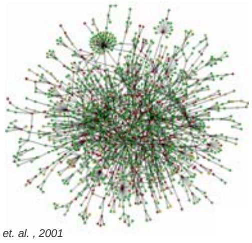
Atomic Resolution Modeling of Large Macromolecular Assemblies

Haim J. Wolfson
School of Computer Science
Tel Aviv University

Complexes as functional modules of the cell

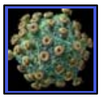
Protein-Protein interaction network



Jeong et. al. , 2001

Complexes

ATP synthase

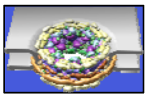


Virus

26S proteasome

Chaperonin

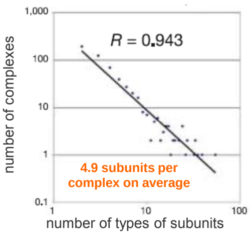
…



Nuclear pore complex

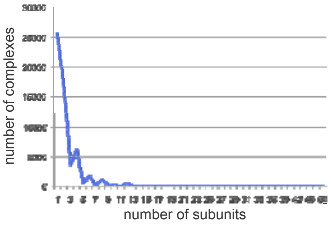
Protein complex size statistics

distribution of complex size in yeast



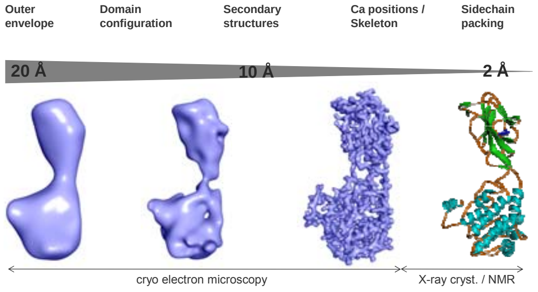
Krogan et. al., Nature,2006

protein data bank



There are thousands of biologically relevant macromolecular complexes whose structures are yet to be characterized.

Experimental techniques for Protein Structure determination



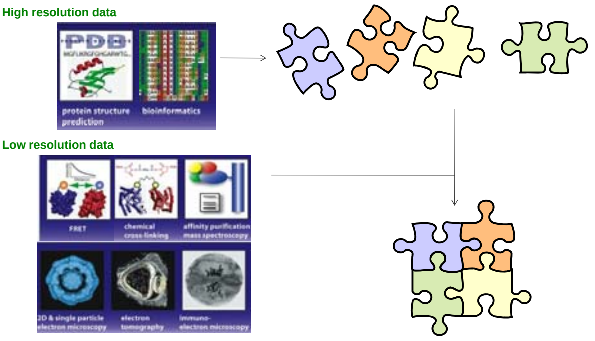
Use hybrid methods to bridge the resolution gaps

Analogy

Multi-molecular assembly is analogous to the solution of 3D puzzles –a classical spatial Pattern Discovery task.

High resolution data

Low resolution data


PDB
protein structure prediction
bioinformatics
FRET
chemical cross-linking
affinity purification mass spectroscopy
2D & single particle electron microscopy
electron tomography
immuno-electron microscopy

Puzzle Assembly in Computer Vision and Robotics

H. Wolfson et al., Solving jigsaw puzzles by computer

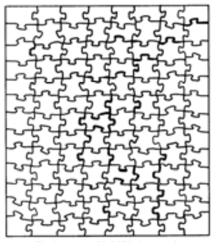


Fig. 1. An assembled 104-piece puzzle.

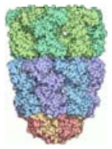
Circa 1986

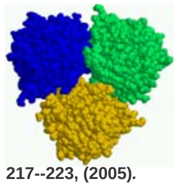
Additional Low Resolution Data Sources

- FRET
- Existence of di-sulfide bonds
- MasSpec (e.g.distance constraints by chemical cross linking).
- SAXS
- Interaction Data (Y2H, gene fusion, similarity with known complexes, etc.)
- and more…

SPECIAL FREQUENT CASE:

Structure Prediction of (cyclically) Symmetric Multi-Molecular Assemblies



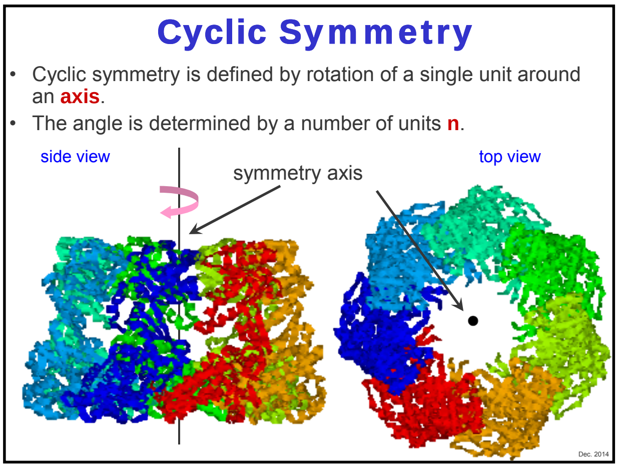


D. Schneidman-Duhovny et al., Proteins, 60, 217--223, (2005).

D. Schneidman-Duhovny et al., NAR 33 (web server issue), W363—W367, (2005).

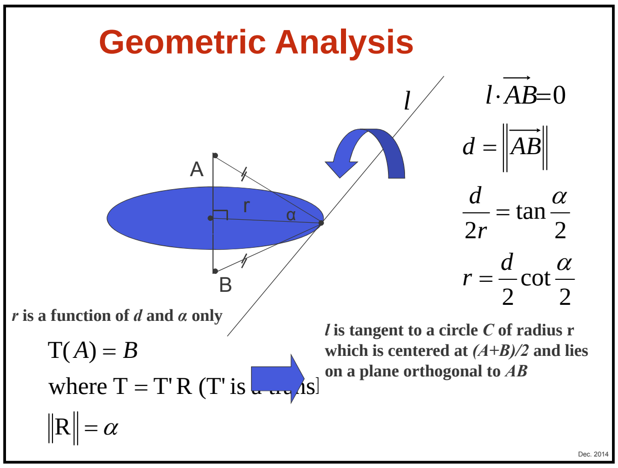
Exploiting the Symmetry Constraints

- A trivial “naïve” approach – perform “regular” multimolecular docking and discard non-symmetric solutions.
- A more sophisticated approach – use the symmetry constraints as an integral part of the algorithm to reduce complexity and improve accuracy.
- Observation – if point A in the protein is matched after the symmetry rotation to point B, one can detect a plane to which the symmetry axis is perpendicular and its location is restricted to a known circle in that plane.



Cyclic Symmetry

- Cyclic symmetry is defined by rotation of a single unit around an **axis**.
- The angle is determined by a number of units **n**.



Geometric Analysis

$$l \cdot \overrightarrow{AB} = 0$$

$$d = \left\| \overrightarrow{AB} \right\|$$

$$\frac{d}{2r} = \tan\frac{\alpha}{2}$$

$$r = \frac{d}{2}\cot\frac{\alpha}{2}$$

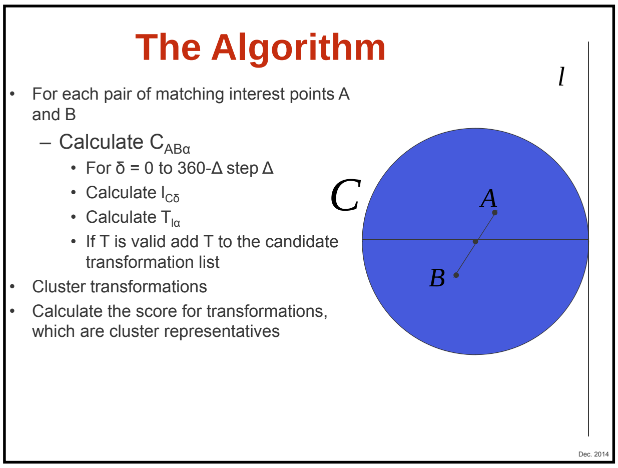
***r* is a function of *d* and *α* only**

$$\mathrm{T}(A) = B$$

where $\mathrm{T} = \mathrm{T'R}$ (T' is a transl

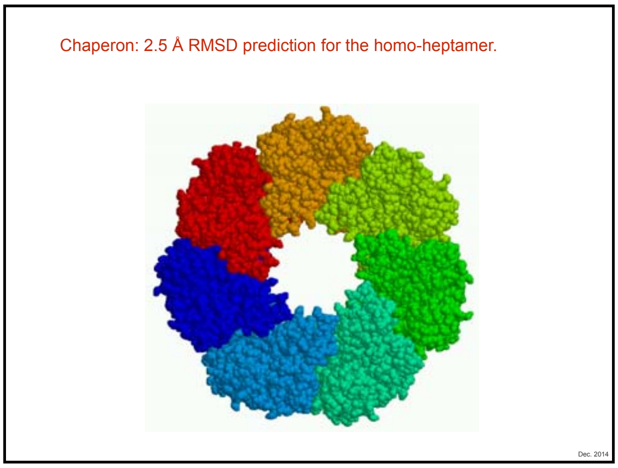
$$\|\mathrm{R}\| = \alpha$$

l* is tangent to a circle *C* of radius r which is centered at *(A+B)/2* and lies on a plane orthogonal to *AB



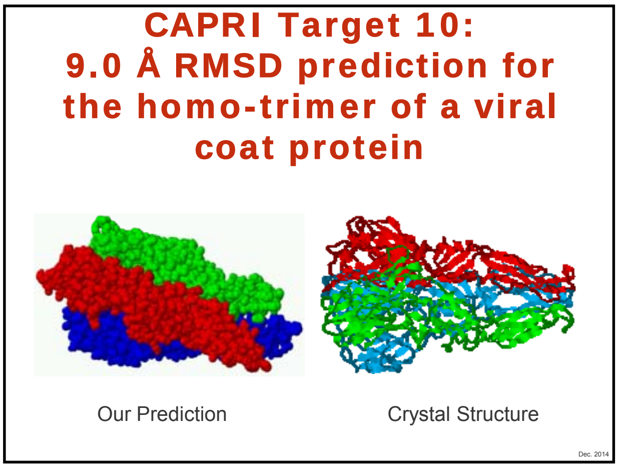
The Algorithm

- For each pair of matching interest points A and B
  - Calculate $C_{AB\alpha}$
    - For δ = 0 to 360-Δ step Δ
    - Calculate $l_{C\delta}$
    - Calculate $T_{l\alpha}$
    - If T is valid add T to the candidate transformation list
- Cluster transformations
- Calculate the score for transformations, which are cluster representatives



Chaperon: 2.5 Å RMSD prediction for the homo-heptamer.

Dec. 2014

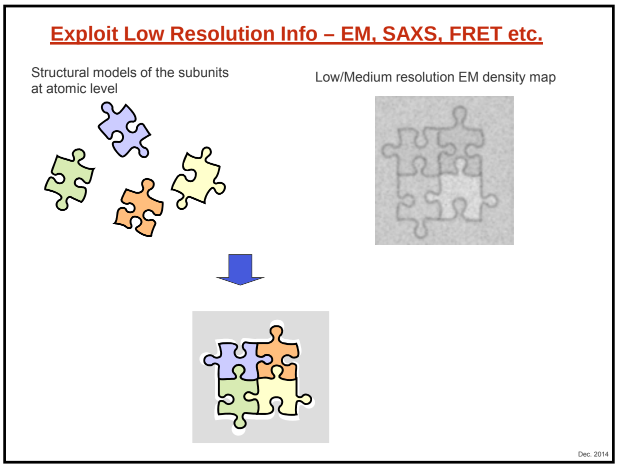


CAPRI Target 10: 9.0 Å RMSD prediction for the homo-trimer of a viral coat protein

Our Prediction

Crystal Structure

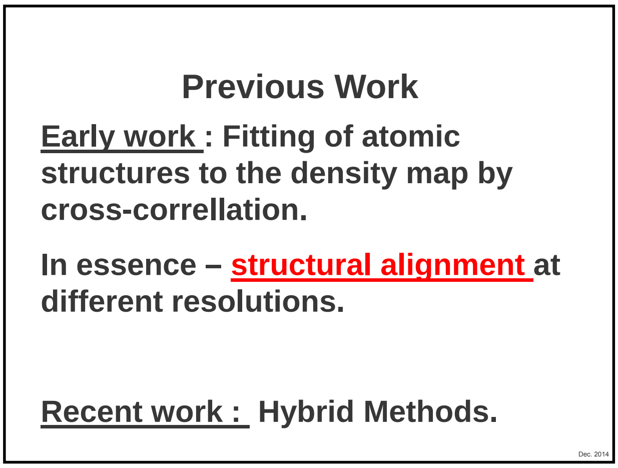
Dec. 2014



Exploit Low Resolution Info – EM, SAXS, FRET etc.

Structural models of the subunits at atomic level
Low/Medium resolution EM density map

Dec. 2014



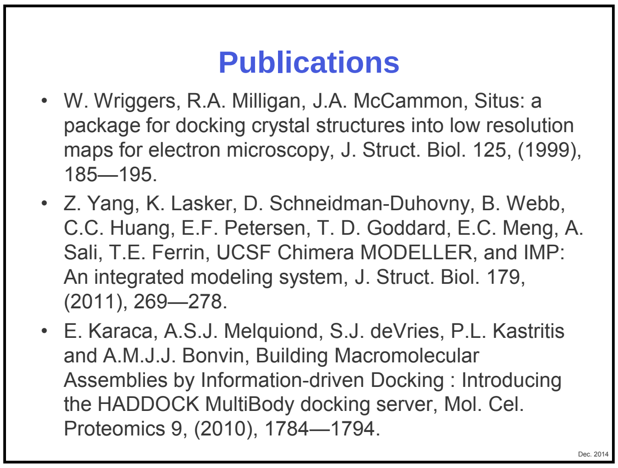
Previous Work

Early work : Fitting of atomic structures to the density map by cross-corellation.

In essence – structural alignment at different resolutions.

Recent work : Hybrid Methods.

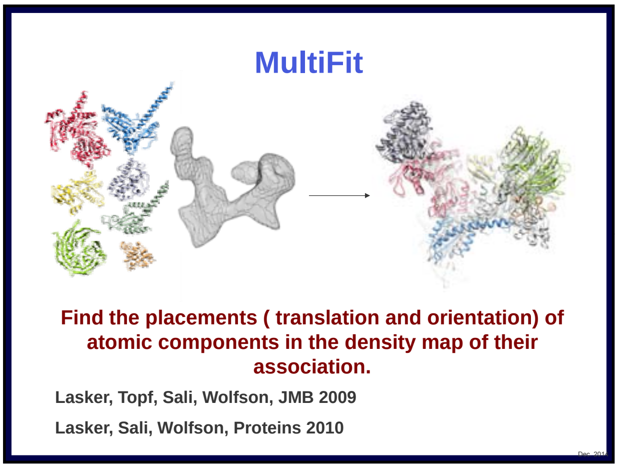
Dec. 2014



Publications

- W. Wriggers, R.A. Milligan, J.A. McCammon, Situs: a package for docking crystal structures into low resolution maps for electron microscopy, J. Struct. Biol. 125, (1999), 185—195.
- Z. Yang, K. Lasker, D. Schneidman-Duhovny, B. Webb, C.C. Huang, E.F. Petersen, T. D. Goddard, E.C. Meng, A. Sali, T.E. Ferrin, UCSF Chimera MODELLER, and IMP: An integrated modeling system, J. Struct. Biol. 179, (2011), 269—278.
- E. Karaca, A.S.J. Melquiond, S.J. deVries, P.L. Kastritis and A.M.J.J. Bonvin, Building Macromolecular Assemblies by Information-driven Docking : Introducing the HADDOCK MultiBody docking server, Mol. Cel. Proteomics 9, (2010), 1784—1794.

Dec. 2014



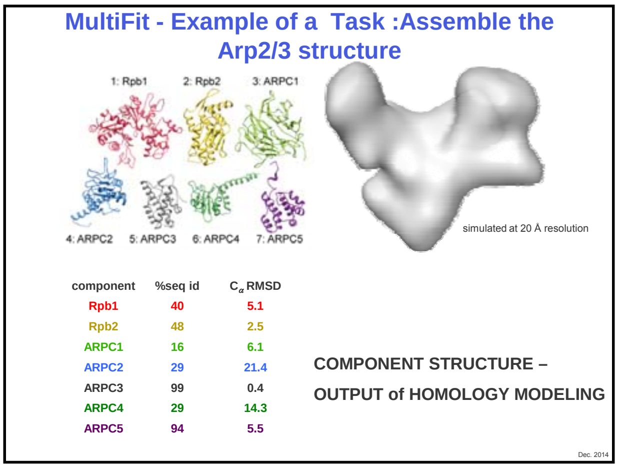
MultiFit

Find the placements (translation and orientation) of atomic components in the density map of their association.

Lasker, Topf, Sali, Wolfson, JMB 2009

Lasker, Sali, Wolfson, Proteins 2010

Dec. 2014



MultiFit - Example of a Task :Assemble the Arp2/3 structure

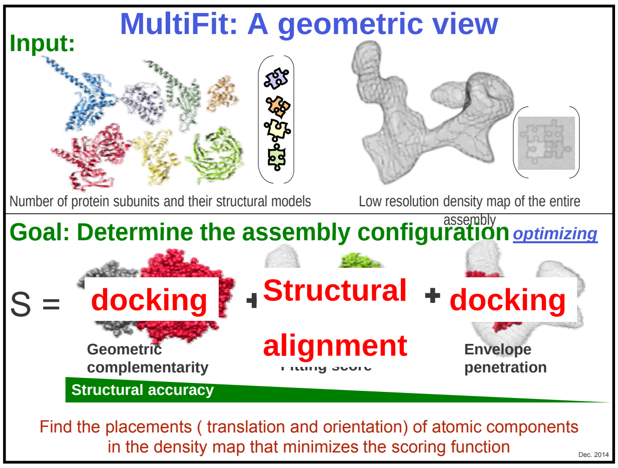
simulated at 20 Å resolution

component	%seq id	C_α RMSD
Rpb1	40	5.1
Rpb2	48	2.5
ARPC1	16	6.1
ARPC2	29	21.4
ARPC3	99	0.4
ARPC4	29	14.3
ARPC5	94	5.5

COMPONENT STRUCTURE –

OUTPUT of HOMOLOGY MODELING

Dec. 2014



MultiFit: A geometric view

Input:

Number of protein subunits and their structural models

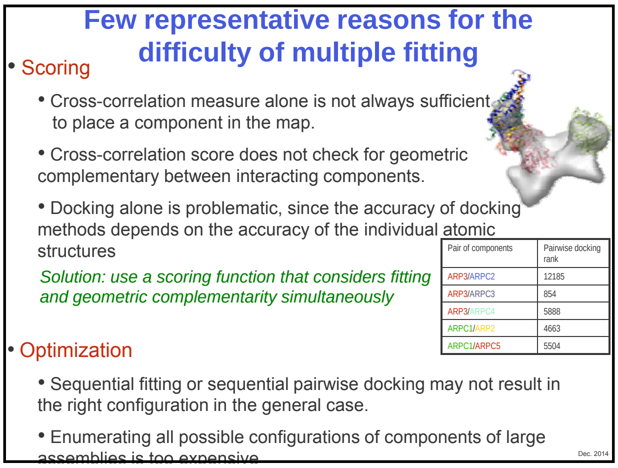
Low resolution density map of the entire assembly

Goal: Determine the assembly configuration *optimizing*

S = docking + Structural alignment + docking

Find the placements (translation and orientation) of atomic components in the density map that minimizes the scoring function

Dec. 2014



Few representative reasons for the difficulty of multiple fitting

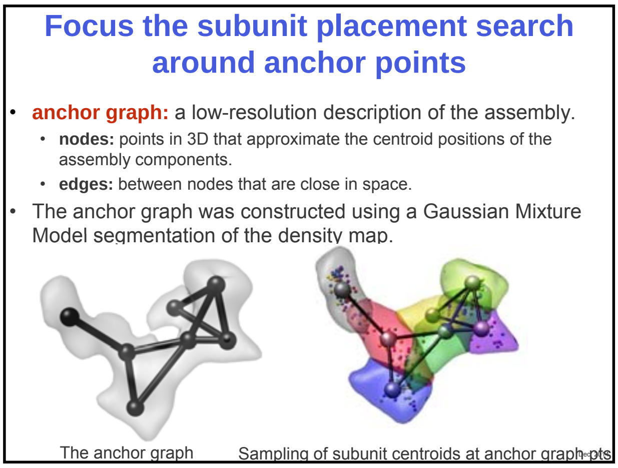
- Scoring
  - Cross-correlation measure alone is not always sufficient to place a component in the map.
  - Cross-correlation score does not check for geometric complementary between interacting components.
  - Docking alone is problematic, since the accuracy of docking methods depends on the accuracy of the individual atomic structures

Pair of components	Pairwise docking rank
ARP3/ARPC2	12185
ARP3/ARPC3	854
ARP3/ARPC4	5888
ARPC1/ARP2	4663
ARPC1/ARPC5	5504

Solution: use a scoring function that considers fitting and geometric complementarity simultaneously

- Optimization
  - Sequential fitting or sequential pairwise docking may not result in the right configuration in the general case.
  - Enumerating all possible configurations of components of large assemblies is too expensive

Dec. 2014



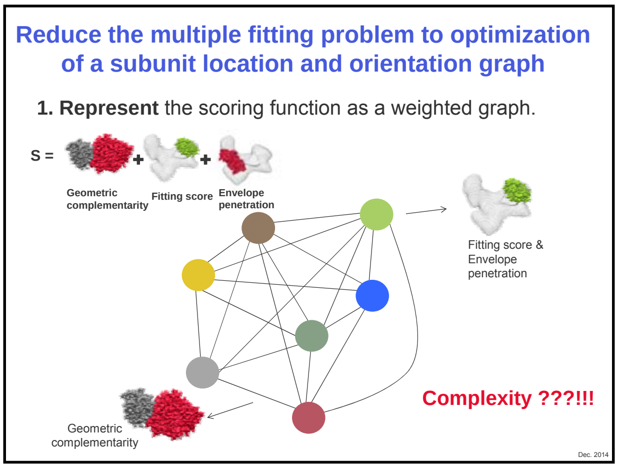
Focus the subunit placement search around anchor points

- **anchor graph:** a low-resolution description of the assembly.
  - **nodes:** points in 3D that approximate the centroid positions of the assembly components.
  - **edges:** between nodes that are close in space.
- The anchor graph was constructed using a Gaussian Mixture Model segmentation of the density map.

The anchor graph

Sampling of subunit centroids at anchor graph pts

Dec. 2014



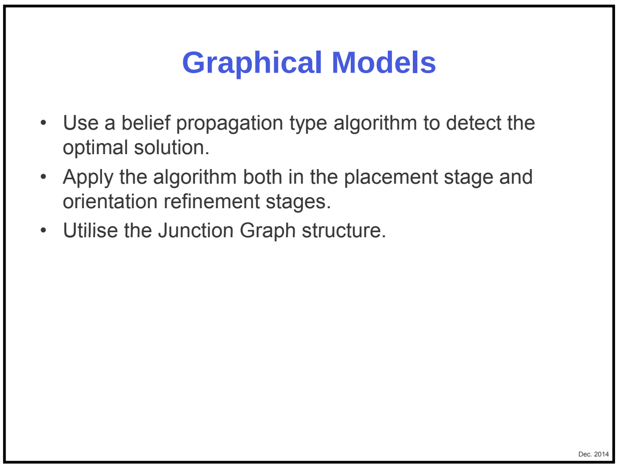
Reduce the multiple fitting problem to optimization of a subunit location and orientation graph

1. **Represent** the scoring function as a weighted graph.

S =
+
+
Geometric complementarity
Fitting score
Envelope penetration
Fitting score & Envelope penetration
Geometric complementarity

Complexity ???!!!

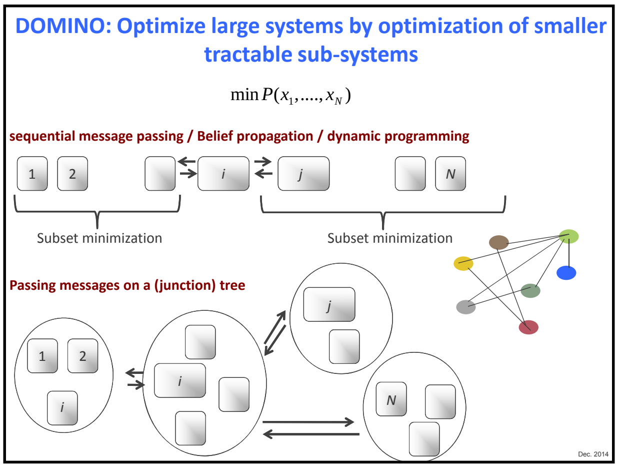
Dec. 2014



Graphical Models

- Use a belief propagation type algorithm to detect the optimal solution.
- Apply the algorithm both in the placement stage and orientation refinement stages.
- Utilise the Junction Graph structure.

Dec. 2014



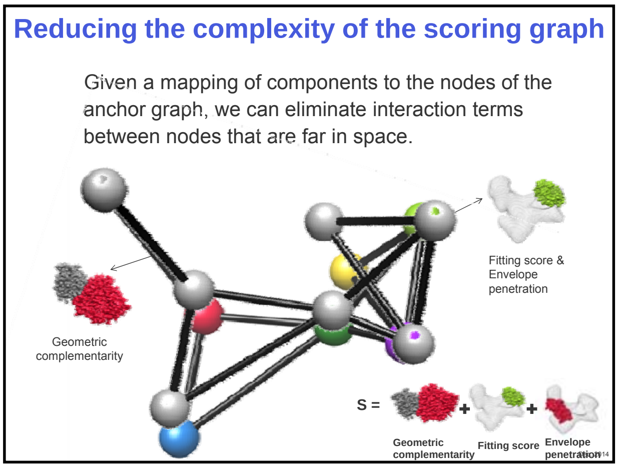
DOMINO: Optimize large systems by optimization of smaller tractable sub-systems

$$\min P(x_1, \ldots, x_N)$$

sequential message passing / Belief propagation / dynamic programming

Passing messages on a (junction) tree

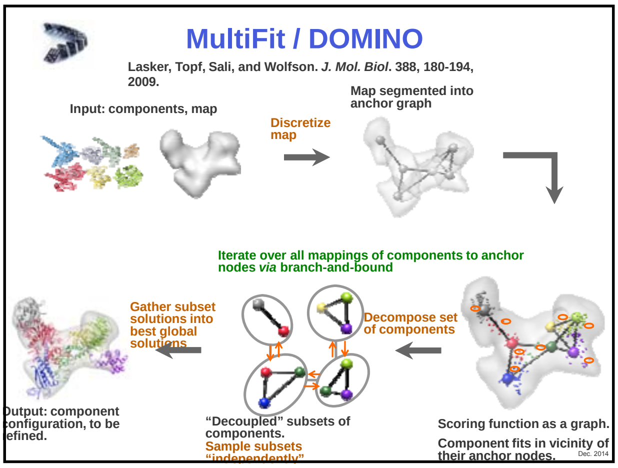
Dec. 2014



Reducing the complexity of the scoring graph

Given a mapping of components to the nodes of the anchor graph, we can eliminate interaction terms between nodes that are far in space.

Dec. 2014



MultiFit / DOMINO

Lasker, Topf, Sali, and Wolfson. *J. Mol. Biol*. 388, 180-194, 2009.

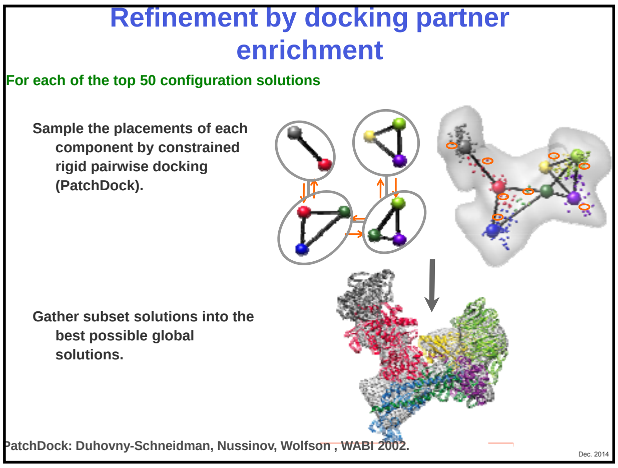
Output: component configuration, to be efined.

"Decoupled" subsets of components. Sample subsets "independently"

Scoring function as a graph.

Component fits in vicinity of their anchor nodes.

Dec. 2014



Refinement by docking partner enrichment

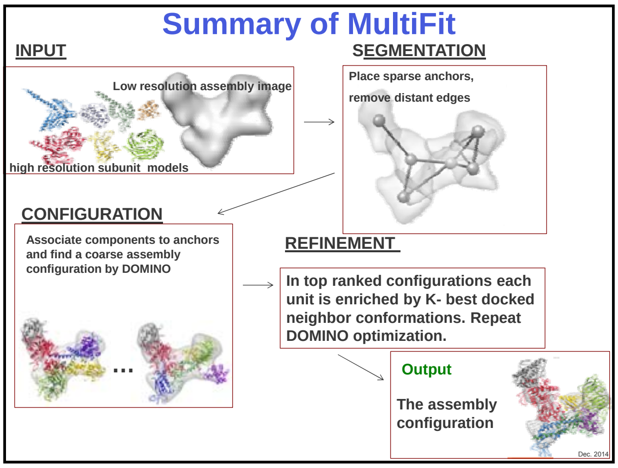
For each of the top 50 configuration solutions

Sample the placements of each component by constrained rigid pairwise docking (PatchDock).

Gather subset solutions into the best possible global solutions.

PatchDock: Duhovny-Schneidman, Nussinov, Wolfson , WABI 2002.

Dec. 2014



Summary of MultiFit

INPUT

Low resolution assembly image

high resolution subunit models

SEGMENTATION

Place sparse anchors,

remove distant edges

CONFIGURATION

Associate components to anchors and find a coarse assembly configuration by DOMINO

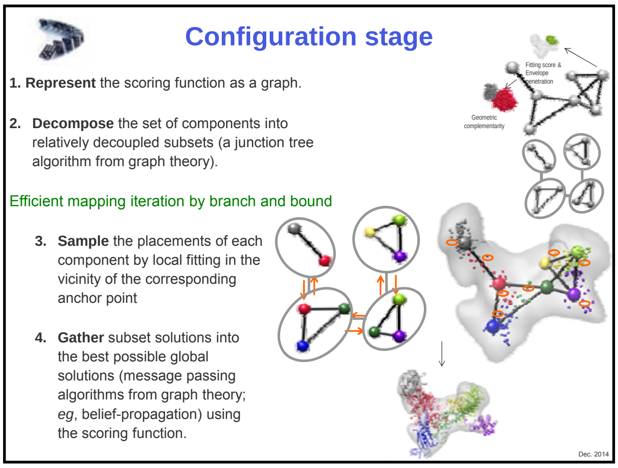
REFINEMENT

In top ranked configurations each unit is enriched by K- best docked neighbor conformations. Repeat DOMINO optimization.

Output

The assembly configuration

Dec. 2014



Configuration stage

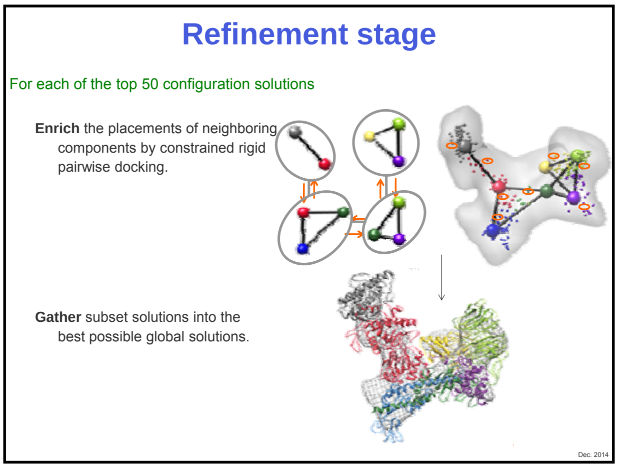
1. **Represent** the scoring function as a graph.
2. **Decompose** the set of components into relatively decoupled subsets (a junction tree algorithm from graph theory).

Efficient mapping iteration by branch and bound

3. **Sample** the placements of each component by local fitting in the vicinity of the corresponding anchor point
4. **Gather** subset solutions into the best possible global solutions (message passing algorithms from graph theory; *eg*, belief-propagation) using the scoring function.

Fitting score & Envelope penetration
Geometric complementarity

Dec. 2014



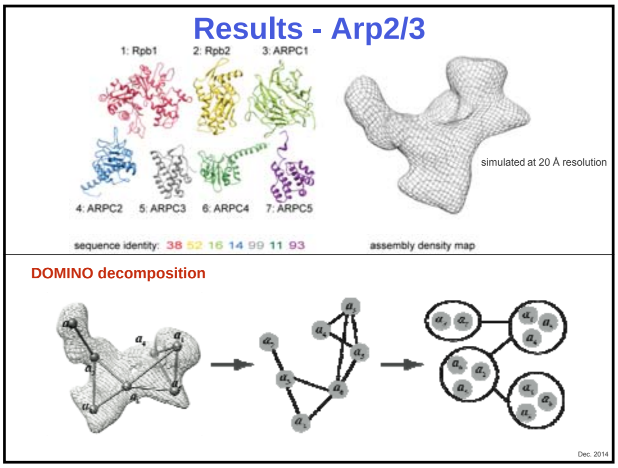
Refinement stage

For each of the top 50 configuration solutions

Enrich the placements of neighboring components by constrained rigid pairwise docking.

Gather subset solutions into the best possible global solutions.

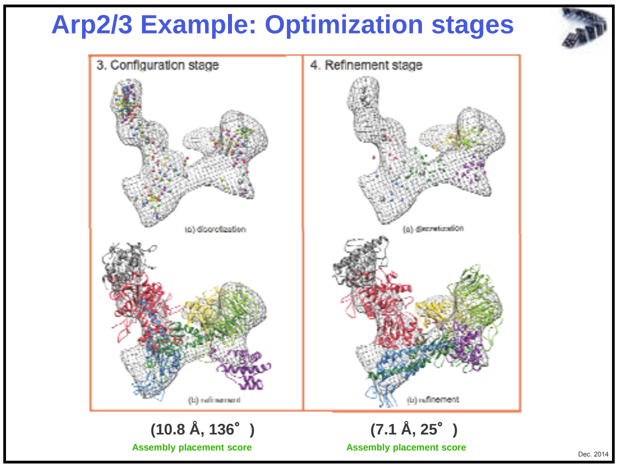
Dec. 2014



Results - Arp2/3

DOMINO decomposition

Dec. 2014

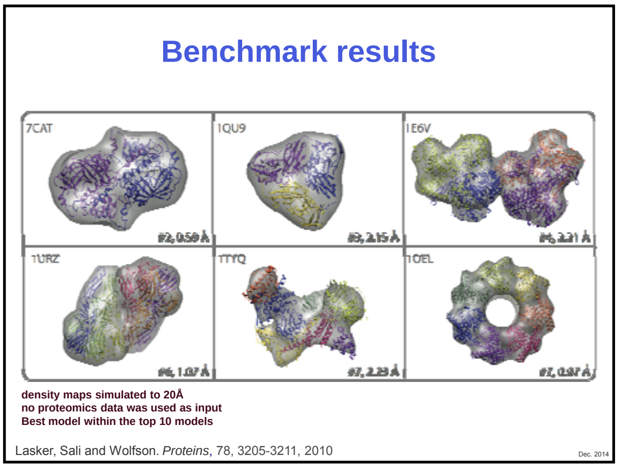


Arp2/3 Example: Optimization stages

(10.8 Å, 136°) Assembly placement score

(7.1 Å, 25°) Assembly placement score

Dec. 2014

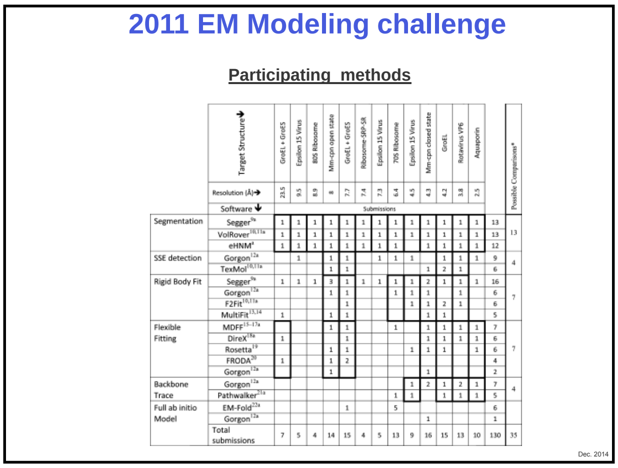


Benchmark results

density maps simulated to 20Å
no proteomics data was used as input
Best model within the top 10 models

Lasker, Sali and Wolfson. *Proteins*, 78, 3205-3211, 2010

Dec. 2014

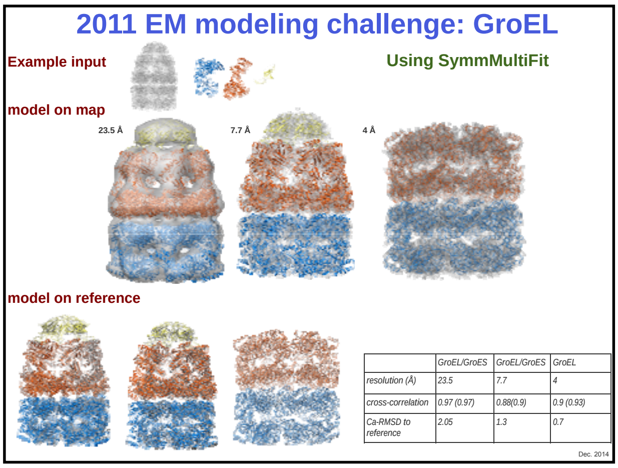


2011 EM Modeling challenge

Participating methods

	Target Structure →	GroEL + GroES	Epsilon 15 Virus	80S Ribosome	Mm-cpn open state	GroEL + GroES	Ribosome-SRP-SR	Epsilon 15 Virus	70S Ribosome	Epsilon 15 Virus	Mm-cpn closed state	GroEL	Rotavirus VP6	Aquaporin		Possible Comparisons*
	Resolution (Å) →	23.5	9.5	8.9	8	7.7	7.4	7.3	6.4	4.5	4.3	4.2	3.8	2.5		
	Software ↓	Submissions														
Segmentation	Segger[9a]	1	1	1	1	1	1	1	1	1	1	1	1	1	13	13
	VolRover[10,11a]	1	1	1	1	1	1	1	1	1	1	1	1	1	13	
	eHNM[a]	1	1	1	1	1	1	1	1		1	1	1	1	12	
SSE detection	Gorgon[12a]		1		1	1		1	1	1		1	1	1	9	4
	TexMol[10,11a]				1	1					1	2	1		6	
Rigid Body Fit	Segger[9a]	1	1	1	3	1	1	1	1	1	2	1	1	1	16	7
	Gorgon[12a]				1	1			1	1	1		1		6	
	F2Fit[10,11a]					1				1	1	2	1		6	
	MultiFit[13,14]	1			1	1					1	1			5	
Flexible Fitting	MDFF[15–17a]				1	1			1		1	1	1	1	7	7
	DireX[18a]	1				1					1	1	1	1	6	
	Rosetta[19]				1	1				1	1	1		1	6	
	FRODA[20]	1			1	2									4	
	Gorgon[12a]				1						1				2	
Backbone Trace	Gorgon[12a]									1	2	1	2	1	7	4
	Pathwalker[21a]								1	1		1	1	1	5	
Full ab initio Model	EM-Fold[22a]					1			5						6	
	Gorgon[12a]										1				1	
	Total submissions	7	5	4	14	15	4	5	13	9	16	15	13	10	130	35

Dec. 2014



2011 EM modeling challenge: GroEL

Example input

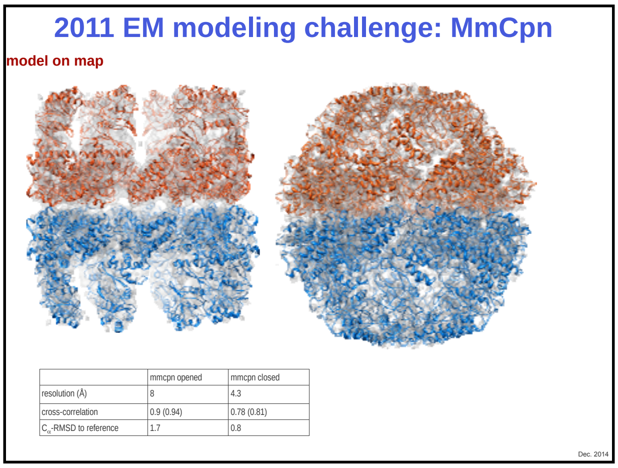
Using SymmMultiFit

model on map

model on reference

	GroEL/GroES	GroEL/GroES	GroEL
resolution (Å)	23.5	7.7	4
cross-correlation	0.97 (0.97)	0.88(0.9)	0.9 (0.93)
Ca-RMSD to reference	2.05	1.3	0.7

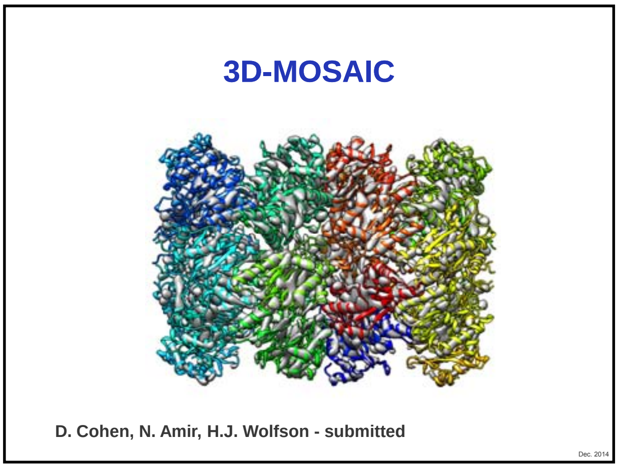
Dec. 2014



2011 EM modeling challenge: MmCpn

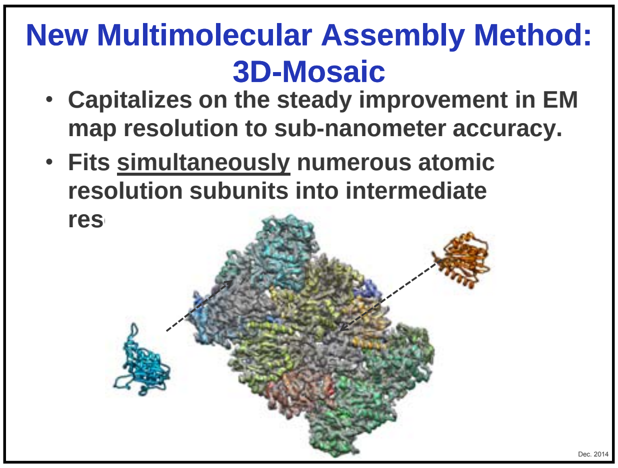
model on map

	mmcpn opened	mmcpn closed
resolution (Å)	8	4.3
cross-correlation	0.9 (0.94)	0.78 (0.81)
C_α-RMSD to reference	1.7	0.8



3D-MOSAIC

D. Cohen, N. Amir, H.J. Wolfson - submitted



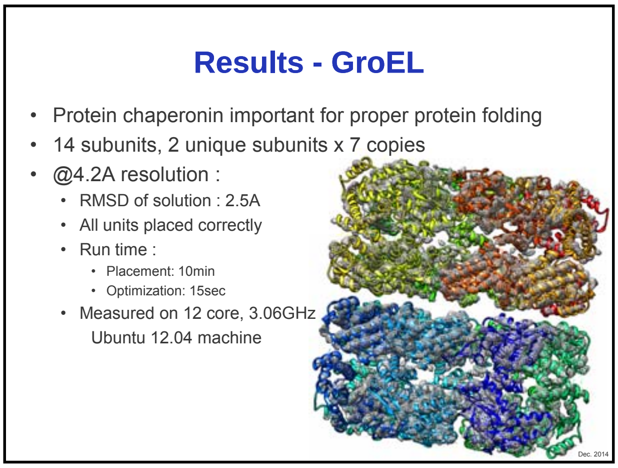
New Multimolecular Assembly Method: 3D-Mosaic

- Capitalizes on the steady improvement in EM map resolution to sub-nanometer accuracy.
- Fits simultaneously numerous atomic resolution subunits into intermediate res

Advantages of 3D-Mosaic

- Requires *no prior segmentation* of the EM map.
- Handles *“missing”* subunits.
- *Highly efficient* handling of a large number of *multiple structurally homologous copies* of complex subunits.
- Efficient new method for integrative *simultaneous* modeling of large multi-molecular assemblies by formulating the optimization task as an Integer Linear Program (ILP).
- Incorporates both EM and X-link information into the same framework.

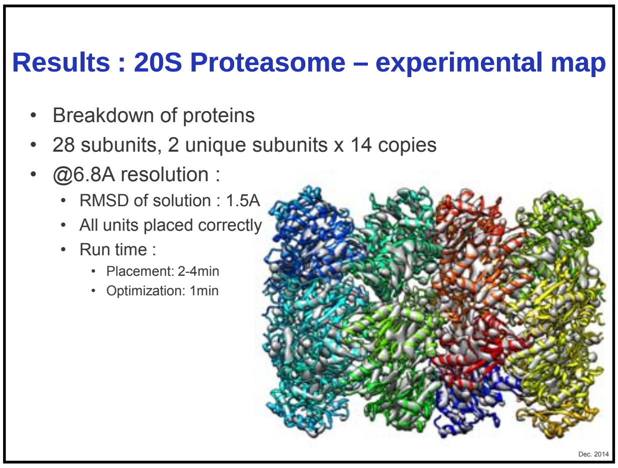
D. Cohen, N. Amir, H.J. Wolfson, 3D-MOSAIC: An efficient method for integrative modeling of large multimolecular assemblies, (to be submitted).



Results - GroEL

- Protein chaperonin important for proper protein folding
- 14 subunits, 2 unique subunits x 7 copies
- @4.2A resolution :
  - RMSD of solution : 2.5A
  - All units placed correctly
  - Run time :
    - Placement: 10min
    - Optimization: 15sec
  - Measured on 12 core, 3.06GHz Ubuntu 12.04 machine

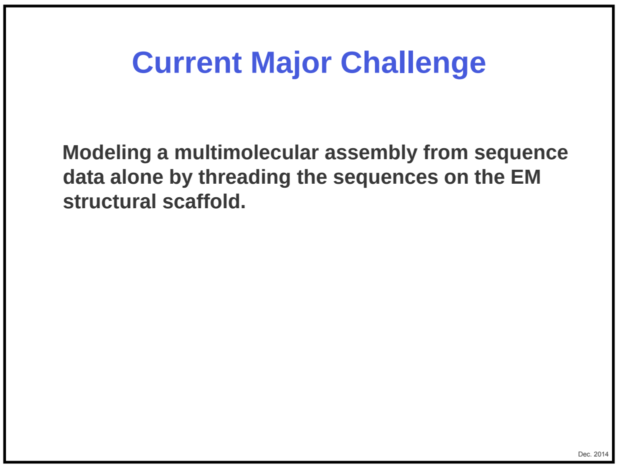
Dec. 2014



Results : 20S Proteasome – experimental map

- Breakdown of proteins
- 28 subunits, 2 unique subunits x 14 copies
- @6.8A resolution :
  - RMSD of solution : 1.5A
  - All units placed correctly
  - Run time :
    - Placement: 2-4min
    - Optimization: 1min

Dec. 2014



Current Major Challenge

Modeling a multimolecular assembly from sequence data alone by threading the sequences on the EM structural scaffold.

Dec. 2014