# Protein Structure Prediction

# Energetics of protein structure

# What is an atom?

- Classical mechanics: a solid object
- Defined by its position (x,y,z), its shape (usually a ball) and its mass
- May carry an electric charge (positive or negative), usually partial (less than an electron)

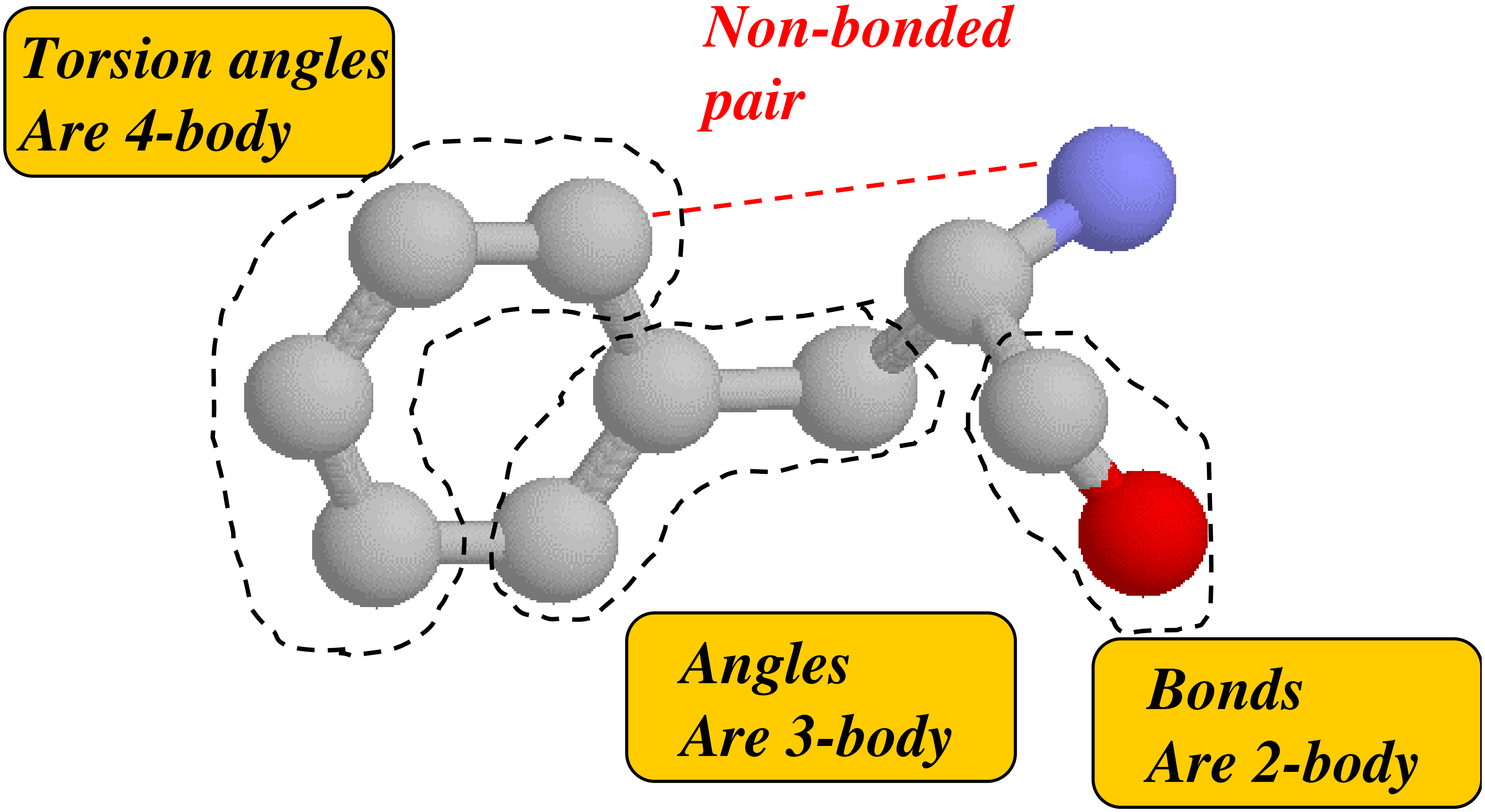
# Example of atom definitions: CHARMM

```
MASS    20 C      12.01100 C ! carbonyl C, peptide backbone
MASS    21 CA     12.01100 C ! aromatic C
MASS    22 CT1    12.01100 C ! aliphatic sp3 C for CH
MASS    23 CT2    12.01100 C ! aliphatic sp3 C for CH2
MASS    24 CT3    12.01100 C ! aliphatic sp3 C for CH3
MASS    25 CPH1   12.01100 C ! his CG and CD2 carbons
MASS    26 CPH2   12.01100 C ! his CE1 carbon
MASS    27 CPT    12.01100 C ! trp C between rings
MASS    28 CY     12.01100 C ! TRP C in pyrrole ring
```

# Example of residue definition: CHARMM

```
RESI ALA          0.00
GROUP
ATOM N    NH1    -0.47  !     |
ATOM HN   H       0.31  !  HN-N
ATOM CA   CT1     0.07  !     |     HB1
ATOM HA   HB      0.09  !     |    /
GROUP                   !  HA-CA--CB-HB2
ATOM CB   CT3    -0.27  !     |    \
ATOM HB1  HA      0.09  !     |     HB3
ATOM HB2  HA      0.09  !   O=C
ATOM HB3  HA      0.09  !     |
GROUP                   !
ATOM C    C       0.51
ATOM O    O      -0.51
BOND CB CA  N  HN  N  CA
BOND C  CA  C  +N  CA HA   CB HB1  CB HB2  CB HB3
DOUBLE O  C
```

# Atomic interactions



***Torsion angles Are 4-body***

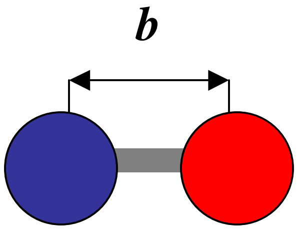
***Non-bonded pair***

***Angles Are 3-body***

***Bonds Are 2-body***

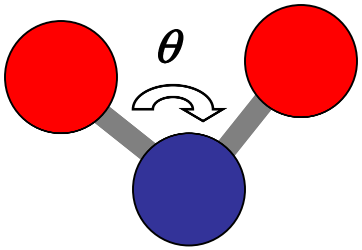
# Forces between atoms

## *Strong bonded interactions*



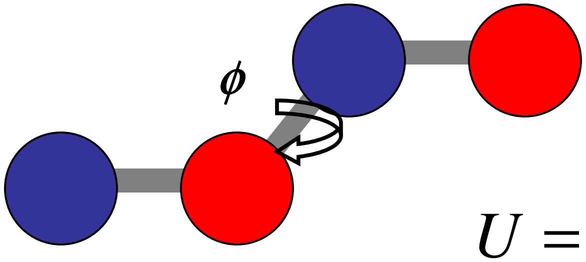
$$U = K(b - b_0)^2$$

All chemical bonds



$$U = K(\theta - \theta_0)^2$$

Angle between chemical bonds

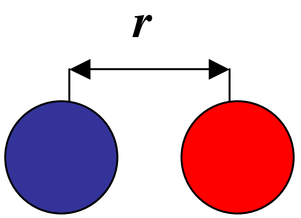


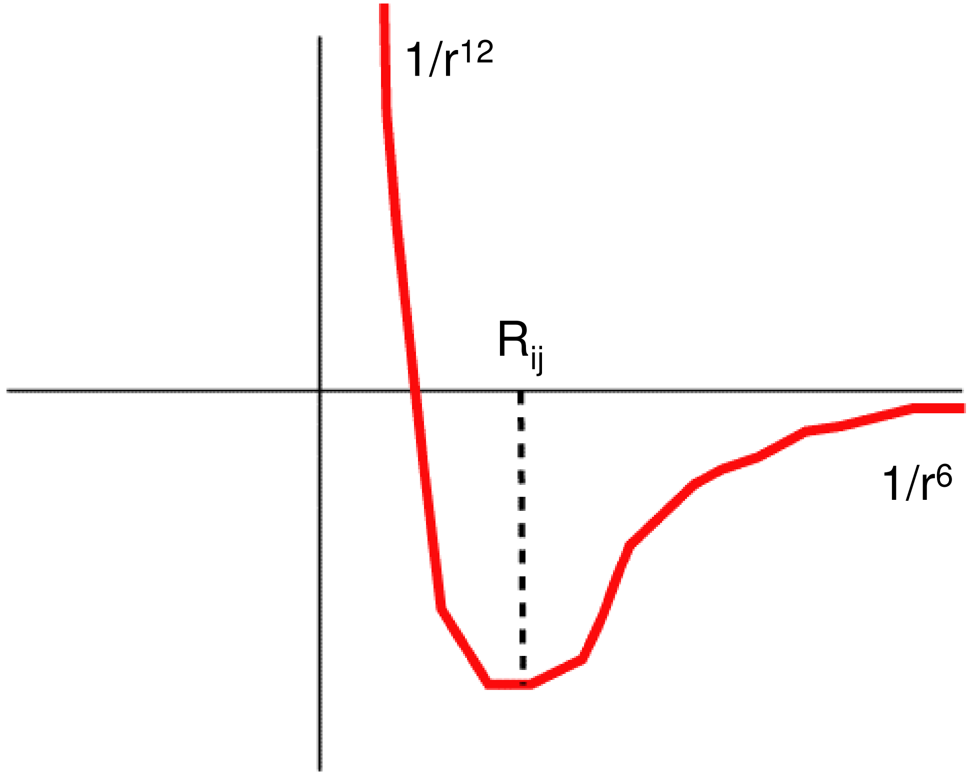
$$U = K(1 - \cos(n\phi))$$

Preferred conformations for Torsion angles:

- $\omega$ angle of the main chain
- $\chi$ angles of the sidechains (aromatic, …)

# Forces between atoms: vdW interactions



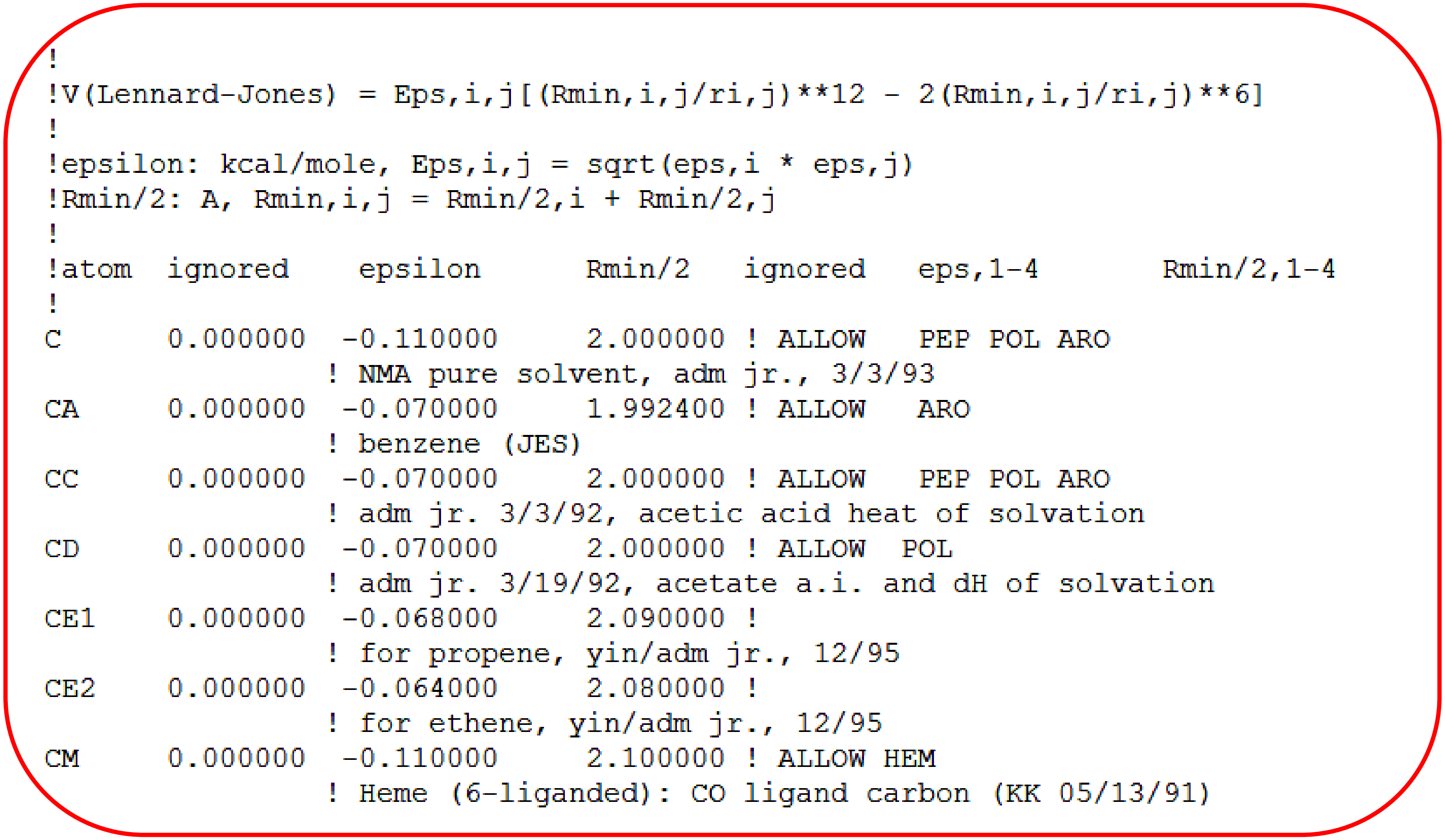


*Lennard-Jones potential*

$$E_{LJ}(r) = \varepsilon_{ij}\left(\left(\frac{R_{ij}}{r}\right)^{12} - 2\left(\frac{R_{ij}}{r}\right)^{6}\right)$$

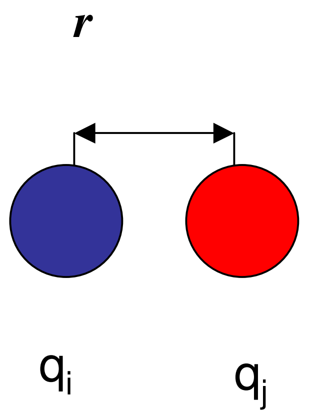
$$R_{ij} = \frac{R_i + R_j}{2}; \quad \varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$$

# Example: LJ parameters in CHARMM



```
!
!V(Lennard-Jones) = Eps,i,j[(Rmin,i,j/ri,j)**12 - 2(Rmin,i,j/ri,j)**6]
!
!epsilon: kcal/mole, Eps,i,j = sqrt(eps,i * eps,j)
!Rmin/2: A, Rmin,i,j = Rmin/2,i + Rmin/2,j
!
!atom  ignored    epsilon      Rmin/2   ignored   eps,1-4       Rmin/2,1-4
!
C      0.000000  -0.110000     2.000000 ! ALLOW   PEP POL ARO
                ! NMA pure solvent, adm jr., 3/3/93
CA     0.000000  -0.070000     1.992400 ! ALLOW   ARO
                ! benzene (JES)
CC     0.000000  -0.070000     2.000000 ! ALLOW   PEP POL ARO
                ! adm jr. 3/3/92, acetic acid heat of solvation
CD     0.000000  -0.070000     2.000000 ! ALLOW  POL
                ! adm jr. 3/19/92, acetate a.i. and dH of solvation
CE1    0.000000  -0.068000     2.090000 !
                ! for propene, yin/adm jr., 12/95
CE2    0.000000  -0.064000     2.080000 !
                ! for ethene, yin/adm jr., 12/95
CM     0.000000  -0.110000     2.100000 ! ALLOW HEM
                ! Heme (6-liganded): CO ligand carbon (KK 05/13/91)
```

# Forces between atoms: Electrostatics interactions



*Coulomb potential*

$$E(r) = \frac{1}{4\pi\varepsilon_0\varepsilon}\frac{q_i q_j}{r}$$

# Some Common force fields in Computational Biology

ENCAD (Michael Levitt, Stanford)

AMBER (Peter Kollman, UCSF; David Case, Scripps)

CHARMM (Martin Karplus, Harvard)

OPLS (Bill Jorgensen, Yale)

MM2/MM3/MM4 (Norman Allinger, U. Georgia)

ECEPP (Harold Scheraga, Cornell)

GROMOS (Van Gunsteren, ETH, Zurich)

*Michael Levitt. The birth of computational structural biology. Nature Structural Biology, 8, 392-393 (2001)*

# Homology Modeling

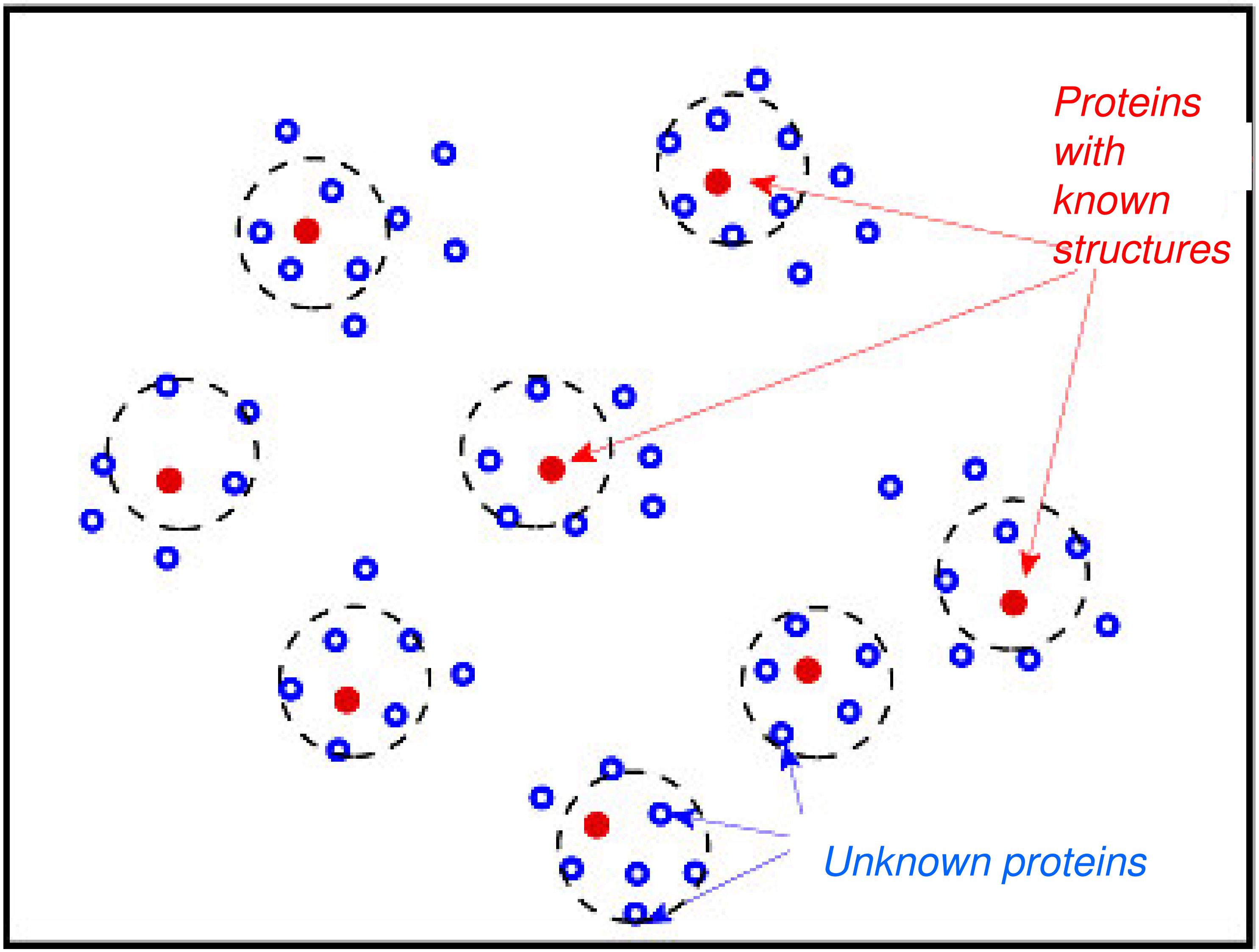
# Structural Genomics project

- *Aim to solve the structure of all proteins: this is too much work experimentally!*

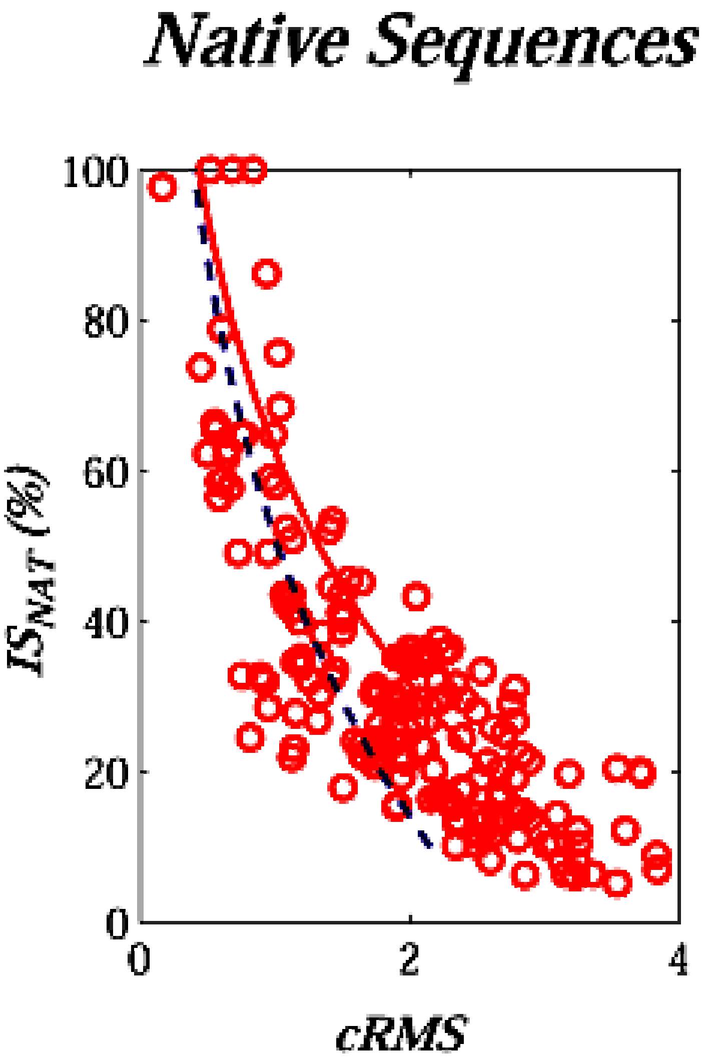
- *Solve enough structures so that the remaining structures can be inferred from those experimental structures*

- *The number of experimental structures needed depend on our abilities to generate a model*.

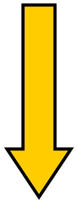
# Structural Genomics



# Homology Modeling: why it works

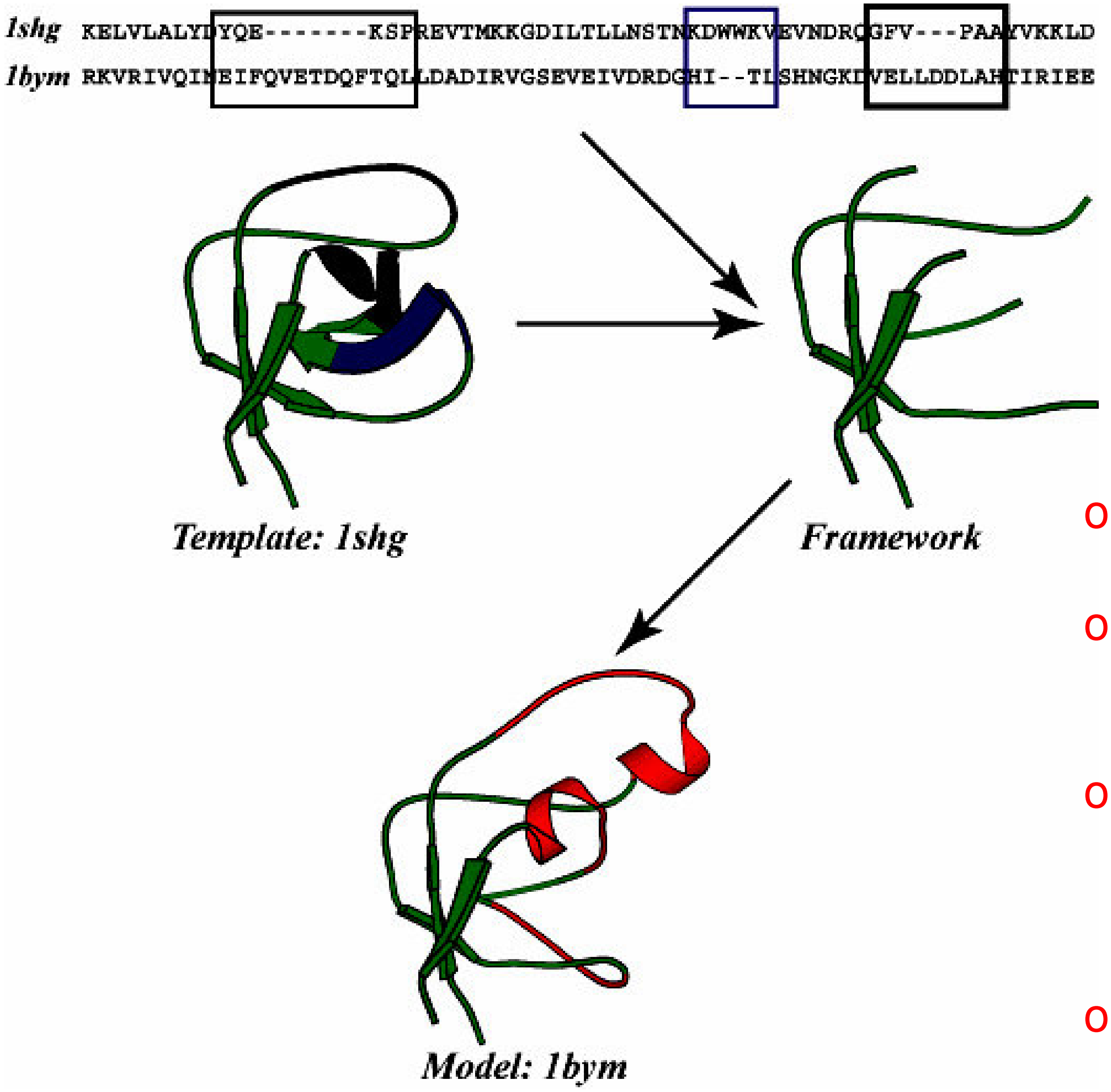


*High sequence identity*



*High structure similarity*

# Homology Modeling: How it works



```
1shg KELVLALYDYQE-------KSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFV---PAAYVKKLD
1bym RKVRIVQINEIFQVETDQFTQLLDADIRVGSEVEIVDRDGHI--TLSHNGKDVELLDDLAHTIRIEE
```

- *Find template*
- *Align target sequence with template*
- *Generate model:*
  - *- add loops*
  - *- add sidechains*
- *Refine model*

# Fold Recognition

*Homology modeling refers to the easy case when the template structure can be identified using BLAST alone.*

What to do when BLAST fails to identify a template?

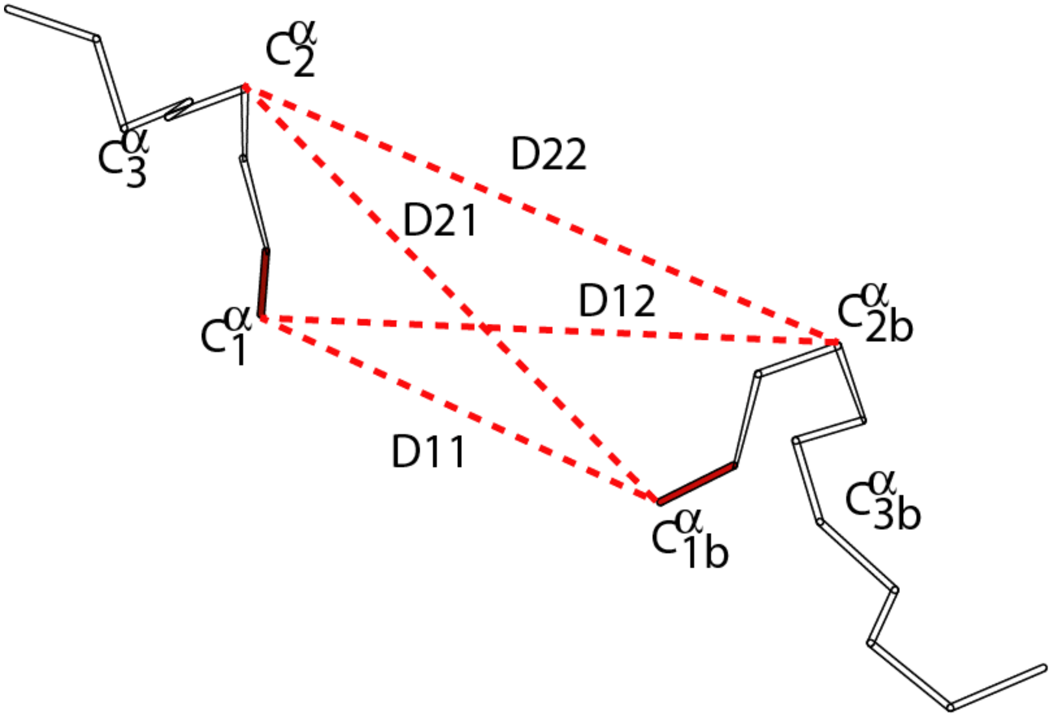
- *Use more sophisticated sequence methods*
  - Profile-based BLAST: PSIBLAST
  - Hidden Markov Models (HMM)

- *Use secondary structure prediction to guide the selection of a template, or to validate a template*

- *Use threading programs: sequence-structure alignments*

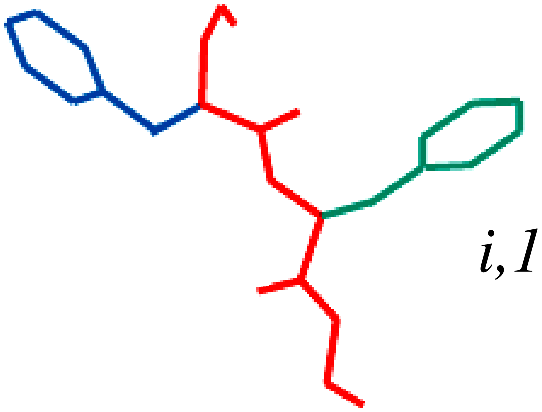
- *Use all of these methods! Meta-servers: http://bioinfo.pl/Meta*

# Loops: A database approach

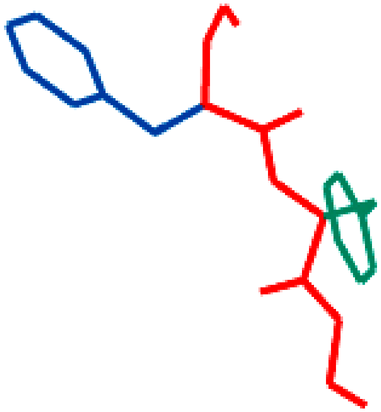


*Scan database and search protein fragments with correct number of residues and correct end-to-end distances*

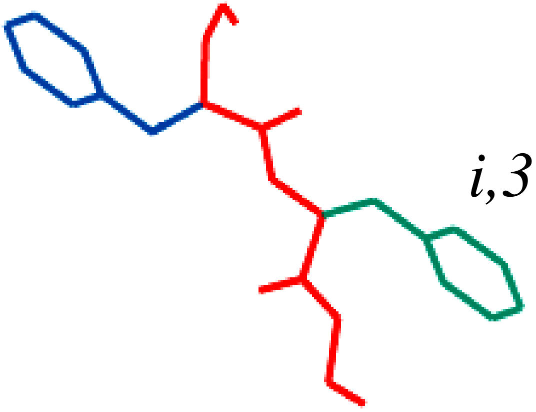
# Self-Consistent Mean-Field Sampling



*i,1*

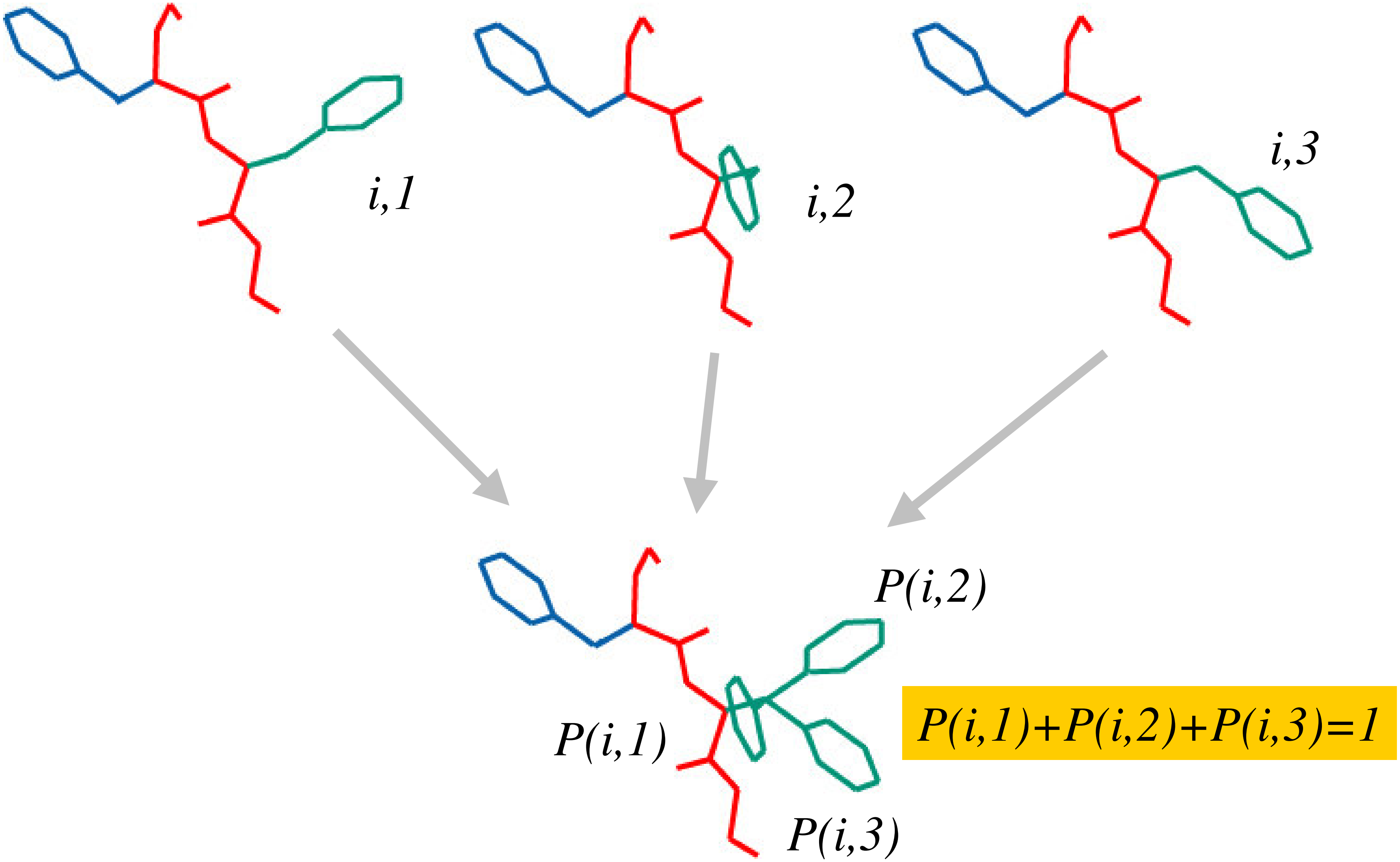


*i,2*



*i,3*

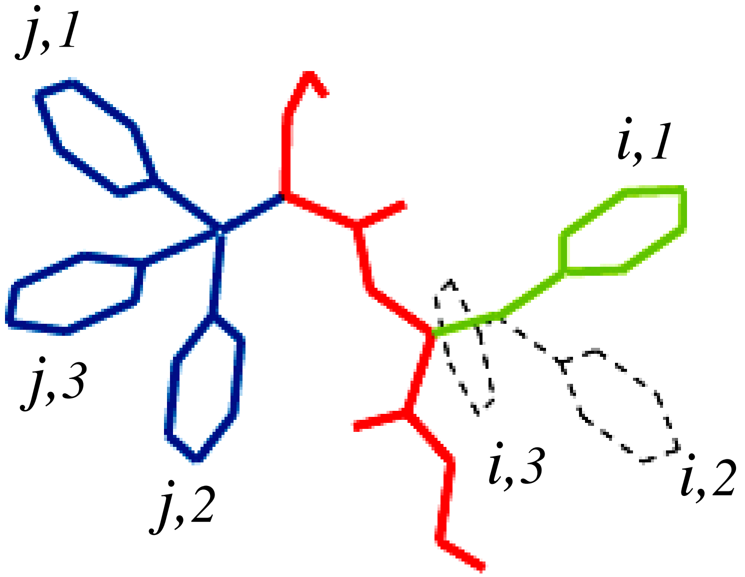
# Self-Consistent Mean-Field Sampling



$P(i,1)+P(i,2)+P(i,3)=1$

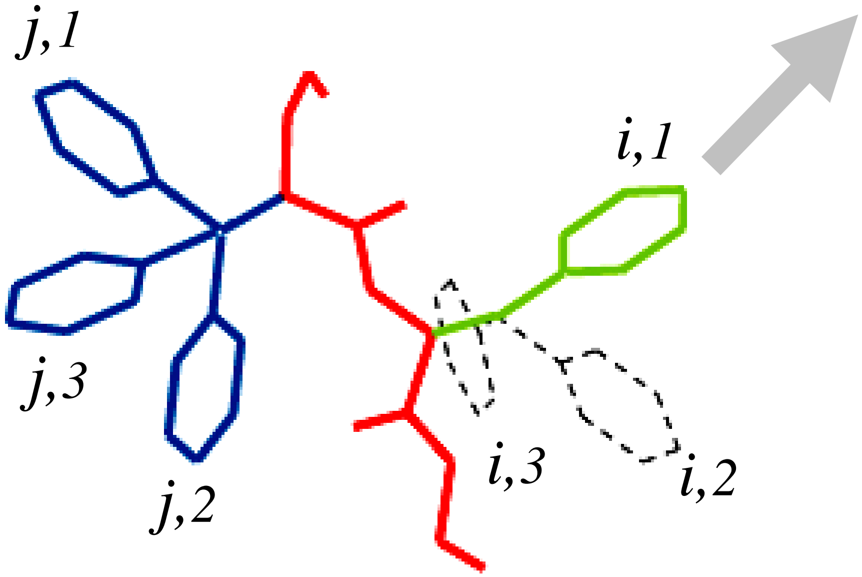
# Self-Consistent Mean-Field Sampling

*Multicopy Protein*



# Self-Consistent Mean-Field Sampling

*Multicopy Protein*

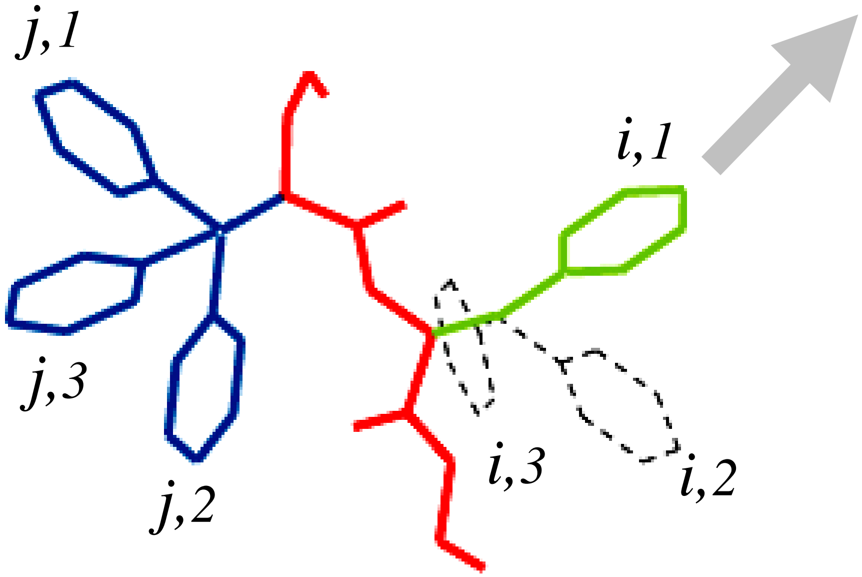


*Mean-Field Energy*

$$E(i,k) = U(i,k) + U(i,k,Backbone) + \sum_{j=1,j\neq i}^{N} \sum_{l=1}^{Nrot(j)} P(j,l)U(i,k,j,l)$$

# Self-Consistent Mean-Field Sampling

*Multicopy Protein*



*Mean-Field Energy*

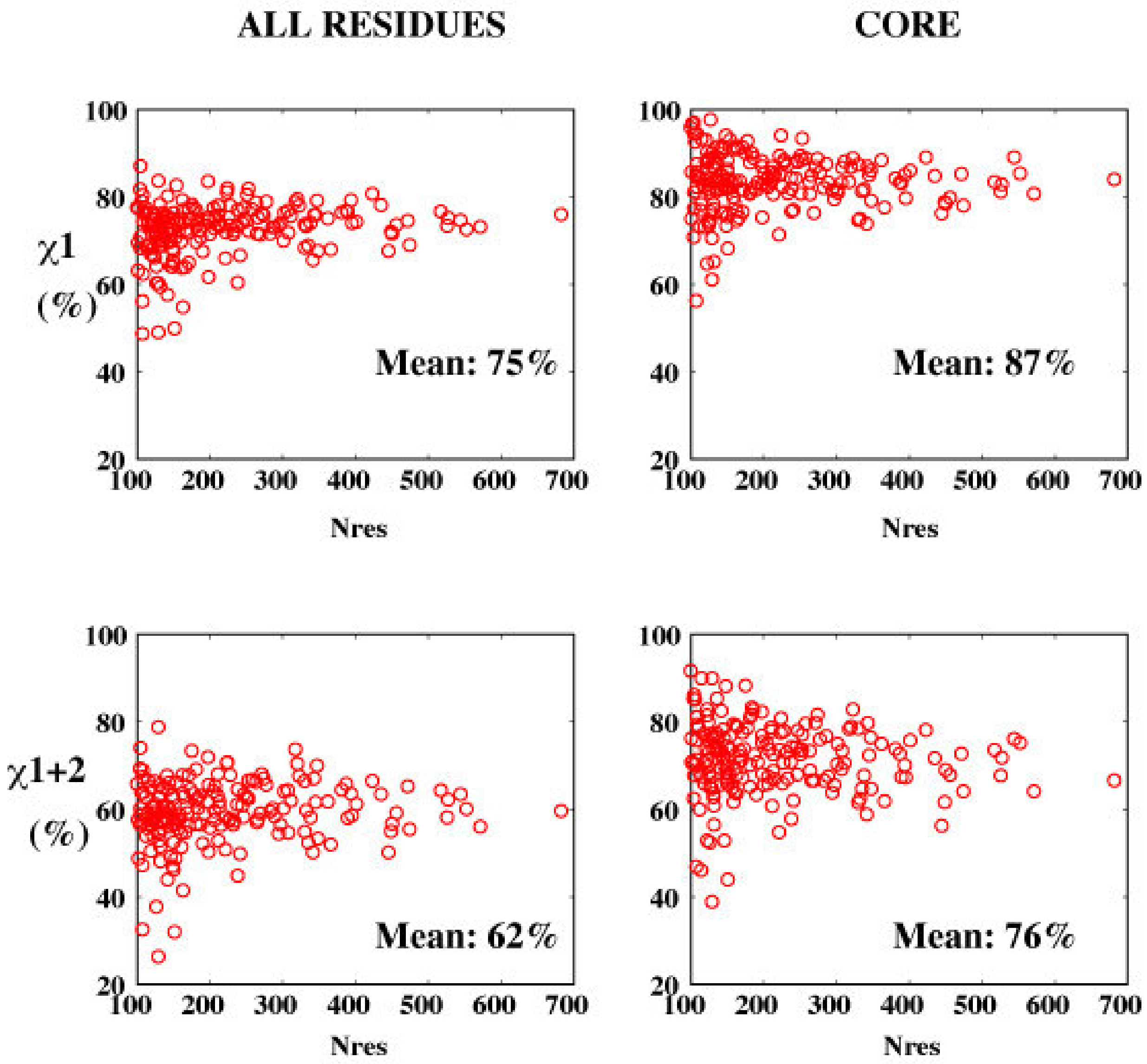
$$E(i,k) = U(i,k) + U(i,k,Backbone) + \sum_{j=1,j\neq i}^{N} \sum_{l=1}^{Nrot(j)} P(j,l)U(i,k,j,l)$$

*Update Cycle*

$$P_{new}(i,k) = \frac{exp\left[-\frac{E(i,k)}{RT}\right]}{\sum_{l=1}^{Nrot(i)} exp\left[-\frac{E(i,l)}{RT}\right]}$$

***(Koehl and Delarue, J. Mol. Biol., 239:249-275 (1994))***

# Self-Consistent Mean-Field Sampling



ALL RESIDUES

CORE

# Refinement ?

CASP5 assessors, homology modeling category:

> "We are forced to draw the disappointing conclusion that, similarly to what observed in previous editions of the experiment, no model resulted to be closer to the target structure than the template to any significant extent."

The consensus is not to refine the model, as refinement usually pulls the model away from the native structure!!

# The CASP experiment

- *CASP= Critical Assessment of Structure Prediction*
- *Started in 1994, based on an idea from John Moult (Moult, Pederson, Judson, Fidelis, Proteins, 23:2-5 (1995))*
- *First run in 1994; now runs regularly every second year (CASP6 was held last december)*

# The CASP experiment: how it works

*1) Sequences of target proteins are made available to CASP participants in June-July of a CASP year*

*- the structure of the target protein is know, but not yet released in the PDB, or even accessible*

*2) CASP participants have between 2 weeks and 2 months over the summer of a CASP year to generate up to 5 models for each of the target they are interested in.*

*3) Model structures are assessed against experimental structure*

*4) CASP participants meet in December to discuss results*

# CASP Statistics

| Experiment | # of Targets | # of predictors | # of 3D models |
|---|---|---|---|
| CASP1 | 33 | 35 | 100 |
| CASP2 | 42 | 72 | 947 |
| CASP3 | 43 | 61 | 1256 |
| CASP4 | 43 | 111 | 5150 |
| CASP5 | 67 | 175 | 22909 |
| CASP6 | 87 | 166 | 28965 |

# CASP

*Three categories at CASP*

- Homology (or comparative) modeling

- Fold recognition

- Ab initio prediction

*CASP dynamics:*

- Real deadlines; pressure: positive, or negative?

- Competition?

- Influence on science ?

Venclovas, Zemla, Fidelis, Moult. Assessment of progress over the CASP experiments.
Proteins, 53:585-595 (2003)

# Homology Modeling: Practical guide

*Approach 1: Manual*

- Submit target sequence to BLAST; identify potential templates

- For each template:

  - Generate alignment between target and template (Smith-Waterman + manual correction)

  - Build framework

  - build loop + sidechain

  - assess model (stereochemistry, …)

# Homology Modeling: Practical guide

*Approach 2:* Submit target sequence to automatic servers

- *Fully automatic*:
  - 3D-Jigsaw : http://www.bmm.icnet.uk/servers/3djigsaw/
  - EsyPred3D: http://www.fundp.ac.be/urbm/bioinfo/esypred/
  - SwissModel: http://swissmodel.expasy.org//SWISS-MODEL.html
- *Fold recognition*:
  - PHYRE: http://www.sbg.bio.ic.ac.uk/~phyre/
- *Useful sites:*
  - Meta server: http://bioinfo.pl/Meta
  - PredictProtein: http://cubic.bioc.columbia.edu/predictprotein/

# Secondary Structure Prediction

- *Given a protein sequence $a_1a_2\ldots a_N$, secondary structure prediction aims at defining the state of each amino acid ai as being either H (helix), E (extended=strand), or O (other) (Some methods have 4 states: H, E, T for turns, and O for other).*
- *The quality of secondary structure prediction is measured with a "3-state accuracy" score, or $Q_3$. $Q_3$ is the percent of residues that match "reality" (X-ray structure).*

# Secondary Structure Assignment

Determine Secondary Structure positions in known protein structures using DSSP or STRIDE:

1. Kabsch and Sander. Dictionary of Secondary Structure in Proteins: pattern recognition of hydrogen-bonded and geometrical features.
   Biopolymer 22: 2571-2637 (1983) (DSSP)
2. Frischman and Argos. Knowledge-based secondary structure assignments.
   Proteins, 23:566-571 (1995) (STRIDE)

# Early methods for Secondary Structure Prediction

- *Chou and Fasman*

  (Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974)

- *GOR*

  (Garnier, Osguthorpe and Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol., 120:97-120, 1978)

# Chou and Fasman

- *Start by computing amino acids propensities to belong to a given type of secondary structure:*

$$\frac{P(i/Helix)}{P(i)} \qquad \frac{P(i/Beta)}{P(i)} \qquad \frac{P(i/Turn)}{P(i)}$$

Propensities > 1 mean that the residue type I is likely to be found in the Corresponding secondary structure type.

# Chou and Fasman

| Amino Acid | α-Helix | β-Sheet | Turn | |
|---|---|---|---|---|
| Ala | 1.29 | 0.90 | 0.78 | Favors α-Helix |
| Cys | 1.11 | 0.74 | 0.80 | |
| Leu | 1.30 | 1.02 | 0.59 | |
| Met | 1.47 | 0.97 | 0.39 | |
| Glu | 1.44 | 0.75 | 1.00 | |
| Gln | 1.27 | 0.80 | 0.97 | |
| His | 1.22 | 1.08 | 0.69 | |
| Lys | 1.23 | 0.77 | 0.96 | |
| Val | 0.91 | 1.49 | 0.47 | Favors β-strand |
| Ile | 0.97 | 1.45 | 0.51 | |
| Phe | 1.07 | 1.32 | 0.58 | |
| Tyr | 0.72 | 1.25 | 1.05 | |
| Trp | 0.99 | 1.14 | 0.75 | |
| Thr | 0.82 | 1.21 | 1.03 | |
| Gly | 0.56 | 0.92 | 1.64 | Favors turn |
| Ser | 0.82 | 0.95 | 1.33 | |
| Asp | 1.04 | 0.72 | 1.41 | |
| Asn | 0.90 | 0.76 | 1.23 | |
| Pro | 0.52 | 0.64 | 1.91 | |
| Arg | 0.96 | 0.99 | 0.88 | |

# Chou and Fasman

*Predicting helices:*

- find nucleation site: 4 out of 6 contiguous residues with $P(\alpha)>1$
- extension: extend helix in both directions until a set of 4 contiguous residues has an average $P(\alpha) < 1$ (breaker)
- if average $P(\alpha)$ over whole region is >1, it is predicted to be helical

Predicting strands:

- find nucleation site: 3 out of 5 contiguous residues with $P(\beta)>1$
- extension: extend strand in both directions until a set of 4 contiguous residues has an average $P(\beta) < 1$ (breaker)
- if average $P(\beta)$ over whole region is >1, it is predicted to be a strand

# Chou and Fasman

*Position-specific parameters for turn:*

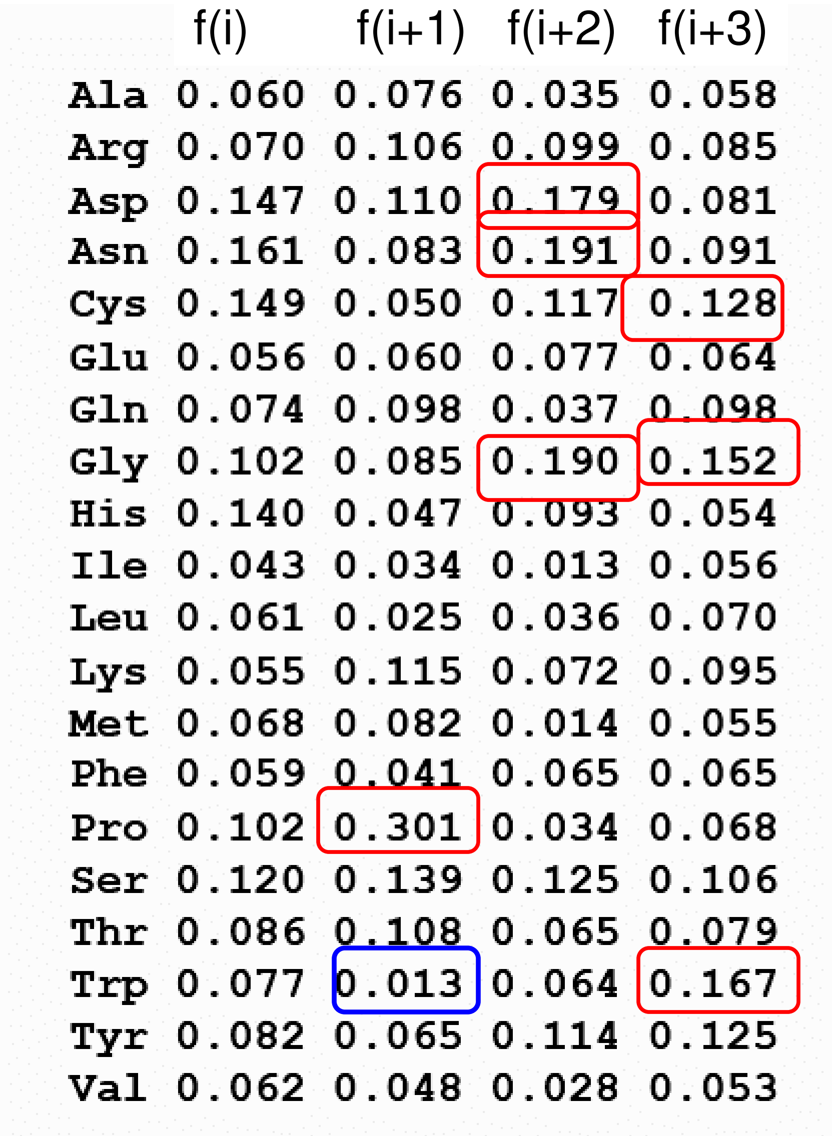
Each position has distinct amino acid preferences.

Examples:

-At position 2, Pro is highly preferred; Trp is disfavored

-At position 3, Asp, Asn and Gly are preferred

-At position 4, Trp, Gly and Cys preferred



|  | f(i) | f(i+1) | f(i+2) | f(i+3) |
|---|---|---|---|---|
| Ala | 0.060 | 0.076 | 0.035 | 0.058 |
| Arg | 0.070 | 0.106 | 0.099 | 0.085 |
| Asp | 0.147 | 0.110 | 0.179 | 0.081 |
| Asn | 0.161 | 0.083 | 0.191 | 0.091 |
| Cys | 0.149 | 0.050 | 0.117 | 0.128 |
| Glu | 0.056 | 0.060 | 0.077 | 0.064 |
| Gln | 0.074 | 0.098 | 0.037 | 0.098 |
| Gly | 0.102 | 0.085 | 0.190 | 0.152 |
| His | 0.140 | 0.047 | 0.093 | 0.054 |
| Ile | 0.043 | 0.034 | 0.013 | 0.056 |
| Leu | 0.061 | 0.025 | 0.036 | 0.070 |
| Lys | 0.055 | 0.115 | 0.072 | 0.095 |
| Met | 0.068 | 0.082 | 0.014 | 0.055 |
| Phe | 0.059 | 0.041 | 0.065 | 0.065 |
| Pro | 0.102 | 0.301 | 0.034 | 0.068 |
| Ser | 0.120 | 0.139 | 0.125 | 0.106 |
| Thr | 0.086 | 0.108 | 0.065 | 0.079 |
| Trp | 0.077 | 0.013 | 0.064 | 0.167 |
| Tyr | 0.082 | 0.065 | 0.114 | 0.125 |
| Val | 0.062 | 0.048 | 0.028 | 0.053 |

# Chou and Fasman

*Predicting turns*:

- for each tetrapeptide starting at residue i, compute:
  - $P_{Turn}$ (average propensity over all 4 residues)
  - $F = f(i)*f(i+1)*f(i+2)*f(i+3)$

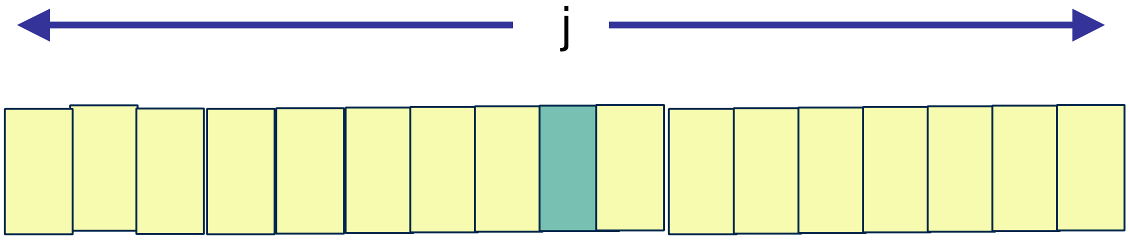
- if $P_{Turn} > P\alpha$ and $P_{Turn} > P\beta$ and $P_{Turn} > 1$ and $F>0.000075$ tetrapeptide is considered a turn.

## Chou and Fasman prediction:

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

# The GOR method

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.



A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries.
Build similar tables for strands and turns.

*GOR simplification:*
The predicted state of AAj is calculated as the sum of the position-dependent propensities of all residues around AAj.

GOR can be used at : http://abs.cit.nih.gov/gor/ (current version is GOR IV)

# Accuracy

- Both Chou and Fasman and GOR have been assessed and their accuracy is estimated to be Q3=60-65%.

# Secondary Structure Prediction

-*Available servers*:

- JPRED : http://www.compbio.dundee.ac.uk/~www-jpred/
- PHD: http://cubic.bioc.columbia.edu/predictprotein/
- PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/
- NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
- Chou and Fassman: http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

-*Interesting paper:*

- *Rost and Eyrich. EVA: Large-scale analysis of secondary structure prediction. Proteins 5:192-199 (2001)*

# Protein Structure Prediction

- *One popular model for protein folding assumes a sequence of events:*
  - Hydrophobic collapse
  - Local interactions stabilize secondary structures
  - Secondary structures interact to form motifs
  - Motifs aggregate to form tertiary structure

# Protein Structure Prediction

*A physics-based approach:*

- find conformation of protein corresponding to a thermodynamics minimum (free energy minimum)

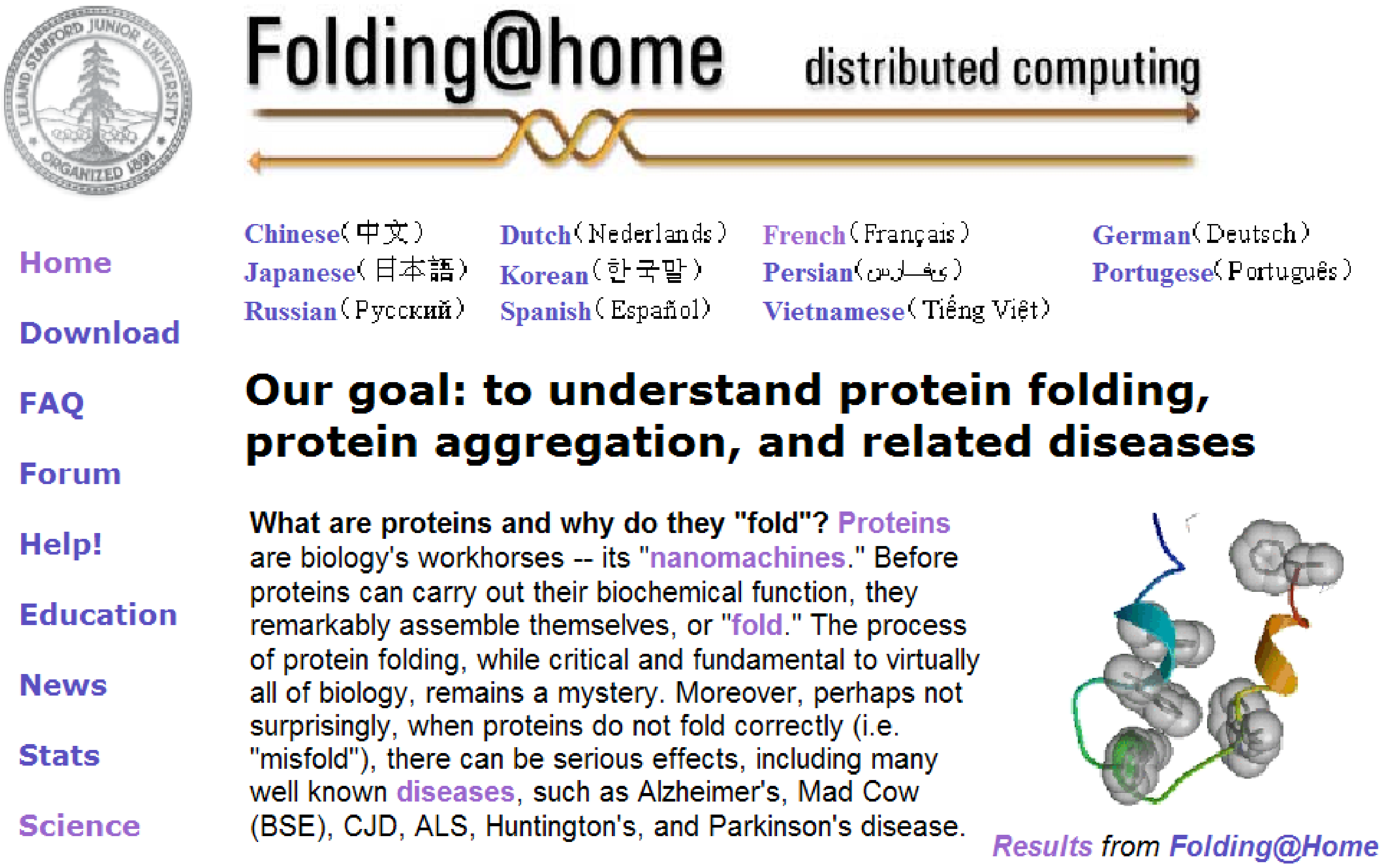
- cannot minimize internal energy alone!
  Needs to include solvent

- simulate folding…a very long process!

Folding time are in the ms to second time range
Folding simulations at best run 1 ns in one day…

# The Folding @ Home initiative

*(Vijay Pande, Stanford University)*



Folding@home distributed computing

Chinese(中文) Dutch(Nederlands) French(Français) German(Deutsch)
Japanese(日本語) Korean(한국말) Persian(فارسی) Portugese(Português)
Russian(Русский) Spanish(Español) Vietnamese(Tiếng Việt)

- Home
- Download
- FAQ
- Forum
- Help!
- Education
- News
- Stats
- Science

**Our goal: to understand protein folding, protein aggregation, and related diseases**

**What are proteins and why do they "fold"?** Proteins are biology's workhorses -- its "nanomachines." Before proteins can carry out their biochemical function, they remarkably assemble themselves, or "fold." The process of protein folding, while critical and fundamental to virtually all of biology, remains a mystery. Moreover, perhaps not surprisingly, when proteins do not fold correctly (i.e. "misfold"), there can be serious effects, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, and Parkinson's disease.

*Results from Folding@Home*

*http://folding.stanford.edu/*

# The Folding @ Home initiative

**What does Folding@Home do?** Folding@Home is a distributed computing project which studies protein folding, misfolding, aggregation, and related diseases. We use novel computational methods and large scale distributed computing, to simulate timescales thousands to millions of times longer than previously achieved. This has allowed us to simulate folding for the first time, and to now direct our approach to examine folding related disease.

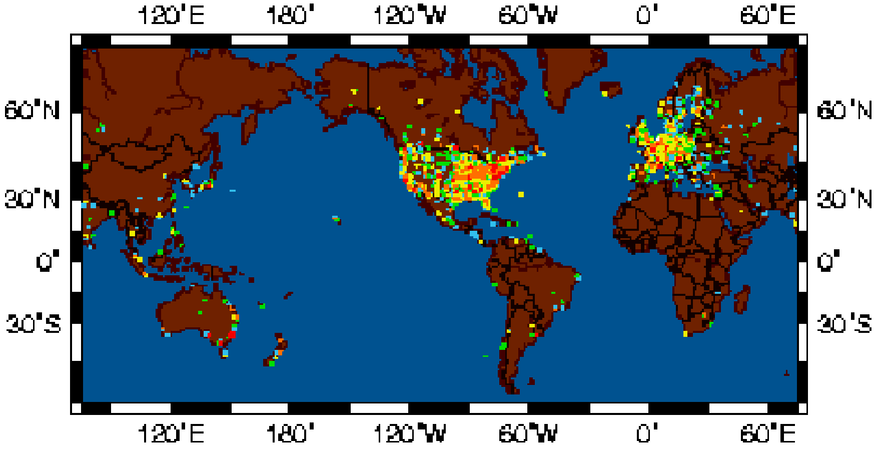
F@H exhibit exploratorium

See Prof. Pande's lecture on F@H at Xerox PARC parc

**How can you help?** You can help our project by downloading and running our client software. Our algorithms are designed such that for every computer that joins the project, we get a commensurate increase in simulation speed.

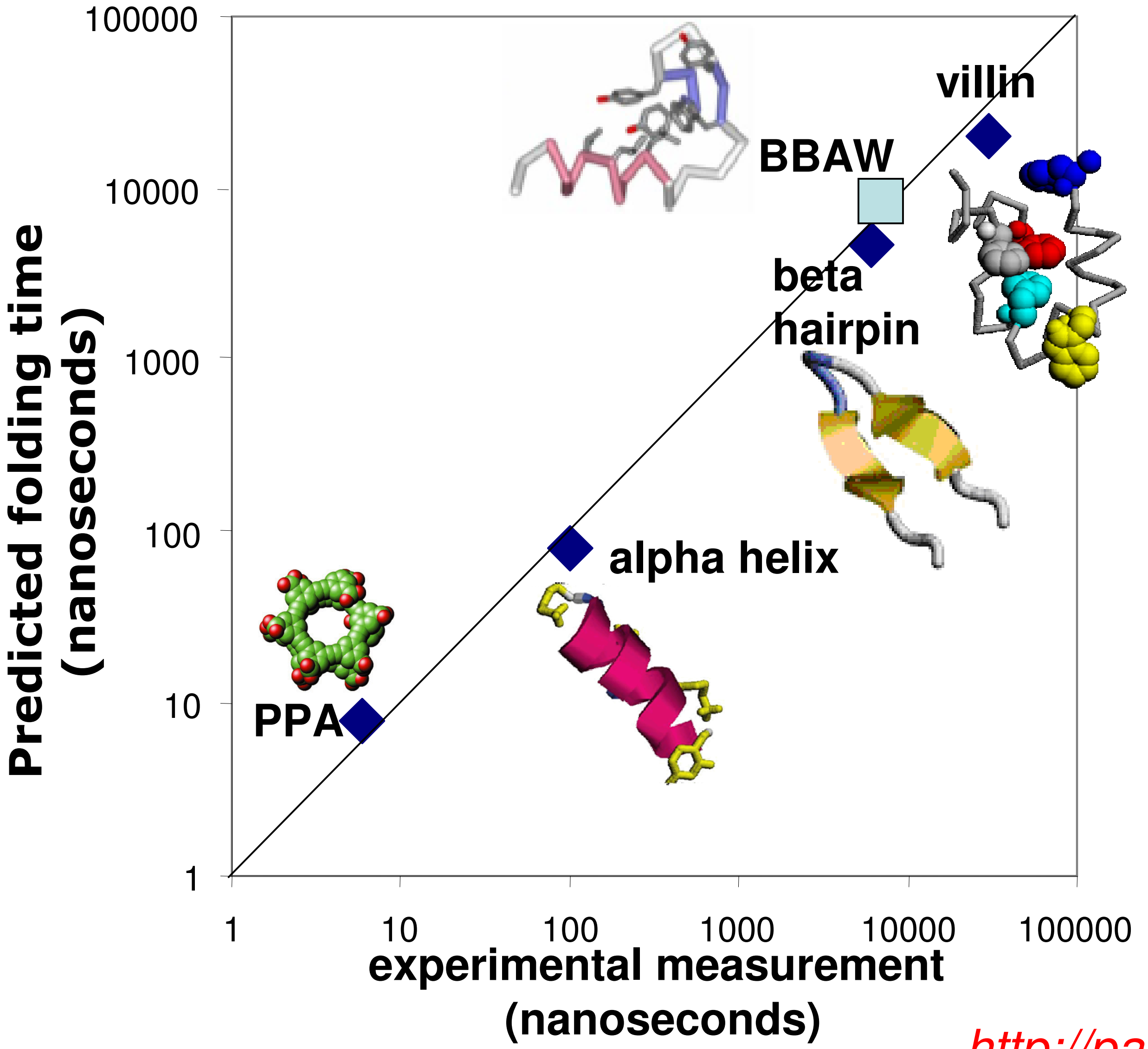
One can also help by donating funds to the project, via Stanford University.

**What have we done so far?** We have had several successes. You can read about them on our Science page, Results section, or go directly to our press and papers page.



*Since October 1, 2000, over 1,000,000 CPUs* ***throughout the world*** *have participated in Folding@Home. Each additional CPU gives us an added boost in performance, allowing us to tackle more difficult problems or solve existing research faster or more accurately.*

# Folding @ Home: Results



## Experiments:

**villin:**
Raleigh, et al,
SUNY, Stony Brook

**BBAW:**
Gruebele, et al, UIUC

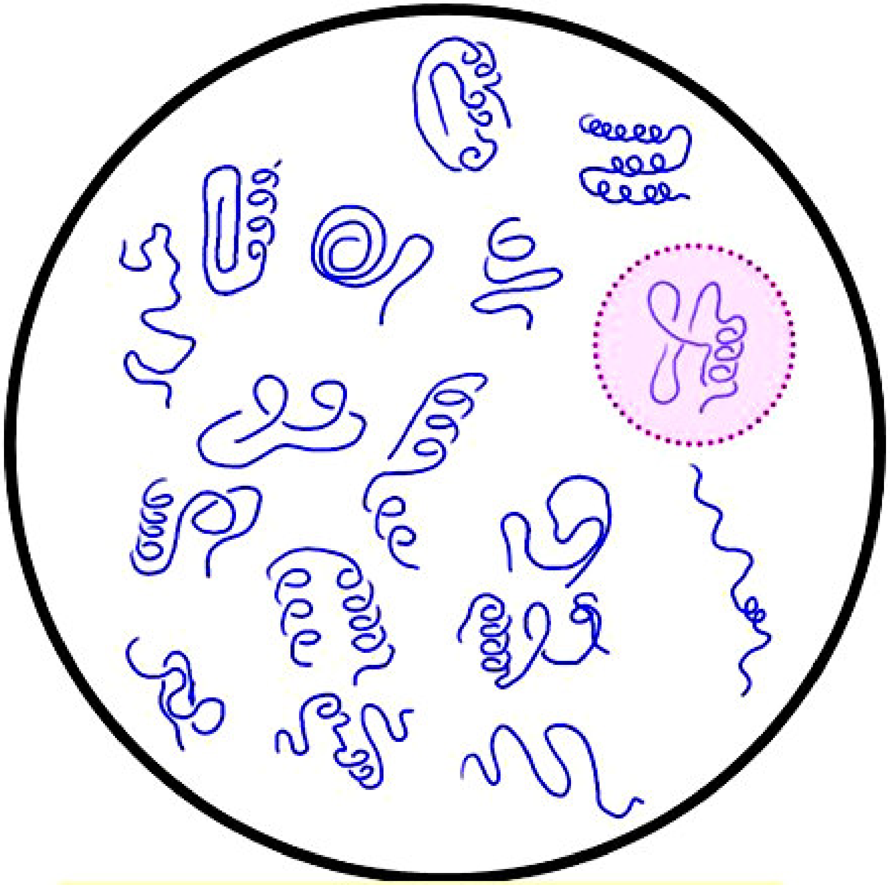
**beta hairpin:**
Eaton, et al, NIH

**alpha helix:**
Eaton, et al, NIH

**PPA:**
Gruebele, et al, UIUC

*http://pande.stanford.edu/*

# Protein Structure Prediction



DECOYS:

Generate a large number of possible shapes

DISCRIMINATION:

Select the correct, native-like fold

***Need good decoy structures***

***Need a good energy function***