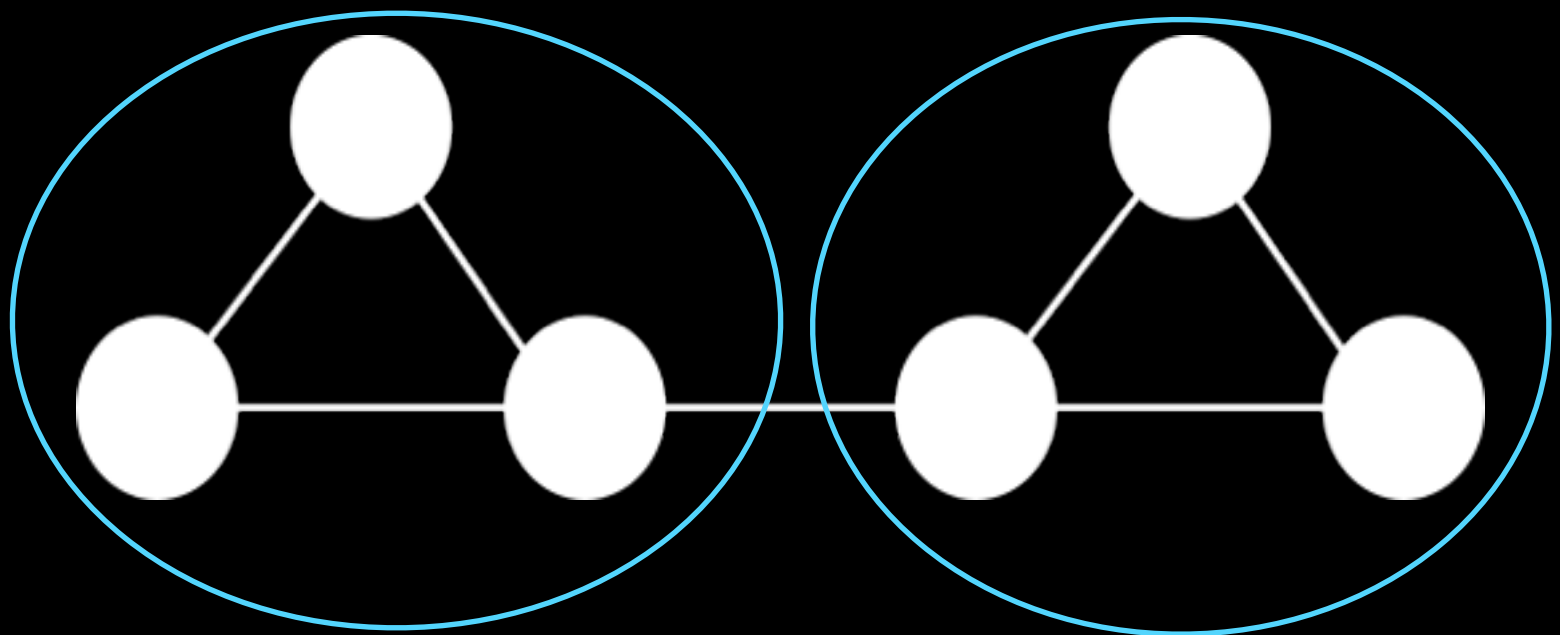


The protein behaves like an ideal chain in poor solvent.
There are no persistent long-range contacts in the protein.

A Sethi, et al., Biophys. J., 2012

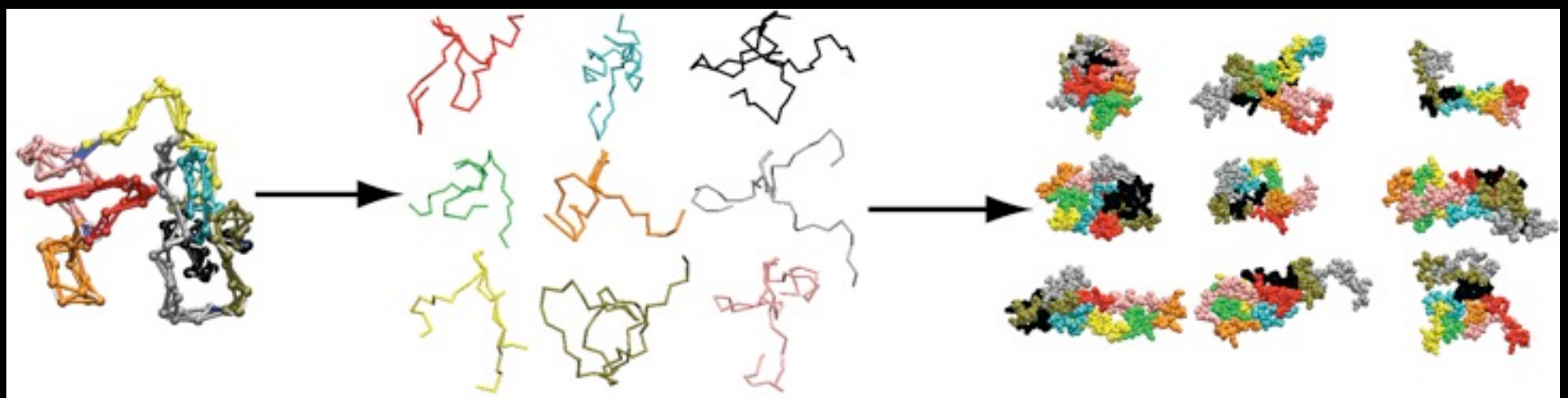
Signaling Cascades: Disordered Regions

Communities are modules within the network.



Signaling Cascades: Disordered Regions

The heterogeneous nature of the compact states may make it possible to split α-synuclein into smaller peptides

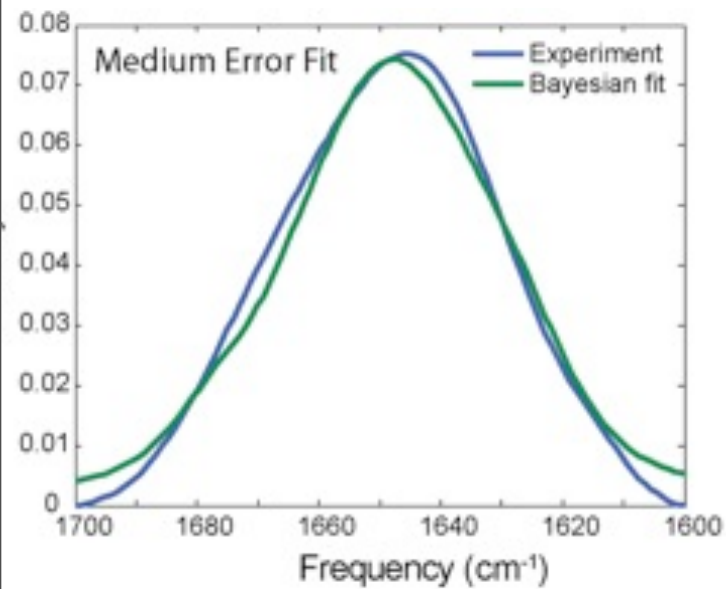


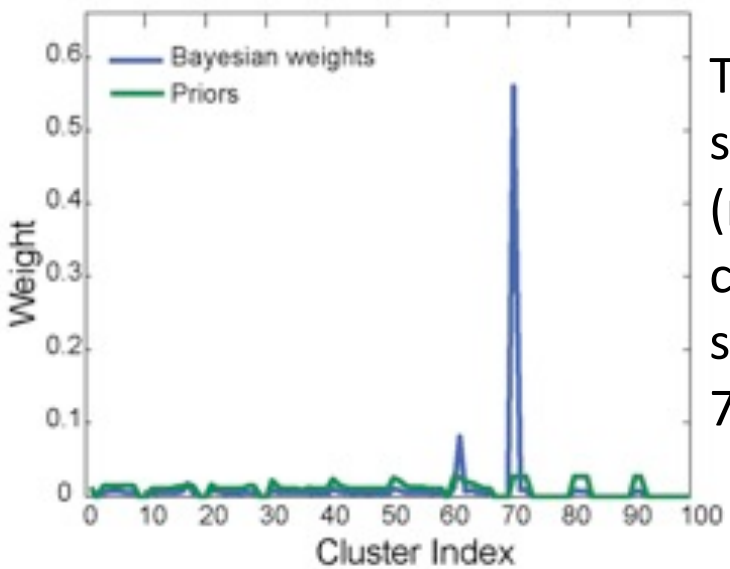
Nine communities in combined trajectory.

The communities are made of sequentially contiguous residues.

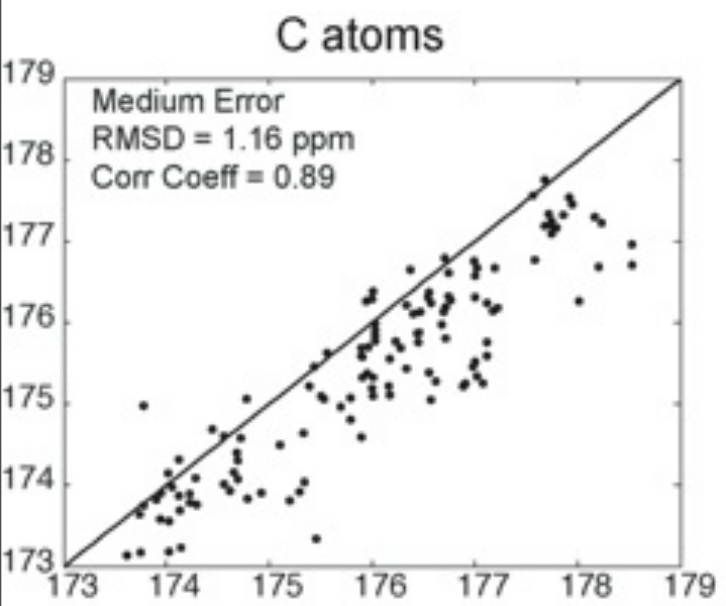
A Sethi, et al., Biophys. J., 2012

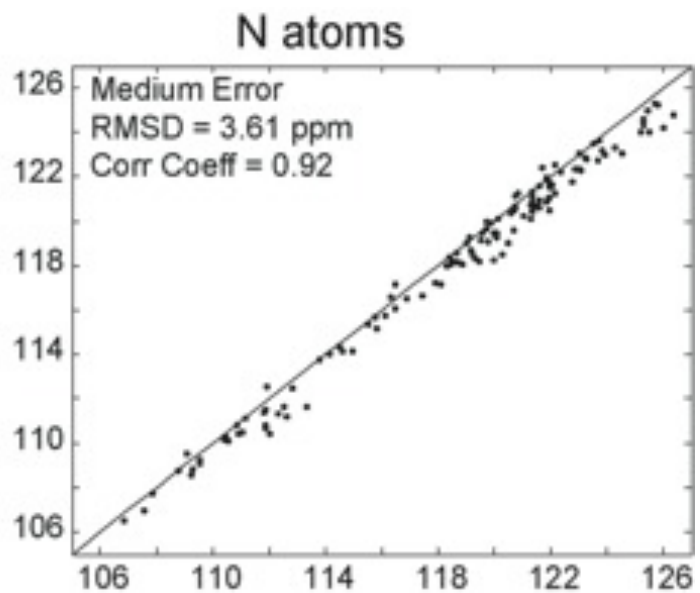
Bayesian Statistics of Intrinscially Disordered Proteins





The Bayesian ensemble of α-synuclein is semi-collapsed (mean of 35 Angstroms) and contains very low residual secondary structure (mean of 7% helicity)





The predicted chemical shifts for the Bayesian ensemble are consistent with experimental observations

Sethi et al., Chem. Phys., In Press.

Conclusions

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

The distribution of the multivalent complexes further increases the effective equilibrium constants for complexes of different stoichiometry to form.

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

The distribution of the multivalent complexes further increases the effective equilibrium constants for complexes of different stoichiometry to form.

IDPs can exist in a heterogeneous collapsed state in aqueous environment.

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

The distribution of the multivalent complexes further increases the effective equilibrium constants for complexes of different stoichiometry to form.

IDPs can exist in a heterogeneous collapsed state in aqueous environment.

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

The distribution of the multivalent complexes further increases the effective equilibrium constants for complexes of different stoichiometry to form.

IDPs can exist in a heterogeneous collapsed state in aqueous environment.

The collapsed states are stabilized by contacts that remain for hundreds of nanoseconds.

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

The distribution of the multivalent complexes further increases the effective equilibrium constants for complexes of different stoichiometry to form.

IDPs can exist in a heterogeneous collapsed state in aqueous environment.

The collapsed states are stabilized by contacts that remain for hundreds of nanoseconds.

Conclusions

The local concentration in multivalent complexes can be quantified using a combination of polymer models and MD simulations.

The distribution of the multivalent complexes further increases the effective equilibrium constants for complexes of different stoichiometry to form.

IDPs can exist in a heterogeneous collapsed state in aqueous environment.

The collapsed states are stabilized by contacts that remain for hundreds of nanoseconds.

The heterogeneity in the collapsed states might make it possible to divide α−Synuclein into smaller fragments.

Acknowledgements

Mentors:

S. Gnanakaran (LANL)

Zan Schulten (UIUC)

Byron Goldstein (LANL)

Collaborators:

Carl Woese (UIUC) - aaRS and ribosome

Dung Vu (LANL) - IDP

Martin Gruebele (UIUC) - RNA unfolding

V Jo Davisson (Purdue) - Allostery

Cynthia Derdeyn (Emory) - HIV

Bette Korber (LANL) - HIV

Shishir Chundawat (GLBRC) - Biofuels

Peter Goodwin (LANL) - Biofuels

Bridget Wilson (UNM) - Signaling

Chang-Shung Tung (LANL) - Signaling

Patrick O'Donoghue (Yale) - aaRS

Rommie Amaro (UCSD) - Allostery

Elijah Roberts (JHU) - Ribosome

John Eargle (UIUC) - aaRS

Jianhui Tian (ORNL) - IDP/HIV