1 Principles of Protein Docking

Biological Importance of Protein-Protein Interactions

- Protein interactions (PPIs) are central to many biological systems
  - Humans have about 30,000 proteins, each having about 5 PPIs
  - Understanding PPIs could lead to immense scientific advances

Protein-protein interactions as therapeutic drug targets
- Small “drug” molecules often inhibit or interfere with PPIs
The CAPRI Blind Docking Experiment

- CAPRI = Critical Assessment of PRedicted Interactions
- http://www.ebi.ac.uk/msd-srv/capri/
- Given the unbound structure, predict the unpublished 3D complex...

T8 = nidogen/laminin
T9 = LiCT dimer
T10 = TEV trimer
T11-12 = cohesin/dockerin
T13 = Fab/SAG1
T14 = PP1/δ MYPT1
T15 = colicin/ImmD
T18 = Xylanase/TAXI
T19 = Fab/bovine prion

- T11, T14, T19 involved homology model-building step...
- T15-T17 cancelled: solutions were on-line & found by Google !!

ICM Docking – Multi-Start Pseudo-Brownian Search

- Stick pins in protein surfaces at 15Å intervals
- For each pair of pins, find minimum energy (6 rotations for each):
  \[ E = E_{HVW} + E_{CVW} + 2.16E_{el} + 4.35E_{hp} + 0.20E_{solv} \]
- Often gives good results, but is computationally expensive

Fernández-Recio, Abagyan (2004), J Mol Biol, 335, 843–865

PatchDock – Docking by Geometric Hashing

- Use "MS" program to calculate mesh surfaces for each protein
- Divide the mesh into convex “caps”, concave “pits”, and flat “belts”
- For docking, match pairs of concave/convex, and flat/any ...
- ... then test for steric clashes between rest of surfaces
- The method is fast (minutes/seconds), and gave good results in CAPRI

Duhovný et al. (2002), LNCS 2452, 185–200
Schneidman-Duhovny et al. (2005), NAR, 33, W363–W367
Connolly (1983), J Appl Cryst, 16, 548–558

Protein Docking Using Fast Fourier Transforms

- Conventional approaches digitise proteins into 3D Cartesian grids...
- ...and use FFTs to calculated TRANSLATIONAL correlations:
  \[ C(\Delta x, \Delta y, \Delta z) = \sum_{x,y,z} A(x, y, z) \times B(x + \Delta x, y + \Delta y, z + \Delta z) \]
- BUT for docking, have to repeat for many rotations - expensive!
- Conventional grid-based FFT docking = SEVERAL CPU-HOURS

Katchalski-Katzir et al. (1992) PNAS, 89 2195–2199
Spherical Polar Fourier Correlations

Protein Docking Using Polar Fourier Correlations
- Rigid docking can be considered as a largely ROTATIONAL problem
- This means we should use ANGULAR coordinate systems
- With FIVE rotations, we should get a good speed-up?

Some Theory – 2D Spherical Harmonic Surfaces
- Spherical harmonics (SHs) are classical “special functions”
- SHs are products of Legendre polynomials and circular functions:
  - Real SHs: \( y_{lm}(\theta, \phi) = P_{lm}(\theta) \cos m\phi + P_{lm}(\theta) \sin m\phi \)
  - Complex SHs: \( Y_{lm}(\theta, \phi) = P_{lm}(\theta) e^{im\phi} \)
  - Orthogonal: \( \int y_{lm} y_{kl} d\Omega = \int Y_{lm} Y_{kl} d\Omega = \delta_{lk} \delta_{mj} \)
  - Rotation: \( y_{lm}(\theta', \phi') = \sum_{j} R_{jm}(\alpha, \beta, \gamma) y_{lj}(\theta, \phi) \)

Spherical Harmonic Molecular Surfaces
- Use spherical harmonics (SHs) as orthogonal shape “building blocks”
- Reals SHs \( y_{lm}(\theta, \phi) \), and coefficients \( a_{lm} \)
- Encode distance from origin as SH series:
  \( r(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta, \phi) \)
- Calculate coefficients by numerical integration
- Good for shape-matching, not so good for docking...

Ritchie and Kemp (1999), J. Comp. Chem. 20, 383–395
Docking Needs 3D Polar Fourier Representation
- Special orthonormal Laguerre-Gaussian radial functions, $R_n(r)$
  
  $R_n(r) = N_n^q e^{-\rho^2/2} \rho^{l+1/2} L_{n-l-1}^l(\rho)$; $\rho = r^2 / q$, $q = 20$.

Set up 3D rotational FFT as a series of matrix multiplications:

$$\sigma(l) = \begin{cases} 1; & \zeta \in \text{surface skin} \\ 0; & \text{otherwise} \end{cases} \quad \tau(l) = \begin{cases} 1; & \zeta \in \text{protein atom} \\ 0; & \text{otherwise} \end{cases}$$

Polar Fourier polynomial:

$$\sigma(l) = \sum_{n=1}^{N} \sum_{l=0}^{l-n-1} \sum_{m=-l}^{l} a_{nlm}^* R_n(r) y_{in}(\theta, \phi)$$

Analytic translations:

$$a_{nlm}^* = \sum_{m'} T_{nlml'}(R) a_{n'mm}$$  \hspace{1cm} (1)

Protein Docking Using SPF Density Functions

$$\tau(r) = \int (\sigma_A(z_A) \tau_B(z_B) + \tau_A(z_A) \sigma_B(z_B)) dV$$

Penalty Factor: $Q = 11$

Search:
- 6D space = 1 distance + 5 Euler rotations: $(R, \beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B)$

Hex SPF Correlation Example – 3D Rotational FFTs
- Set up 3D rotational FFT as a series of matrix multiplications:
  - Rotate:
    $$a_{nlm}^l = \sum_{t=-1}^1 R_{nt}^l(0, \beta_A, \gamma_A) a_{nt}$$
  - Translate:
    $$\bar{a}_{nlm}^l = \sum_{k,j} T_{nljk}^l(R) a_{nkjm}$$
  - Real to complex:
    $$A_{nlm} = \sum_t a_{ntlm} U_{tlm}, \quad B_{nlm} = \sum_t b_{ntlm} U_{tlm}$$
  - Multiply:
    $$C_{muv} = \sum_{n,l} A_{nlm}^* B_{nlm} N_{lm}$$
  - 3D FFT:
    $$S(\alpha_B, \beta_B, \gamma_B) = \sum_{muv} C_{muv} e^{-i(m \alpha_B + 2u \beta_B + v \gamma_B)}$$

On one CPU, docking takes from 15 to 30 minutes...
Exploiting Prior Knowledge in SPF Docking

- Knowing just one key residue can reduce search space enormously...
- This accelerates calculation and helps to reduce false-positives...

Docking Very Large Molecules Using Multi-Sampling

- Example: docking an antibody to the VP2 viral surface protein

Hex Protein Docking Example – CAPRI Target 3

- Example: best prediction for CAPRI Target 3 – Hemagglutinin/HC63

GPU Implementation – Perform Multiple FFTs

- Calculate multiple 1D FFTs of the form:
  \[ S_{AB}(\alpha_B) = \sum_m e^{-im\alpha_B} \sum_n A_{nlm}^a(R, \beta_A, \gamma_A) \times B_{nlm}^b(\beta_B, \gamma_B) \]
- Cross-multiply transformed A with rotated B coefficients
- Perform batch of 1D FFTs using cuFFT and save best orientations

- 3D FFTs in \((\alpha_B, \beta_B, \gamma_B)\) can be calculated in a similar way...
Results – Multiple GPUs and CPUs

- With Multi-threading, we can use all available GPUs and CPUs
- Best performance: use 2 GPUs alone, or 6 CPUs plus 2 GPUs
- 2 GPUs => 6D docking in about 15 sec – important for large-scale!

“Hex” and “HexServer”

- Hex: interactive docking (~ 40,000 downloads) – http://hex.loria.fr/
- Hexserver (~ 1,000 docking jobs/month) – http://hexserver.loria.fr/

3

Docking Symmetrical Protein Complexes

Symmetry in Protein Quaternary Structures

Many protein complexes have quaternary symmetry

- How? – Two or more asymmetric monomers related by symmetry operators

Symmetry reduces the number of degrees of freedom (D)

- So, symmetrical complexes should be easy for CAPRI dockers?
3D-Complex Shows Many Symmetrical Structures in PDB

Levy et al. (2008), Nature, 453, 1262–1265

Existing Symmetry Docking Approaches

- Post-Filter 3D FFT Grid
  - Molfit ($D_2$): Berchanski et al. (2003), Proteins, 53, 817–829
  - Cluspro ($C_n$, $D_2$, $D_3$): Comeau et al. (2005), JSB, 150, 233-244

- Symmetry-Constrained 3D FFT / Geometric Hashing
  - M-ZDOCK ($C_n$): Pierce et al. (2005), Bioinformatics, 21, 1472–1478
  - SymmDock ($C_n$): Schneidman et al. (2005), Proteins, 60, 224–231

- MD-based Energy-Minimisation
  - Rosetta3 (any): Andr´e et al. (2007), PNAS, 104, 17676–17661
  - Haddock ($C_n$, $D_n$): Karaca et al. (2010), Mol Cell Prot, 9, 1784–1794

Coordinate Operators and Docking Equations

Describe search space using operators
- Rotation: $\hat{R}(\alpha, \beta, \gamma)$
- Translation: $\hat{T}_z(R)$

Describe interaction as an “equation”
- $\hat{R}(0, \beta_A, \gamma_A)A(\ell) \leftrightarrow \hat{T}_z(R)\hat{R}(\alpha_B, \beta_B, \gamma_B)B(\ell)$

Can re-write this in many ways...
- $\hat{T}_z(R)^{-1}\hat{R}(0, \beta_A, \gamma_A)A(\ell) \leftrightarrow \hat{R}(\alpha_B, \beta_B, \gamma_B)B(\ell)$

In SPF basis, score as an overlap integral
- $S_{AB} = \int \left[ \hat{T}_z(R)^{-1}\hat{R}(0, \beta_A, \gamma_A)A(\ell) \right]^* \left[ \hat{R}(\alpha_B, \beta_B, \gamma_B)B(\ell) \right]$

The Docking Equation for $C_n$

SPF translations are easiest in the $z$ direction
- So let $y$ be principal symmetry axis, $\omega$ be symmetry angle

- $\hat{R}_y(\omega_j, 1)\hat{T}_z(D)\hat{R}(\alpha, \beta, \gamma)B(\ell) \leftrightarrow R_y(\omega_j)\hat{T}_z(D)\hat{R}(\alpha, \beta, \gamma)A(\ell)$
  - $0 \leq \alpha < \pi, 0 \leq \beta < \pi, 0 \leq \gamma < 2\pi$ (variable Euler angles)
  - $0 \leq D \leq 50$ Ångstrom (variable inter-molecular distance)
  - $\omega_j = 2\pi j/n$ (fixed by symmetry; just need $j = 0$ and $j = 1$)
Technical Note – Working Near the Origin

- With Polar Fourier correlations, it is best to work near the origin
- Computational frame (left) for FFT
- Symmetry frame (right) for results

One-Dimensional FFT Docking with \( C_n \) Symmetry

Re-order the docking equation to collect terms in \( \alpha \)

- \( \hat{T}_s(D)^{-1} \tilde{R}(0, \beta', \gamma')A(\zeta) \leftrightarrow R_s(\alpha')^{-1} \tilde{R}_s(0, \beta', \gamma')B(\zeta) \)

Apply operators to SPF expansion coefficients

- \( B_{\text{num}} = \sum_{nm} D_{nm}^{(0)}(0, \beta', \gamma')B_{\text{num}} \)
- \( A_{\text{num}} = \sum_{nj} \hat{T}_{n}^{(r)}(-D)B_{nj} \)

Expand the \( \alpha \) and \( \omega \) rotations in a similar way

- \( S(\alpha'; \omega, D, \beta', \gamma') = \sum_{mpl} A_{mpl}^{(e)}e^{-i(p-\beta)\alpha'} d_{mpl}^{(f)}(\omega)B_{mpl}^{(r)} \)

Scale onto \( 2\pi (\alpha'' = 2\alpha') \) and collect coeffs for 1D FFT in \( \alpha'' \)

- \( S(\alpha''; \omega, D, \beta', \gamma') = \sum_{t} C_t e^{-it\alpha''} \)

C\textsubscript{3} Example (Rigid-Body Reconstruction)

1F7O: Feline immuno-deficiency virus DUTP pyrophosphatase

- 107 AA, 2.2 Å resolution
- SPF expansions to \( N=30 \)
- 0.8 Å steps in \( D \)
- 2.8° steps in \( \alpha \) (by FFT)
- 7.5° steps in (\( \beta, \gamma \))

Performance Comparison (dual 6-core Intel X5650 2.67 GHz)

- SAM: 1st solution, 2.82 Å RMSD, 48 seconds
- M-ZDOCK: 1st solution, 2.33 Å RMSD, 4641 seconds
- SymmDock: 1st solution, 2.32 Å RMSD, 14 seconds

Building \( D_n \) Complexes From Two \( C_n \) Solutions

- For \( D_n \), define 2 more operators: \( \hat{T}_s(E) \) and \( \hat{R}_s(\eta) \)

- Also need allow a possible flip of one \( C_n \) plane relative to the other...

- Then dock \( k \) pairs of trial \( C_3 \) pseudo-molecules (\( i = 0 \) or 1 for flips):
  - \( P_{ui}(\zeta) = A_{ui}(\zeta) + B_{ui}(\zeta) + C_{ui}(\zeta) \)
  - (just do by brute-force search in \( E \) and \( \eta \), no FFT here)
Making Higher Symmetries From $C_3$ Trimers

$T$, $O$, and $I$ all have multiple $C_3$ axes

- But like $D_n$, we still have only 4+2 degrees of freedom
- So make positioning operators from geometry of platonic solids...

- Put a $C_3$ pseudo-molecule at two vertices and score as before...

Results – Selected 3D-Complex Examples

- All except 2 solutions are rank-1, RMSD < 3 Å
- D5/116w: rank-5, 1.3/3.6 Å RMSD; D8/1q3r: rank-25, 3.6/10.8 Å RMSD

Results – All 3D-Complex Structures

- Tested on (nearly) ALL point-group symmetry structures in 3D-Complex

- Main limitation is size of monomer (approx 500 residue limit)
- About 89% of 3D-Complex structures have monomers ≤ 500 residues...

4

The CASP/CAPRI Round 30 Targets
The CASP Protein Structure Prediction Experiment

What is CASP?
- CASP = Critical Assessment of (Protein) Structure Prediction
  - http://predictioncenter.org/
- Given a protein sequence, predict the unpublished 3D structure

Organisation
- Crystallographers deposit new 3D structures with CASP
- CASP organisers release sequences of targets to predictors
- CASP predictors may submit up to 5 models per target
- Any method is allowed — homology template / ab inito...
- Model quality is assessed later by a variety of methods...

CASP-12/CAPRI-30
- Many CASP targets were symmetric homo-oligomers
- CAPRI participants invited to dock the CASP models...

Using KBDOCK to Find Symmetric DDI Templates

- PfamScan to find Pfam family
- Select “homo” DDIs...
- Then, eye-ball for symmetry

Ghoorah et al. (2014), Nucleic Acids Research, 42, D389–D395

Using Kpax to Select a “Representative” CASP Model

- Download the “Stage-2” CASP models...
- Find centre model using Kpax all-vs-all structural alignments (e.g., T85)

kpax -all T0813/*

Zhang_Server_TS4

Ritchie et al. (2012), Bioinformatics, 28, 3274–3281

The Round 30 Targets That We Attempted

<table>
<thead>
<tr>
<th>Target</th>
<th>Category</th>
<th>CASP Stage-2</th>
<th>Kpax</th>
<th>Kpax/SAM</th>
<th>CAPRI Top-10</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>T79</td>
<td>DIFFICULT</td>
<td>FAS603_TS1</td>
<td>JRCO</td>
<td>1 / 0</td>
<td>M01 / KBDOCK</td>
<td>MEDIUM</td>
</tr>
<tr>
<td>T80</td>
<td>EASY</td>
<td>FALCON MANUAL_X_TS2</td>
<td>20GA 3NYU</td>
<td>1-2 / 3-10</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T81</td>
<td>HETERO</td>
<td>STRINGS_TS1</td>
<td>2G2N</td>
<td>1-2 / 3-10 (*)</td>
<td>M01 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T82</td>
<td>EASY</td>
<td>FALCON_ENVFold_TS4</td>
<td>3ED8</td>
<td>1 / 2-10</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T83</td>
<td>CANCELED</td>
<td></td>
<td></td>
<td></td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T84</td>
<td>EASY</td>
<td>PhynX_TS1</td>
<td>4BI5</td>
<td>1 / 2-10</td>
<td>M01 / KBDOCK</td>
<td>MEDIUM</td>
</tr>
<tr>
<td>T85</td>
<td>EASY</td>
<td>Zhang_Server_TS4</td>
<td>2FKK 3GCP</td>
<td>1-2 / 3-10</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T86</td>
<td>DIFFICULT</td>
<td>FALCON_ENVFold_TS5</td>
<td>43HU 30Y 3ER7</td>
<td>1-3 / 4-10</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T87</td>
<td>EASY</td>
<td>FALCON_ENVFold_TS1</td>
<td>20OR</td>
<td>0 / 1-10</td>
<td>M01 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T88</td>
<td>DIFFICULT</td>
<td>PhynX_TS1</td>
<td>2X1U</td>
<td>1 / 2-10</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T89</td>
<td>HETERO</td>
<td></td>
<td></td>
<td>0 / 1-10 (*)</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
<tr>
<td>T90</td>
<td>EASY</td>
<td></td>
<td></td>
<td>0 / 1-10</td>
<td>M07 / SAM</td>
<td>ACCEPTABLE</td>
</tr>
</tbody>
</table>

* used Hex instead of SAM

- We were the only predictor group to get acceptable models for T86
- Overall, we modeled 4 / 9 symmetric homo-dimers (2 KBDOCK, 2 SAM)
- But our models for T84 and T87 were rejected due to steric clashes...
**Our Best Solution With KBDOCK – T80**
- T80–M01: $F_{\text{nat}} = 0.493$, LRMSD = 3.77 Å, IRMSD = 1.97 Å

**Our Best Solution With SAM – T86**
- T86–M09: $F_{\text{nat}} = 0.667$, LRMSD = 8.97 Å, IRMSD = 2.55 Å

**Question:** Is “LRMSD” really appropriate for symmetry targets?
- No doubt, RMSD(Receptor+Ligand) would give a smaller number ...

**Does NAMD Energy Minimisation Help?**
- After CAPRI, we used NAMD to minimise the top 50 SAM solutions
  - (still working with CASP centre models here)

<table>
<thead>
<tr>
<th>Target</th>
<th>Centre Model</th>
<th>Rank</th>
<th>SAM RMSD</th>
<th>Rank</th>
<th>NAMD RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T69/C2</td>
<td>SAM-T08-server_TS3</td>
<td>8</td>
<td>3.18</td>
<td>1</td>
<td>1.01</td>
</tr>
<tr>
<td>T72/C2</td>
<td>3D-Jigsaw-V5.1_TS1</td>
<td>13</td>
<td>5.55</td>
<td>1</td>
<td>6.18</td>
</tr>
<tr>
<td>T75/C2</td>
<td>SAM-T08-server_TS4</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T77/C2</td>
<td>FALCON_EnvFold_TS3</td>
<td>–</td>
<td>–</td>
<td>3</td>
<td>3.57</td>
</tr>
<tr>
<td>T80/C2</td>
<td>MULTICOM-CLUSTER_TS1</td>
<td>32</td>
<td>5.75</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T86/C2</td>
<td>FALCON_EnvFold_TS5</td>
<td>32</td>
<td>8.06</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T87/C2</td>
<td>FALCON_TOPO_TS5</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>0.58</td>
</tr>
<tr>
<td>T94/C2</td>
<td>3D-Jigsaw-V5.1_TS3</td>
<td>–</td>
<td>–</td>
<td>9</td>
<td>2.75</td>
</tr>
<tr>
<td>T70/D2</td>
<td>FALCON_TOPO_TS1</td>
<td>–</td>
<td>–</td>
<td>19</td>
<td>9.70</td>
</tr>
<tr>
<td>T71/D2</td>
<td>MULTICOM-CLUSTER_TS1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T73/D2</td>
<td>3D-Jigsaw-V5.1_TS2</td>
<td>–</td>
<td>–</td>
<td>5</td>
<td>3.60</td>
</tr>
</tbody>
</table>

- Often get more good solutions after NAMD
- But surprisingly, NAMD fails to improve T80 and T86

**So, How Many CASP Models are Dockable by SAM?**
- SAM Energy v’s RMSD for top 10 solutions for all attempted CASP models

- => CASP community can nearly always produce “dockable” monomers
- BUT it would be expensive to minimise many CASP models with NAMD...
Conclusions

SAM
- New “SAM” code for FFT-based symmetry docking
  - (every model is perfectly symmetrical)

NAMD
- Using NAMD to minimise top SAM solutions helps, but is expensive

CASP / CAPRI
- Many recent CASP / CAPRI targets were symmetric multimers
  - (surprisingly many symmetric complexes exist in Nature)

The future?
- Many challenges remain in protein docking
  - e.g. docking very large non-symmetrical structures
  - e.g. exploiting new low resolution data from cryo-EM

Thank You!

Acknowledgments
Sergei Grudinin
Marwa El Houasli

Program and papers:
http://hex.loria.fr/
http://sam.loria.fr/