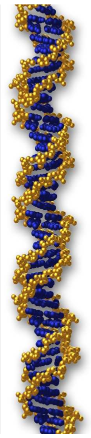


UMDNJ ROBERT WOOD JOHNSON MEDICAL SCHOOL
University of Medicine & Dentistry of New Jersey

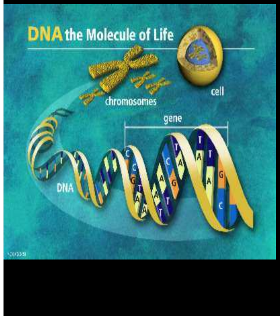
Computational Strategies for Drug Screening & Discovery

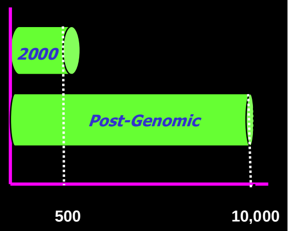
Bill Welsh
UMDNJ-Robert Wood Johnson Medical School
UMDNJ Informatics Institute
welshwj@UMDNJ.edu

Overview of Lecture

- Introduction
  - Opportunities & Challenges in Drug Discovery
  - Computational Approaches
- Receptor-based Computational Methods
- Ligand-based Computational Methods

Fruits of the Genomics Revolution





Number of gene products targeted by drugs

Drug Targets and Mechanisms of Drug Action

- Enzymes – inhibitors (reversible, irreversible)
- Receptors – agonists and antagonists
- Ion Channels – blockers
- Transporters – uptake inhibitors
- DNA – intercalating agents, minor groove binders, antisense drugs

"Needle in a Haystack"

- Estimated 10^{200} compounds could be made
- 28 million compounds currently registered (CAS)
- Drug company biologists screen up to 1 million compounds against target using ultra-high throughput technology
- Chemists select 50-100 compounds for follow-up
- Chemists work on these compounds, developing new, more potent compounds
- Pharmacologists test compounds for pharmacokinetic and toxicological profiles
- 1-2 compounds are selected as potential drugs

How are Most Drugs Discovered ?

By serendipity (propecia, penicillin, etc...)

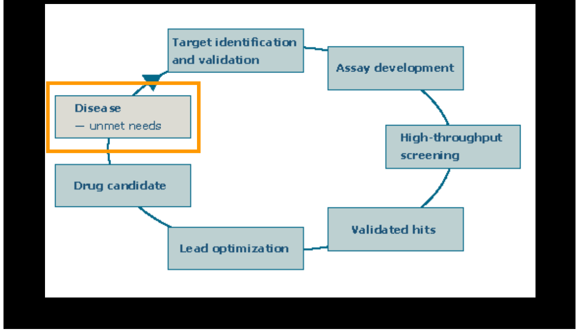
by structure-activity relationships (most)

from natural products (aspirin, digitalis, taxol)

by rational design (since the 80's)

by systematic screening (since the 90's)

Drug Discovery Cycle



Target identification and validation

Assay development

High-throughput screening

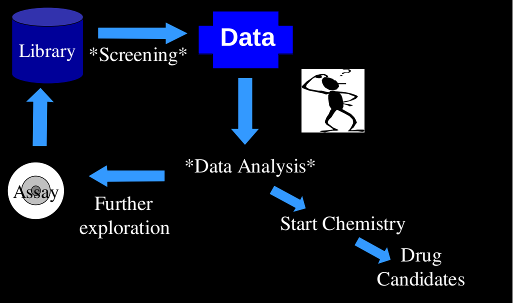
Validated hits

Lead optimization

Drug candidate

Disease
— unmet needs

Early-Stage Drug Discovery Process



Library

Screening

Data

Data Analysis

Assay

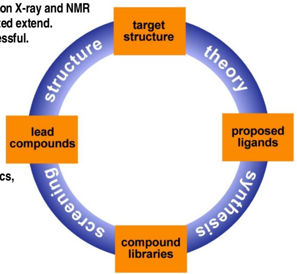
Further exploration

Start Chemistry

Drug Candidates

The Drug Design Cycle

Target identification (genetics – molecular biology - bioinformatics)
De novo design requires detailed knowledge of
Target structure determination X-ray and NMR
Homology modelling to limited extend.
Homology >50% to be successful.



target structure

theory

proposed ligands

synthesis

compound libraries

screening

lead compounds

structure

Design Cycle (left):
Lead development
Computer aided design
and molecular modeling
Fine tuning, pharmacokinetics,
bioavailablity

Substrate characterisation
Computer based screening
Automated site mapping and compound modelling

Lead compound discovery and testing
Biological Assays (high through-put screening)

Target Identification & Lead Discovery

- Identify target (e.g., enzyme, receptor, ion channel, transporter)
- Determine DNA and protein sequence
- Elucidate structure and function of protein
- Prove therapeutic concept in animals ("knock-outs")
- Develop assay for high-throughput molecular screen
- Mass screening and/or directed synthesis program
- Select one or more lead structures

Lead Optimization -> Drug Development

- Determine 3D structure of target receptor complexed with leads
- Molecular modeling- design and refinement of new leads
- Synthesis and biological testing of new leads
- Optimization of selectivity, bioavailability, and pharmacokinetics
- Pharmaceutical formulation
- Preclinical and clinical development
- Drug approval and market introduction

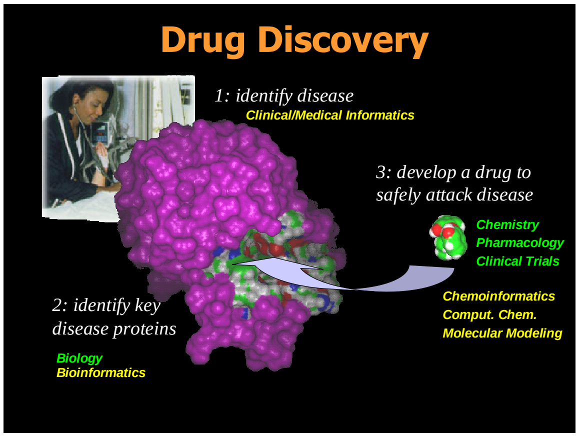
Reasons for Failure in Drug Discovery

- Poor pharmacokinetics (poor ADME profile in humans, metabolite problems)
- Poor clinical activity (doesn't work in humans)
- Unacceptable side effects, toxicity (drug, metabolites, poor selectivity)
- Poor market strategy (won't earn revenues, profit)

ADME: Adsorption, Distribution, Metabolism, Excretion

New Strategies in Drug Design

- Design of inhibitors from structure of substrate (peptidomimetics)
- Computer-aided design of ligands
  - Ø Receptor-based (Structure-based) design
  - Ø Ligand-based design
- Pharmacophore hypotheses
- Combinatorial design of ligands
- Virtual screening for desirable properties: drug-like, bioavailability (e.g., Lipinski's *Rule of Five*)



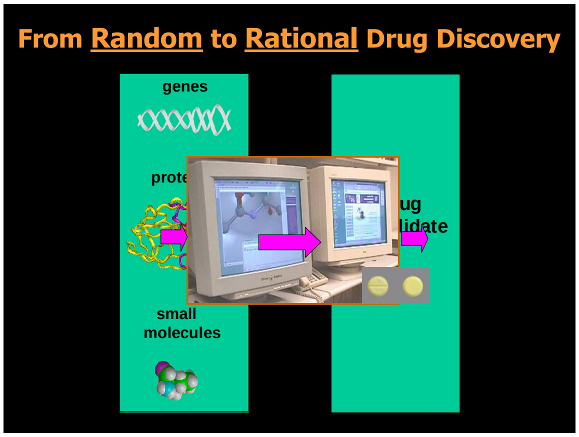
Drug Discovery

1: identify disease
Clinical/Medical Informatics

3: develop a drug to safely attack disease
Chemistry
Pharmacology
Clinical Trials

Chemoinformatics
Comput. Chem.
Molecular Modeling

2: identify key disease proteins
Biology
Bioinformatics



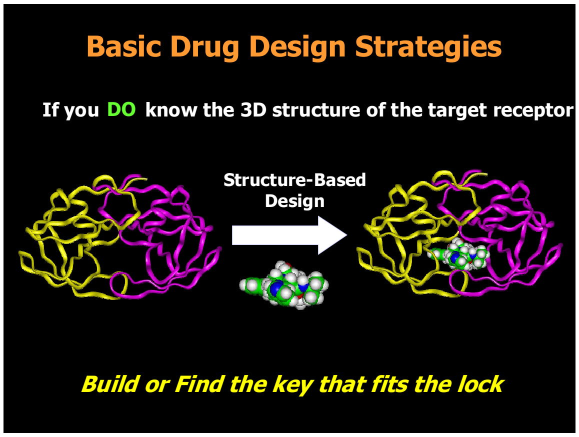
From Random to Rational Drug Discovery

genes

prote

ug
lidate

small molecules

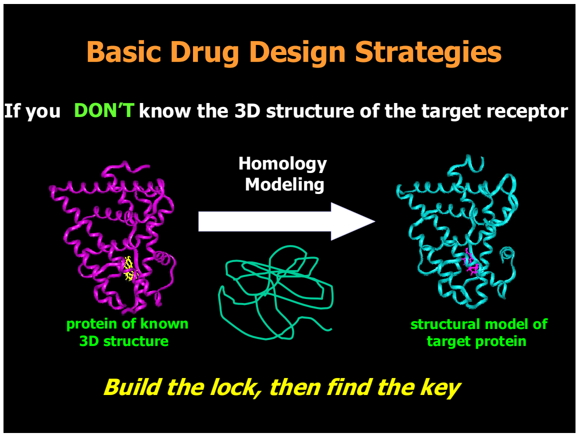


Basic Drug Design Strategies

If you DO know the 3D structure of the target receptor

Structure-Based Design

Build or Find the key that fits the lock



Basic Drug Design Strategies

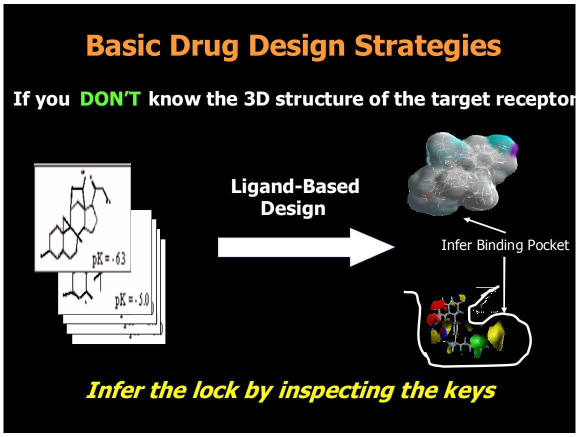
If you DON'T know the 3D structure of the target receptor

Homology Modeling

protein of known 3D structure

structural model of target protein

Build the lock, then find the key



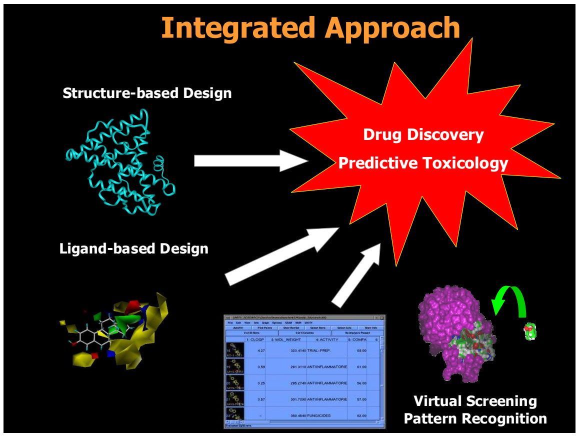
Basic Drug Design Strategies

If you DON'T know the 3D structure of the target receptor

Ligand-Based Design

Infer Binding Pocket

Infer the lock by inspecting the keys



Integrated Approach

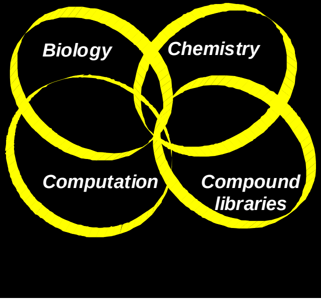
Structure-based Design

Drug Discovery
Predictive Toxicology

Ligand-based Design

Virtual Screening
Pattern Recognition

Research Paradigm: Integration of Technologies



Biology

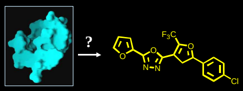
Chemistry

Computation

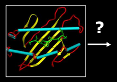
Compound libraries

Receptor-based Methods

Receptor-Based Drug Design and Virtual Screening

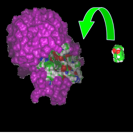


1. de novo design



ΔG_{bind} (IC50, Ki)

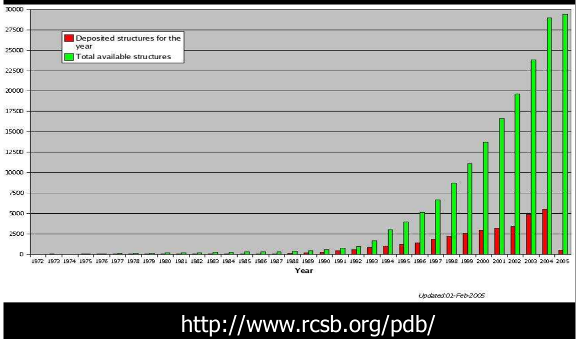
2. Free energy calculations



3. Virtual Screening (docking & scoring)

Protein Data Bank

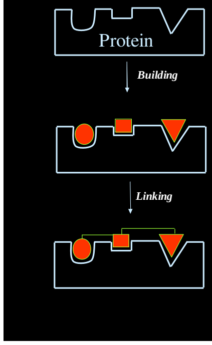
worldwide repository for the processing and distribution of 3-D biological macromolecular structure data



Updated:01-Feb-2005

http://www.rcsb.org/pdb/

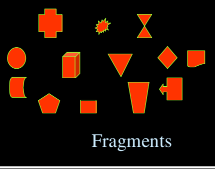
Lead Finding: *de novo* Design



Protein

Building

Linking

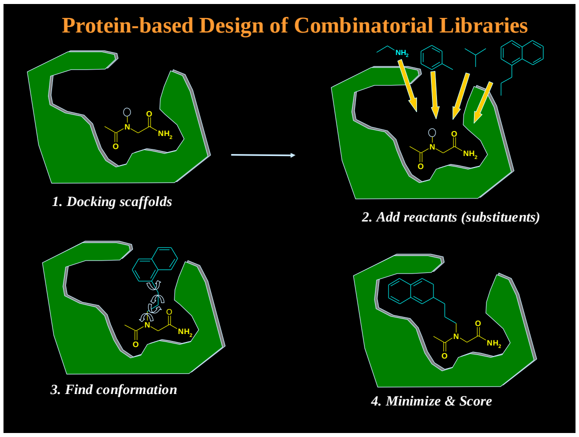


Fragments

PROs:
Find molecules fast

CONs:
Combinatorial Explosion
Many Synthetically Unfeasible Molecules

Protein-based Design of Combinatorial Libraries



1. Docking scaffolds

2. Add reactants (substituents)

3. Find conformation

4. Minimize & Score

Predicting binding affinities (energies)

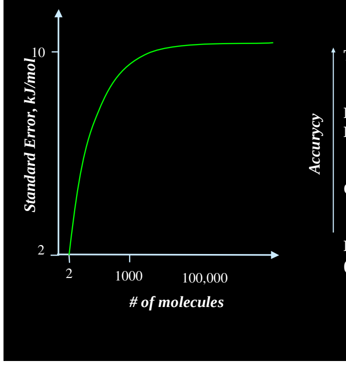
ü 3D database searching
ü docking
ü protein-ligand simulations
ü QSAR studies

Question: Can informatics methods reliably predict reasonable drug candidates?

Which molecules do you propose for chemical synthesis?

Probably the most challenging issue in pharmaceutical computational chemistry

Predicting binding free energies



Thermodynamic methods (<20)

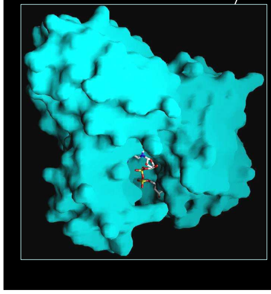
Protein-Ligand Interaction Energies (<50)

QSAR, 3D-QSAR (<200)

Docking & scoring methods (>100,000)

Free Energy Scoring: Predicting binding affinities

predict the binding affinity of a ligand for its host protein from the 3D-structure of the protein-ligand complex ?



?

$$-\Delta G_{binding} = RT.\log K_D$$

Binding free energy

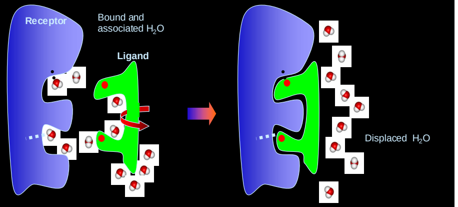
gas constant

Temperature

Equilibrium dissociation constant

$\Delta G_{binding} = f\ (Interactions)$

Key Steps in Ligand-Receptor Binding



Affinity: ΔG = ΔH -TΔS

Upon complex formation:

- water molecules are released
- receptor and ligand loose degrees of freedom
- interactions between ligand and receptor

complication: mutual compensation of enthalpy and entropy

Thermodynamics of Ligand-Receptor Binding

ΔG = ΔH – T(ΔS)

Dictum: ΔG must be negative for spontaneous process

Four Possible Scenarios

	ΔH	ΔS	negative ΔG?	Prognosis
1)	(-)	(+)	always	always spontaneous
2)	(+)	(-)	never	never spontaneous
3)	(+)	(+)	if T(ΔS) > ΔH	favorable as T ↑
4)	(-)	(-)	if T(ΔS) < ΔH	favorable as T ↓

Thermodynamics of Ligand-Receptor Binding

Ligand (aq) + Receptor (aq) à Ligand-Receptor (aq)

ΔG = ΔH – T(ΔS)

Dictum: ΔG must be negative for spontaneous process

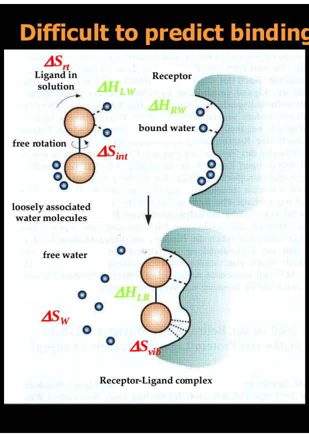
Multi-Step Process	ΔH	ΔS
ligand desolvation	unfavorable	favorable
receptor desolvation	unfavorable	favorable
drug adopts binding conformation	typically unfavorable	unfavorable
receptor adopts binding conformation	unfavorable	unfavorable
ligand binds to receptor	*hopefully favorable*	unfavorable

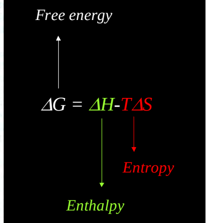
Our Goal: maximize the favorable and minimize the unfavorable steps

But How??

- **Minimize unfavorable desolvation enthalpy**
  - Ø **Ligand can't be too hydrophilic**
  - Ø **Ligand can't have too many H-bonding atoms/groups**
- **Maximize favorable desolvation entropy**
  - Ø **Ligand should fill receptor binding site, to displace all water molecules**
- **Minimize enthalpy cost to adopt "binding conformation"**
  - Ø **Ligand should bind in low-energy conformation**
  - Ø **Shape of ligand should correspond to enzyme's transition-state**
- **Minimize entropy cost to adopt "binding conformation"**
  - Ø **Ligand should be fairly rigid, but not too rigid (most drugs are semi-rigid)**
  - Ø **Shape of ligand should complement shape of receptor's binding site (pre-assembly concept)**
- **Maximize ligand-receptor binding enthalpy**
  - Ø **Hydrophobic surfaces of ligand should touch hydrophobic surfaces of receptor**
  - Ø **Hydrophilic surfaces of ligand should touch hydrophilic surfaces of receptor**
  - Ø **H-bond donors/acceptors of ligand and receptor should be complementary**

Difficult to predict binding affinities




Free energy

$\Delta G = \Delta H - T\Delta S$

Entropy

Enthalpy

Interesting Relationship Between $\Delta G_{binding}$ and $K_{binding}$

$$(\Delta G^o)_{binding} = -RT*\ln K_{binding} = (-\ 1.42\ kcal/mol) * \log K_{binding}$$

at physiological temp (37°C)

Now consider Ligand A and Ligand B binding to the same receptor.

How do various K_B/K_A ratios translate to $\Delta(\Delta G^o)_{binding}$ values?

$$\boxed{\Delta(\Delta G^o)_{binding} = (-1.42\ kcal/mol) * \log(K_B/K_A)_{binding}}$$

(K_B/K_A)	$\Delta(\Delta G^o)_{binding}$
10	1.42 kcal/mol
10^2	2.84
10^3	4.26

Interpretation: $K_{binding}$ is very sensitive to differences in binding energies. Loss of single H-bond (~5 kcal/mol) translates to >1000-fold effect on $K_{binding}$.

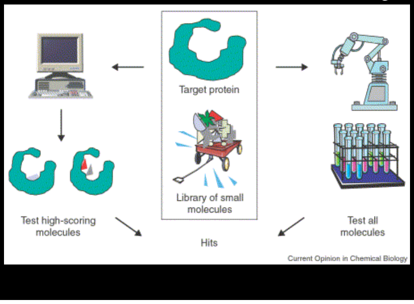
Virtual Screening (VS)

- **Need to prioritize the many molecules that *could* be tested**
- **Increasingly sophisticated level of filtering to maximize the numbers of potential leads**
  - Ø **"Drugability" considerations**
  - Ø **3D substructure searching once possible pharmacophoric patterns have been identified**
  - Ø **Ligand-based VS: Similarity searching (both 2D and 3D) using initial weak leads**
  - Ø **Structure-based VS: Docking once the 3D structure of the biological target is available (*docking & scoring*)**

Virtual Screening ("docking & scoring) Methods

Virtual Screening
(Computational Docking & Scoring)

High Throughput
Screening



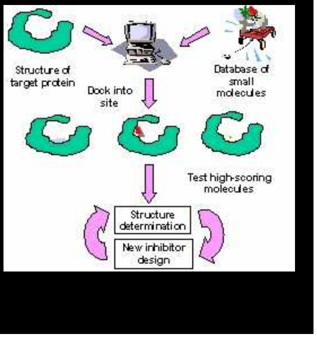
Current Opinion in Chemical Biology

Virtual Screening

Rapid computational mining of small-molecule databases is central to generating new drug leads

The algorithms must be able to handle tens of thousands of molecules

Requires delicate balance between *speed & accuracy*


Structure of target protein
Database of small molecules
Dock into site
Test high-scoring molecules
Structure determination
New inhibitor design

Virtual Screening ("docking & scoring")

(1) Docking - What is it?
- Why is it of interest to us?

(2) Basic principles
- Rigid vs flexible docking

(3) New approach to the problem
- Knowledge-based flexible docking
- Two-step scoring

Docking: what is it?

Given two molecules with 3D conformations in atomic detail

1. Do the molecules bind to each other?
2. If yes, what does the molecule/molecule complex look like (docking problem)?

⬇

Goal: Reproduce the experimental pose of ligand in the binding site

3. How strong is the binding affinity? (scoring problem)

⬇

Drug Discovery

Docking & Scoring Problem

Given the molecular structures of the small-molecule compound and the targeted protein receptor

- Do the molecules bind to each other?
- If yes, what does the ligand-receptor complex look like? (docking problem)

⬇

Reproduce the experimental pose of the ligand within the receptor binding pocket

- How strong is their binding affinity (scoring problem)

⬇

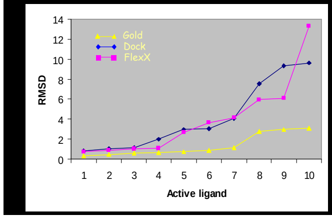
Hit, Lead

Docking Methodology: Evaluation

Evaluation of different methods:

- ***Gold*: genetic algorithm**
- ***FlexX*: incremental docking**
- ***Dock*: fast shape matching**

Docking results for 10 ligands of thymidine kinase were compared with the known complex structures



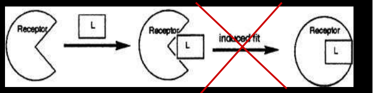
Best results for *Gold*, which finds a solution for all 10 ligands

Docking Programs

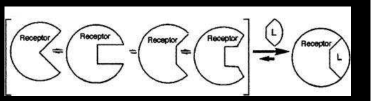
- **DOCK**
  - Developed in Tak Kuntz's group at UCSF - http://www.cmpharm.ucsf.edu/kuntz/dock.html
  - Shape algorithm
  - Recent versions allow for ligand flexibility
- **GOLD**
  - Developed at Sheffield University, distributed by CCDC http://www.ccdc.cam.ac.uk/
  - Uses genetic algorithm
  - Flexible ligand
- **FLEXX**
  - Distributed by Tripos – http://www.tripos.com
  - Flexible ligand
- **FRED**
  - By OpenEye Scientific – http://www.openeye.com
  - Rigid, but able to use multiple, well chosen conformers
  - Very fast
- **AUTODOCK**
  - Scripps Lab http://www.scripps.edu/pub/olson-web/doc/autodock/
  - Uses Genetic Algorithm
- **LIGANDFIT**
  - Accelrys http://www.accelrys.com/cerius2/c2ligandfit.html

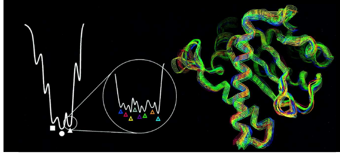
Rigid vs Flexible Docking Methods

Rigid Docking



Flexible Docking





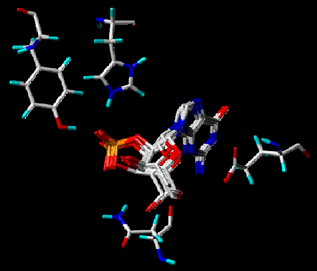
Docking algorithms

- **Require 3D atomic structure for protein, and 3D structure for compound ("ligand")**
- **May require initial rough positioning for the ligand**
- **Will use an optimization method to try and find the best rotation and translation of the ligand in the protein, for optimal binding affinity**

GOLD algorithm

- **Uses a genetic algorithm for optimization**
- **Can output multiple solutions (i.e. output multiple final population members)**
- **Full ligand and partial protein flexibility**
- **Fitness function combination of four elements:**
  - **protein-ligand hydrogen bond energy (*external H-bond*)**
  - **protein-ligand van der Waals (vdw) energy (*external vdw*)**
  - **ligand internal vdw energy (*internal vdw*)**
  - **ligand torsional strain energy (*internal torsion*)**

Sample GOLD output



GMP into RNaseT1

FRED

- Docking is exhaustive
  Unlike most docking programs FRED does not use stochastic sampling to dock ligand. Rather it begins with the set of all possible orientations (to a resolution of one Angstrom, by default) of each conformer near the receptor site and selects the docked position of the ligand from this set.
- Speed
  FRED docks typically docks from 7 to 15 conformers per second on a single PIII-800Mhz CPU.
- Multi-processor
  FRED fully supports PVM (Parallel Virtual Machine) on linux and sgi platforms. This allows FRED to take advantage of multiple processors on muliple machines while still returning a single centralized set of output.
- Multiple scoring fuctions
  FRED currently supports Chemscore, PLP, ScreenScore and Gaussian shape scoring. Scoring with ZAP (a PB solver written by OpenEye Scientific Software) is comming in the near future.
- Alternative docking positions for ligands
  FRED returns alternative docked poses for each ligand as well as the top scoring ligand.
- Graphic preping of receptor site (with VIDA)
  While FRED is fully functional as a command line program, our graphics program VIDA has a FRED wizard which can be used to setup the receptor site for fred.

FlexX

- **Publicly available at http://cartan.gmd.de/flexx/**

Docking & Scoring References

- **Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins, Paul S. Charifson, Joseph J. Corkery, Mark A. Murcko, and W. Patrick Walters, *J. Med. Chem.*** 1999, ***42,* 5100-5109**
- **Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations, Caterina Bissantz, Gerd Folkers, and Didier Rognan, *J. Med. Chem.*** 2000, ***43,* 4759-4767**

Consensus Opinion

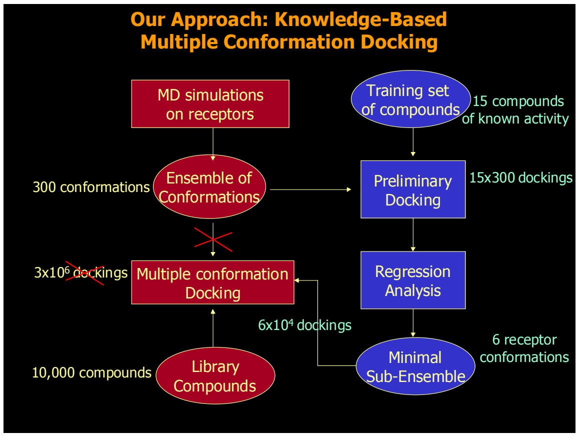
Rigid or Single-Conformation Docking/Scoring Methods (*GOLD*)

Fast, but Inaccurate

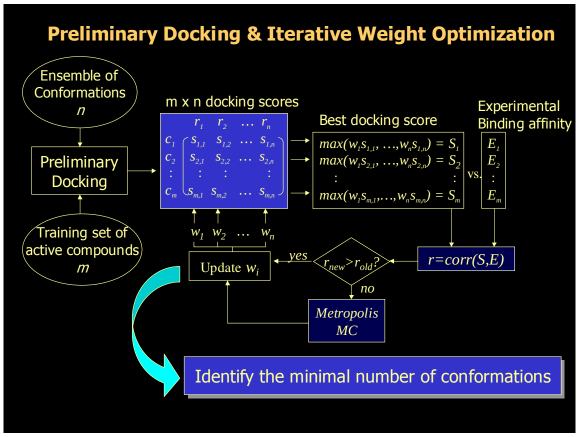
Multiple-Conformation Docking/Scoring Methods

Accurate, but Too Slow

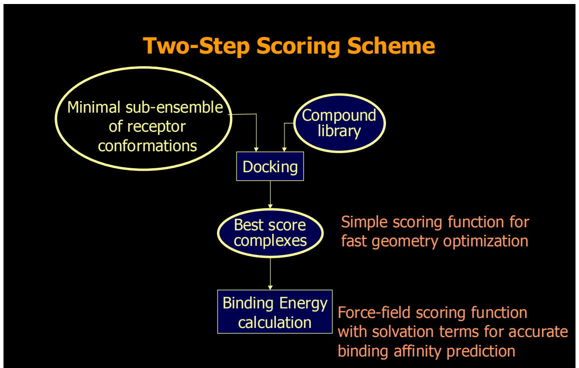
Our Approach: Knowledge-Based Multiple Conformation Docking



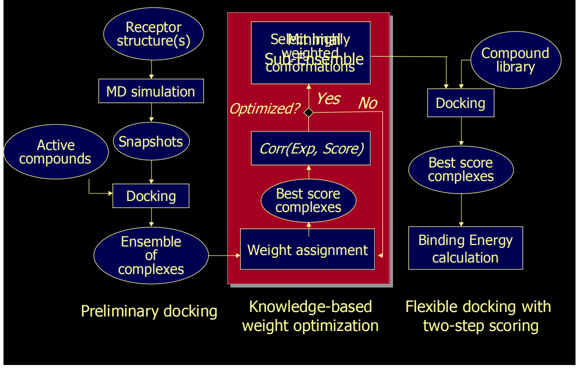
Preliminary Docking & Iterative Weight Optimization


Ensemble of Conformations
n
Preliminary Docking
Training set of active compounds
m
m x n docking scores
r_1 r_2 … r_n
c_1 $s_{1,1}$ $s_{1,2}$ … $s_{1,n}$
c_2 $s_{2,1}$ $s_{2,2}$ … $s_{2,n}$
c_m $s_{m,1}$ $s_{m,2}$ … $s_{m,n}$
Best docking score
$max(w_1 s_{1,1}, \dots, w_n s_{1,n}) = S_1$
$max(w_1 s_{2,1}, \dots, w_n s_{2,n}) = S_2$
$max(w_1 s_{m,1}, \dots, w_n s_{m,n}) = S_m$
vs.
Experimental Binding affinity
E_1 E_2 … E_m
w_1 w_2 … w_n
Update w_i
yes
$r_{new} > r_{old}$?
no
$r = corr(S,E)$
Metropolis MC
Identify the minimal number of conformations

Two-Step Scoring Scheme


Minimal sub-ensemble of receptor conformations
Compound library
Docking
Best score complexes
Simple scoring function for fast geometry optimization
Binding Energy calculation
Force-field scoring function with solvation terms for accurate binding affinity prediction

Knowledge-based flexible docking with two-step scoring



Example: ERα

- Nuclear hormone receptor superfamily

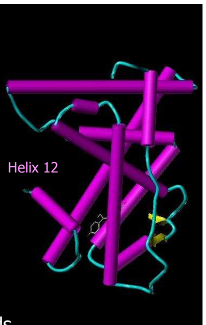
  Associated with numerous diseases:
  breast cancer, osteoporosis,
  endometrial cancer, prostate hypertrophy

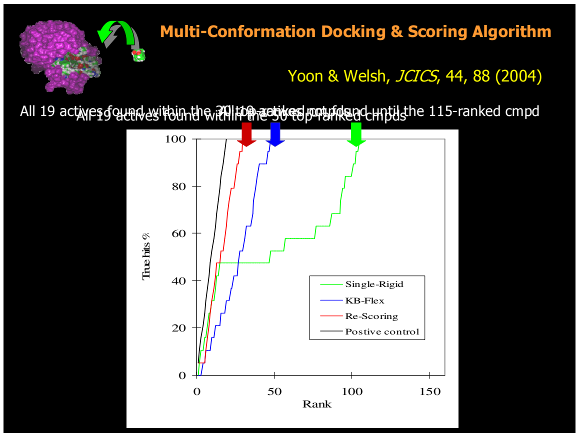
- Natural ERα ligand - estrogen

- Xenoestrogens - phytoestrogens, etc.

- Environmental chemicals - pesticides, PCBs, etc.

Demands fast screening methods





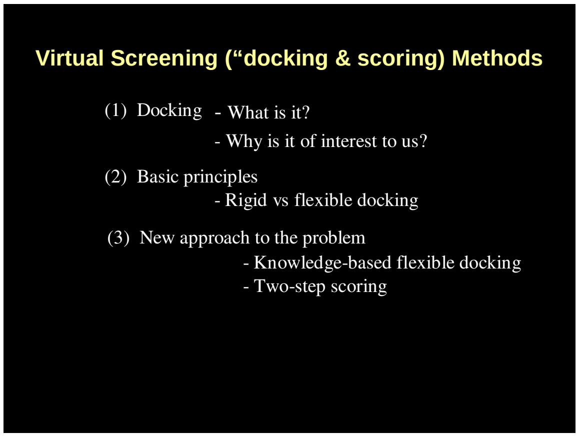
Multi-Conformation Docking & Scoring Algorithm

Yoon & Welsh, *JCICS*, 44, 88 (2004)

All 19 actives found within the 30 top-ranked cmpds

All 19 actives found within the 50 top-ranked cmpds

All 19 actives not found until the 115-ranked cmpd

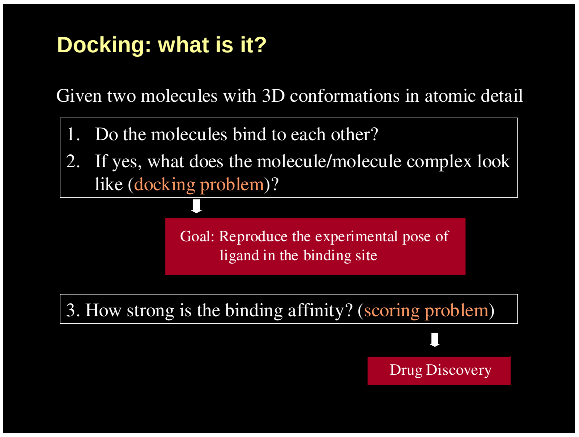


Virtual Screening (“docking & scoring) Methods

(1) Docking - What is it?
- Why is it of interest to us?

(2) Basic principles
- Rigid vs flexible docking

(3) New approach to the problem
- Knowledge-based flexible docking
- Two-step scoring



Docking: what is it?

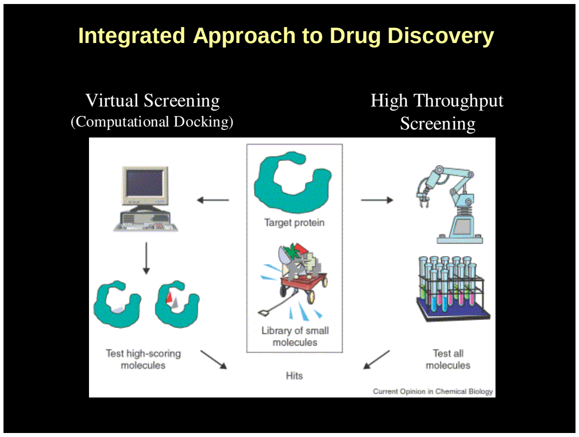
Given two molecules with 3D conformations in atomic detail

1. Do the molecules bind to each other?
2. If yes, what does the molecule/molecule complex look like (docking problem)?

Goal: Reproduce the experimental pose of ligand in the binding site

3. How strong is the binding affinity? (scoring problem)

Drug Discovery

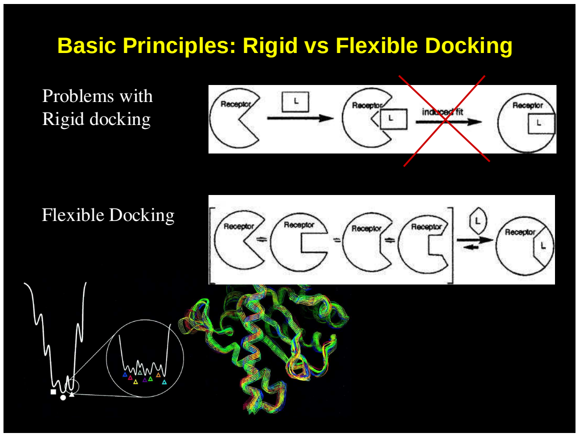


Integrated Approach to Drug Discovery

Virtual Screening
(Computational Docking)

High Throughput Screening

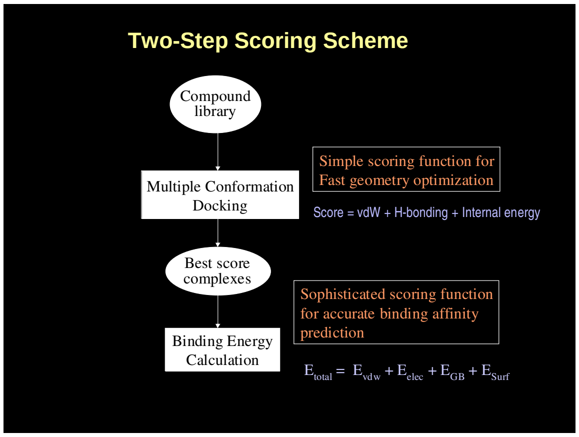
Target protein
Library of small molecules
Test high-scoring molecules
Hits
Test all molecules
Current Opinion in Chemical Biology



Basic Principles: Rigid vs Flexible Docking

Problems with Rigid docking

Flexible Docking



Two-Step Scoring Scheme

Compound library

Multiple Conformation Docking

Simple scoring function for Fast geometry optimization

Score = vdW + H-bonding + Internal energy

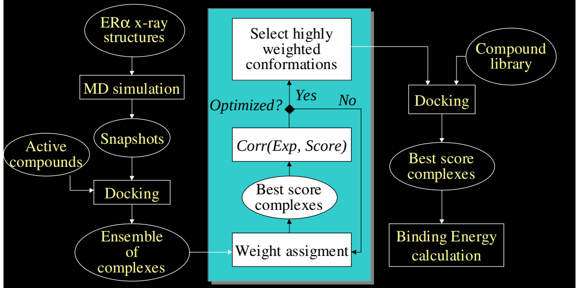
Best score complexes

Binding Energy Calculation

Sophisticated scoring function for accurate binding affinity prediction

$E_{total} = E_{vdw} + E_{elec} + E_{GB} + E_{Surf}$

Knowledge-Based Flexible Docking With Two-Step Scoring

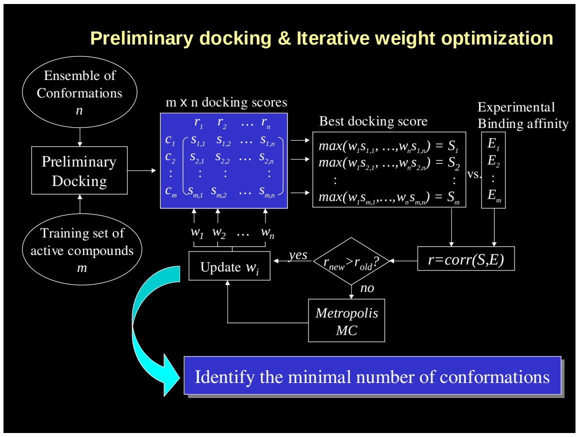


A. Preliminary multiple conformation docking

B. Iterative weight optimization

C. Flexible docking with two-step scoring

Preliminary docking & Iterative weight optimization


Ensemble of Conformations n
Preliminary Docking
Training set of active compounds m
m x n docking scores
r1 r2 … rn
c1 s1,1 s1,2 … s1,n
c2 s2,1 s2,2 … s2,n
cm sm,1 sm,2 … sm,n
Best docking score
max(w1s1,1, …,wns1,n) = S1
max(w1s2,1, …,wns2,n) = S2
max(w1sm,1,…,wnsm,n) = Sm
vs.
Experimental Binding affinity
E1
E2
Em
w1 w2 … wn
Update wi
yes
rnew>rold?
no
r=corr(S,E)
Metropolis MC
Identify the minimal number of conformations

Population Weight Optimization by Metropolis MC

Initialize Population weight

$$\sum_{i=1}^{N} w_i = 1$$

Population Weight Update

$$w_{new} = w_{old} + (2\xi - 1) \times \delta x_{\max}$$

Pearson's Correlation Coefficient

$$r = \frac{\sum \left(Ex_j - \langle Ex \rangle\right)\left(S_{\max,\, j} - \langle S_{\max} \rangle\right)}{\sqrt{\sum \left(Ex_j - \langle Ex \rangle\right)^2 \sum \left(S_{\max,\, j} - \langle S_{\max} \rangle\right)^2}}$$

Metropolis criteria

$$rand(0,1) \le \exp(-\Delta r / C)$$

Weighted Docking Score

$$S_{i,j} = w_i s_{ij}$$

$$S_{j,\max} = \max(S_{1,j}, S_{2,j}, \ldots, S_{N,j})$$

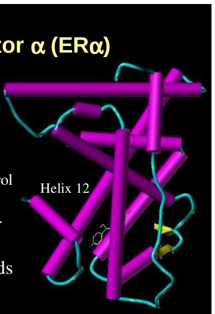
Test system: Estrogen Receptor α (ERα)

- Nuclear hormone receptor superfamily

- Associated with numerous diseases
  ex) breast cancer, osteoporosis,
  endometrial cancer, prostate hypertrophy

- Natural ERα ligands – estrogen, diethylstilbestrol
  Xenoestrogens – phytoestrogens, …
  Environmental chemicals – pesticides, PCBs, …

↳ Demanding fast screening methods

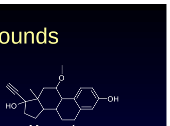


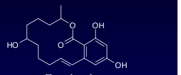
ERα x-ray structures used in MD

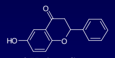
	PDB ID	Ligand	Resolution (Å)
With agonists	3ERD	Diethylstilbestrol	2.03
	1QKU	Estradiol	3.20
	1L2I	Diethyl Tetrahydrochrysene Diol	1.95

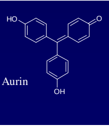
Structural diversity of 15 active compounds

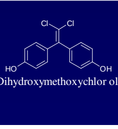
Class	Name	Log(RBA)
Steroids	3-methylestriol	-1.65
	Ethynylestradiol	2.28
	Moxestrol	1.14
	5α-Antrostane-3β, 17β-diol	-0.92
Alkylphenolic	nonylphenol	-1.53
	4-Ethylphenol	-4.17
Diphenyl derivatives	Bisphenol A	-2.11
	3-Phenylphenol	-3.44
Organochlorines	2',3',4',5'-Tetrachloro-4-biphenylol	-0.64
	Dihydroxymethoxychlor olefin	0.42
Alkyl hydroxy benzoate	Ethyl 4-hydroxybenzoate	-3.22
Benzophenone	4,4'-Dihydroxybenzophenone	-2.46
Others	Aurin	-1.49
	Zearalenol	1.63
	6-Hydroxyflavanone	-3.05

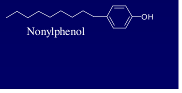

Moxestrol


Zearalenol

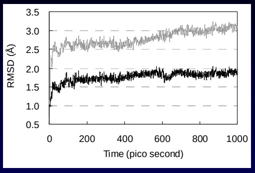

6-Hydroxyflavanone


Aurin


Dihydroxymethoxychlor olefin


Nonylphenol

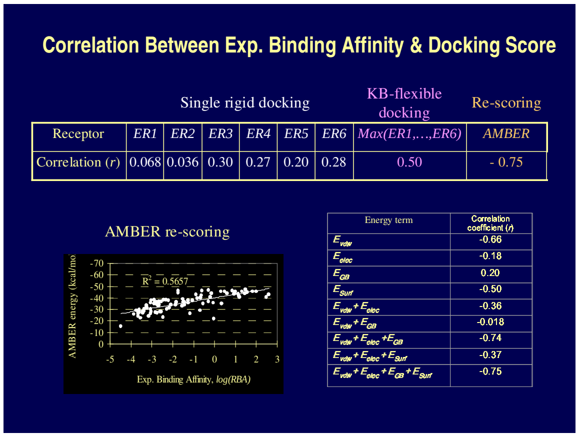
Generation of Multiple Conformations



MD trajectory of ERα (PDB ID: 3ERD)

Identification of Minimal Subset of ERα Conformations

Selected conformations	*ER1*	*ER2*	*ER3*	*ER4*	*ER5*	*ER6*
Sampling position	3erd 490 ps	3erd 840 ps	1l2i 410 ps	1l2i 830 ps	1qku 700 ps	1qku 940 ps
Correlation (*r*)	*max(ER1,..,ER6) = 0.94*					

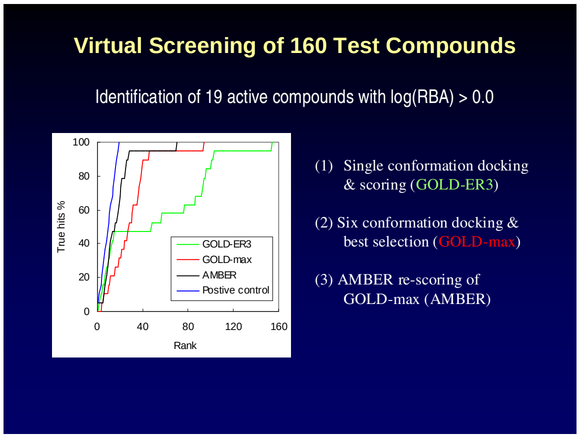


Correlation Between Exp. Binding Affinity & Docking Score

	Single rigid docking						KB-flexible docking	Re-scoring
Receptor	*ER1*	*ER2*	*ER3*	*ER4*	*ER5*	*ER6*	*Max(ER1,…,ER6)*	*AMBER*
Correlation (*r*)	0.068	0.036	0.30	0.27	0.20	0.28	0.50	- 0.75

AMBER re-scoring

Energy term	Correlation coefficient (*r*)
E_{vdw}	-0.66
E_{elec}	-0.18
E_{GB}	0.20
E_{Surf}	-0.50
$E_{vdw} + E_{elec}$	-0.36
$E_{vdw} + E_{GB}$	-0.018
$E_{vdw} + E_{elec} + E_{GB}$	-0.74
$E_{vdw} + E_{elec} + E_{Surf}$	-0.37
$E_{vdw} + E_{elec} + E_{GB} + E_{Surf}$	-0.75



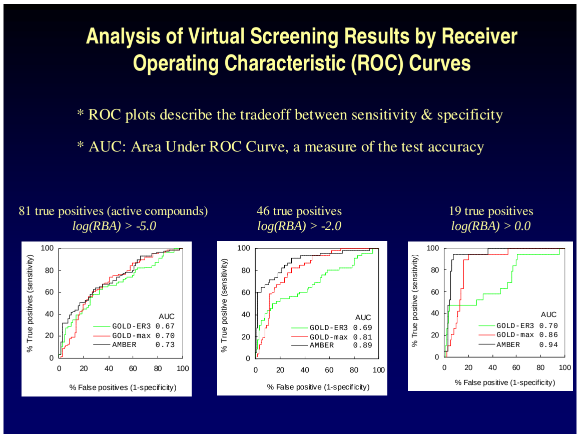
Virtual Screening of 160 Test Compounds

Identification of 19 active compounds with log(RBA) > 0.0

(1) Single conformation docking & scoring (GOLD-ER3)

(2) Six conformation docking & best selection (GOLD-max)

(3) AMBER re-scoring of GOLD-max (AMBER)



Analysis of Virtual Screening Results by Receiver Operating Characteristic (ROC) Curves

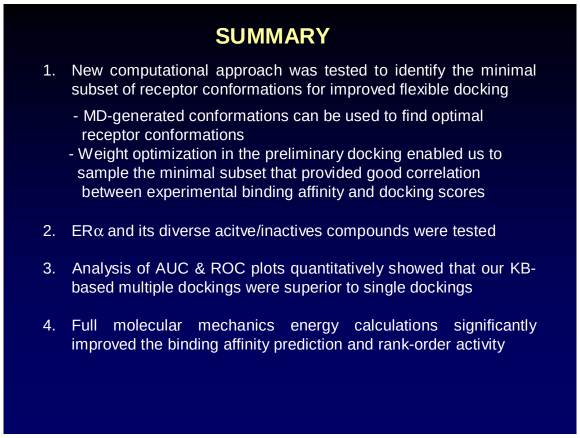
* ROC plots describe the tradeoff between sensitivity & specificity

* AUC: Area Under ROC Curve, a measure of the test accuracy

81 true positives (active compounds)
log(RBA) > -5.0

46 true positives
log(RBA) > -2.0

19 true positives
log(RBA) > 0.0

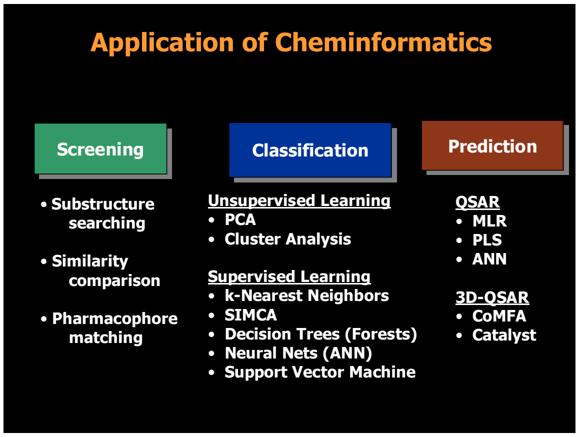


SUMMARY

1. New computational approach was tested to identify the minimal subset of receptor conformations for improved flexible docking
   - MD-generated conformations can be used to find optimal receptor conformations
   - Weight optimization in the preliminary docking enabled us to sample the minimal subset that provided good correlation between experimental binding affinity and docking scores
2. ERα and its diverse acitve/inactives compounds were tested
3. Analysis of AUC & ROC plots quantitatively showed that our KB-based multiple dockings were superior to single dockings
4. Full molecular mechanics energy calculations significantly improved the binding affinity prediction and rank-order activity

Ligand-based Methods

Cheminformatics



Application of Cheminformatics

Screening

- **Substructure searching**
- **Similarity comparison**
- **Pharmacophore matching**

Classification

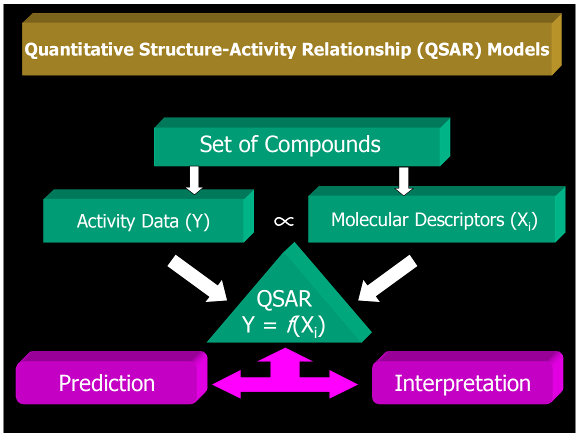
Unsupervised Learning
- **PCA**
- **Cluster Analysis**

Supervised Learning
- **k-Nearest Neighbors**
- **SIMCA**
- **Decision Trees (Forests)**
- **Neural Nets (ANN)**
- **Support Vector Machine**

Prediction

QSAR
- **MLR**
- **PLS**
- **ANN**

3D-QSAR
- **CoMFA**
- **Catalyst**



Quantitative Structure-Activity Relationship (QSAR) Models

Set of Compounds

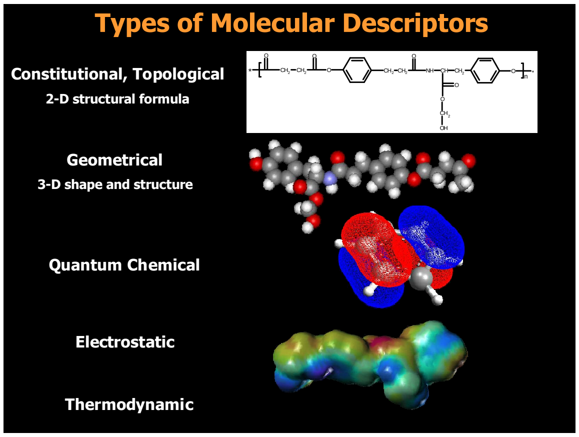
Activity Data (Y) ∝ Molecular Descriptors (X_i)

QSAR

$Y = f(X_i)$

Prediction

Interpretation



Types of Molecular Descriptors

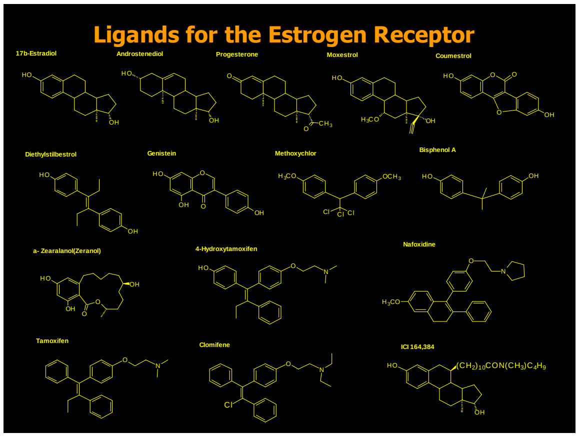
Constitutional, Topological
2-D structural formula

Geometrical
3-D shape and structure

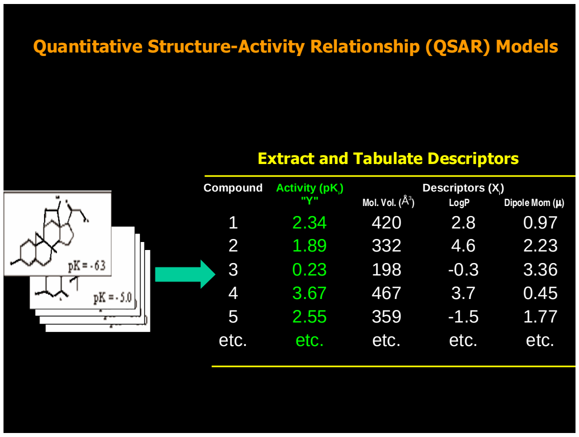
Quantum Chemical

Electrostatic

Thermodynamic



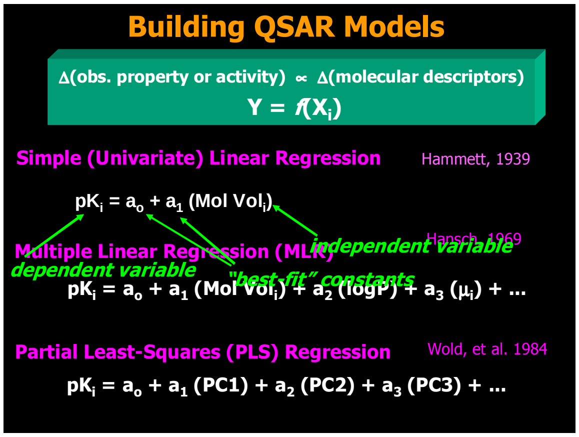
Ligands for the Estrogen Receptor



Quantitative Structure-Activity Relationship (QSAR) Models

Extract and Tabulate Descriptors

Compound	Activity (pK_i) "Y"	Descriptors (X_i) Mol. Vol. ($Å^3$)	LogP	Dipole Mom (µ)
1	2.34	420	2.8	0.97
2	1.89	332	4.6	2.23
3	0.23	198	-0.3	3.36
4	3.67	467	3.7	0.45
5	2.55	359	-1.5	1.77
etc.	etc.	etc.	etc.	etc.



Building QSAR Models

Δ(obs. property or activity) ∝ Δ(molecular descriptors)

$Y = f(X_i)$

Simple (Univariate) Linear Regression Hammett, 1939

$$pK_i = a_o + a_1\,(\text{Mol Vol}_i)$$

dependent variable

"best-fit" constants

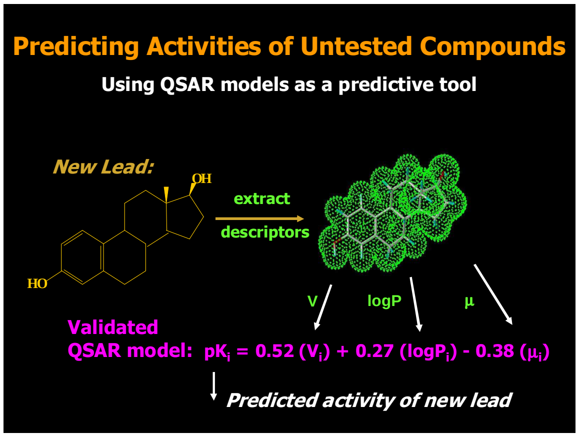
independent variable

Multiple Linear Regression (MLR) Hansch, 1969

$$pK_i = a_o + a_1\,(\text{Mol Vol}_i) + a_2\,(\text{logP}) + a_3\,(\mu_i) + \ldots$$

Partial Least-Squares (PLS) Regression Wold, et al. 1984

$$pK_i = a_o + a_1\,(PC1) + a_2\,(PC2) + a_3\,(PC3) + \ldots$$



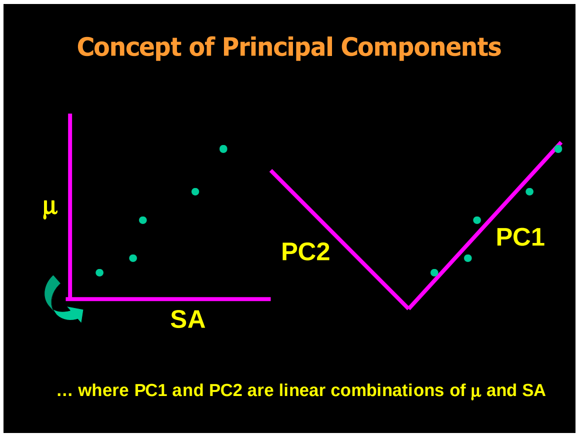
Predicting Activities of Untested Compounds

Using QSAR models as a predictive tool

New Lead:

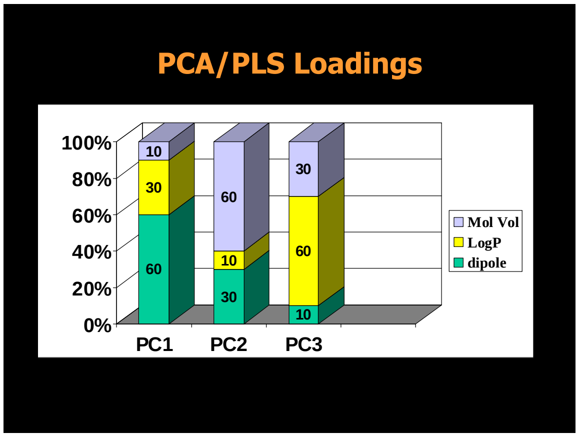
Validated QSAR model: $pK_i = 0.52\,(V_i) + 0.27\,(\text{logP}_i) - 0.38\,(\mu_i)$

Predicted activity of new lead

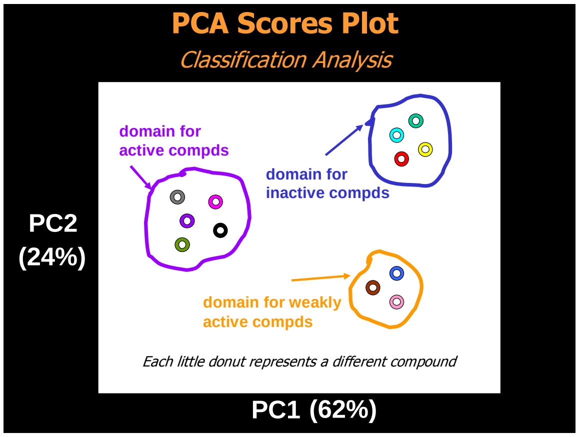


Concept of Principal Components

… where PC1 and PC2 are linear combinations of μ and SA



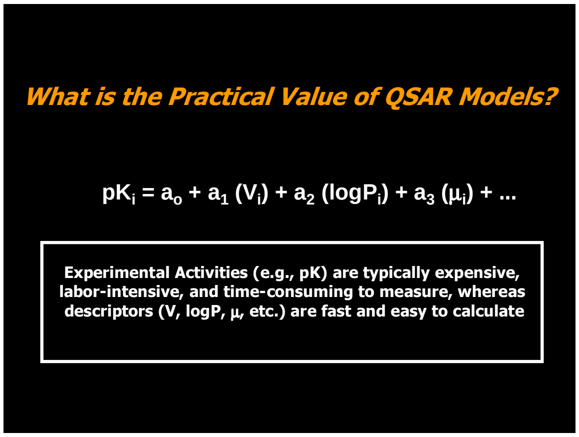
PCA/PLS Loadings



PCA Scores Plot

Classification Analysis

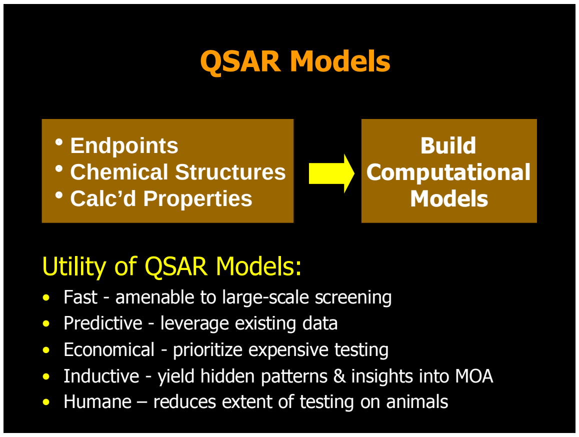
Each little donut represents a different compound



What is the Practical Value of QSAR Models?

$$pK_i = a_o + a_1 (V_i) + a_2 (\log P_i) + a_3 (\mu_i) + \ldots$$

Experimental Activities (e.g., pK) are typically expensive, labor-intensive, and time-consuming to measure, whereas descriptors (V, logP, μ, etc.) are fast and easy to calculate



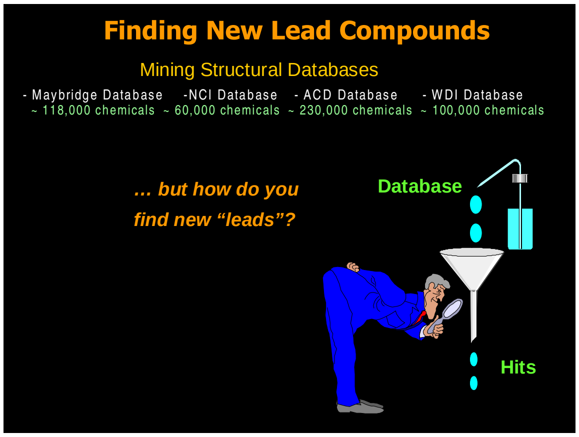
QSAR Models

- Endpoints
- Chemical Structures
- Calc'd Properties

→ Build Computational Models

Utility of QSAR Models:

- Fast - amenable to large-scale screening
- Predictive - leverage existing data
- Economical - prioritize expensive testing
- Inductive - yield hidden patterns & insights into MOA
- Humane – reduces extent of testing on animals



Finding New Lead Compounds

Mining Structural Databases

- Maybridge Database	-NCI Database	- ACD Database	- WDI Database
~ 118,000 chemicals	~ 60,000 chemicals	~ 230,000 chemicals	~ 100,000 chemicals

… but how do you find new "leads"?

DRUG-LIKE BEHAVIOR

The Lipinski "Rule of Five" [1]

Ø Molecular Weight ≤ 500 (opt = ~350)
Ø # Hydrogen Bond Acceptors ≤10 (opt = ~5)
Ø # Hydrogen Bond Donors ≤ 5 (opt = ~2)
Ø -2 < cLog P < 5 (opt = ~3.0)
Ø # Rotatable Bonds ≤ 5

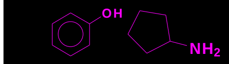
1: C. Lipinski et al, Adv. Drug. Del. Rev, 23, 3-25 (1997)

Requirements for Orally Active Drugs - Pharmacokinetics -

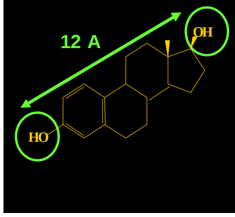
- Aqueous solubility
- Membrane passive permeability
- Cytochrome P450 activities
- Plasma protein binding
- Efflux pumping and active transport

Ligand-Based VS of Small-Molecule Structural Databases

1. (Sub)structure Searching



2. Pharmacophore Matching



3. Property Search:
   Similar Molecular Features
   (e.g., Vol, SA, μ, … hundreds more)

4. Filtering: Lipinski's "Rule of 5"

   Oral Drug-like molecules share the following characteristics:

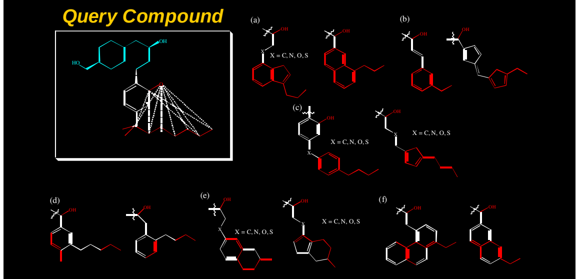
   1) Maximum of 5 H-bond donors
   2) Maximum of 10 H-bond acceptors
   3) Molecular Weight < 500
   4) LogP < 5

   C. A. Lipinski, et al., Adv Drug Delivery Reviews, 23, 3 (1997)

5. Apply QSAR Models

6. Molecular (Dis)Similarity

Mining Structural Databases

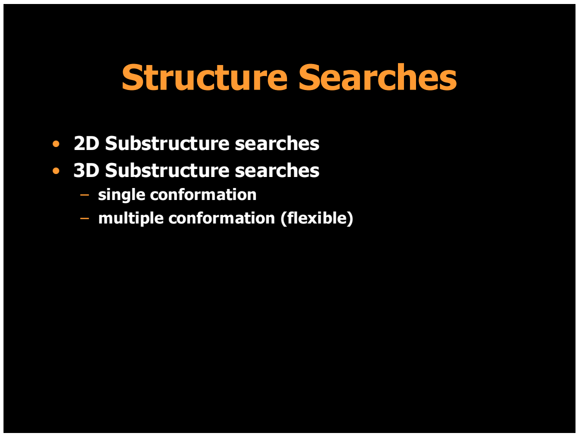


Molecular Similarity

§ Widely used all over drug discovery process
§ Sample applications:
  Ø Assessing diversity of a chemical dataset
  Ø Picking representative dataset from compound library
  Ø Given a compound and a compound library, identifying subset
    of similar compounds
  Ø Organizing library compounds for screening and analysis
    - Major step: sort into chemical families based on molecular similarity

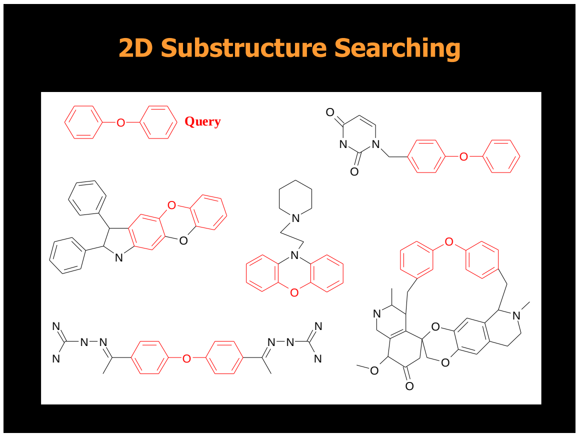
Technology Employed

§ Compound representation methods
  Ø Fingerprints/bit vectors, graph-based, …
  Ø 2D-keys vs 3D-keys, fragment vs distance based, …
§ Similarity and distance measures
  Ø Tanimoto, Euclidean, …, graph-based, …
§ Clustering methods
§ Classification methods
§ Substructure searching/(sub)graph matching

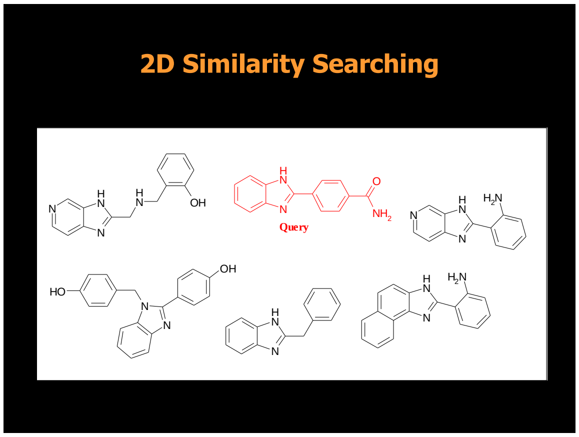


Structure Searches

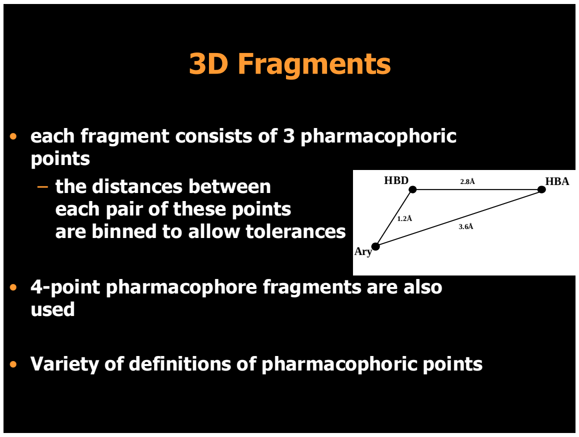
- 2D Substructure searches
- 3D Substructure searches
  - single conformation
  - multiple conformation (flexible)



2D Substructure Searching

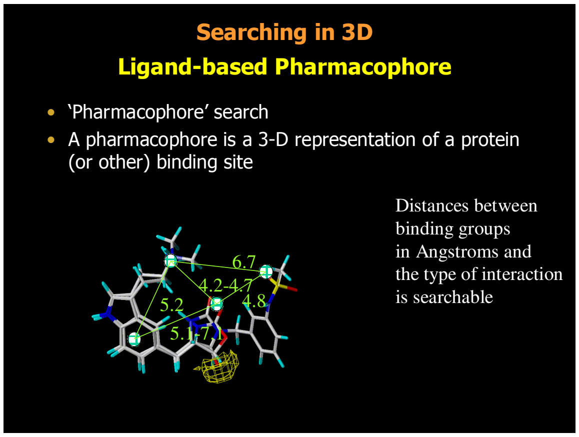


2D Similarity Searching



3D Fragments

- each fragment consists of 3 pharmacophoric points
  - the distances between each pair of these points are binned to allow tolerances
- 4-point pharmacophore fragments are also used
- Variety of definitions of pharmacophoric points

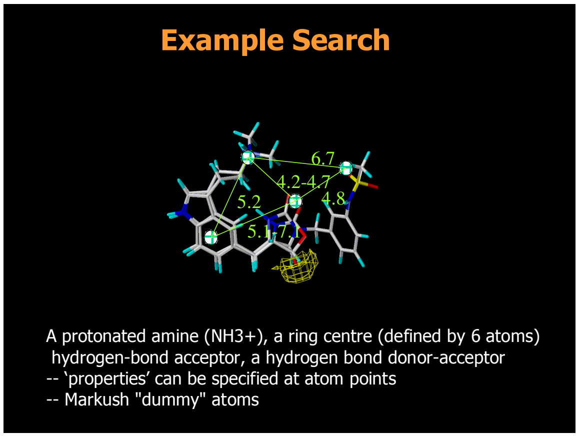


Searching in 3D

Ligand-based Pharmacophore

- 'Pharmacophore' search
- A pharmacophore is a 3-D representation of a protein (or other) binding site

Distances between binding groups in Angstroms and the type of interaction is searchable

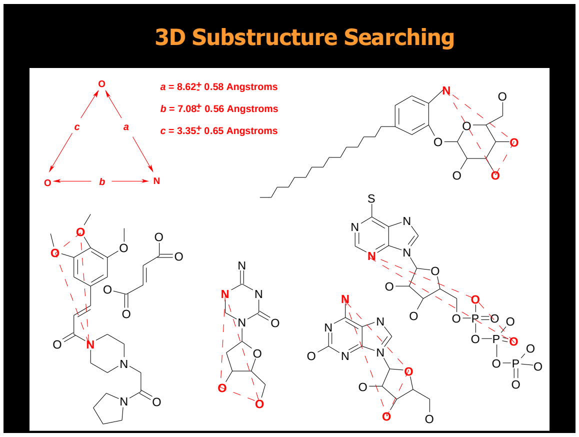


Example Search

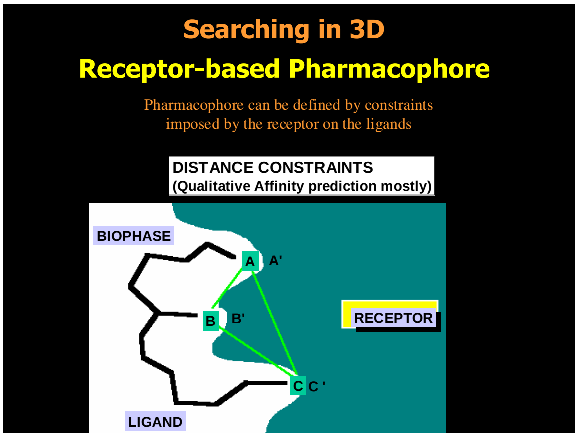
A protonated amine (NH3+), a ring centre (defined by 6 atoms) hydrogen-bond acceptor, a hydrogen bond donor-acceptor

-- 'properties' can be specified at atom points

-- Markush "dummy" atoms



3D Substructure Searching

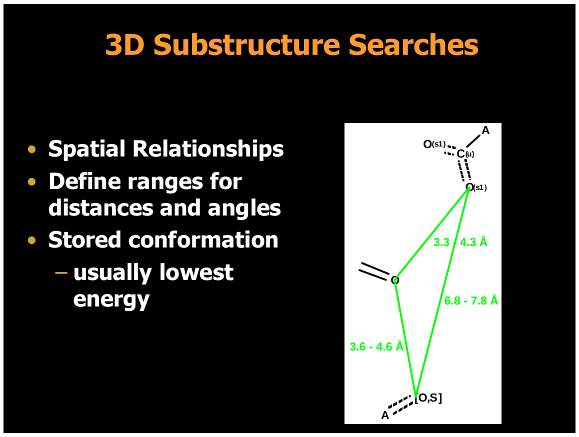


Searching in 3D

Receptor-based Pharmacophore

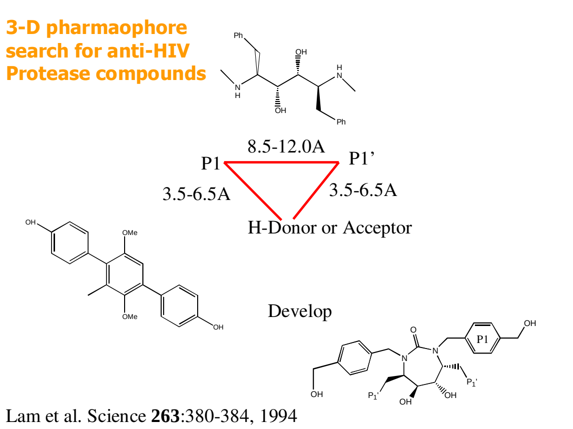
Pharmacophore can be defined by constraints imposed by the receptor on the ligands

DISTANCE CONSTRAINTS
(Qualitative Affinity prediction mostly)



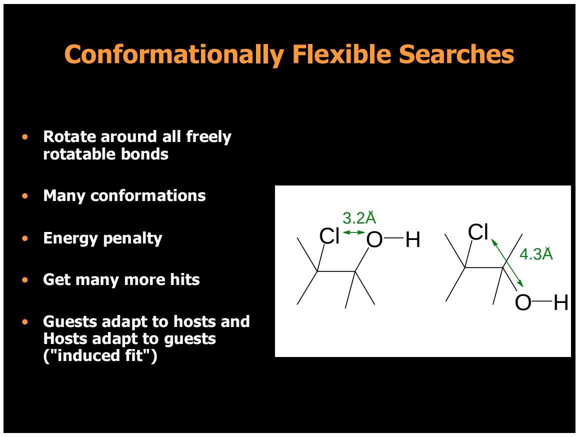
3D Substructure Searches

- Spatial Relationships
- Define ranges for distances and angles
- Stored conformation
  - usually lowest energy



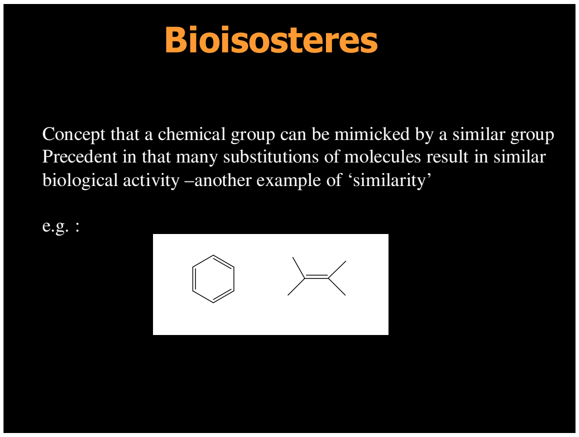
3-D pharmaophore search for anti-HIV Protease compounds

Lam et al. Science **263**:380-384, 1994



Conformationally Flexible Searches

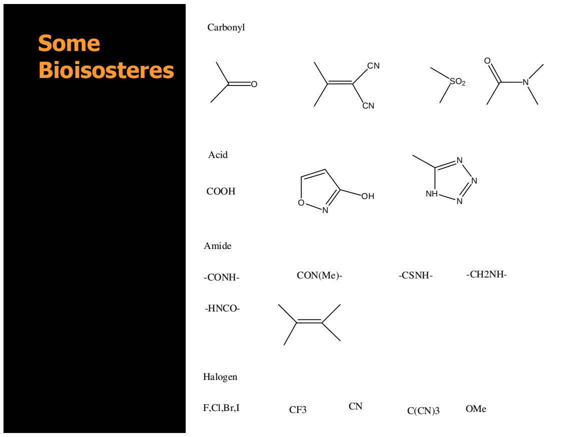
- Rotate around all freely rotatable bonds
- Many conformations
- Energy penalty
- Get many more hits
- Guests adapt to hosts and Hosts adapt to guests ("induced fit")



Bioisosteres

Concept that a chemical group can be mimicked by a similar group
Precedent in that many substitutions of molecules result in similar biological activity –another example of ‘similarity’

e.g. :



Some Bioisosteres

Carbonyl

Acid

COOH

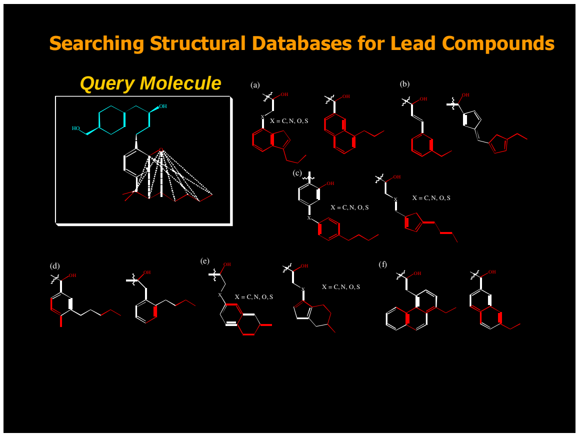
Amide

-CONH- CON(Me)- -CSNH- -CH2NH-

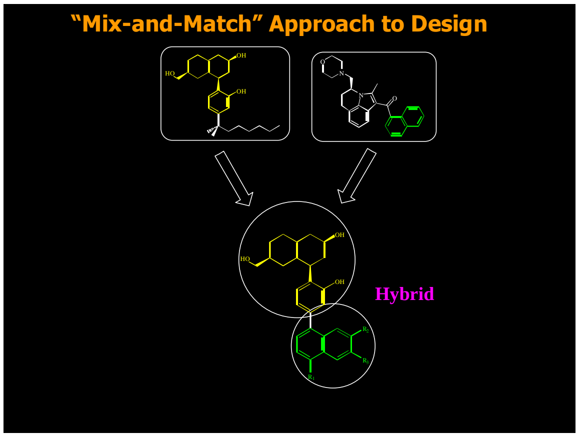
-HNCO-

Halogen

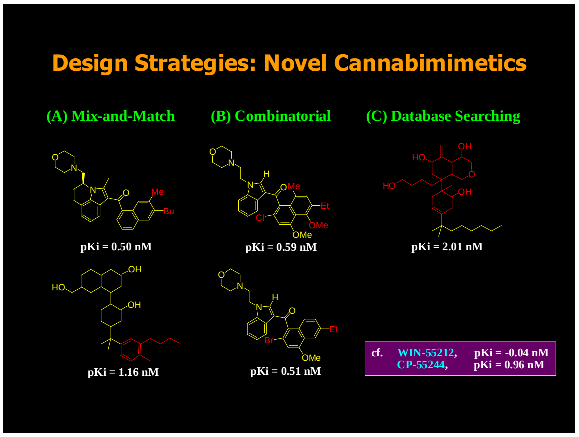
F,Cl,Br,I CF3 CN C(CN)3 OMe



Searching Structural Databases for Lead Compounds



"Mix-and-Match" Approach to Design



Design Strategies: Novel Cannabimimetics

(A) Mix-and-Match (B) Combinatorial (C) Database Searching

pKi = 0.50 nM

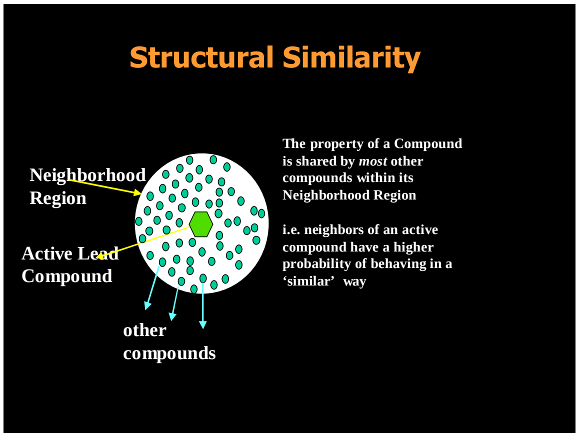
pKi = 0.59 nM

pKi = 2.01 nM

pKi = 1.16 nM

pKi = 0.51 nM

cf. WIN-55212, pKi = -0.04 nM
CP-55244, pKi = 0.96 nM



Structural Similarity

The property of a Compound is shared by *most* other compounds within its Neighborhood Region

i.e. neighbors of an active compound have a higher probability of behaving in a 'similar' way

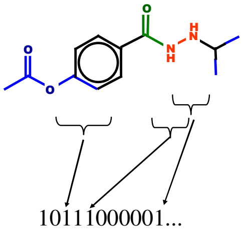
Numerical Similarity Measures

- Calculate some numerical measure of similarity between molecules
- Query structure is a "target" molecule
- Database structures can be ranked in decreasing order of similarity to target
  - find all molecules within threshold similarity to target
  - find N most similar molecules to target

A fingerprint is a 'molecular bar code' for a molecule

- Used because
  - shows neighborhood behavior
  - does not require structural conformation or alignment
  - fast searching method
- Fingerprint method used is CRC algorithm
- Advantages/disadvantages
  - 'valid' similarity in wide range of biological assay
  - easy to calculate
  - difficult to understand
  - not specific to one area

Substructure Keys

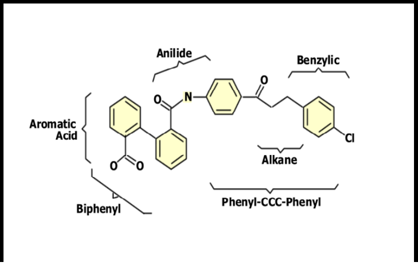


Dictionary of Keys

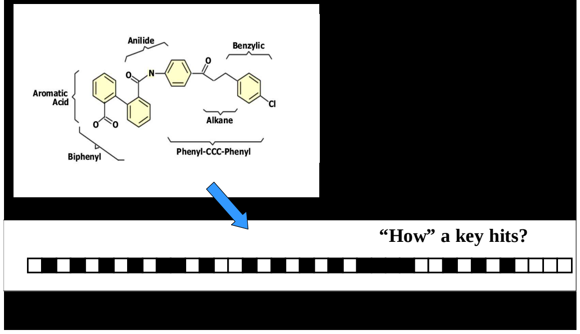
N-N
O-C(-N)-C
CH_3-Ar-CH_3
C-N-N
N-Ar-Ar-O
N-C-O
N-Ar-O
OH > 1
CH_3 > 1
N > 1
NH
...

Substructural Keys

- Compounds are multi-domain:
  - multiple occurrences of a key/substructure
  - members of more than one chemical family



"Bit Strings" of Substructure Keys



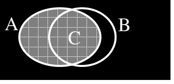
"How" a key hits?

Similarity from Fingerprints

- similarity measures are most commonly calculated from structure fingerprints
  - count the bits that are "on" in both molecules ("C")
  - count the bits that are "on" in each molecule separately ("D")

struct A:	0001010001010100010101001110100	13 bits	
struct B:	0000000010010100100100001110000	8 bits	
A & B = C:	0000000000010100000100001110000	6 bits	$A \cap B$
A or B = D:	0001010011010100110101001110100	15 bits	$A \cup B$

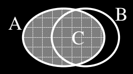
  - similarity coefficient can be calculated from A, B and C



Tanimoto Coefficient

- similarity = C/D
- similarity = $\dfrac{C}{A + B - C}$

  = 6 / (13 + 8 – 6) = 0.4
- the number of bits set in both molecules ("C") divided by the number of bits set in either molecule ("D")
- The Tanimoto Coefficient is the most commonly used similarity coefficient in chemical informatics
  - also called the Jaccard coefficient

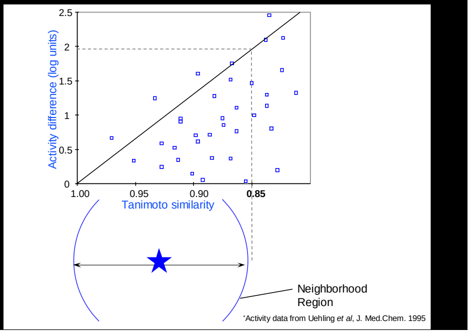


$$T = \frac{(A \cap B)}{(A \cup B)}$$

Values above 0.85 are usually significant.

Neighborhood Behavior

How well do 2D fingerprints measure neighborhood behavior in 20 literature datasets?



*Activity data from Uehling *et al*, J. Med.Chem. 1995

Selection of Representative Compounds From Virtual Libraries

From all the molecules in a Chemical Library,

choose a diverse but representative subset to study

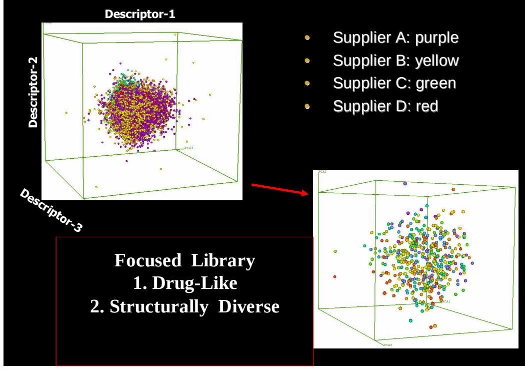
↓

Run Biological Assays only on Representative Subset,

thereby saving Time, Money, Resources and Labor

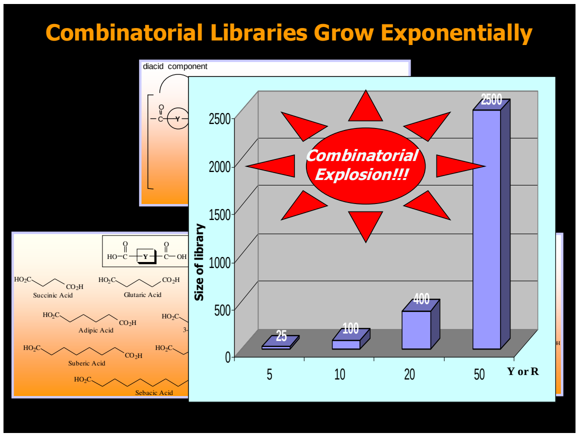
Chemicals Mapped in Descriptor Space

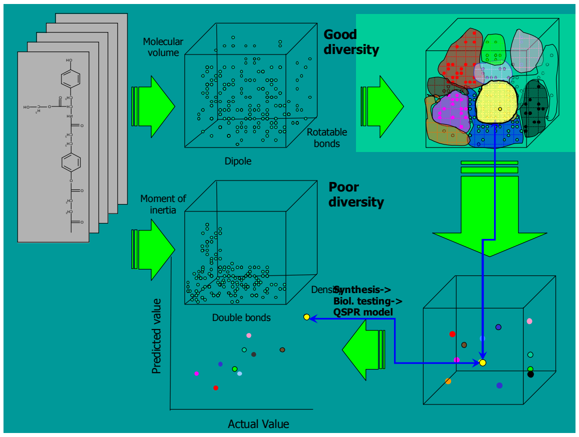
- Supplier A: purple
- Supplier B: yellow
- Supplier C: green
- Supplier D: red



Focused Library
1. Drug-Like
2. Structurally Diverse

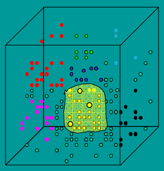
Combinatorial Libraries Grow Exponentially





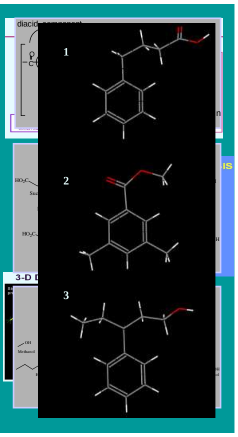
From Models to Rational Design and Synthesis

From QSPR models, select those molecular features that are associated with optimal performance property

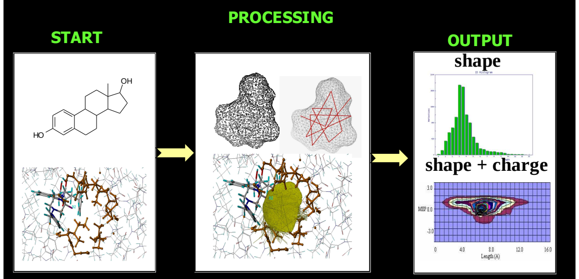


Synthesize known molecules within cluster

Design and synthesize new scaffolds within cluster



Shape Signatures Tool



START

PROCESSING

OUTPUT

shape

shape + charge

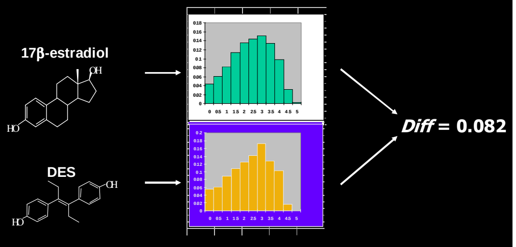
Small molecule or Protein binding pocket

Generation of the Shape Signature

1D and 2D Shape Signatures

Shape Signatures

compares molecules based on similarity in shape and electrostatic properties, rather than on chemical structure



17β-estradiol

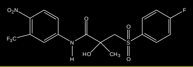
DES

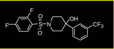
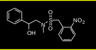
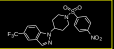
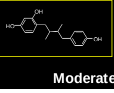
Diff = 0.082

Small value of *Diff* indicates that 17β-estradiol and DES are "similar", even though they belong to different chemical classes.

Predictive Capability of *Shape Signatures*

Bicalutamide (AR antagonist) as query against multiple databases



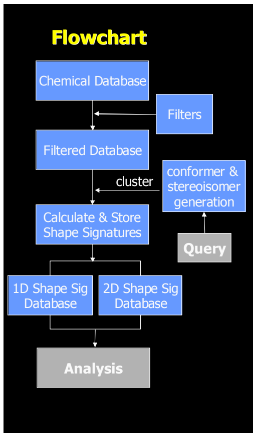
Source	Entry ID	1D Score	2D Score	Structure
LeadQuest	T_535051	0.047	0.204	
LeadQuest	T_409658	0.058	0.305	
ASINEX	ASN_6365557	0.063	0.482	
EDC	EDC 130_042	0.069	0.480	

show AR affinity

Moderate affinity to AR

Chem. Res. Toxicol. 16, 1338-1358 (2003)

Flowchart



Chemical Database

Filters

Filtered Database

cluster

conformer & stereoisomer generation

Calculate & Store Shape Signatures

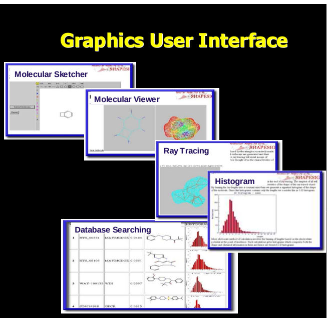
Query

1D Shape Sig Database

2D Shape Sig Database

Analysis

Graphics User Interface

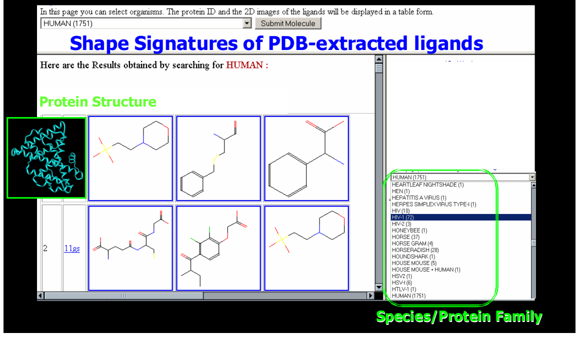


Shape Signature Databases

- Current Database >2 million cmpds
- Directed kinase and GPCR databases
- PDB-extracted ligand database

PDB-based *Shape Signatures* Database

Protein Data Bank (PDB): World Repository of 30K Protein-Ligand Crystal Structures (http://www.rcsb.org/pdb/)



In this page you can select organisms. The protein ID and the 2D images of the ligands will be displayed in a table form.

HUMAN (1751) Submit Molecule

Shape Signatures of PDB-extracted ligands

Here are the Results obtained by searching for HUMAN :

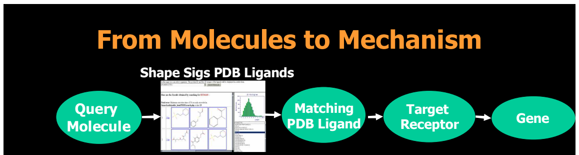
Protein Structure

2 11gs

HUMAN (1751)
HEARTLEAF NIGHTSHADE (1)
HEN (1)
HEPATITIS A VIRUS (1)
HERPES SIMPLEX VIRUS TYPE-I (1)
HIV (10)
HIV-1 (72)
HIV-2 (3)
HONEYBEE (1)
HORSE (37)
HORSE GRAM (4)
HORSERADISH (28)
HOUNDSHARK (1)
HOUSE MOUSE (5)
HOUSE MOUSE + HUMAN (1)
HSV2 (1)
HSV-I (6)
HTLV-1 (1)
HUMAN (1751)

Species/Protein Family

From Molecules to Mechanism



Shape Sigs PDB Ligands

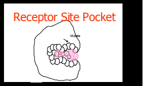
Query Molecule

Matching PDB Ligand

Target Receptor

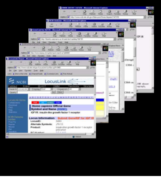
Gene

- KEGG Metabolic Pathways:
  - http://www.genome.ad.jp/kegg/metabolism.html
- EMP - Enzymes and Metabolic Pathways:
  - http://emp.mcs.anl.gov/
- WIT - Metabolic Reconstruction:
  - http://wit.mcs.anl.gov/WIT2/
- UM-BBD - Microbial Biocatalysis/Biodegradatation:
  -- http://umbbd.ahc.umn.edu/
- EcoCyc - *E. coli* Genes and Metabolism:
  - http://www.ecocyc.org/
- Metalgen - Genes and Metabolism:
  - http://indigo.genetique.uvsq.fr/
- Boehringer Mannheim - Biochemical Pathways:
  - http://www.expasy.org/cgi-bin/search-biochem-index



Receptor Site Pocket

Public Databases



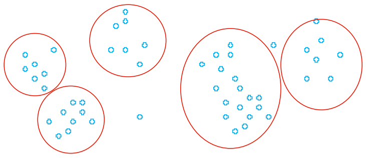
Key Features of *Shape Signatures*

- **Uses molecular shape and features (e.g. surface charge), thus find hits missed by techniques that search on chemical (sub)structure alone**
- **Many uses: scaffold hopping (crossing chemical families), predictive toxicology, *inverse* structure-based drug design**
- **Fast; simple input; easy to use, update, and expand; very compact; infinitely expandable**
- **Works for organics and organometallics, neutral or charged**
- **Applicable in ligand-based mode (ligand-ligand *similarity*) and receptor-based mode (ligand-receptor *complementarity*)**

Cluster analysis

- Refers to a group of statistical methods used for identifying groups ("clusters") of similar items in a multi-dimensional space
- Three popular methods of cluster-analysis:

  Ward's, K-means and Jarvis-Patrick
- Require a measure of distance or similarity between items



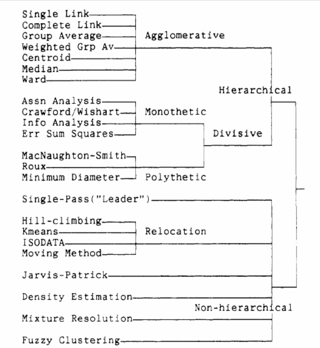
Cluster analysis applied to chemical information

- Three main uses:
  - Grouping compounds into series, particularly helpful in analyzing large datsets (i.e. 1,000 series easier to analyze than 50,000 arbitrary compounds)
  - Grouping structures which are likely to have similar biological activity, the premise being that if several compounds in a cluster are active, others are likely to be active too
  - Picking small sets of "representative compounds" from large datasets
- Common measures of similarity and distance – Tanimoto and Euclidean
- By incorporating these fingerprint-based methods, we can use standard cluster-analysis techniques for finding groups of similar structures in a dataset

Kinds of cluster analysis used in chemoinformatics

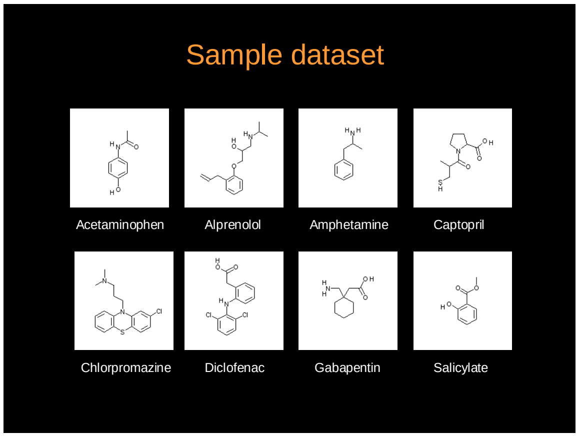
- Hierarchical
  - Agglomerative (e.g. Wards)
  - Divisive
- Non-hierarchical
  - Single-pass
  - Nearest Neighbor (e.g. Jarvis-Patrick)
  - Relocation (e.g. K-means)
- "New" methods
  - ROCK, CURE, CLARA, Chamelion

Clustering methods



Hierarchical Clustering - Agglomerative

- Starts with each compound in its own cluster
- The two most similar clusters are merged
- The process repeats (creating a "tree") until all items are merged into one cluster
- *Wards* uses Euclidean Distance to measure similarity between items. Clusters of more than one compound are represented by an "mean" fingerprint



Sample dataset

Acetaminophen

Alprenolol

Amphetamine

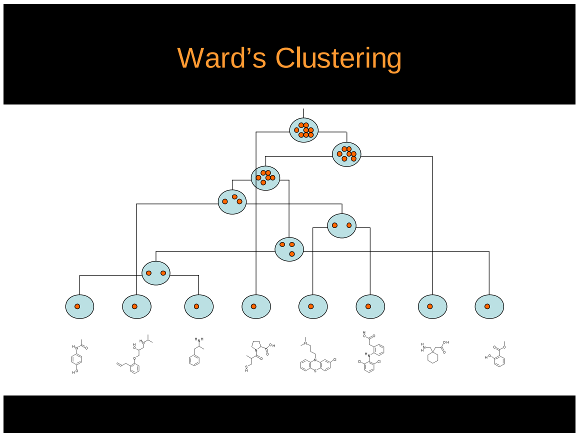
Captopril

Chlorpromazine

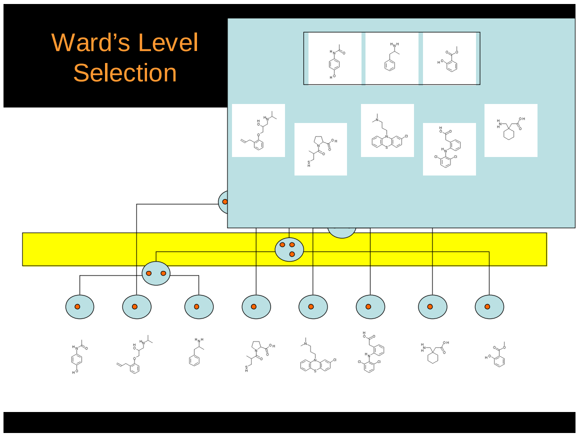
Diclofenac

Gabapentin

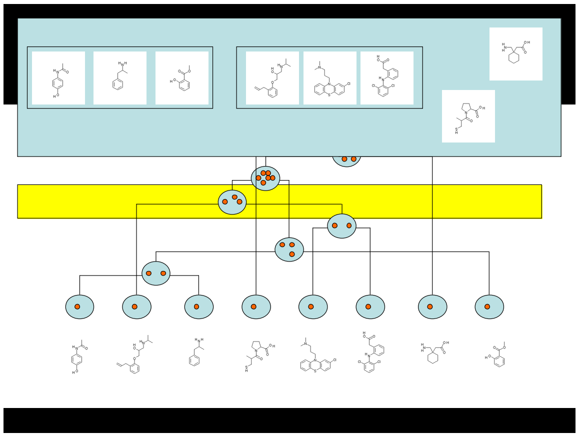
Salicylate

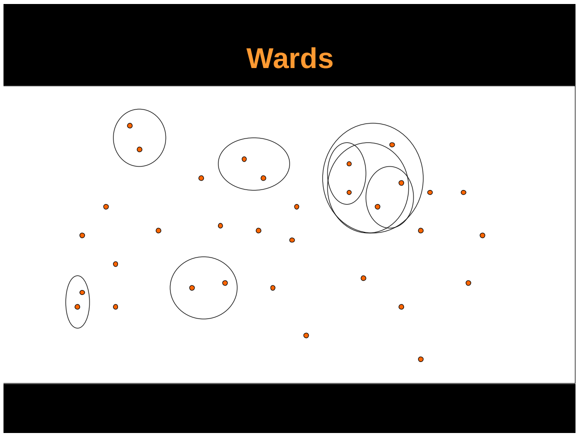


Ward's Clustering

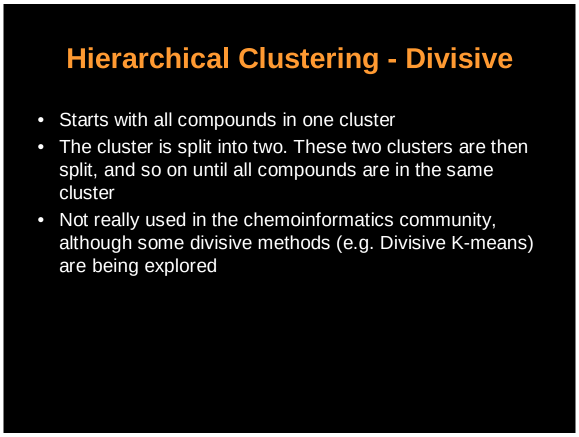


Ward's Level Selection





Wards



Hierarchical Clustering - Divisive

- Starts with all compounds in one cluster
- The cluster is split into two. These two clusters are then split, and so on until all compounds are in the same cluster
- Not really used in the chemoinformatics community, although some divisive methods (e.g. Divisive K-means) are being explored

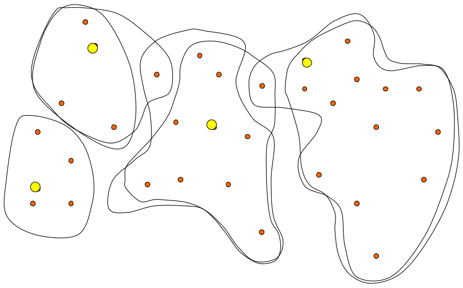
Jarvis-Patrick

- For each compound in a dataset, the *J* nearest neighbors (i.e. other compounds in the dataset that are the most similar) are identified.
- Compounds are then placed in the same cluster if they:
  - Are in each others' list of *J*-nearest neighbors
  - *K* of their *J* nearest neighbors are in common
- Requires that *J* and *K* be predefined
- Usually uses Tanimoto as measure of similarity
- Very fast, but clusterings generally not as good as other methods

K-means clustering (Relocation)

- Pick a random set of initial cluster "centroids"
- Place each of the items into the nearest cluster
- Recalculate centroids
- Repeat, until no more items change cluster

K-means



K-means

- Need to decide number of clusters beforehand
- Much faster than Wards
- Generally requires a few (3-50) iterations to settle
- Less likely to produce "singletons" than Wards => you have 'stragglers' in clusters

"New" methods

- Most work was done on clustering methods in the 60's and 70's. Then not much was done until the 90's when a bunch of new methods were developed as a result of the needs of data mining
- These are generally able to handle oddly-shaped clusters better than their older counterparts
- Still yet to be evaluated for chemoinformatics
- Examples: ROCK, CURE, Chameleon
- See Downs & Barnard 2002 paper for more information

Current consensus on Clustering

- Wards provides the most accurate clustering, but is time consuming – O(N^2)
- There are multiple ways to choose a level from a Ward's hierarchy
- K-means is much faster than Wards – O(N) – but not quite as effective
- Jarvis-Patrick still used especially for very large datasets
- A number of new methods have been introduced into the data mining community in the last 10 years, and these are under investigation for use in Chemoinformatics applications

Cluster analysis - General References

- Chemical Similarity Searching, P. Willett, J.M. Barnard, G.M.Downs, *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996
- Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures, *J. Chem. Inf. Comput. Sci*, **1992**, *36*, 644-649
- Clustering methods and their uses in Computational Chemistry, G.M.Downs and J. M. Barnard, *Reviews in Computational Chemistry*, **2002**, *18*, 1-40
- Gaussian mixture clustering and imputation of microarray data, M Ouyang, WJ Welsh, P Georgopoulos, *Bioinformatics*, **2004**, *20*, 917-923

Cluster analysis - Application

- Separating Actives and Inactives
  - Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection, R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comput. Sci.*, **1996,** *36*, 572-584.
- Finding series
  - Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping using Wards Clustering, D.J. Wild, J. Blankley, *J. Chem. Inf. Comput. Sci.*, **2000**, 40, 155-162.

Diversity Analysis

- Arose in the late 1990’s in response to the following needs:
  - There was much interest as to how well the corporate collections held by pharmas “covered” possible chemistry / drug space
  - *Combinatorial Chemistry* experiments were producing many new compounds, and people wanted to know if these compounds added anything new to their corporate collections, i.e. if they made the datasets more diverse, or just replicated what was already in there
  - Libraries of thousands of compounds became available for purchase – are they worth the money?

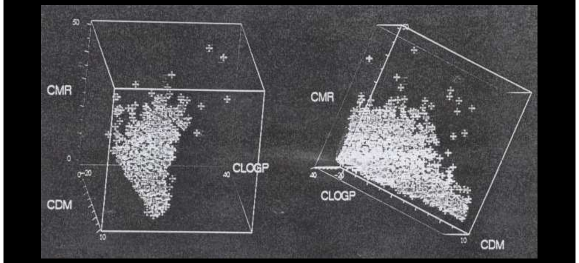
“Descriptor Space”

- If you chose a descriptor set (e.g. of *n* fingerprint bits), the “descriptor space” represents the space created if you plot each of the descriptors as a separate dimension
- E.g. if we just had two descriptors (mol.wt. and LogP), our descriptor space would be:

“Descriptor Space”

- People began to talk about “*Chemistry Space”* and “*Drug Space”*:
  - Chemistry space – if you made all the possible compounds that could theoretically be made, the chemistry space represents the regions of a multi-dimensional descriptor space (as defined by a given descriptor set) that would be occupied
  - Drug space – the regions of the chemistry space that would be inhabited by drug molecules
- So questions began to be asked such as “how much of chemistry space does our corporate collection cover?”; “how could we cover more?”; “what about drug space?” etc.

Simple descriptor space for corporate collection



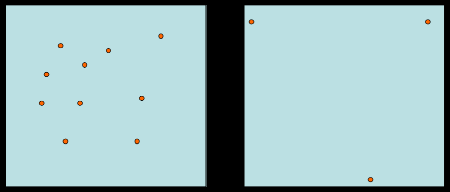
“Diversity”

- Thus, companies wanted to increase the “diversity” of their corporate collections, i.e. make them cover more chemistry and / or drug space.
- The hope then is that you have a better chance of finding a “hit” in a high-throughput screen, etc.

Measuring Diversity of a set of compounds - Mean dissimilarity method

- Calculate the *Mean Inter-molecular Similarity* of all the pairs of molecules in the set, e.g. using the tanimoto coefficient:
- Mean Dissimilarity = (1 – MIMS)
- Gives a measure of *relative* diversity, i.e. how different the molecules are to each other. Doesn’t say how much “space” is covered by the molecules

Which is the most “diverse”?

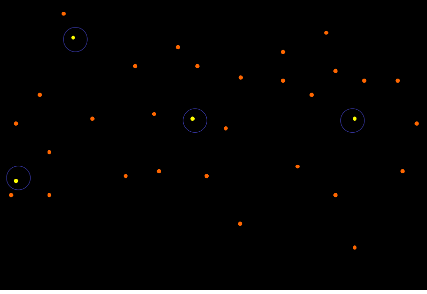


Picking a “representative set”

- Find a small subset of compounds from a larger set which “represents” the large set
- We can then, e.g. only screen the small subset, on the assumption that we’re “covering the chemistry space” of the large set

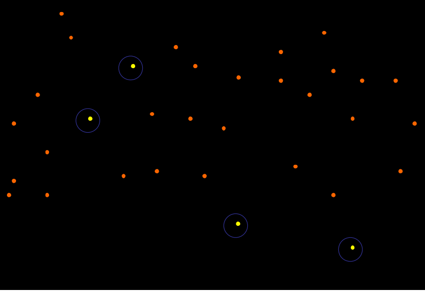
Picking a “representative set”

- E.g. by clustering, and picking compounds nearest the cluster centroids:



Picking a “representative set”

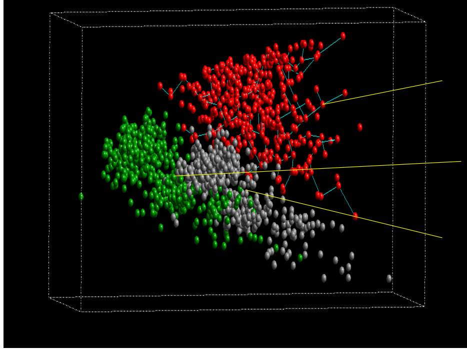
- E.g. pick the set which Maximizes the Minimum distance between representatives



Comparing sets of compounds

- How diverse is this set compared to this other set?
  - You can compare Mean dissimilarity
  - Comparing with a large, general dataset (e.g. World Drugs Index) can give a measure of how a dataset compares in diversity to a large, general collection, which approaches "coverage"
- How different are these two sets of compounds?
  - Calculate individual diversity measures, then the diversity measure when combined. How much does the diversity go up?
  - BUT: May not be accurately reflected by mean dissimilarity

Comparing sets of compounds



Modern QSAR

- Use computational statistical and machine-learning methods to build "models of activity" to predict activity of unknown compounds (2D or 3D)
- Models are trained using compounds where activity is known
- Examples:
  - Linear and Multiple regression
  - Principal Component Analysis
  - Recursive partitioning
  - Neural Networks
  - Support Vector Machines
  - Genetic Algorithms
  - Bayesian analysis
  - Version Spaces
- See NetSci QSAR articles in http://www.netsci.org/Science/Compchem/

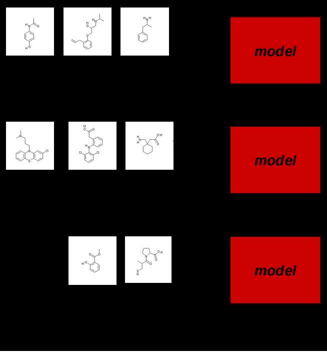
Building models of activity

- Most methods assume a single response variable (e.g. activity) and multiple descriptor variables (e.g. fingerprint bits, properties).
- *Linear methods* (e.g. Hansch, Free Wilson) assume that the activity varies linearly with the descriptor values that affect it
- *Non-linear methods* do not make this assumption, and thus are generally the most useful.

Building models of activity

- Most nonlinear methods use three phases:
- A ***training phase*** where the models are presented with sets of descriptors and known responses (e.g. fingerprint bits and known activities for a set of compounds)
- A ***validation phase*** where the trained model is tested on compounds with known activity, but where the activity isn't presented to the model
- A ***predictive phase*** where the model is used to predict activity of unknown compounds

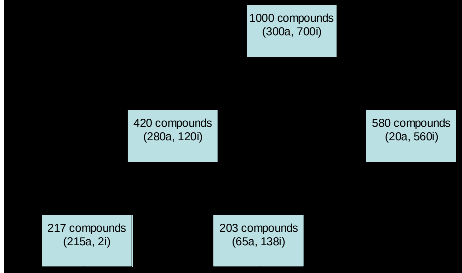
Model development phases



Recursive Partitioning

- One of the first methods to be applied to large datasets (e.g. using HTS data)
- When trained, RP recursively splits a dataset into two subsets, based on the values of a particular descriptor. It splits based on the descriptors and their values that best discriminate between actives and inactives
- The criterion used for splitting can then be used predictively – the predicted activity is usually the average of the set into which it falls

Recursive Partitioning



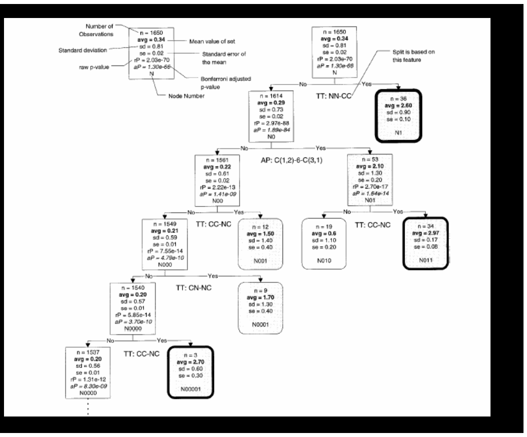
1000 compounds (300a, 700i)

420 compounds (280a, 120i)

580 compounds (20a, 560i)

217 compounds (215a, 2i)

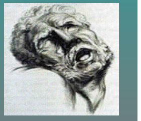
203 compounds (65a, 138i)



Ligand-Based Drug Design

Opioid Receptor Active Compounds

The Opioids for Treating Pain



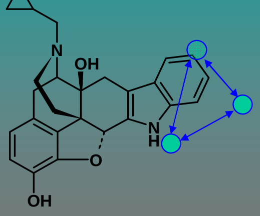
Ø Powerful analgesics like Morphine
- codeine, methadone, fentanyl, etc.
- three related receptors: δ, κ, μ
- morphine prefers μ over δ and κ

Ø So, what's wrong with the opioids?
- respiratory depression
- nausea, vomiting, constipation
- addictive

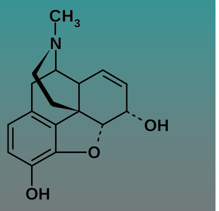
Ø Our Solution
- find a new molecule that prefers δ over μ and κ
- *okay ... but how?*

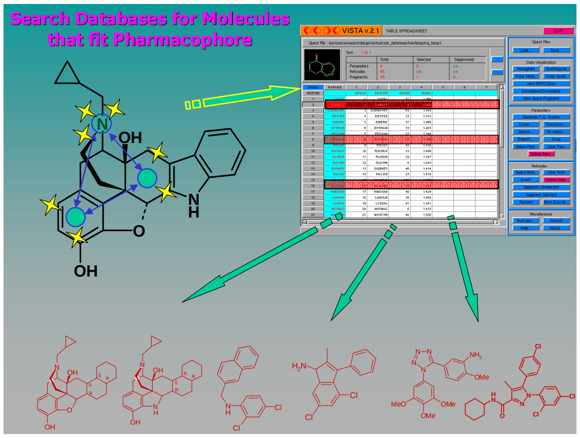
Identify Common Features (Pharmacophore)

Naltrindole



Morphine





Search Databases for Molecules that fit Pharmacophore

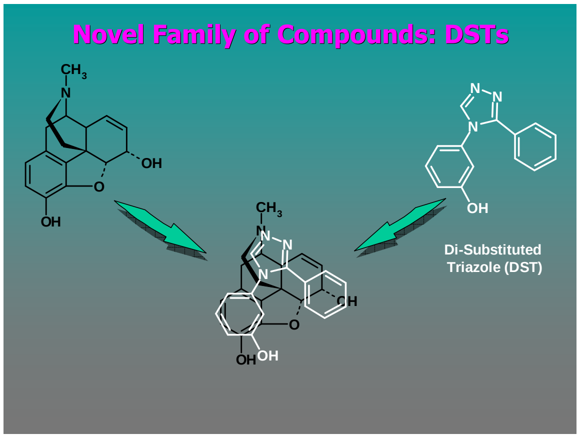
VISTA v.2.1 TABLE SPREADSHEET

Quest File : /usr/users/research/large/vucha/csds_data/searches/temp/cq_temp1

Test : 1 of 1

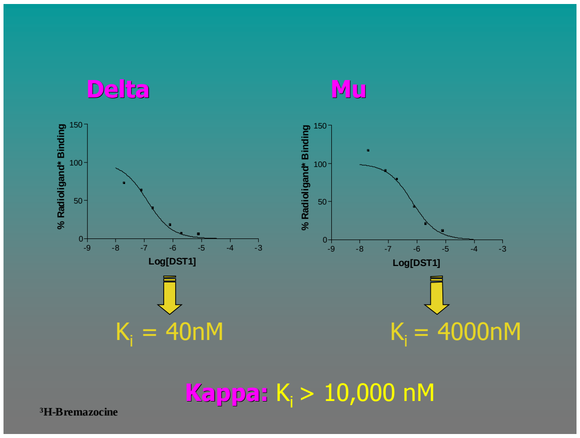
	Total	Selected	Suppressed
Parameters	4	0	n/a
Refcodes	45	n/a	n/a
Fragments	45	1	0

<Home>	PARAMS	1	2	3	4	5	6	7
REFCOD		NFRAG	REFCOD	*COOR	*DENC			
1	CIXSIJ	1	CIXSIJ	51	1.560			
2	CUHRAV	2	CUHRAV	0	0.000			
3	CUHRAV01	3	CUHRAV01	63	1.343			
4	DESTER	4	DESTER	33	1.533			
5	EMERIC	5	EMERIC	57	1.268			
6	EPSHAM	6	EPSHAM	53	1.283			
7	FECVAN	7	FECVAN	23	1.340			
8	[illegible]	8	[illegible]	57	[illegible]			
9	FISCEP	9	FISCEP	42	1.532			
10	FUCWUU	10	FUCWUU	53	1.499			
11	FUJROR	11	FUJROR	52	1.347			
12	GUAVIN	12	GUAVIN	0	1.245			
13	GUDNEY	13	GUDNEY	40	1.414			
14	HALSUI	14	HALSUI	25	1.410			
15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]			
16	[illegible]	16	[illegible]	[illegible]	[illegible]			
17	HMOXAN	17	HMOXAN	40	1.429			
18	LANUUR	18	LANUUR	58	1.482			
19	LVSOTA	19	LVSOTA	91	1.381			
20	MSTMAC	20	MSTMAC	0	1.433			
21	[illegible]	21	MXSCYN	40	1.520			



Novel Family of Compounds: DSTs

Di-Substituted Triazole (DST)



Delta

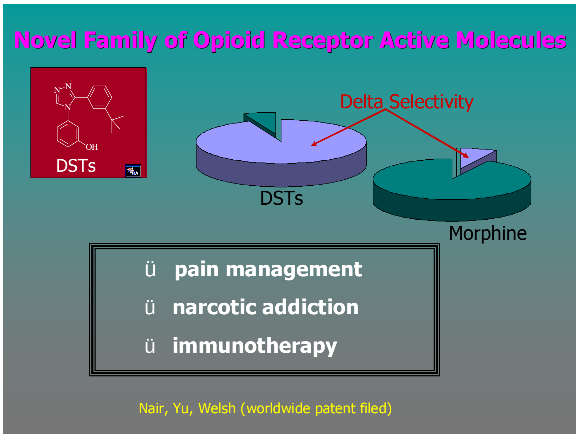
Mu

K_i = 40nM

K_i = 4000nM

Kappa: K_i > 10,000 nM

^{3}H-Bremazocine



Novel Family of Opioid Receptor Active Molecules

DSTs

Delta Selectivity

DSTs

Morphine

ü **pain management**

ü **narcotic addiction**

ü **immunotherapy**

Nair, Yu, Welsh (worldwide patent filed)