Protein Docking and Molecular Shape Recognition using Polar Fourier Correlations

Dave Ritchie
INRIA Nancy – Grand Est
Protein Docking – To Predict Protein-Protein Interactions

- Protein-protein interactions (PPIs) define the “machinery” of life

- Humans have about 30,000 proteins, each having about 5 PPIs
- Understanding PPIs could lead to immense scientific advances
- Controlling PPIs could have huge therapeutic benefits (new drug molecules)
Why is Protein Docking Difficult?

- Protein docking = predicting protein interactions at the molecular level

- If proteins are rigid => six-dimensional search space
- But proteins are flexible => multi-dimensional space!
- Modeling protein-protein interactions accurately is difficult!
Protein Docking Using Fast Fourier Transforms

- Conventional approaches digitise proteins into 3D Cartesian grids...

Katchalski-Katzir et al. (1992) PNAS, 89 2195–2199

...and use FFTs to calculate 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!

- POLAR coords allow ROTATIONAL nature of problem to be exploited
Some Theory – 2D Spherical Harmonic 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 to order $L$:
  $r(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta, \phi)$

- Calculate coefficients by numerical integration

- ROTATIONS: $a'_{lm} = \sum_{m'=-l}^{l} R_{mm'}^{(l)}(\alpha, \beta, \gamma) a_{lm}$

- Good for shape-matching, not so good for docking...

Ritchie and Kemp (1999), J. Comp. Chem. 20 383–395
Docking Needs a 3D “Spherical Polar Fourier” Representation

• Need to introduce special orthonormal Laguerre-Gaussian radial functions, $R_{nl}(r)$

• $R_{nl}(r) = N_{nl}^{(q)} e^{-\rho/2} \rho^{l/2} L_{n-l-1}^{(l+1/2)}(\rho); \quad \rho = r^2/q, \quad q = 20.$

• Surface Skin: $\sigma(r) = \begin{cases} 1; & r \in \text{surface skin} \\ 0; & \text{otherwise} \end{cases}$

• Interior: $\tau(r) = \begin{cases} 1; & r \in \text{protein atom} \\ 0; & \text{otherwise} \end{cases}$

• Parametrise as: $\sigma(r) = \sum_{n=1}^{N} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} a_{nlm} \sigma R_{nl}(r) y_{lm}(\theta, \phi)$

• TRANSLATIONS: $a_{nlm}^{\sigma'} = \sum_{n'l'} T_{nl,n'l'}^{(|m|)}(R) a_{n'l'm}^{\sigma}$
SPF Protein Shape-Density Reconstruction

Interior density:  \[ \tau(r) = \sum_{nlm} a_{n\ell m}^\tau R_{n\ell}(r) y_{\ell m}(\theta, \phi) \]

<table>
<thead>
<tr>
<th>Image</th>
<th>Order</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Gaussians</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>N = 16</td>
<td>1,496</td>
</tr>
<tr>
<td>C</td>
<td>N = 25</td>
<td>5,525</td>
</tr>
<tr>
<td>D</td>
<td>N = 30</td>
<td>9,455</td>
</tr>
</tbody>
</table>

Protein Docking Using SPF Density Functions

\[ \int (\sigma_A(r_A) \tau_B(r_B) + \tau_A(r_A) \sigma_B(r_B)) \, dV \]

Favourable:

\[ \int \tau_A(r_A) \tau_B(r_B) \, dV \]

Unfavourable:

\[ S_{AB} = \int (\sigma_A \tau_B + \tau_A \sigma_B - Q \tau_A \tau_B) \, dV \quad \text{Penalty Factor: } Q = 11 \]

Score:

Orthogonality:

\[ S_{AB} = \sum_{nlm} (a^\sigma_{nlm} b^\tau_{nlm} + a^\tau_{nlm} (b^\sigma_{nlm} - Q b^\tau_{nlm})) \]

Search:

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

The CAPRI Experiment (Critical Assessment of PRedicted Interactions)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Software</th>
<th>Algorithm</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abagyan</td>
<td>ICM</td>
<td>FF</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>***</td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Camacho</td>
<td>CHARMM</td>
<td>FF</td>
<td>*</td>
<td></td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Eisenstein</td>
<td>MolFit</td>
<td>FFT</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Sternberg</td>
<td>FTDOCK</td>
<td>FFT</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>Ten Eyck</td>
<td>DOT</td>
<td>FFT</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>Gray</td>
<td></td>
<td>MC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>**</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Ritchie</td>
<td>Hex</td>
<td>SPF</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weng</td>
<td>ZDOCK</td>
<td>FFT</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>Wolfson</td>
<td>BUDDA/PPD</td>
<td>GH</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Bates</td>
<td>Guided Docking</td>
<td>FF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>Palma</td>
<td>BIGGER</td>
<td>GF</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>Gardiner</td>
<td>GAPDOCK</td>
<td>GA</td>
<td>*</td>
<td>*</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Olson</td>
<td>Surfdock</td>
<td>SH</td>
<td>*</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Valencia</td>
<td>ANN</td>
<td></td>
<td>*</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vakser</td>
<td>GRAMM</td>
<td>FFT</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* low, ** medium, *** high accuracy prediction; — no prediction

Hex Protein Docking Example – CAPRI Target 3

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


Best Hex Orientation for Target 6 – Amylase/AMD9

- CAPRI “high accuracy” (Ligand RMSD ≤ 1Å)
### Subsequent CAPRI Targets 8 – 19

<table>
<thead>
<tr>
<th>Target</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>T8</td>
<td>Nidogen- γ3 - Laminin</td>
<td>U/U</td>
</tr>
<tr>
<td>T9</td>
<td>LiCT homodimer</td>
<td>build from monomer – 12Å RMS deviation</td>
</tr>
<tr>
<td>T10</td>
<td>TBEV trimer</td>
<td>build from monomer – 11Å RMS deviation</td>
</tr>
<tr>
<td>T11</td>
<td>Cohesin - dockerin</td>
<td>U/U; model-build dockerin</td>
</tr>
<tr>
<td>T12</td>
<td>Cohesin - dockerin</td>
<td>U/B</td>
</tr>
<tr>
<td>T13</td>
<td>SAG1 - antibody Fab</td>
<td>SAG1 conformational change: 10Å RMS</td>
</tr>
<tr>
<td>T14</td>
<td>MYPT1 - PP1 δ</td>
<td>U/U; model-build PP1 α → PP1 δ</td>
</tr>
<tr>
<td>T18</td>
<td>TAXI - xylanase</td>
<td>U/B</td>
</tr>
<tr>
<td>T19</td>
<td>Ovine prion - antibody Fab</td>
<td>model-build prion</td>
</tr>
</tbody>
</table>

- T15-T17 cancelled: solutions were on-line & found by Google !!!
- T11, T14, T19 involved homology model-building step...
### CAPRI Results: Targets 8–19 (2003 – 2005)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Software</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
<th>T13</th>
<th>T14</th>
<th>T18</th>
<th>T19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abagyan</td>
<td>ICM</td>
<td>**</td>
<td>*</td>
<td></td>
<td>***</td>
<td>*</td>
<td>***</td>
<td>*</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Wolfson</td>
<td>PatchDock</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>_</td>
<td>**</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Weng</td>
<td>ZDOCK/RDOCK</td>
<td>**</td>
<td>*</td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Bates</td>
<td>FTDOCK</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baker</td>
<td>RosettaDock</td>
<td>-</td>
<td>**</td>
<td></td>
<td>***</td>
<td>**</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>Camacho</td>
<td>SmoothDock</td>
<td>**</td>
<td>**</td>
<td></td>
<td>***</td>
<td>***</td>
<td>**</td>
<td>**</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Gray</td>
<td>RosettaDock</td>
<td>***</td>
<td>_</td>
<td></td>
<td>***</td>
<td>**</td>
<td>***</td>
<td></td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>Bonvin</td>
<td>Haddock</td>
<td>-</td>
<td>**</td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Comeau</td>
<td>ClusPro</td>
<td>**</td>
<td>**</td>
<td>***</td>
<td></td>
<td>*</td>
<td>***</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Sternberg</td>
<td>3D-DOCK</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Eisenstein</td>
<td>MolFit</td>
<td>***</td>
<td>*</td>
<td>***</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ritchie</td>
<td>Hex</td>
<td></td>
<td>**</td>
<td>***</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhou</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td>***</td>
<td>**</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ten Eyck</td>
<td>DOT</td>
<td></td>
<td></td>
<td></td>
<td>***</td>
<td>***</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zacharias</td>
<td>ATTRACT</td>
<td>**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>***</td>
<td></td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>Valencia</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td>_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vakser</td>
<td>GRAMM</td>
<td>-</td>
<td>-</td>
<td></td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Umeyama</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td>***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaznesssis</td>
<td></td>
<td>-</td>
<td>-</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fano</td>
<td>Grid-Hex</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Hex” and “HexServer”

- **Hexserver** – [http://hexserver.loria.fr/](http://hexserver.loria.fr/) – about 1,000 docking jobs per month


...  

Macindoe et al. (2010), Nucleic acids Research, 38 W445–W449
Inside Hex – High Order FFTs, Multi-threading on GPUs

• The SPF gives an analytic way to calculate TRANSLATIONAL + ROTATIONAL correlations:

In particular:

\[ S_{AB} = \sum_{jsmlvrt} \Lambda_{js}^{rm} T_{js,lv}^{(m)}(R) \Lambda_{lv}^{tm} e^{-i(r\beta_A-s\gamma_A+m\alpha_B+t\beta_B+v\gamma_B)} \]

• This allows high order FFTs to be used – 1D, 3D, and 5D

• It also allows calculations to be easily ported to modern GPUs

• Up to 512 arithmetic “cores”
• Up to 6 Gb memory
• Easy API with C++ syntax
• Grid of threads model ("SIMT")

• BUT – for best results, need to understand the hardware...

Ritchie, Kozakov, Vajda (2008), Bioinformatics 24 1865–1873
Ritchie and Venkatraman (2010), Bioinformatics, 26, 2398–2405
CUDA Device Architecture

• Typically 8–16 multiprocessor blocks, each with 16 thread units

- NB. only a very small amount of fast shared memory is available
- NB. global memory is ~ 80x slower than shared memory
- Strategy: aim for “high arithmetic intensity” in shared memory
CUDA Programming Example - Matrix Multiplication

- Matrix multiplication $C = A \times B$
- Each thread is responsible for calculating one element: $C[i,k]$ 

\[
\begin{array}{c}
\begin{array}{c}
\text{C} \\
\text{by}
\end{array}
\end{array}
= \begin{array}{c}
\begin{array}{c}
\text{A} \\
\text{tx}
\end{array}
\end{array} \times \begin{array}{c}
\begin{array}{c}
\text{B} \\
\text{ty}
\end{array}
\end{array}
\]

- Conventional algorithm: rows and columns
  - $C[i,k] = A[i] \times B[k]$
  
- Thread-block algorithm working on TILES

- A tile size of 16x16 is just right!
- Threads co-operate by reading & sharing tiles of A & B
- Multi-processor launches multiple blocks to compute all of C
- Executing thread-blocks concurrently hides global memory latency
GPU Implementation – Perform Multiple FFTs

• Next, calculate multiple 1D FFTs of the form:

\[ S_{AB}(\alpha_B) = \sum_m e^{-im\alpha_B} \sum_{nl} A_{nlm}^\sigma(R, \beta_A, \gamma_A) \times B_{nlm}^\tau(\beta_B, \gamma_B) \]

4. On GPU, cross-multiply transformed A with rotated B coefficients (as above)

5. On GPU, 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 as many GPUs and CPUs as are available

• For best performance: use 2 GPUs alone, or 6 CPUs plus 2 GPUs

• With 2 GPUs, docking takes only about 15 seconds – very important for large-scale!
Speed Comparison with ZDOCK and PIPER

- Hex: 52000 x 812 rotations, 50 translations (0.8 Å steps)
- ZDOCK: 54000 x 6 deg rotations, 92Å 3D grid (1.2Å cells)
- PIPER: 54000 x 6 deg rotations, 128Å 3D grid (1.0Å cells)
- Hardware: GTX 285 (240 cores, 1.48 GHz)

<table>
<thead>
<tr>
<th>FFT</th>
<th>ZDOCK</th>
<th>PIPER†</th>
<th>PIPER‡</th>
<th>Hex</th>
<th>Hex</th>
<th>Hex‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>7,172</td>
<td>468,625</td>
<td>26,372</td>
<td>224</td>
<td>60</td>
<td>84</td>
</tr>
<tr>
<td>(3D)*</td>
<td>(1,195)</td>
<td>(42,602)</td>
<td>(2,398)</td>
<td>224</td>
<td>60</td>
<td>84</td>
</tr>
<tr>
<td>1D</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>676</td>
<td>243</td>
<td>15</td>
</tr>
</tbody>
</table>

# execution times in seconds
* (times scaled to two-term potential, as in Hex)

- What’s next?
  - Better energy functions & constraints...
  - Using homology templates...
  - Modeling flexibility...
  - Multi-component complexes...
Conclusions and Future Prospects

(+) Rigid-body docking on a GPU now takes only a few seconds:
   • This was implemented using only 5 or 6 GPU kernels

(−) Modeling protein flexibility during docking is still difficult

• With SPF correlations, high-throughput shape comparison is now feasible:
  • All-vs-all docking ?
  • Electron-microscopy density fitting ?
  • Assembling multi-component machines ?

(?) The challenge for this decade – “the structural interactome”
Acknowledgments

BBSRC 1996 – 2000
EPSRC 2000 – 2006

ANR 2009 – 2010
ANR 2011 – 2014

Anisah Ghoorah
Matthieu Chavent
Violeta Pérez-Nuño
Vishwesh Venkatraman
Lazaros Mavridis

People, Papers, Programs:  http://www.loria.fr/~ritchied/