Introduction: Machine learning for structure-based virtual screening

CS/CME/Biophys/BMI 371

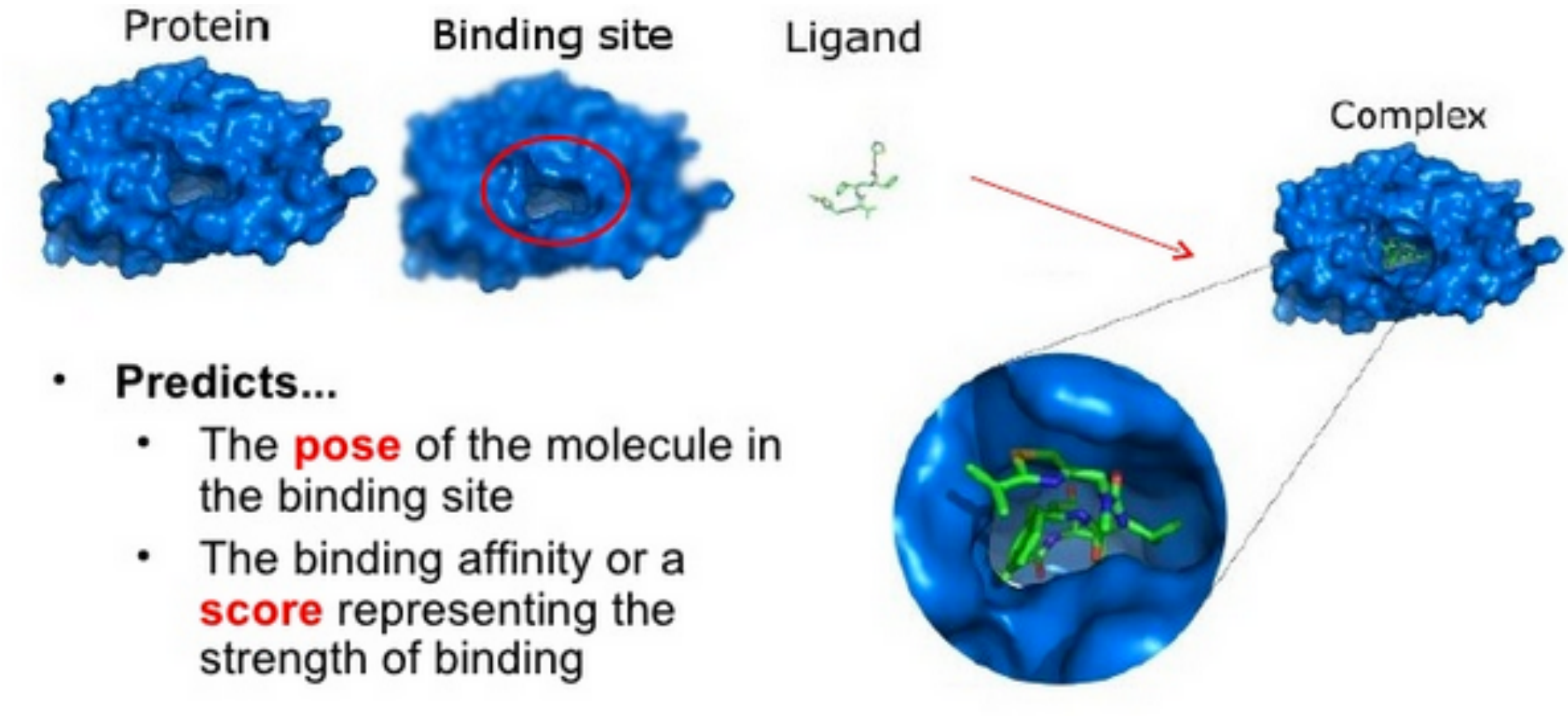
Jan. 25, 2017

Ron Dror

Virtual screening

- Virtual screening: Identifying drug candidates by considering large numbers of possible ligands
  - A ligand is any molecule that might bind to a protein
- Virtual screening is an alternative to experimental high-throughput screening (done by robots)
- Once a candidate is identified, it undergoes an extensive optimization process in which it is modified chemically to improve its properties
  - This optimization is a big part of drug discovery

Ligand docking: standard approach to virtual screening



http://www.slideshare.net/baoilleach/proteinligand-docking-13581869

Note that predicting binding *pose* (i.e., where each atom of the ligand ends up) is very important in its own right, particularly for the ligand optimization process

Ligand docking software

Program	Country of Origin	Year Published
AADS	India	2011
ADAM	Japan	1994
AutoDock	USA	1990
AutoDock Vina	USA	2010
BetaDock	South Korea	2011
DARWIN	USA	2000
DIVALI	USA	1995
DOCK	USA	1988
DockVision	Canada	1992
EADock	Switzerland	2007
eHiTS	UK	2006
EUDOC	USA	2001
FDS	UK	2003
FlexE	Germany	2001
FlexX	Germany	1996
FLIPDock	USA	2007
FLOG	USA	1994
FRED	UK	2003
FTDOCK	UK	1997
GEMDOCK	Taiwan	2004
Glide	USA	2004
GOLD	UK	1995
Hammerhead	USA	1996
ICM-Dock	USA	1997
Lead finder	Canada	2008
LigandFit	USA	2003
LigDockCSA	South Korea	2011
LIGIN	Germany	1996
LUDI	Germany	1992
MADAMM	Portugal	2009
MCDOCK	USA	1999
MDock	USA	2007
MolDock	Denmark	2006
MS-DOCK	France	2008
ParDOCK	India	2007
PhDOCK	USA	2003
PLANTS	Germany	2006
PRO_LEADS	UK	1998
PRODOCK	USA	1999
ProPose	Germany	2004
PSI-DOCK	China	2006
PSO@AUTODOCK	Germany	2007
PythDock	South Korea	2011
Q-Dock	USA	2008
QXP	USA	1997
rDock	UK	2013
SANDOCK	UK	1998
SFDOCK	China	1999
SODOCK	Taiwan	2007
SOFTDocking	USA	1991
Surflex	USA	2003
SYSDOC	USA	1994
VoteDock	Poland	2011
YUCCA	USA	2005

Most popular (based on citations 2001–2011):

AutoDock
GOLD
DOCK
FlexX
Glide
FTDOCK
QXP

Sousa et al., Current Medicinal Chemistry 2013

http://en.wikipedia.org/wiki/Docking_(molecular)

So what's the problem with ligand docking?

- Ligand docking is a physics-based heuristic approach with two key components
  - A *scoring function* that very roughly approximates the binding affinity (i.e., binding strength) of a ligand to a protein given a binding pose
  - A *search method* that searches for the best-scoring binding pose for a given ligand
- Accuracy is poor!

What can we do?

- Try a machine learning approach
  - Forget the physics and treat this instead as a statistical pattern recognition problem
  - Learn from existing structures of protein-ligand complexes (and, perhaps, data on affinities of many ligands to proteins)
  - Monday's papers
- Improve the physics-based approaches
  - Next Wednesday's papers

Background material

- Ligand docking slides from CS/CME/BioE/ Biophys/BMI 279:
  - http://web.stanford.edu/class/cs279/lectures/lecture7.pdf
- Deep learning tutorial (Andrew Ng's group)
  - http://ufldl.stanford.edu/tutorial/