Evaluation & Comparison of Virtual Screening Tools

(focussing on docking)

Romano T. Kroemer

Molecular Modeling & Design
Pharmacia - Italy

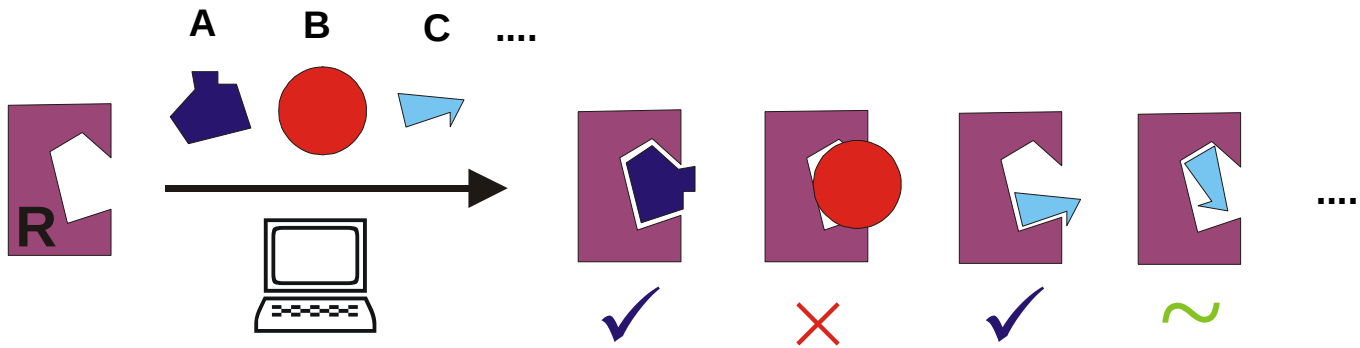
PHARMACIA

Overview

- Introduction
- Evaluation: Pose Prediction
- Evaluation: Affinity Prediction
- Development of Scoring Functions
- Conclusions

PHARMACIA

Introduction: Docking



"Pose" Prediction (*same* molecule):
C_1 (correct orientation) **better** than C_2

Affinity Prediction (*different* molecules):
A makes **better** interactions with R than C_1
C_1 **better** than **B**

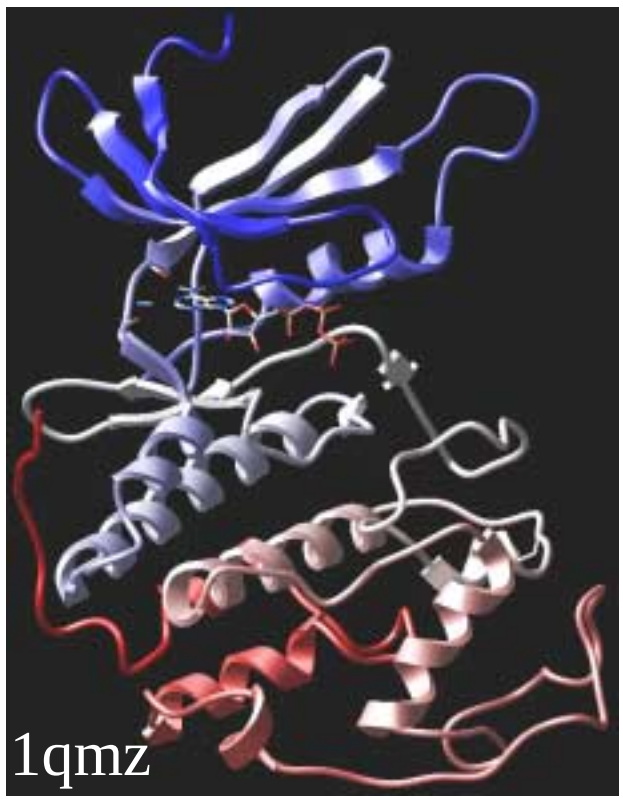
⇒ **Rank Order** molecules according to predicted affinity

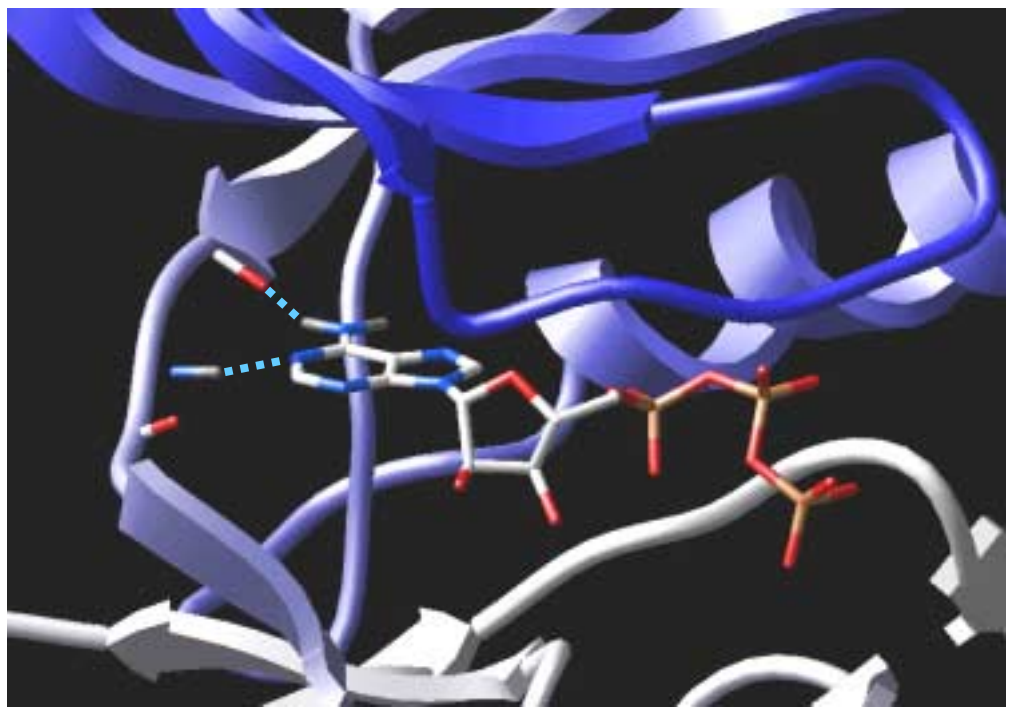
Scoring Function: Interaction Energies, Solvation/Desolvation, Entropy, ...

PHARMACIA

Evaluation: Pose Prediction

- **Data Set 1: 20 CDK2 – Inhibitor Complexes (public)**
- **Docking Programs: QXP, NWU-Dock, ICM, LigandFit**

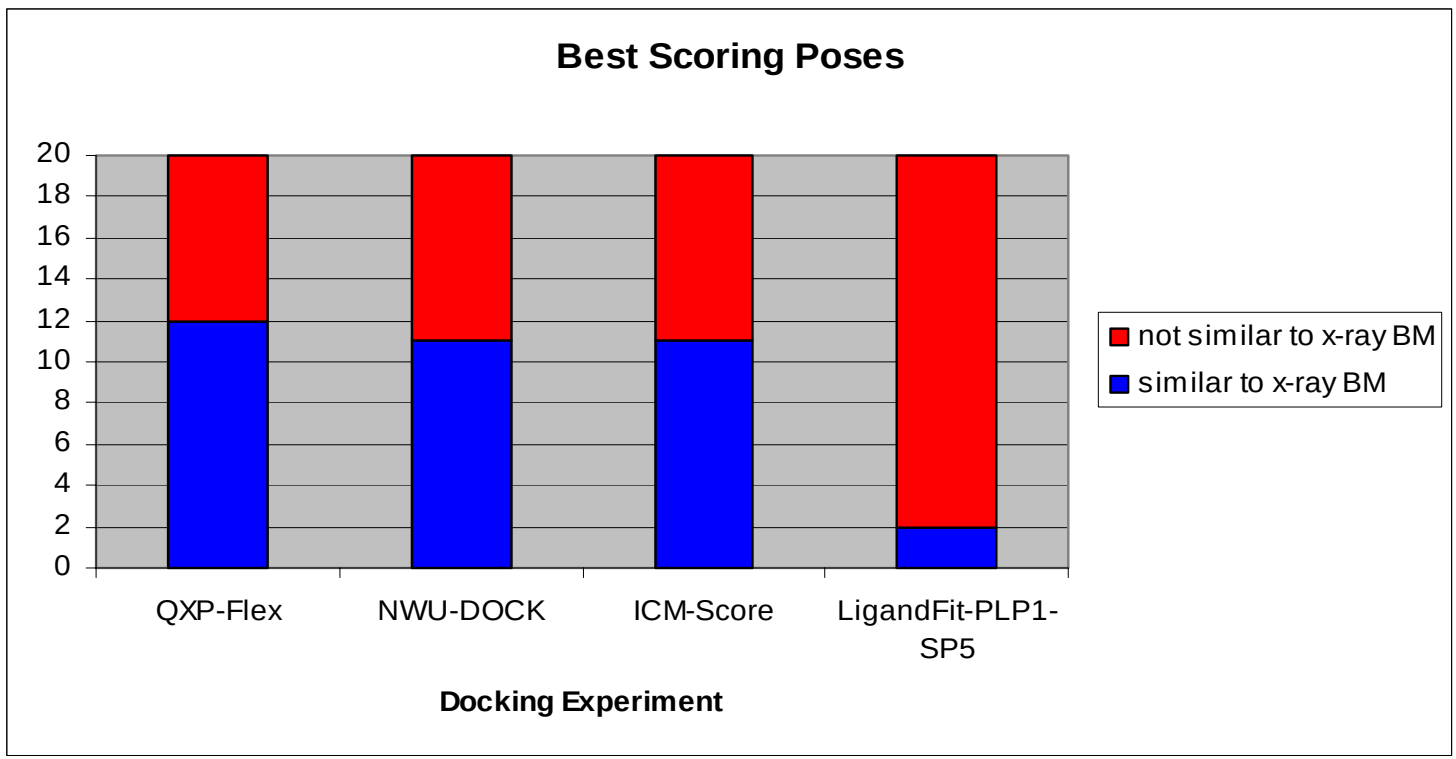




PHARMACIA

Evaluation: Pose Prediction

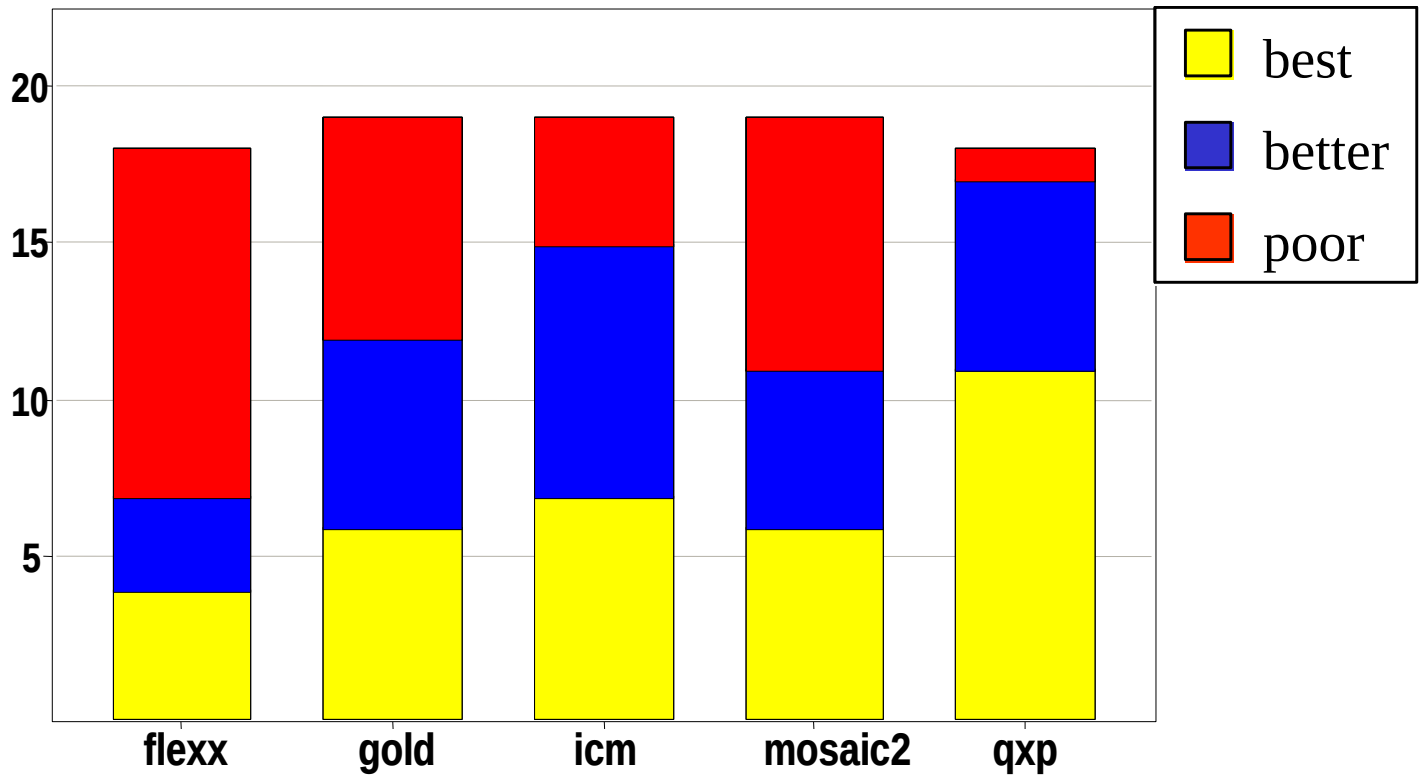
- **Data Set 1: 20 CDK2 – Inhibitor Complexes (public)**
- **Docking Programs: QXP, NWU-Dock, ICM, LigandFit**



PHARMACIA

Evaluation: Pose Prediction

- **Data Set 2: 20 Complexes of Different Proteins (public)**
- **Variety of Docking Programs**



PHARMACIA

Evaluation: Pose Prediction

- **Data Set 3: 654 CDK2 inhibitors (in-house, 18 classes)**
- **Docking Programs: QXP, ICM, LigandFit**

Program	% Scaffold RMS < 1.0 Å
QXP (2001)	60
QXP (2002)	31
QXP (mcdock+)	76 *
ICM	46
LigandFit	
DockScore	2
LigScore1	18
LigScore2	20
PLP1	24
PLP2	26
PMF	10
Ludi	15
Best RMS (20 conf.)	40

PHARMACIA

Evaluation: Affinity Prediction

- **Data set 3: 654 CDK2 inhibitors**
- **IC50 range: 0.004 μM to 10 μM**
- **Docking programs: QXP, ICM, LigandFit**
- **Standard (built-in) scoring functions**
- **Ordered scores: 10 bins à 65 (last 69)**

"actives" < 0.1 μM (219 compounds)

"inactives" ≥ 0.1 μM (435 compounds)

Enrichment:

- **Homogeneous / random: 22 (21.9) actives / bin**

⇒ **22 + 0 actives = 0 % enrichment**

22 + 22 actives = 100 % enrichment (2 –fold)

$$E(\%) = \frac{Na - Na_{random}}{Na_{random}} \cdot 100$$

$$E = \frac{Na}{Na_{random}}$$

		QXP		ICM		LigandFit	
	Bin	+	%E	+	%E	+	%E
10% best scores →	1	+10	46	+11	50	+ 6	27
	2	+ 5	23	+ 2	9	+10	46
	3	+ 6	27	+ 8	36	+ 9	41
	4	+ 1	5	– 5		+ 3	14
	5	– 2		– 4		+ 4	18
	⋮						

PHARMACIA

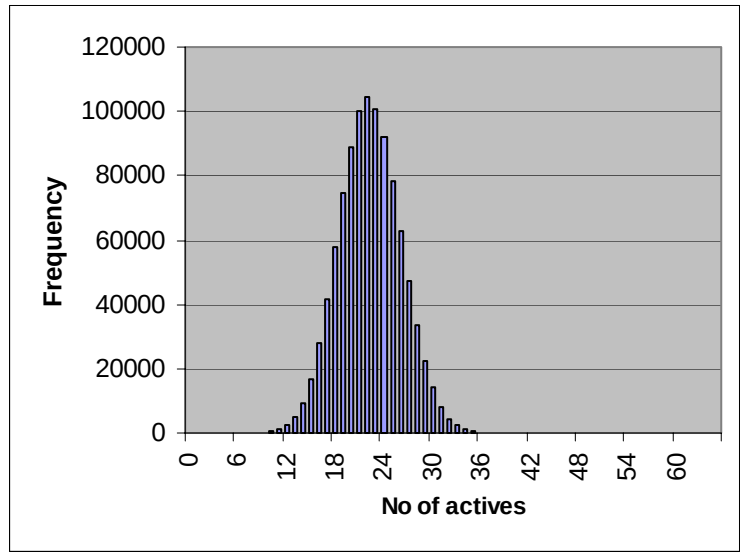
Evaluation: Affinity Prediction

Signal or not?

- **Pick randomly 65 cpds (10^6 x)**
- **Count actives**

→ **Peak at 22**

1 σ = 3.823
2 σ = 7.646
3 σ = 11.469



Bin	QXP +	QXP %E	ICM +	ICM %E	LigandFit +	LigandFit %E
1	+10	46	+11	50	+ 6	27
2	+ 5	23	+ 2	9	+10	46
3	+ 6	27	+ 8	36	+ 9	41
4	+ 1	5	– 5		+ 3	14
5	– 2		– 4		+ 4	18
⋮						

PHARMACIA

Development: Data Sets

- **Data Set 3: 654 CDK2 inhibitors**
- **IC50 range: 0.004 μM to 10 μM**
- **>10 different scaffolds**
- **X-ray structures for at least one of each scaffold**
- **Binding modes for other molecules by analogy**

PHARMACIA

Development: Descriptors

- SASA polar/apolar for protein/ligands
- ΔSASA polar/apolar upon binding
- Conformational *S* protein (Murphy et al. (1992) JMB, 227, 293)
- Conf. *S* ligands: No rotatable bonds (in-house)
- FF-terms from QXP:

  E_{lig}, E_{she}, E_{vdW}, E_{est}, E_{cnt} , E_{vW+}, N_{hph}, N_{hbd}
- Cerius2 compound specific descriptors, e.g.:
  - Total charge
  - "area"
  - MW
  - "density" = MW / Vol.
  - No H-bond donors / acceptors
  - logP (logP, AlogP)
  - MR
  - Dipole
  - Radius of gyration

PHARMACIA

Development: Methodology

- Use X-ray data, generate analogous poses by hand
- Minimize with QXP
- Calculate interaction specific descriptors
  (e.g. QXP FF-terms, ΔSASA)
- Calculate compound specific descriptors
- Descriptor filtering: cross-correlation matrix, PCA, ...
- Derive models using:
  - stepwise LR
  - GFA
  - PLS
  - NN
- Calculate enrichment, q^2, ...

PHARMACIA

Development: Results

Data Set 3: 654 CDK2 inhibitors

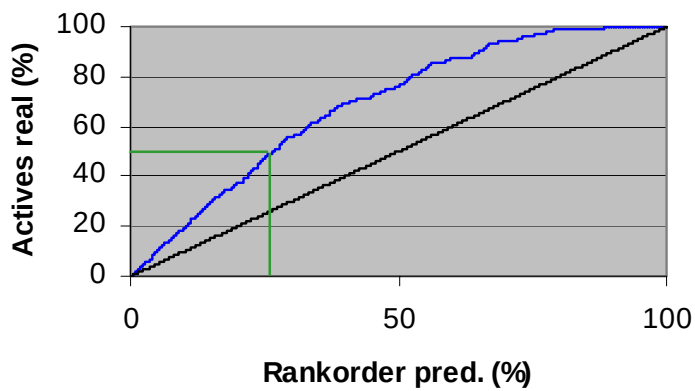
- **IC50 range: 0.004 μM to 10 μM**
- **Enrichment (linear model):**

Bin	+	% E	th. max.
1	+ 21	95	195
2	+ 17	77	
3	+ 21	95	
4	+ 8	36	
5	- 7		
⋮			

- **r^2 = 0.25, q^2 = 0.23 (LOO)**
- **26 % screened → 50 % actives**
- **3 σ = 11.469**

Model / Scoring Function:

- **backwards stepwise LR**
- **Descriptors:**
  - **entropic terms**
  - **SASA related**
  - **compound specific**
- **Total of 8 descriptors**



PHARMACIA

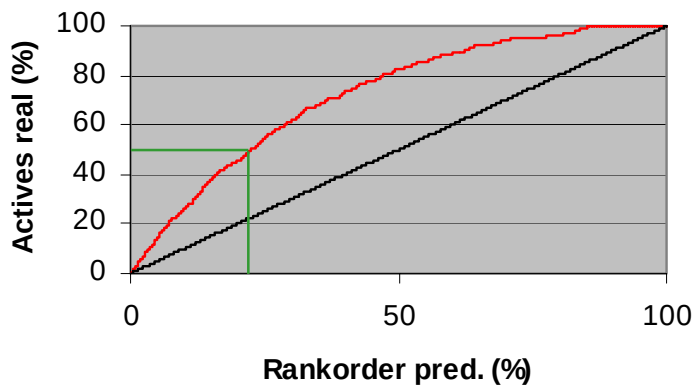
Development: Results

Data Set 3: 654 CDK2 inhibitors
- **IC50 range: 0.004 μM to 10 μM**
- **Enrichment (NN model):**

Bin	+	% E	th. max.
1	+ 36	164	195
2	+ 21	95	
3	+ 14	64	
4	+ 4	18	
5	- 1		
⋮			

- **22 % screened→ 50 % actives**
- **3 σ = 11.469**

Model / Scoring Function:
- **PCA on descriptors**
- **5 PCs retained**
- **back-propagate feed forward**
- **Input – 1 hidden – output NN**



Test:
- **Test set: 65 compounds**
- **20 % screened→ 55 % actives**

PHARMACIA

Conclusions

- Not all docking programs are equal
- Huge variations, depending on:
  - Conditions of docking run
  - Scoring schemes
- Validation!
- Global $\leftrightarrow$ Local
- Correctly Docked $\leftrightarrow$ Incorrectly Docked
- User defined scoring functions

The Heroes

- Fabrizio Giordanetto

- Anna Vulpetti
- Jean-Yves Trosset
- Pieter Stouten

- John Irvin
- Brian Shoichet
- Colin McMartin
- Shashidhar Rao
- Maxim Totrov

- Joe McDonald
- Doug Rohrer