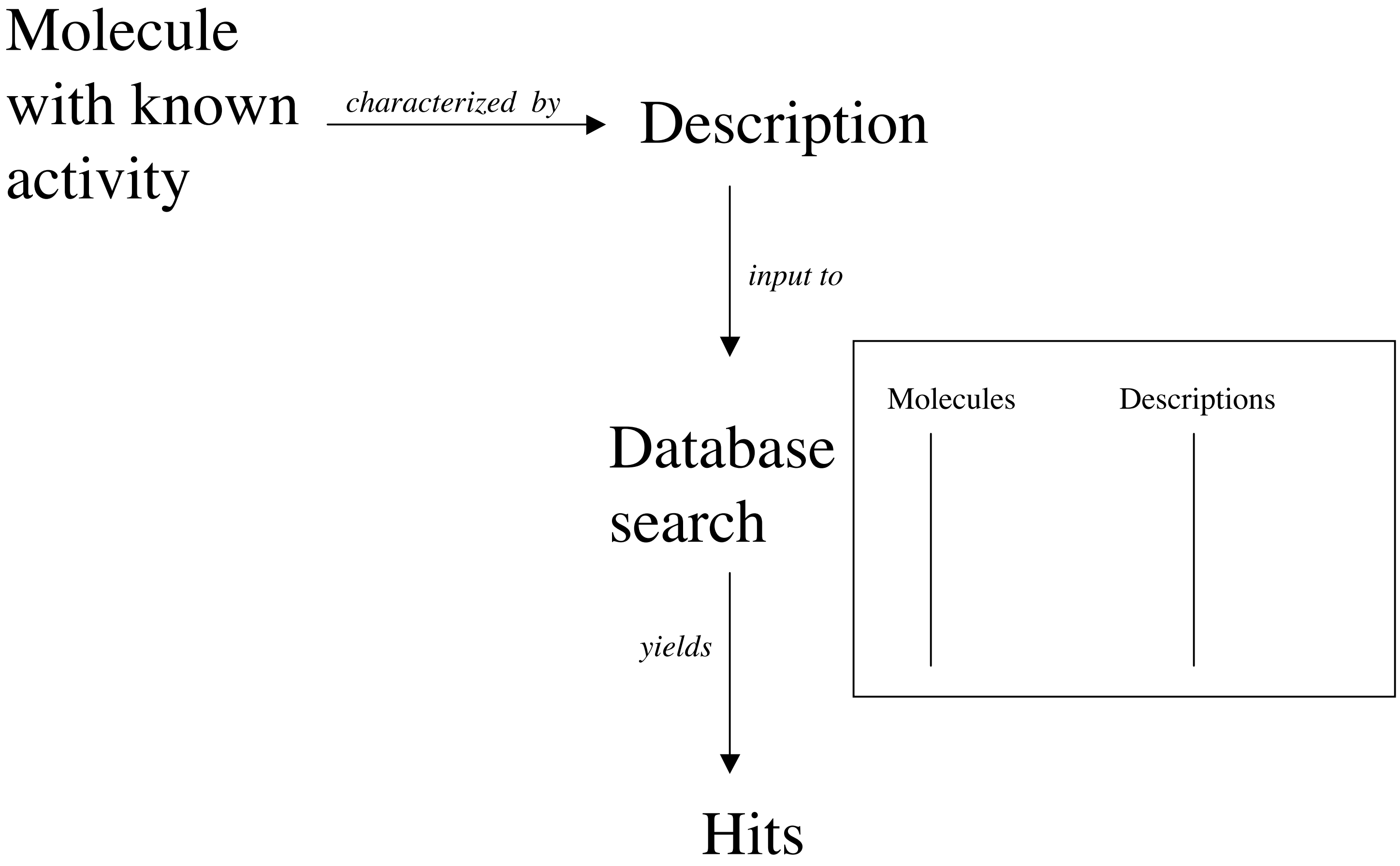
Similarity Searching and QSAR

January 17, 2006

Basic Idea

Characterize molecule in a way that hopefully captures cause of its activity



Molecule with known activity

characterized by

Description

input to

Database search

Molecules

Descriptions

yields

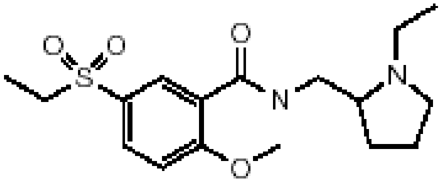
Hits

Descriptors

1D

$C_{17}H_{26}N_2O_4S$ → molecular mass

2D



→ number of aromatic bonds;
molecular connectivity index;
logP(o/w)

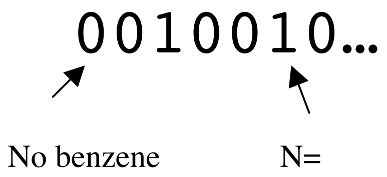
3D

→ van der Waals volume;
solvent-accessible surface area

Bajorath, 2002

More Descriptors

- Molecular fingerprints
  - Each bit associated with a given feature



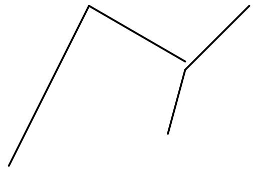
0010010…

No benzene

N=

  - Tanimoto coefficient $T = \frac{b_c}{b_1 + b_2 - b_c}$

- Graphs
  - Vertices represent atoms
  - Edges represent bonds



3D Pharmacophores

- Part of ligand that binds
- Set of ligand features (chemical or structural) together with distances between them
- *N*-point pharmacophore means pharmacophore defined by *N* features and the distances between them (N=3,4,5 common). How search a DB?
  - If we enumerate all pharmacophores from a set of features and distances, then can construct a pharmacophore fingerprint where each bit represents presence or absence of given pharmacophore

QSAR

- A structure activity relationship (SAR) relates chemical structure with biological activity. With computation, make it quantitative.
- In QSAR, derive a function *f* that satisfies $a=f(\mathbf{x})$, where *a* is the activity of the molecule and **x** is a vector of properties of the molecule
- Uses
  - Not really to find a brand new molecule
  - Gain insights into what aspects of a compound are important in its activity
  - Help decide whether a series of compounds can be further optimized

Just Machine Learning

- Need a training set
- Feature selection
  - Knowledge
  - Forward stepping
  - Backward stepping
- Relationship
  - Regression
  - SVM
- Cross-validation
- Beware overfitting, poor training data

Docking

Basics

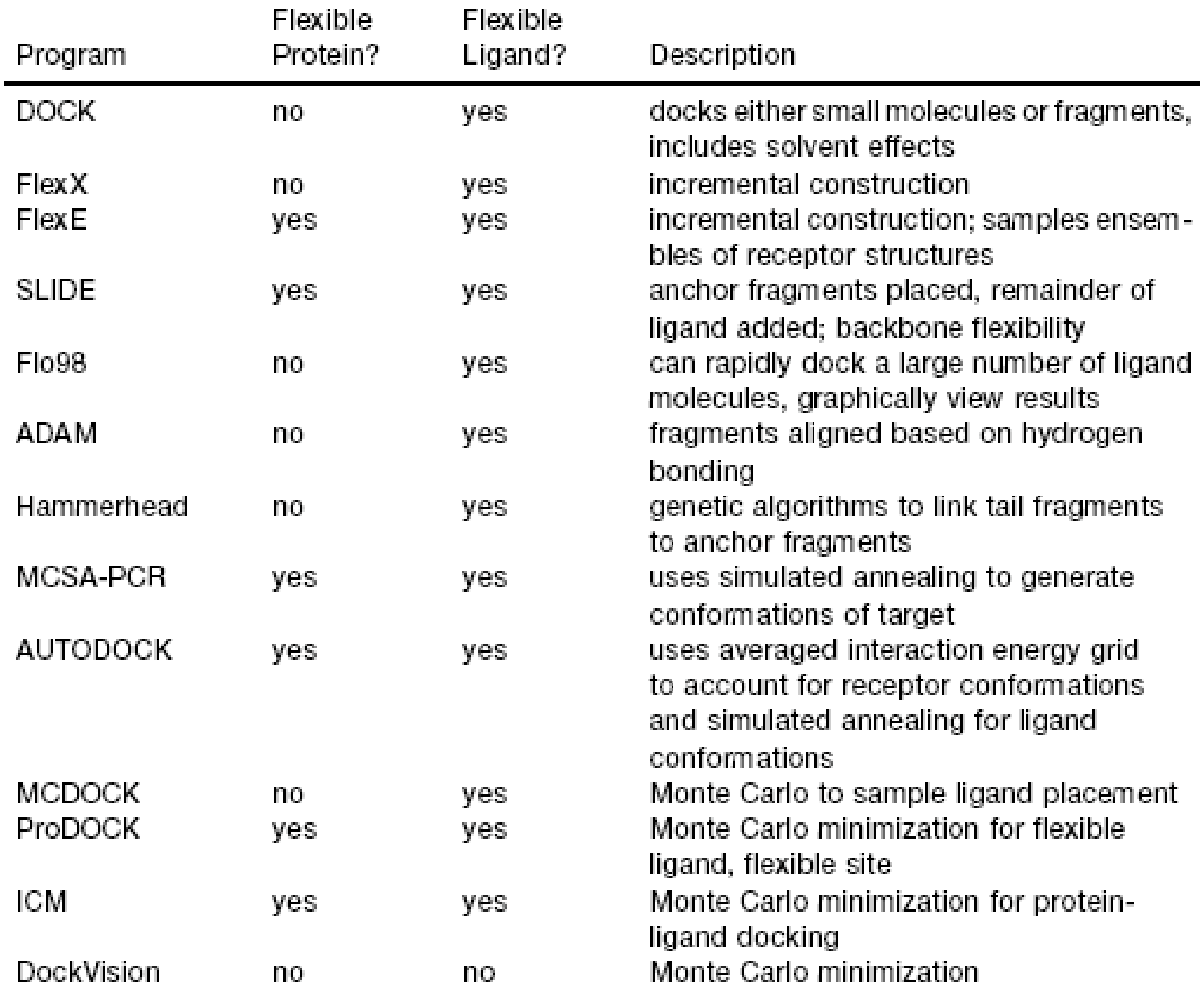
- Problem
  - Given protein and ligands, how do the ligands bind to the protein (where’s the binding pocket, what shape does the ligand take…)?
  - How well do they bind?

- Purpose
  - Prediction of binding conformations
  - Screening databases
  - Ranking ligand affinities

Choices

- What's flexible and what's rigid?
  - At first protein and ligand both rigid, now more flexibility allowed
- Sampling method
- Scoring method

Implementations



Program	Flexible Protein?	Flexible Ligand?	Description
DOCK	no	yes	docks either small molecules or fragments, includes solvent effects
FlexX	no	yes	incremental construction
FlexE	yes	yes	incremental construction; samples ensembles of receptor structures
SLIDE	yes	yes	anchor fragments placed, remainder of ligand added; backbone flexibility
Flo98	no	yes	can rapidly dock a large number of ligand molecules, graphically view results
ADAM	no	yes	fragments aligned based on hydrogen bonding
Hammerhead	no	yes	genetic algorithms to link tail fragments to anchor fragments
MCSA-PCR	yes	yes	uses simulated annealing to generate conformations of target
AUTODOCK	yes	yes	uses averaged interaction energy grid to account for receptor conformations and simulated annealing for ligand conformations
MCDOCK	no	yes	Monte Carlo to sample ligand placement
ProDOCK	yes	yes	Monte Carlo minimization for flexible ligand, flexible site
ICM	yes	yes	Monte Carlo minimization for protein-ligand docking
DockVision	no	no	Monte Carlo minimization

Next Week Readings

- A Critical Assessment of Docking Programs and Scoring Functions (Warren and GSK coworkers)
- Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine (Jain)
- Ligand-Based Structural Hypotheses for Virtual Screening (Jain)