Integrative Structure Modeling

Marc A. Marti-Renom

centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

cnag

Summary

- **CONCEPTS (10’)**
    - ➡ Data groups
    - ➡ Stages
    - ➡ Advantages

- **EXAMPLES (remaining!)**
    - ➡ Proteins
    - ➡ Complexes of proteins
    - ➡ Genomes

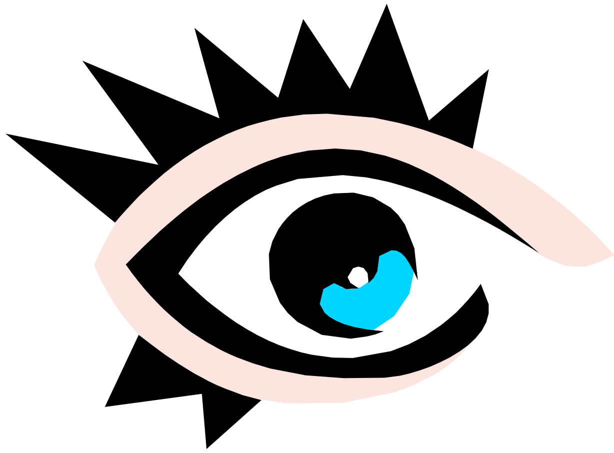
DISCLAIMER!

IMP

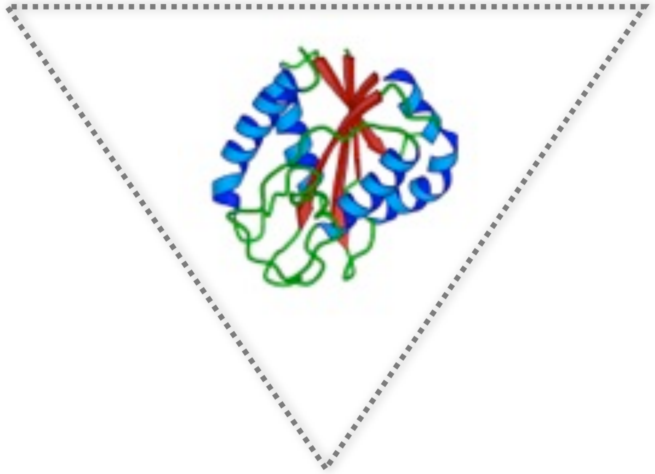
Integrative Modeling Platform

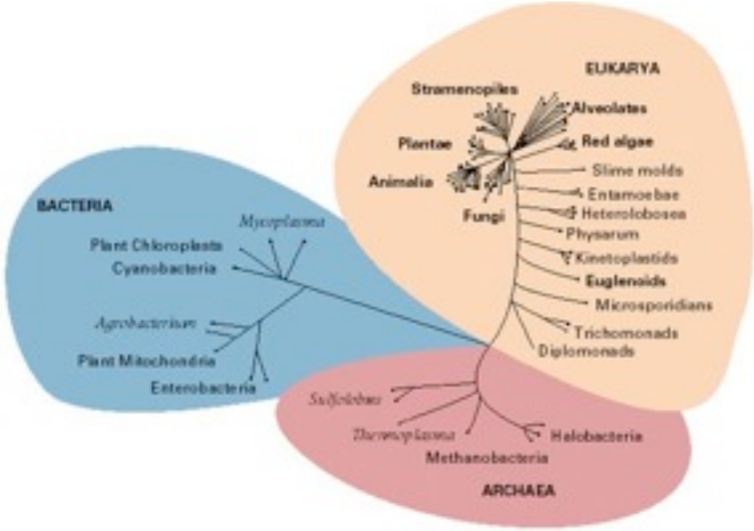
http://integrativemodeling.org

Data groups

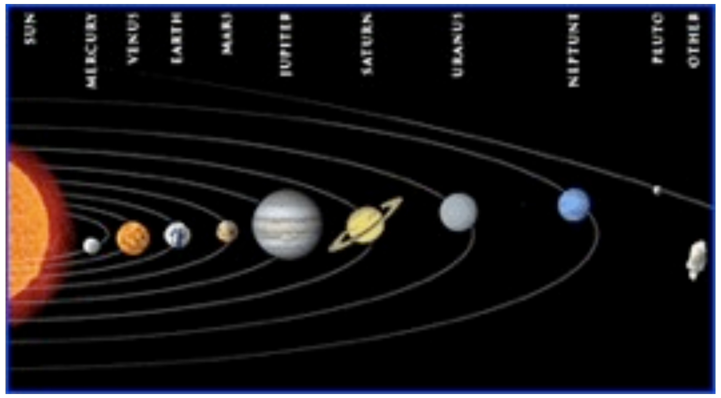


Experimental observations





Statistical rules



Laws of physics

Stages

Stage 1: Gathering Information. Information is collected in the form of data from wet lab experiments, as well as statistical tendencies such as atomic statistical potentials, physical laws such as molecular mechanics force fields, and any other feature that can be converted into a score for use to assess features of a structural model.

Stage 2: Choosing How To Represent And Evaluate Models. The resolution of the representation depends on the quantity and resolution of the available information and should be commensurate with the resolution of the final models: different parts of a model may be represented at different resolutions, and one part of the model may be represented at several different resolutions simultaneously. The scoring function evaluates whether or not a given model is consistent with the input information, taking into account the uncertainty in the information.

Stage 3: Finding Models That Score Well. The search for models that score well is performed using any of a variety of sampling and optimization schemes (such as the Monte Carlo method). There may be many models that score well if the data are incomplete or none if the data are inconsistent due to errors or unconsidered states of the assembly.

Stage 4: Analyzing Resulting Models and Information. The ensemble of good-scoring models needs to be clustered and analyzed to ascertain their precision and accuracy, and to check for inconsistent information. Analysis can also suggest what are likely to be the most informative experiments to perform in the next iteration.

Integrative modeling iterates through these stages until a satisfactory model is built. Many iterations of the cycle may be required, given the need to gather more data as well as to resolve errors and inconsistent data.

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., et al. (2012). *PLoS Biology*, *10*(1), e1001244

Advantages

Using New Information. Integrative modeling makes it easy to take advantage of new information and new types of information, resulting in a low barrier for using incremental information that is generally not applied to structure characterization. Even when a single data type is relatively uninformative, multiple types can give a surprisingly complete picture of an assembly [9,10].

Maximizing Accuracy, Precision and Completeness. Integrative models fit multiple types of information, and can thus be more accurate, precise, and complete than models based on the individual sources.

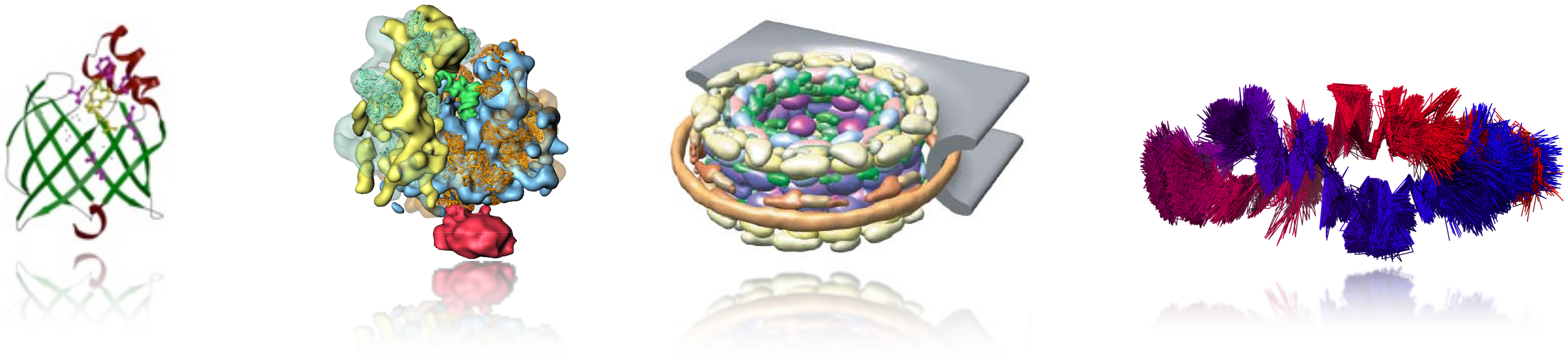
Understanding and Assessing the Models. By exhaustively sampling the space of models fitting the information, integrative modeling can find all models fitting the information, not only one. A full sampling of the models of a structure can improve the understanding of its function [49]. Because the data are encoded in scoring functions and the full set of models can be found, integrative modeling facilitates assessing the input information and output models in terms of precision and accuracy.

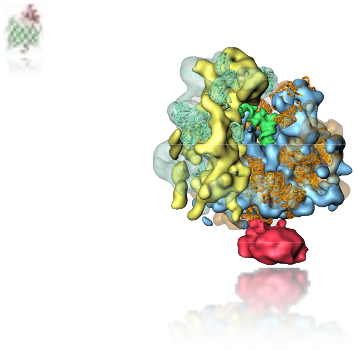
Planning Experiments. Integrative modeling provides feedback to guide future experiments, by computationally testing the impact of hypothetical datasets. As a result, experiments can be chosen to best improve our knowledge of the assembly.

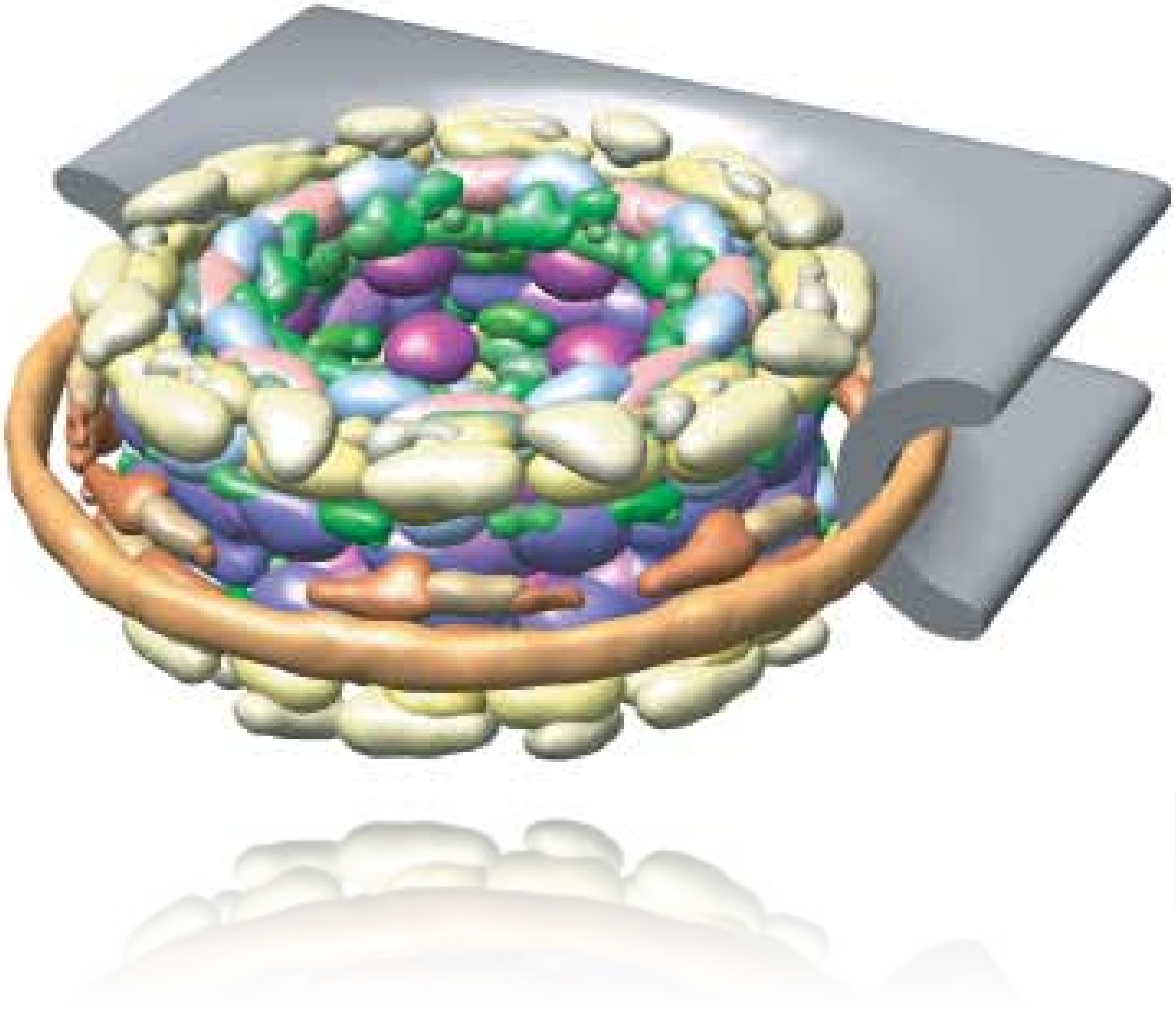
Understanding and Assessing Experimental Accuracy. Data errors present a challenge for all methods of model building. Integrative modeling can detect inconsistent data as no models will exist that fit all the data. In addition, integrative modeling facilitates the application of more sophisticated methods for error estimation, such as Inferential Structure Determination [16].

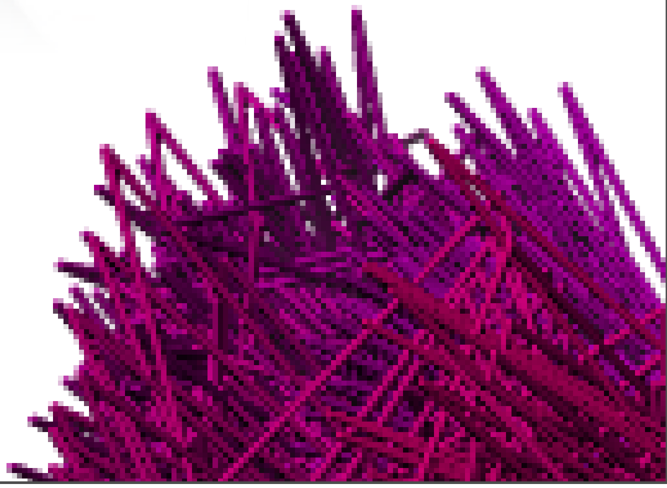
Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., et al. (2012). *PLoS Biology*, *10*(1), e1001244

Data integration

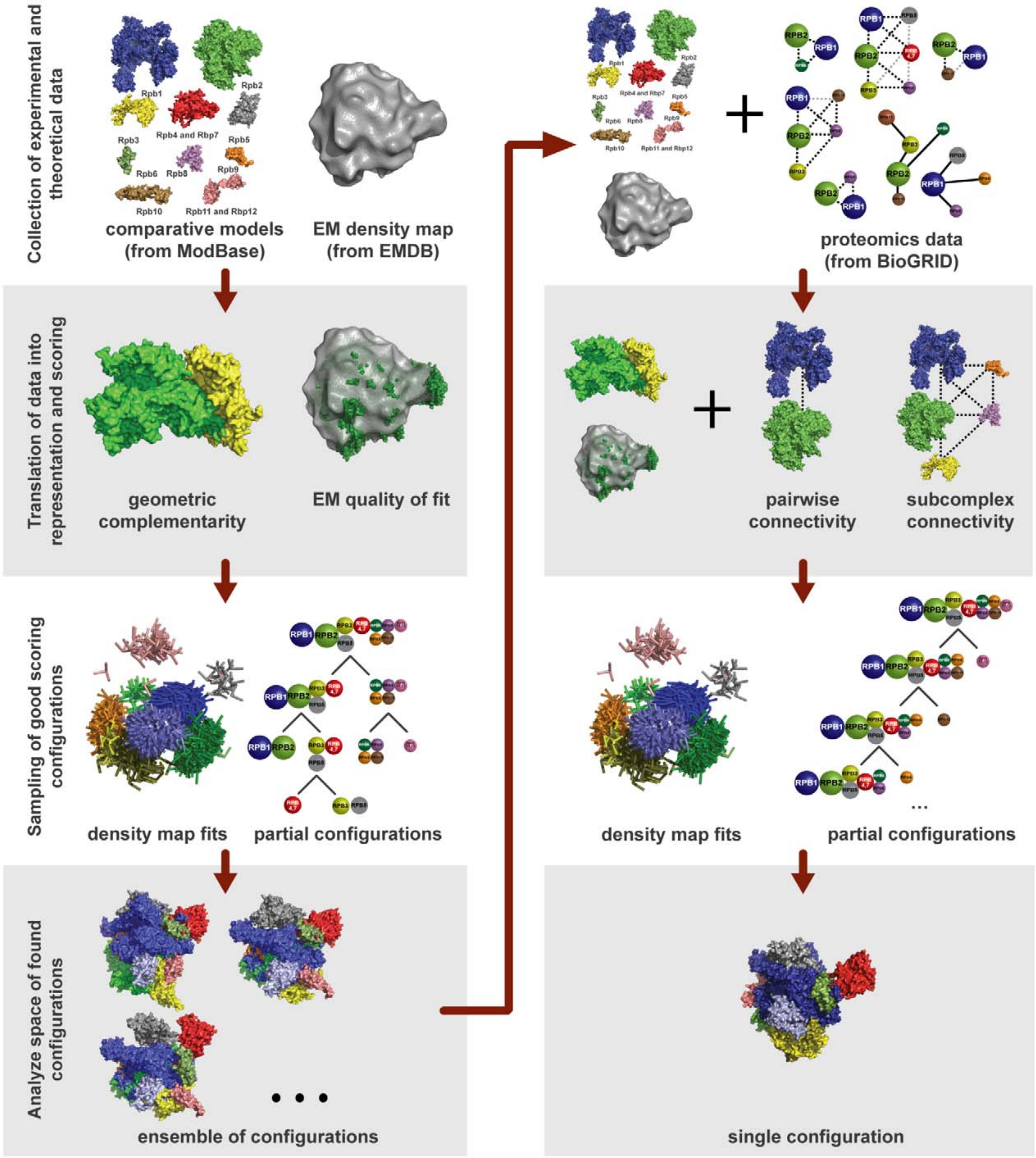








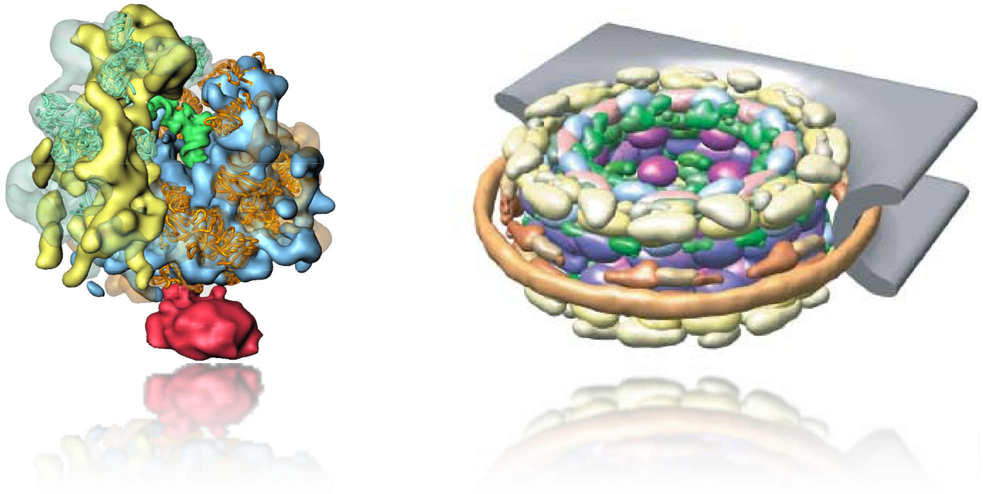
“Toy” example...



Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., et al. (2012). *PLoS Biology*, *10*(1), e1001244

“Real” examples

PROTEINS

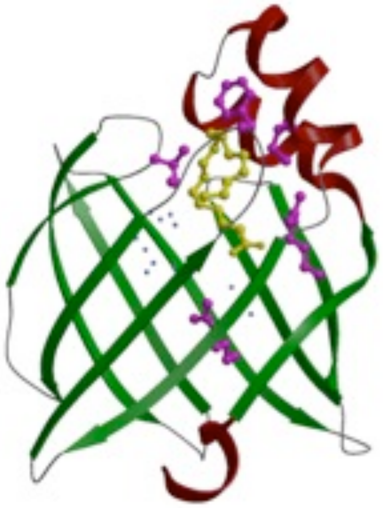


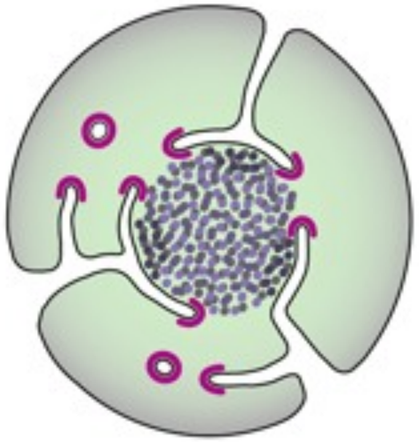
COMPLEXES

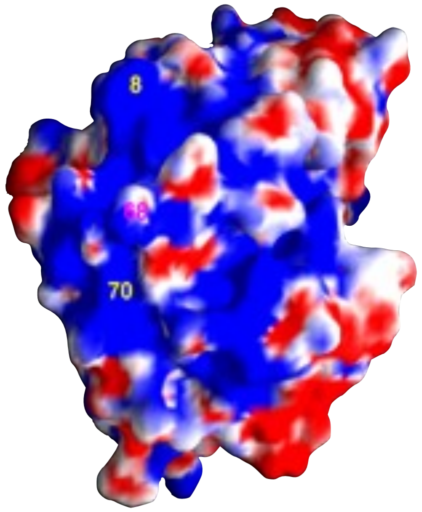
GENOMES

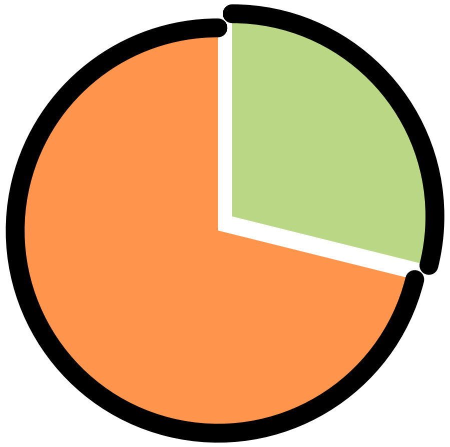
PROTEINS

single *data type*



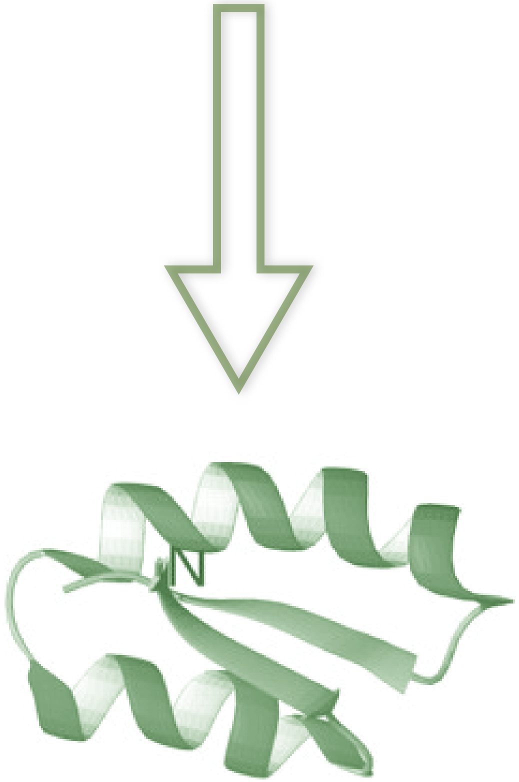


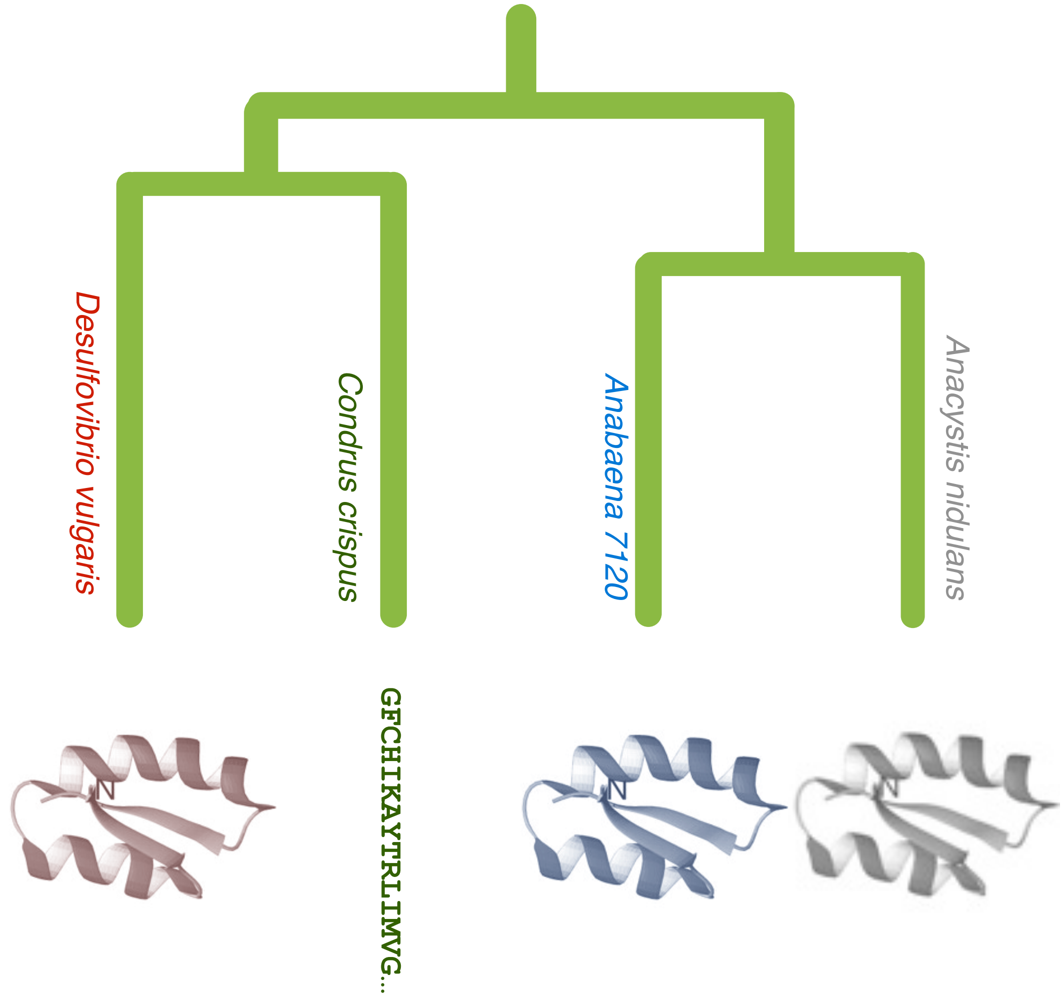




Principles of protein structure

GFCHIKAYTRLIMVG...





Folding (physics)

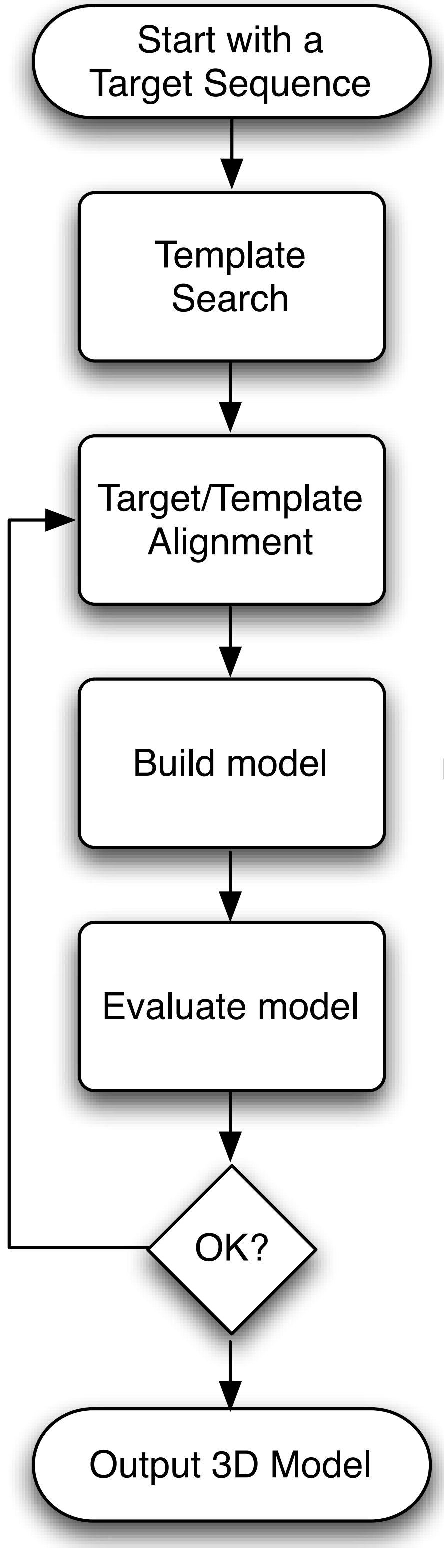
Ab initio prediction

Evolution (rules)

Threading
Comparative Modeling

D. Baker & A. Sali. Science 294, 93, 2001.

Comparative modeling by satisfaction of spatial restraints



Start with a Target Sequence

Template Search

Target/Template Alignment

Build model

Evaluate model

OK?

Output 3D Model

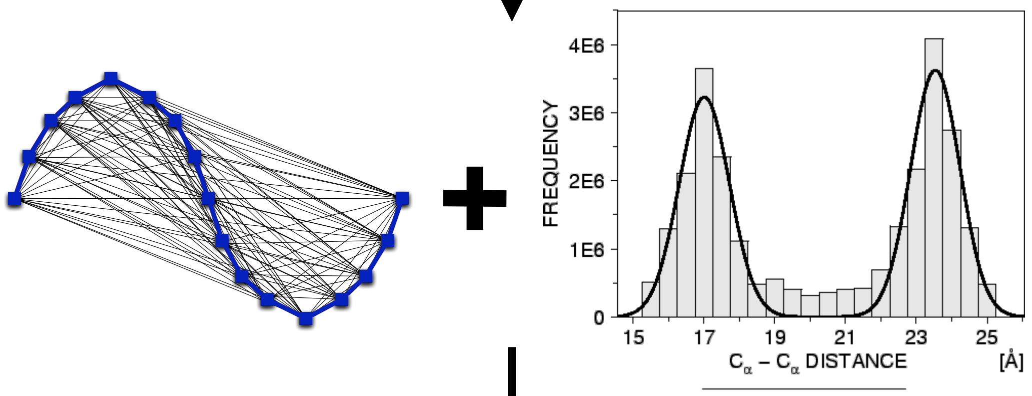
Given an alignment...

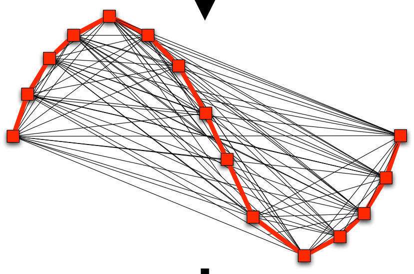
extract spatial features from the template(s) and statistics from known structures

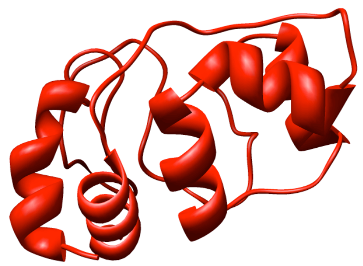
apply these features as restraints on your target sequence

optimize to find the best solution for the restraints to produce your 3D model

MSVIPKR--GNCEQTSE
ASILPKRLFGNCEQTSD

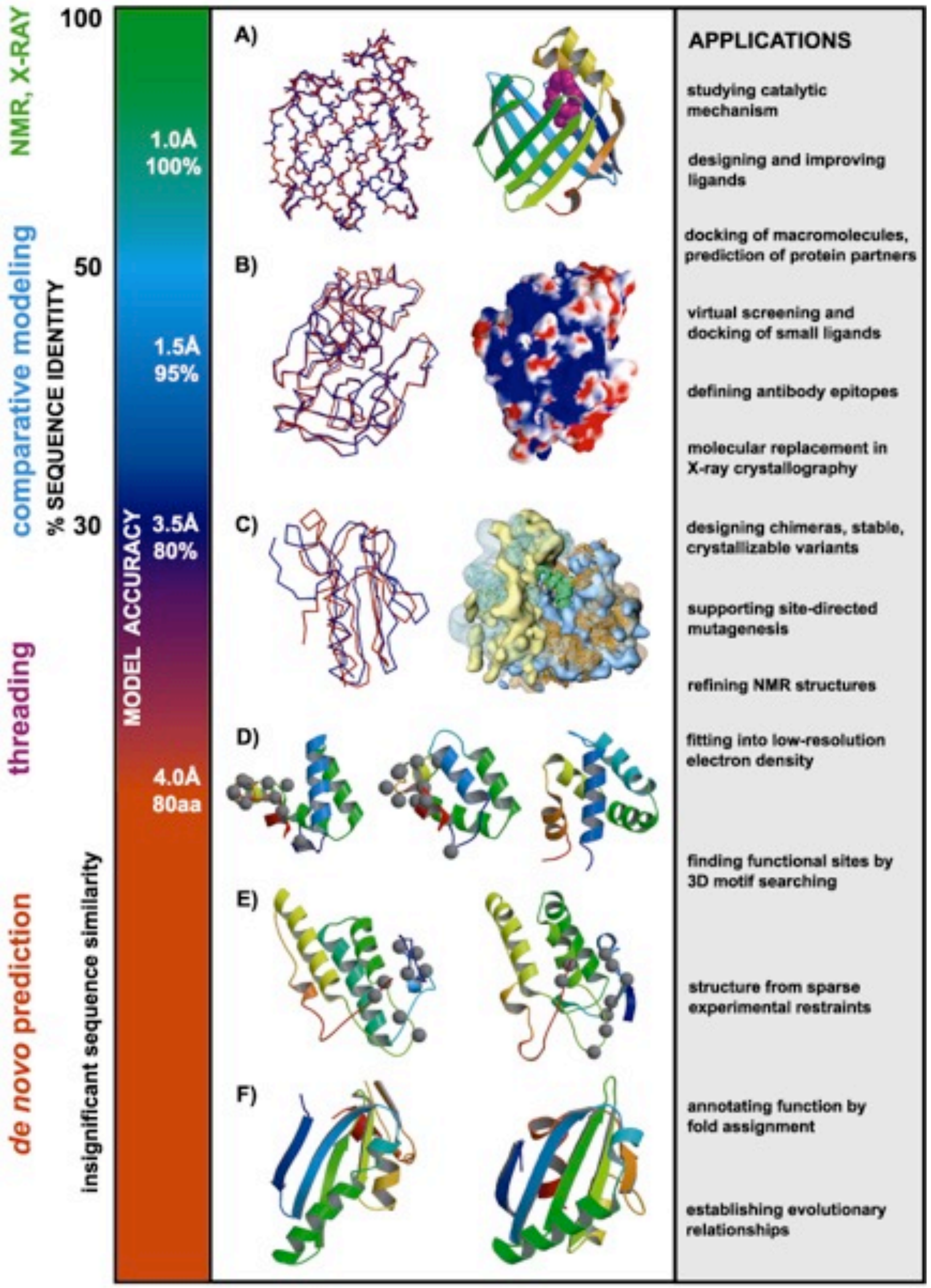






A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Utility of protein structure models, despite errors



D. Baker & A. Sali. Science 294, 93, 2001.

What is the physiological ligand of Brain Lipid-Binding Protein?

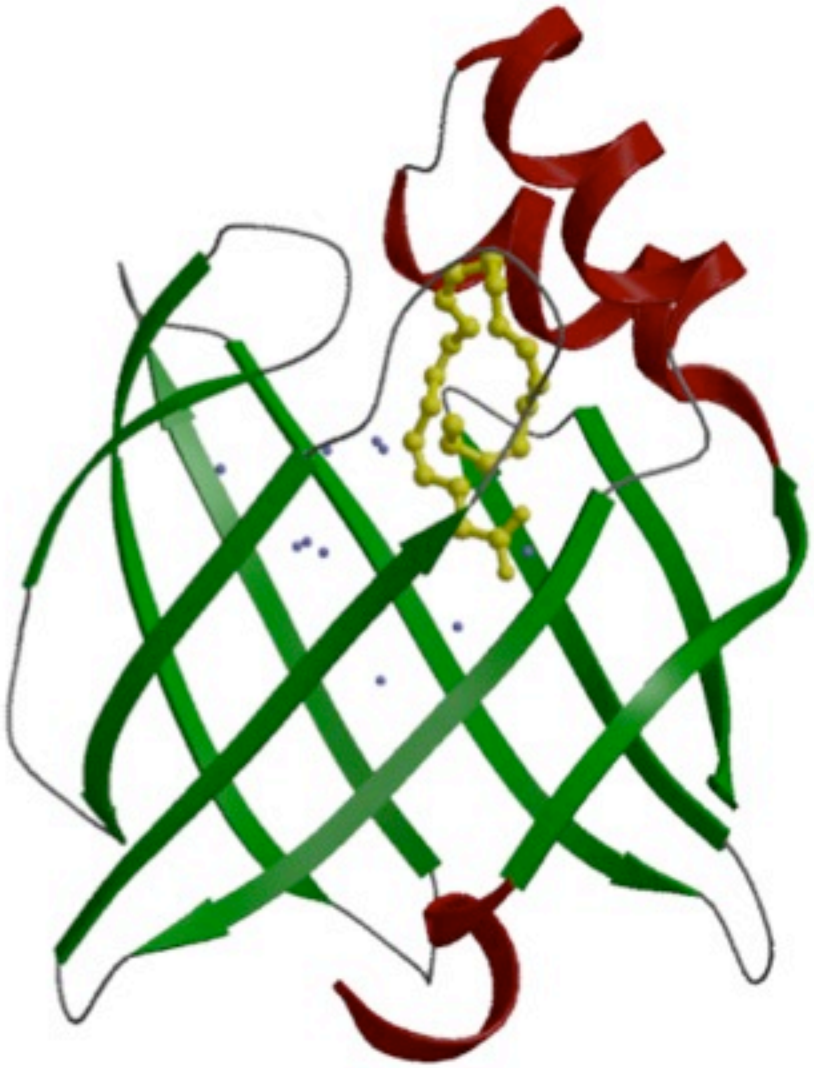
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is not filled

BLBP/docosahexaenoic acid

Cavity is filled



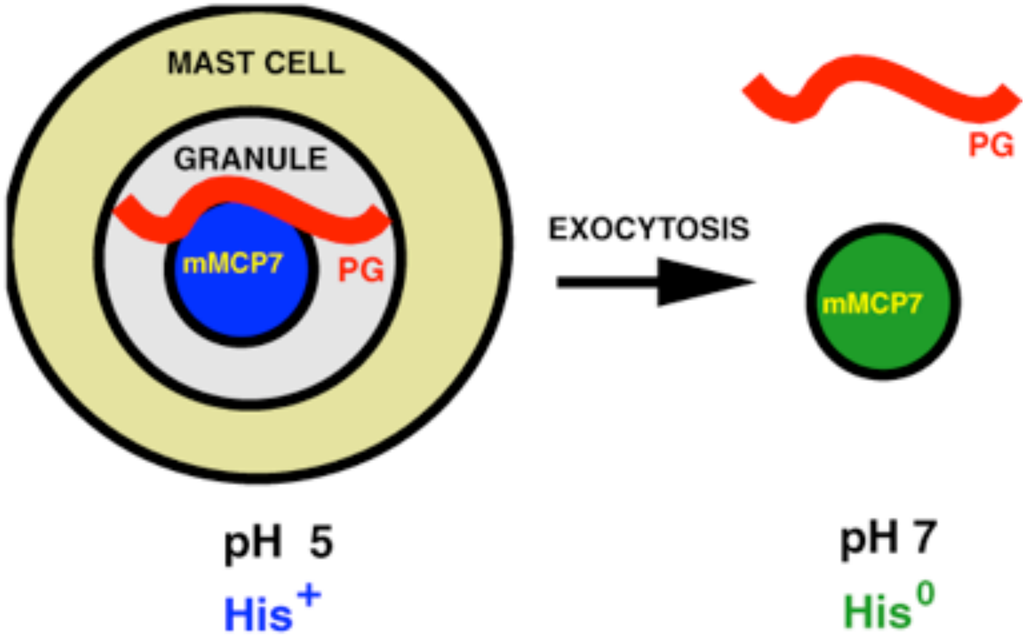
1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

L. Xu, R. Sánchez, A. Šali, N. Heintz, J. Biol. Chem. 271, 24711, 1996.

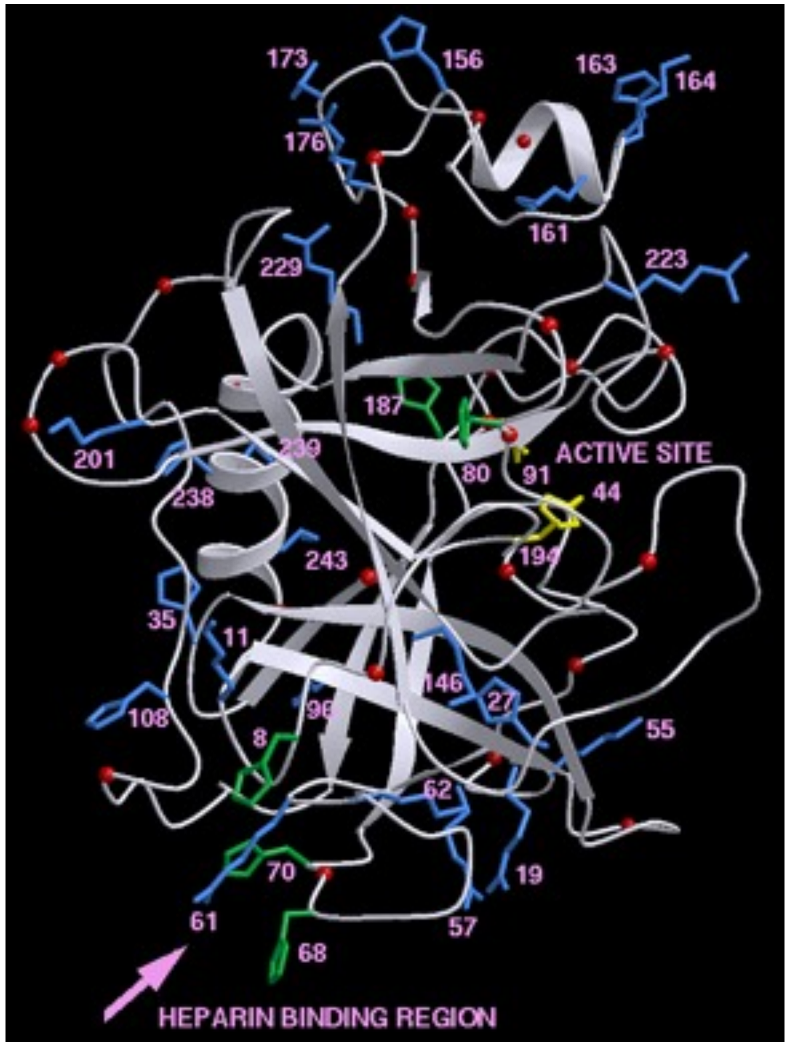
Do mast cell proteases bind proteoglycans? Where? When?

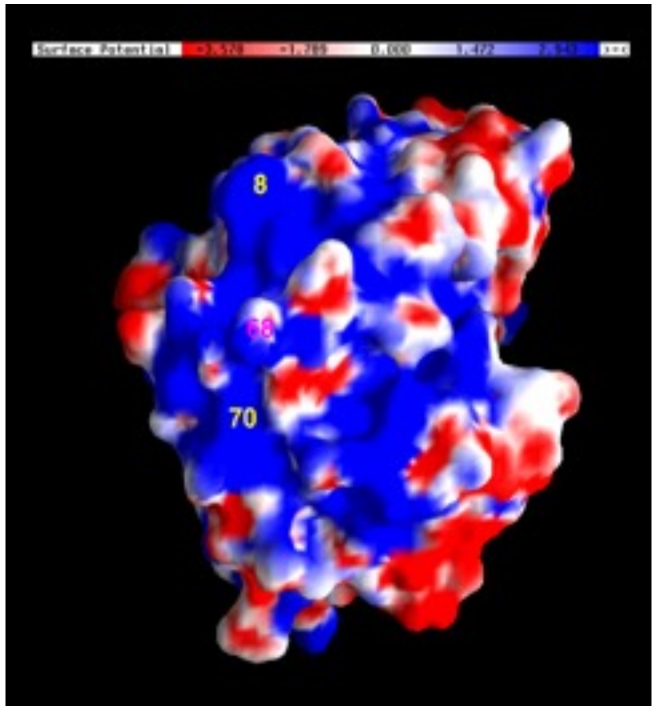
Predicting features of a model that are not present in the template

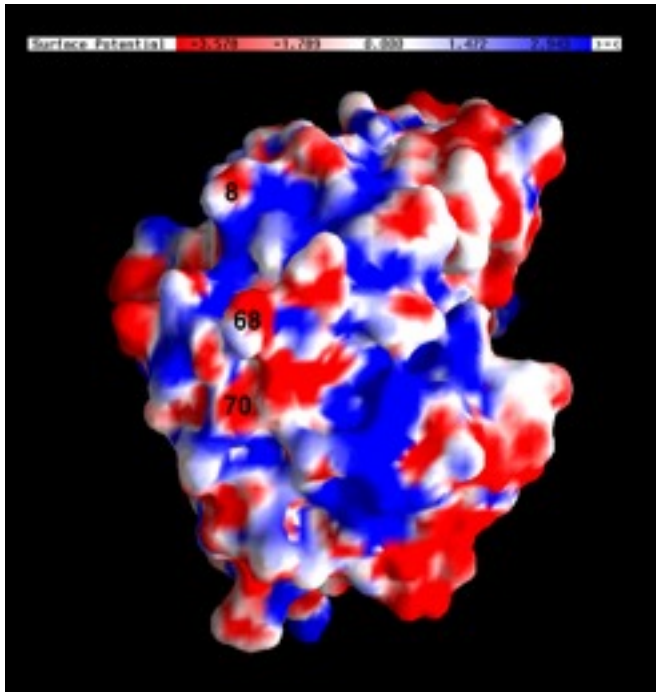
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.





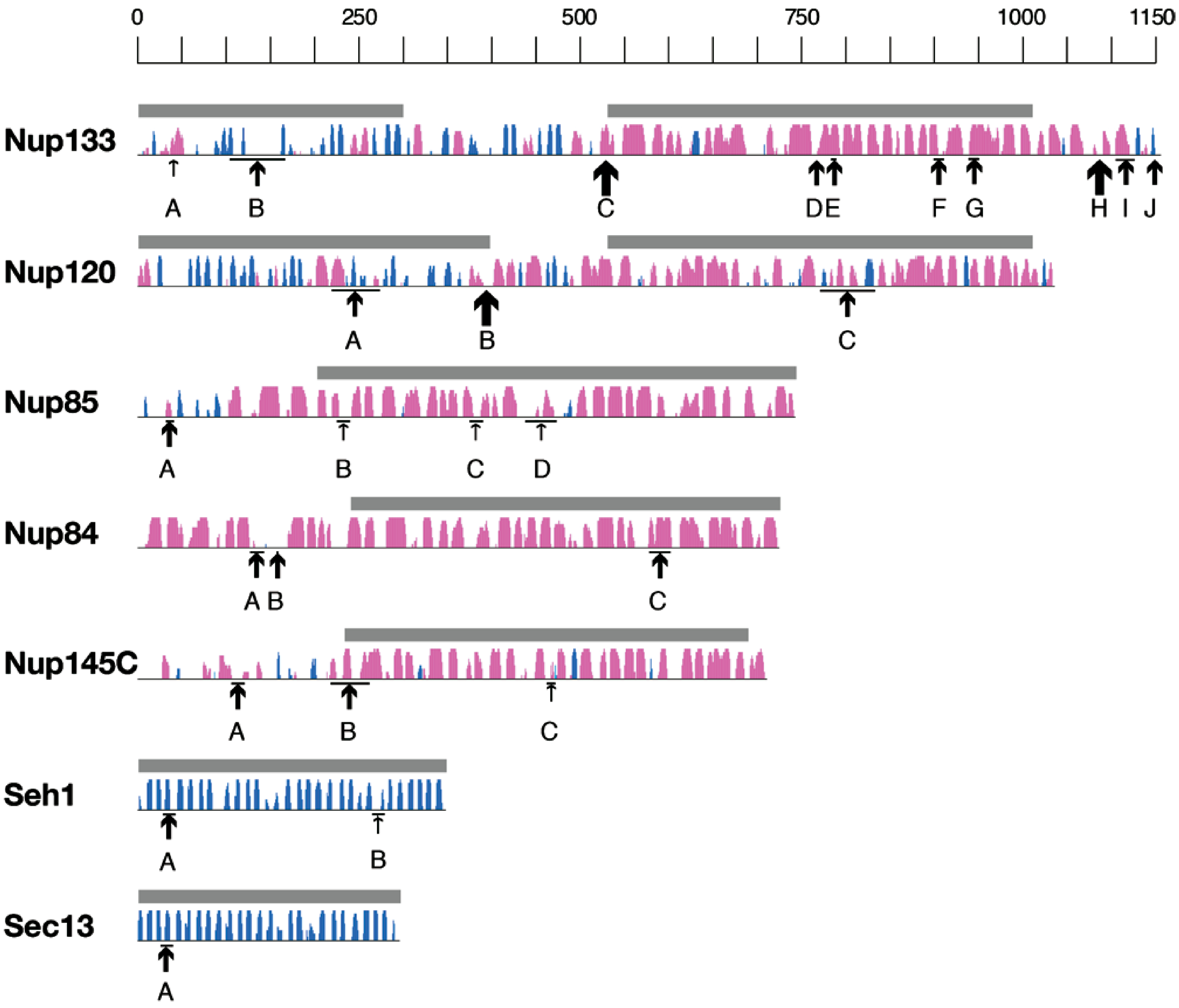


Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

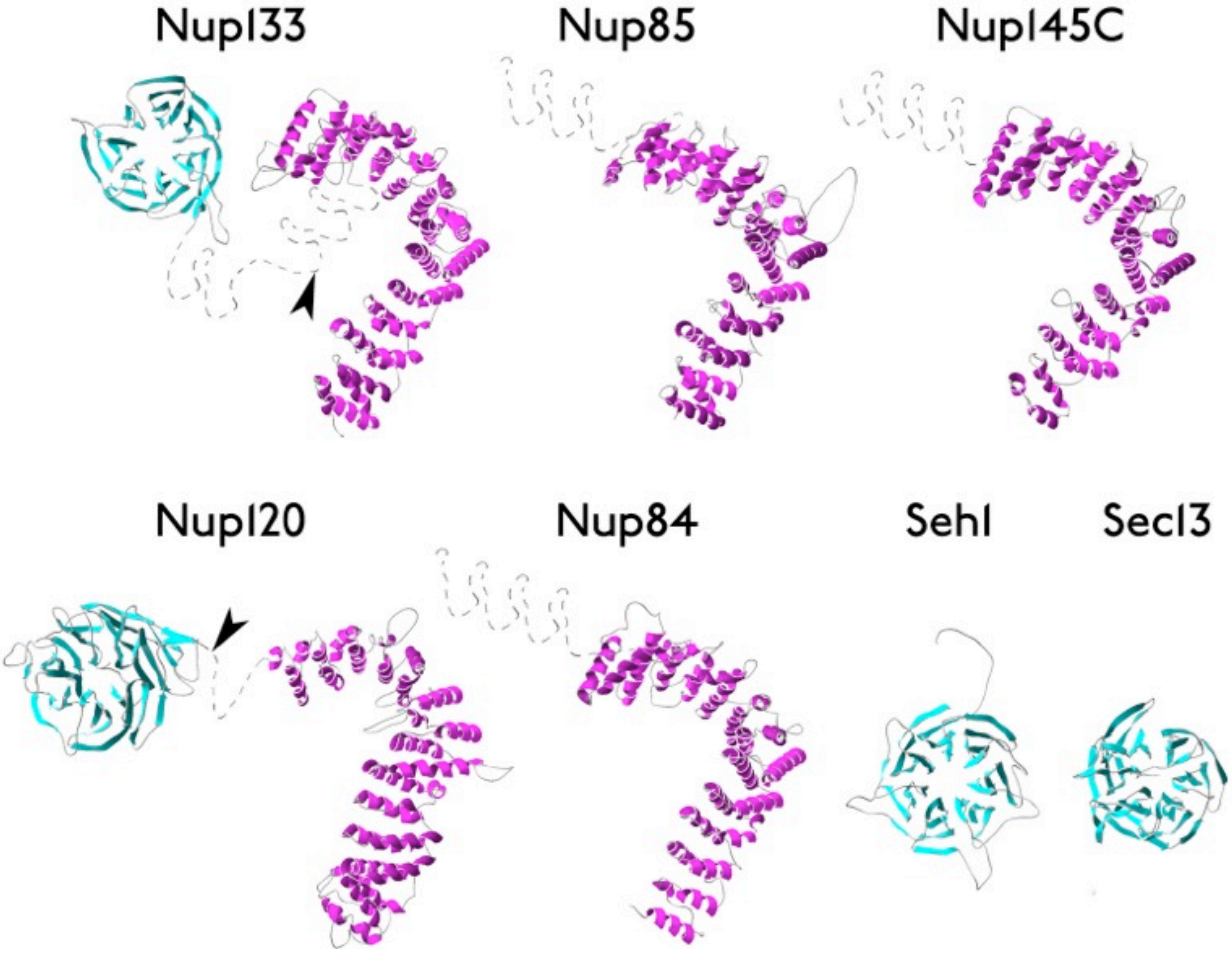
mGenThreader + SALIGN + MOULDER

D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B.T. Chait, A. Sali, M.P. Rout.
Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture.
PLOS Biology **2(12):**e380, 2004

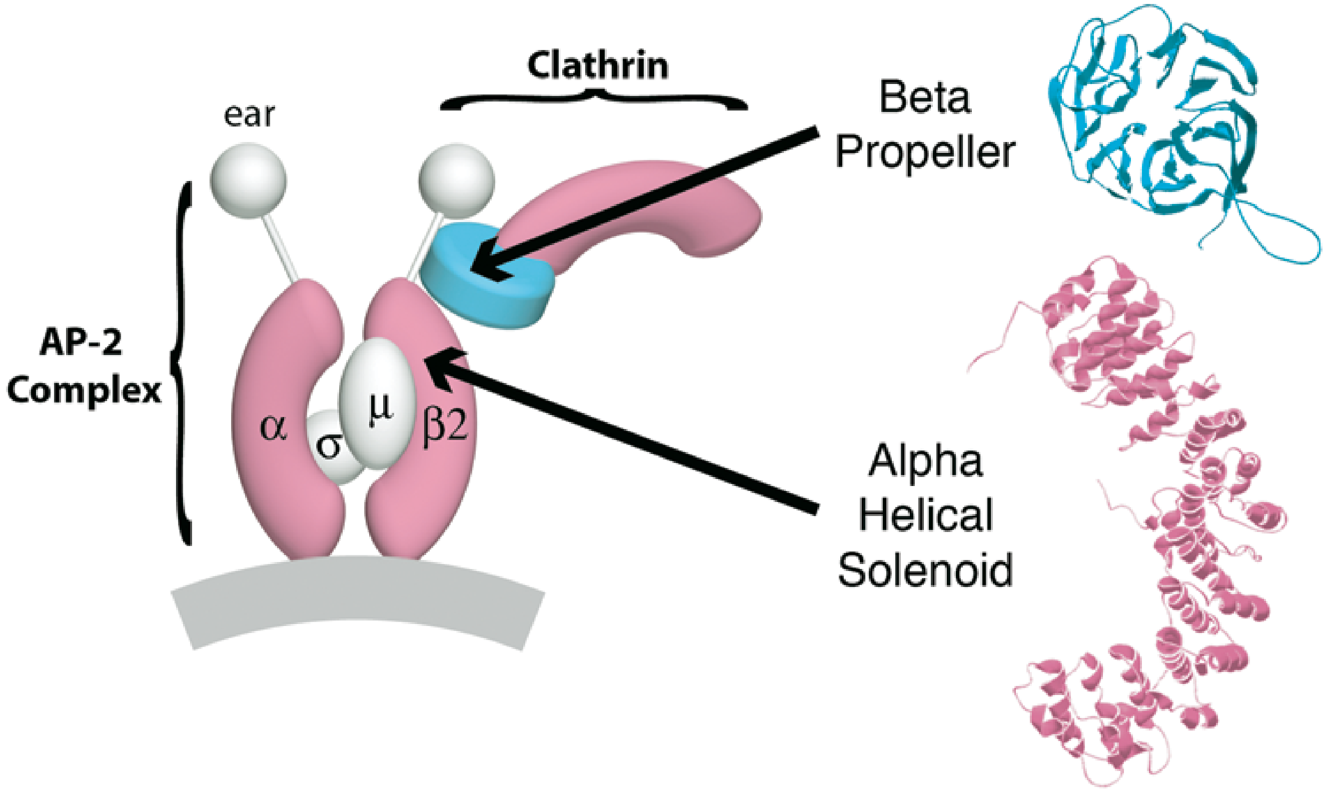
yNup84 complex proteins



All Nucleoporins in the Nup84 Complex are Predicted to Contain β-Propeller and/or α-Solenoid Folds



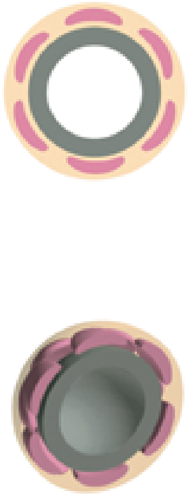
NPC and Coated Vesicles Share the β-**Propeller and** α-**Solenoid** Folds and Associate with Membranes



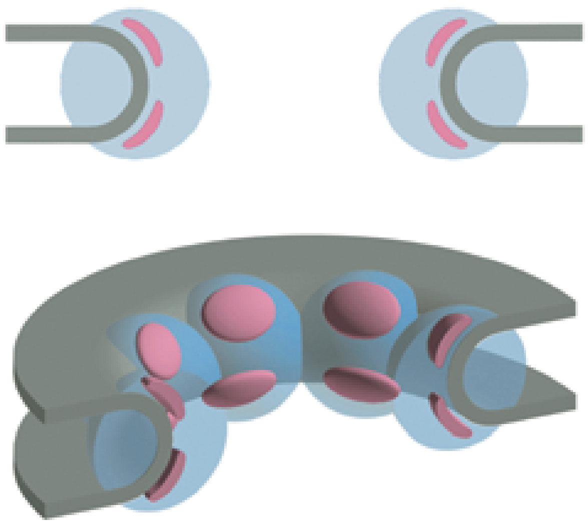
FOLDS	VARIANTS			
	CCVs	*COPI*	*COPII*	*NPC*
Beta Propeller	Clathrin-N	α-subunit-N β'-subunit-N	Sec13 (Seh1) Sec31-N	Sec13 Seh1 Nup133-N Nup120-N
Alpha Helical Solenoid	Clathrin-C α/γ/δ/ε-subunit β-subunits	α-subunit-C β'-subunit-C β-subunit γ-subunit	Sec31-C	Nup133-C Nup120-C Nup85 Nup84 Nup145C

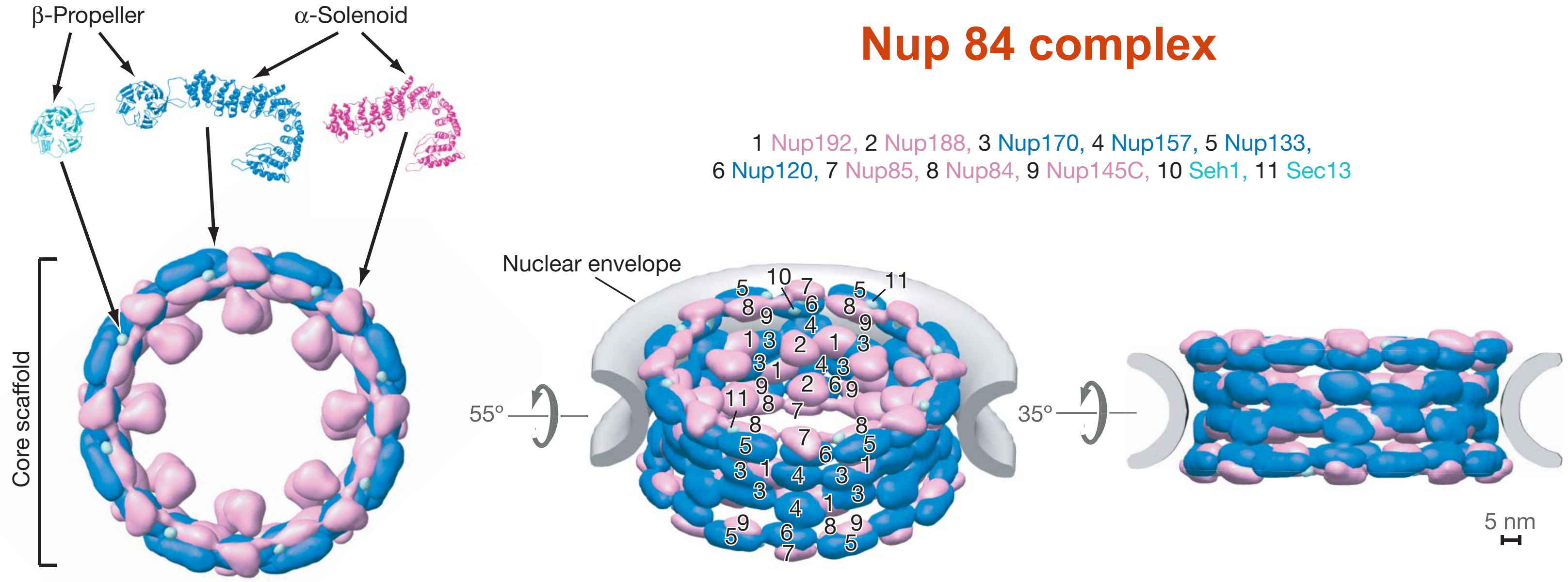
NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle



NPC model



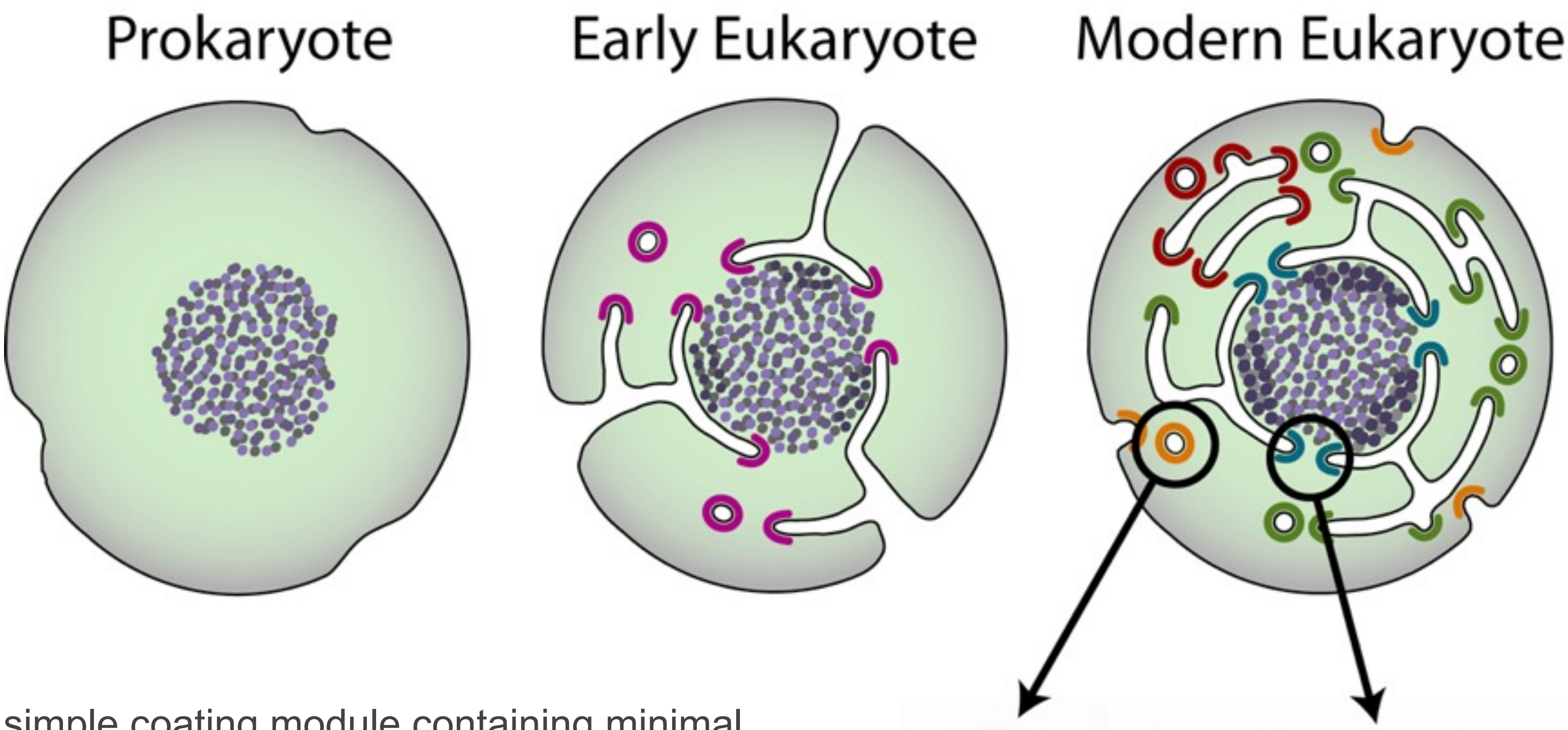


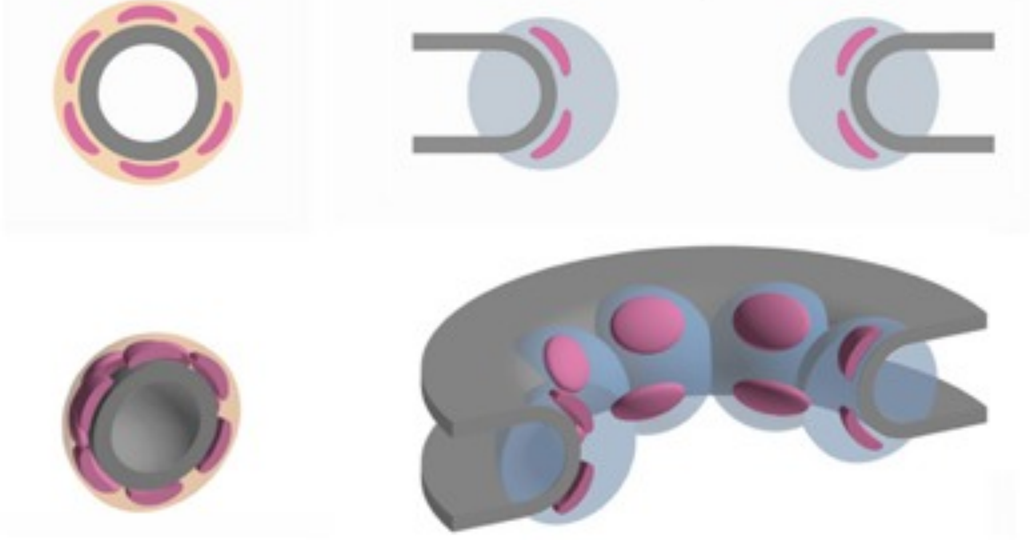
Nup 84 complex

Alber et al. The molecular architecture of the nuclear pore complex. Nature (2007) vol. 450 (7170) pp. 695-701

A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?

The proto-coatomer hypothesis



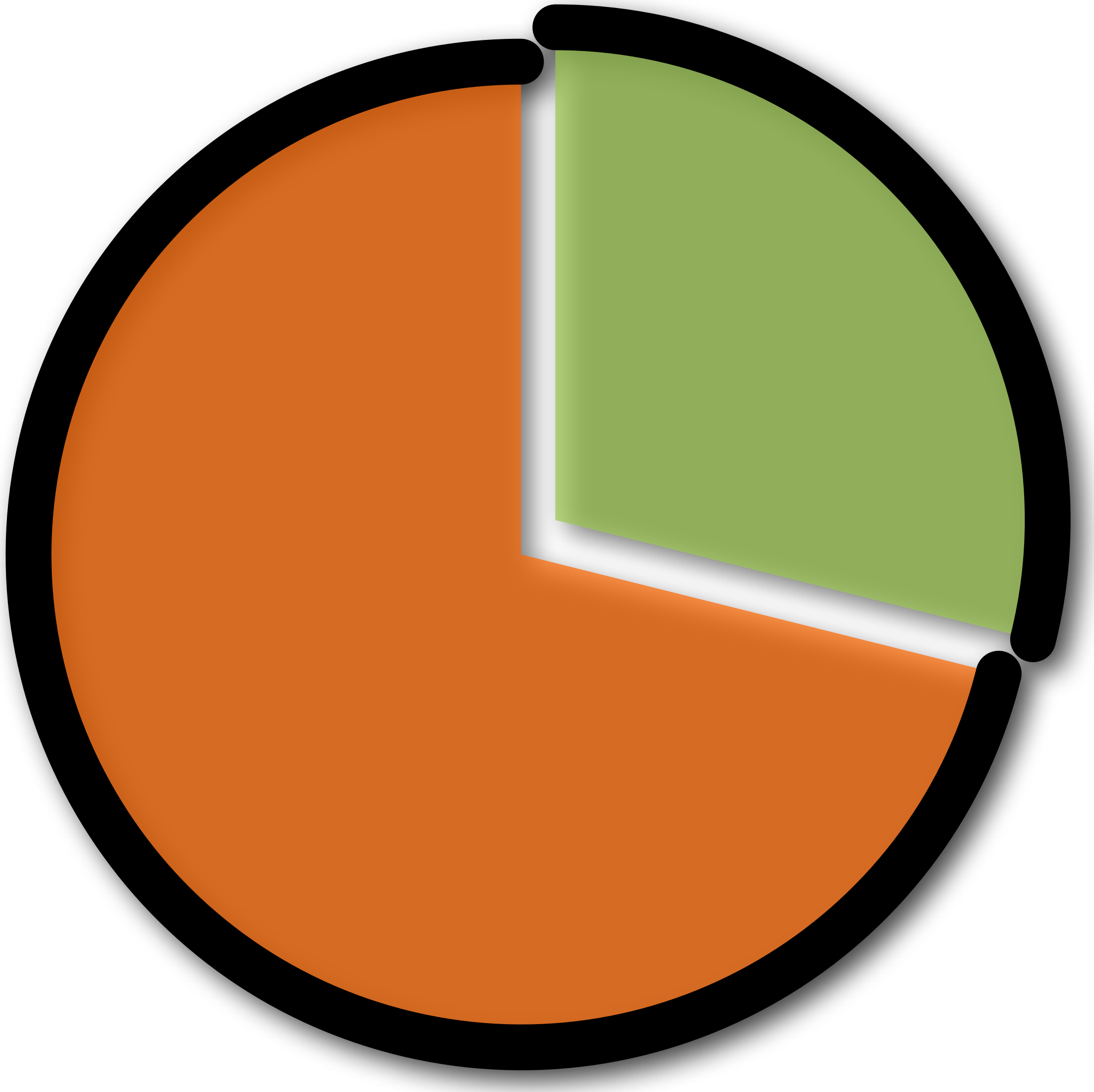


A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

The progenitor of the NPC arose from a membrane-coating module that wrapped extensions of an early ER around the cell's chromatin.

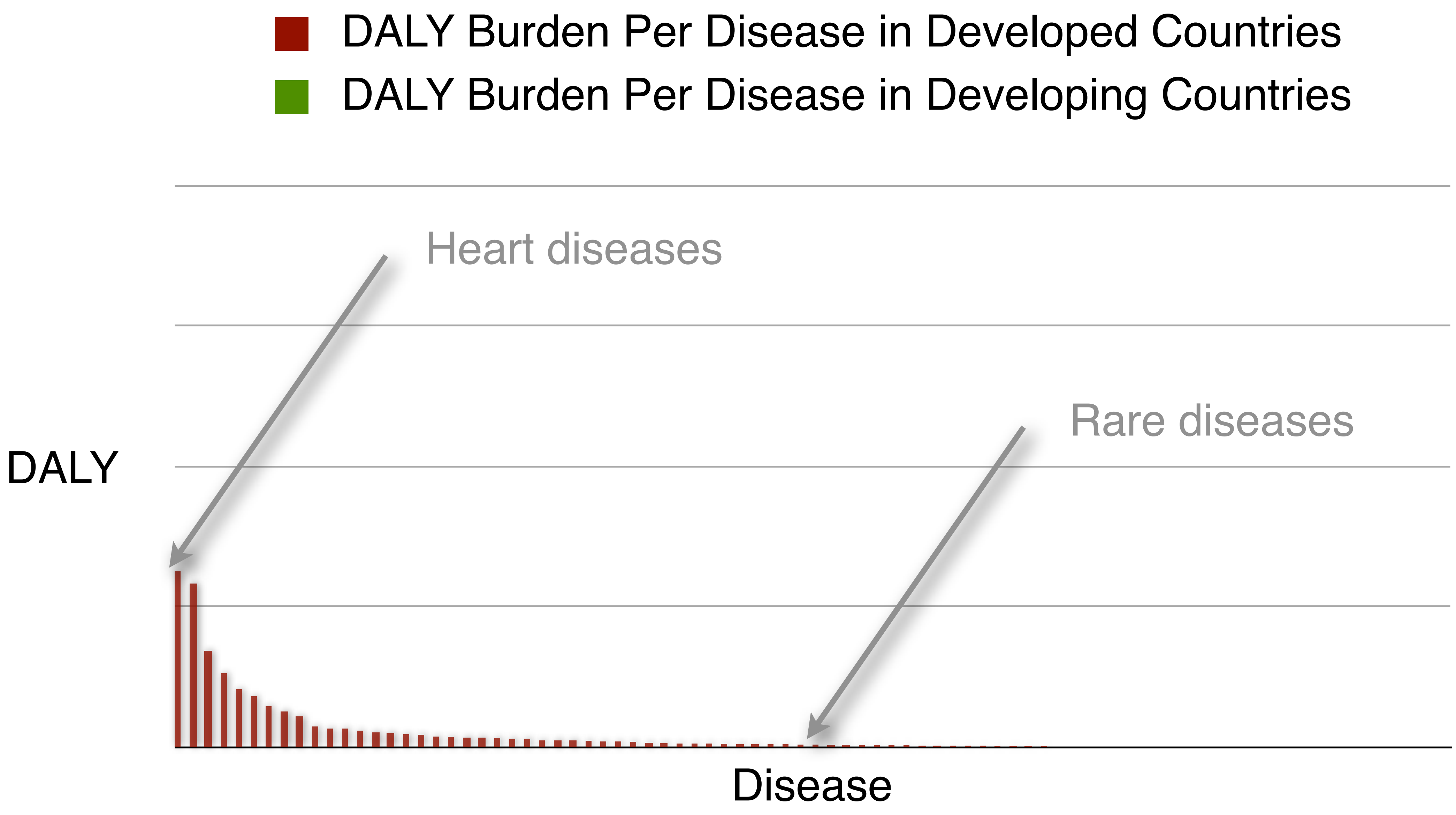
Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.



http://www.tropicaldisease.org

Need is High in the Tail



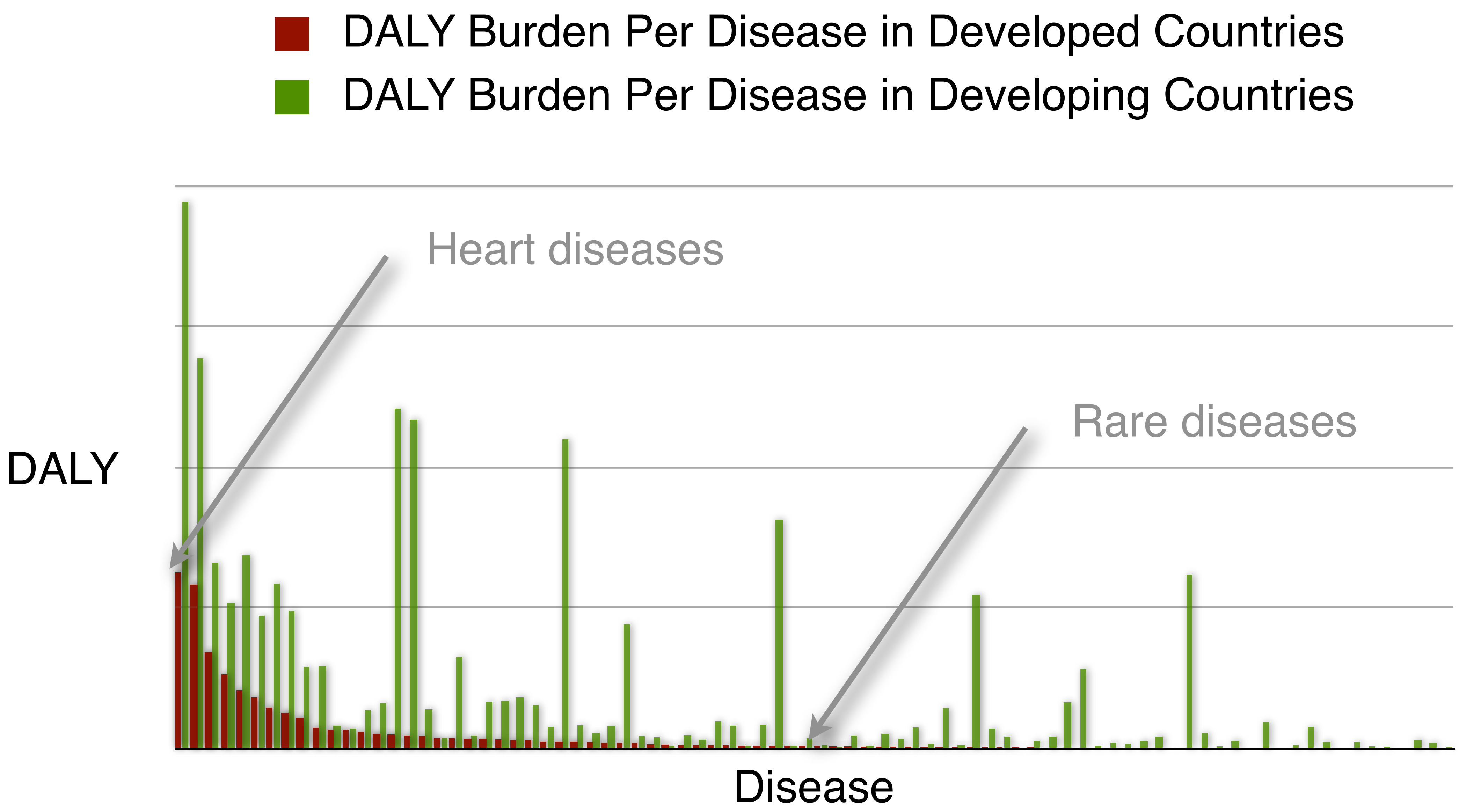
Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life years

DALY is not a perfect measure of market size, but is certainly a good measure for importance.

DALYs for a disease are the sum of the years of life lost due to premature mortality (YLL) in the population and the years lost due to disability (YLD) for incident cases of the health condition. The DALY is a health gap measure that extends the concept of potential years of life lost due to premature death (PYLL) to include equivalent years of 'healthy' life lost in states of less than full health, broadly termed disability. One DALY represents the loss of one year of equivalent full health.

Need is High in the Tail



Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life years

DALY is not a perfect measure of market size, but is certainly a good measure for importance.

DALYs for a disease are the sum of the years of life lost due to premature mortality (YLL) in the population and the years lost due to disability (YLD) for incident cases of the health condition. The DALY is a health gap measure that extends the concept of potential years of life lost due to premature death (PYLL) to include equivalent years of 'healthy' life lost in states of less than full health, broadly termed disability. One DALY represents the loss of one year of equivalent full health.

“Unprofitable” Diseases and Global DALY (in 1000’s)

Malaria*	**46,486**
Tetanus	7,074
Lymphatic filariasis*	**5,777**
Syphilis	4,200
Trachoma	2,329
Leishmaniasis*	**2,090**
Ascariasis	1,817
Schistosomiasis*	**1,702**
Trypanosomiasis*	**1,525**

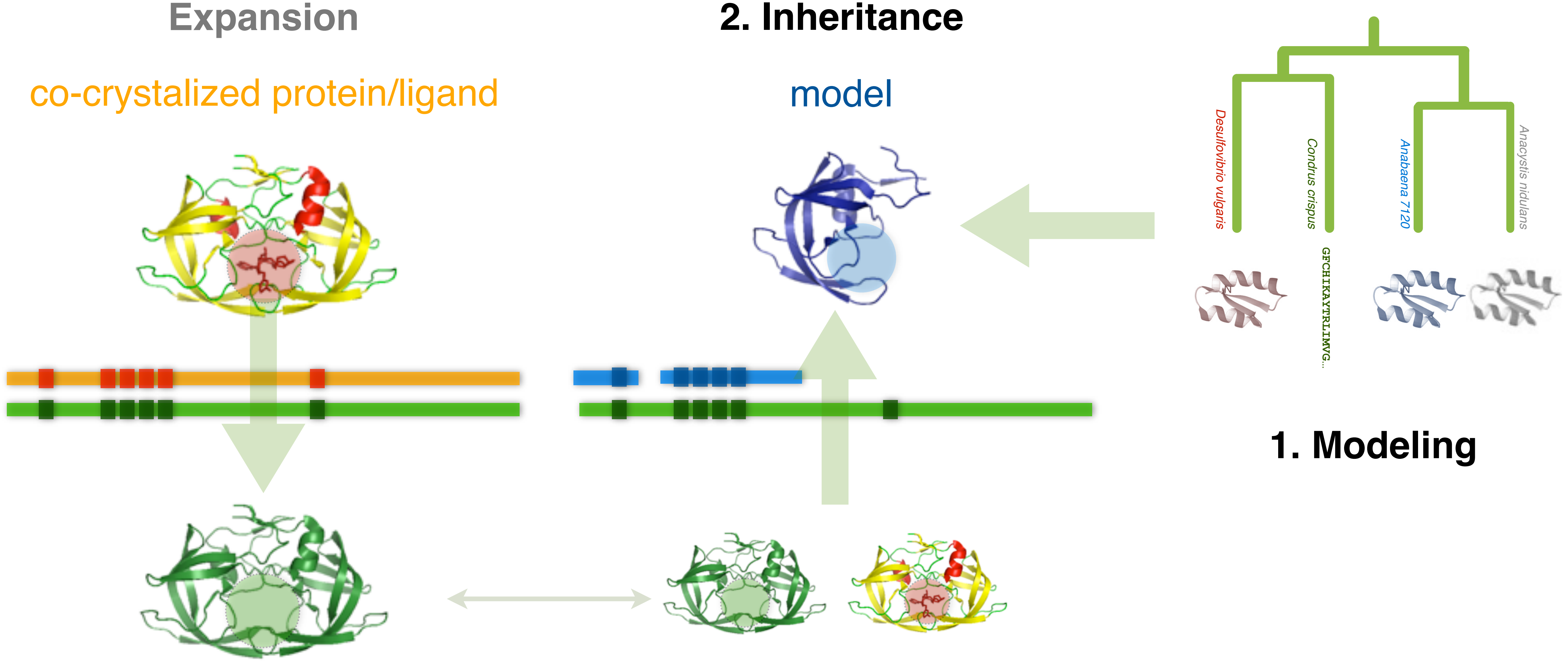
Trichuriasis	1,006
Japanese encephalitis	709
Chagas Disease*	**667**
Dengue*	**616**
Onchocerciasis*	**484**
Leprosy*	**199**
Diphtheria	185
Poliomyelitise	151
Hookworm disease	59

Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life year in 1000’s.

* Officially listed in the WHO Tropical Disease Research disease portfolio.

Comparative docking



Expansion

co-crystalized protein/ligand

2. Inheritance

model

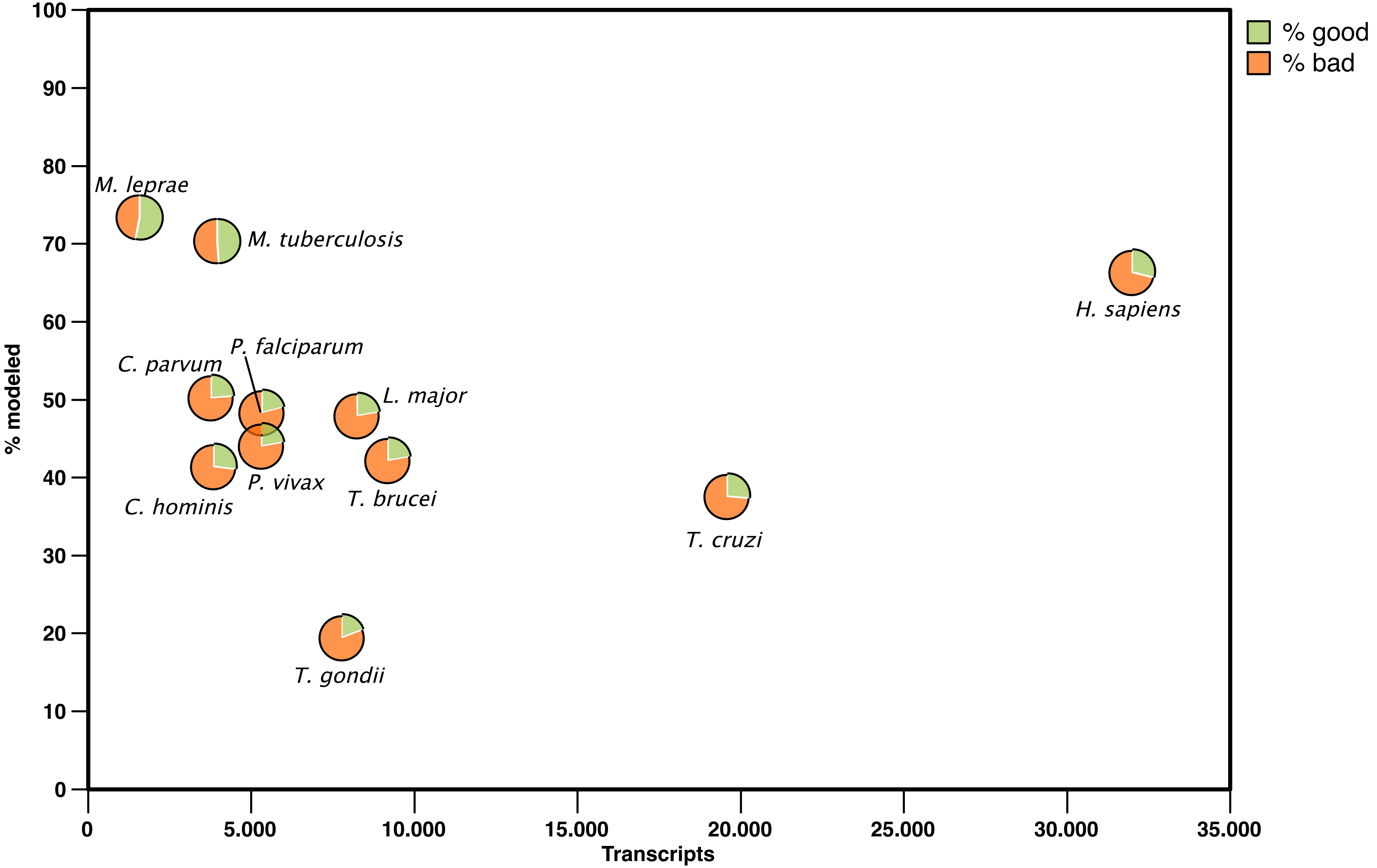
1. Modeling

crystalized protein

template

Modeling Genomes

data from models generated by ModPipe (Eswar, Pieper & Sali)



A good model has MPQS of 1.0 or higher

Summary table

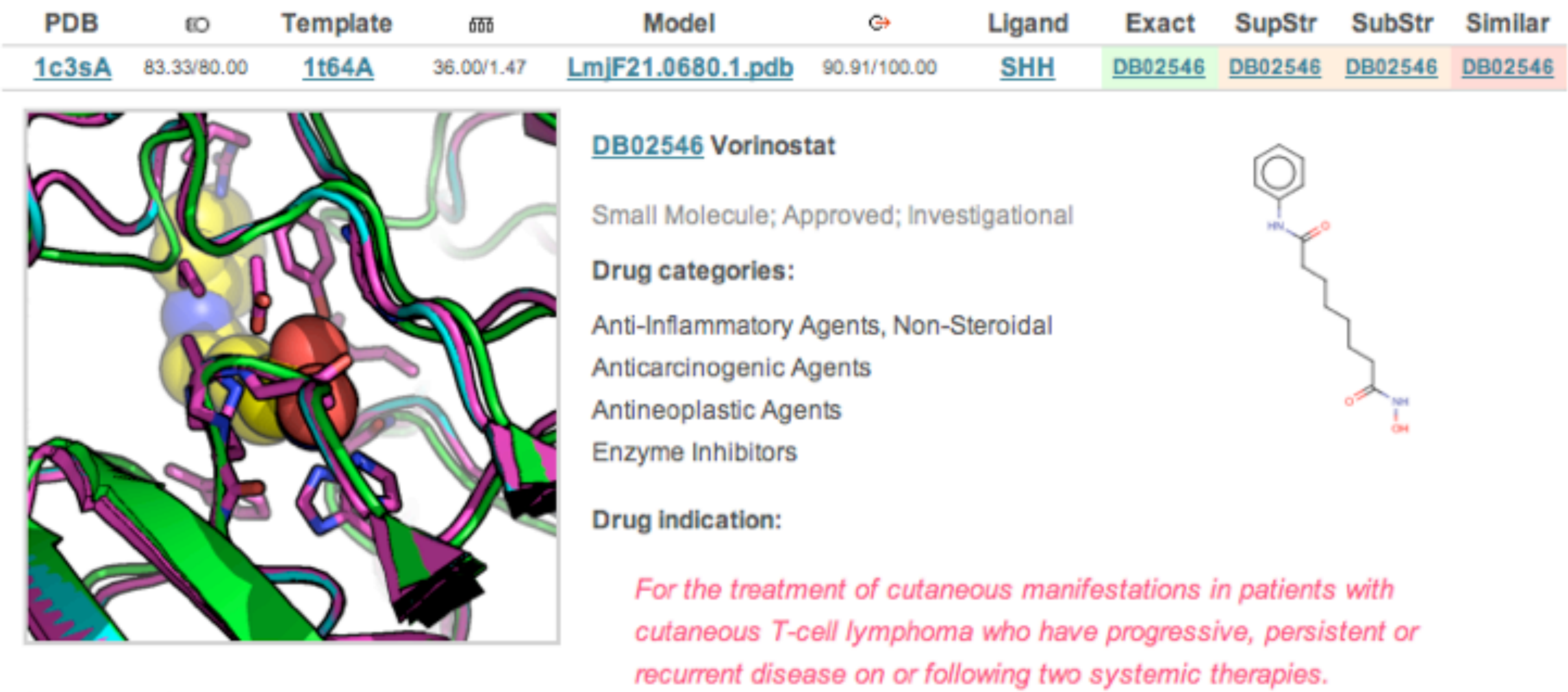
models with inherited ligands

29,271 targets with good models, 297 inherited a ligand/substance similar to a known drug in DrugBank

	Transcripts	Modeled targets	Selected models	Inherited ligands	Similar to a drug	Drugs
C. hominis	3,886	1,614	666	197	20	13
C. parvum	3,806	1,918	742	232	24	13
L. major	8,274	3,975	1,409	478	43	20
M. leprae	1,605	1,178	893	310	25	6
M. tuberculosis	3,991	2,808	1,608	365	30	10
P. falciparum	5,363	2,599	818	284	28	13
P. vivax	5,342	2,359	822	268	24	13
T. brucei	7,793	1,530	300	138	13	6
T. cruzi	19,607	7,390	3,070	769	51	28
T. gondii	9,210	3,900	1,386	458	39	21
TOTAL	**68,877**	**29,271**	**11,714**	**3,499**	**297**	**143**

L. major Histone deacetylase 2 + Vorinostat

Template 1t64A a human HDAC8 protein.



PDB		Template		Model		Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/80.00	1t64A	36.00/1.47	LmjF21.0680.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546

DB02546 **Vorinostat**

Small Molecule; Approved; Investigational

Drug categories:

Anti-Inflammatory Agents, Non-Steroidal

Anticarcinogenic Agents

Antineoplastic Agents

Enzyme Inhibitors

Drug indication:

For the treatment of cutaneous manifestations in patients with cutaneous T-cell lymphoma who have progressive, persistent or recurrent disease on or following two systemic therapies.

L. major Histone deacetylase 2 + Vorinostat

Literature

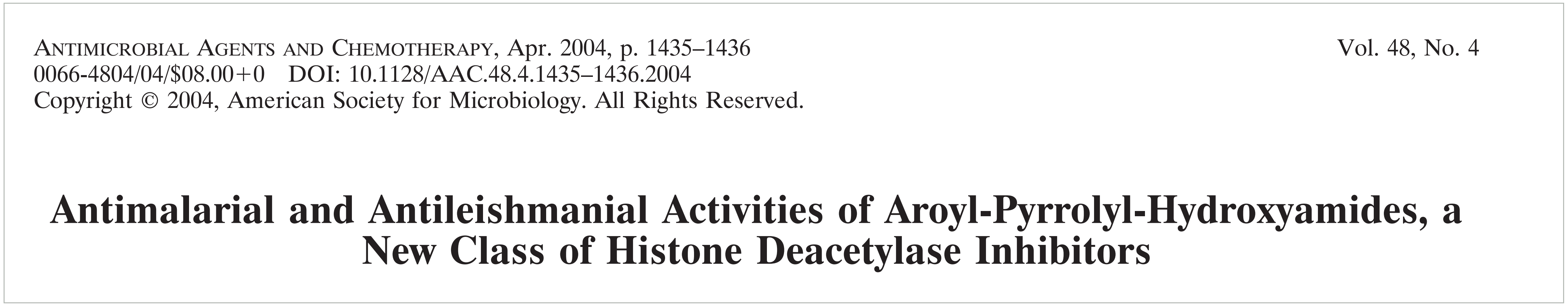
Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13143–13147, November 1996
Medical Sciences

Apicidin: A novel antiprotozoal agent that inhibits parasite histone deacetylase

(cyclic tetrapeptide/Apicomplexa/antiparasitic/malaria/coccidiosis)

Sandra J. Darkin-Rattray[*,†], Anne M. Gurnett[*], Robert W. Myers[*], Paula M. Dulski[*],
Tami M. Crumley[*], John J. Allocco[*], Christine Cannova[*], Peter T. Meinke[‡], Steven L. Colletti[‡],
Maria A. Bednarek[‡], Sheo B. Singh[§], Michael A. Goetz[§], Anne W. Dombrowski[§],
Jon D. Polishook[§], and Dennis M. Schmatz[*]

Departments of [*]Parasite Biochemistry and Cell Biology, [‡]Medicinal Chemistry, and [§]Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

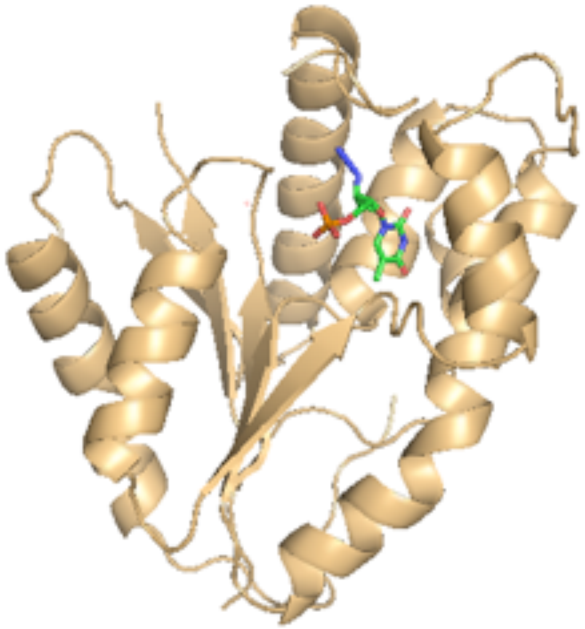


Antimicrobial Agents and Chemotherapy, Apr. 2004, p. 1435–1436 Vol. 48, No. 4
0066-4804/04/$08.00+0 DOI: 10.1128/AAC.48.4.1435–1436.2004

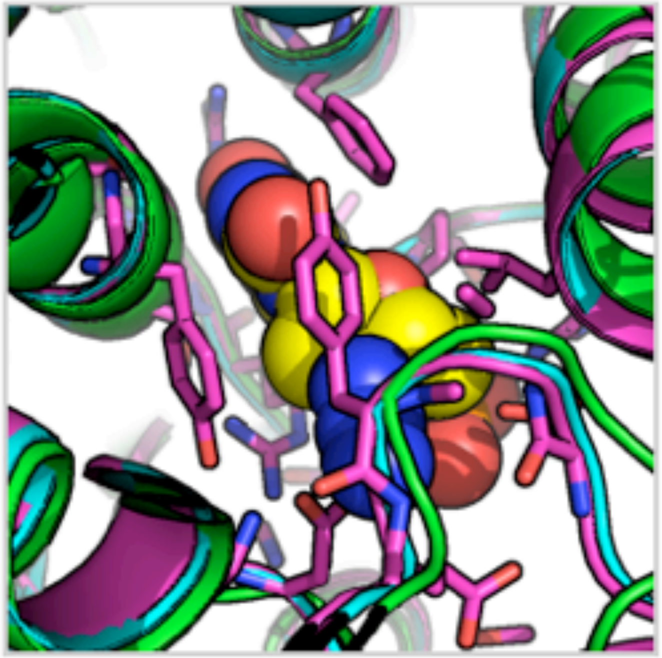
Antimalarial and Antileishmanial Activities of Aroyl-Pyrrolyl-Hydroxyamides, a New Class of Histone Deacetylase Inhibitors

P. falciparum tymidylate kinase + zidovudine

Template 3tmkA a yeast tymidylate kinase.



PDB		Template		Model		Ligand	Exact	SupStr	SubStr	Similar
2tmkB	100.00/100.00	3tmkA	41.00/1.49	PFL2465c.2.pdb	82.61/100.00	ATM		DB00495		DB00495



DB00495 Zidovudine

Small Molecule; Approved

Drug categories:

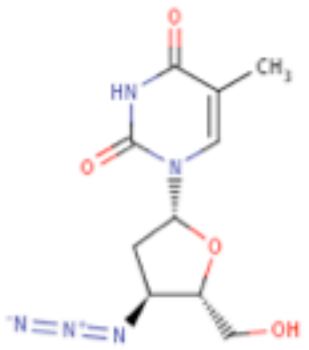
Anti-HIV Agents

Antimetabolites

Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

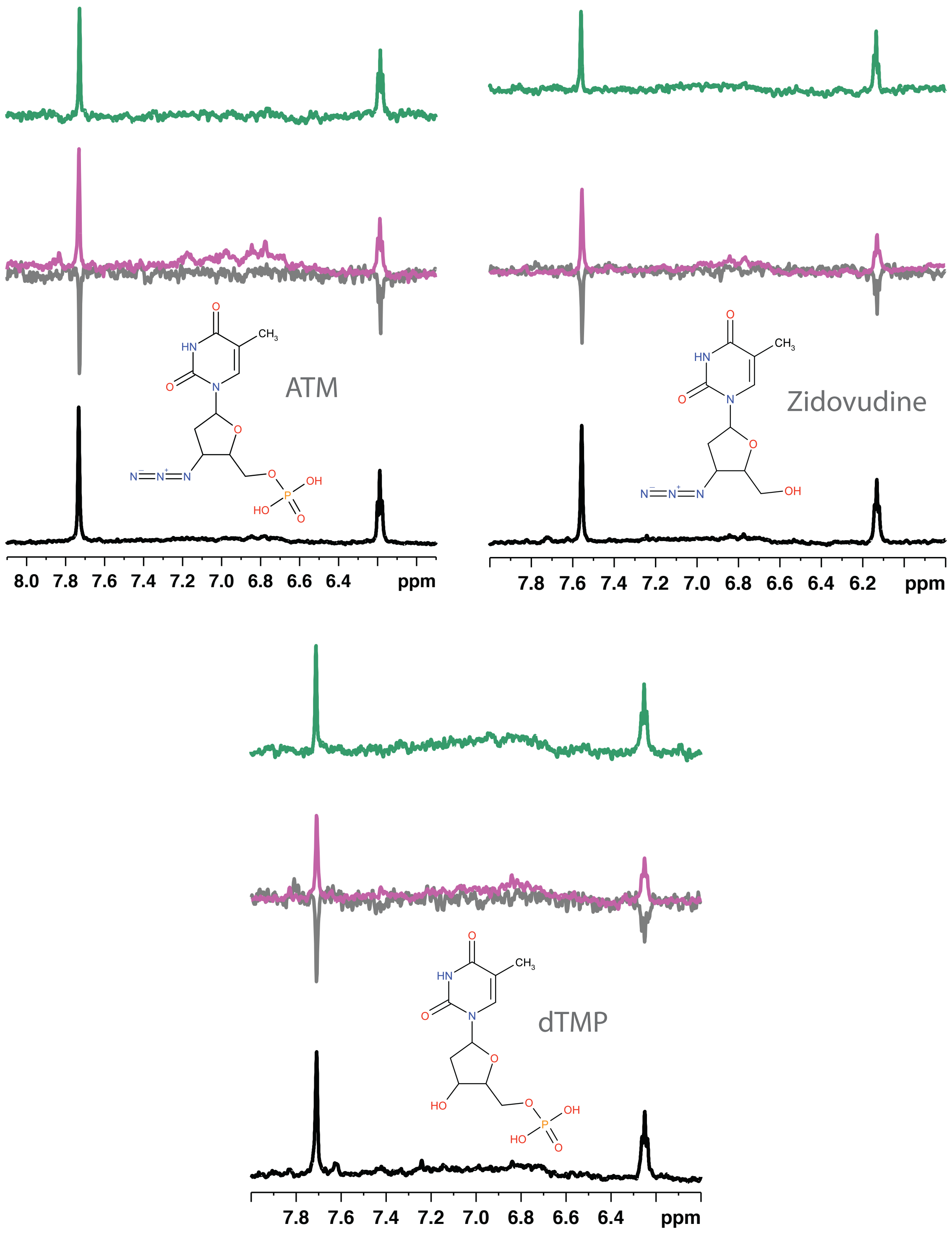
Drug Indication:

For the treatment of human immunovirus (HIV) infections.



P. falciparum tymydilate kinase + zidovudine

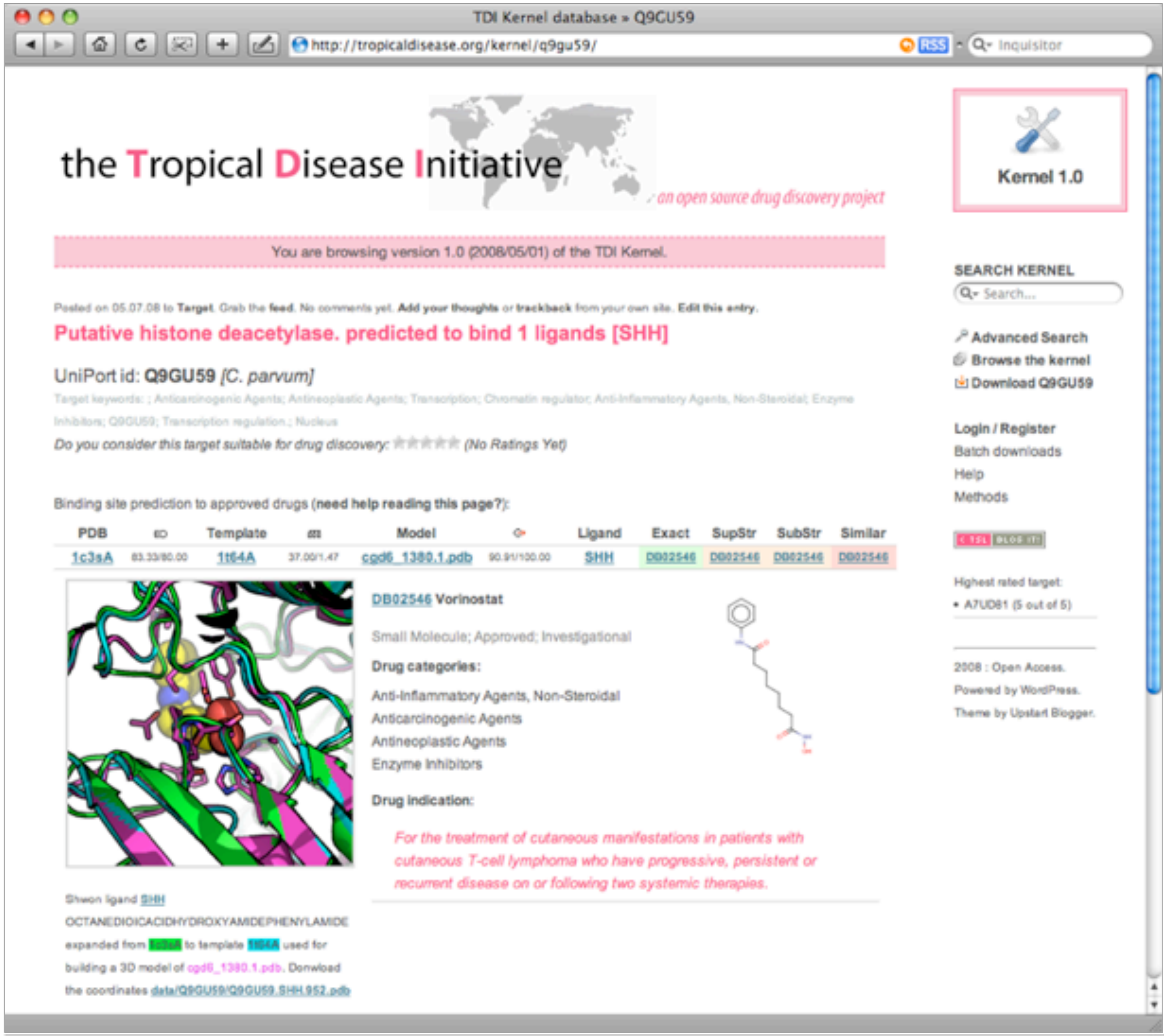
*NMR **Water-LOGSY** and **STD** experiments*



Leticia Ortí, Rodrigo J. Carbajo, and Antonio Pineda-Lucena

TDI’s kernel

http://tropicaldisease.org/kernel



You are browsing version 1.0 (2008/05/01) of the TDI Kernel.

Posted on 05.07.08 to **Target**. Grab the **feed**. No comments yet. **Add your thoughts** or **trackback** from your own site. **Edit this entry**.

Putative histone deacetylase. predicted to bind 1 ligands [SHH]

UniPort id: **Q9GU59** *[C. parvum]*

Target keywords: ; Anticarcinogenic Agents; Antineoplastic Agents; Transcription; Chromatin regulator; Anti-Inflammatory Agents, Non-Steroidal; Enzyme Inhibitors; Q9GU59; Transcription regulation.; Nucleus

Do you consider this target suitable for drug discovery: ☆☆☆☆☆ (No Ratings Yet)

Binding site prediction to approved drugs (**need help reading this page?**):

PDB		Template		Model		Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/80.00	1t64A	37.00/1.47	cgd6_1380.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546

DB02546 **Vorinostat**

Small Molecule; Approved; Investigational

Drug categories:

Anti-Inflammatory Agents, Non-Steroidal
Anticarcinogenic Agents
Antineoplastic Agents
Enzyme Inhibitors

Drug indication:

For the treatment of cutaneous manifestations in patients with cutaneous T-cell lymphoma who have progressive, persistent or recurrent disease on or following two systemic therapies.

Shwon ligand SHH OCTANEDIOICACIDHYDROXYAMIDEPHENYLAMIDE expanded from 1c3sA to template 1t64A used for building a 3D model of cgd6_1380.1.pdb. Donwload the coordinates data/Q9GU59/Q9GU59.SHH.952.pdb

SEARCH KERNEL

Search...

- Advanced Search
- Browse the kernel
- Download Q9GU59

Login / Register
Batch downloads
Help
Methods

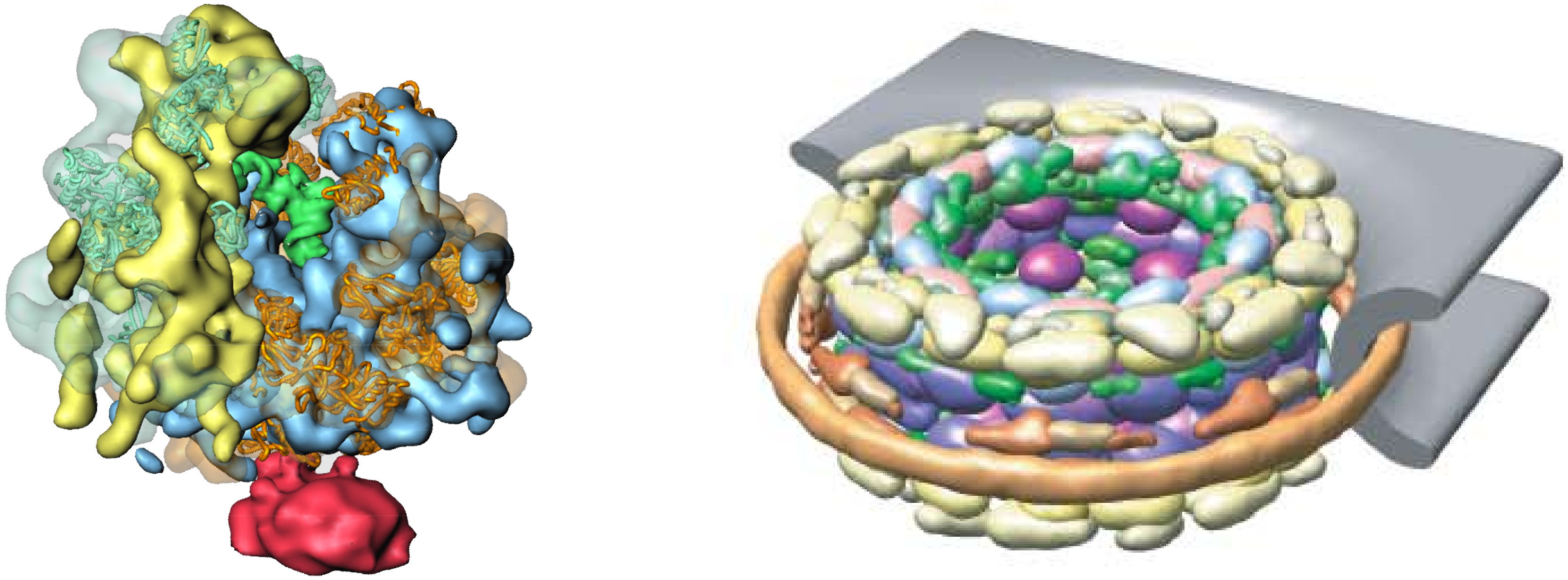
Highest rated target:

- A7UD81 (5 out of 5)

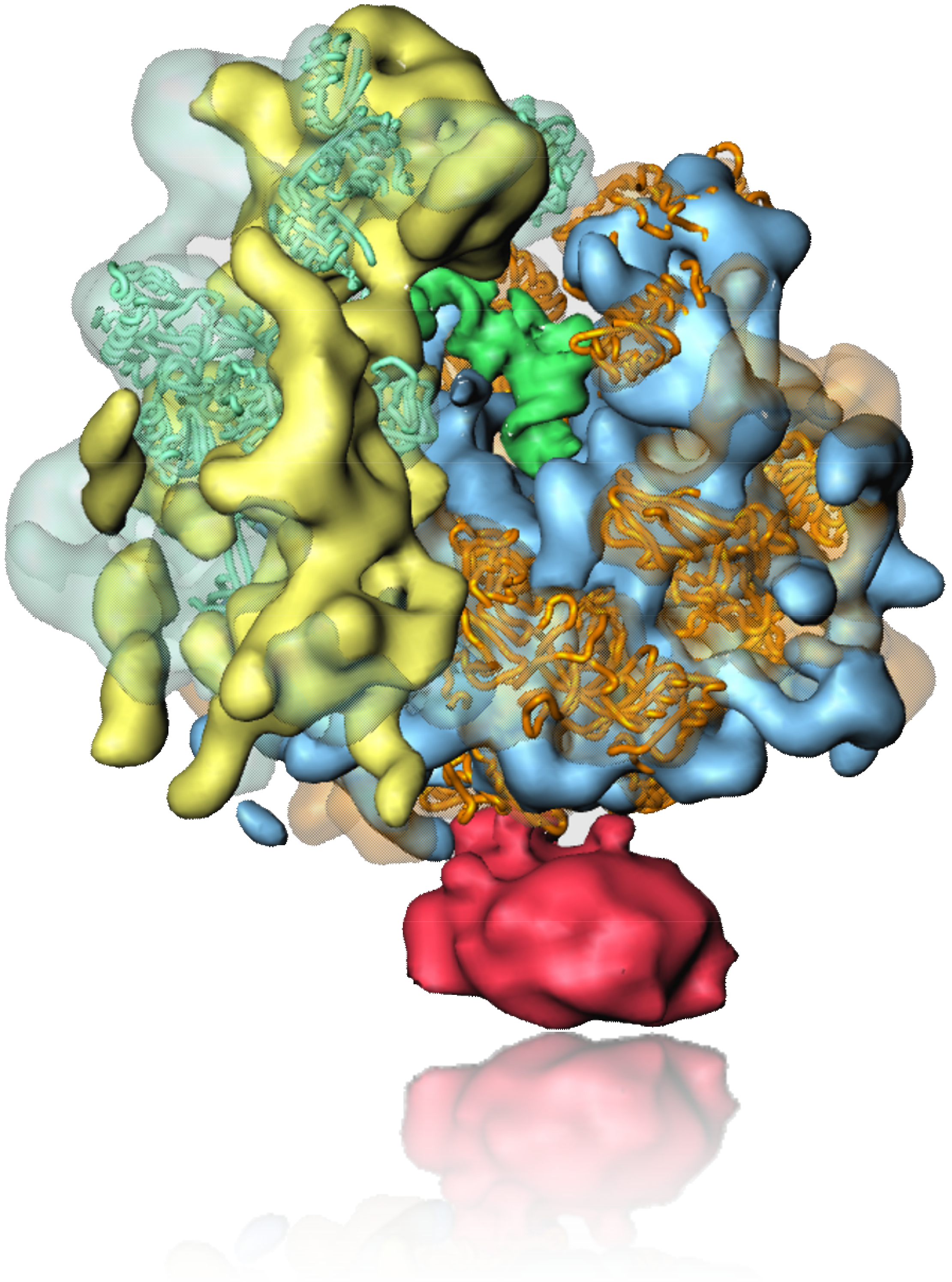
2008 : Open Access.
Powered by WordPress.
Theme by Upstart Blogger.

COMPLEXES

multiple *data types*



S. cerevisiae ribosome



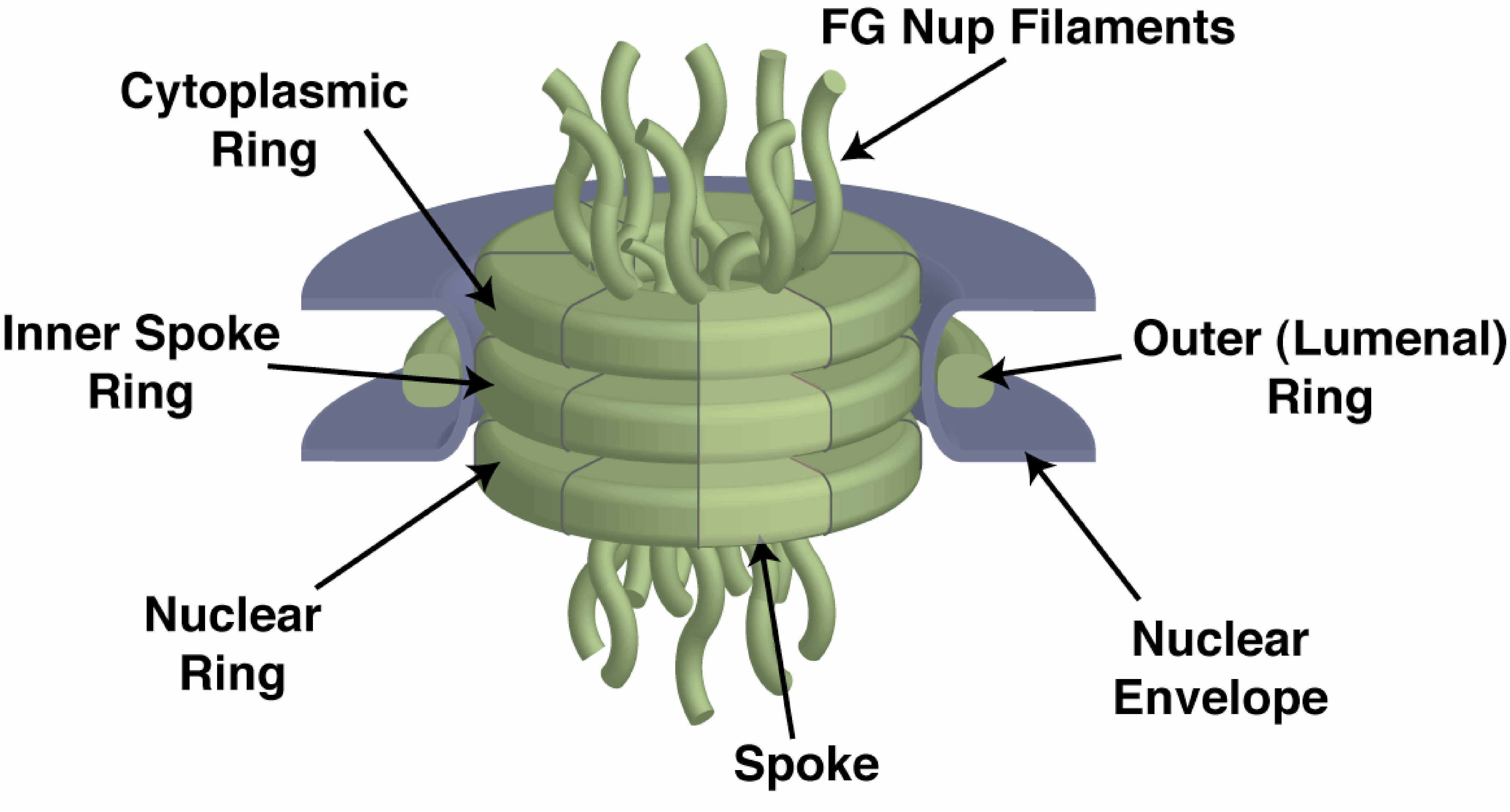
Fitting of comparative models into 15Å cryo- electron density map.

43 proteins could be modeled on 20-56% seq.id. to a known structure.

The modeled fraction of the proteins ranges from 34-99%.

C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, J. Frank. Cell 107, 361-372, 2001.

The NPC



Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). *Nature*, *450*(7170), 695–701

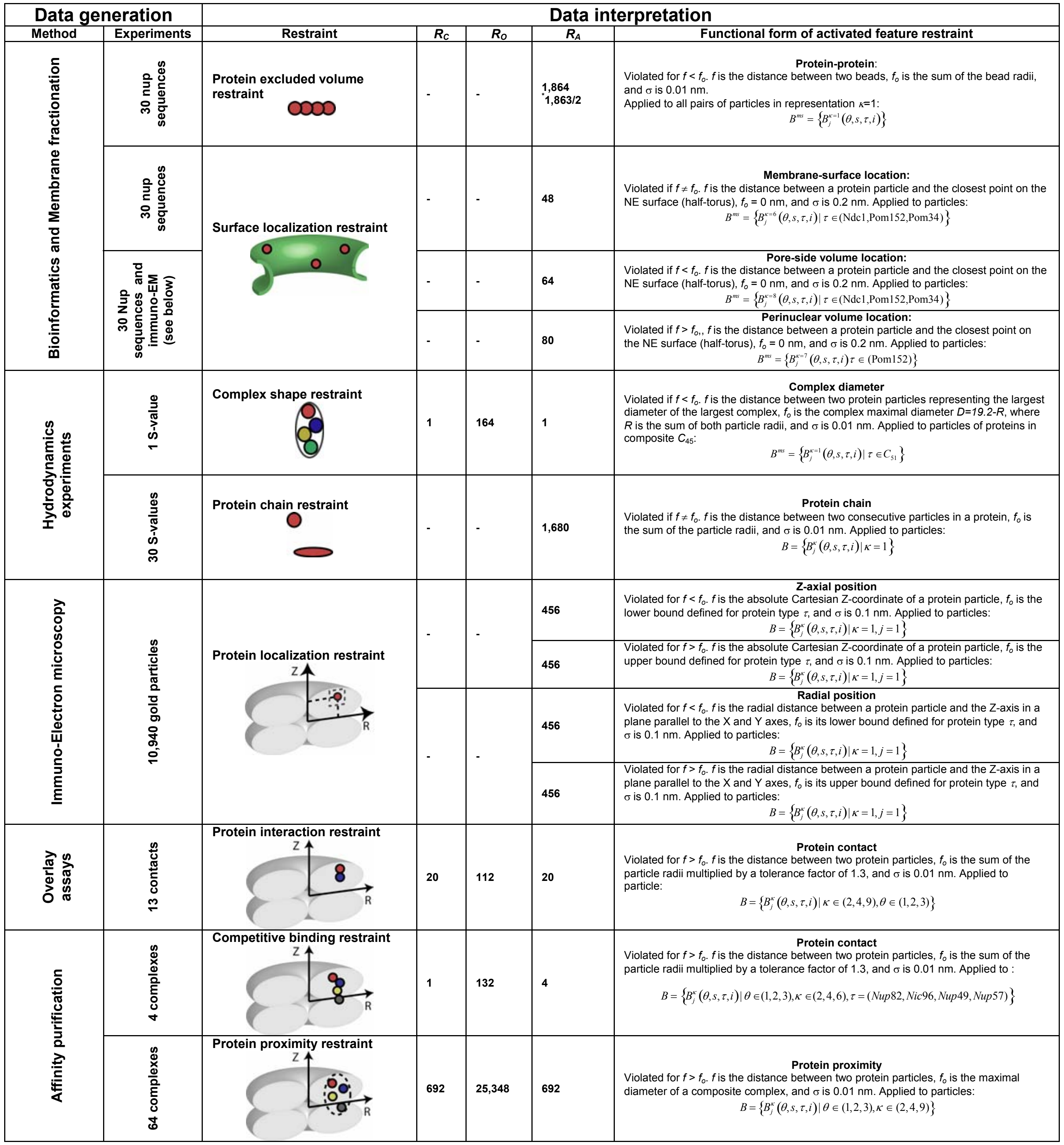
Representation

436 proteins!

τ	N_τ^1	N_τ^2	κ	$\{B_j^\kappa\}$	n_κ	r
Nup192	1	1	1,2,5	●●	2	3.0
			3	-	1	-
Nup188	1	1	1,2,5	●●	2	3.0
			3	-	1	-
Nup170	1	1	1,2,5	●●	2	2.9
			3	-	1	-
Nup157	1	1	1,2,5	●●●	3	2.5
			3	-	1	-
Nup133	1	1	1,2,5	●●	2	2.7
			3	-	1	-
Nup120	1	1	1,2,5	●●	2	2.6
			3	-	1	-
Nup85	1	1	1,2,5	●●●	3	2.0
			3	-	1	-
Nup84	1	1	1,2,5	●●●	3	2.0
			3	-	1	-
Nup145C	1	1	1,2,5	●●	2	2.3
			3	-	1	-
Seh1	1	1	1,2,3,5	●	1	2.2
Sec13	1	1	1,2,3,5	●	1	2.1
Gle2	1	1	1,2,3,5	●	1	2.3
Nic96	2	2	1,2,5	●●	2	2.4
			3	-	1	-
Nup82	1	1	1,2,5	●●	2	2.3
			3	-	1	-
Nup1	0	1	1,5	●●●●●●●●●	9	1.5
			2	●●○○○○○○○	2	1.5
			3	-	1	-
			4	○○●●●●●●●	7	1.5
Nsp1	2	2	1,5	●●●●●●●●●●●●	12	1.3
			2	○○○○○○○○○●●●	3	1.3
			3	-	1	-
			4	●●●●●●●●●○○○	9	1.3
Gle1	1	0	1,2,5	●●	2	2.1
			3	-	1	-
Nup60	0	1	1,5	●●●●	4	1.6
			2,3	●○○○	1	1.6
			4	○●●●	3	1.6
Nup59	1	1	1,5	●●●●	4	1.6
			2	○○●●	2	1.6
			3	-	1	-
			4	●●○○	2	1.6
Nup57	1	1	1,5	●●●	3	1.8
			2,3	○○●	1	1.8
			4	●●○	2	1.8
Nup53	1	1	1,5	●●●	3	1.7
			2,3	○○●	1	1.7
			4	●●○	2	1.7
Nup145N	0	2	1,5	●●●●●●	6	1.5
			2,3	○○○○○●	1	1.5

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). *Nature*, *450*(7170), 695–701

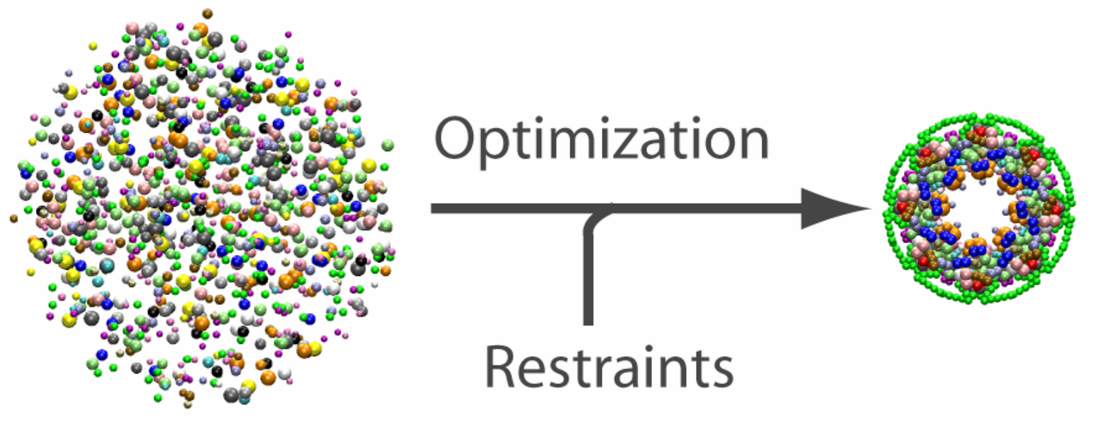
Scoring

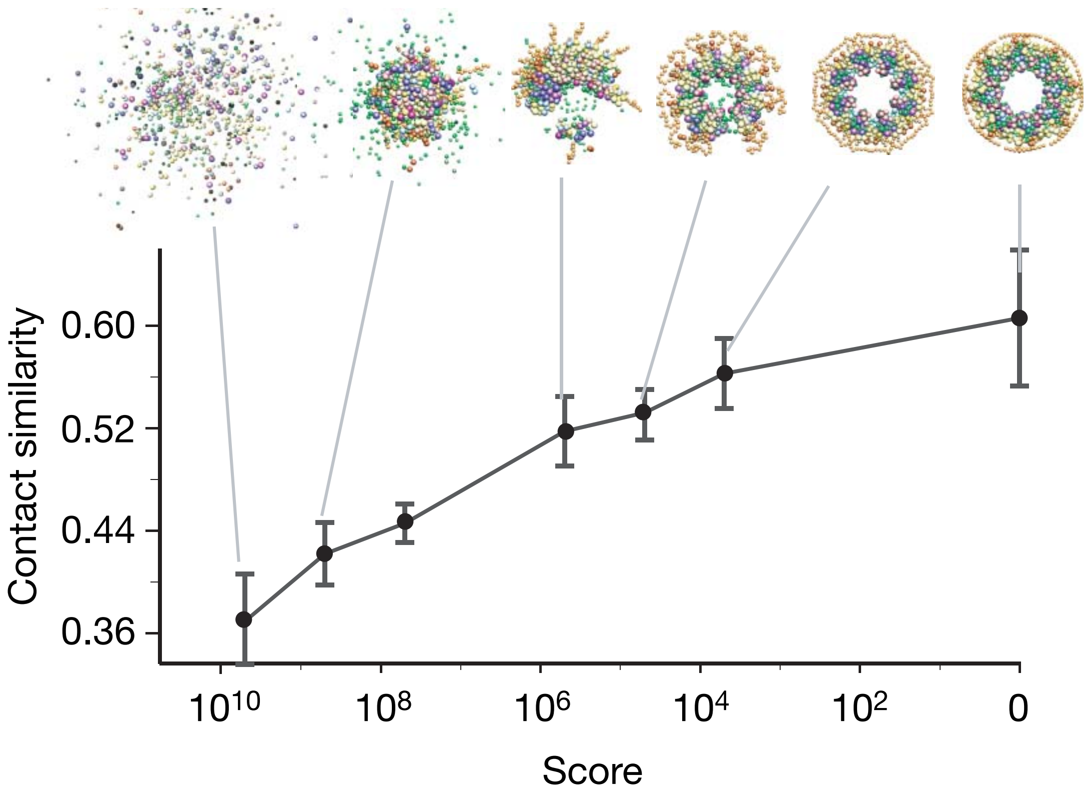


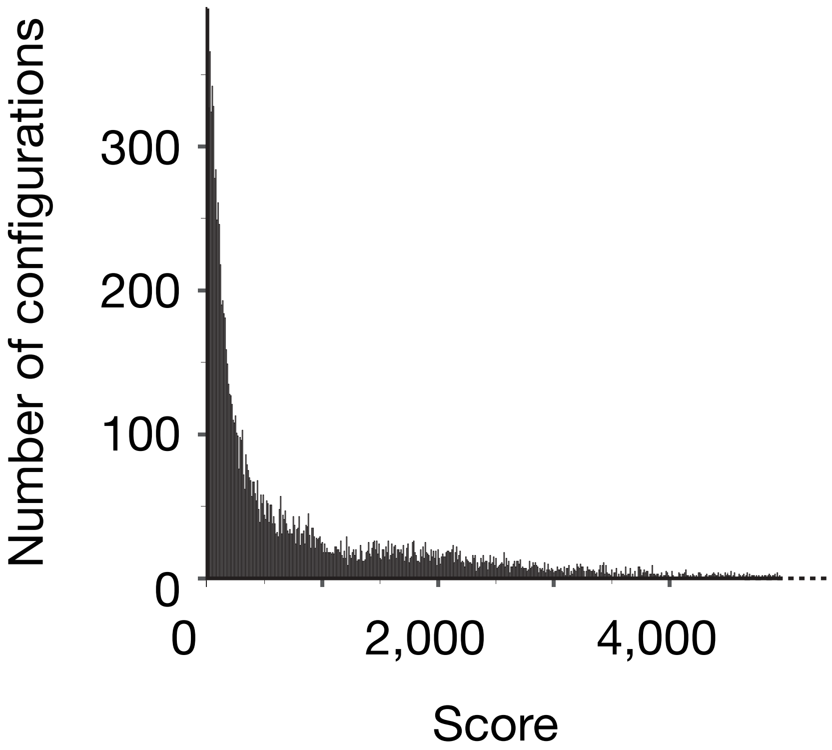
Data generation		Data interpretation				
Method	**Experiments**	**Restraint**	R_C	R_O	R_A	**Functional form of activated feature restraint**
Bioinformatics and Membrane fractionation	**30 nup sequences**	**Protein excluded volume restraint**	-	-	**1,864** ***1,863/2**	**Protein-protein**: Violated for $f < f_o$. f is the distance between two beads, f_o is the sum of the bead radii, and σ is 0.01 nm. Applied to all pairs of particles in representation κ=1: $B^{ms} = \left\{B_j^{\kappa=1}(\theta,s,\tau,i)\right\}$
	30 nup sequences	**Surface localization restraint**	-	-	**48**	**Membrane-surface location:** Violated if $f \neq f_o$. f is the distance between a protein particle and the closest point on the NE surface (half-torus), f_o = 0 nm, and σ is 0.2 nm. Applied to particles: $B^{ms} = \left\{B_j^{\kappa=6}(\theta,s,\tau,i) \mid \tau \in (\mathrm{Ndc1,Pom152,Pom34})\right\}$
	30 Nup sequences and immuno-EM (see below)		-	-	**64**	**Pore-side volume location:** Violated if $f < f_o$. f is the distance between a protein particle and the closest point on the NE surface (half-torus), f_o = 0 nm, and σ is 0.2 nm. Applied to particles: $B^{ms} = \left\{B_j^{\kappa=8}(\theta,s,\tau,i) \mid \tau \in (\mathrm{Ndc1,Pom152,Pom34})\right\}$
			-	-	**80**	**Perinuclear volume location:** Violated if $f > f_o$,, f is the distance between a protein particle and the closest point on the NE surface (half-torus), f_o = 0 nm, and σ is 0.2 nm. Applied to particles: $B^{ms} = \left\{B_j^{\kappa=7}(\theta,s,\tau,i)\tau \in (\mathrm{Pom152})\right\}$
Hydrodynamics experiments	**1 S-value**	**Complex shape restraint**	**1**	**164**	**1**	**Complex diameter** Violated if $f < f_o$. f is the distance between two protein particles representing the largest diameter of the largest complex, f_o is the complex maximal diameter $D=19.2-R$, where R is the sum of both particle radii, and σ is 0.01 nm. Applied to particles of proteins in composite C_{45}: $B^{ms} = \left\{B_j^{\kappa=1}(\theta,s,\tau,i) \mid \tau \in C_{51}\right\}$
	30 S-values	**Protein chain restraint**	-	-	**1,680**	**Protein chain** Violated if $f \neq f_o$. f is the distance between two consecutive particles in a protein, f_o is the sum of the particle radii, and σ is 0.01 nm. Applied to particles: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \kappa = 1\right\}$
Immuno-Electron microscopy	**10,940 gold particles**	**Protein localization restraint**	-	-	**456**	**Z-axial position** Violated for $f < f_o$. f is the absolute Cartesian Z-coordinate of a protein particle, f_o is the lower bound defined for protein type τ, and σ is 0.1 nm. Applied to particles: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \kappa = 1, j = 1\right\}$
					456	Violated for $f > f_o$. f is the absolute Cartesian Z-coordinate of a protein particle, f_o is the upper bound defined for protein type τ, and σ is 0.1 nm. Applied to particles: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \kappa = 1, j = 1\right\}$
			-	-	**456**	**Radial position** Violated for $f < f_o$. f is the radial distance between a protein particle and the Z-axis in a plane parallel to the X and Y axes, f_o is its lower bound defined for protein type τ, and σ is 0.1 nm. Applied to particles: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \kappa = 1, j = 1\right\}$
					456	Violated for $f > f_o$. f is the radial distance between a protein particle and the Z-axis in a plane parallel to the X and Y axes, f_o is its upper bound defined for protein type τ, and σ is 0.1 nm. Applied to particles: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \kappa = 1, j = 1\right\}$
Overlay assays	**13 contacts**	**Protein interaction restraint**	**20**	**112**	**20**	**Protein contact** Violated for $f > f_o$. f is the distance between two protein particles, f_o is the sum of the particle radii multiplied by a tolerance factor of 1.3, and σ is 0.01 nm. Applied to particle: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \kappa \in (2,4,9), \theta \in (1,2,3)\right\}$
Affinity purification	**4 complexes**	**Competitive binding restraint**	**1**	**132**	**4**	**Protein contact** Violated for $f > f_o$. f is the distance between two protein particles, f_o is the sum of the particle radii multiplied by a tolerance factor of 1.3, and σ is 0.01 nm. Applied to : $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \theta \in (1,2,3), \kappa \in (2,4,6), \tau = (Nup82, Nic96, Nup49, Nup57)\right\}$
	64 complexes	**Protein proximity restraint**	**692**	**25,348**	**692**	**Protein proximity** Violated for $f > f_o$. f is the distance between two protein particles, f_o is the maximal diameter of a composite complex, and σ is 0.01 nm. Applied to particles: $B = \left\{B_j^{\kappa}(\theta,s,\tau,i) \mid \theta \in (1,2,3), \kappa \in (2,4,9)\right\}$

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). *Nature*, *450*(7170), 695–701

Optimization

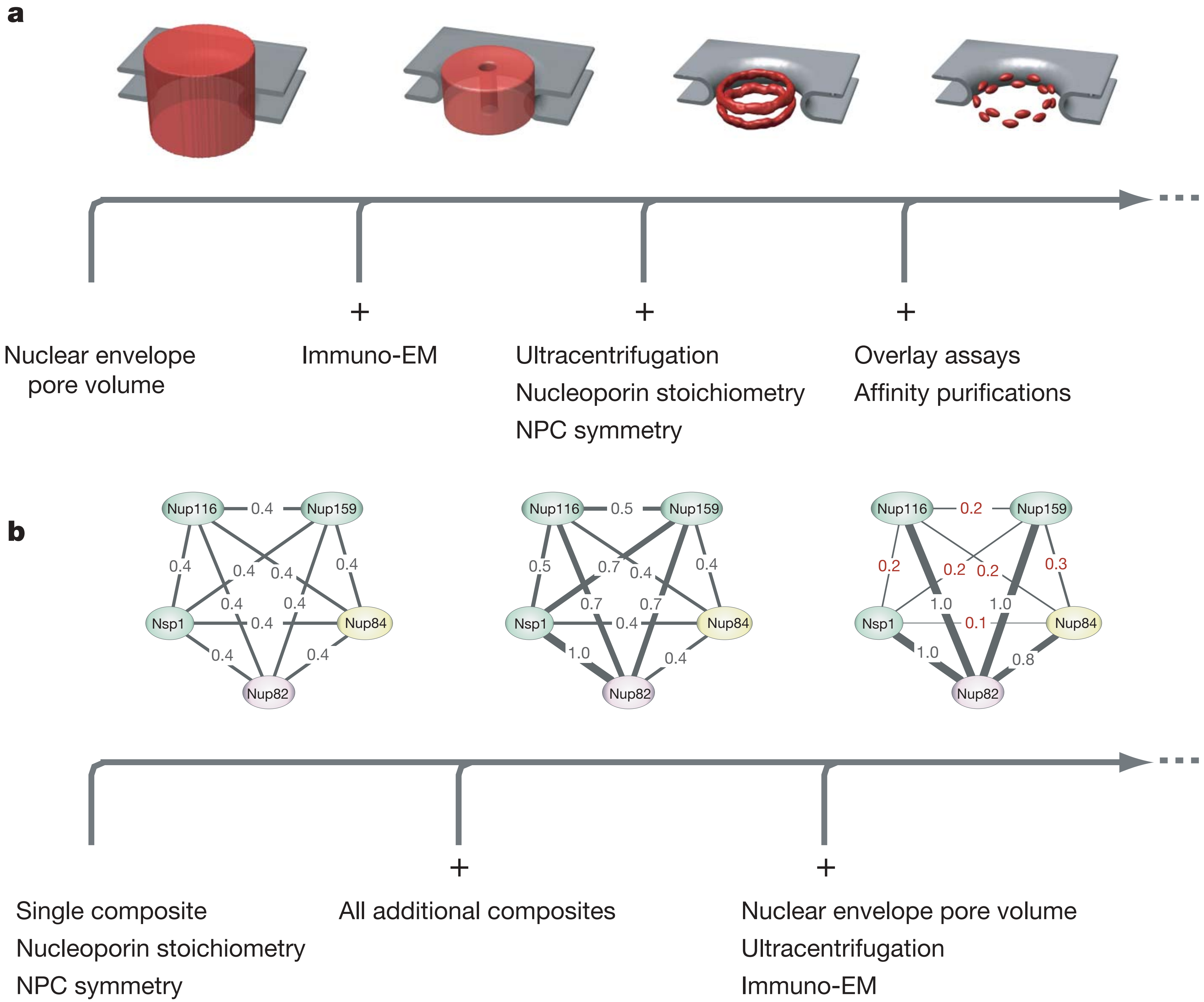






Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). *Nature*, *450*(7170), 695–701

Integrating data



Nuclear envelope pore volume

\+ Immuno-EM

\+ Ultracentrifugation
Nucleoporin stoichiometry
NPC symmetry

\+ Overlay assays
Affinity purifications

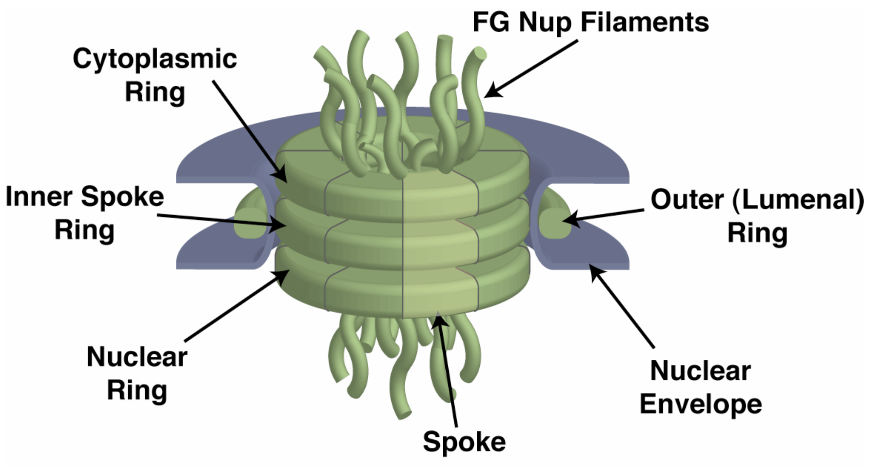
Single composite
Nucleoporin stoichiometry
NPC symmetry

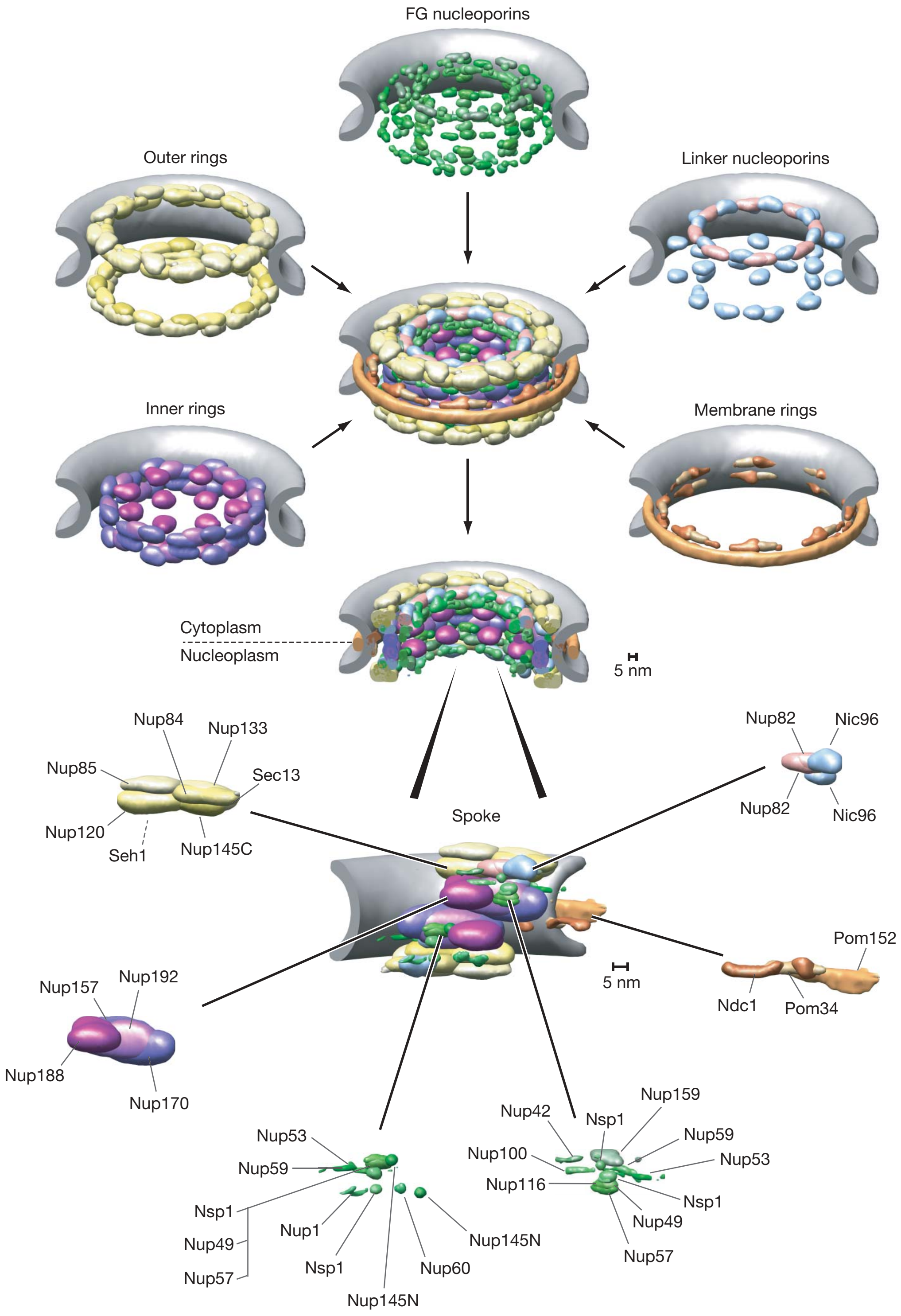
\+ All additional composites

\+ Nuclear envelope pore volume
Ultracentrifugation
Immuno-EM

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). *Nature*, *450*(7170), 695–701

The STRUCTURE of NPC

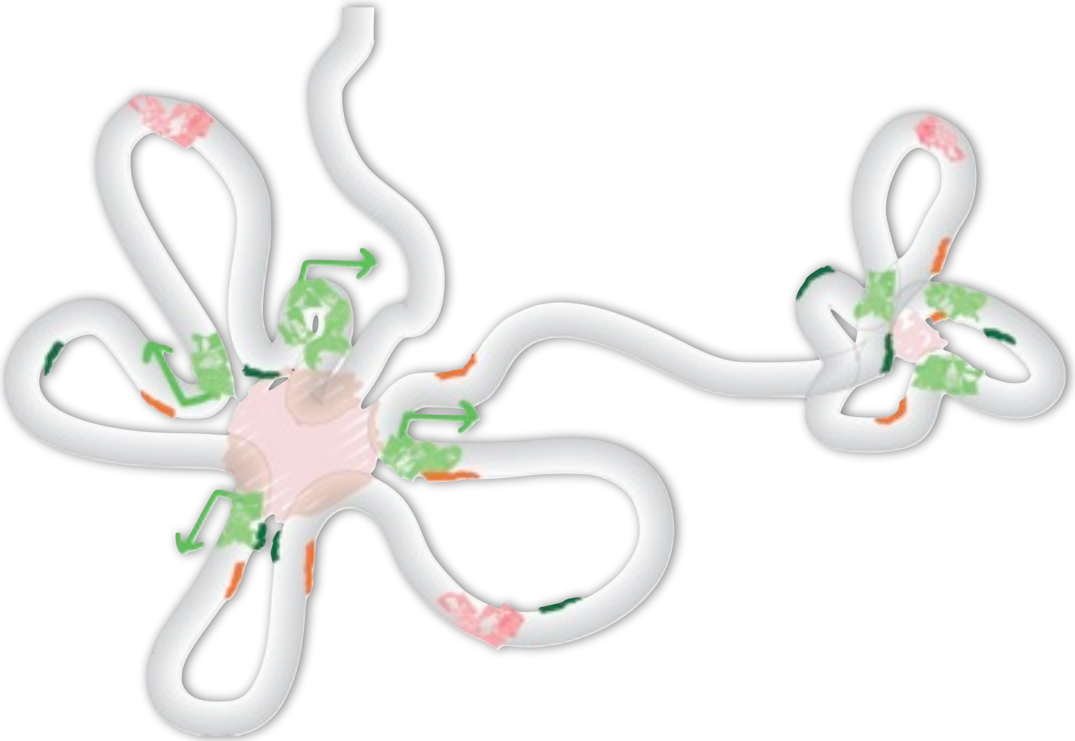


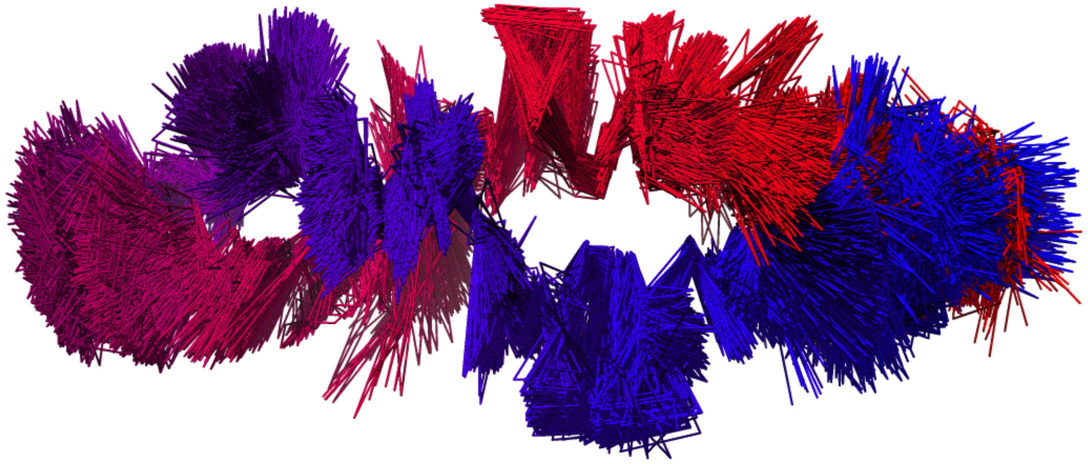


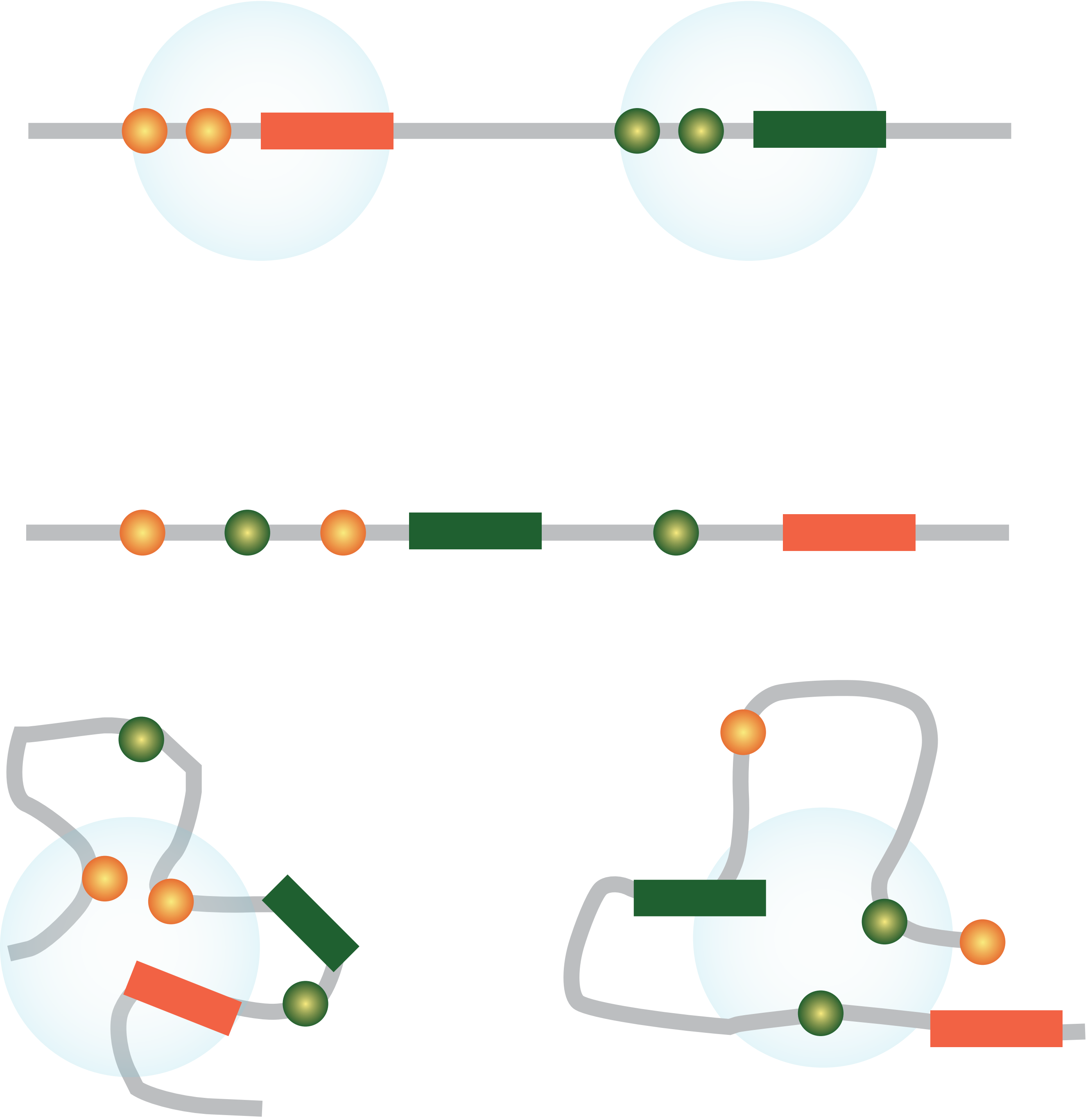
Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). *Nature*, *450*(7170), 695–701

GENOMES

limited *data types*



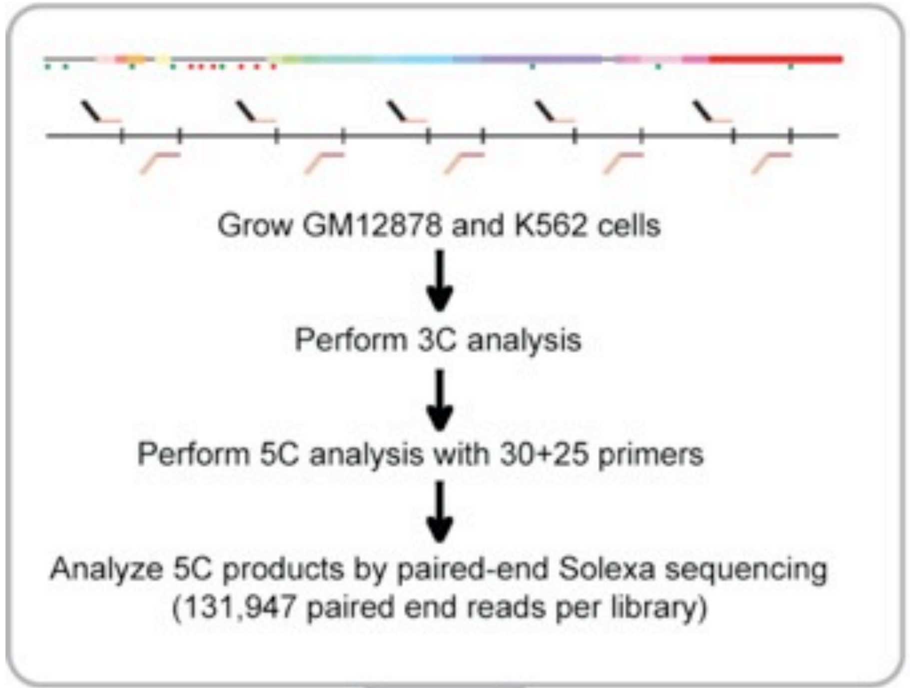




Simple genomes

Complex genomes

Experiments


Grow GM12878 and K562 cells

↓

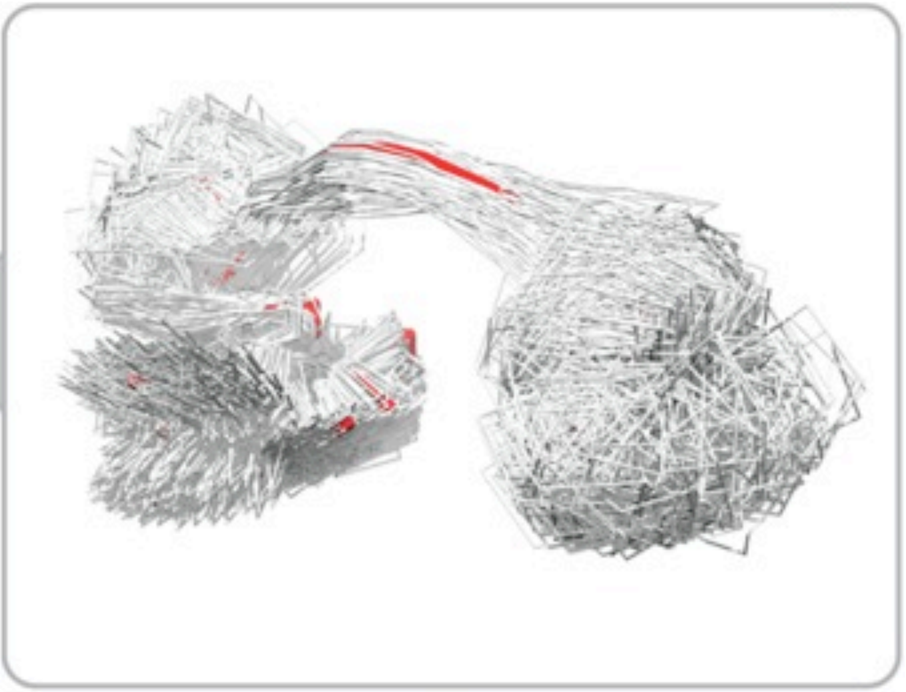
Perform 3C analysis

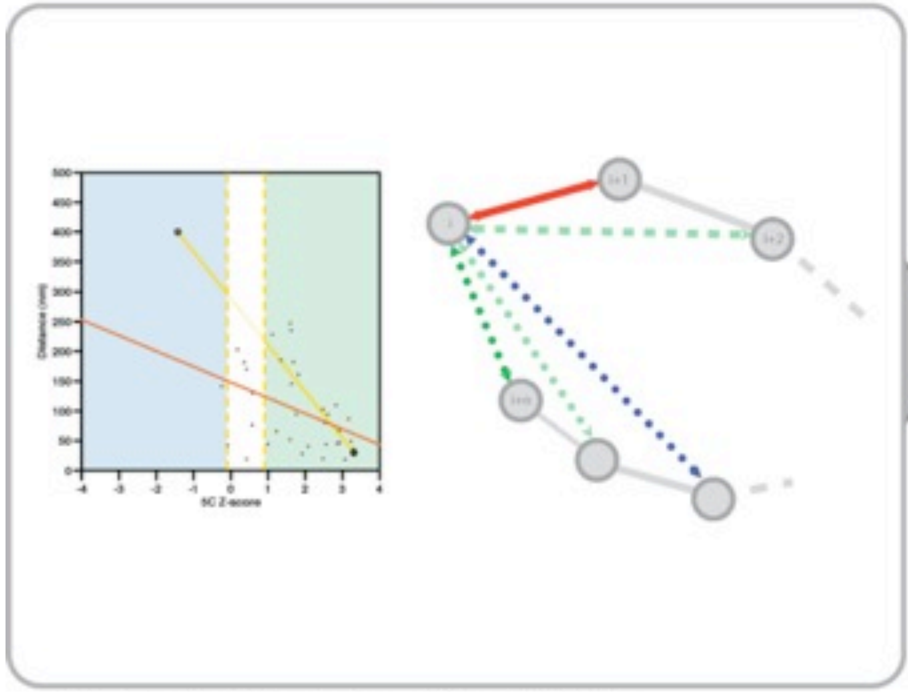
↓

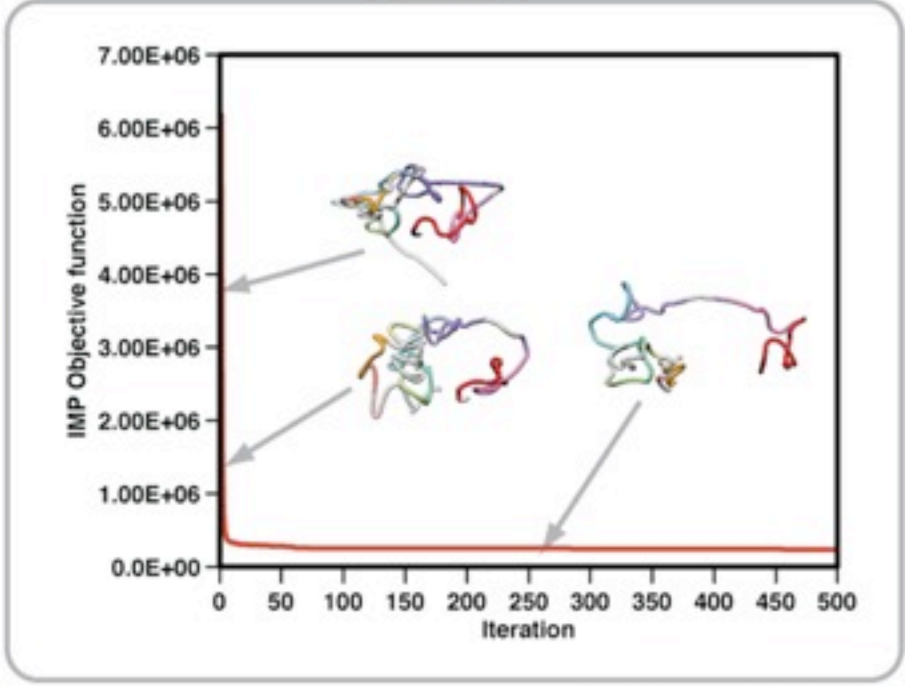
Perform 5C analysis with 30+25 primers

↓

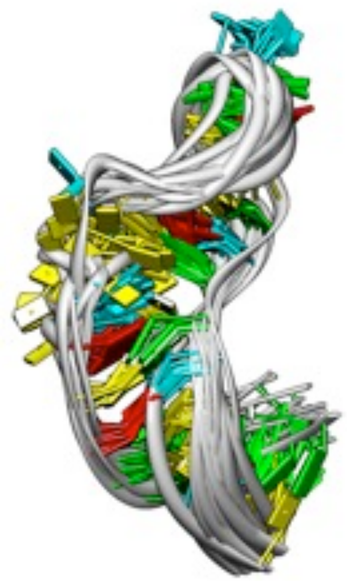
Analyze 5C products by paired-end Solexa sequencing
(131,947 paired end reads per library)

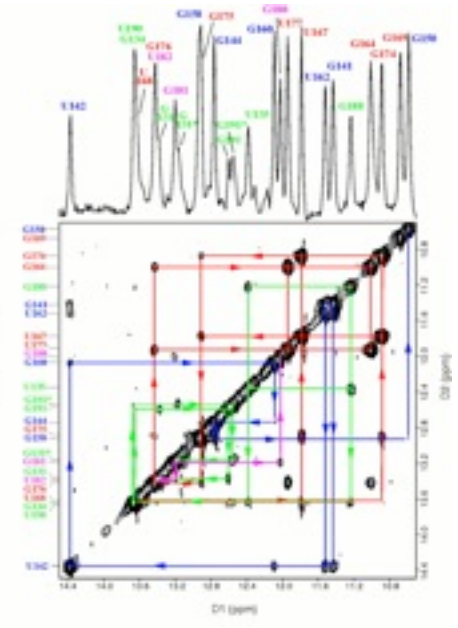




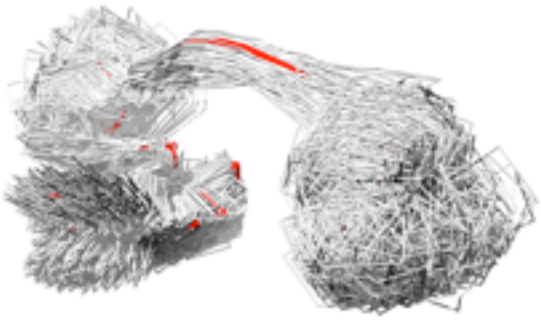


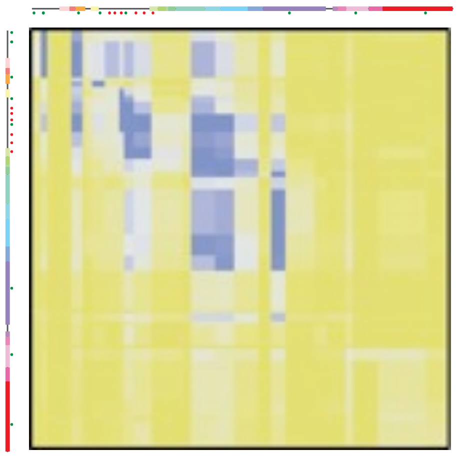
Computation





Biomolecular structure determination
2D-NOESY data

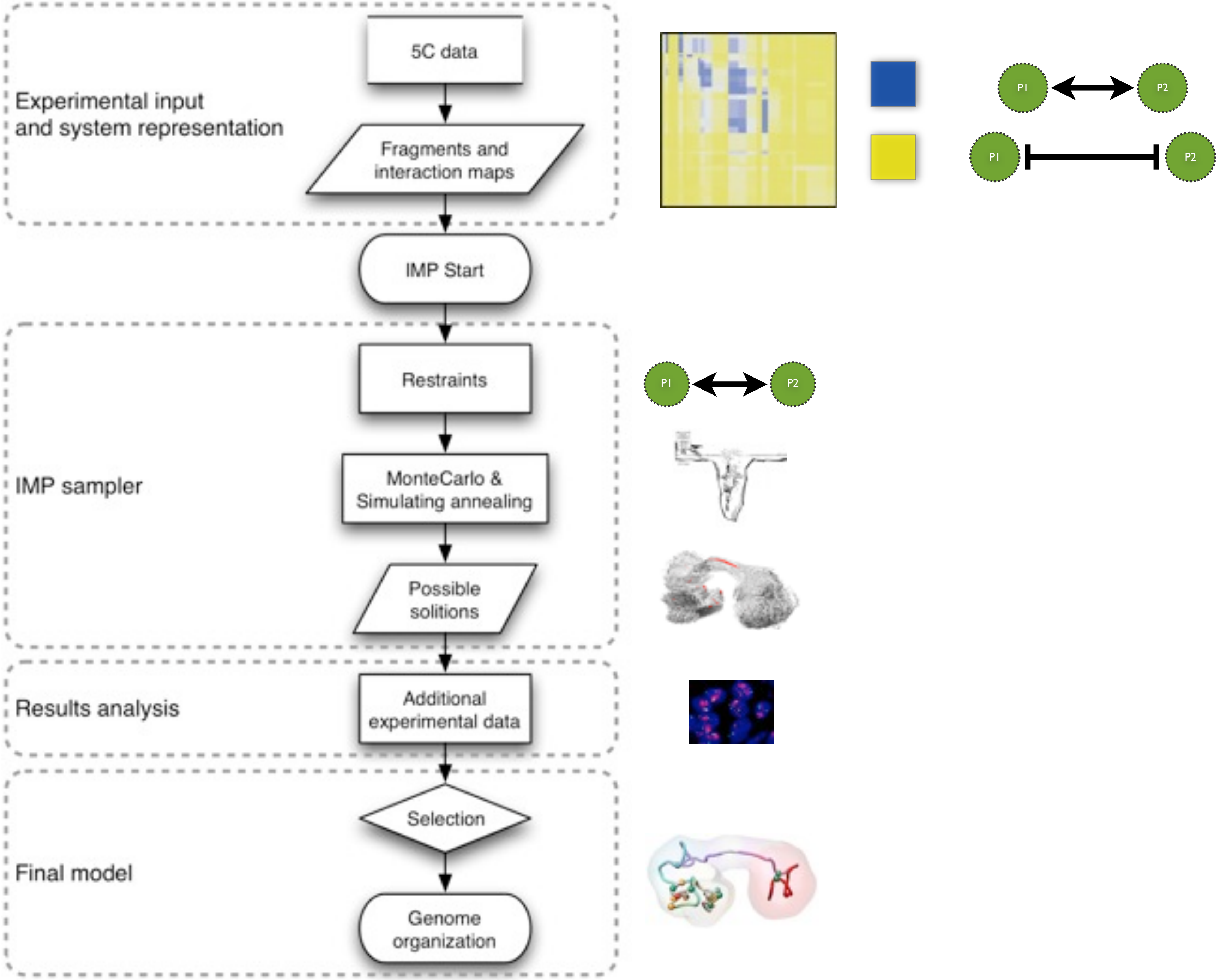




Chromosome structure determination
5C data

Integrative Modeling

http://www.integrativemodeling.org



Experimental input and system representation

5C data

Fragments and interaction maps

IMP Start

IMP sampler

Restraints

MonteCarlo & Simulating annealing

Possible solitions

Results analysis

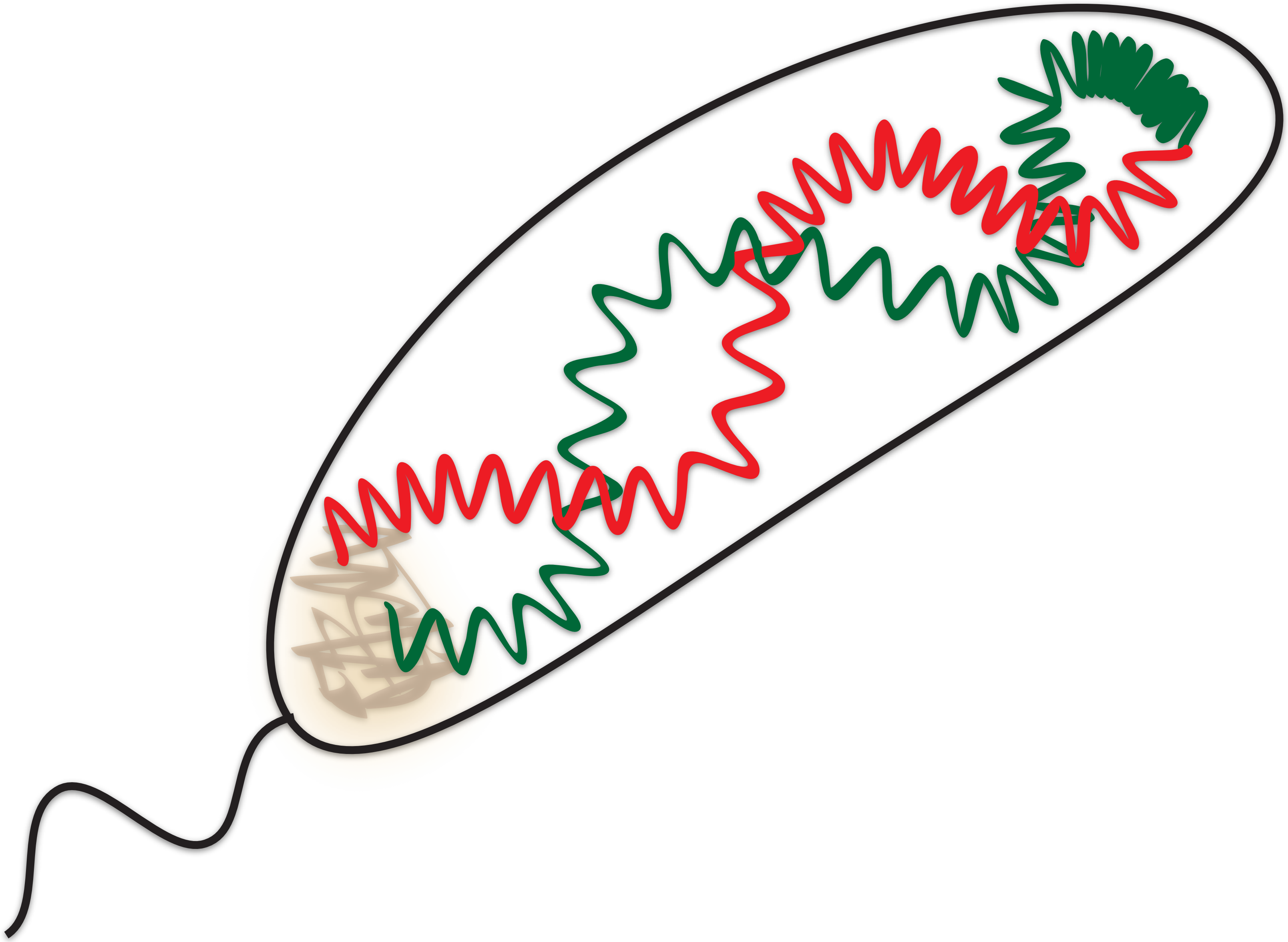
Additional experimental data

Final model

Selection

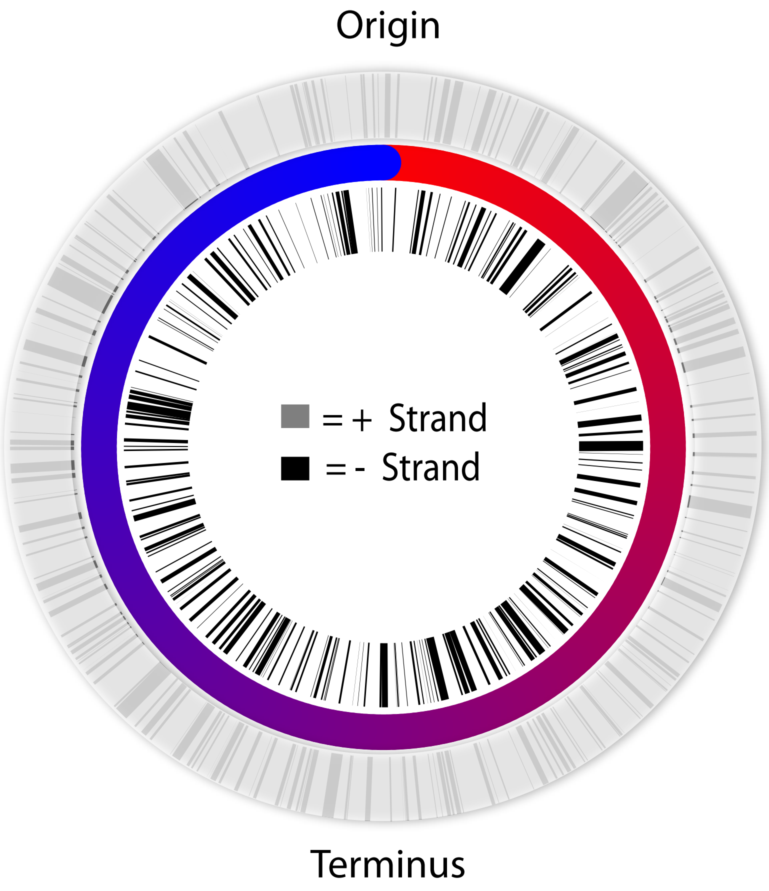
Genome organization

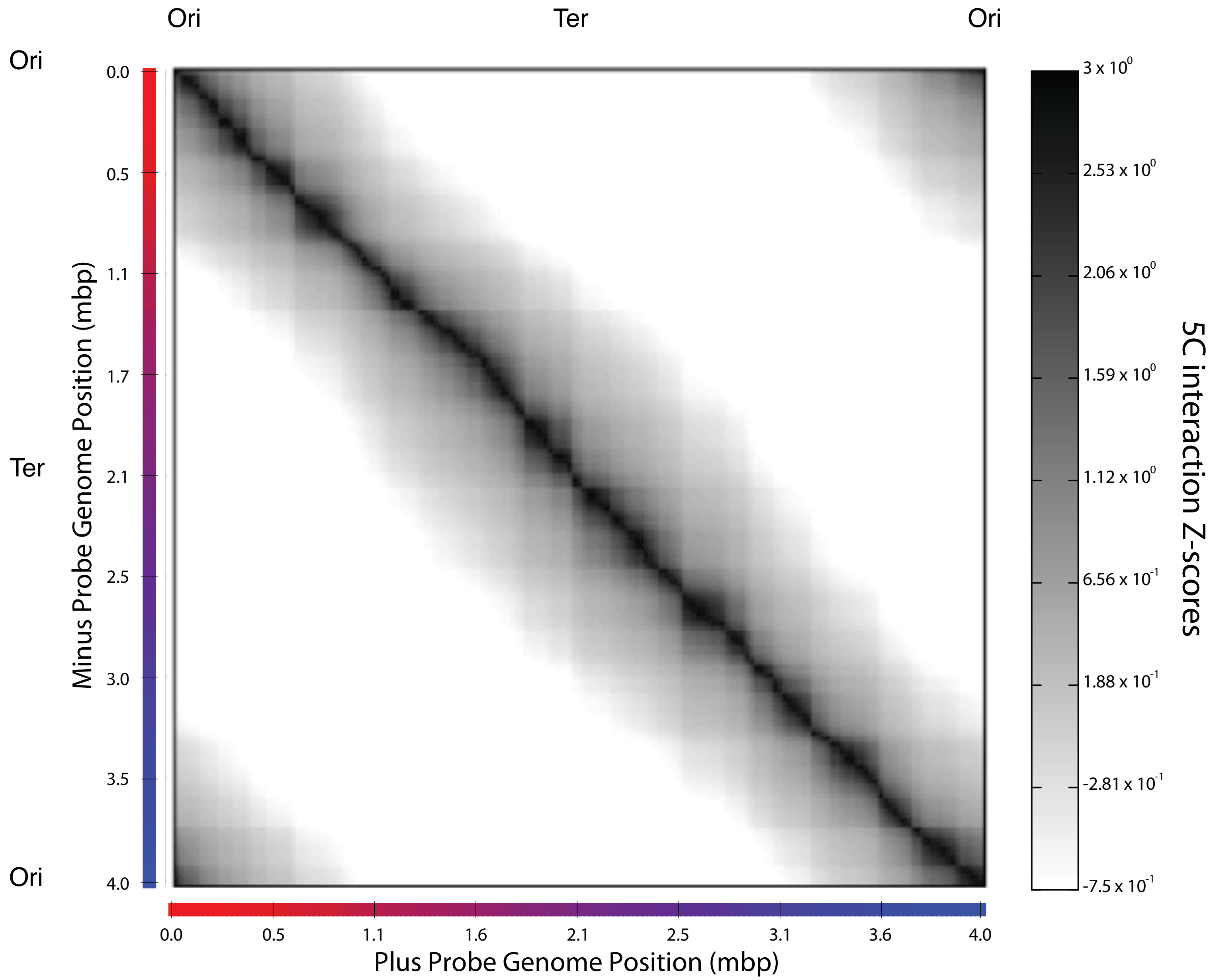
Caulobacter crescentus genome



The 3D architecture of Caulobacter Crescentus

4,016,942 bp & 3,767 genes





169 5C primers on + strand

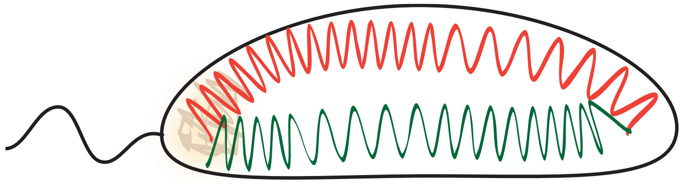
170 5C primers on – strand

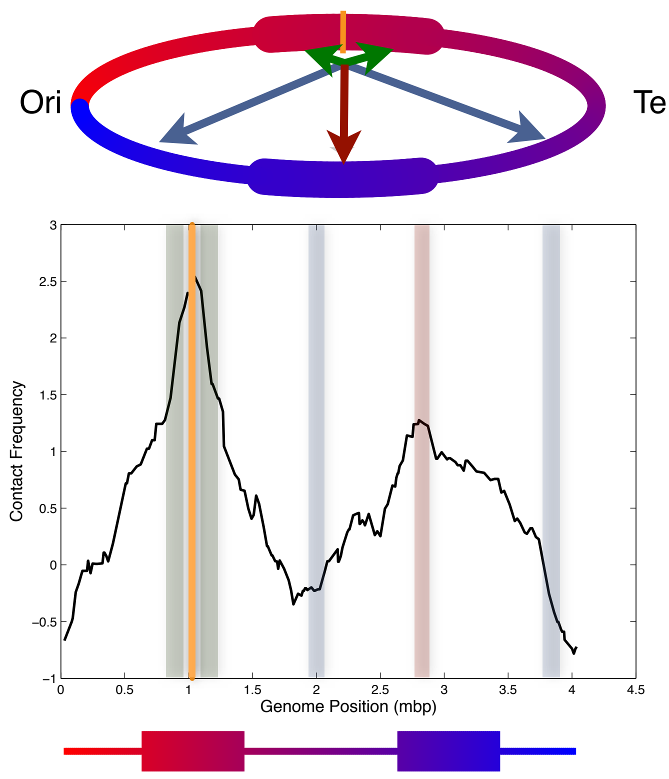
28,730 chromatin interactions

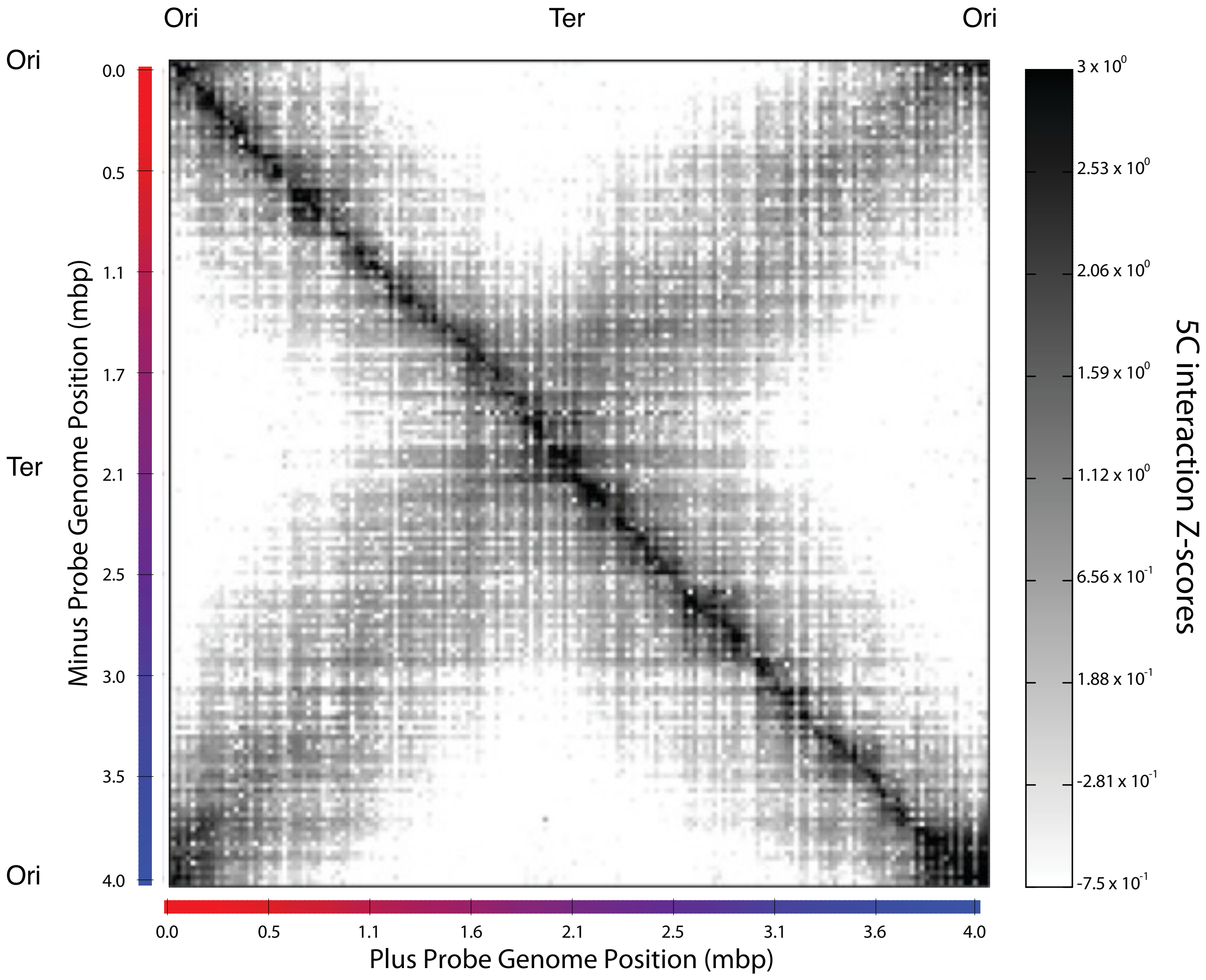
~13Kb

5C interaction matrix

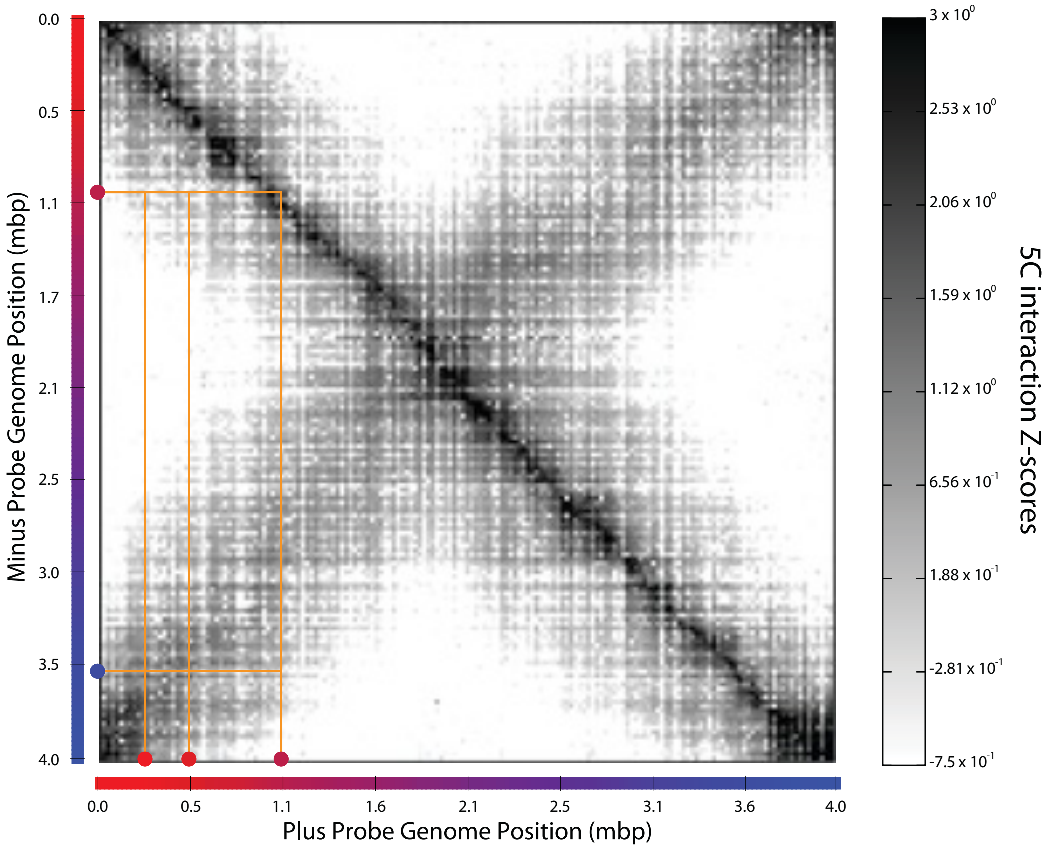
ELLIPSOID for Caulobacter cresentus



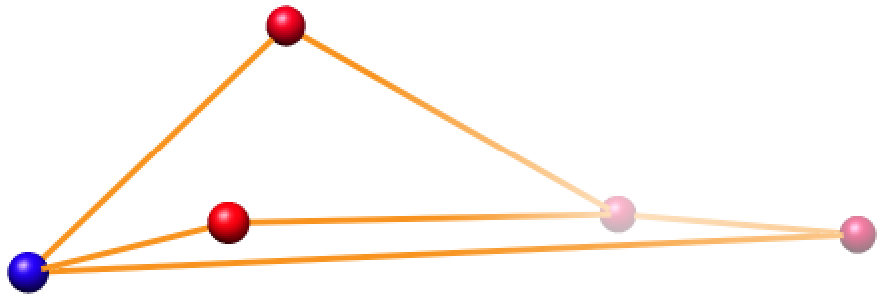


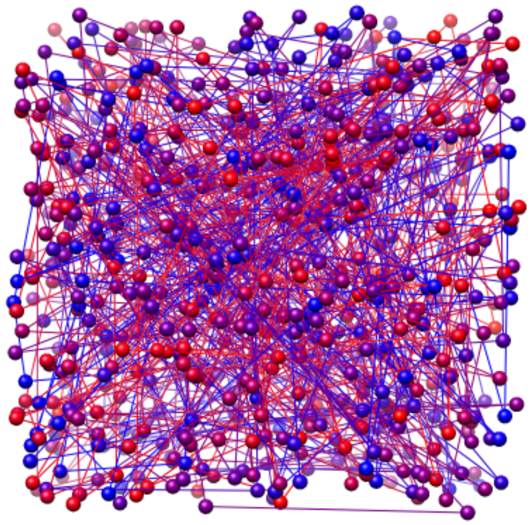


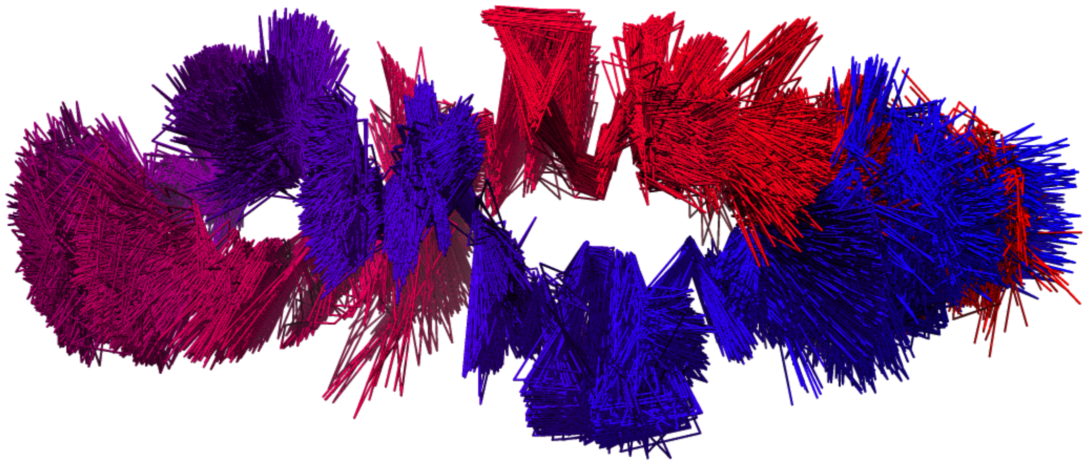
3D model building with the 5C + IMP approach



339 mers





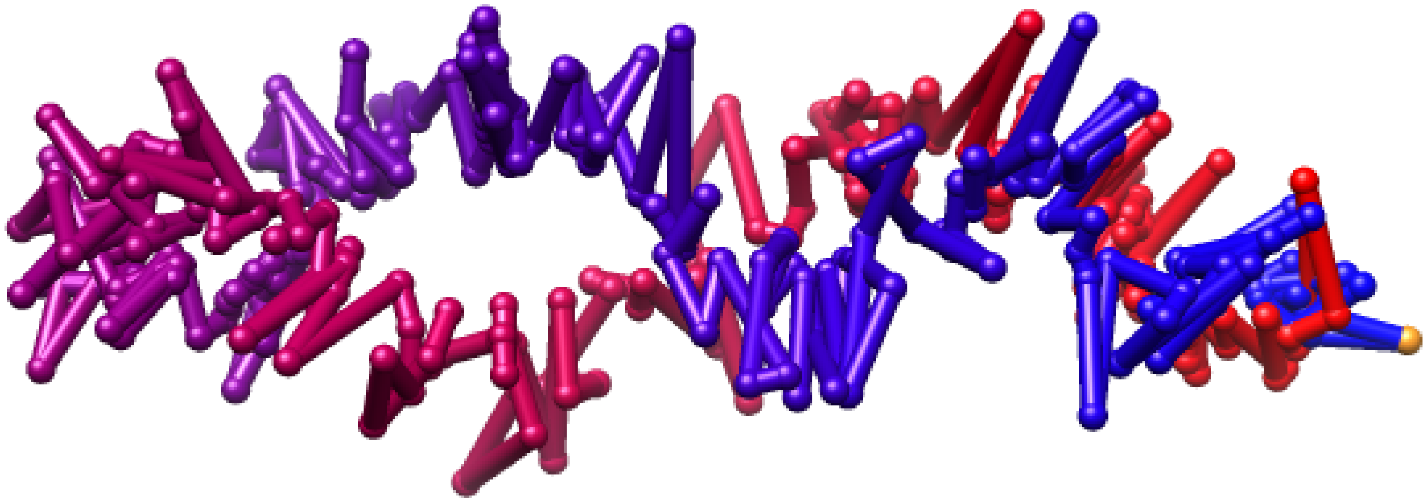


Genome organization in Caulobacter crescentus

Arms are helical

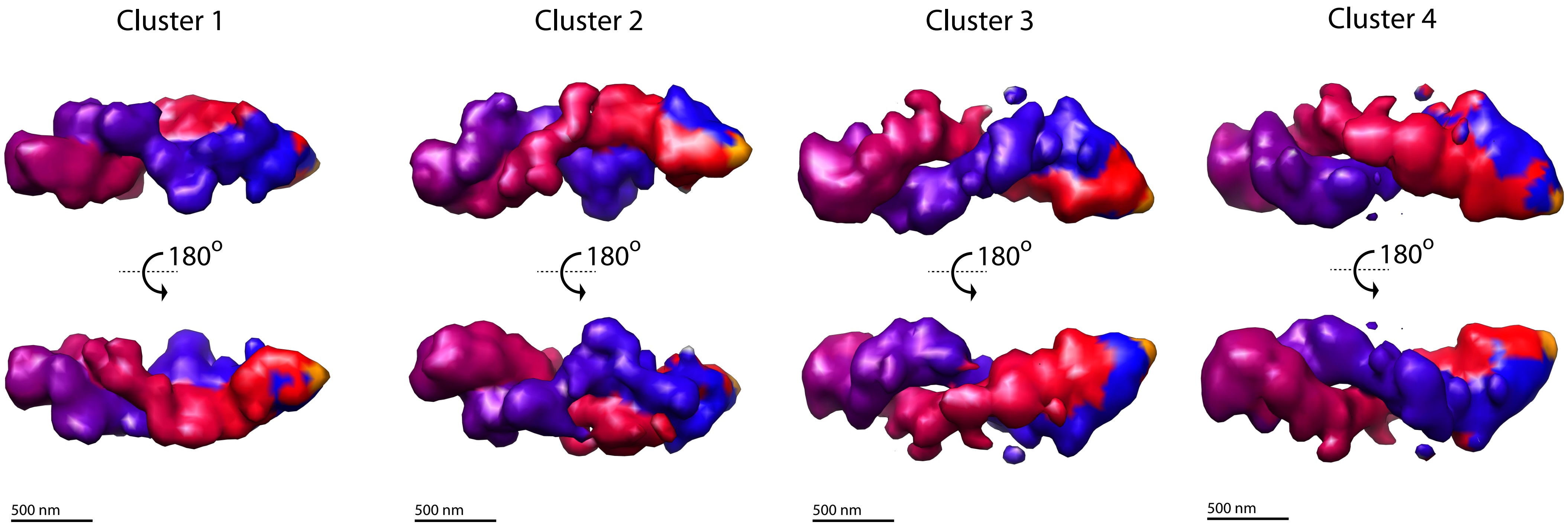
Resolution

dif site 47±17Kb from Ter



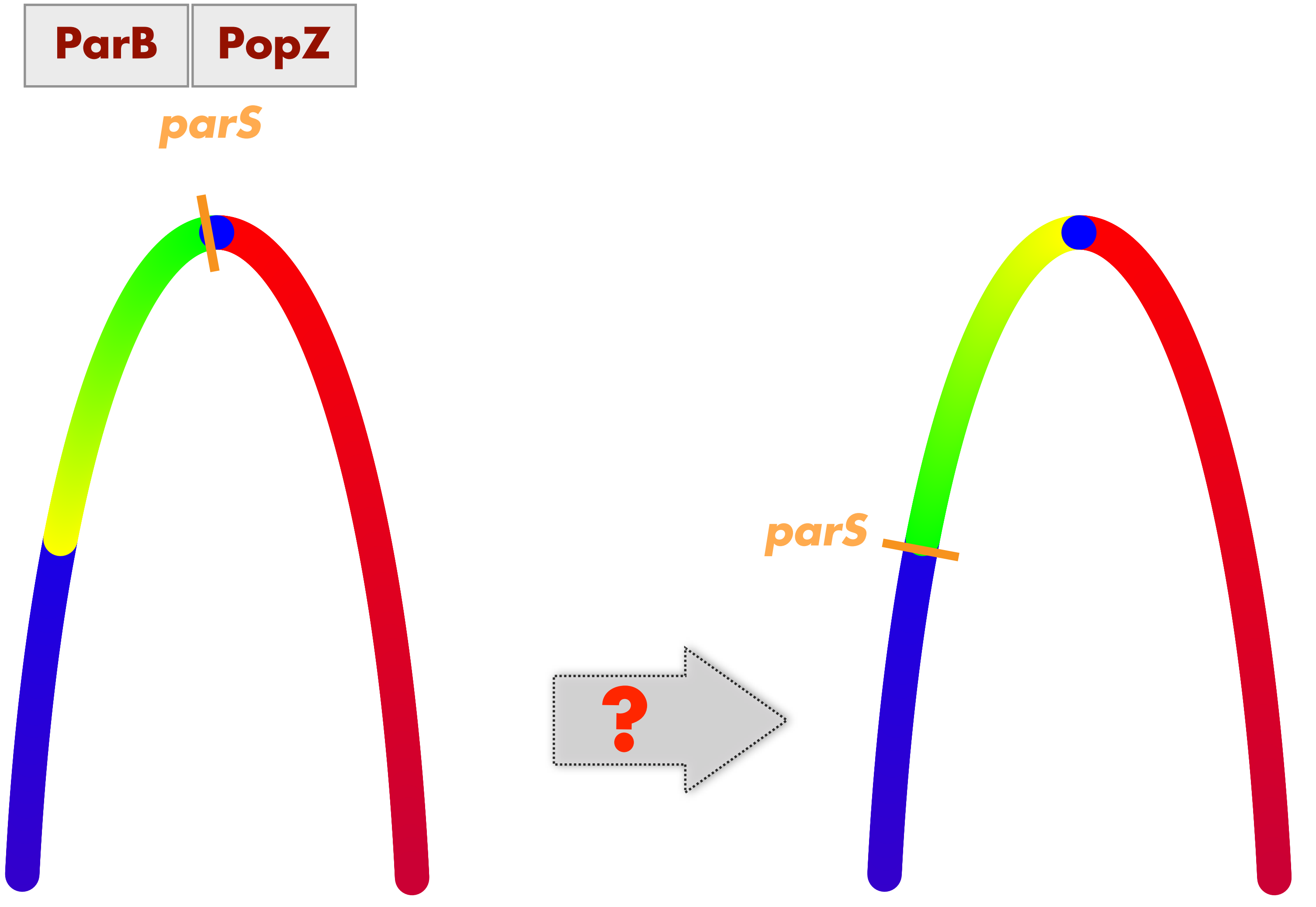
Centromer-like

parS sites 25±17Kb from Ori



MIRRORS!

Moving the parS sites 400 Kb away from Ori



ParB **PopZ**

parS

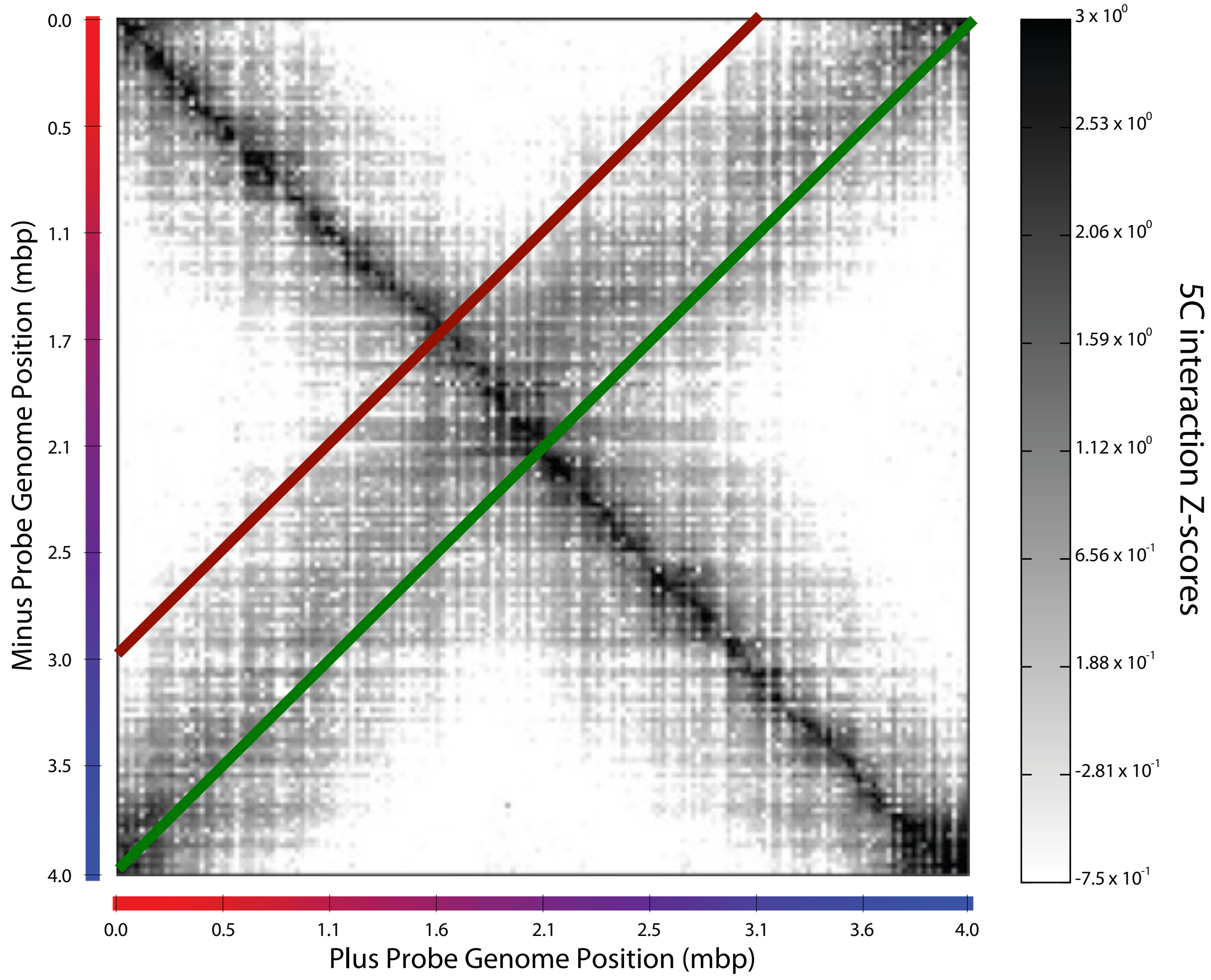
parS

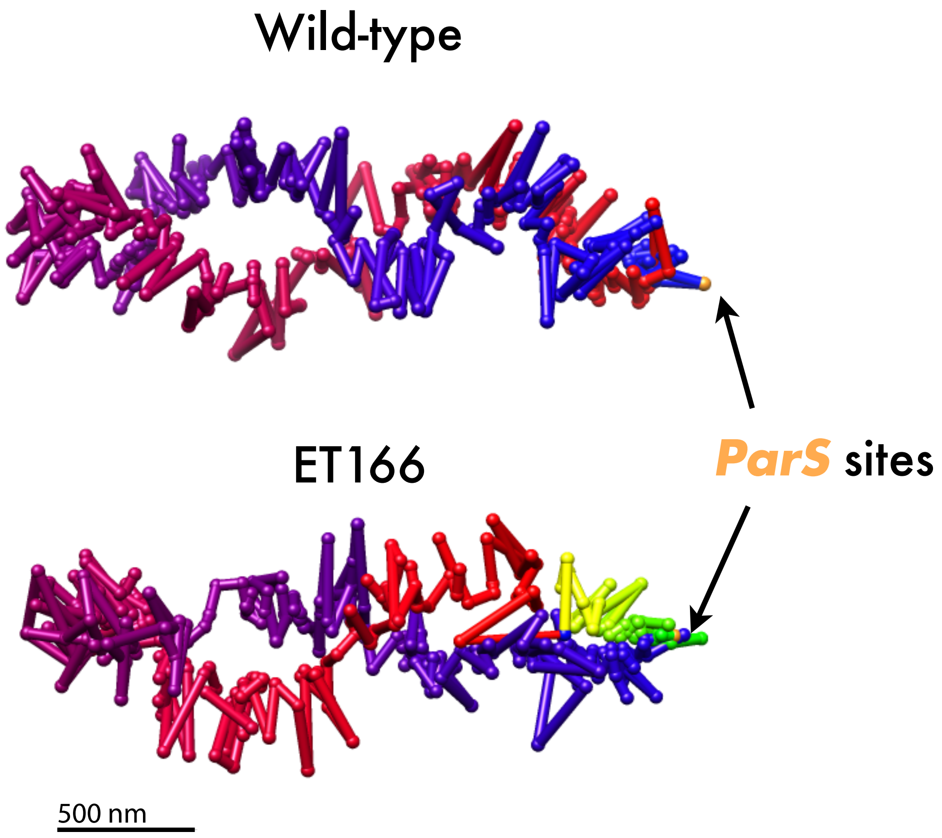
?

Wild-type

ET166

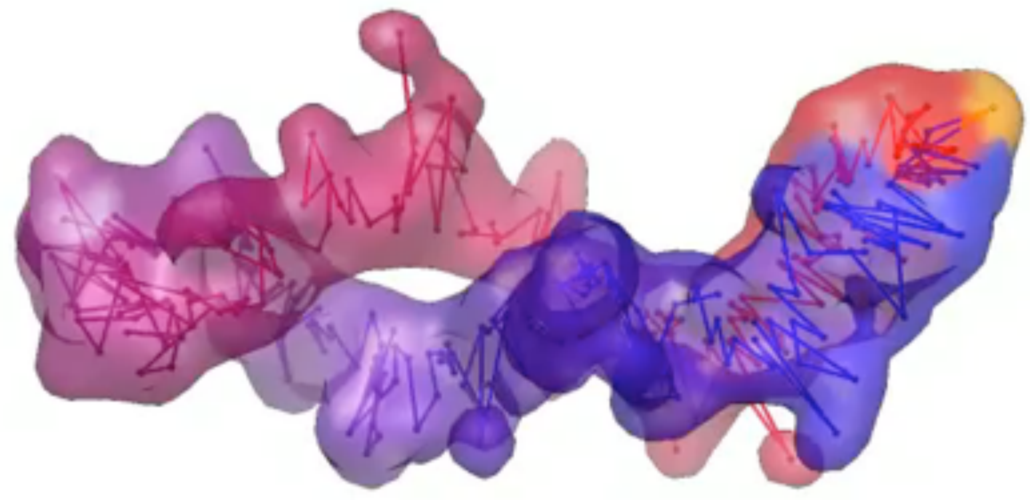
Moving the parS sites results in whole genome rotation!

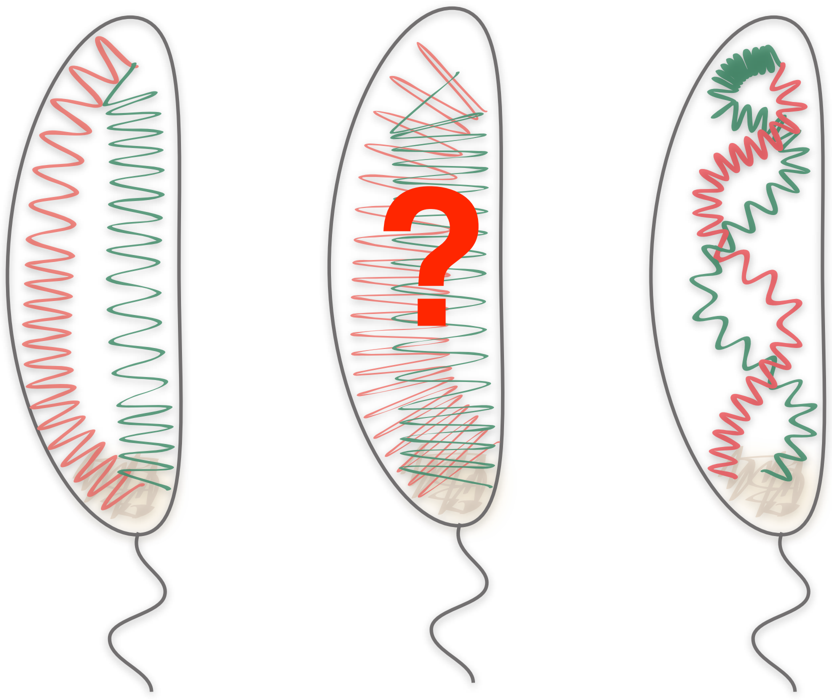


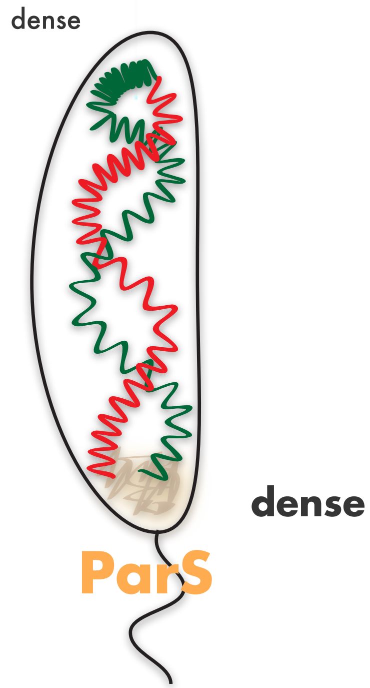


Arms are **STILL** helical

Genome architecture in Caulobacter





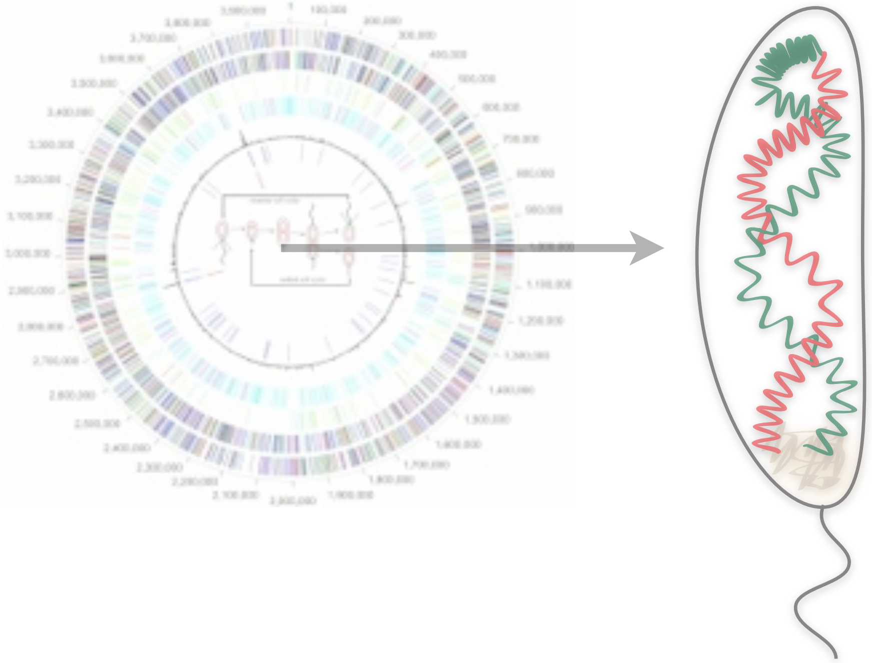


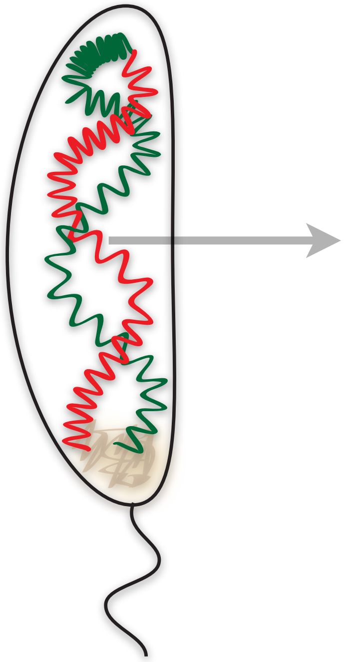
M.A. Umbarger, *et al.* **Molecular Cell** (2011) **44**:252–264

From Sequence to Function

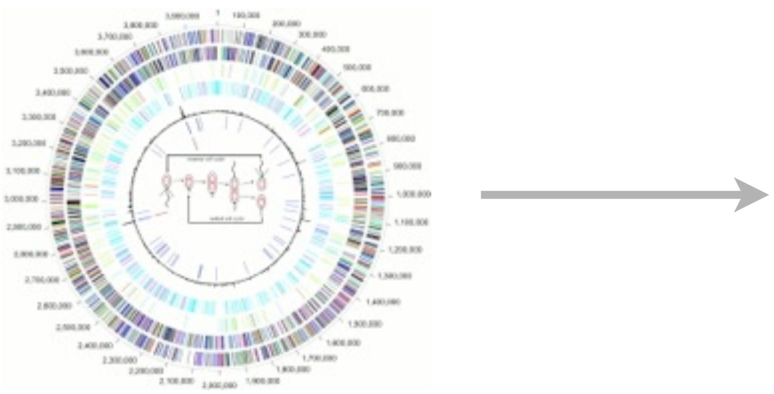
5C + IMP

Technology





Hypothesis

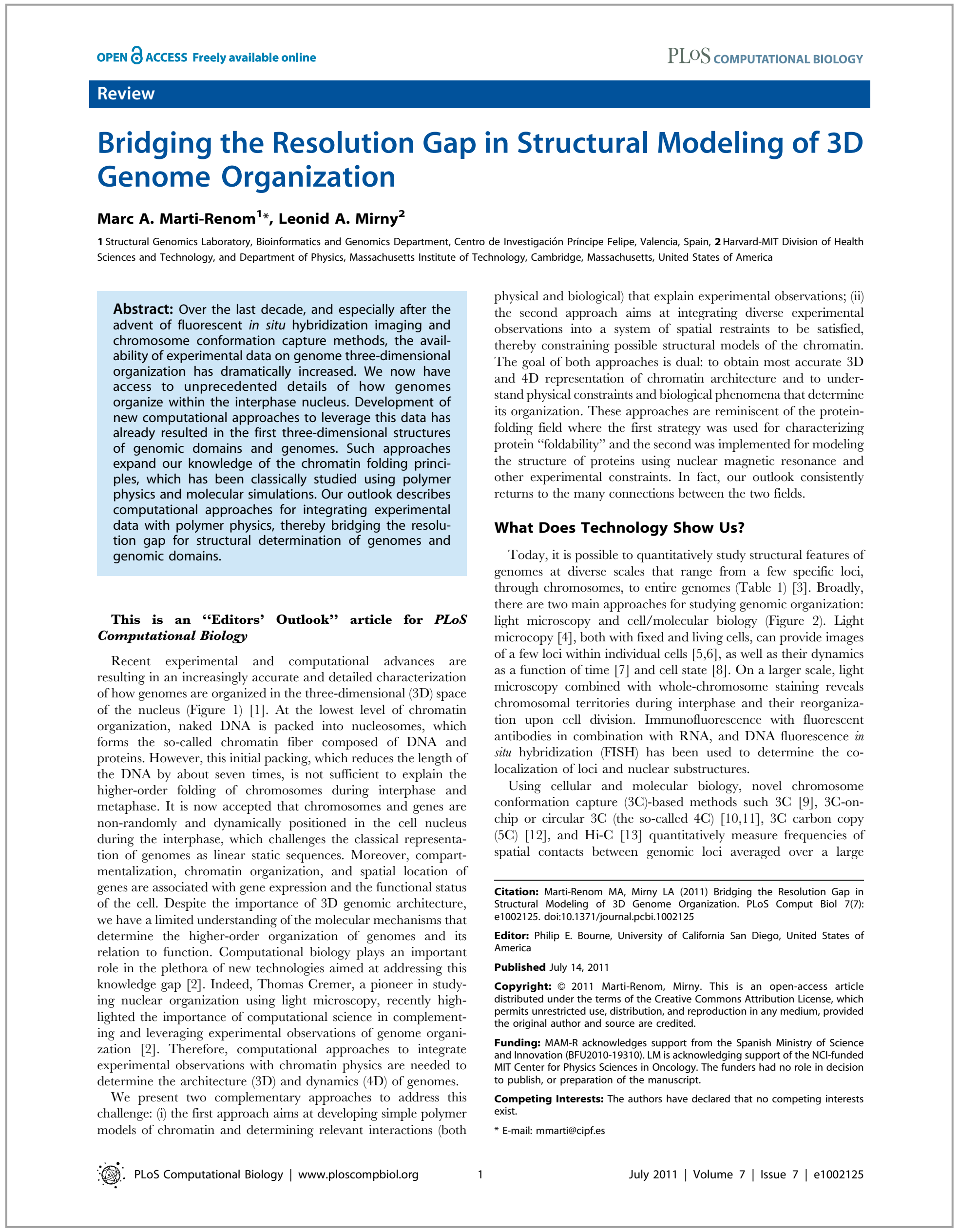


Function!

D. Baù and M.A. Marti-Renom **Chromosome Res** (2011) 19:25-35.

PLoS CB Outlook

Marti-Renom MA, Mirny LA (2011) PLoS Comput Biol 7(7): e1002125.



Review

Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization

Marc A. Marti-Renom[1]*, **Leonid A. Mirny**[2]

1 Structural Genomics Laboratory, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, **2** Harvard-MIT Division of Health Sciences and Technology, and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract: Over the last decade, and especially after the advent of fluorescent *in situ* hybridization imaging and chromosome conformation capture methods, the availability of experimental data on genome three-dimensional organization has dramatically increased. We now have access to unprecedented details of how genomes organize within the interphase nucleus. Development of new computational approaches to leverage this data has already resulted in the first three-dimensional structures of genomic domains and genomes. Such approaches expand our knowledge of the chromatin folding principles, which has been classically studied using polymer physics and molecular simulations. Our outlook describes computational approaches for integrating experimental data with polymer physics, thereby bridging the resolution gap for structural determination of genomes and genomic domains.

This is an ''Editors' Outlook'' article for *PLoS Computational Biology*

Recent experimental and computational advances are resulting in an increasingly accurate and detailed characterization of how genomes are organized in the three-dimensional (3D) space of the nucleus (Figure 1) [1]. At the lowest level of chromatin organization, naked DNA is packed into nucleosomes, which forms the so-called chromatin fiber composed of DNA and proteins. However, this initial packing, which reduces the length of the DNA by about seven times, is not sufficient to explain the higher-order folding of chromosomes during interphase and metaphase. It is now accepted that chromosomes and genes are non-randomly and dynamically positioned in the cell nucleus during the interphase, which challenges the classical representation of genomes as linear static sequences. Moreover, compartmentalization, chromatin organization, and spatial location of genes are associated with gene expression and the functional status of the cell. Despite the importance of 3D genomic architecture, we have a limited understanding of the molecular mechanisms that determine the higher-order organization of genomes and its relation to function. Computational biology plays an important role in the plethora of new technologies aimed at addressing this knowledge gap [2]. Indeed, Thomas Cremer, a pioneer in studying nuclear organization using light microscopy, recently highlighted the importance of computational science in complementing and leveraging experimental observations of genome organization [2]. Therefore, computational approaches to integrate experimental observations with chromatin physics are needed to determine the architecture (3D) and dynamics (4D) of genomes.

We present two complementary approaches to address this challenge: (i) the first approach aims at developing simple polymer models of chromatin and determining relevant interactions (both physical and biological) that explain experimental observations; (ii) the second approach aims at integrating diverse experimental observations into a system of spatial restraints to be satisfied, thereby constraining possible structural models of the chromatin. The goal of both approaches is dual: to obtain most accurate 3D and 4D representation of chromatin architecture and to understand physical constraints and biological phenomena that determine its organization. These approaches are reminiscent of the protein-folding field where the first strategy was used for characterizing protein ''foldability'' and the second was implemented for modeling the structure of proteins using nuclear magnetic resonance and other experimental constraints. In fact, our outlook consistently returns to the many connections between the two fields.

What Does Technology Show Us?

Today, it is possible to quantitatively study structural features of genomes at diverse scales that range from a few specific loci, through chromosomes, to entire genomes (Table 1) [3]. Broadly, there are two main approaches for studying genomic organization: light microscopy and cell/molecular biology (Figure 2). Light microcopy [4], both with fixed and living cells, can provide images of a few loci within individual cells [5,6], as well as their dynamics as a function of time [7] and cell state [8]. On a larger scale, light microscopy combined with whole-chromosome staining reveals chromosomal territories during interphase and their reorganization upon cell division. Immunofluorescence with fluorescent antibodies in combination with RNA, and DNA fluorescence *in situ* hybridization (FISH) has been used to determine the co-localization of loci and nuclear substructures.

Using cellular and molecular biology, novel chromosome conformation capture (3C)-based methods such 3C [9], 3C-on-chip or circular 3C (the so-called 4C) [10,11], 3C carbon copy (5C) [12], and Hi-C [13] quantitatively measure frequencies of spatial contacts between genomic loci averaged over a large

---

Citation: Marti-Renom MA, Mirny LA (2011) Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization. PLoS Comput Biol 7(7): e1002125. doi:10.1371/journal.pcbi.1002125

Editor: Philip E. Bourne, University of California San Diego, United States of America

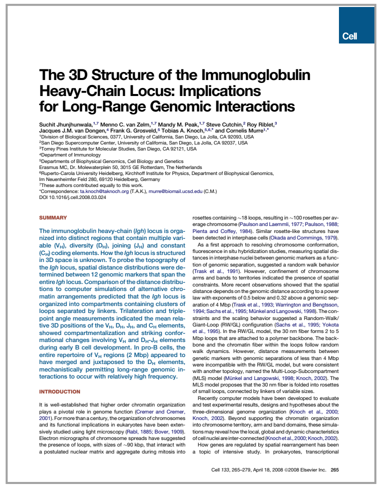
Published July 14, 2011

Funding: MAM-R acknowledges support from the Spanish Ministry of Science and Innovation (BFU2010-19310). LM is acknowledging support of the NCI-funded MIT Center for Physics Sciences in Oncology. The funders had no role in decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

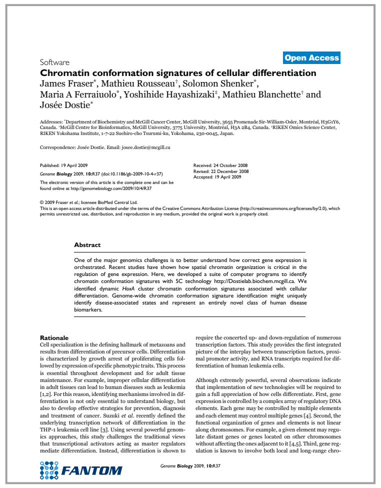
* E-mail: mmarti@cipf.es

MURRE
Cell (2008) **133**:265-79



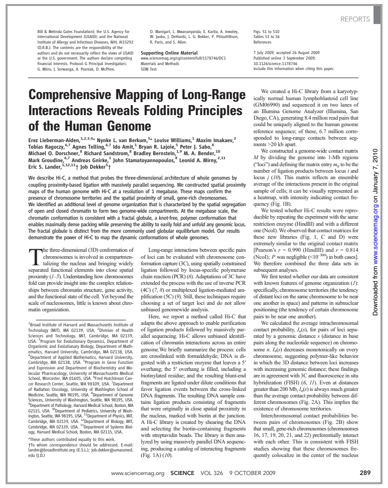
The 3D Structure of the Immunoglobulin Heavy-Chain Locus: Implications for Long-Range Genomic Interactions

DOSTIE/BLANCHETTE
Genome Biol (2009) **10**: R37



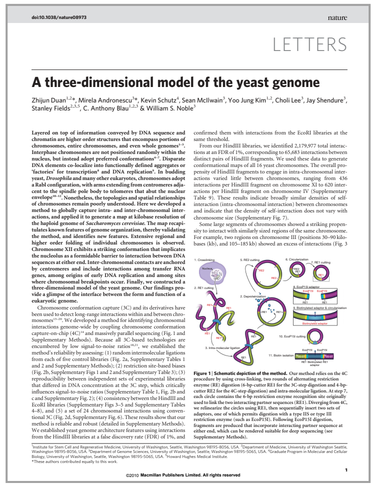
Chromatin conformation signatures of cellular differentiation

DEKKER/LANDER/MIRNY
Science (2009) **326**:289-93



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

NOBLE
Nature (2010) **465**: 363-7

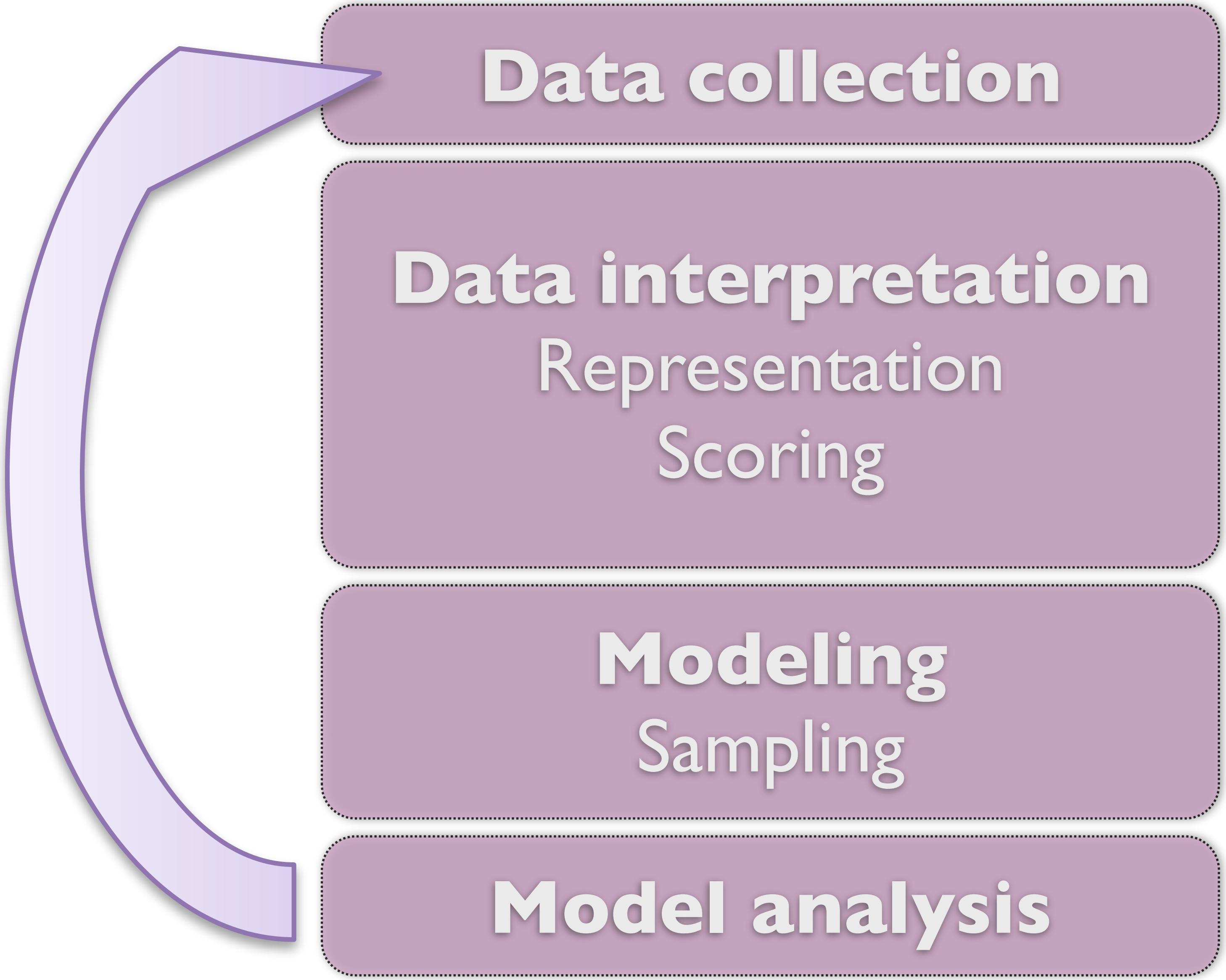


A three-dimensional model of the yeast genome

DEKKER/MARTI-RENOM
NSMB (2011) **18**:107-14

The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules

Take home message



Data collection

Data interpretation

Representation

Scoring

Modeling

Sampling

Model analysis

Acknowledgments

http://marciuslab.org
http://cnag.cat
http://integrativemodeling.org

OPEN POSITIONS!
Starting spring 2012

COMPARATIVE MODELING
Andrej Sali
M. S. Madhusudhan
Narayanan Eswar
Min-Yi Shen
Ursula Pieper
Ben Webb
Maya Topf (Birbeck College)

MODEL ASSESSMENT
David Eramian
Min-Yi Shen
Damien Devos

FUNCTIONAL ANNOTATION
Andrea Rossi (Rinat-Pfizer)
Fred Davis (Janelia Fram)

FUNDING
CNAG
MINECO
Era-Net Pathogenomics
HFSP

MODEL ASSESSMENT
Francisco Melo (CU)
Alejandro Panjkovich (CU)

NMR
Antonio Pineda-Lucena
Leticia Ortí
Rodrigo J. Carbajo

MAMMOTH
Angel R. Ortiz

3D Genomes
George Church (Harvard)
Job Dekker (UMASS)
Jeane Lawrence (UMASS)
Lucy Shapiro (Stanford)

BIOLOGY
Jeff Friedman (RU)
James Hudsped (RU)
Partho Ghosh (UCSD)
Alvaro Monteiro (Cornell U)
Stephen Krilis (St.George H)

Tropical Disease Initiative
Stephen Maurer (UC Berkeley)
Arti Rai (Duke U)
Andrej Sali (UCSF)
Ginger Taylor (TSL)
Matthew Todd (U Sydney)

CCPR Functional Proteomics
Patsy Babbitt (UCSF)
Fred Cohen (UCSF)
Ken Dill (UCSF)
Tom Ferrin (UCSF)
John Irwin (UCSF)
Matt Jacobson (UCSF)
Tack Kuntz (UCSF)
Andrej Sali (UCSF)
Brian Shoichet (UCSF)
Chris Voigt (UCSF)

EVA
Burkhard Rost (Columbia U)
Alfonso Valencia (CNB/UAM)

GeMoA
LLuís Ballell (GSK)
Brigitte Gicquel (IP)
Olivier Neyrolles (IPBS)
Marc A. Marti-Renom (CNAG)
Matthias Wilmanns (EMBL)