Structural Bioinformatics

Yasaman Karami

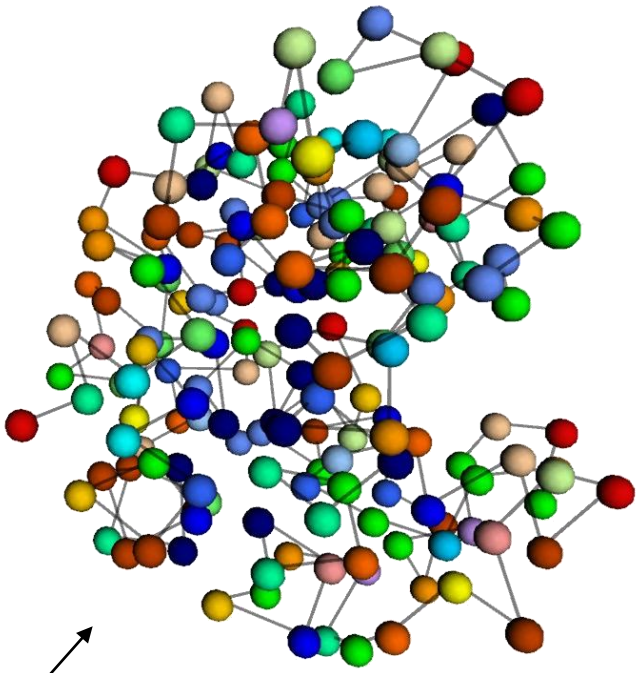
Master BIM-BMC Semestre 3, 2014-2015

Laboratoire de Biologie Computationnelle et Quantitative (LCQB)
e-documents: http://www.lcqb.upmc.fr/laine/STRUCT
e-mail: elodie.laine@upmc.fr

Protein Structure Prediction (PSP)

- The goal is to predict the (complex) 3D structure (and some sub-features) of a protein from its amino acid sequence (a 1D object)

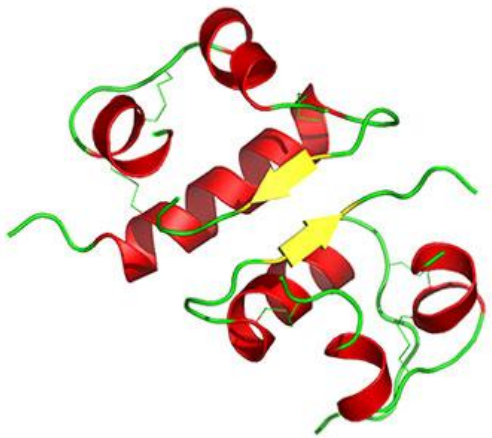
Primary Sequence



3D Structure

The importance of PSP

- The function of a protein inside a cell depends on its 3D structure
- Protein folding procedure to understand their physical, chemical and biological features
- According to Science, the problem remains one of the top 125 outstanding issues in modern science.[1]
- Some of the most successful methods have a reasonable probability of predicting the folds of small, single-domain proteins within 1.5 angstroms over the entire structure.[2]
- Drug design (using protein structures and docking methods)
- Elimination of some diseases (like Alzheimer)

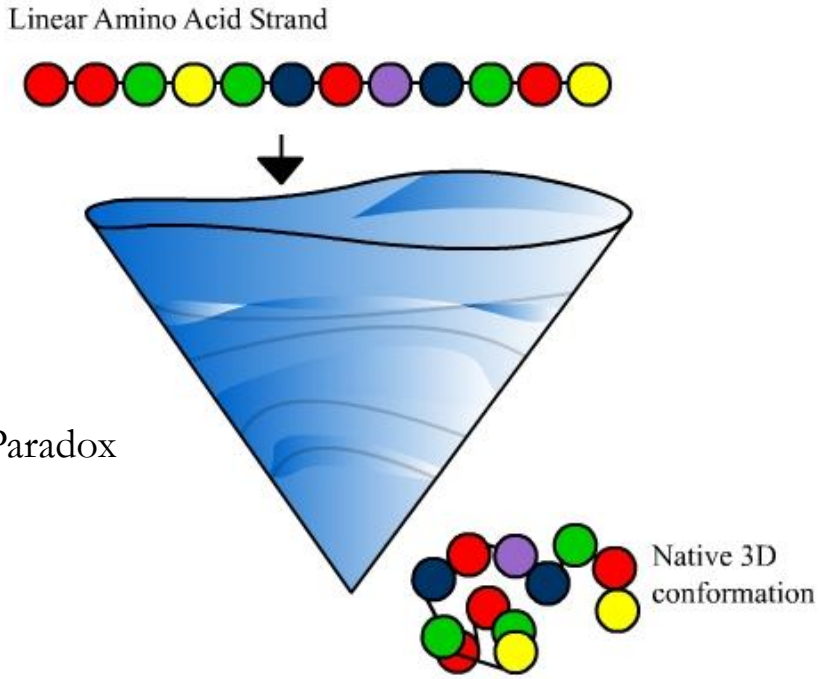


Insulin 3D structure

Science 2005, 309:78-102.
Dile et al.,Current Opinion in Structural Biology 2007, 17:342–346.

The difficulties

- The huge number of possible conformations (Levinthal paradox)
- Physical features of proteins are still unknown.
- Some proteins could fold to multiple structures.
- The impact of nature and other macromolecules on folding procedure.



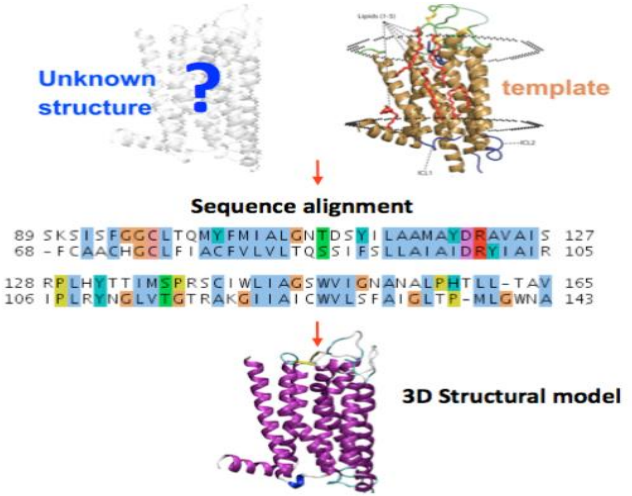
Levinthal Paradox

Methods

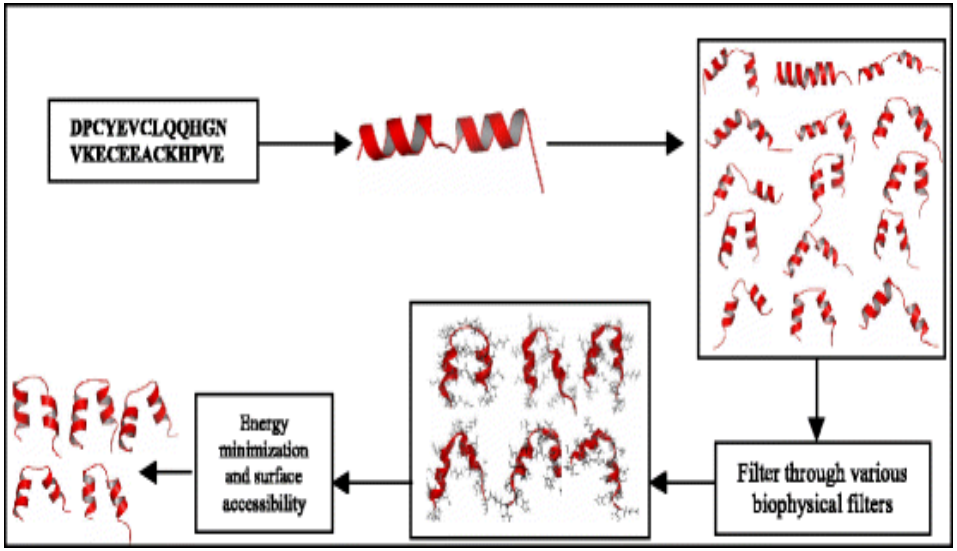
- Three categories:

1. Homology modeling/Comparative modeling
2. Threading
3. Ab-initio

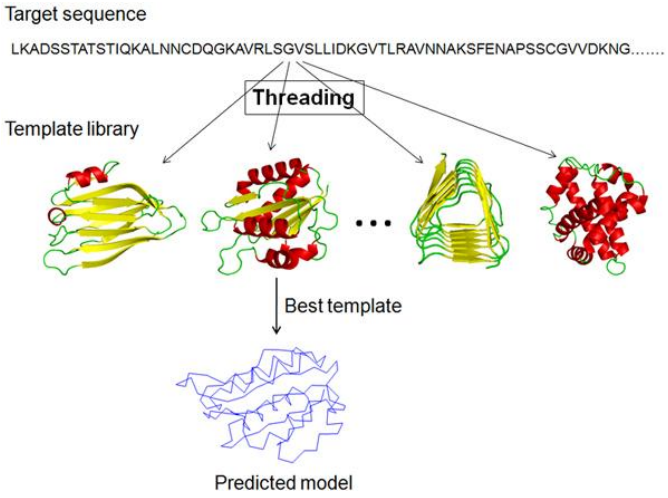
1



3



2



Anfinsen Dogma

- Championed by the Nobel Prize Laureate
- At the environmental conditions (temperature, solvent concentration and composition, etc.) at which folding occurs, the native structure is a unique, stable and kinetically accessible minimum of the free energy.

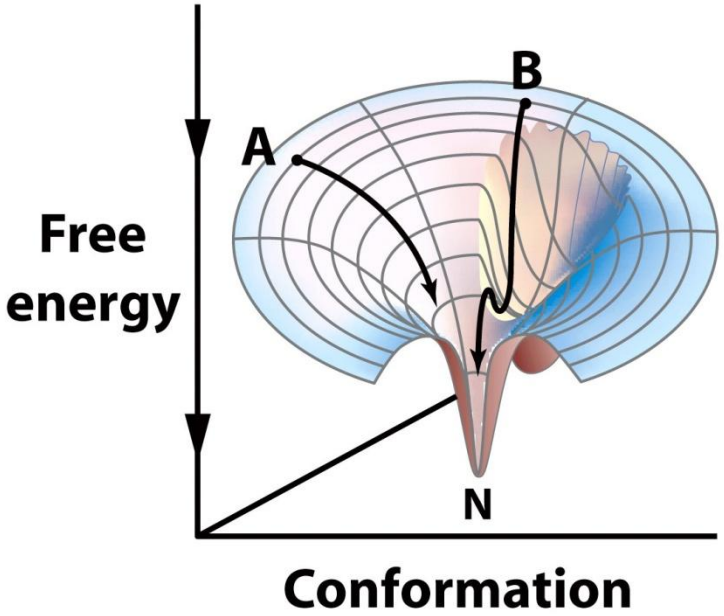


Figure 4-30a Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

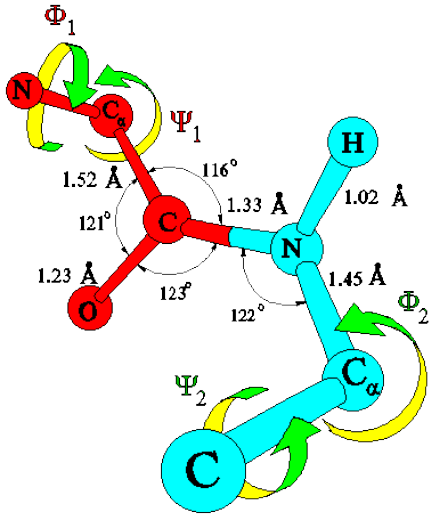
- Optimization problem
  - Evolutionary Algorithms (Genetic Algorithm)
  - Bio-inspired Algorithms (Ant Colony Algorithm, etc.)

Steps of ab-initio prediction

1. The conformation representation

The main degrees of freedom in forming the 3D trace of the polypeptide chain are the two dihedral angles on each side of the Ca atom.

- φ (phi, involving the backbone atoms C'-N-Cα-C')
- ψ (psi, involving the backbone atoms N-Cα-C'-N)
- ω (omega, involving the backbone atoms Cα-C'-N-Cα)

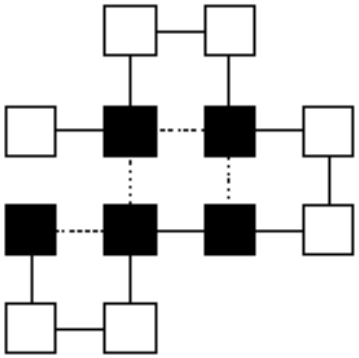


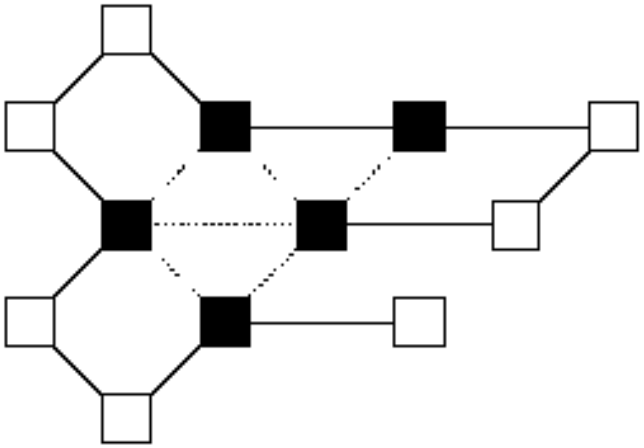
φ controls the C'-C' distance, ψ controls the N-N distance and ω controls the Cα-Cαdistance.

a dihedral or torsion angle is the angle between two planes

HP Protein Folding Model

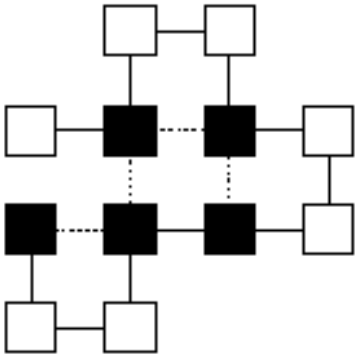
- Amino acid chains (proteins) are represented as connected beads on a 2D or 3D lattice
- HP: hydrophobic – hydrophilic property
- Hydrophobic amino acids can form a hydrophobic core energy potential

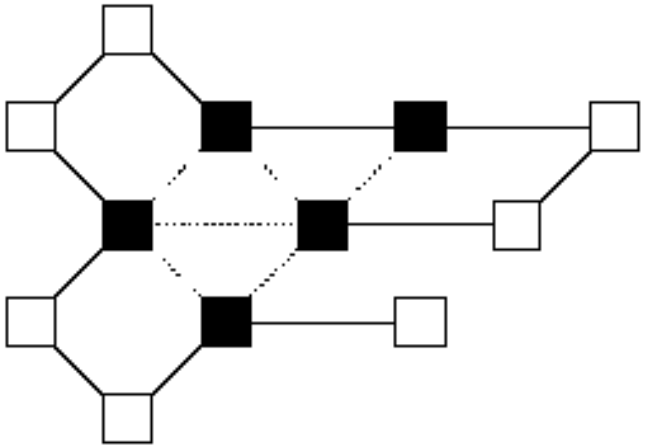




HP Protein Folding Model

- Model adds energy value to each hydrophobic pair that are adjacent on lattice AND not consecutive in the sequence





- Goal: *find low energy configurations!*

Encodings for Internal Coordinates

- Proteins are represented using internal coordinates (vs. Cartesian)
- Absolute vs. Relative encoding
- Absolute Encoding: specifies an absolute direction

  cubic lattice: $\{U,D,L,R,F,B\}^{n-1}$
- Relative Encoding: specifies direction relative to the previous amino acid

  cubic lattice: $\{U,D,L,R,F\}^{n-1}$

Steps of ab-initio prediction (cont.)

2. Force field or energy function

a force field refers to the form and parameters of mathematical functions used to describe the potential energy of a system of particles (typically molecules and atoms)

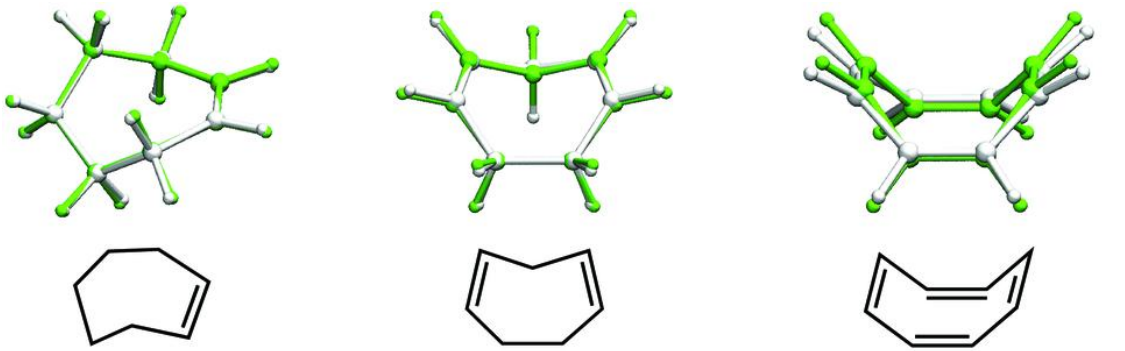
- Bonded terms: atoms that are linked by covalent bonds
- Non-bonded (non-covalent) terms: the long-range electrostatic and van der Waals forces

$$E(R) = \sum_{bonds} B(R) + \sum_{angles} A(R) + \sum_{torsions} T(R) + \sum_{non-bonded} N(R)$$

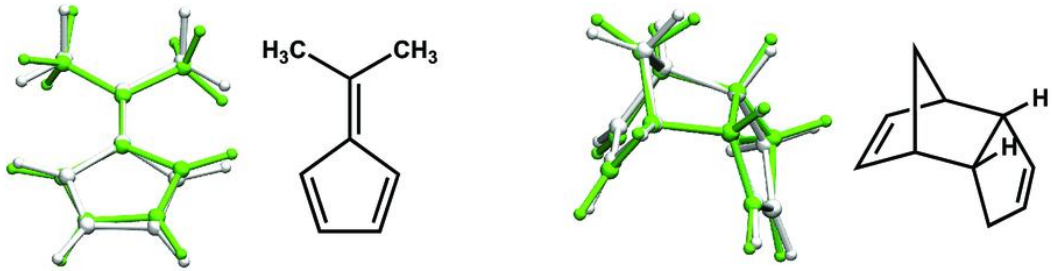
Steps of ab-initio prediction (cont.)

3. Evaluation of predicted structure

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} |r_{ai} - r_{bi}|^2}{n}}$$

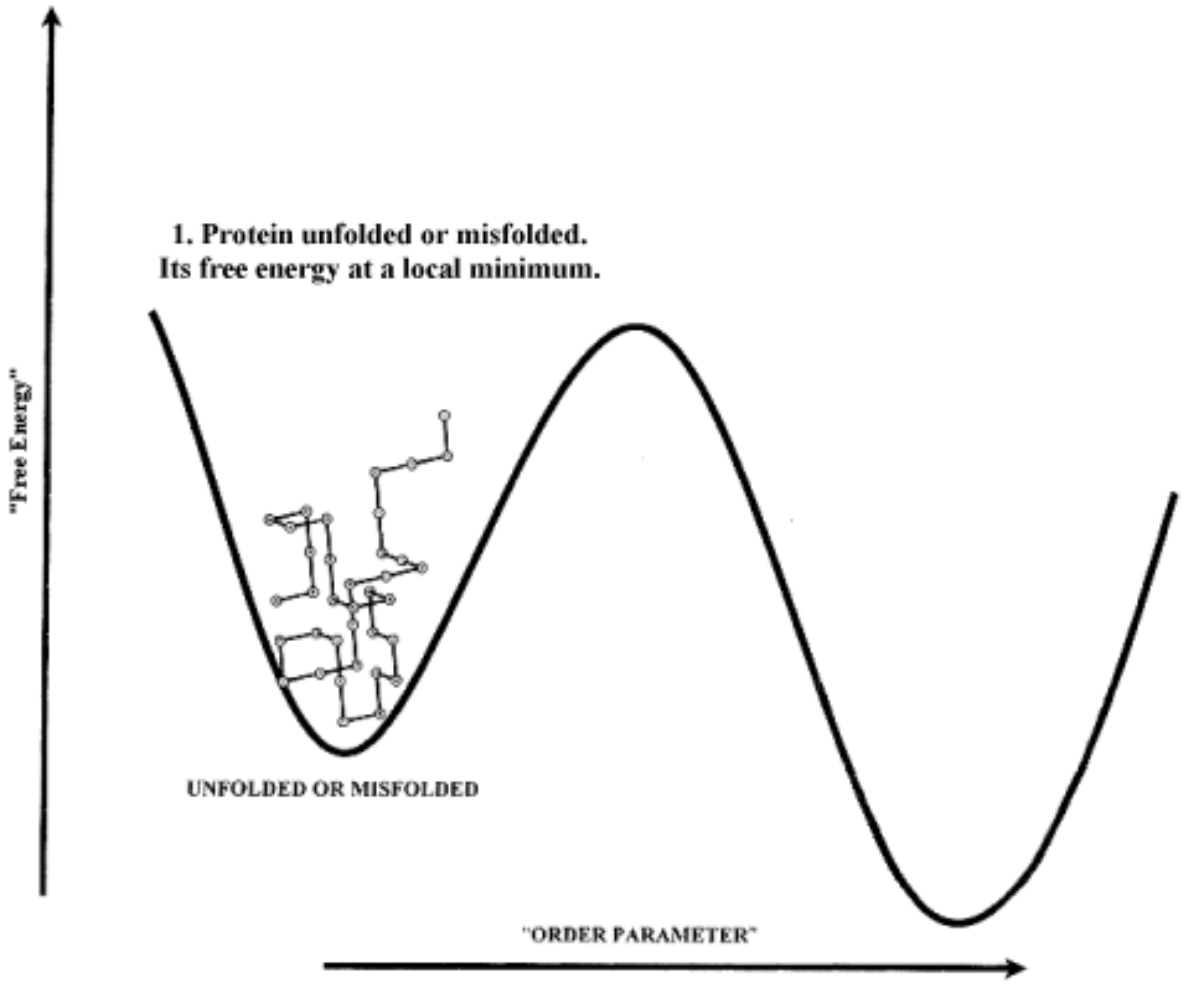


cycloheptene (2.36) 1,3-cycloheptadiene (0.16) cyclooctatetraene (3.29)

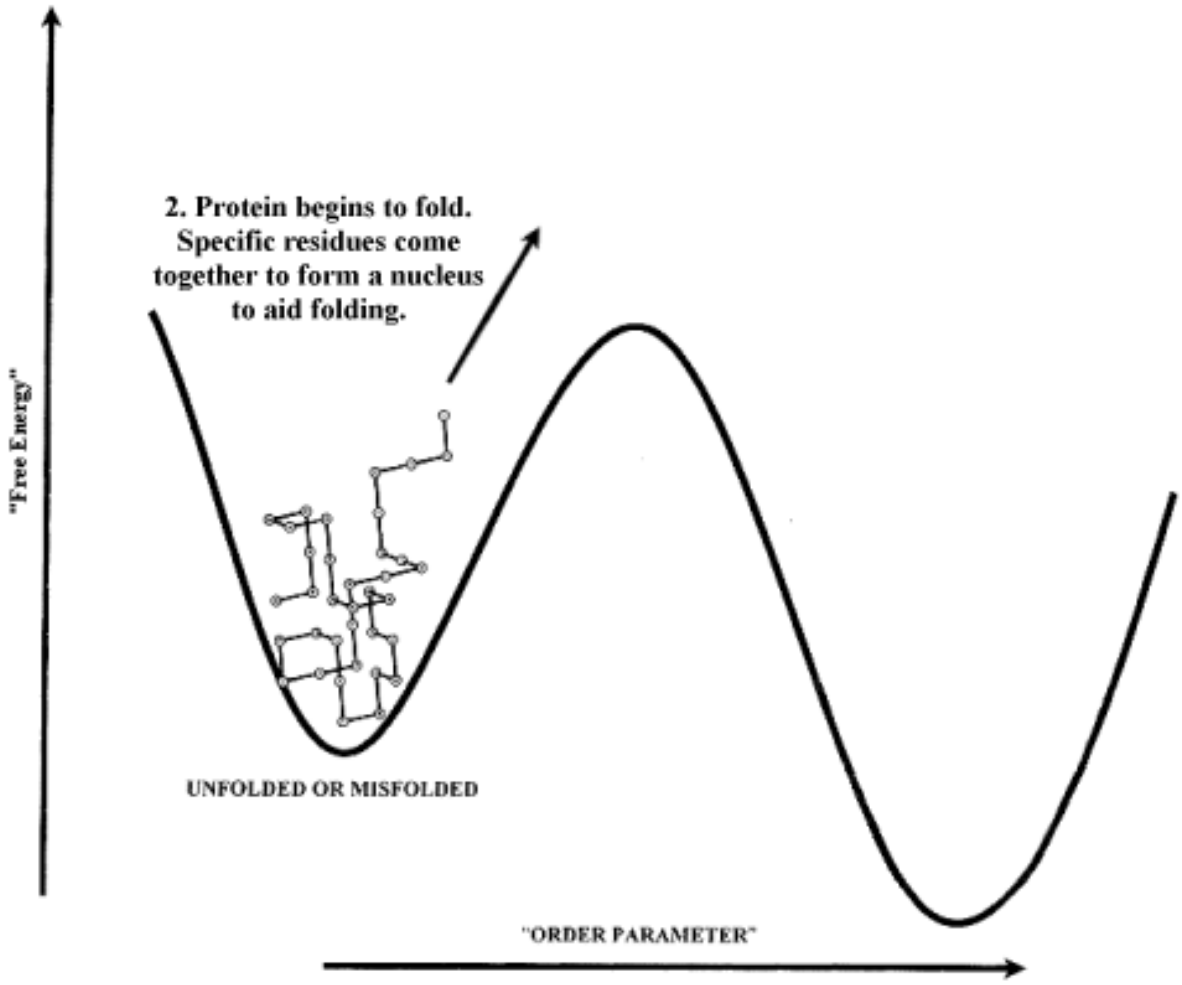


dimethylfulvene (1.79) dicyclopentadiene (0.22)

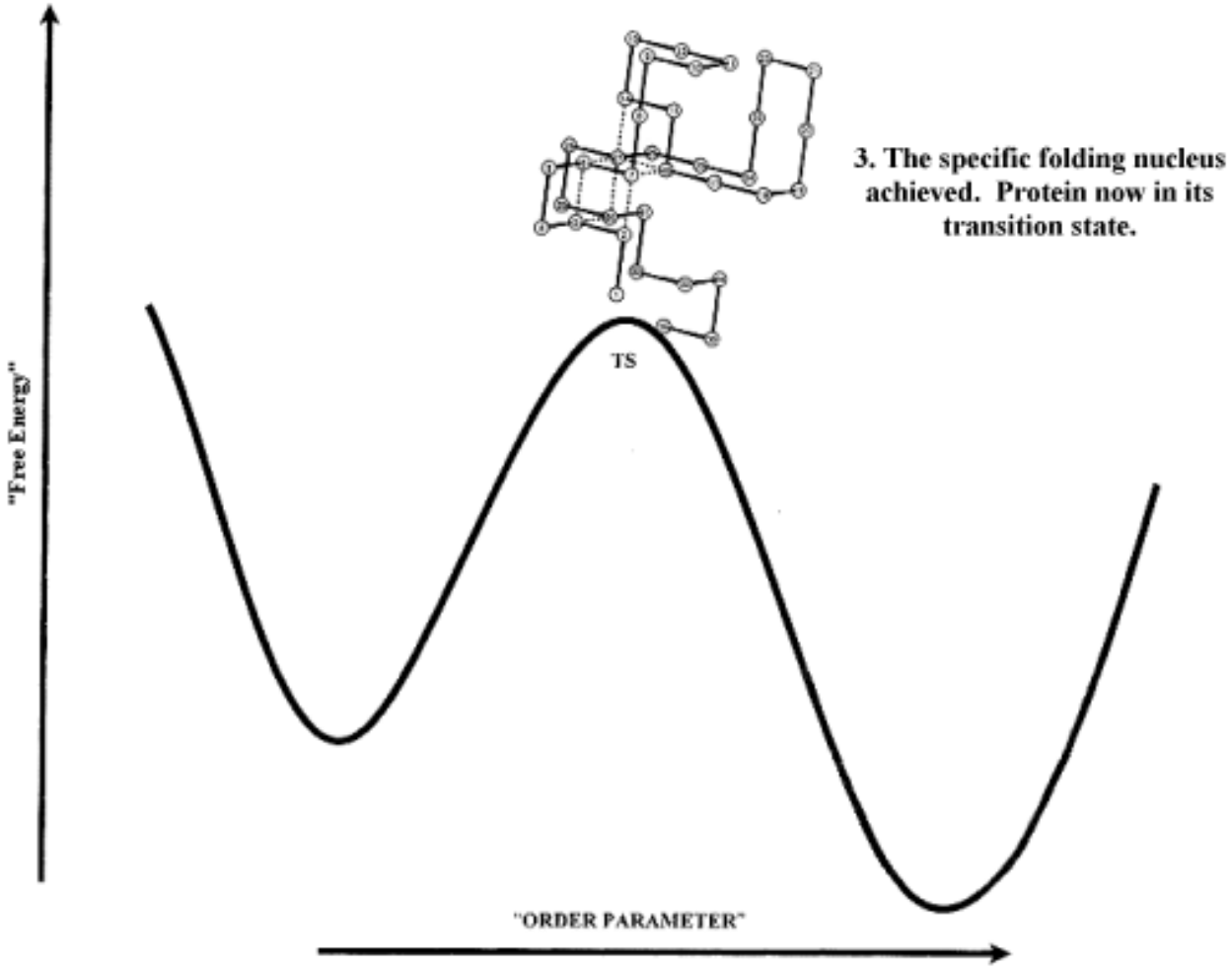
PSP using optimization algorithm



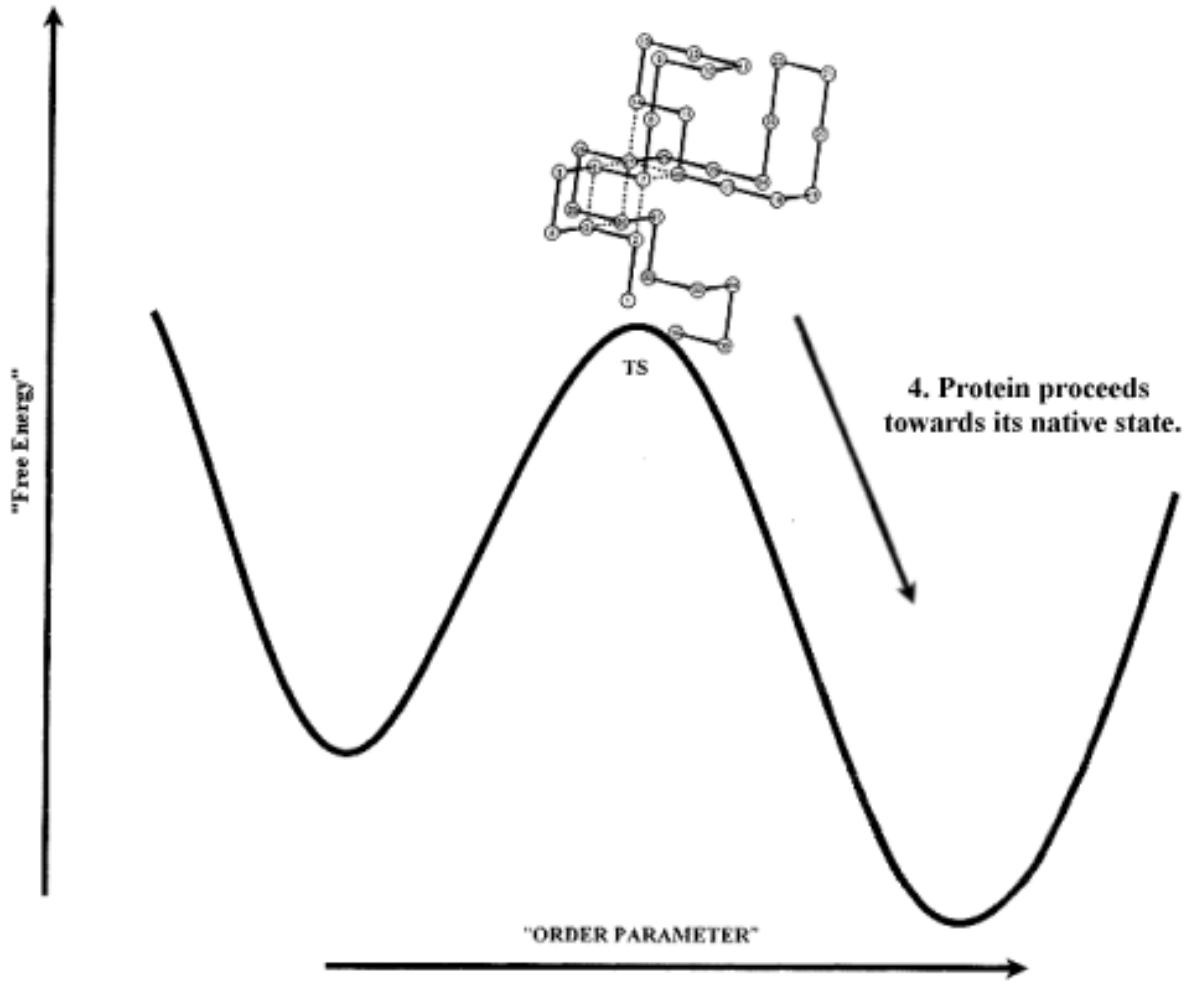
PSP using optimization algorithm



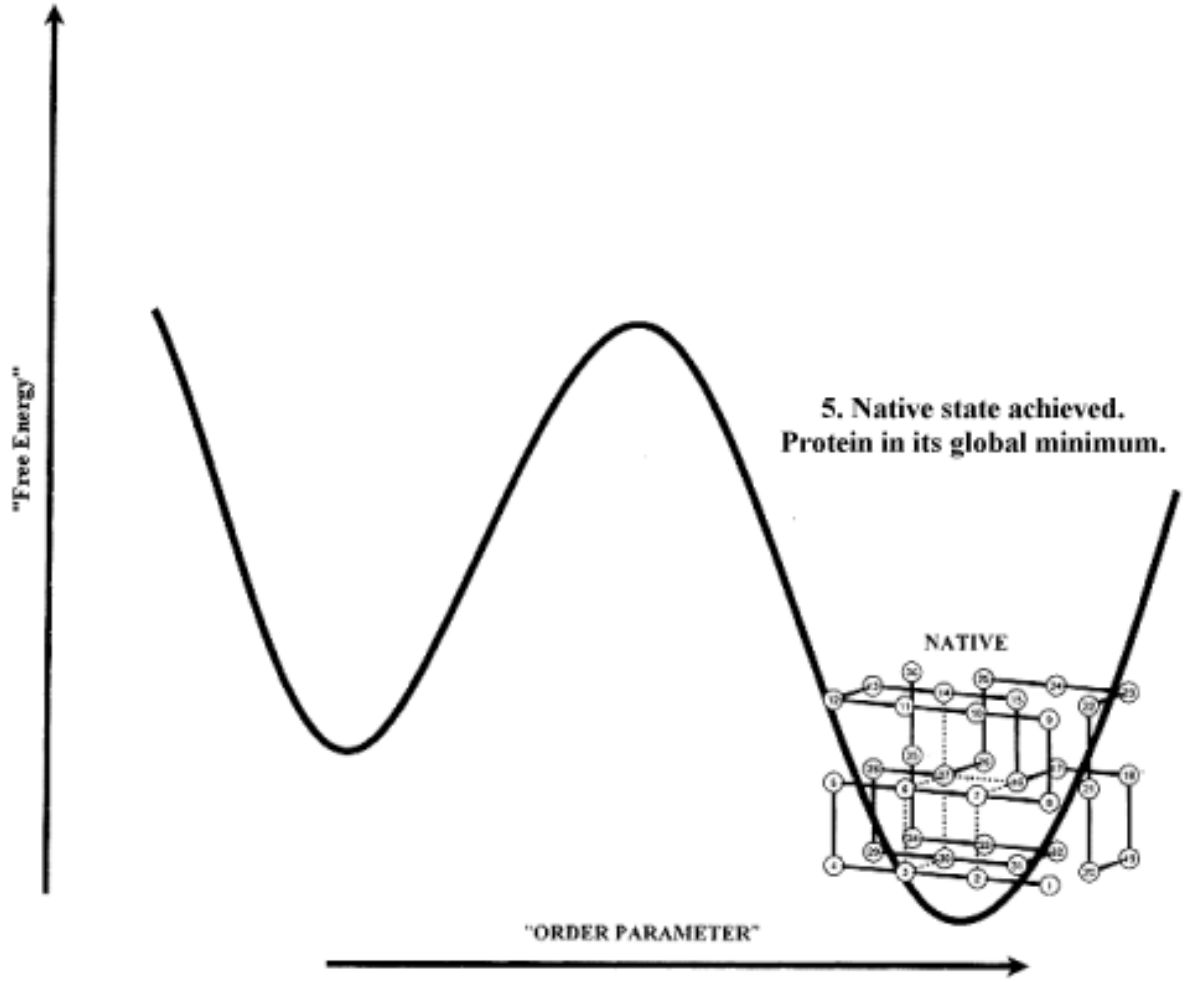
PSP using optimization algorithm



PSP using optimization algorithm

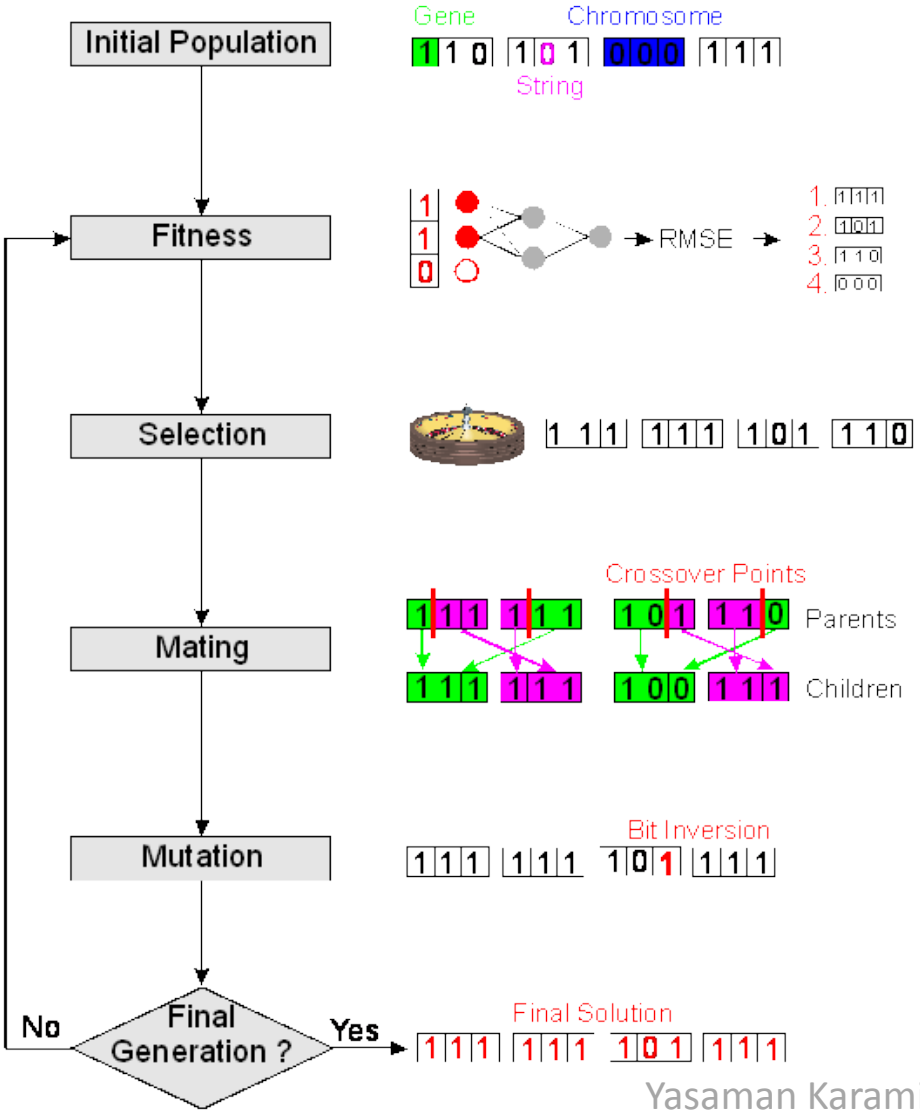


PSP using optimization algorithm



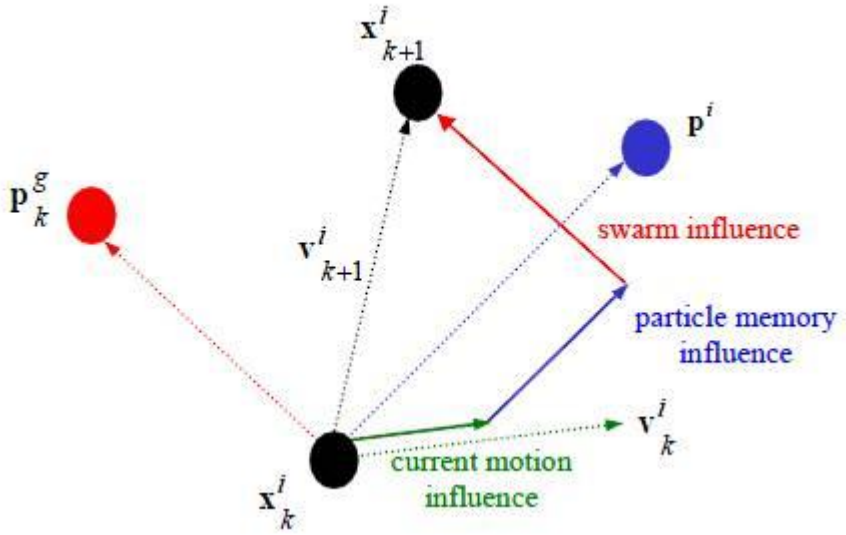
Genetic Algorithm (GA)

- Genetic Algorithm (based on evolution process in human body):
  - Chromosome or conformation
  - Coding method
  - Fitness function
  - Selection
  - Cross over/Mating
  - Mutation



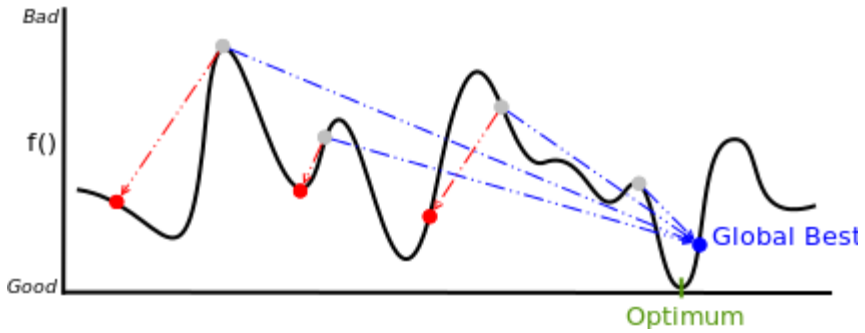
Particle Swarm Optimization (PSO)

- stylized representation of the movement of organisms in a bird flock or fish school.
- The local and global best position and velocity is updated permanently.



Hassan et al., American Institute of Aeronautics and Astronautics (2005).

Reid et al., Evolutionary Computation (CEC), IEEE, 2014.

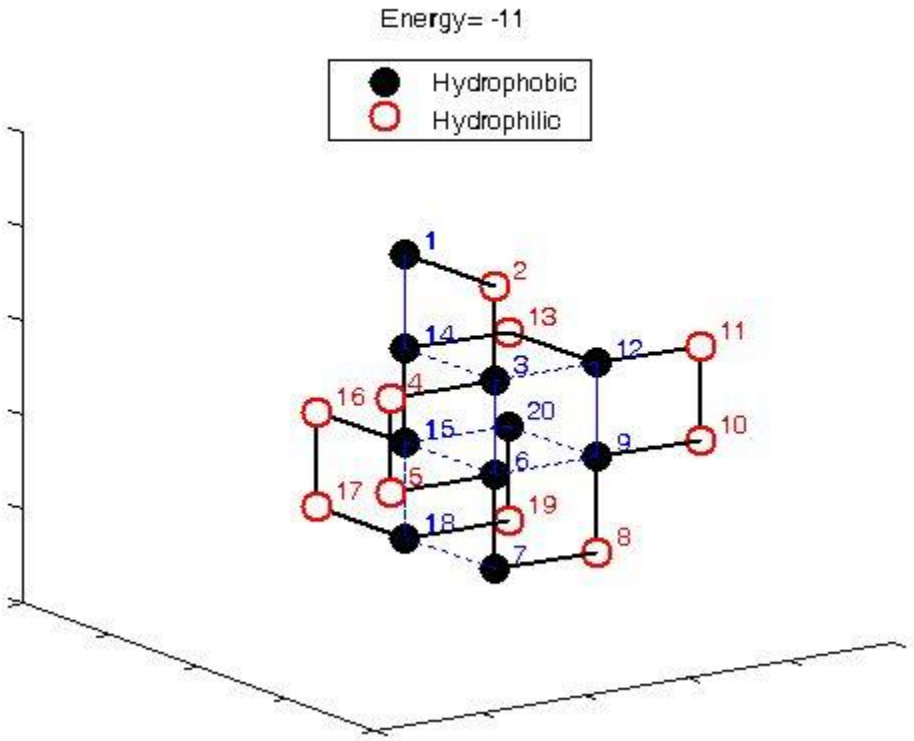


$$V_{id}^{k+1} = V_{id}^{k} + c_1 \times rand(.) \times \left(P_{id} - x_{id}^{k}\right) + c_2 \times rand(.) \times \left(P_{gd} - x_{id}^{k}\right)$$

$$X_{id}^{k+} = X_{id}^{k} + V_{id}^{k+1}$$

HP lattice model

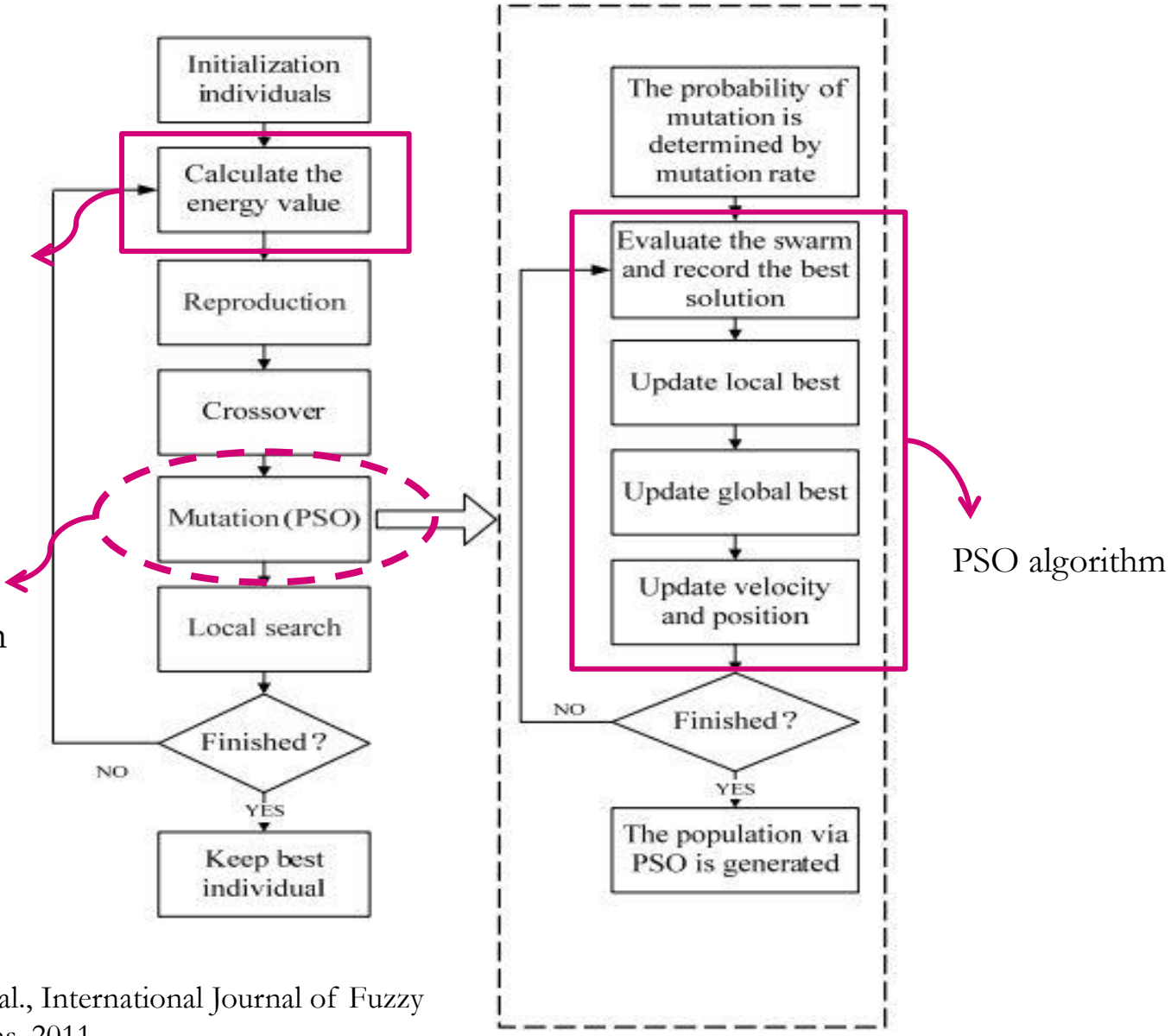
- HP lattice simplification:
  - The number of H-H neighboring connections is computed and multiplied by -1.



combination of GA and PSO on HP lattice

using hydrophobic -hydrophilic feature of amino acids. (very simple and inaccurate)

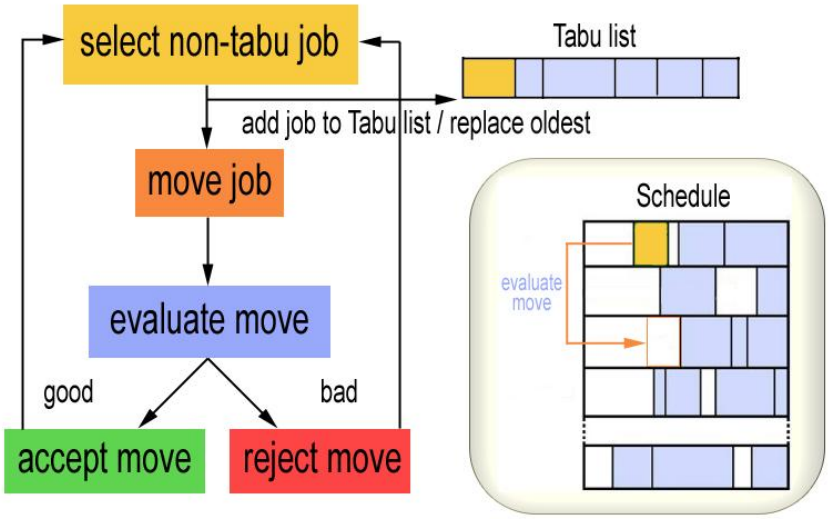
PSO is used to prevent the collision produced by mutation operator



Lin et al., International Journal of Fuzzy Systems, 2011.

Tabu Search

- Algorithm
  - TS is a local neighborhood search algorithm, leads the next search based on a flexible memory function. It could search quickly but is weak to find the global optimum point.
  - In this work fitness function of the mutated children are calculated, then they are placed in the tabu list with their energy values. This list is sorted increasingly based on energy values and used to find the final mutated children. Crossover is implemented in TS the same as mutation with some differences; energy of offspring are compared with a threshold then placed in the tabu list.

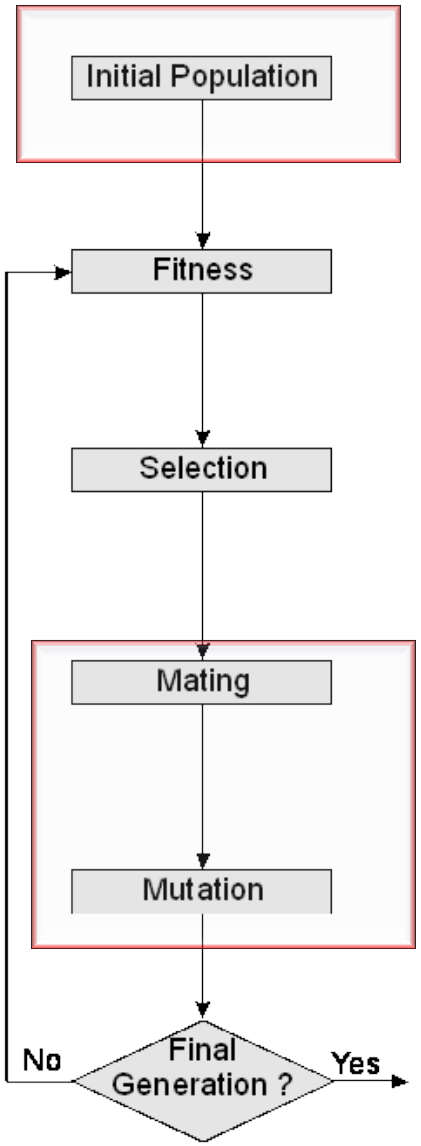


AB off-lattice

- AB off-lattice model is also based on hydrophobicity and divides the amino acids to two groups, A and B.
- A stands for hydrophobic and B stands for hydrophilic amino acids.
- In addition to the property of hydrophobicity, bend angles and the distance between amino acids should be considered to calculate the energy value.
- This model is more accurate than HP lattice, also the calculations to find the energy value are simpler.

$$E = \sum_{i=2}^{n-1} \frac{1}{4}(1 - \cos\theta_i) + \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} 4\left[r_{ij}^{-12} - c(\xi_i + \xi_j) r_{ij}^{-6}\right]$$

combination of GA and TS on AB off-lattice



AB off-lattice

Using Tabu Search method

Zhang et al., BMC Systems Biology, 2010.

Multi-objective optimization

- The main idea:
  - finding the native structure of a given protein is not equivalent to "finding a native state needle in a conformational space haystack" but, instead, should be more like "finding a set of equivalent needles in a haystack".

- CHARMM force field:

$$E_{CHARMM} = E_1 + E_2 + E_3 + E_4 + E_5 + E_6 + E_7$$

$$E_1 = \sum_{bonds} k_b(b - b_0)^2, \qquad E_2 = \sum_{UB} k_{UB}(S - S_0)^2, \qquad E_3 = \sum_{angles} k_\theta(\theta - \theta_0)^2,$$

$$E_4 = \sum_{torsions} k_x[1 + \cos(\eta x - \delta)], \; E_5 = \sum_{UB} k_{imp}(\varphi - \varphi_0)^2,$$

$$E_6 = \sum_{non-bond} \varepsilon_{ij}\left[\left(\frac{R\,min_{ij}}{r_{ij}}\right)^{12} - \left(\frac{R\,min_{ij}}{r_{ij}}\right)^{6}\right], \qquad E_7 = \frac{q_i q_j}{e r_{ij}}$$

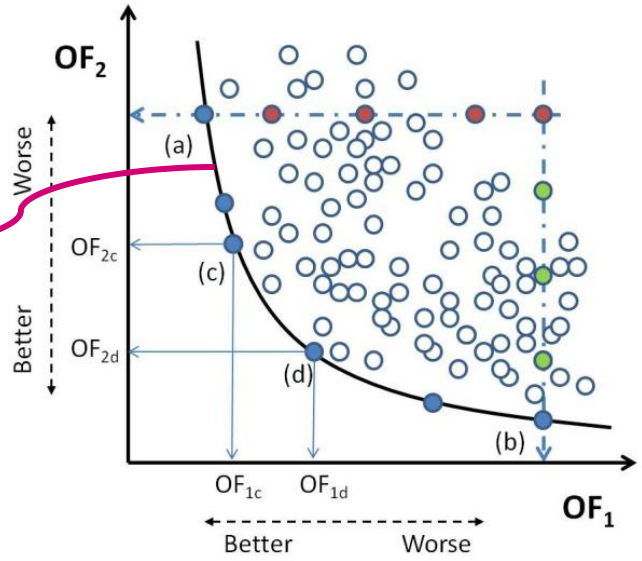
Multi-objective optimization

- Two objective function:
  - Optimization using Pareto front

$$F_1 = E_{bond}(A_{bond}, C_{bond}) = \sum_{k=1}^{5} E_k$$

$$F_2 = E_{non-bond}(A_{non-bond}, C_{non-bond}) = \sum_{k=6}^{7} E_k$$

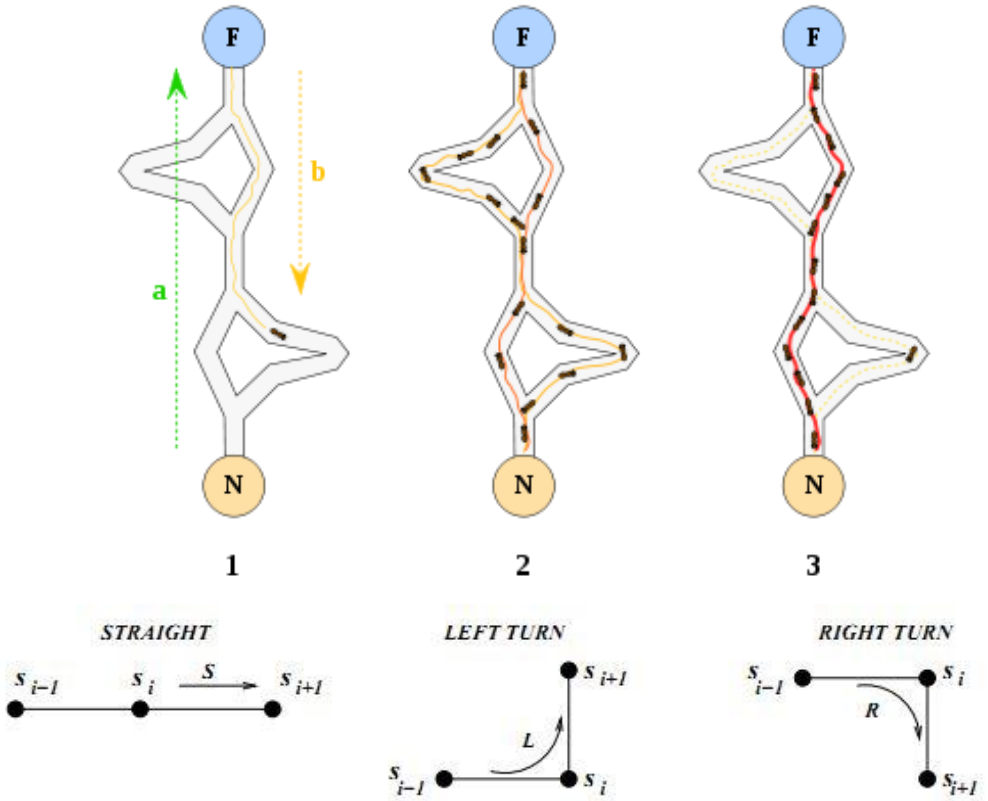
Pareto front line



Cutello et al., J. R. Soc. Interface, 2005.

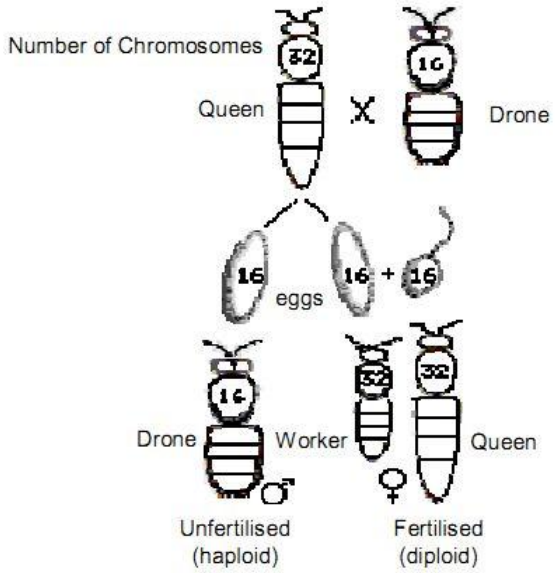
Ant Colony Optimization (ACO)

- Ant colony algorithm is inspired from the ants' social life and food foraging.
- Ants communicate with each other and find food source through pheromone trails.



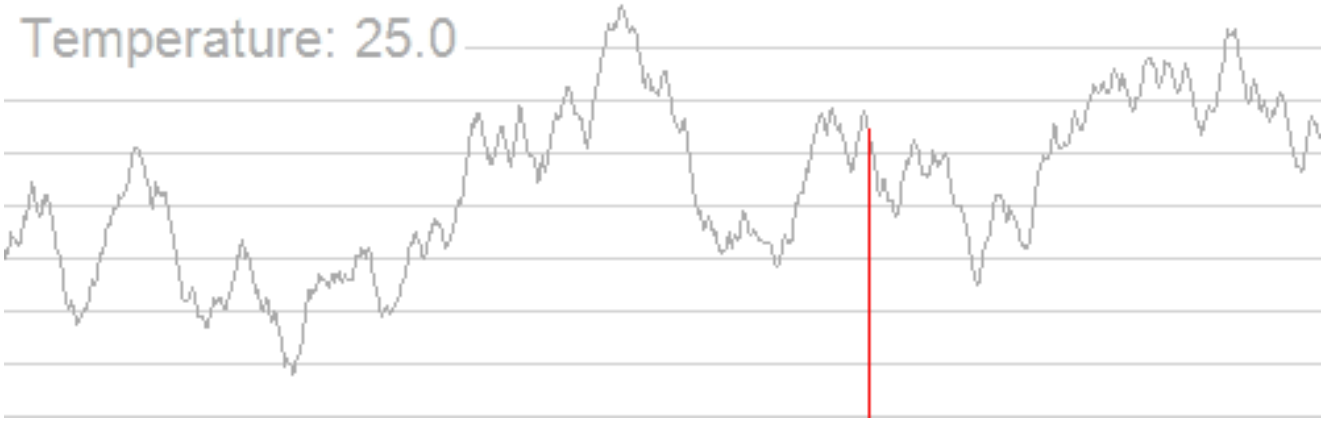
Honey Bee Optimization (HBO)

- It is based on marriage in honey bees and inspired from honey bee's reproduction process.
- Honey bee colony consists of a queen, drones, broods and workers.
- Workers are considered as local search algorithms that improve the results and queen as the answer.
- At the end, the improved and mutated brood will be considered as the queen if it has the better fitness value.
- Childs will be produced from crossover of the queen and drones.



Simulated Annealing (SA)

- SA is a generic probabilistic meta-heuristic for the global optimization problem of locating a good approximation to the global optimum of a given function in a large search space.



Simulated Annealing used for hill climbing

Conclusion

- 3D structure gives clues to function determination
- 3D structure determination is difficult, slow and expensive
- The use of a near-optimal meta-heuristic algorithm, such as a Genetic Algorithm, is one of the most promising optimization methods as it explores minimal number of potential structures.
- Conformation representation and energy models are very important
- A powerful exploration of the conformational space can be achieved by using Optimization algorithms.