

# Methods for “finding” biomolecular complexes

Nathan Baker

[baker@biochem.wustl.edu](mailto:baker@biochem.wustl.edu)

Bio 5476

Spring 2005

# Introduction

- A significant portion of biology is built on the paradigm  
sequence → structure → function
- As we sequence more genomes and get more structural information, the next challenge will be to predict interactions and binding for two or more biomolecules (nucleic acids, proteins, peptides, drugs or other small molecules)

# Introduction

- The questions we are interested in are:
  - Do two biomolecules bind each other?
  - If so, how do they bind?
  - What is the binding free energy or affinity?
- The goals we have are:
  - Searching for lead compounds
  - Estimating effect of modifications
  - General understanding of binding
  - ...

# Rationale

- The ability to predict the binding site and binding affinity of a drug or compound is immensely valuable in the area of pharmaceutical design
- Most (if not all) drug companies use computational methods as one of the first methods of screening
- Computer-aided drug design is a more daunting task, but there are several examples of drugs developed with a significant contribution from computational methods

# Examples

- **Tacrine** – inhibits acetylcholinesterase and boost acetylcholine levels (for treating Alzheimer's disease)
- **Relenza** – targets influenza
- **Invirase, Norvir, Crixivan** – Various HIV protease inhibitors
- **Celebrex** – inhibits Cox-2 enzyme which causes inflammation

# Docking

- *Docking* refers to a computational scheme that tries to find the best binding orientation between two biomolecules where the starting point is the atomic coordinates of the two molecules
- Additional data may be provided (biochemical, mutational, conservation, etc.) and this can significantly improve the performance, however this extra information is not required

# Bound vs. Unbound Docking

- The simplest problem is the “**bound**” docking problem. The goal in this case is to reproduce a known complex where the starting point is atomic structures from a co-crystal
- The “**unbound**” docking problem is significantly more difficult task. Here the starting point is structure in their unbound conformation, perhaps the native structure, perhaps a modeled structure, etc.

# Approach to the Problem

- One of the first suggestions on how to tackle this problem was given by Crick who postulated that “knobs” could be fitted into “holes” on the protein surface
- Lee & Richards had already described how to treat the van der Waals surface of a protein, however the van der Waals surface was not appropriate for docking purposes (too many crevices)
- Connolly showed how to calculate the molecular surface (and also the solvent accessible surface) which allowed for many docking programs to start being developed in the mid 1980's

# A Matter of Size

- For two proteins, the docking problem is very difficult since the search space (all possible relative conformations) is extremely large
- In the case of a small molecule (drug, peptide or ligand) binding to a protein, we have a chance of exploring the conformational space, at least for the small molecule
- Now we want to consider the case where we have limited or no *a priori* knowledge and the mode of binding

# Docking Methodology

- All small molecule docking programs have three main components
  - Representation of system
  - The search algorithm
  - The scoring or energy evaluation routine
- In order to do a good job of docking, you need to search efficiently and evaluate the energy accurately

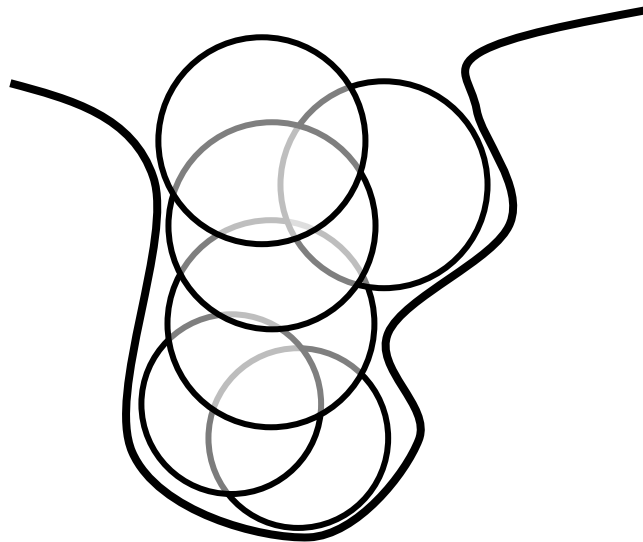
Representation of the System  
*or*  
Site Characterization

# Representation of Phase Space

- The type or choice of representation for a system is really a reflection of the type of energy evaluation or scoring function that will be used
- If we would choose the most straightforward or logical representation of the atomic coordinates of the of the two biomolecules, we would simply use a molecular mechanics force field such as Amber, Charmm or OPLS
- However, this may not be the best choice in terms of computational efficiency or practicality

# DOCK

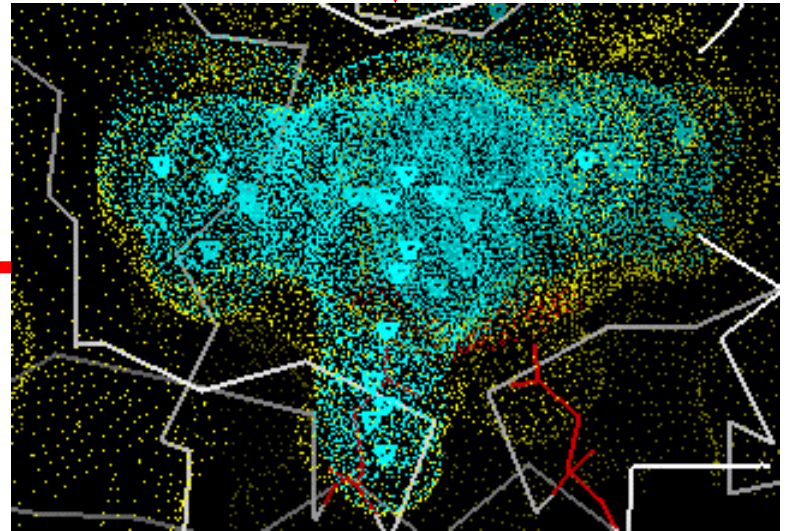
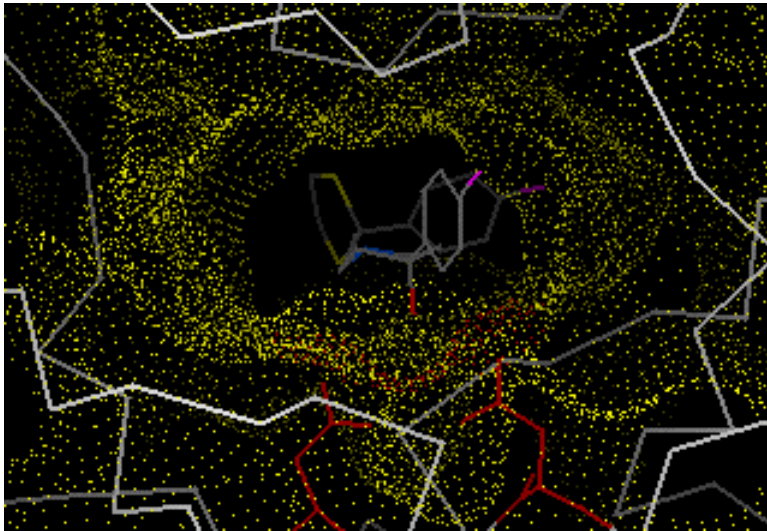
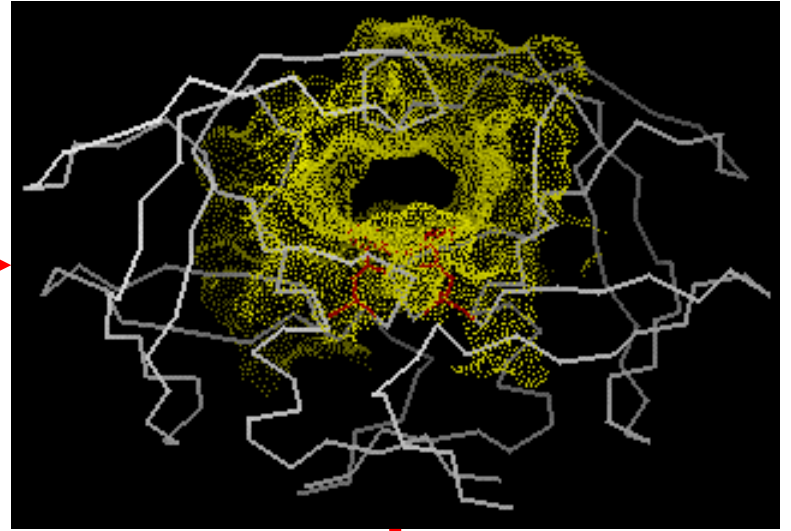
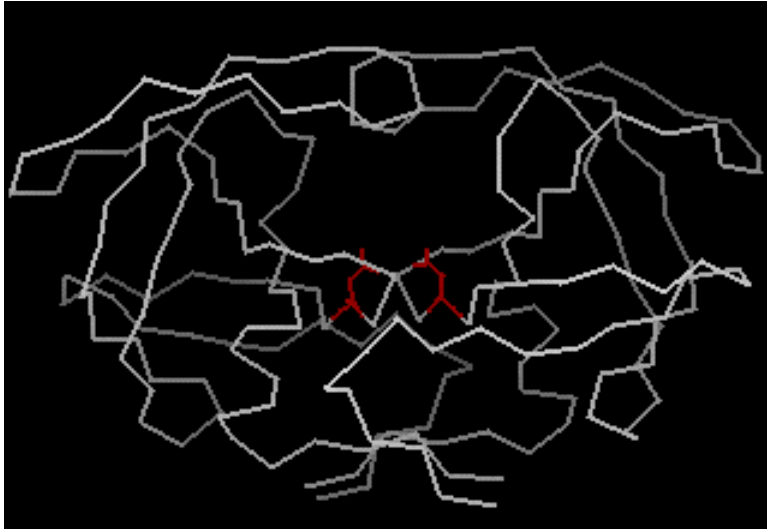
- The DOCK program is from the Kuntz group at UCSF
- It was the first docking program developed in 1982
- It represents the (negative image of the) binding site as a collection of overlapping spheres



# DOCK

- This method of a negative image is targeted at finding complexes with a high degree of shape complementarity
- The ligand is fit into the image by a least squares fitting of the atomic positions to the sphere centers
- Creating the negative image is obviously not a problem with a unique solution. Hence, factors such as the sphere radius and center-to-center distance of the spheres must be carefully controlled.

# DOCK

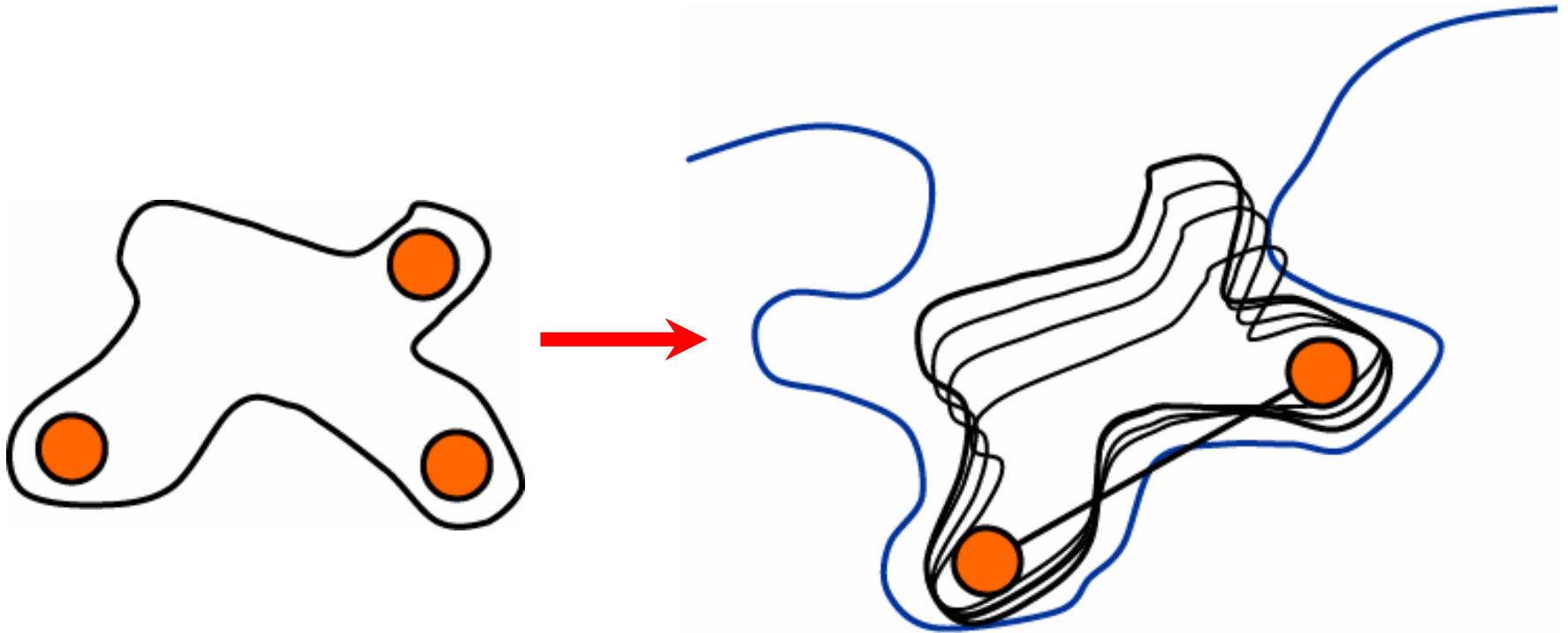


# DOCK Extensions

- Dock 4.0 has recently been extended to include:
  - Several different scoring schemes
  - Ligand flexibility (via incremental construction and fragment joining)
  - Chemical properties of receptor (each sphere assume a chemical characteristic)

# CLIX

- CLIX uses a chemical description of the receptor and distance constraints on the atoms



# ESCHER

- ESCHER uses the solvent accessible surface from a Connolly algorithm
- This surface is cut into 1.5 Å slabs that are transformed into polygons and used for rigid docking (again image matching)

G. Ausiello, G. Cesareni, M. Helmer Citterich, "ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure", *Proteins*, **28**:556-567 (1997)

# Search Methods

# Search Basics

- One of the more difficult tasks in computational docking is simply enumerating the number of ways two molecules can be put together  
(3 translational plus 3 rotational degrees of freedom)
- The size of this search space grows exponentially as we increase the size of the molecules, however this is still only for rigid structures

# Search Difficulties

- If we allow flexibility of one or both of the binding partners, this quickly becomes an intractable problem
- If we manage to solve/circumvent the problem of flexibility, we then want to be able to screen large databases of structures or drugs, so our troubles are again compounded

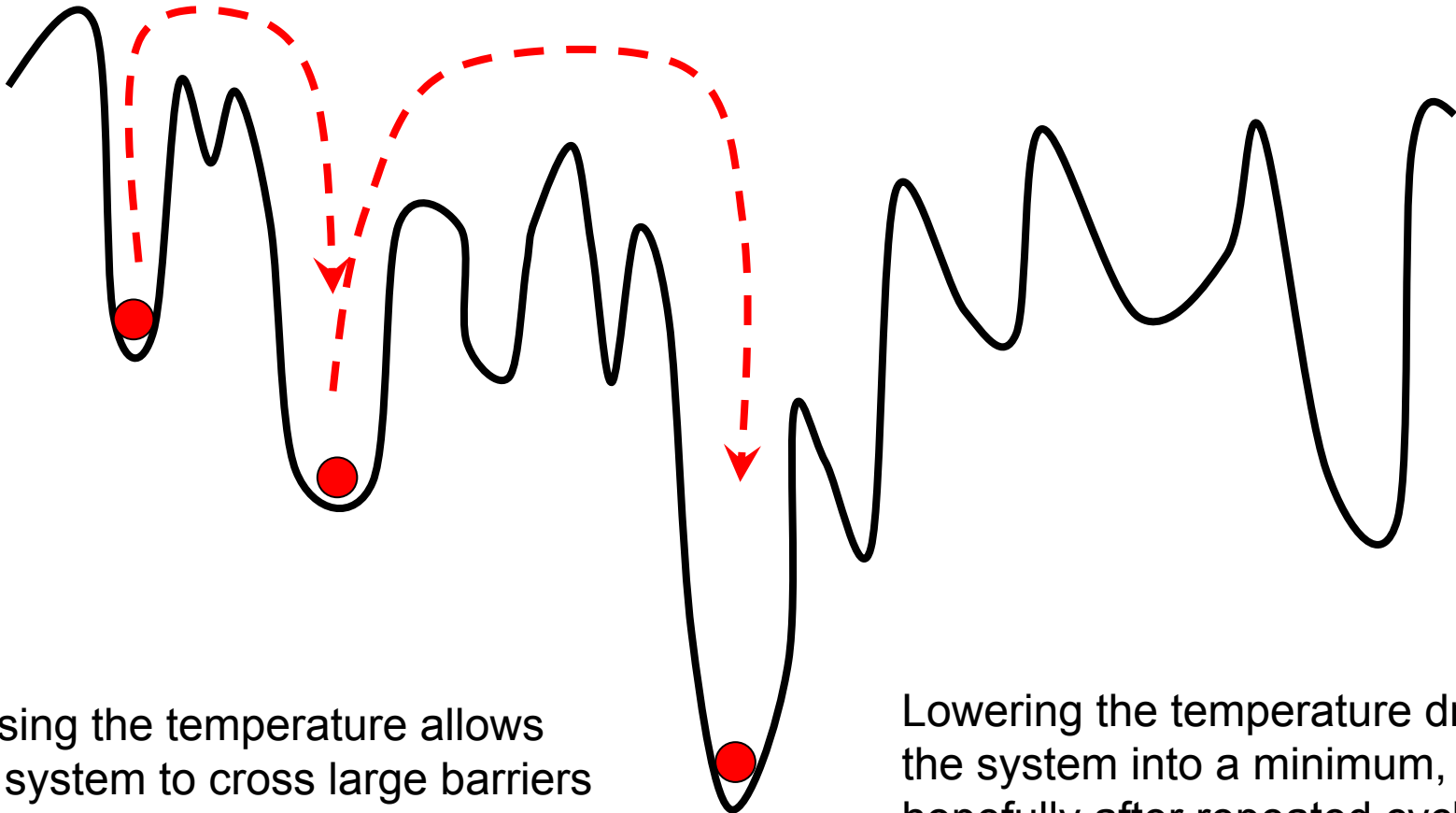
# Monte Carlo Simulated Annealing

- Also known as Metropolis Monte Carlo
- The basic steps are
  1. The ligand performs a random walk around the protein
  2. At each step, a random displacement, rotation, etc. is applied and the energy is evaluated and compared to the previous energy
  3. If the new energy is lower, the step is accepted
  4. If the energy is higher, the step is accepted with a probability  $\exp(\Delta E/kT)$

# Simulated Annealing

- If we were to perform this at a constant temperature, it would be basic Monte Carlo
- In this case, after a specified number of steps, the temperature is lowered and the search repeated
- As the temperature continues to go down, steps which increase the energy become less likely and the system moves to the minimum (or minima)

# Simulated Annealing

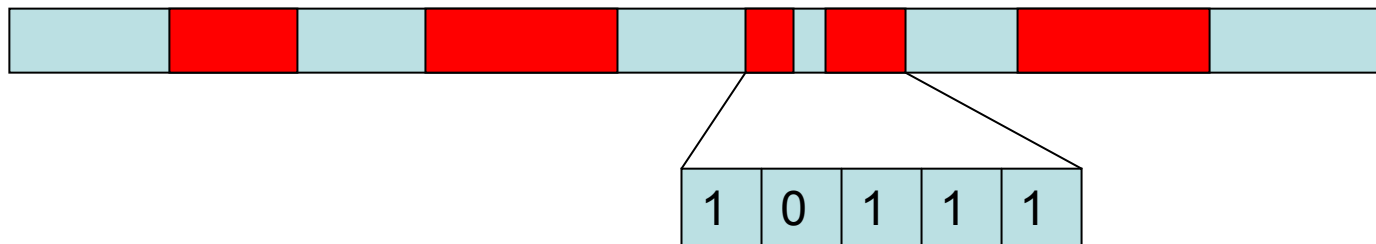


Raising the temperature allows the system to cross large barriers and explore the full phase space

Lowering the temperature drives the system into a minimum, and hopefully after repeated cycles, the global minimum

# Genetic Algorithms

- Genetic algorithms belong to a class of stochastic search methods, but rather than operating on a single solution, they function on a population
- As implied by the name, you must encode your solution, structure or problem as a genome or chromosome
- The translation, orientation and conformation of the ligand is encoded (the *state variables*)



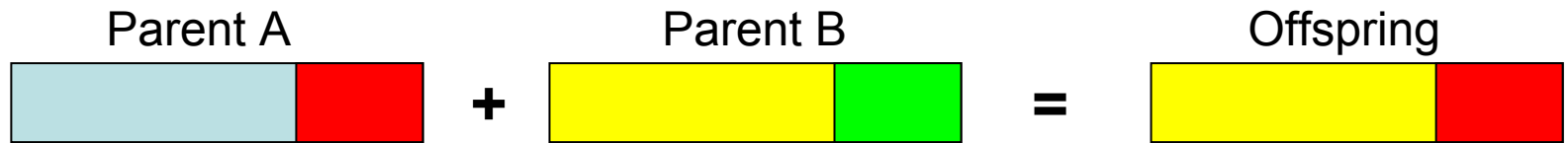
# Genetic Algorithms

- The algorithm starts by generating a population of genomes and then applies crossover and mutation to the individuals to create a new population
- The “fitness” of each structure has to be evaluated, in our case by our estimate of the binding free energy
- The best member or members survive to the next generation
- This procedure is repeated for some number of generations or energy evaluations

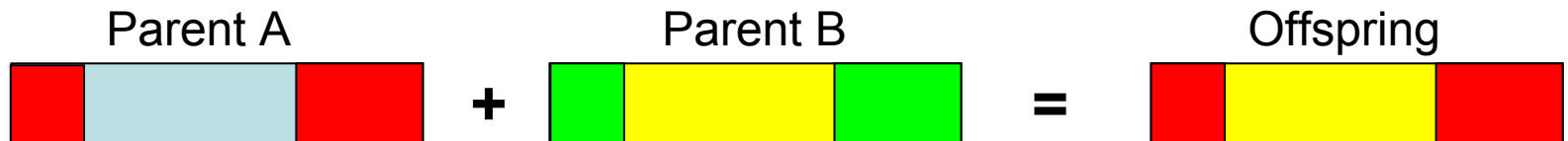
# Crossover and Mutation

- There are several possible methods of crossover

## Single Point Crossover

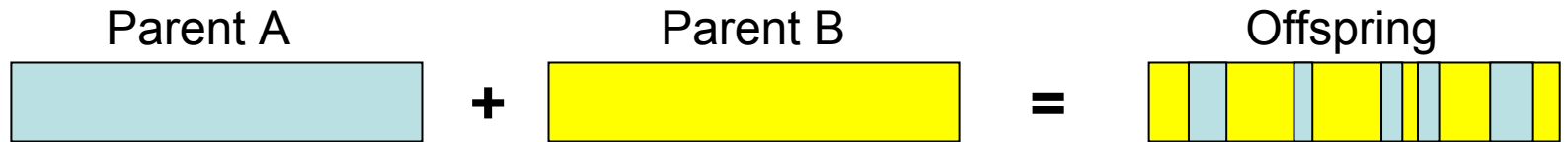


## Two Point Crossover



# Crossover and Mutation

## Uniform Crossover



- For mutation, selected bits are simply inverted



# Efficiency

- Random mutations are also possible (perhaps the orientation or position is shifted by some random amount)
- Binary crossover and mutation can introduce inefficiencies into the algorithm since they can easily drive the system away from the region of interest
- For this reason, genetic searches are often combined with local searches (producing a Lamarckian Genetic Algorithm)

# Local Searches

- In a Lamarckian GA, the genetic representations of the ligands starting point is replaced with the results of the local search
- The mapping of the local search *back* on to the genetic representation is a biologically unrealistic task, but it works well and makes the algorithm more efficient
- One of the more common local search methods is the Solis and Wets algorithm

# Solis and Wets Algorithm

Starting point  $x$

Set bias vector  $b$  to 0

Initialize  $\rho$

While max iterations not exceeded:

Add deviate  $d$  to each dimension  
(from distribution with width  $\rho$ )

If new solution is better:

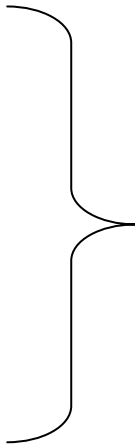
$success++$   
     $b = 0.4 d + 0.2 b$

else

$failures++$   
     $b = b - 0.4 d$

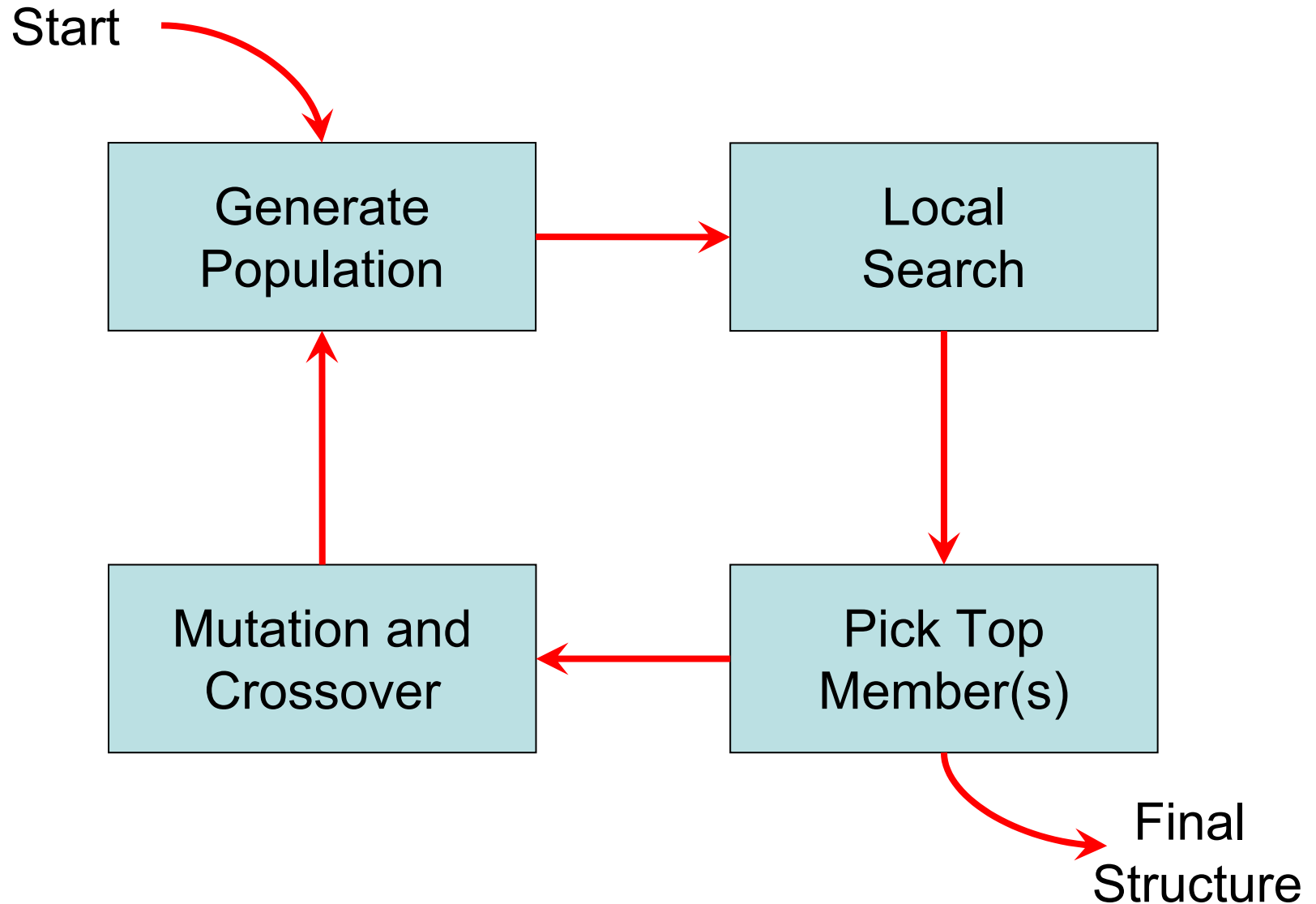
If too much success: increase  $\rho$

If too many failures: decrease  $\rho$



Adaptive step size  
which biases the  
search in the  
direction of success

# LGA Scheme



# Approaches to Flexibility

- A relatively simple molecule with 10 rotatable bonds has more than  $10^9$  possible conformation if we only consider 6 possible positions for each bond
- Monte Carlo, Simulated Annealing and Genetic Algorithm can help navigate this vast space
- Other methods have been developed to again circumvent this problem

# Flexibility

- Some algorithms (call Place & Join algorithms) break the ligand up into pieces, dock the individual pieces, and try and reconnect the bound conformations
- FlexX uses a library of precomputed, minimized geometries from the Cambridge database with up to 12 minima per bond. Sets of alternative fragments are selected by choosing single or multiple pieces in combination
- Flexible docking via molecular dynamics with minimization can handle arbitrary flexibility, however it is extremely slow

# Hybrid Methods

- Some of the newer methods use a hybrid scheme of a quick docking using a GA or other scheme followed by molecular dynamics to refine the prediction
- These methods are likely to be some of the best once they are correctly parameterized, etc.

# The Third Component: Scoring

- Last class you were introduced to the three basic requirements of any docking program
- We covered some of the most common representations of the system as well as the most commonly used search methods
- All of these search methods involve evaluating the “fitness” or “energy” of a given binding conformation
- In order to do this effectively, we must have a good scoring function that can give an accurate estimate of the binding free energy (or relative free energy)

# Practical Considerations

- If we have developed an efficient search algorithm, it may produce  $10^9$  or more potential solutions
- Many solutions can be immediately eliminated due to atomic clashes or other obvious problems, but we still must evaluate the fitness of a large number of structures
- For this reason, our scoring function must not only accurate, but it must be fast and efficient

# Scoring Accuracy

- If we were scoring a single ligand-protein complex, we could adopt much more sophisticated methods to arrive at an accurate value for the binding free energy
- Requiring true accuracy in a scoring function is not a realistic expectation, however there are two features that a good scoring function should possess
- When docking a database of compounds, a good scoring function should
  - Give the best rank to the “true” bound structure
  - Give the correct relative rank of each ligand in the database
  - And again, it must be able to do these things relatively quickly

# Types of Scoring Functions

- There are several types of scoring functions that we will discuss
  - First Principles Methods
  - Semiempirical methods
  - Empirical methods
  - Knowledge based potentials

# First Principles Methods

- First principles methods are typically based on molecular mechanics force fields which model the interactions between atoms through a combination of bonded and non-bonded interactions

$$\begin{aligned} E_{tot} = & \sum_{bonds} \frac{k}{2} (l - l_0)^2 + \sum_{angles} \frac{k}{2} (\theta - \theta_o)^2 \\ & + \sum_{torsions} \frac{V_n}{2} (1 + \cos(n\omega + \gamma)) + \sum_{atoms} 4\epsilon \left( \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right) \\ & + \sum_{atoms} \frac{q_i q_j}{4\pi\epsilon r} + \sum_{don/acc} \frac{A}{r^{12}} - \frac{B}{r^{10}} \end{aligned}$$

# First principle methods

- Although these methods can be very accurate, they are really only providing an accurate measure of the enthalpic contribution to binding

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

- As well, these methods can be very time consuming due to the large number of terms and the complexity of some of the terms (involving  $1/r^{12}$  or cosine terms)

# Simplifications

- Many docking programs use a simplified form of first principles methods
  - DOCK uses just a van der Waals term and a Coulombic term (rigid docking does not require bonded interactions)
  - AutoDock uses a more complete form of the potential but using pre-computed grids (more on this later)

# Semiempirical Methods

- Some semiempirical methods such as the Linear Interaction Energy method (LIE) use the same Coulombic and van der Waals terms, but in linear combination

$$\Delta G = \alpha (\Delta V_{\text{vdw}}) + \beta (\Delta V_{\text{Coul}})$$

- The coefficients  $\alpha$  and  $\beta$  need to be fit from experimental data
- Since electrostatic effects typically have a linear response, it was found that  $\beta=1/2$  leaving only  $\alpha$  to fit for, but this is only true for charged ligands
- Additional terms such as a surface area term have also been added requiring that all three coefficients be fit (but the surface area term is not really semiempirical)

# Empirical Methods

- There are many empirical scoring functions that have been developed (LUDI is one example)
- These methods do not use theoretically derived components but instead rely on structural descriptors which represent the physical interactions in forming the complex
- The weights for each of these components are calculated by extensive fitting to experimental data
- With sufficient fitting these potentials can function fairly well, but there is a limited amount of available data, and the diversity amongst the ligands is quite large

# Knowledge Based Potentials

- In place of deriving and fitting empirical or semiempirical potentials with experimental data, there has been significant development in what are termed “knowledge based potentials”
- These potentials have their basis in statistics and are derived from looking at the interatomic contact preferences of atoms in known structures
- The advantage of knowledge based potentials is that the statistical spread tends to make these “soft” potentials unlike those based on molecular mechanics
- This mimics the effect of flexibility without the added overhead of calculating multiple conformations and extensive sampling

# Knowledge Based Potentials

- Once a probability  $g_{ij}$  is calculated for a given pair of atoms  $i$  and  $j$ , the pseudo pair potential is simply

$$\Delta W_{ij} \sim -\ln g_{ij}/g_{\text{ref}}$$

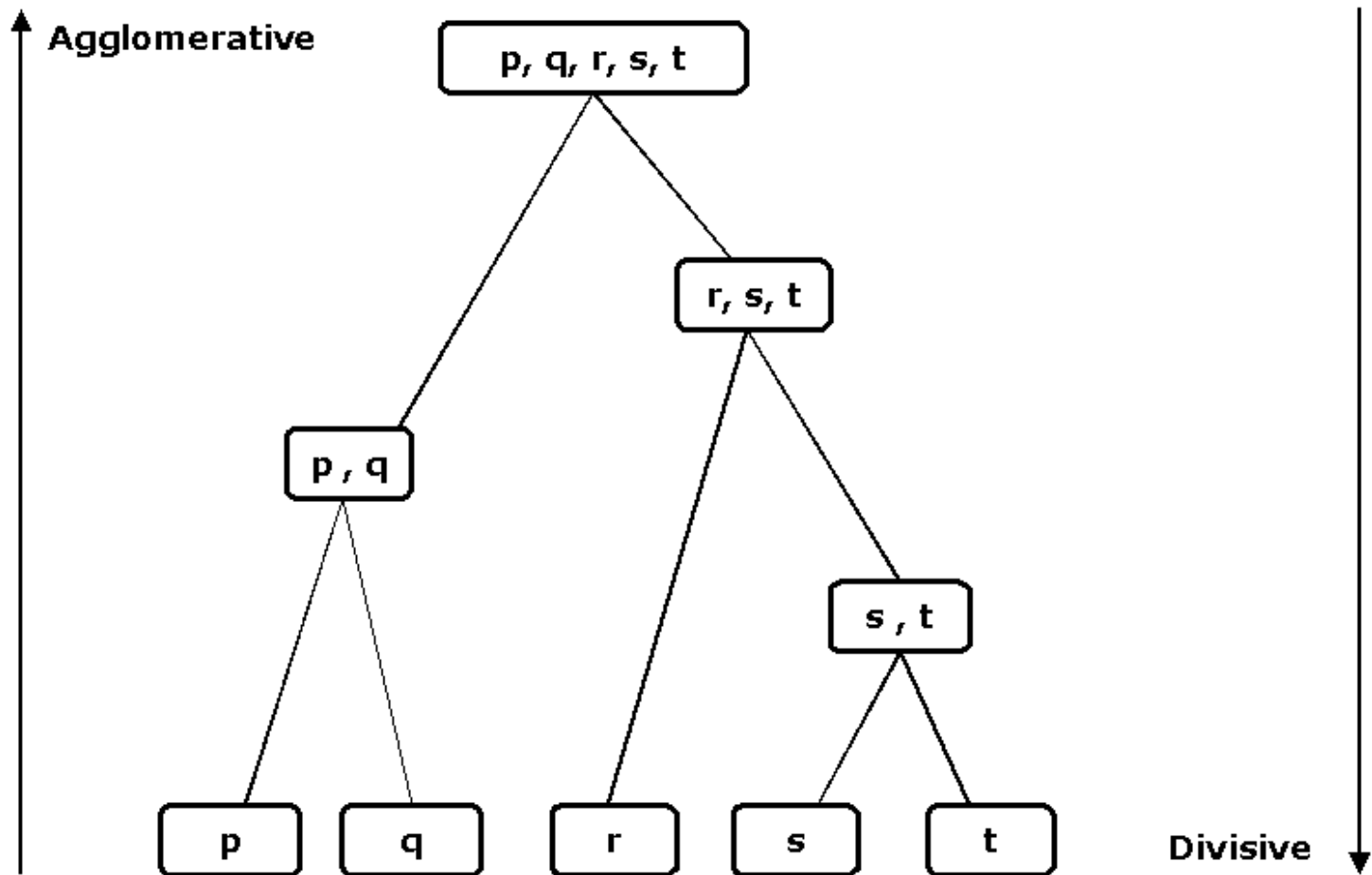
- Again a large amount of data is needed to get good statistics, but there is an abundance of structural data
- Using a training set of 697 structures, the correct binding affinity of a set of 77 structures was able to be predicted with fairly high correlation ( $r^2 = 0.61$ )
- Using family specific data, this correlation can be higher (between 0.7 and 0.8)
- Drugscore, one of the more successful scoring functions was derived this way using 1376 crystal structures and a basis set of 17 atom types (see Gohlke and Klebe, “Statistical potentials and scoring functions applied to protein-ligand binding”, *Curr. Opin. Struct. Biol.* (2001) **11**: 231-235)

# Clustering

- No docking algorithm can produce a single, trustworthy structure for the bound complex, but instead they produce an ensemble of predictions
- Each predicted structure has an associated energy, but we need to consider both the binding free energy (or enthalpy as the case may be) as well as the relative population
- By clustering our data based on some “distance” criteria, we can gain some sense of similarity between predictions
- The distance can be any of a number of similarity measures, but for 3D structures, RMSD is the standard choice

# Clustering

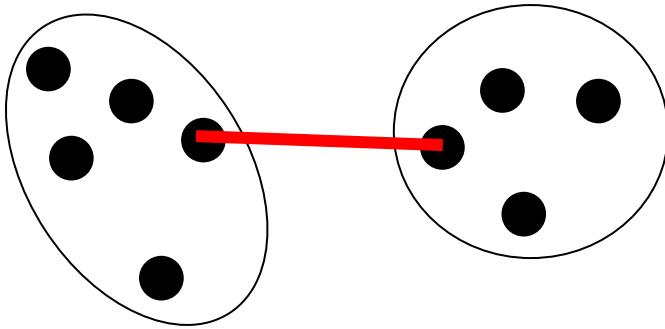
- One of the more commonly used methods is hierarchical clustering



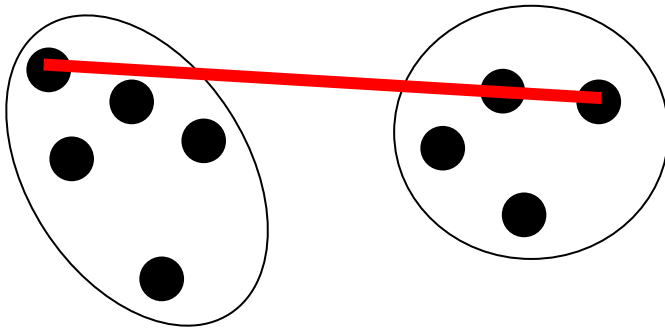
# Agglomerative Clustering

- In the agglomerative hierarchical clustering method, all structures begin as individual clusters, and the closest clusters are subsequently (iteratively) merged together
- There are again several different methods of measuring the distance between clusters
  - Simple linkage
  - Complete linkage
  - Group average

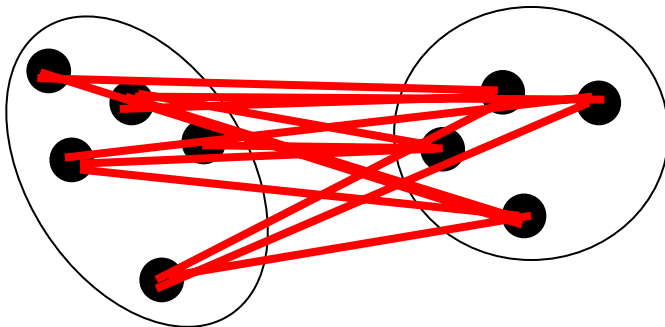
# Linkage Criterion



Simple linkage selects the distance as the gives the minimum distance between any two members



Complete linkage selects the maximum distance between any two members



Group average approaches vary in effectiveness depending on the nature of the data

# Protein-Protein Docking Methods

- The approaches to protein-protein docking have a lot in common with small molecule docking
- The methods are still based on the combination of search algorithm coupled to a scoring function
- The scoring functions are essentially the same (since we are still dealing with atomic interactions), however the major problem is that the conformational space we now need to search is massive

# FFT Methods

- A clever method developed by Katchalski-Katzir et al. uses FFT methods to search the translational and rotational degrees of freedom
- H. A. Gabb, R. M Jackson and M. Sternberg  
“Modelling protein docking using shape complementarity, electrostatics and biochemical information”  
J. Mol Biol. (1997) 272:106-120

# CAPRI

- Just like the CASP competition in the protein folding field, there is a bi-annual competition capped CAPRI: the **C**ritical **A**ssessment of **P**redicted **I**nteractions
- J. Janin et al.  
[“CAPRI: a Critical Assessment of Predicted Interactions”](#)  
Proteins (2003) 52:2-9
- Mendez et al.  
[“Assessment of blind predictions of protein-protein interactions: Current status of docking methods”](#)  
Proteins (2003) 52:51-67

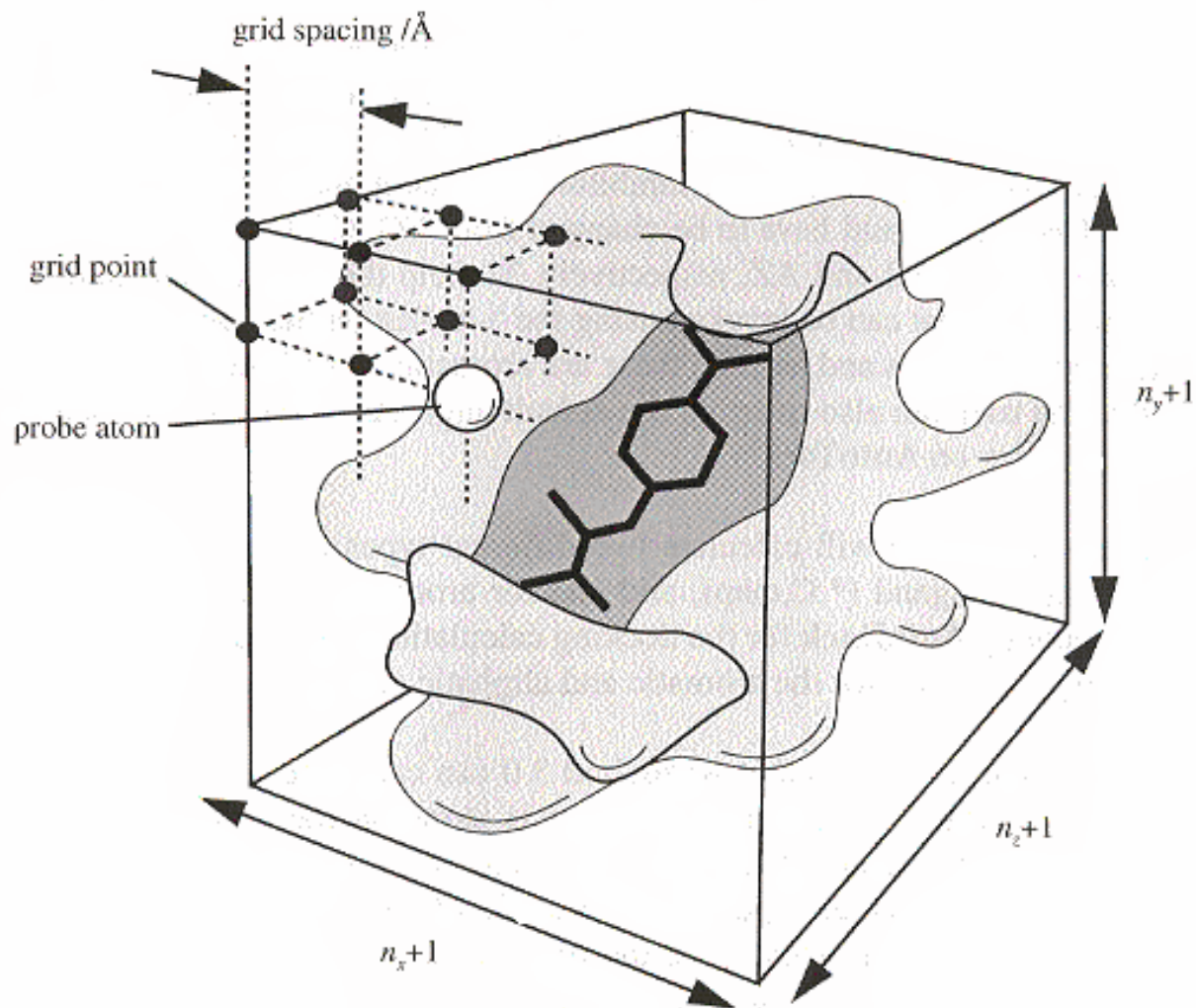
# Autodock

- Autodock uses pre-calculated affinity maps for each atom type in the substrate molecule, usually C, N, O and H, plus an electrostatic map
- These grids include energetic contributions from all the usual sources

$$\Delta G = \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right) + \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + \Delta G_{tor} + \sum_{i_C, j} S_i V_j e^{(-r_{ij}^2 / 2\sigma^2)}$$

# Grid Maps

Each type of atom is placed at each individual grid point and the change in free energy is calculated



# Other Energetic Notes

- Also includes torsional energy (if the ligand is flexible)
- Entropic contributions due to rotatable bonds
- Desolvation is calculated, but only for carbon atoms

# Search Methods

- Autodock has three search methods
  - Monte Carlo simulated annealing
  - A global genetic algorithm search
  - A local Solis and Wets search
- Combining the last two search methods gives the Lamarckian Genetic Algorithm which is what we will use

# Methodology

- Require structures for both ligand and macromolecule
- Add charges to both structures (create a *pdbq* file)
  - For proteins or polypeptides we use the Kollman united atom model
  - For drugs and other molecule, we use Gasteiger charges

# Methodology

- For the macromolecule, solvation needs to be calculated (producing a *pdbs* file)
- Based on the atom types in the ligand, the appropriate maps need to be calculated for the macromolecule
- Flexible bonds (if any) need to be assigned to the ligand

# Performing the Docking

- Once the ligand and maps are prepared, you perform docking by performing a series of runs (say 10 to 50)
- Each run proceeds until you reach a maximum number of energy evaluations (1 to 10 million) and the final structure is stored
- The results of all the runs are clustered based on RMSD and energy – the lowest energy structure forms the center of the first cluster