V3 DockStar: overcome limitations of CombDock

DockStar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes

Naama Amir*, Dan Cohen and Haim J. Wolfson*
Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

2 subtasks for generation of macromolecular complex structures:

(a) Identify the protein-protein interaction graph between the individual subunits.
   This can be done e.g. based on data from MS and chemical cross-linking.

(b) Detect a globally consistent pose of the subunits, so that
   - there are no steric clashes between them and
   - the binding energy of the whole complex is optimized.
Chemical cross-linking

(a) Cross-linking reaction using a chemical cross-linking reagent. These molecules have a certain length, have two reactive groups at both ends of the molecule and may covalently bind either to cysteine or lysine residues of a single protein or of two proteins.

(b) enzymatic digestion of the proteins to peptides,

(c) enrichment of cross-linked peptides,

(d) analysis of cross-linked peptides by LC-MS/MS,

(e) data analysis.

StarDock

- MS of intact protein complexes and their subcomplexes (→TAP-MS) can determine the stoichiometry of the complex subunits and deduce the interaction graph of the multimolecular complex.

- Chemical cross-linking combined with MS provides distance constraints between surface residues both on the same and on neighboring subunits.

This provides information both for the detection of the interaction graph as well as constraints on the relative spatial poses of neighboring subunits.

Amir et al., Bioinformatics 31, 2801 (2015)
Example: refining the 3D structure of S26 proteasome

Low resolution EM structure

Chemical cross-links for the *S. pombe* and *S. cerevisiae* 26S proteasomes.

55 (21) pairs of cross-linked lysines from the *S. pombe* (*S. cerevisiae*) 26S proteasome subunits.

Multiple edges between a pair of subunits indicate multiple cross-linked lysine pairs.

Atomistic structure generated

StarDock: Generate transformation sets

Assume that the interaction graph is known (task a).

Now we will generate for each subunit a set of candidate rigid transformations.

Select as anchor subunit the subunit having most neighbors in the multimolecular assembly interaction graph.
All other subunits which are known to interact with the anchor are then docked to it.
This requires a star shaped spanning tree topology of the interaction graph.

Pairwise docking is carried out by PatchDock, which optimizes shape complementarity, while satisfying maximal distance constraints between residues of neighboring subunits from cross-linking (details not important here).

The top 1000 PatchDock transformations are refined, rescored and re-ranked by the FiberDock tool → pairwise scores

Amir et al., Bioinformatics 31, 2801 (2015)
Let
- \( P_i (0 \leq i < n) \) be subunit \( i \),
- \( T(P_i) \) be the set of candidate transformations for subunit \( P_i \) received from the previous stage.
- \( T_{i,r} \) be a particular transformation \( r \) of subunit \( P_i \).
- \( S(T_{i,r}, T_{j,s}) \) be the pairwise interaction score of subunits \( P_i \) and \( P_j \) transformed by \( T_{i,r} \) and \( T_{j,s} \), respectively (obtained by pairwise docking before).

The **globally optimal solution** \( \text{Sol} \) includes one transformation per subunit and maximizes the score(\( \text{Sol} \)) defined as:

\[
\text{score(\( \text{Sol} \))} = \sum_{T_{i,r}, T_{j,s} \in \text{Sol} \cap i \neq j} S(T_{i,r}, T_{j,s})
\]

Amir et al., Bioinformatics 31, 2801 (2015)
DockStar: Select best global solution

This optimization task can be formulated as the following graph theoretic problem:

Let $G = (V,E)$ be an undirected $n$-partite graph with a partition of the vertex set

$V = V_0 \cup \ldots \cup V_{n-1}$,

so that each transformation $T_{i,r} \in T(P_i)$ corresponds to a vertex $u_{i,r} \in V_i$.

(Each $V_i$ contains all transformations $r$ of subunit $P_i$ as its vertices $u_{i,r}$).

Each pair of vertices is joined by an edge:

$$E = \{(u_{i,r}, v_{j,s})|u_{i,r} \in V_i; v_{j,s} \in V_j; i \neq j\}$$

with the weight

$$w(u_{i,r}, v_{j,s}) = S(T_{i,r}, T_{j,s}) \quad \forall (u_{i,r}, v_{j,s}) \in E$$

The optimal solution is achieved by choosing one vertex per $V_i$ that maximizes the edge-weight of the induced sub-graph.

Amir et al., Bioinformatics 31, 2801 (2015)
Formulate Integer Linear Program (ILP)

This graph theoretic task can be formulated as an ILP. Define a variable $X_{i,r}$ for each vertex $u_{i,r} \in V$ and a variable $Y_{i,r,j,s}$ for each edge $e(u_{i,r}, v_{j,s}) \in E$ as follows:

$$X_{i,r} = \begin{cases} 1 & \text{if } u_{i,r} \text{ is chosen} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i,r,j,s} = \begin{cases} 1 & \text{if both } u_{i,r} \text{ and } v_{j,s} \text{ are chosen} \\ 0 & \text{otherwise} \end{cases}$$

The ILP **objective function** is:

Maximize \[ \text{score(Sol)} = \sum_{(u_{i,r}, v_{j,s}) \in E} w(u_{i,r}, v_{j,s}) Y_{i,r,j,s} \]

Subject to the constraints:

$$\sum_{u_{i,r} \in V_i} X_{i,r} = 1 \quad \forall i, 0 \leq i < n$$

$$\sum_{u_{i,r} \in V_i} Y_{i,r,j,s} = X_{j,s} \quad \forall j, s, i, \quad j \neq i$$

The objective function is exactly the edge-weight of the chosen sub-graph. The first constraint ensures that exactly one transformation is chosen for each subunit. The second constraint ensures that an edge is chosen if and only if both vertices that it connects are chosen as well. The ILP step was solved by the CPLEX 12.5 package.

Amir et al., Bioinformatics 31, 2801 (2015)
The ILP method outputs one single highest scoring global solution.

To retrieve additional high scoring solutions, the ILP step is applied iteratively to find a solution that maximizes the objective function and was not chosen before.

For this, a **linear constraint** is used (see paper by Amir et al.).
ILP formulation – arbitrary complexes

Sofar we considered complexes having a **star shaped spanning tree**, where an **anchor** subunit, which interacts with all the other subunits, can be chosen. However, this is a special case.

**Arbitrary complexes** are divided into overlapping sub-complexes, each with a star shaped spanning tree, which are solved separately as above.

(A) A complex interaction graph that is **not star shaped**. Therefore, the complex is divided to 2 sub-complexes B and C and each sub-complex structure is solved separately. The transformation set for each subunit is generated by docking the subunit to the "anchor" subunit.

In (B) the anchor is represented by the red vertex and in (C) by the green. For each sub-complex a set of solutions is generated. Then, top solutions of these sub-complexes are integrated to create the 3D structure of the whole complex.

Amir et al., Bioinformatics 31, 2801 (2015)
### Table 1. Summary of the DockStar's results

<table>
<thead>
<tr>
<th>Target complex</th>
<th>Bound/unbound</th>
<th>Subunits number</th>
<th>Rank</th>
<th>Global Ca-RMSD(^a)</th>
<th>Number of contacts(^b)</th>
<th>Quality of predicted contacts(^c)</th>
<th>Run time HH:MM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP2A</td>
<td>Bound</td>
<td>3</td>
<td>1</td>
<td>0.68</td>
<td>2</td>
<td>high 0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Unbound</td>
<td>3</td>
<td>1</td>
<td>6.9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Beef liver</td>
<td>Bound</td>
<td>4</td>
<td>1</td>
<td>0.85</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Catalase</td>
<td>Unbound</td>
<td>4</td>
<td>1</td>
<td>2.7</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNA polII</td>
<td>Bound</td>
<td>11</td>
<td>1</td>
<td>7.9</td>
<td>10</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Unbound</td>
<td>11</td>
<td>3</td>
<td>4.8</td>
<td>10</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Yeast exosome</td>
<td>Bound</td>
<td>10</td>
<td>1</td>
<td>5.1</td>
<td>9</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Unbound</td>
<td>10</td>
<td>12</td>
<td>6.0</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\)Global Ca-RMSD between the predicted and the native assemblies including only predictions with lenient to high quality.

\(^b\)Number of contacts in the spanning tree of the complex interaction graph.

\(^c\)Predicted interfaces in the target complex that are of lenient to high quality.

---

**Fig. 1.** The predicted models of the bound cases (coloured by chains) superimposed on the correct complex structures taken from the PDB (grey). (A) PP2A (A(yellow), B(blue), C(red)), (B) The Beef Liver Catalase (A(yellow), B(blue), C(red), D(green)), (C) RNA polymerase II (Rbp1(blue), Rbp2(cyan), Rbp3(light blue), Rbp5(purple), Rbp6(green), Rbp7(pink), Rbp8(yellow), Rbp9(dark green), Rbp10(orange), Rbp11(brown), Rbp12(red)), (D) The Yeast Exosome (Rrp45(blue), Rrp41(cyan), Rrp43(light blue), Rrp46(green), Rrp42(purple), Mrp3(pink), Rrp40(red), Rrp4(orange), Cst4(yellow), Dis3(dark green)). (E) The predicted order of chains in the model of the TRiC/CCT Chaperonin: Z(red) Q(blue) H(yellow) E(light blue) B(pink) D(grey) A(green) G(purple)
Mosaic-3D

Input:
(1) high-resolution 3D **structures** of a representative of each protein involved in forming the complex
(2) information on the **stoichiometry** of the complex.
(3) information on pairwise **interfaces** that provide the presumed binding modes in the complex.

Output:
3D-MOSAIC assembles the complex in an iterative tree-based greedy fashion.

Similar to CombDock, each node represents a monomer attached in a particular orientation.

Mosaic-3D

The algorithm starts from a seed monomer with the largest number of interfaces.

In each iteration, new child solutions are generated by adding an additional monomer to each of the parent solutions retained from the previous iteration.

A new monomer of a particular protein type $p$ can be attached to the complex $r$ of a previous stage, if
i) the number of occurrences of $p$ in the parent solution has not yet reached its maximum multiplicity,
ii) $r$ has unoccupied interfaces for an interaction with $p$.
iii) The new monomer does not lead to severe steric clashes with other monomers already present in the parent solution.

The new child monomer is scored according to the number of interfaces it has with all ancestor monomers already present in the complex.

After each iteration: cluster solutions based on $C_{\alpha}$-RMSD
Finally: optimize symmetry
Workflow 3D-Mosaic

Assembly of homo-hexameric hemocyanin from *Panulirus interruptus* (1HCY.pdb).

In each iteration, new monomers can be attached to all previously retained solutions.

If a matching interface is found, the complex match score increases and the corresponding complex might be ranked further up in the list of solutions (green double-tilted arrows).

Solutions similar to better-ranked ones or yielding severe steric clashes are discarded.

Mosaic-3D

Examples of complexes and corresponding topology graphs for hard cases:
(a) ring-like topology of T4 lysozyme hexamer (3SBA),
(b) cage-like topology of pyruvate dehydrogenase E2 60-mer core complex (1B5S),
(c) inovirus coat protein filament (2C0W) composed of helical monomers,
(d) human cystatin C complex (1R4C) forming interchain β-sheets.
Different node colors correspond to different protein types, different edge colors to different binding modes.

On a diverse benchmark set of 308 homo and heteromeric complexes containing 6 to 60 monomers, the mean fraction of correctly reconstructed benchmark complexes during crossvalidation was 78.1%.

Summary

Our current atomistic understanding of how large macromolecular machines work is mainly based on results from protein crystallography. These discoveries were rewarded with several Nobel Prizes in Chemistry and Medicine.

Recent breakthrough: new detectors for EM that improve its resolution down to atomic resolution.

Ideal for structural characterization of large multi-protein complexes: combination of methods in structural biology:
- X-ray crystallography and NMR for high-resolution structures of single proteins and pieces of protein complexes
- (cryo) EM to determine high- to medium-resolution structures of entire protein complexes
- stained EM for still pictures at medium-resolution of cellular organells and
- (cryo) electron tomography for three-dimensional reconstructions of biological cells and for identification of the individual components.

2.4 Fitting atomistic structures into EM maps

Atomistic structure of a part of the complex

Coarse EM structure of the whole complex

- same resolution for both structures
- exhaustive search with scoring
- choose best pose(s)
The procedure

**blurring:** adapt resolution

**searching:** test all displacements and orientations

**scoring:** find combination with maximum overlap
Step 1: blurring the picture

Mathematically:
convolution of (exact) atomistic structure \( f(x) \) with experimental resolution \( g(x) \)

\[
\tilde{f}(x) = (f \otimes g)(x) = \int dz \ f(z) \ g(x - z)
\]

- original data seen through the imaging apparatus
- original signal
- "kernel" => what is the image of a single point? (=delta signal)

Often: blurring with gaussian
Put it on a grid

Discretize:
2.5 Fourier Transformation

Forward

\[ F(k) = \int_{-\infty}^{\infty} dx \ e^{-ikx} \ f(x) \]

and inverse Fourier transform

\[ f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \ e^{ikx} \ F(k) \]

with \[ e^{ikx} = \cos(kx) + i \sin(kx) \]

\[ \Rightarrow \text{convert between real and frequency (Fourier) space} \]

short distances \( \Leftrightarrow \) high frequencies

long distances \( \Leftrightarrow \) low frequencies
Shift of the Argument

\[ FT[f(x + \Delta x)] = \int_{-\infty}^{\infty} dx \ e^{-ikx} \ f(x + \Delta x) \]
\[ = \int dy \ e^{-ik(y-\Delta x)} \ f(y) \]
\[ = e^{ik \Delta x} \int dy \ e^{-iky} \ f(y) \]
\[ = e^{ik \Delta x} \ FT[f(x)] \]

Variable transformation:
y = x + \Delta x

cChange name of integration variable
back from y to x
Convolution

\[ \tilde{f}(x) = (f \otimes g)(x) = \int dz \, f(z) \, g(x-z) \]

Apply FT on both sides:

\[ FT[\tilde{f}(x)] = FT[(f \otimes g)(x)] = \]

\[ = \]

\[ = \]

\[ = FT[f(x)] \cdot FT[g(x)]\]

Integration in real space is replaced by simple multiplication in Fourier space.

But FTs need to be computed.

What is more efficient?

If the same width \( g(x) \) is used for multiple displaced datasets

\( \Rightarrow \) do \( FT[g(x)] \) only once
Fourier on a Grid

On a finite grid:

\[ F_k = \sum_{j=0}^{N-1} e^{-2\pi i j k/N} f_j \]

\[ f_k = \frac{1}{N} \sum_{j=0}^{N-1} e^{+2\pi i j k/N} F_j \]

\[ \Rightarrow \text{maximum wavelength} = \text{length of grid} \]

\[ \Rightarrow \text{minimum wavelength} = \text{grid spacing} \]

\[ \Rightarrow \text{sum instead of integral} \]
2.5.5 FFT by Danielson and Lanczos (1942)

Danielson and Lanczos showed that a discrete Fourier transform of length \( N \) can be rewritten as the sum of two discrete Fourier transforms, each of length \( N/2 \).

One of the two is formed from the even-numbered points of the original \( N \), the other from the odd-numbered points.

\[
F_k = \sum_{j=0}^{N-1} e^{-2\pi i k \frac{j}{N}} f_j
\]

\[
= \sum_{j=0}^{\frac{N}{2}-1} e^{-2\pi i k \frac{2j}{N}} f_{2j} + \sum_{j=0}^{\frac{N}{2}-1} e^{-2\pi i k \frac{2j+1}{N}} f_{2j+1}
\]

\[
= \sum_{j=0}^{\frac{N}{2}-1} e^{-2\pi i k \frac{j}{2}} f_{2j} + e^{-2\pi i k \frac{1}{N}} \sum_{j=0}^{\frac{N}{2}-1} e^{-2\pi i k \frac{j}{2}} f_{2j+1}
\]

\[
= F_k^e + e^{-2\pi i k \frac{1}{N}} F_k^o
\]

\( F_k^e \) : \( k \)-th component of the Fourier transform of length \( N/2 \) formed from the even components of the original \( f_j \)'s

\( F_k^o \) : \( k \)-th component of the Fourier transform of length \( N/2 \) formed from the odd components of the original \( f_j \)'s
FFT by Danielson and Lanczos (1942)

The wonderful property of the Danielson-Lanczos-Lemma is that it can be used recursively.

Having reduced the problem of computing $F_k$ to that of computing $F_k^e$ and $F_k^o$, we can do the same reduction of $F_k^e$ to the problem of computing the transform of its $N/4$ even-numbered input data and $N/4$ odd-numbered data.

We can continue applying the DL-Lemma until we have subdivided the data all the way down to transforms of length 1.

What is the Fourier transform of length one? It is just the identity operation that copies its one input number into its one output slot.

For every pattern of $\log_2 N$ e‘s and o‘s, there is a one-point transform that is just one of the input numbers $f_n$

$$F_k^{eoeooee...oeo} = f_n \quad \text{for some } n$$
FFT by Danielson and Lanczos (1942)

The next trick is to figure out which value of \( n \) corresponds to which pattern of e‘s and o‘s in

\[
F_k^{eoeoeo...oe} = f_n
\]

Answer: reverse the pattern of e‘s and o‘s, then let e = 0 and o = 1, and you will have, in binary the value of \( n \).

This works because the successive subdividisions of the data into even and odd are tests of successive low-order (least significant) bits of \( n \).

Thus, computing a FFT can be done efficiently in \( O(N \log(N)) \) time.
Discretization and Convolution

For practical applications:
=> first put atomistic data onto the grid, then blur with FFT

discretized hi-res data  blurring kernel  low-res image
Step 3: Scoring the Overlap

Most simple case:
- apply density threshold and count overlapping voxels
- displace images relative to each other, recount
=> find displacement with maximum overlap

In matrix form with displacements $x, y$:

$$c(x, y) = \sum_{l=1}^{N} \sum_{m=1}^{N} a_{l,m} b_{l+x,m+y}$$
Cross Correlation

Generalization: maximize cross correlation of grided densities with respect to displacement (and orientation)

\[
C_{x,y,z} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} \times b_{l+x,m+y,n+z}
\]

Note: maximize the cross-correlation \(\iff\) minimize the squared difference

On a grid with \(N^3\) gridpoints \(\implies\) \(N^3\) possible displacements
\(\implies\) runtime \(O(N^6)\)

Further complication: the convolution

\[
C'_{x,y,z} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} \times (g \otimes b_{l+x,m+y,n+z})
\]
Correlation and Fourier

Apply Fourier transformation to both sides of

\[ C_{x,y,z} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} \times b_{l+x,m+y,n+z} \]

=> matrix multiplication

\[ FT[C] = FT[A]^* \times FT[B] \]

Runtime of 3D FFT = \( O(N^3 \log^3(N)) \) \(<\ N^6 \)
=> all possible displacements tested simultaneously

Note: FT[A] only calculated once initially
   => two FFTs per orientation
   => scan orientation via Euler angles

<= Step 2: exhaustive search
Include convolution

Maximize

\[ C_{x,y,z} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} \times (g \otimes b_{l+x,m+y,n+z}) \]

In Fourier space:

Insert convolution

\[ FT[f \otimes g] = FT[f] \times FT[g] \]

Into correlation:

\[ FT[C] = FT[A]^* \times FT[G \otimes B] \]

\[ = FT[A]^* \times (FT[G] \times FT[B]) \]

\[ = (FT[A]^* \times FT[G]) \times FT[B] \]

can be precomputed

2 FFTs + 1 matrix multiplication
2.7 Katchalski-Kazir algorithm

Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques
(protein–protein interaction/surface complementarity/macromolecular complex prediction/molecular docking)

Ephraim Katchalski-Katzir†‡, Isaac Shariv§, Miriam Eisenstein¶, Asher A. Friesem§, Claude Aflalo‖, and Ilya A. Vakser†

Departments of †Membrane Research and Biophysics, ‡Electronics, ¶Structural Biology, and †Biochemistry, Weizmann Institute of Science, Rehovot 76100, Israel

Contributed by Ephraim Katchalski-Katzir, October 24, 1991

Developed for protein-ligand docking
<= same techniques applicable for docking "on the inside"
Next, to distinguish between the surface and the interior of each molecule, we retain the value of 1 for the grid points along a thin surface layer only and assign other values to the internal grid points. The resulting functions thus become

\[ a_{l,m,n} = \begin{cases} 
1 & \text{on the surface of the molecule} \\
\rho & \text{inside the molecule} \\
0 & \text{outside the molecule,}
\end{cases} \quad [2a] \]

and

\[ b_{l,m,n} = \begin{cases} 
1 & \text{on the surface of the molecule} \\
\delta & \text{inside the molecule} \\
0 & \text{outside the molecule,}
\end{cases} \quad [2b] \]

where the surface is defined here as a boundary layer of finite width between the inside and the outside of the molecule. The parameters \( \rho \) and \( \delta \) describe the value of the points inside the molecules, and all points outside are set to zero. Two-

Typical values: \( \rho = -15, \ \delta = 1 \)

\( \Rightarrow \) penalty for overlap of volumes
Docking the hemoglobin dimer

2D cross sections at \( l = 46 \) (\( N = 90 \))

Correlation at \( \alpha = 0 \)

- a) no contact
- b) limited contact
- c) overlap (black area)
- d) good geometric match

highest peak corresponds to native dimer arrangement
The algorithm

The entire procedure described above can be summarized by the following steps:

(i) derive $\mathbf{a}$ from atomic coordinates of molecule $\mathbf{a}$ (Eq. 2),
(ii) $A^* = [\text{DFT}(\mathbf{a})]^*$ (Eq. 4),
(iii) derive $\mathbf{b}$ from atomic coordinates of molecule $\mathbf{b}$ (Eq. 2),
(iv) $B = \text{DFT}(\mathbf{b})$ (Eq. 4),
(v) $C = A^*.B$ (Eq. 5),
(vi) $\mathbf{c} = \text{IFT}(C)$ (Eq. 6),
(vii) look for a sharp positive peak of $\mathbf{c}$,
(viii) rotate molecule $\mathbf{b}$ to a new orientation,
(ix) repeat steps iii–viii and end when the orientations scan is completed, and
(x) sort all of the peaks by their height.

Each high and sharp peak found by this procedure indicates geometric match and thus represents a potential complex. The relative position and orientation of the molecules within each such complex can readily be derived from the

Katchalski-Kazir et al. 1992

Algorithm has become a workhorse for docking and density fitting.
Problem I: limited contrast

Docking of the RecA helicase monomer into simulated EM density of the hexamer at 15 Å resolution (exhaustive 6D search with 5Å / 9° steps plus off-lattice optimization) => multiple fits with similar correlations

2.6 Laplace filter

Evaluate
\[ \nabla^2 = \frac{d^2}{dx^2} + \frac{d^2}{dy^2} + \frac{d^2}{dz^2} \]

on a grid:
\[ \nabla^2 a_{l,m,n} = -6a_{l,m,n} + a_{l+1,m,n} + a_{l-1,m,n} + a_{l,m+1,n} + a_{l,m-1,n} + a_{l,m,n+1} + a_{l,m,n-1} \]

Correlation:
\[ C_{x,y,z} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} (\nabla^2 \otimes a_{l,m,n}) \times (\nabla^2 \otimes g \otimes b_{l+x,m+y,n+z}) \]
**Enhanced contrast ➔ better fit**

With the density alone:

With the Laplacian filter:

The big picture

Wriggers, Chacón, Structure 9 (2001) 779
Problem 2: more efficient search

Observations:
• many displacements can be excluded a priori (FFT alg. calculates them all)
• FFT idea makes more sense for rotations (no simple limit on rotations)
Masked displacements

Search space for displacements =
(inside of the target molecule) – (extent of the probe)
Rotational search

Express densities in spherical harmonics on "onion shells"

\[
\rho_{\text{low}}(r, \beta, \lambda) = \sum_{l=0}^{B-1} \sum_{m=-l}^{l} C_{lm}^{\text{low}}(r) Y_{lm}(\beta, \lambda) \quad \rho_{\text{high}}(r, \beta, \lambda) = \sum_{l=0}^{B-1} \sum_{m=-l}^{l} C_{lm}^{\text{high}}(r) Y_{lm}(\beta, \lambda),
\]

<table>
<thead>
<tr>
<th>(Y_{lm})</th>
<th>(l = 0)</th>
<th>(l = 1)</th>
<th>(l = 2)</th>
<th>(l = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m = -3)</td>
<td></td>
<td></td>
<td>(\sqrt{2/3} \sin^3 \phi \ e^{-3\iota \phi})</td>
<td></td>
</tr>
<tr>
<td>(m = -2)</td>
<td>(\sqrt{\tfrac{15}{4\pi}} \sin \phi \ e^{-2\iota \phi})</td>
<td>(\sqrt{\tfrac{10}{3\pi}} \sin^2 \phi \ e^{-2\iota \phi})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(m = -1)</td>
<td>(\sqrt{\tfrac{3}{4\pi}} \sin \phi \ e^{-\iota \phi})</td>
<td>(\sqrt{\tfrac{15}{4\pi}} \sin \phi \cos \phi \ e^{-\iota \phi})</td>
<td>(\sqrt{\tfrac{27}{4\pi}} \sin \phi (5 \cos^2 \phi - 1) \ e^{-\iota \phi})</td>
<td></td>
</tr>
<tr>
<td>(m = 0)</td>
<td>(\sqrt{\tfrac{1}{4\pi}} \cos \phi)</td>
<td>(\sqrt{\tfrac{5}{16\pi}} (3 \cos^2 \phi - 1))</td>
<td>(\sqrt{\tfrac{5}{16\pi}} (5 \cos^3 \phi - 3 \cos \phi))</td>
<td></td>
</tr>
<tr>
<td>(m = 1)</td>
<td>(\sqrt{\tfrac{3}{4\pi}} \sin \phi \ e^{\iota \phi})</td>
<td>(\sqrt{\tfrac{15}{4\pi}} \sin \phi \cos \phi \ e^{\iota \phi})</td>
<td>(\sqrt{\tfrac{27}{4\pi}} \sin \phi (5 \cos^2 \phi - 1) \ e^{\iota \phi})</td>
<td></td>
</tr>
<tr>
<td>(m = 2)</td>
<td>(\sqrt{\tfrac{15}{4\pi}} \sin^2 \phi \ e^{2\iota \phi})</td>
<td>(\sqrt{\tfrac{105}{4\pi}} \sin^2 \phi \cos \phi \ e^{2\iota \phi})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(m = 3)</td>
<td></td>
<td></td>
<td>(\sqrt{\tfrac{35}{4\pi}} \sin^3 \phi \ e^{3\iota \phi})</td>
<td></td>
</tr>
</tbody>
</table>

Correlation for all orientations at a given displacement:

\[
C(R) = F T_{m, h, m'}^{-1} \left[ \sum_{l} d_{m h}^{l} d_{l m'}^{l} \int_{0}^{\infty} C_{lm}^{\text{low}}(r) C_{lm'}^{\text{high}}(r) r^2 dr \right].
\]

**Known Fourier coefficients of spherical harmonics \(Y_{lm}\).**
Accuracy

Registration accuracy on simulated EM maps of 28 structures for bandwidths (number of angular sampling points) of $B = 16, 24, 32$ (11°, 8°, ~6°) compared to Wriggers’ COLORES (situs package – Katchalski-Katzir algorithm + local Powell optimization)

rmsd with respect to known atomistic structure of target.
**Performance**

**Table 1.** Timing results, in seconds, obtained with the benchmark described in Figure 1

<table>
<thead>
<tr>
<th></th>
<th>Sampling B/°</th>
<th>Resolution 10Å</th>
<th>Resolution 15Å</th>
<th>Resolution 20Å</th>
<th>Resolution 25Å</th>
<th>Resolution 30Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADP_EM</td>
<td>16/11°</td>
<td>28</td>
<td>31</td>
<td>35</td>
<td>34</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>24/8°</td>
<td>100</td>
<td>108</td>
<td>119</td>
<td>118</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>32/6°</td>
<td>226</td>
<td>220</td>
<td>225</td>
<td>216</td>
<td>221</td>
</tr>
<tr>
<td>FFT search</td>
<td>−/15°</td>
<td>1697</td>
<td>1926</td>
<td>2341</td>
<td>5028</td>
<td>6681</td>
</tr>
<tr>
<td>Powell minim</td>
<td>−/15°</td>
<td>375</td>
<td>918</td>
<td>1747</td>
<td>3739</td>
<td>6597</td>
</tr>
</tbody>
</table>

ADP_EM (Another Docking Platform for EM) is much faster
- only limited spatial region is scanned
- fast evaluation of the orientational correlation via FFT
- spherical harmonics allow for better rotational representation
  => higher accuracy
Some examples

Fig. 2. Docking results with experimental EM data. (A) *E. coli* GroES-ADP7-GroEL-ATP7 from *E. coli* at 23.5 Å (EMD ID 1046, PDB: 1ml5); ADP and ATP GroEL subunits have been docked independently to reconstruct the cis and trans heptameric rings of the complex. For GroES the whole heptamer was used. (B) Docking of 30S and 50S subunits into *E. coli* ribosome map at 14 Å (EMD ID 1046, PDB: 1gix/1gy). Single-molecule docking of prefoldin (C) at 23 Å (Martin-Benito *et al.*, 2002), PDB: 1l6h, and of yeast RNA polymerase II (D) at 15 Å (Craighead *et al.*, 2002), PDB: 1fxk.
Summary

- StarDock
- Mosaic
- Density fitting of low-resolution structures into blurred density maps
  - analogy to FFT protein-protein docking
  - speed up by FFT-transforming the rotational angles