Simulating Receptor-Flexible Small Molecule binding Using AutoDock

S. Ravichandran, Ph.D.
Advanced Biomedical Computing Center (ABCC)
National Cancer InstituteFrederic k
Frederick, MD 21702

10/27/2005

Email: sravi@ncifcrf.gov
Tel: 3018461991
http://nciiris.ncifcrf.gov/~ravichas/docking

ABCC Advanced Biomedical Computing Center
A Center Devoted To Biocomputing

Supercomputing facility

http://www.abcc.ncifcrf.gov

What do we do?

Consultation, training, Research ...

Biomedical Research Groups at ABCC

QM, Molecular Modeling, Bioinformatics, Structural biology

What I will not talk about:

Manual Docking

Incremental Constructs Method
- Docking of fragments

Virtual Docking

Free Energy Methods

- Docking methods always discuss Free Energy
  - Stability is related to F.E
  - Helmholtz (NVT) and Gibbs (NPT)

Goals of Docking

Fitting a small molecular (drug molecule) into a protein

-> Identifying correct posing within the binding site

-> Calculate Activity

Things U need

1) Structures of protein/small molecule

Molecule	Databases	Method
Protein/DNA	PDB	x-ray, NMR
Small Molecules	CSD	x-ray

2) Software:

Program to do docking: Dock, AutoDock Ludi, FlexX etc.

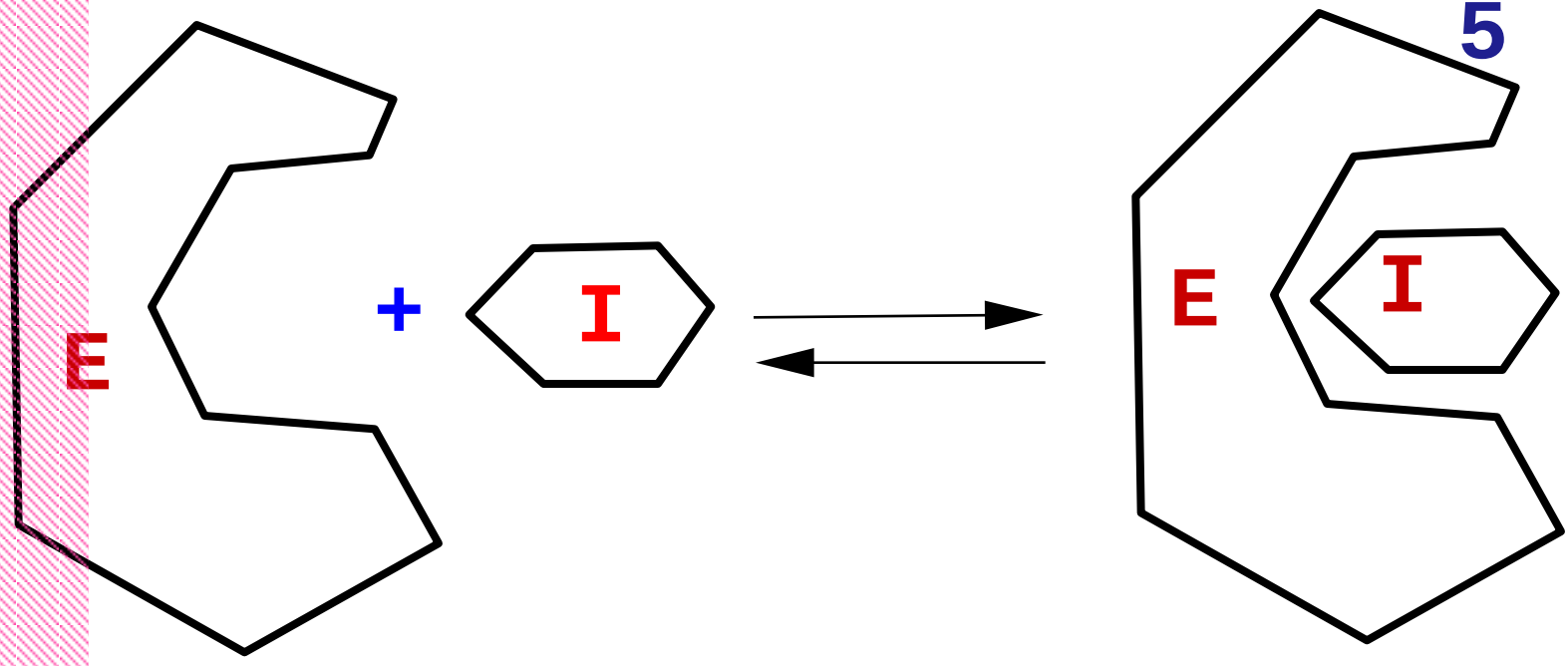
3) Powerful Computers: CPU, disk-space etc.

?s to ask before docking:

- Critical analysis of the protein structures
  Resolution, missing residues,
  High thermal factors ...
  X-ray
  NMR Which model to choose?

- What to do with bound ions, water molecules?
  Remove them? Keep them, Why?

- Crystal structure – Rigid nature–bound ions and water are held in the active site. Are they valid configurations?



$$[E]_{aq} + [I]_{aq} \rightleftharpoons [E-I]_{aq}$$

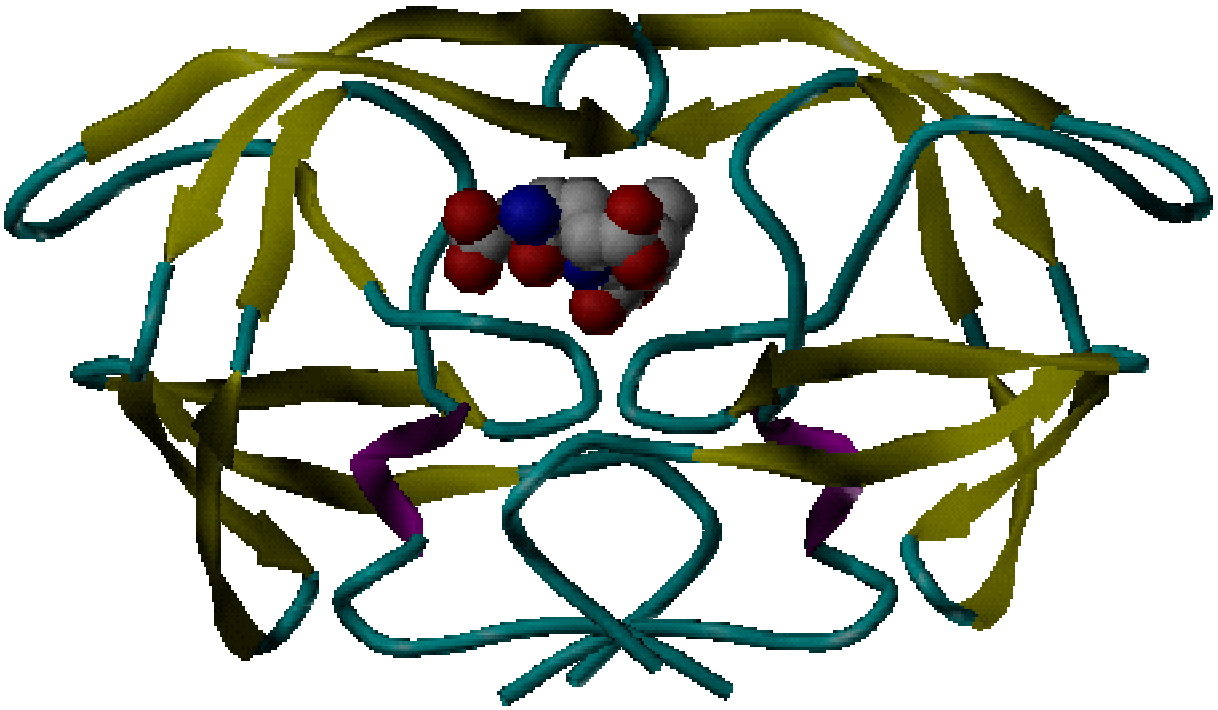
Loosely bound complex

The docking activity gives the proteins the ability to promote and inhibit(or accelerate or prevent) certain chemical reactions.

Importance:

Designing bioactive compounds,
Computer-Aided Drug design
.....
.....
..... and many more

Applications: Drug Design



HIV-Protease complex with tripeptide inhibitor

Things we know:

a) HIV protease: enzyme in the AIDS virus, important for its replication

b) Chemical reaction takes place in protease at an active site

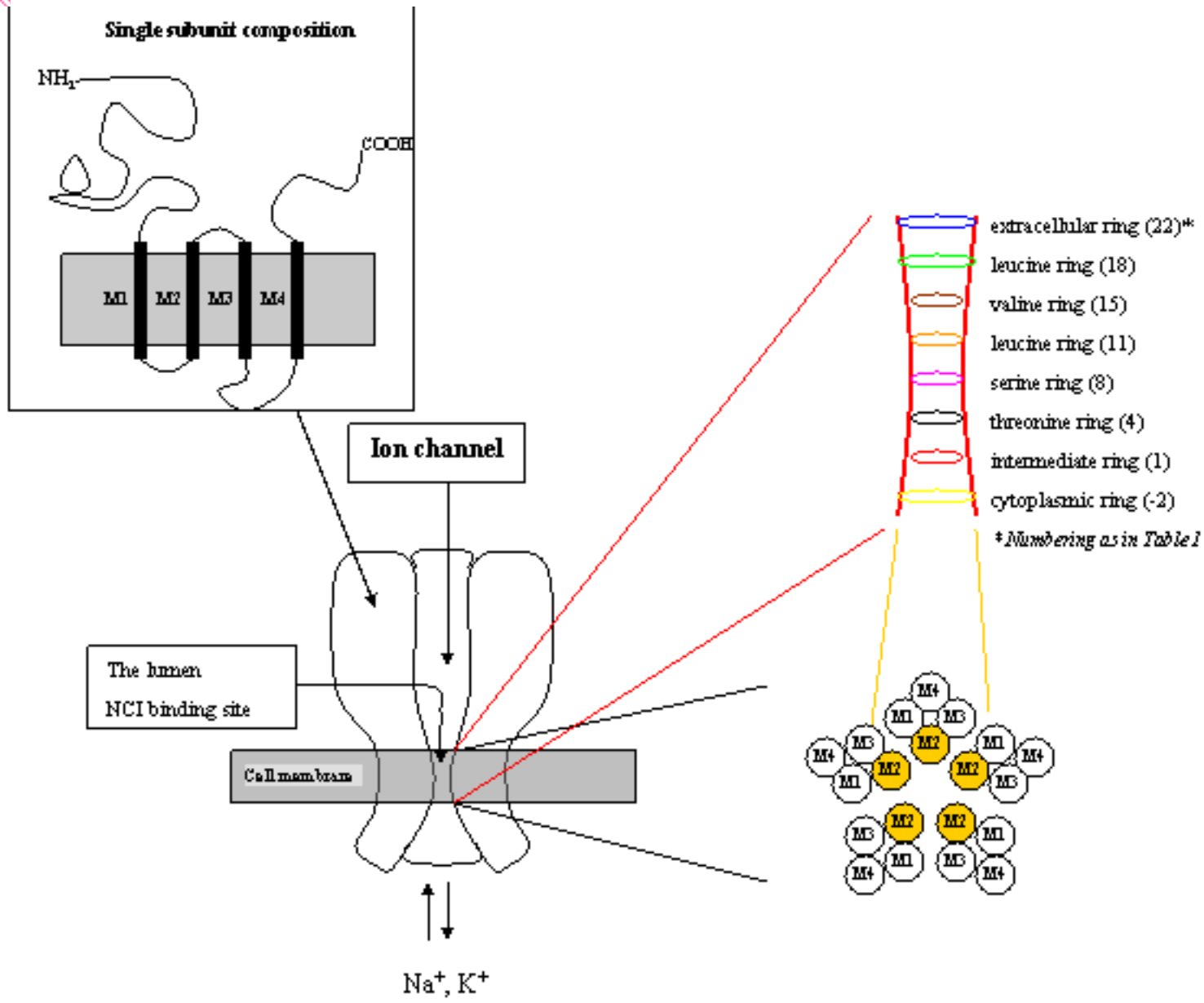
c) Inhibitor drugs bind to the active site and block the functioning

?s we ask:

a) Predicting Binding affinity of the drug

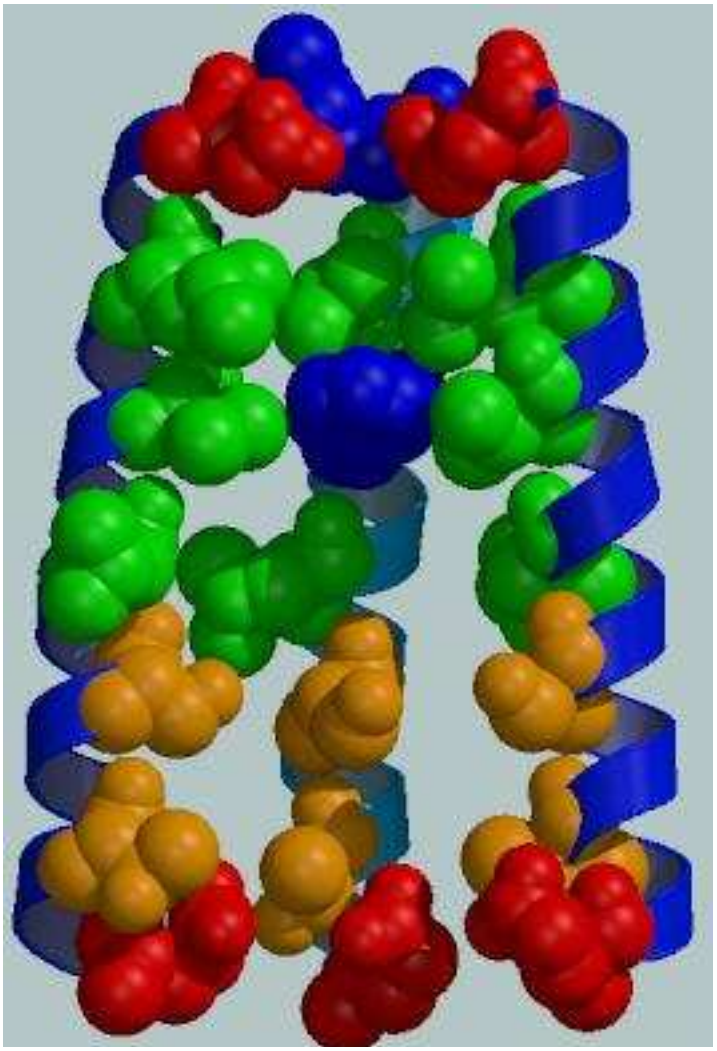
b) What has to be changed for better binding

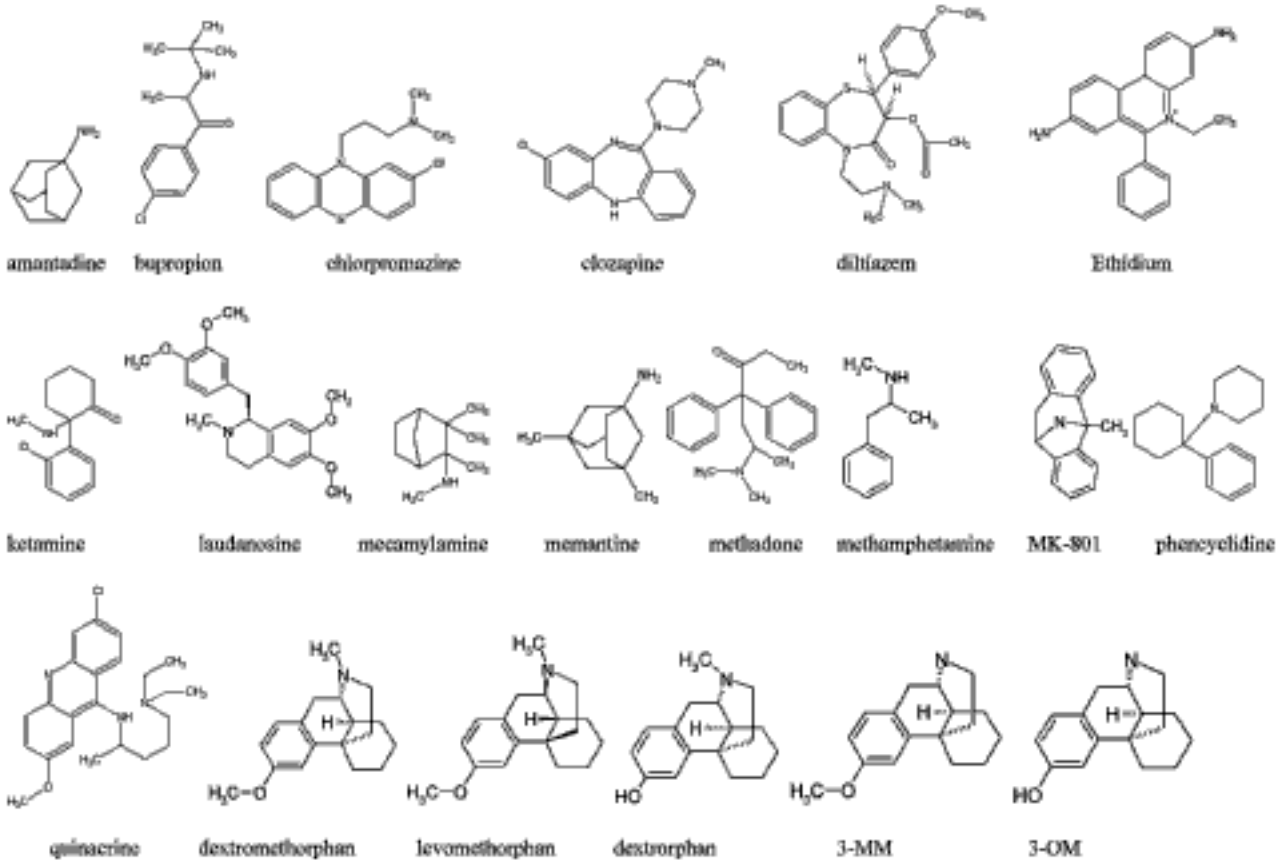
Noncompetitive inhibition of nAChR receptor



Jozwiak K, Ravichandran S, Collins JR and Wainer IW, J. Med. Chem. 2004, Jul 29; 47(16): 4008-21

Jozwiak K, Ravichandran S, Collins JR and Wainer IW, J. Med. Chem. 2004, Jul 29; 47(16): 4008-21





S.Ravichandran, Ph.D., ABCC, NCI-Frederick, MD; sravi@ncifcrf.gov

Problem:

To find molecular binding sites (hot-spots) by computer

Why is it difficult?

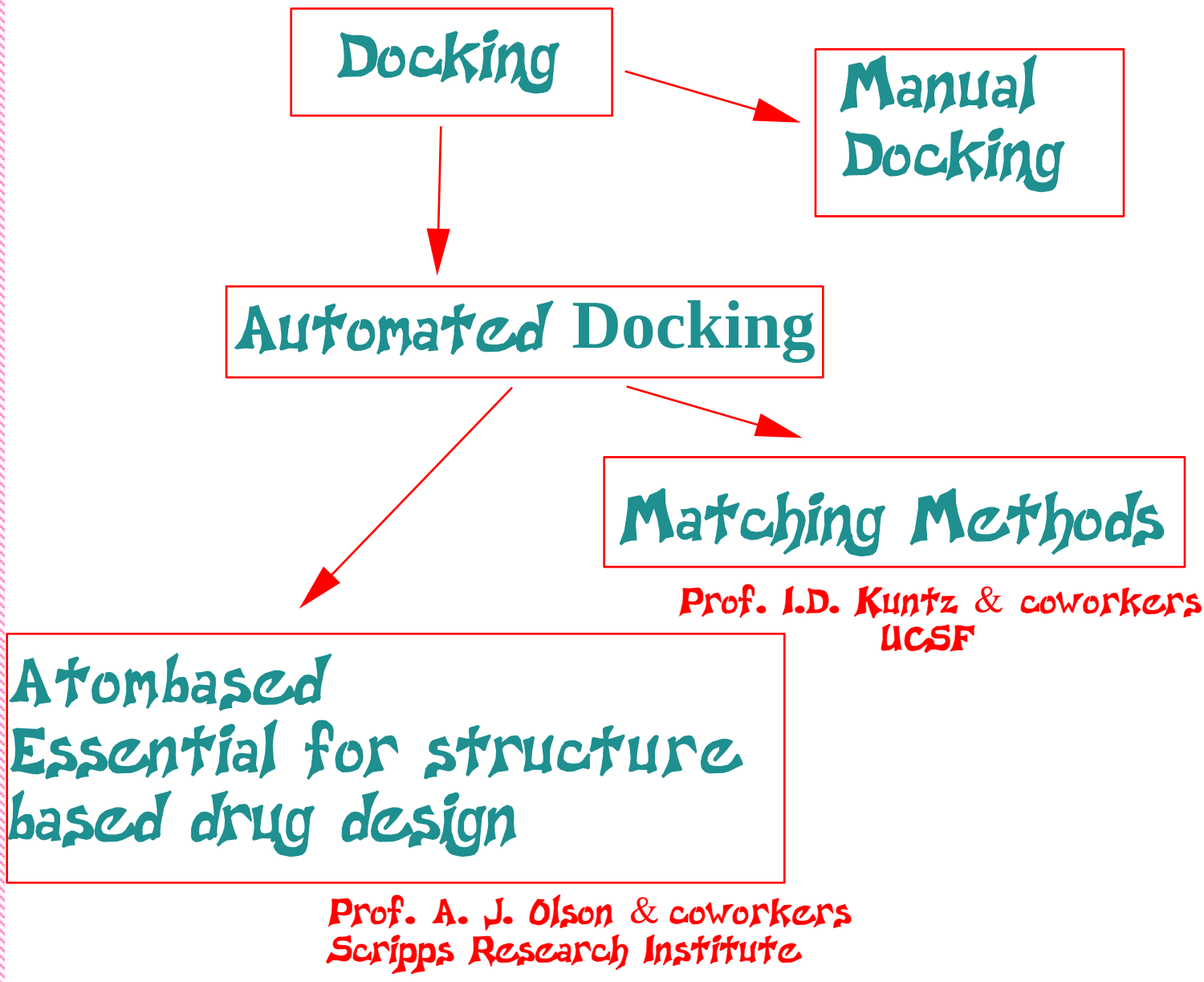
2 Rigid molecules

6 degrees of freedom
(3-rotational, 3 translational)

In addition if you assume the small drug sized molecules to be flexible (say 14 degrees of free dom)

1028 variations

Even difficult with a SUPER COMPUTER!!!!



Atom-based tools are usually slower but better becos of the assumed flexibility of the ligands.

AutoDock

Free for Academics

1990: Drs. David S. Goodsell;
Arthur J. Olson;
Garrett M. Morris;
Ruth Huey; Scott Halliday
RikBelew...

Scripps Research Institute
La Jolla, CA

Orginal Version: f77
Currently: C++
(ver 3.0.5)

AutoDockTools (ADT)

Free for Academics

Michel Sanner & Ruth Huey
Scripps Resarch Institute

GUI to prepare and submit jobs for AutoDock
Code written in Python

Docking Space

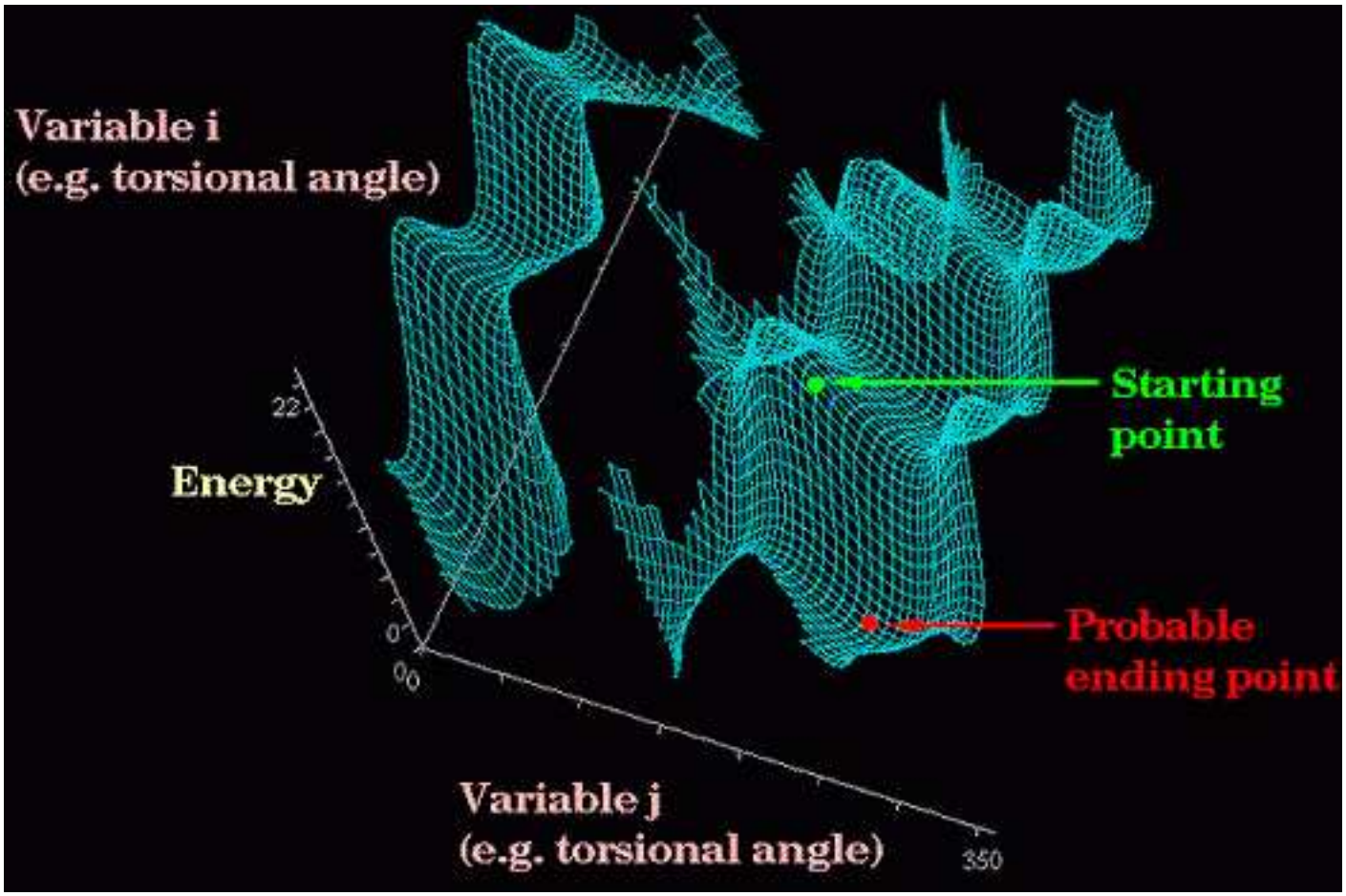


Image taken from NIH MM modeling Page

Stages of Docking

Identifying the Pose (structural Modeling) uses scoring

Ranking uses scoring

Docking Problem:

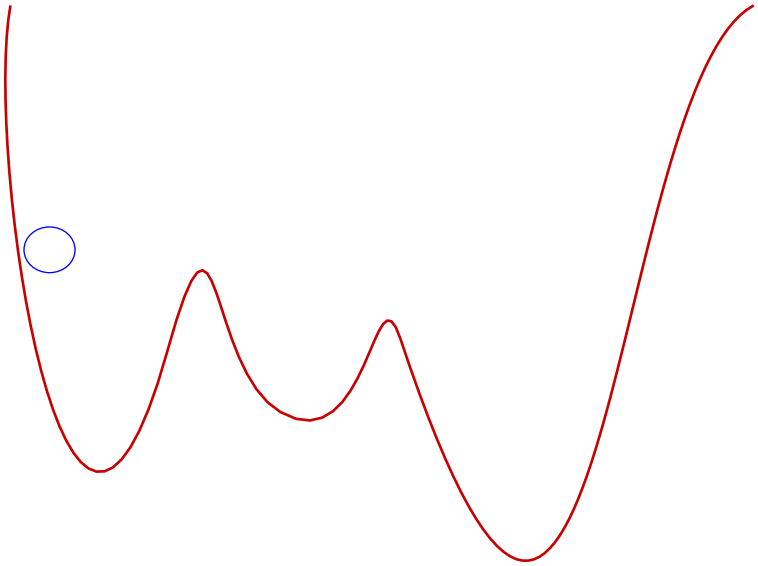
- **Optimization**
  Global minimum or maximimum of a function with following properties:

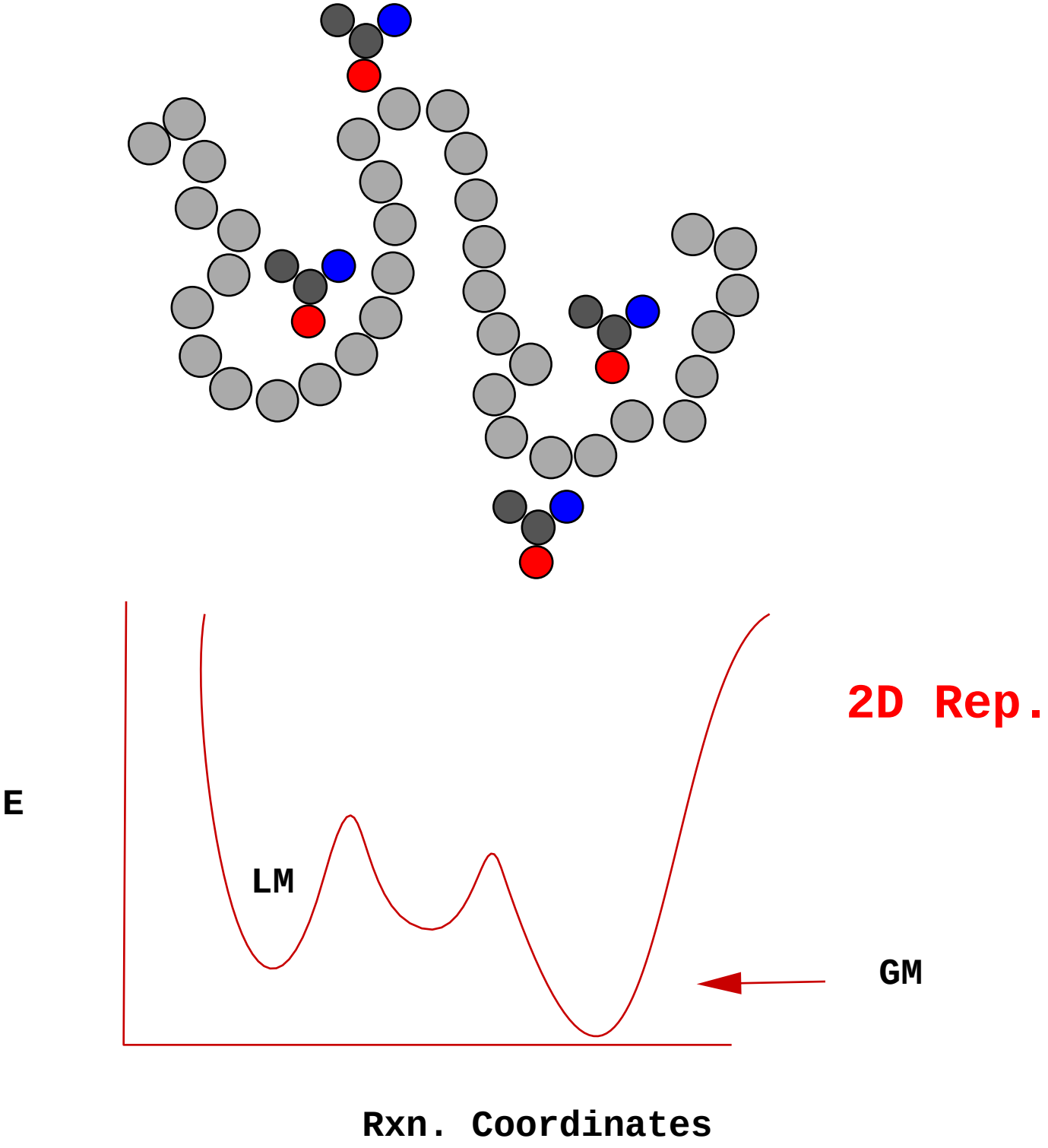
  Continuous function
  Domain Range

Search techniques:

Local: Operation which iteratively improves its estimate of minimum by searching for better solutions in a local neighbourhood of current solution

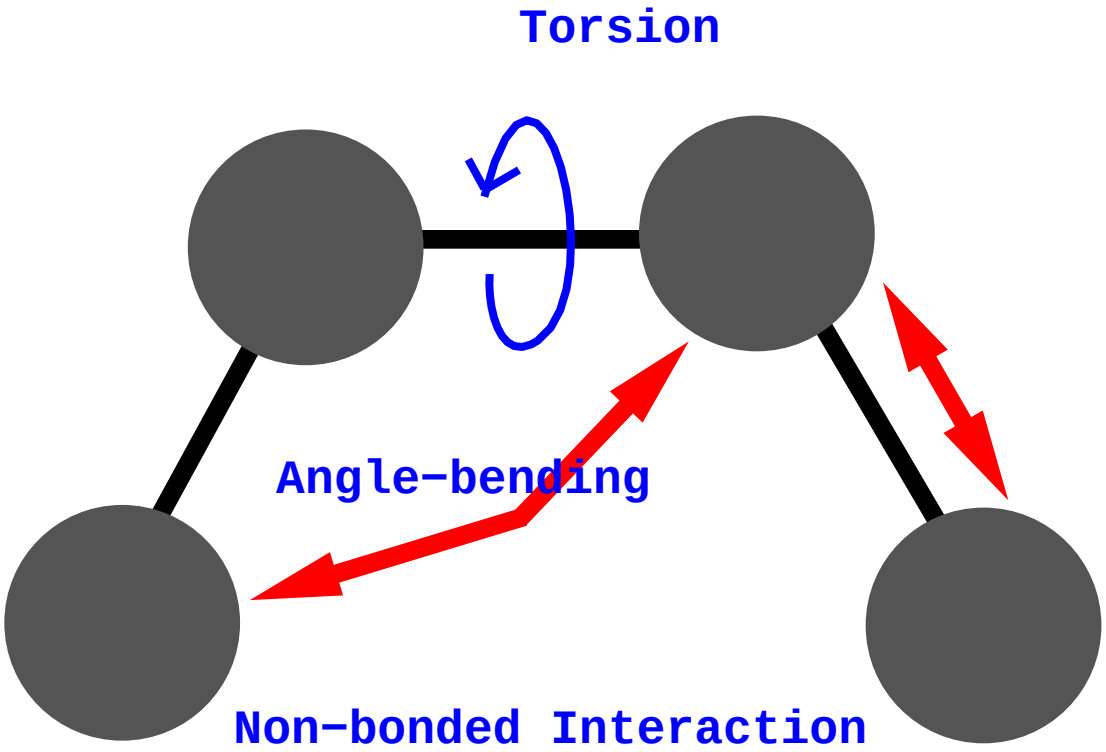
Global: Will perform a sophisticated search across several multiple local minimum




2D Rep.
E
LM
GM
Rxn. Coordinates

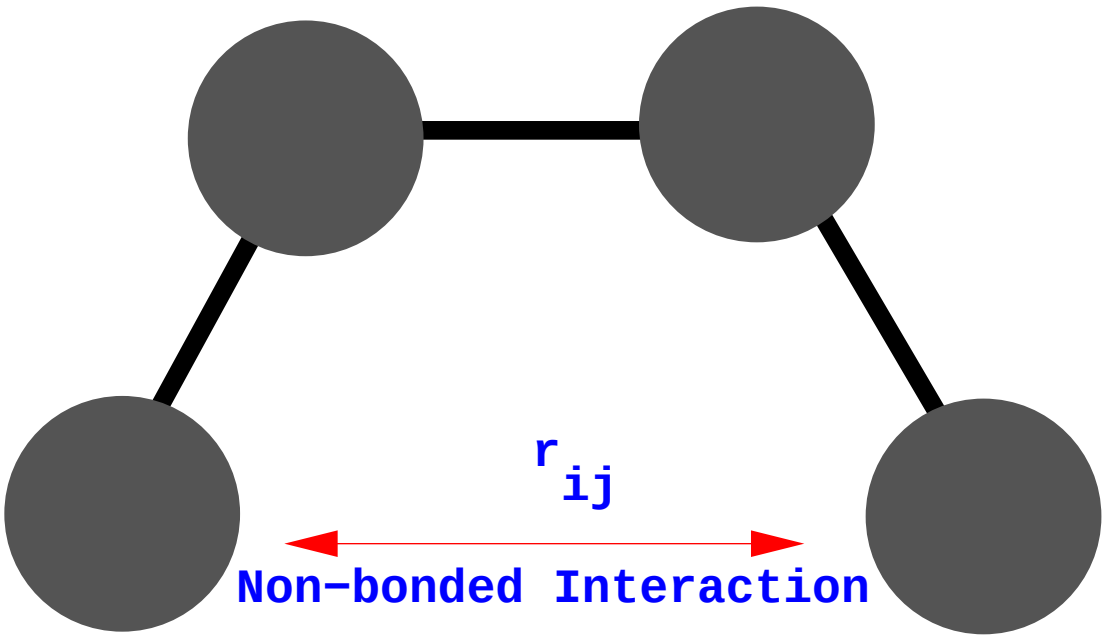
Force Field

FF is the functional form for the independent energy terms + parameters



$$E_{pot} = \Sigma 1/2K_b(b - b_0)^2 + \Sigma 1/2K_\theta(\theta - \theta_0)^2 + \Sigma 1/2K_\phi(1 + \mathrm{Cos}N\phi)^2 + \Sigma 1/2K_\chi(\chi - \chi_0)^2 \Sigma((B/r)^{12} - (A/r)^6) + \Sigma(qq/r)$$

Given a Conformation ---FF---> Energy



$$E = \sum_i \sum_j \frac{-A_{ij}}{r_{ij}^{6}} + \frac{B_{ij}}{r_{ij}^{12}} + \sum_i \sum_j \frac{q_i q_j}{r_{ij}}$$

VDW Electrostatic

$$\text{Energy} = V_{vdw} + V_{coul} + V_{hb}$$

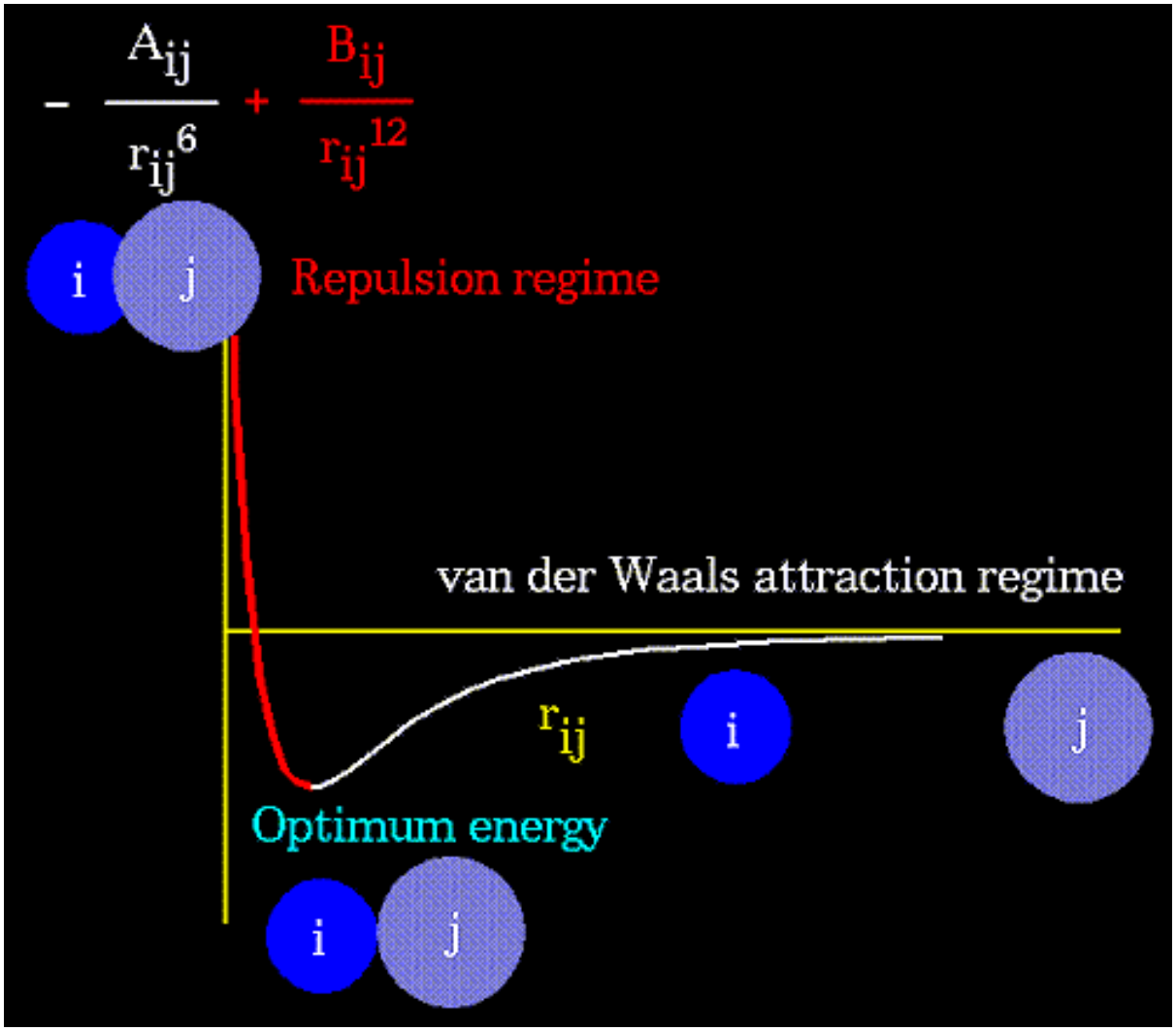
$$V_{vdw} = \sum_{i,j} 4\varepsilon_{ij} \left[(A_{ij}/r_{ij})^{12} - (B_{ij}/r_{ij})^{6} \right]$$

$$V_{coul} = \sum_{i,j} (q_i q_j)/[4\pi\varepsilon(r)\varepsilon_0 r_{ij})$$

Pairwise Additive

(Const dielectric or distance dependent)

H- bonding 12-10 form is used



Scoring:

Identifying the good from the bad

Several Approaches:

- Force-Field
- Empirical
- Knowledge-based

$$\Delta G = -RT \ln K_{eq}$$
$$= -RT \ln (k_{on}/k_{off})$$
$$= \Delta H - T\Delta S$$

Empirical relationship between molecular structure and binding free energy

$$\Delta G = K_{vdw} * V_{vdw} + K_{hb} * V_{hb} + K_{ele} * V_{coul} + K_{tor} * V_{tor} + K_{sol} * V_{sol}$$

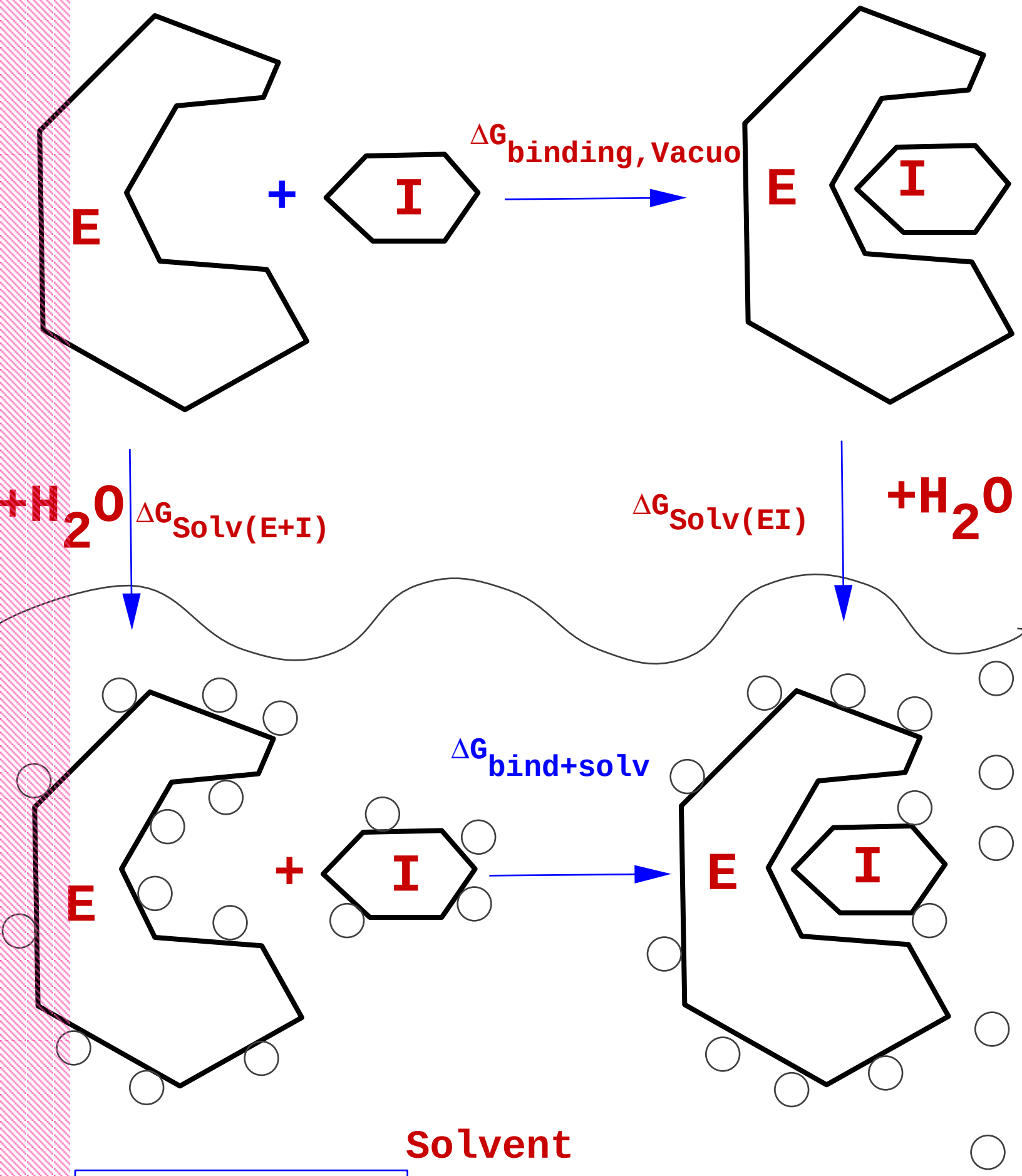
Coefficients are empirically determined using linear regression analysis from a set of Protein-ligand complexes with known binding constants.

Protein–ligand complex	PDB code	Log(K_i)[a]
Concanavalin A / α-methyl-D-mannopyranoside	4cna	2.00
Carboxypeptidase A / glycyl-L-tyrosine	3cpa	3.88
Carboxypeptidase A / phosphonate ZAA=P=(O)F	6cpa	11.52
Cytochrome P-450$_{cam}$ / camphor	2cpp	6.07
Dihydrofolate reductase / methotrexate	4dfr	9.70
α-Thrombin / benzamidine	1dwb	2.92
Endothiapepsin / H-256	zer6	7.22
ε-Thrombin / MQPA	1etr	7.40
ε-Thrombin / NAPAP	1ets	8.52
ε-Thrombin / 4-TAPAP	1ett	6.19
FK506-binding protein (FKBP) / immunosuppressant FK506	1fkf	9.70
D-Galactose / D-glucose binding protein / galactose	2gbp	7.60
Hemagglutinin / sialic acid	4hmg	2.55
HIV-1 Protease / A78791	1hvj	10.46
HIV-1 Protease / MVT101	4hvp	6.15
HIV-1 Protease / acylpepstatine	5hvp	5.96
HIV-1 Protease / XK263	1hvr	9.51
Fatty-acid-binding protein / $C_{15}COOH$	2ifb	5.43
Myoglobin (ferric) / imidazole	1mbi	1.88
McPC603 / phosphocholine	2mcp	5.23
β-Trypsin / benzamidine	3ptb	4.74
Retinol-binding protein / retinol	1rbp	6.72
Thermolysin / Leu-hydroxylamine	4tln	3.72
Thermolysin / phosphoramidon	1tlp	7.55
Thermolysin / *n*-(1-carboxy-3-phenylpropyl)-Leu-Trp	1tmn	7.30
Thermolysin / Cbz-Phe-*p*-Leu-Ala (ZFpLA)	4tmn	10.19
Thermolysin / Cbz-Gly-*p*-Leu-Leu (ZGpLL)	5tmn	8.04
Purine nucleoside phosphorylase (PNP) / guanine	1ulb	5.30
Xylose isomerase / CB3717	2xis	5.82
Triose phosphate isomerase (TIM) / 2-phosphoglycolic acid (PGA)	2ypi	4.82

[a] Adapted from Böhm.[54]

Table from
Morris *et al*, *J. Comp Chem*. 19 (14) 1639–1662 (1998)

Thermodynamic Cycle of E & I binding in both *Vacuo* and solvated phases



S.Ravichandran, ABCC
NCI, sravi@ncifcrf.gov

(Representation based on AutoDock Manual)

Flexibility

Ligand:

Systematic Search

(6 rotatable bonds with 30° increments will result in 2,985,984 confs.)

Random Search

Simulated Annealing, Genetic Algorithm

Simulation Method

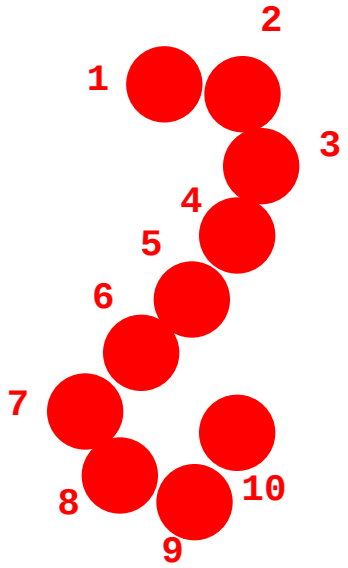
MD

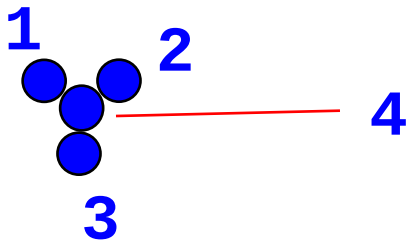
Protein:

Molecular Dynamics

Ensemble of Protein Structures–Docking

Why GRIDS are useful in Docking ?





Model Ligand
(NL = 4)

Model Receptor
(NR = 10)

Energy Calculation:

11 12 13 14 15 16 17 18 19 110

21 22 23 24 25 26 27 28 29 210

31 32 33 34 35 36 37 38 39 310

41 42 43 44 45 46 47 48 49 410

Total Terms = NL * NR

To deal at each simulation step

Example: Biotin has 19 atoms
Receptors of the range of 1000 or more atoms

(1000 X 19 = 19000) terms !!!!

AutoDock we will be doing million steps on GA

so, 19000 * Millon Interactions per Simulation

?

How to avoid this Computational Heaviness without losing information?

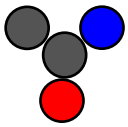
3-D GRID maps with precalculated interaction energies between each ligand atom-type and the receptor (ex.C-map etc.)

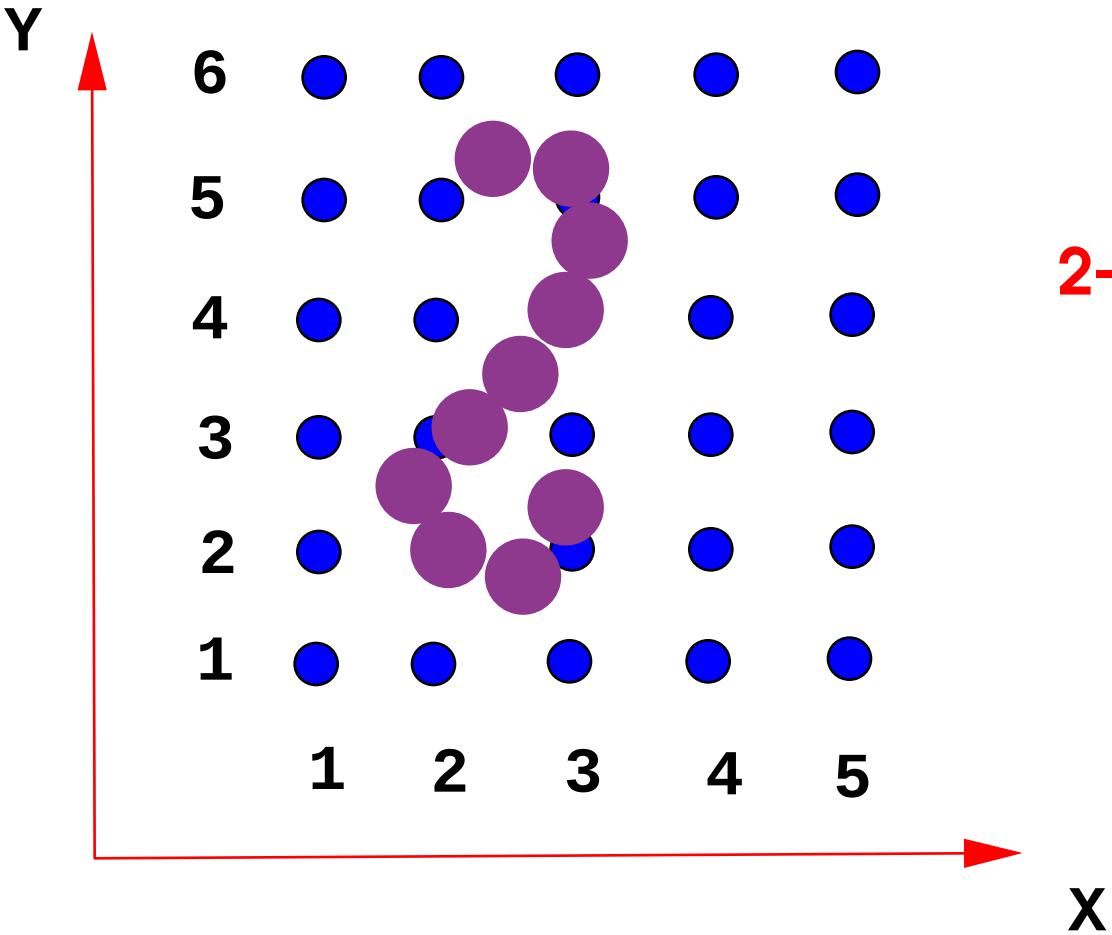
These grids are like LOOKUP-TABLES and used to look-up the interaction energies during docking simulation(s).

The Grid Method makes the energy calculations independent of the size (number) of the Receptor!!

That is all OK, What is a GRID?

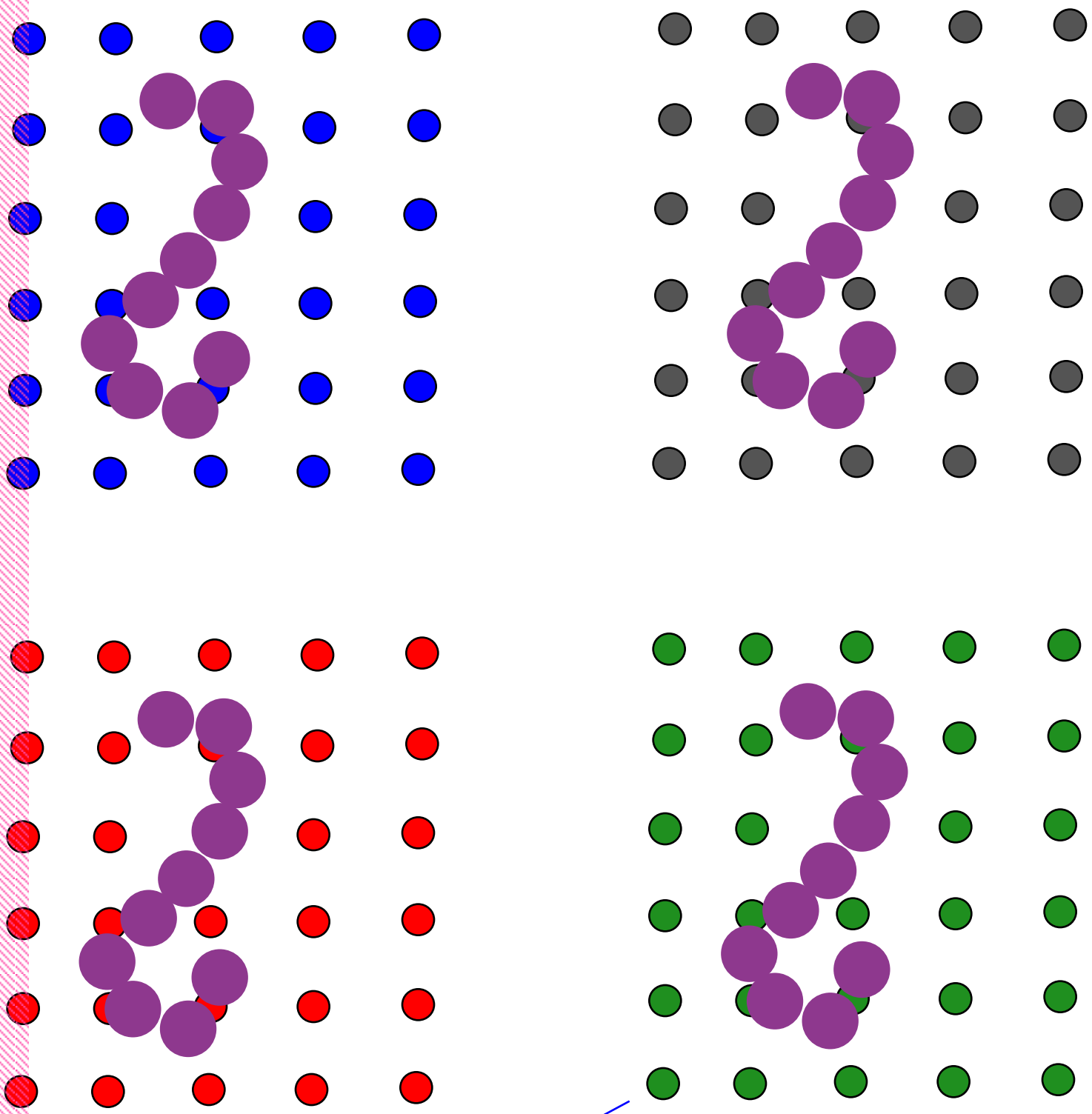
Imagine the ligand has 4 atoms of 3 atom types





2–D Rep.

Possible Maps for the model Ligand



(or one can do a PB calculation to calculate ϕ)

Electron MAP If you want to study Electrostatics?

S.Ravichandran, Ph.D., ABCC, NCI–Frederick, MD; sravi@ncifcrf.gov

Automated Docking

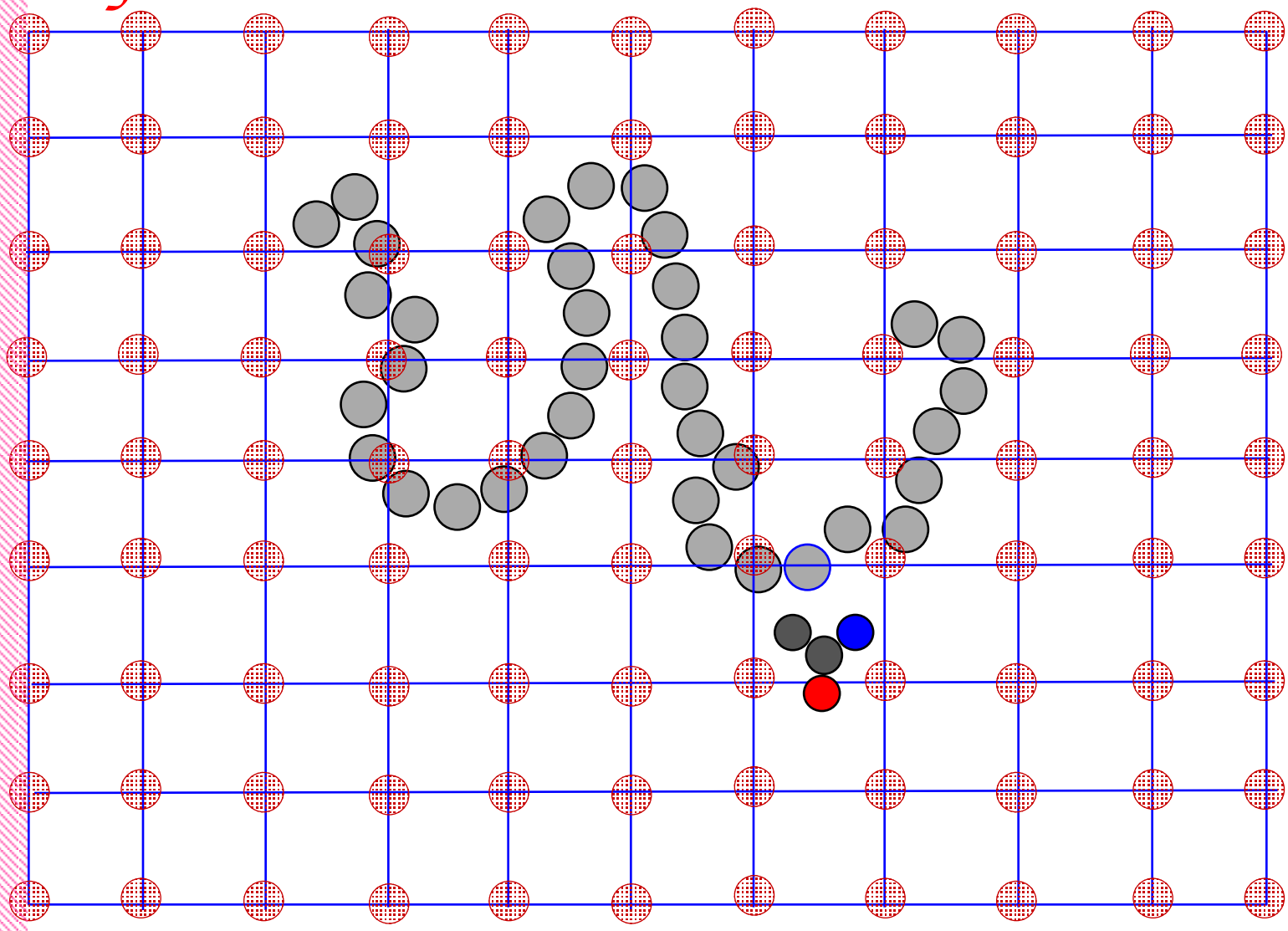
Conflicting Requirements

a) desire for a robust and physically relevant procedure

b) Computation demands at a reasonable level

Static Receptor – Mobile small drug molecule

Grid–based model



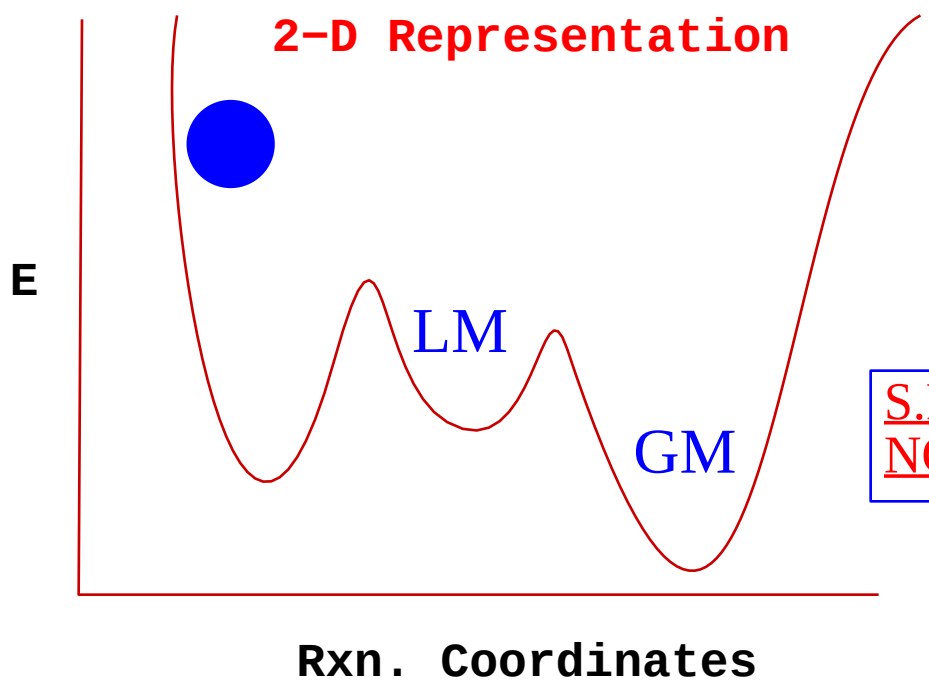
Methods of docking:

a) Simulated Annealing:

b) Genetic Algorithms:

Annealing is a process in which temperature of the substance is reduced (slowly) until the material crystallizes in a single crystal (usually corresponds to global minimum free-energy)

Simulated Annealing: Computational method of mimicking annealing.
SA can be used to do both global and local search. Global at high Temperature and local at low temperatures.



S.Ravichandran, ABCC
NCI, sravi@ncifcrf.gov

Simulated Annealing

Monte Carlo

$T_i = (gT_{i-1})$

Cycle

Random changes to ligand position, orientation and/or conformation

Compute new E_1

Is E_1 is lower than E_0?

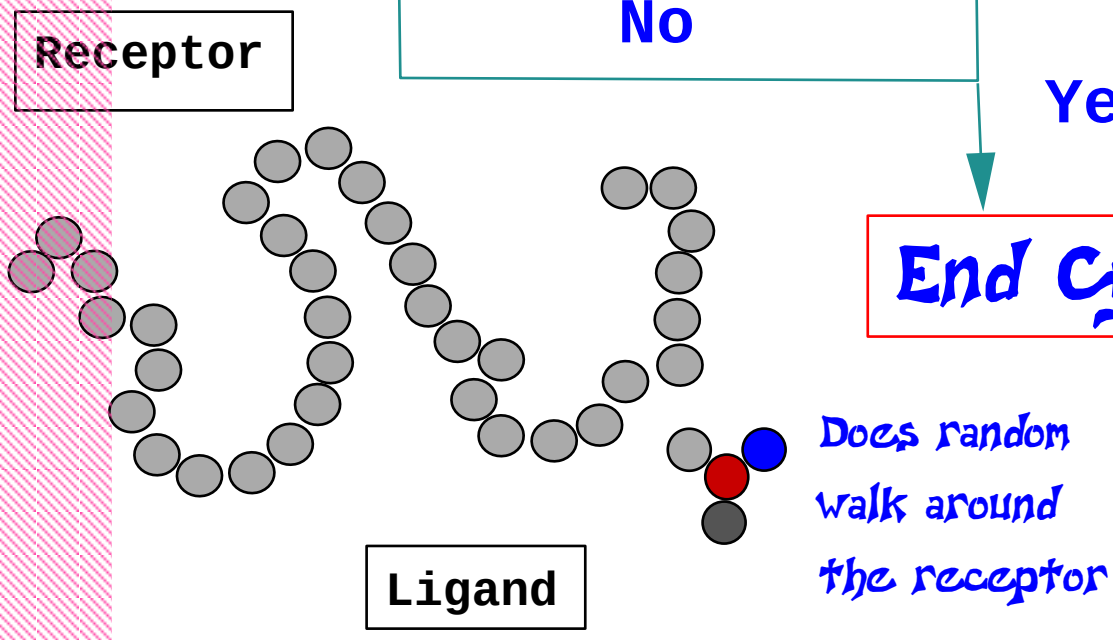
Yes → Accept the changes

No → accepted with prob $e^{(-\Delta E/k_B T)}$

Is the number of acceptances or rejection are within the chosen number?

No → Random changes to ligand position, orientation and/or conformation

Yes → End Cycle



Evolutionary Computing

I. Rachenberg 1960

Genetic Algorithm

(ideas taken from Natural Genetics & Biological Evolution)

Darwin, John Holland, 1975
John Koza, 1992.....

Genetic Programming

Living Organism

CELL

Same set of Chromosomes (strings of DNA)

Serves as a model for the whole organism

Genes--> Protein
--> Trait

Genetic Algorithms are usually used to carry out Global Search.

Hybrid Genetic Algorithm

How big?

Set of solutions (Populations or Chromosomes)

What are chromosomes?

Generate population of n chromosomes

What is fitness?

Evaluate the fitness f(xi)

Crossover how often?
Mutation how often?
Elitsm? Selection?

Create new population

Elitsm *Selection*
Crossover *Mutation*

No

Is Termination Condition (Max # of Energy Eval. or Generations) satisfied?

Yes

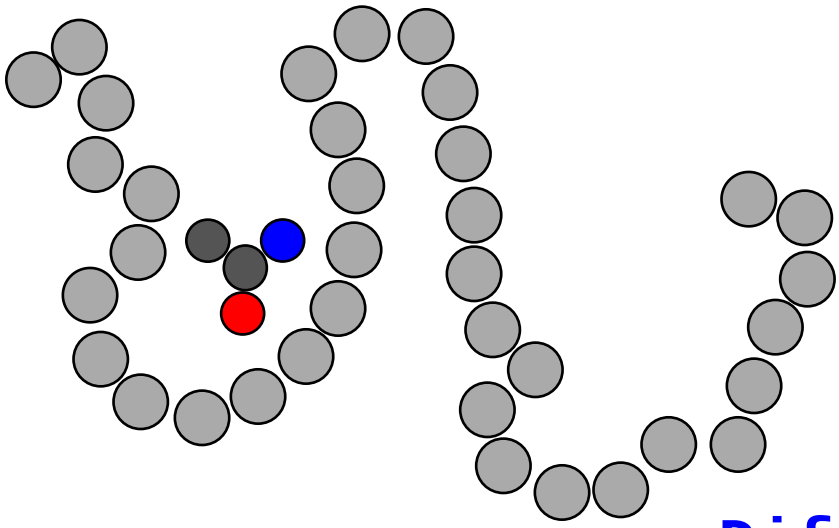
STOP

No Explicit Termination conditions

S.Ravichandran, ABCC
NCI, sravi@ncifcrf.gov

Chromosome 1

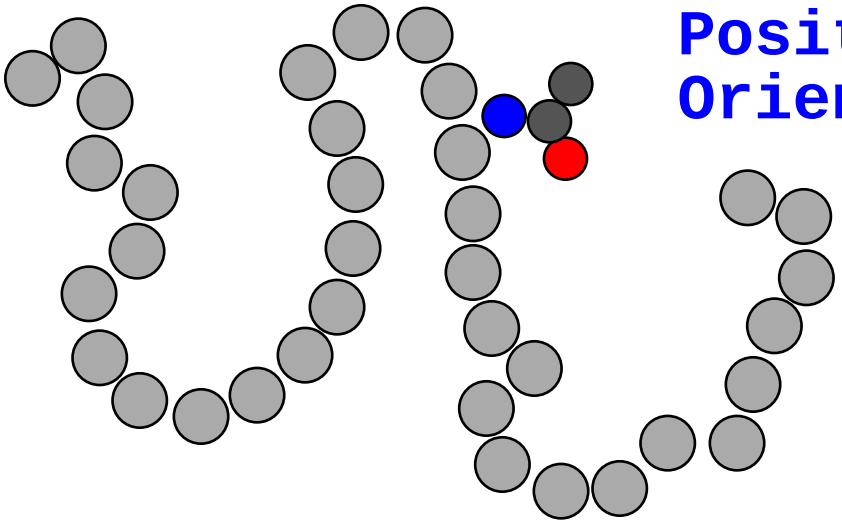
E1



Different Position & Orientation

Chromosome 2

E2



and so on .....

If E1 is lower than E2 probably 1 will eventuvally survive compared to 2

Classical GA represent genome as a fixedlength bit string

x	y	z	q_0	q_1	q_2	q_3	τ_1	τ_2	τ_3	τ_4	τ_5

Transl.Gene **Rot. Gene** **Torsion Gene**

Generation

Mapping & fitness *translates genotypes–phenotypes to calculate E*

Fitness *Which individuals will reproduce based on worst energy individuval*

Crossover *2–point crossover*

Parent 1

+

Parent 2

=

Offspring

Mutation *random changes to variables using Cauchy distribution*

Elitism *How many top individuvals Survive into next generation?*

S.Ravichandran, ABCC
NCI, sravi@ncifcrf.gov

Crossover

2–point crossover

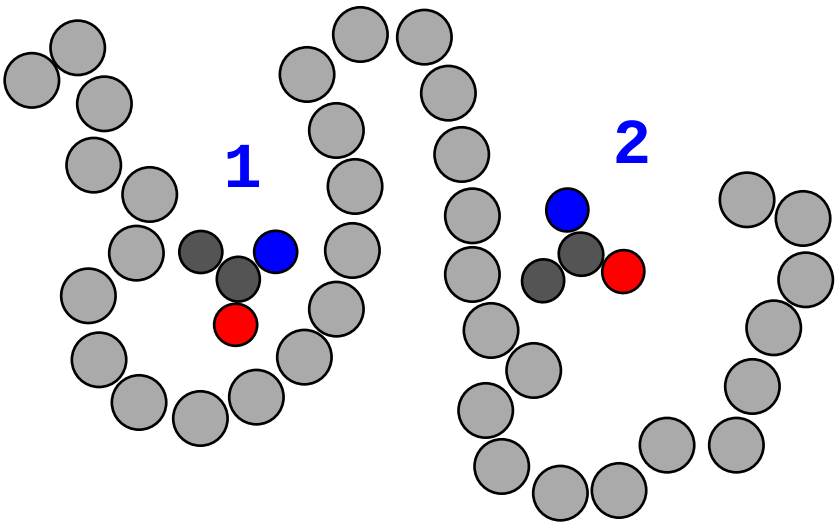
Parent 1

+

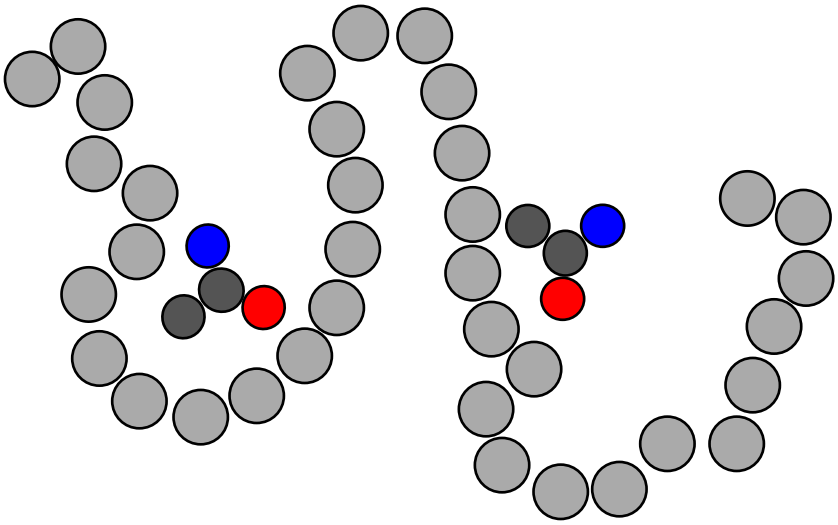
Parent 2

=

Offspring



2 offsprings

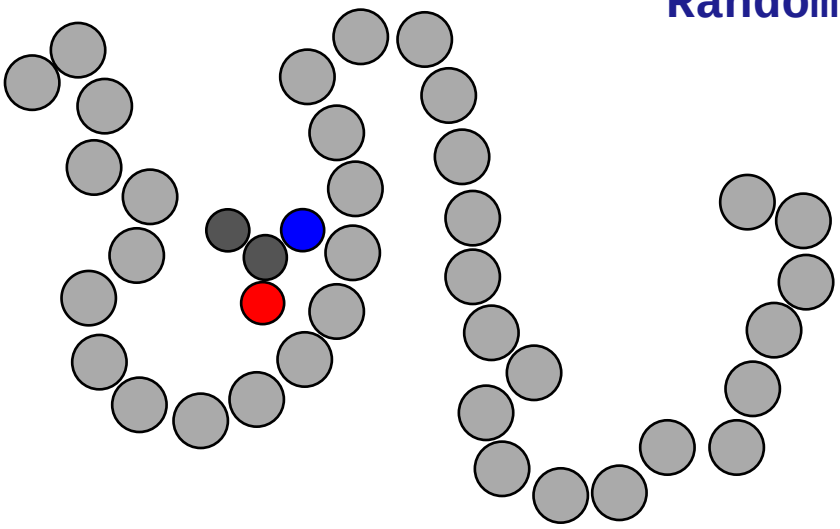


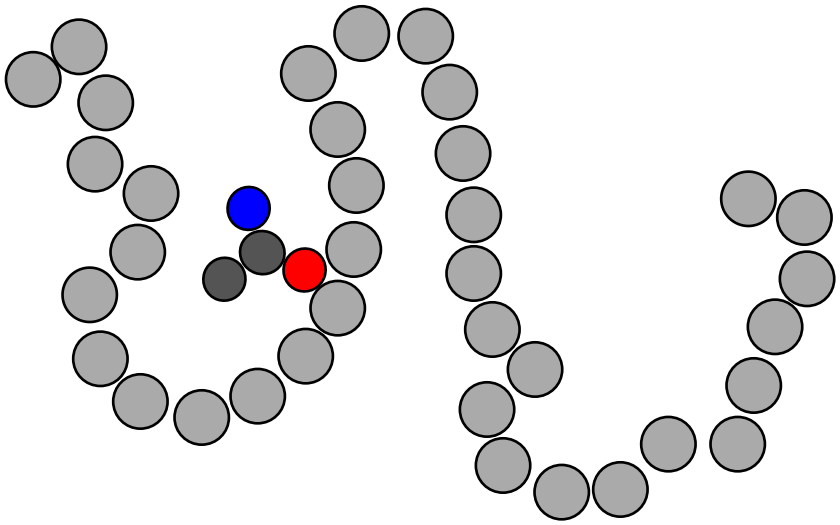
Combines the features from both parents

Mutation

One Parent produces one child

Random Perturbations

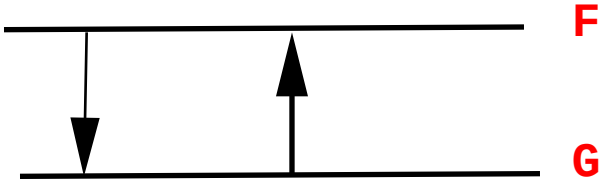




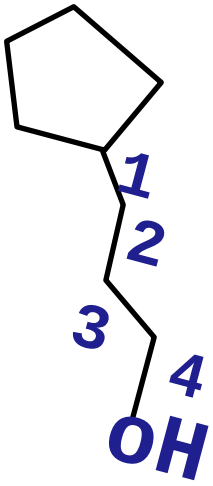
Lamarckian GA

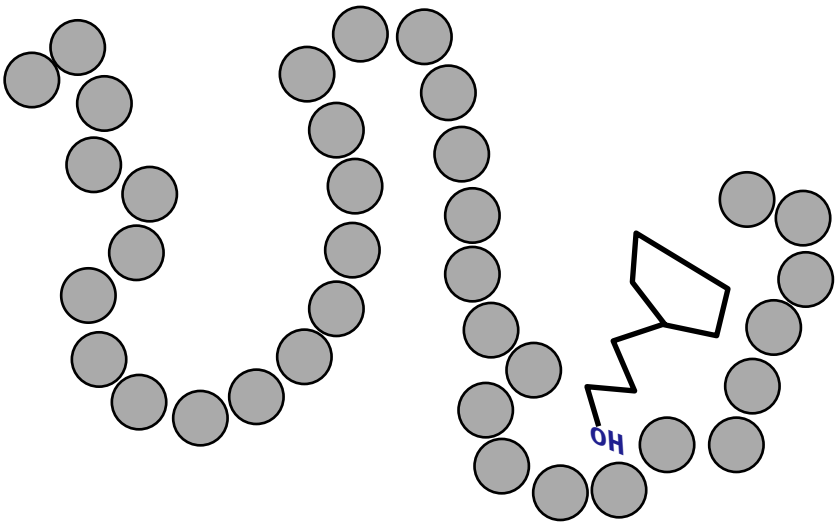
Environmental adaptations of an individual's phenotypic characteristics acquired during lifetime can become heritable traits (genotype)

Lamarck, J.B. Philosphie Zoologique, Flammarion Ed. Paris (94)



Phenotype





Genotype

x	y	z	q_0	q_1	q_2	q_3	τ_1	τ_2	τ_3	τ_4

Issues with GA:

1) Premature convergence

2) Too long for convergence
   How many runs?
   How many energy evaluations?

3) Choice of parameters

Some recommendations:

Cross-over rate:
80-95%

Mutation rate:
0.5-1%

Population size:
100-200
(depends on the problem)

Mark Obitko http://cs.felk.cvut.cz/~xobitko/ga/

```
Initialize population

Repeat

      Evaluate Solutions in the popul-
      ation
      Perform Competitive Search
      Apply Genetic Operators
      Perform Local Search

Until Convergence Criteria Satisfied
```

Pseudo Code for GA

Ph.D. Thesis of
W.E. Hart (1994)

AutoDock Tools (ADT)

Freeware

Dr. Micheal Sanner
Molecular Graphics Lab
Scripps Research Institute
La Jolla, CA

ADT provides GUI interface for setting up and using AutoDock

http://www.scripps.edu/pub/olson-web/people/sanner

1) Sybyl

Receptor:

PDB -> Add essential H -> Fix charges
Kollman United charges -> save as Mol2

Ligand:

PDB -> check for atom types -> add all
-> fix charges -> save as Mol2 file

2) type *adt* at the system prompt

AutoTors

1) Read the ligand molecule
2) Use Autotors to setup the tree and brances
3) Toggle torsion activity
4) Aromatic Carbons C–> A
5) Non–Polar hydrogens (merge or restore)

AutoGpf (Grid Parameter File)

1) Read the macromolecule
2) If the solvation parameters are not added ADT queries whether to add them, if you say yes, it converts the mol2 file to pdbqs file
3) Set Map types
4) Set grid maps
5) Write GPF or edit GPF

AutoDPF (Docking Parameter File)

1) Read the macromolecule

2) Read the ligand

3) Select Docking algorithm

4) Set docking run parameters

5) write or edit DPF file

Start

1) Start AutoGrid

2) Start AutoDock

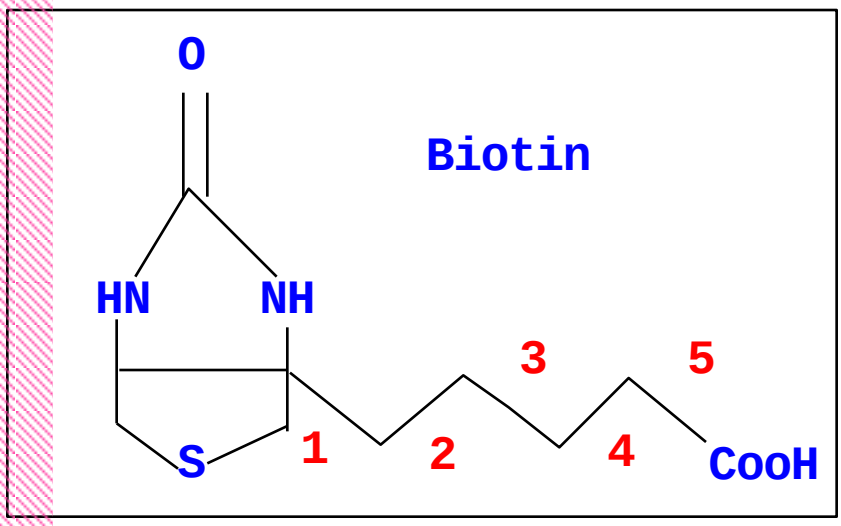
3) Options to submit jobs in other SGIs

4) Cancel the jobs

Test Case

Streptavidin/Biotin (1stp) (2.6 Ang Resol)

Weber et al 1989



Autogrid:

Number of points:

116, 104, 124 (x,y,z),

117 X 104 X 124 = 1508832

Grid Spacing = 0.375 Ang.

Map types C,H,N,S,O & e

CPU Time: 7.53s AutoGrid

Octane SGI workstation

Docking: LGA-LS, Population Size = 50;

Elitism = 1, Cross-over rate = 0.8

rate of mutation = 0.02, GA_num_evals 250000

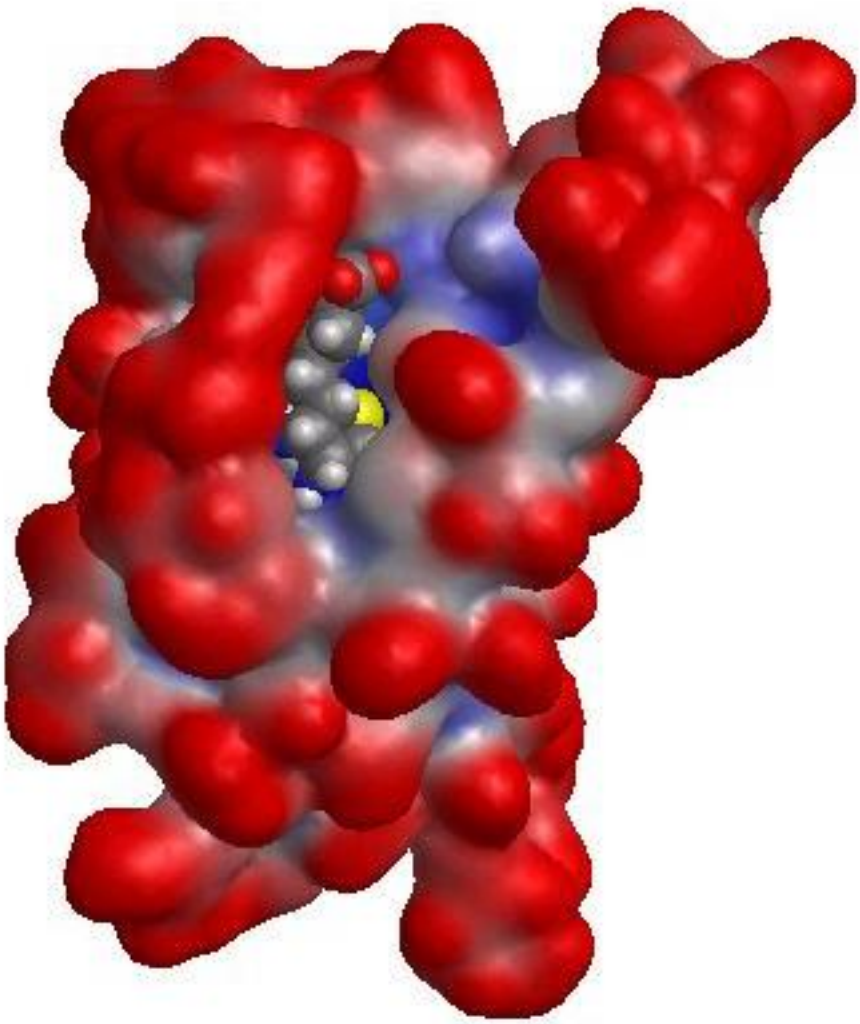
Lowest Docked Energy Mean	Mean Docked En ergy	Number in Clust er	Ref. RMSD	Est Free Ene. of Binding
-10.78	-10.46	6	0.39	9.03

CPU Time: 3m 25.13s

Octane SGI workstation

Streptavidin/Biotin complex (1stp)

P.C. Weber, D.H. Ohlendorf, J.J. Mendolowski, and F.R. Salemme (1992)



Bound Conformation Violet
Lowest Energy Conformation Red

S.Ravichandran, Ph.D., ABCC, NCI–Frederick, MD; sravi@ncifcrf.gov

SAMPLE CLUSTERING HISTOGRAM

Clus-ter Rank	Lowest Docked Energy	Run	Mean Docked Energy	Num in Clus	Histogram 5 10 15 20 25 30 35
1	−10.27	17	−10.21	2	##
2	−10.19	20	−10.12	3	###
3	−10.14	3	−10.14	1	#
4	−10.05	4	−10.05	1	#
5	−9.05	7	−9.05	1	#
6	−8.37	13	−8.37	1	#
7	−8.18	1	−8.18	1	#
8	−8.07	9	−8.07	1	#
9	−7.80	12	−7.80	1	#
10	−7.47	10	−7.47	1	#
11	−7.41	6	−7.41	1	#
12	−7.40	11	−7.40	1	#
13	−7.24	18	−7.24	1	#
14	−7.15	15	−7.15	1	#
15	−7.07	5	−7.07	1	#
16	−6.90	2	−6.90	1	#
17	−6.82	8	−6.82	1	#

Estimated Free Energy of Binding = − 8.11 kcal/mol
Observed = −18.27 kcal/mol

Things to know:

Genetic Algorithm methods are non-deterministic (sucessive runs may not produce the same answer).

Also there is no guarantee that the solution identified by GA will be the best solution.

Useful Links

http://w3.to/autodock

http://nciiris.ncifcrf.gov/~ravichas/docking

Happy Docki