PEPSI-DOCK

A Detailed Data-Driven Protein-Protein Interaction Potential Accelerated By Polar Fourier Correlations

MACARON Workshop- March 21st 2017

Emilie Neveu, Dave Ritchie, Petr Popov, Sergei Grudinin
Nano-D & Capsid INRIA Teams
ABOUT PROTEINS

Definition

protein

(long) chain of amino acids (aa)

side chain

20 possible aa
ABOUT PROTEINS

Definition

protein  (long) chain of amino acids (aa)  20 possible aa

Representation

“cartoon”-like 3D structure

flexible pieces: structures not well-defined

stable pieces: helices, parallel sheets
ABOUT DOCKING

Structure Prediction

Protein - receptor

Protein - ligand

1.  
2.  
3.  

...  

N.

ABOUT DOCKING

Structure Prediction

Protein - receptor

Protein - ligand

1.  

2.  

3.  

…

N.

Why so important?

Influenza virus

Hemagglutinin protein

100 nm

Inhibitor

ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)
ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. **Interaction energy** to score/assess the structures

\[ \Delta G_{\text{bind}} = \Delta H - T\Delta S \]

- enthalpy
- entropy
ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. **Interaction energy** to score/assess the structures

2. **Search algorithm** + set of parameters
ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. **Interaction energy** to score/assess the structures

2. **Search algorithm** + set of parameters

3. **Multilevel approach**: selection of top solutions; restart with higher resolution
ABOUT DOCKING

> 2001 Community-wide experiment: CAPRI (Critical Assessment of PRedicted Interactions)

1. Interaction energy to score/assess the structures

2. Search algorithm + set of parameters

3. Multilevel approach: selection of top solutions; restart with higher resolution

starring

ZDock zdock.umassmed.edu
HexDock hex.loria.fr/hex.php
ClusPro cluspro.bu.edu
AutoDock autodock.scripps.edu
RosettaDock rosie.rosettacommons.org/ligand_docking
DOCK dock.compbio.ucsf.edu
and many others.....
GOAL: To improve the first level: large and global search space
GOAL: To improve the first level: large and global search space

Simple but accurate interaction energy approximation

- SVM-based algorithm to learn the atomistic potentials
- Physically interpretable features:
  - Number densities of site-site pairs at a given distance
- Arbitrarily shaped atomistic distance dependent interaction potentials

**GOAL:** To improve the first level: large and global search space

Simple but accurate interaction energy approximation

**Fast exploration**
- rigid bodies assumption
- spherical Fourier correlation: complexity from $O(N^9)$ to $O(N^6 \log N)$

D.W. Ritchie, D. Kozakov, and S. Vajda, Hex code
GOAL: To improve the first level: large and global search space

Simple but accurate interaction energy approximation

Fast exploration

Sparse representation in Gauss-Laguerre basis
1. Features extraction

2. Sparse Representation

3. Optimisation in Gauss-Laguerre basis

4. Stored 210 atomistic distance dependent potentials

5. From 1D to 3D

6. Fast exploration of the search space

7. Ranked docking predictions
Detailed description of 1-D interactions at the interface

195 native non-redundant complexes from ITScore Training Set
[Zou Lab, University of Missouri Columbia]

1-D native distributions of atom pairs / distance

40 000 generated false complexes

1-D non-native distributions of atom pairs / distance
Detailed description of 1-D interactions at the interface

195 native non-redundant complexes
from ITScore Training Set
[Zou Lab, University of Missouri Columbia]

20 different atom types \( \Rightarrow \) 210 interactions

40,000 generated false complexes

1-D native distributions of atom pairs / distance

1-D non-native distributions of atom pairs / distance
1 & 2 - Features Extraction/ Sparse representation

**Detailed description of 1-D interactions at the interface**

- 195 native non-redundant complexes from ITScore Training Set [Zou Lab, University of Missouri Columbia]
- 40 000 generated false complexes
- 1-D native distributions of atom pairs / distance
- 1-D non-native distributions of atom pairs / distance
- 20 different atom types $\Rightarrow$ 210 interactions

**Sparse representation**

- in a Gauss-Laguerre polynomial basis
- scaled to describe distributions up to 30 Å
- about 6300 geometric features for each native and non-native complex
Optimal discrimination between native and non-native interfaces

convex optimisation problem: Find $w$ and $b^c$ that minimise

$$\min_{w, b^c} \frac{\lambda}{2} \|w\|^2 + \gamma \sum_c \log \left( 1 + e^{y^c(w^T v^c + b^c)} / \gamma \right)$$

- prevents overfitting
- penalises misclassification

features $v^c$; classifier known

- native complexes $y^c = 1$
- associated false complexes $y^c = -1$

hyperplane separator estimated

- $w$ normal vector: 1-D interaction potentials
- $b^c$ margin

Knowledge of Native Protein–Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates.

N+ with O-

4 - 210 atom-atom distance dependent interaction potentials

\[ w_{\text{atom-atom}} \text{ distance-dependent interaction potentials} \]

\[ \text{precision} \]

\[ 210 \text{ interactions} \]

\[ r_{\text{max}} \text{: cut-off distance} \]

\[ \text{regularisation parameter} \]

\[ x_{ij}, y_{ij} \]

\[ 210 \text{ interactions} \]

\[ \text{precision} \]

\[ w_{\text{atom-atom}} \text{ distance-dependent interaction potentials} \]

\[ r_{\text{max}} \text{: cut-off distance} \]

\[ \text{regularisation parameter} \]

\[ x_{ij}, y_{ij} \]

\[ 210 \text{ interactions} \]
Linear sum of atom-atom convolution with potentials and densities

\[ E = \sum_{\text{pairwise interactions } ij} \sum_{R_i} \sum_{L_j} \int \int \int_{V} f_{ij}(x - x_{R_i}) g(x - x_{L_j}) dV \]

Representation with truncated polynomial expansion

\[ \int \int \int_{V} f_{ij}(r) g(r - x_{L_j}) dV = \sum_{nml} \left( R.T.w \right)_{nml} \cdot g_{nml} = f_{nml}^{ij} \]
6 - Exploration of the search space: the Hex engine

**Rigid body assumption**

Energy depends to rigid positions of proteins

\[ E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B) \]

- 1 translation and 5 rotations to adjust
- discretised to enable exhaustive search

\[ R \in [0 : 1 : 40 \text{ Å}] \]
\[ \alpha \in [0 : 7.5 : 360^\circ] \]
\[ (\beta, \gamma) \in [0 : 7.5 : 180^\circ]^2 \]
Rigid body assumption

Energy depends to rigid positions of proteins

\[ E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B) \]

- 1 translation and 5 rotations to adjust
- discretised to enable exhaustive search

\[ R \in [0 : 1 : 40 \text{ Å}] \]
\[ \alpha \in [0 : 7.5 : 360^\circ] \]
\[ (\beta, \gamma) \in [0 : 7.5 : 180^\circ]^2 \]

Fast exhaustive search

Truncated expressions using spherical Fourier correlation

\[ E(R, \beta_A, \gamma_A, \beta_B, \gamma_B, \alpha_B) = \text{DFT}^{-1} \left[ \text{DFT} \left( R_{\beta_A, \gamma_A} T R f_A \right) \cdot \text{DFT} \left( R_{\alpha_B, \beta_B, \gamma_B} g_B \right) \right] \]

complexity from \( O(N^9) \) to \( O(N^6 \log N) \): \( 10^9 \) poses in \( \sim 10 \) min

Accelerating and Focusing Protein-Protein Docking Correlations Using Multi-Dimensional Rotational FFT Generating Functions.
Test on 88 complexes from the Docking Benchmark Set v5.0 for which the separation distance ≤ 30 Å

Docking Benchmark Set = the only existing benchmark to compare different docking algorithms

[Hwang, Vreven, Janin, Weng, 2010]

Comparison on v4.0 Top 10 for I-RMS ≤ 2.5Å

<table>
<thead>
<tr>
<th>Category (Nb. of Complexes)</th>
<th>ZDOCK.</th>
<th>SwarmDock</th>
<th>PEPSI-Dock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy (45)</td>
<td>12</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Medium (15)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Difficult (15)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Running Time of PEPSI-Dock measured on a modern laptop

Docking of $10^9$ poses in less than 10 min on a laptop ~ weeks of a 1 μs MD simulation
A docking automatic algorithm for the first stage of the docking pipeline

» novelty: arbitrarily -shaped + distance-dependent potentials combined with a FFT search sampling technic
A docking automatic algorithm for the first stage of the docking pipeline

- novelty: arbitrarily-shaped + distance-dependent potentials combined with a FFT search sampling technic

TO DO

1. Improve unbound predictions: use other training set
A docking automatic algorithm for the first stage of the docking pipeline

» novelty: arbitrarily -shaped + distance-dependent potentials combined with a FFT search sampling technic

› Bound sets: High-rank predictions

› Large distances ⚠ loss of precision

› Unbound sets: similar results than SwarmDock or ZDOCK

› Adaptation to other types of interactions

TO DO

1. Improve unbound predictions: use other training set
A docking automatic algorithm for the first stage of the docking pipeline

» novelty: arbitrarily-shaped + distance-dependent potentials combined with a FFT search sampling technic

› Bound sets: High-rank predictions

› Large distances ⚠ loss of precision

› Unbound sets: similar results than SwarmDock or ZDOCK

› Adaptation to other types of interactions

TO DO

1. Improve unbound predictions: use other training set

2. Deal with the docking of large proteins: use other sampling
PEPSI-Dock, Neveu et al., Bioinformatics, 2016
PEPSI-Dock, Neveu et al., Bioinformatics, 2016

https://www.samson-connect.net
THANKS

@ Sergei Grudinin
@ Petr Popov
Nano-D team
INRIA Grenoble

@ David Ritchie
Capsid Team
INRIA Nancy
ANR

https://www.samson-connect.net
ANY QUESTION?