Introduction to Bioinformatics

Iosif Vaisman

`Email: ivaisman@gmu.edu`

Protein Modeling Methods

- *Ab initio* methods

- **Energy-based methods**

- **Knowledge-based methods**

Protein Modeling Methods

- *Ab initio* methods:
  solution of a protein folding problem
  search in conformational space
- **Energy-based methods:**
  energy minimization
  molecular simulation
- **Knowledge-based methods:**
  homology modeling
  fold recogniion

Knowledge

Knowledge is a pattern that exceeds certain threshold of interestingness.

Factors that contribute to interestingness:
coverage
confidence
statistical significance
simplicity
unexpectedness
actionability

Knowledge-based methods

Finding patterns in known structures

Deriving rules (usually in the form of PMF)

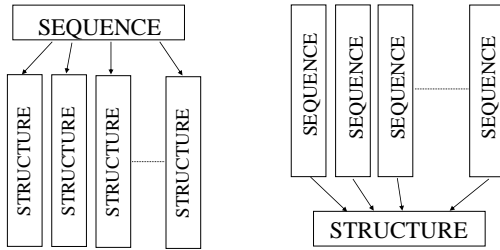Applying the rules

---

Fold Recognition

Pattern searching
sequence patterns
structure patterns
residue composition patterns

Threading
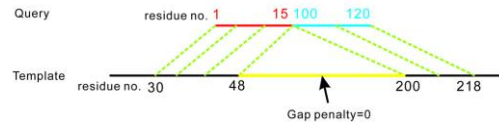sequence-structure compatibility
structure-sequence compatibility

Threading

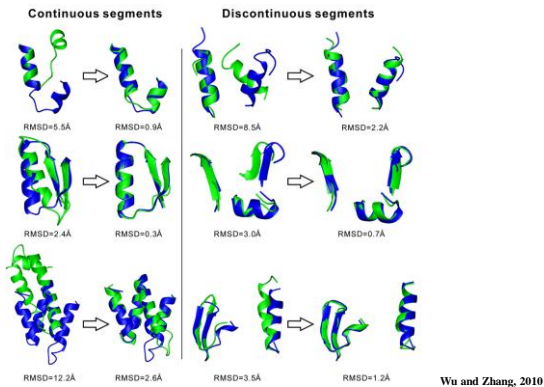Sequence-structure compatibility (fold recognition)

Structure-sequence compatibility (inverse folding)

Segmental Threading

Wu and Zhang, 2010

Segmental Threading

Continuous segments · Discontinuous segments

RMSD=5.5Å RMSD=0.9Å RMSD=8.5Å RMSD=2.2Å

RMSD=2.4Å RMSD=0.3Å RMSD=3.0Å RMSD=0.7Å

RMSD=12.2Å RMSD=2.6Å RMSD=3.5Å RMSD=1.2Å

Wu and Zhang, 2010

Threading

• Only the local environment is taken into account

• Non-local contacts are assumed with generic peptide

• No gaps are allowed in the alignment

Homology Modeling

• Identification of structurally conserved regions (using multiple alignment)

• Backbone construction (based on SCR)

• Loop construction (KB or conformational search)

• Side-chain restoration (KB, rotamer, or MM)

• Structure verification and evaluation

• Structure refinement (energy minimization)

Homology Modeling Programs

Swiss-Model
  (http://swissmodel.expasy.org)

Modeller
  (http://salilab.org/modeller)

CPHmodels
  (http://www.cbs.dtu.dk/services/CPHmodels)

Protein Model Portal (PMP)
  (http://www.proteinmodelportal.org)

Swiss-Model

- Method:
  Knowledge-based approach.
- Requirements:
  At least one known 3D-structure of a related protein.
  Good quality sequence alignements.
- Procedures:
  Superposition of related 3D-structures.
  Generation of a multiple a alignement.
  Generation of a framework for the new sequence.
  Rebuild lacking loops.
  Complete and correct backbone.
  Correct and rebuild side chains.
  Verify model structure quality and check packing.
  Refine structure by energy minimisation and molecular dynamics.

Methods and Programs used by Swiss-Model

- **Sequence Alignment**
  BLAST (Altschul S.F., *J. Mol. Biol.* **215**:403, 1990)
  SIM (Huang, X., Miller, M. *Adv. Appl. Math.***12**:337, 1991)
  ProModII (Peitsch, M.C. *Unpublished,* Server-specific tool)
- **Knowledge Based Protein Modelling**
  ProMod (Peitsch M.C. *Biochem Soc Trans* **24**:274, 1996)
- **Energy Minimisation**
  Gromos96 (van Gunsteren W.F. *http://igc.ethz.ch/gromos/*)
- **Model evaluation**
  Swiss-PdbViewer
  (*http://www.expasy.ch/spdbv/mainpage.html*)

Swiss-Model Request Types

- First Approach mode.
- Optimise mode.
- Combine mode.
- GPCR mode.

Model Confidence Factors

The Model B-factors are determined as follows:

- The number of template structures used for model building.
- The deviation of the model from the template structures.
- The Distance trap value used for framework building.

The Model B-factor is computed as:

85.0 * (1/ # selected template str.) * (Distance trap / 2.5)
and
99.9 for all atoms added during loop and side-chain building

# Structure verification and validation
Bond lengths (Procheck)

```
--------------------------------------------------------------
Bond         | labeling                      | Value | sigma
--------------------------------------------------------------
C-N          | C-NH1      | (except Pro)      | 1.329 | 0.014
             | C-N        | (Pro)             | 1.341 | 0.016
             |            |                   |       |
C-O          | C-O        |                   | 1.231 | 0.020
             |            |                   |       |
Calpha-C     | CH1E-C     | (except Gly)      | 1.525 | 0.021
             | CH2G*-C    | (Gly)             | 1.516 | 0.018
             |            |                   |       |
Calpha-Cbeta | CH1E-CH3E  | (Ala)             | 1.521 | 0.033
             | CH1E-CH1E  | (Ile,Thr,Val)     | 1.540 | 0.027
             | CH1E-CH2E  | (the rest)        | 1.530 | 0.020
             |            |                   |       |
N-Calpha     | NH1-CH1E   | (except Gly,Pro)  | 1.458 | 0.019
             | NH1-CH2G*  | (Gly)             | 1.451 | 0.016
             | N-CH1E     | (Pro)             | 1.466 | 0.015
--------------------------------------------------------------
```

Bond angles (Procheck)

```
--------------------------------------------------------------
Angle        | labeling                      | Value | sigma
--------------------------------------------------------------
C-N-Calpha   | C-NH1-CH1E   | (except Gly,Pro) | 121.7 | 1.8
             | C-NH1-CH2G*  | (Gly)            | 120.6 | 1.7
             | C-N-CH1E     | (Pro)            | 122.6 | 5.0
             |              |                  |       |
Calpha-C-N   | CH1E-C-NH1   | (except Gly,Pro) | 116.2 | 2.0
             | CH2G*-C-NH1  | (Gly)            | 116.4 | 2.1
             | CH1E-C-N     | (Pro)            | 116.9 | 1.5
             |              |                  |       |
Calpha-C-O   | CH1E-C-O     | (except Gly)     | 120.8 | 1.7
             | CH2G*-C-O    | (Gly)            | 120.8 | 2.1
--------------------------------------------------------------
```
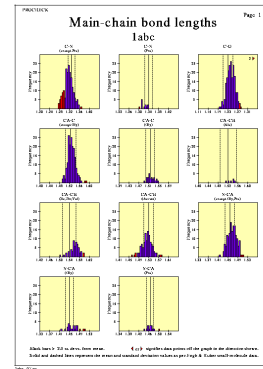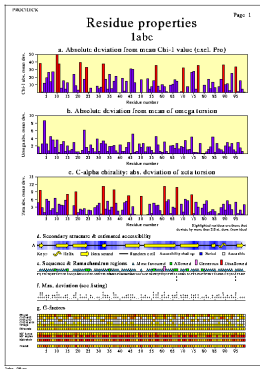
Procheck output
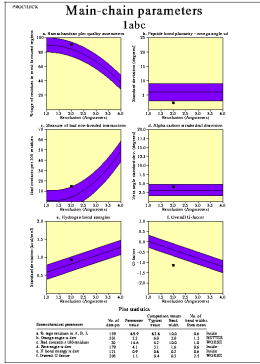
a. Ramachandran plot quality - **percentage** of the protein's residues that are in the **core** regions of the Ramachandran plot.

b. Peptide bond planarity - **standard deviation** of the protein structure's **omega** torsion angles.

c. Bad non-bonded interactions - **number of bad contacts** per **100** residues.

d. Cα tetrahedral distortion - **standard deviation** of the ζ torsion angle (**Cα**, **N**, **C**, and **Cβ**).

e. Main-chain hydrogen bond energy - **standard deviation** of the **hydrogen bond energies** for **main-chain hydrogen bonds**.

f. Overall G-factor - average of different G-factors for each residue in the structure.

Procheck output
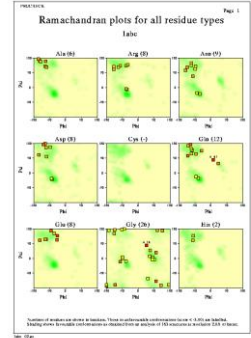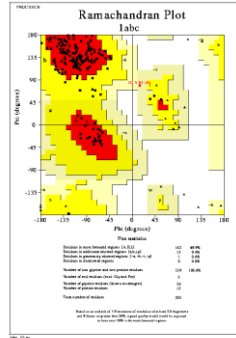
Procheck output
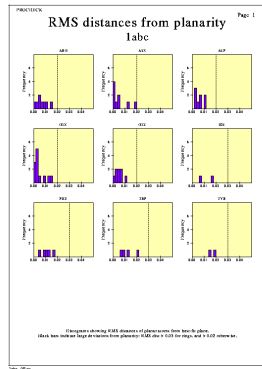
Procheck output

Procheck output - backbone G factors
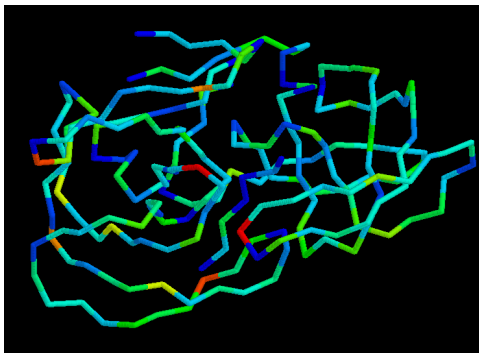
Procheck output - all atom G factors
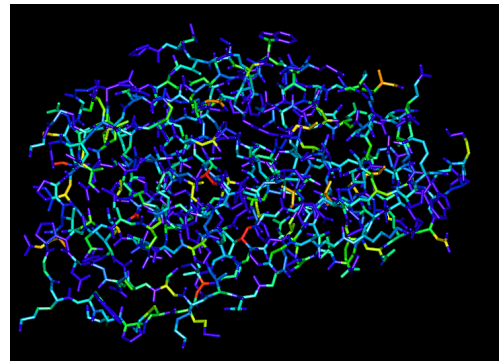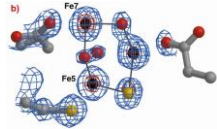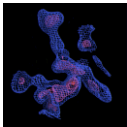
Protein Modeling Methods

- ***Ab initio* methods:**
  solution of a protein folding problem
  search in conformational space
- **Energy-based methods:**
  energy minimization
  molecular simulation
- **Knowledge-based methods:**
  homology modeling
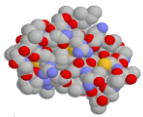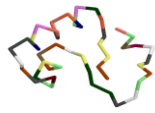 fold recogniion

Potential Energy Functions
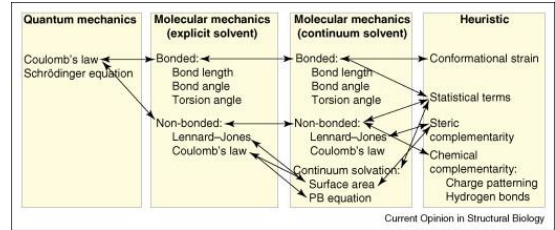
Molecular structure representation

Elementary particles

Atoms
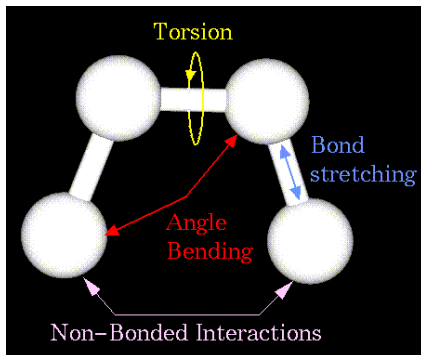

Groups of atoms

Potential Energy Function

$$PEF(R) = \sum_{bonds} K_b \{b(R) - b_{eq}\}^2 + \sum_{angles} K_\theta \{\theta(R) - \theta_{eq}\}^2 +$$

$$\sum_{dihedrals} \frac{K_\phi}{2}\{1 + \cos[n\phi(R) - \gamma]\} +$$

$$\sum_{\substack{non\text{-}bonded \\ atom\ pairs\ i,j}} \left[\frac{A_{ij}}{r_{ij}(R)^{12}} - \frac{B_{ij}}{r_{ij}(R)^{6}} + \frac{q_i q_j}{\varepsilon_r \varepsilon_0 r_{ij}(R)} \right] \quad (1)$$
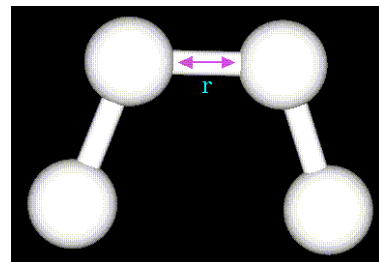
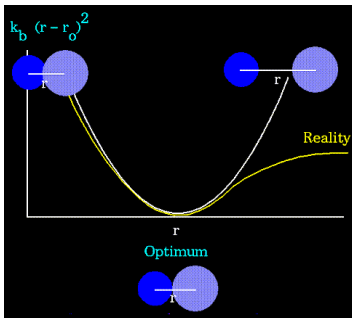Forcefields: AMBER, CHARMM, CVF, ECEPP, GROMOS

Non-Bonded Interactions

Bond length

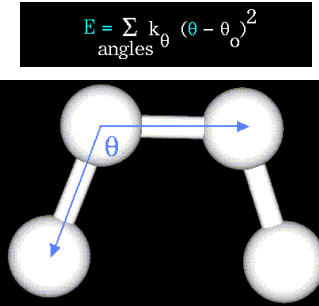$$E = \sum_{bonds} k_b \ (r - r_o)^2$$

Bond length

$$k_b\ (r - r_o)^2$$

Reality

r

Optimum

Bond angle

$$E = \sum_{angles} k_\theta\ (\theta - \theta_o)^2$$

θ

Bond angle

$$k_\theta\ (\theta - \theta_o)^2$$

θ

θ

Optimum
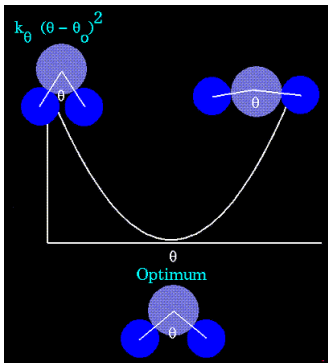
Bond length and angle (parameters)

$$k_\theta\ (\theta - \theta_o)^2 \quad or \quad k_b\ (r - r_o)^2$$

k = 2.0

k = 1.0

k = 0.5

r or θ

Torsional angle

$$E = \sum_{torsions} A\ [1 + \cos(n\tau - \phi)]$$

$$A\ [1 + \cos(n\tau - \phi)]$$

τ
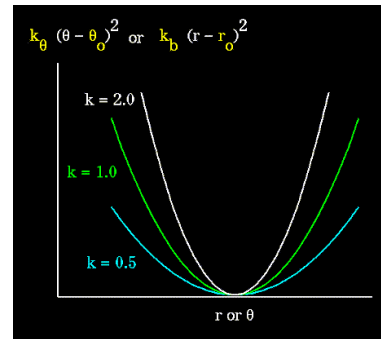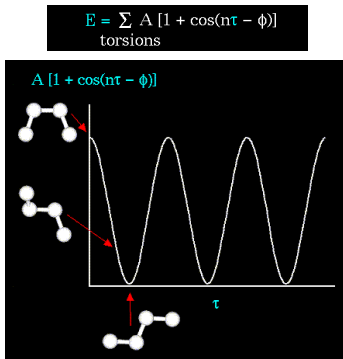
Torsional angle (parameters)

$$A\ [1 + \cos(n\tau - \phi)]$$

A = 2.0, n = 2.0, ϕ = 0.0^o

A = 1.0, n = 1.0, ϕ = 90.0^o

A = 1.0, n = 2.0, ϕ = 0.0^o

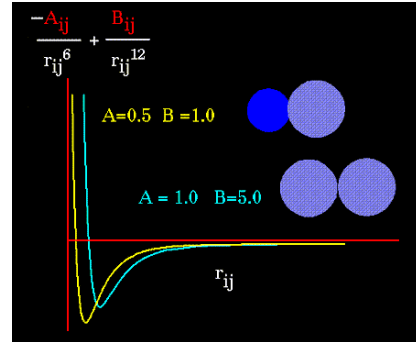τ
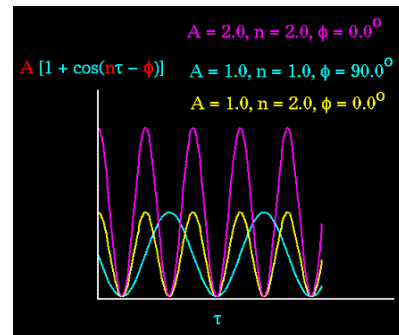
Non-bonded terms

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

van der Waals term   Electrostatic term

$$- \frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}}$$

i j Repulsion regime

van der Waals attraction regime

r_{ij} i   j

Optimum energy

i j
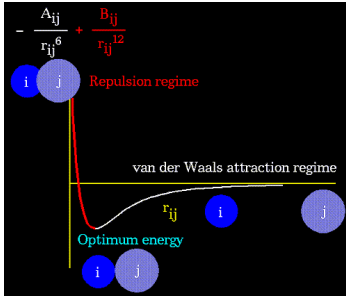
Non-bonded terms (parameters)

$$\frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}}$$

A=0.5 B =1.0

A = 1.0 B=5.0

r_{ij}

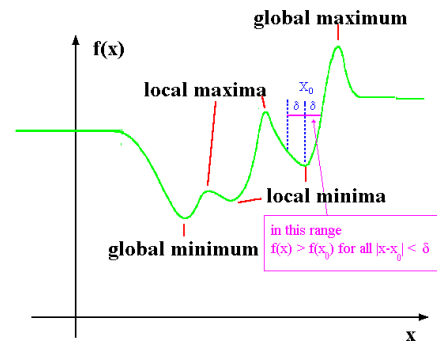## Potential Energy Function

$$PEF(R) = \sum_{bonds} K_b \{b(R) - b_{eq}\}^2 + \sum_{angles} K_\theta \{\theta(R) - \theta_{eq}\}^2 +$$

$$\sum_{dihedrals} \frac{K_\phi}{2} \{1 + \cos[n\phi(R) - \gamma]\} +$$

$$\sum_{\substack{non-bonded \\ atom\ pairs\ i,j}} \left[\frac{A_{ij}}{r_{ij}(R)^{12}} - \frac{B_{ij}}{r_{ij}(R)^6} + \frac{q_i q_j}{\varepsilon_r \varepsilon_0 r_{ij}(R)} \right] \quad (1)$$

Forcefields: AMBER, CHARMM, CVF, ECEPP, GROMOS

Energy Minimazation

f(x)

global maximum

local maxima

x_0

$\delta \mid \delta$

local minima

global minimum

in this range

$f(x) > f(x_0)$ for all $|x - x_0| < \delta$

x

Energy Minimazation

local min

sp   min

sp

global min

X_2

X_1