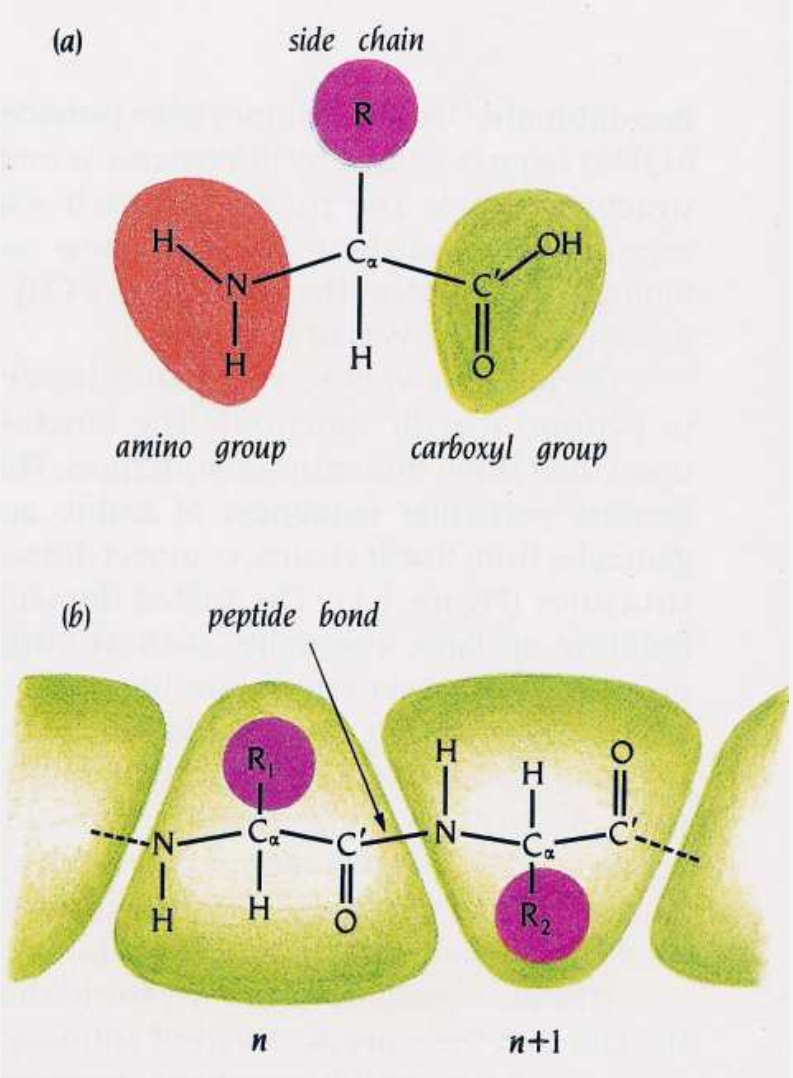
Teaching Computers to Fold Proteins

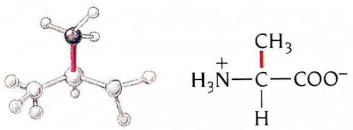
Anders Krogh

The Bioinformatics Centre
University of Copenhagen

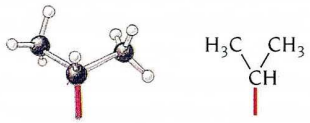
Preprint: Winther & Krogh, arxiv.org/abs/cond-mat/0309497
(see also binf.ku.dk/krogh/)

Proteins consists of amino acids

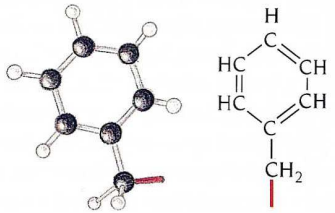




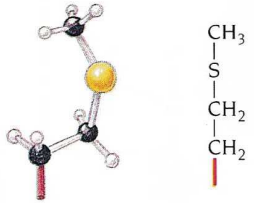
A Ala, Alanine



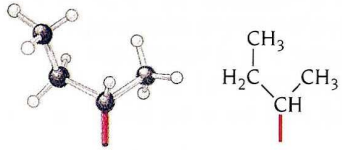
V Val, Valine



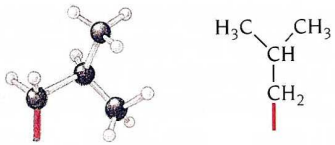
F Phe, Phenylalanine



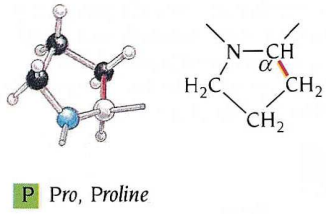
M Met, Methionine



I Ile. Isoleucine

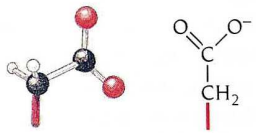


L Leu, Leucine

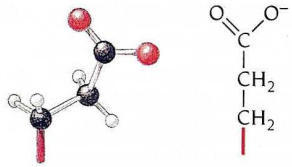


P Pro, Proline

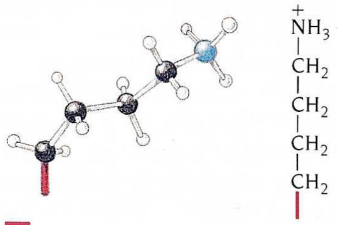
G Gly, Glycine



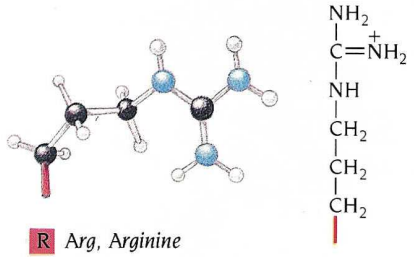
D Asp, Aspartic acid



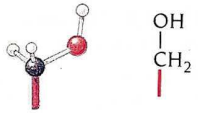
E Glu, Glutamic acid



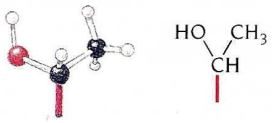
K Lys, Lysine



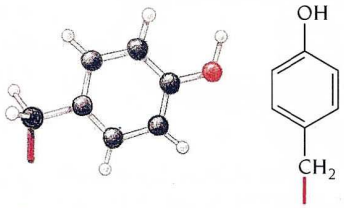
R Arg, Arginine



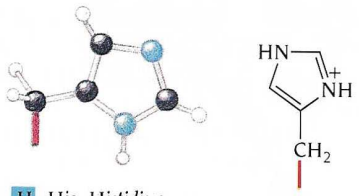
S Ser, Serine



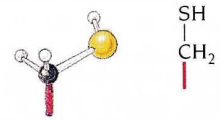
T Thr, Threonine



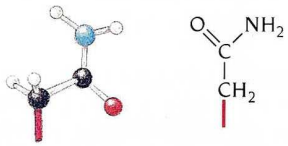
Y Tyr, Tyrosine



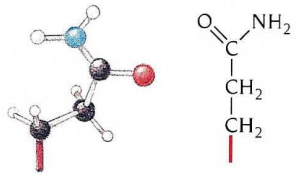
H His, Histidine



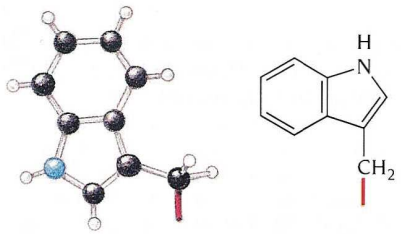
C Cys, Cysteine



N Asn, Asparagine



Q Gln, Glutamine

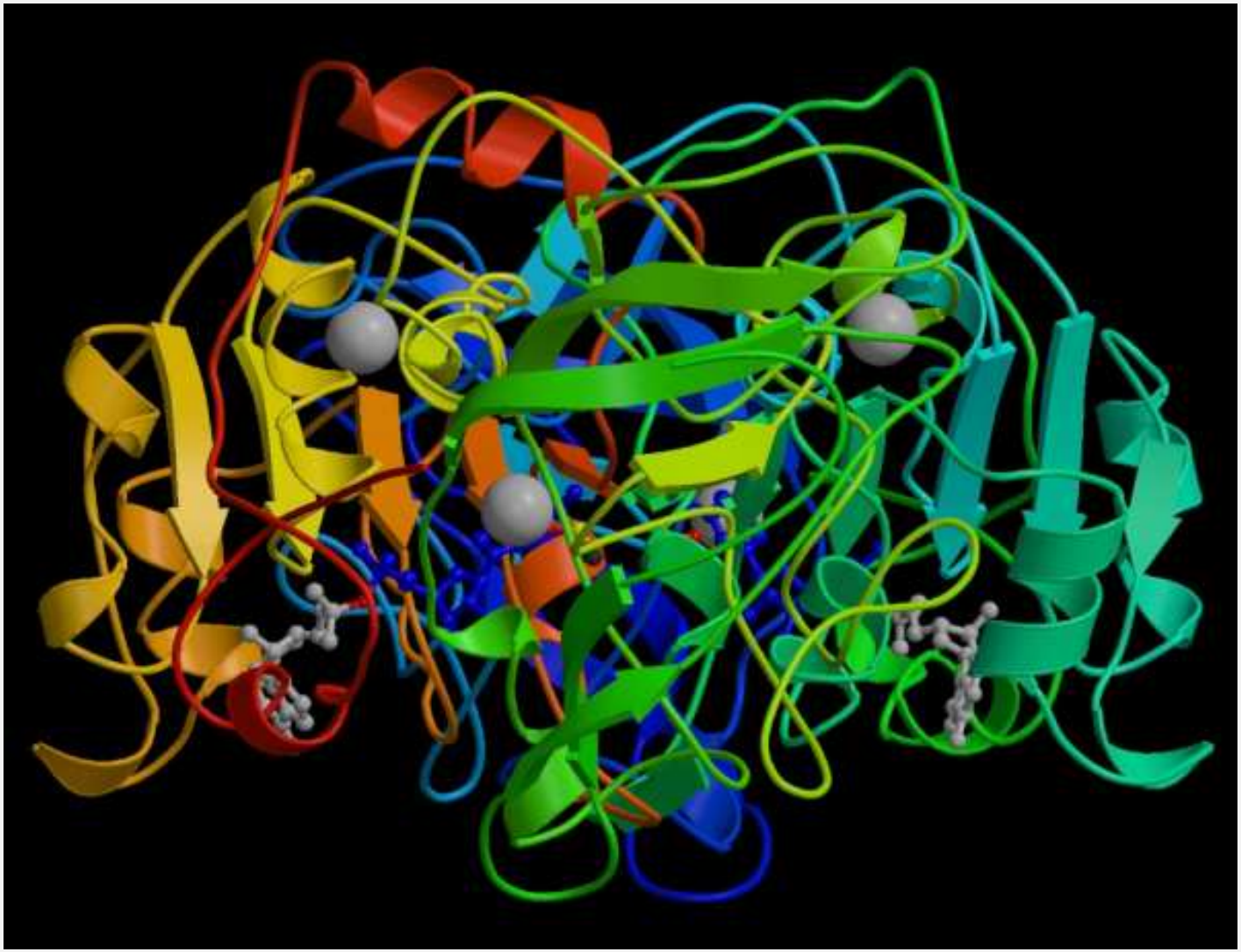


W Trp, Tryptophan

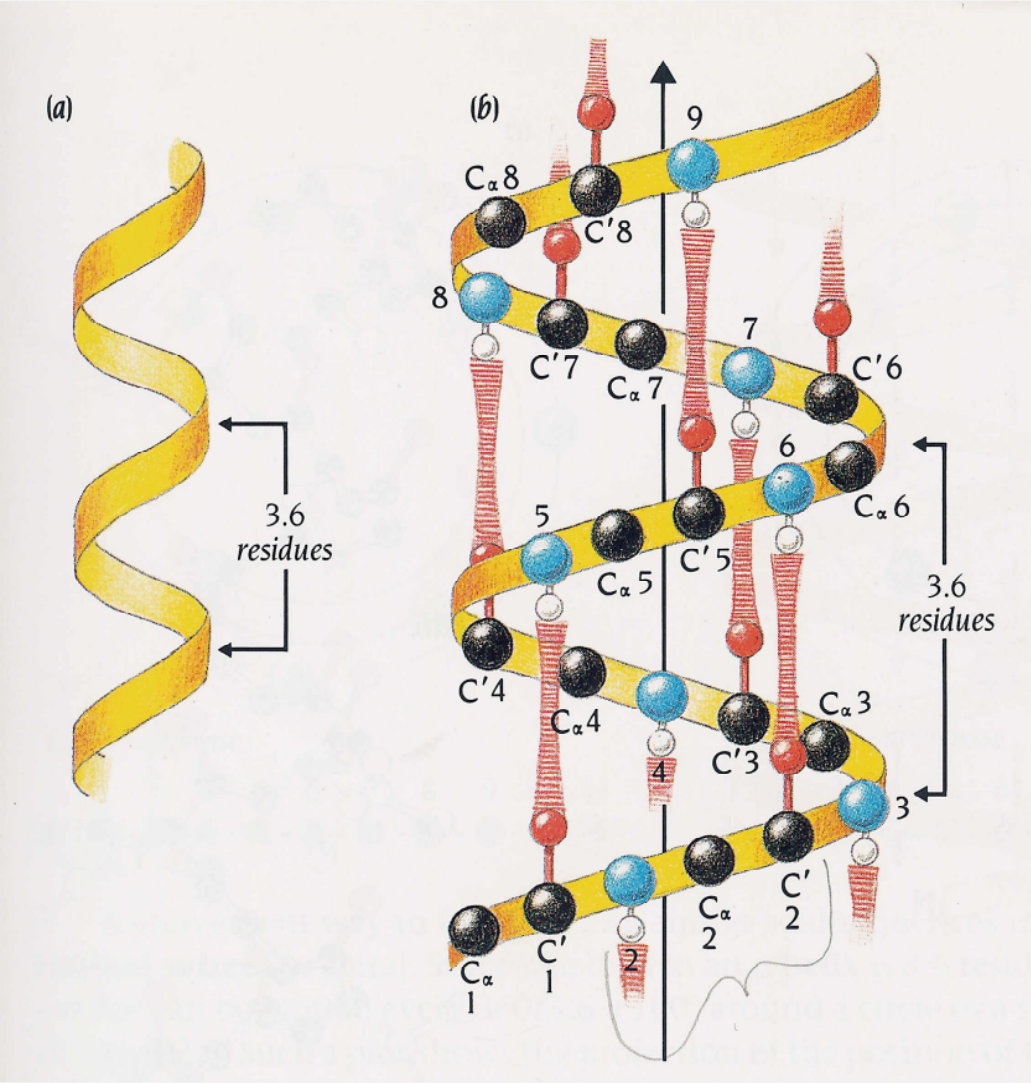
Proteins can be written as sequences of letters

```
>6ADH:A HOLO-LIVER ALCOHOL DEHYDROGENASE
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTP
LPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPR
GTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGY
GSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC
VNPQDYKKPIQEVLTEMSNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFD
LLRSGESIRTILTF
```

But it is the structure that matters



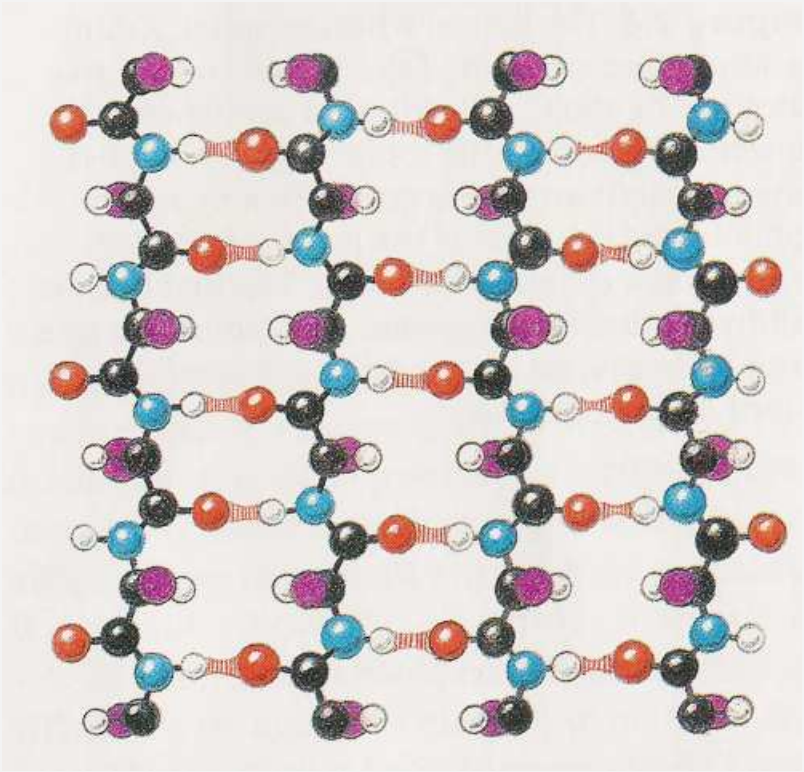
Alpha helices



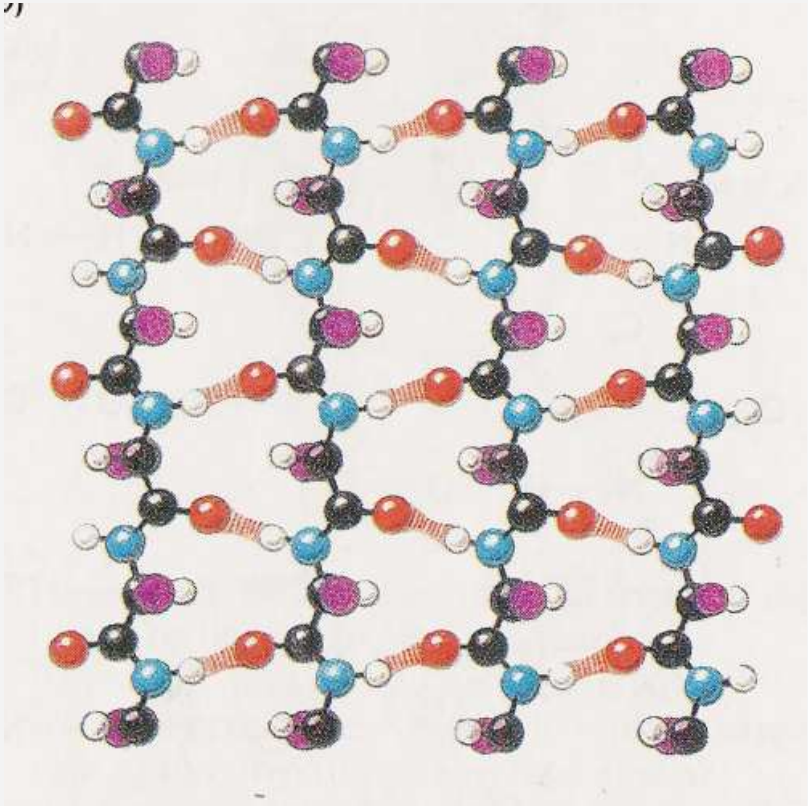
ab006

Beta sheets

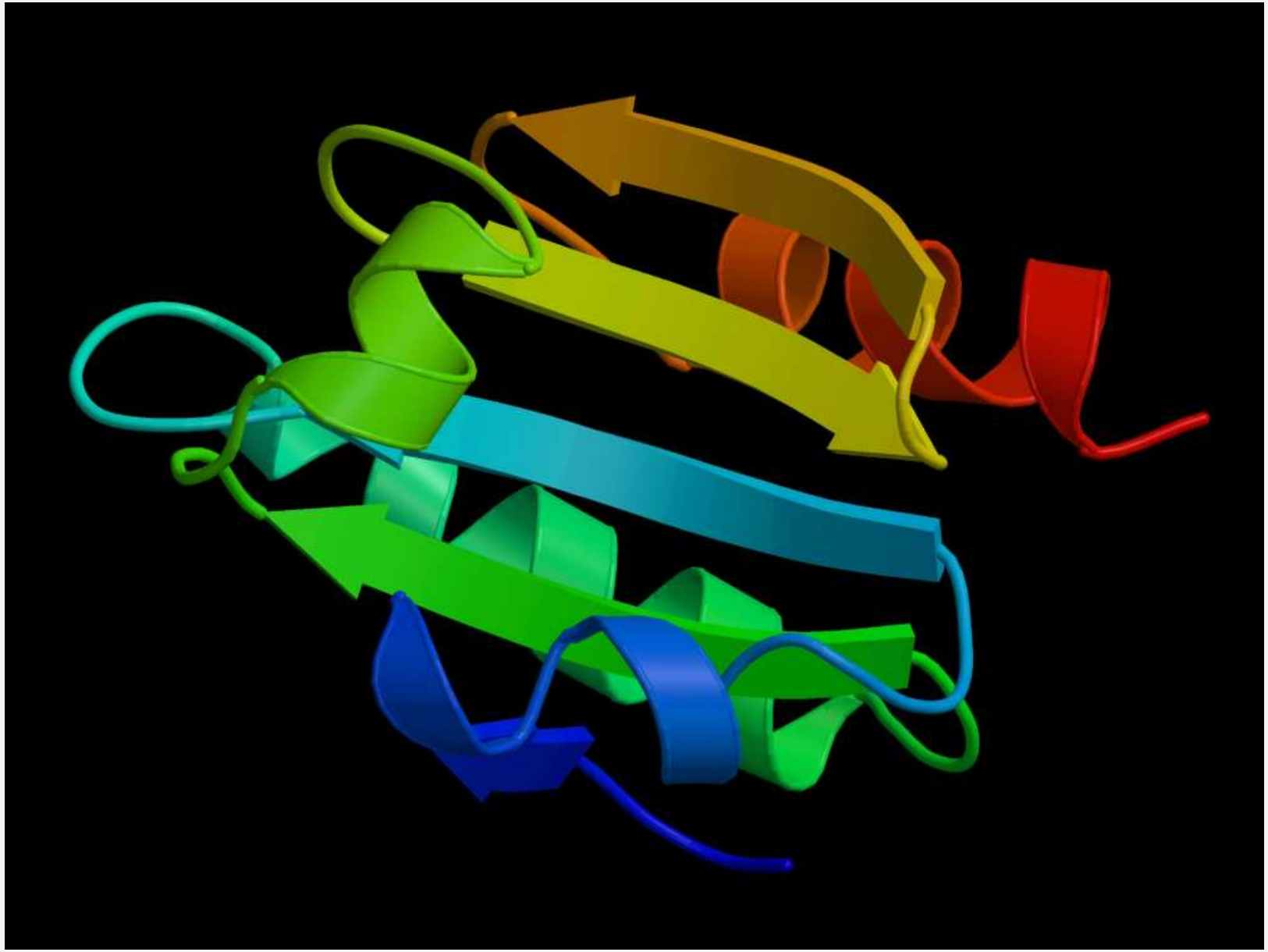
Antiparallel β sheet



Parallel β sheet



1THX



ab008

Experimental structure determination

by X-ray crystallography or NMR is difficult and expensive

The number of known structures $<<$ the number of known sequences

Experimental structure determination

by X-ray crystallography or NMR is difficult and expensive

The number of known structures $<<$ the number of known sequences

$\longrightarrow$ Predict protein structure from sequence

— but this is also difficult

Experimental structure determination

by X-ray crystallography or NMR is difficult and expensive

The number of known structures $<<$ the number of known sequences

$\longrightarrow$ Predict protein structure from sequence

Different approaches:

- Prediction of secondary structure
- Fold recognition

Experimental structure determination

by X-ray crystallography or NMR is difficult and expensive

The number of known structures $<<$ the number of known sequences

$\longrightarrow$ Predict protein structure from sequence

Different approaches:

- Prediction of secondary structure
- Fold recognition
- Abinitio prediction

Why is it difficult to simulate protein folding?

- Too many degrees of freedom
  Only a few nanoseconds can be simulated for reasonably sized proteins with all atoms (+solvent).
- Obtaining correct potentials
  Correct potentials involve quantum mechanical calculations.

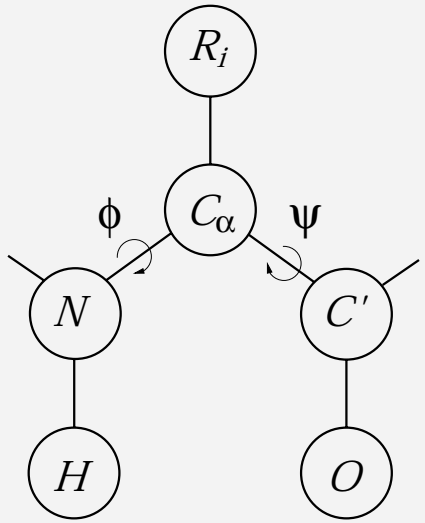
Why is it difficult to simulate protein folding?

- Too many degrees of freedom
  Only a few nanoseconds can be simulated for reasonably sized proteins with all atoms (+solvent).
- Obtaining correct potentials
  Correct potentials involve quantum mechanical calculations.

Solution

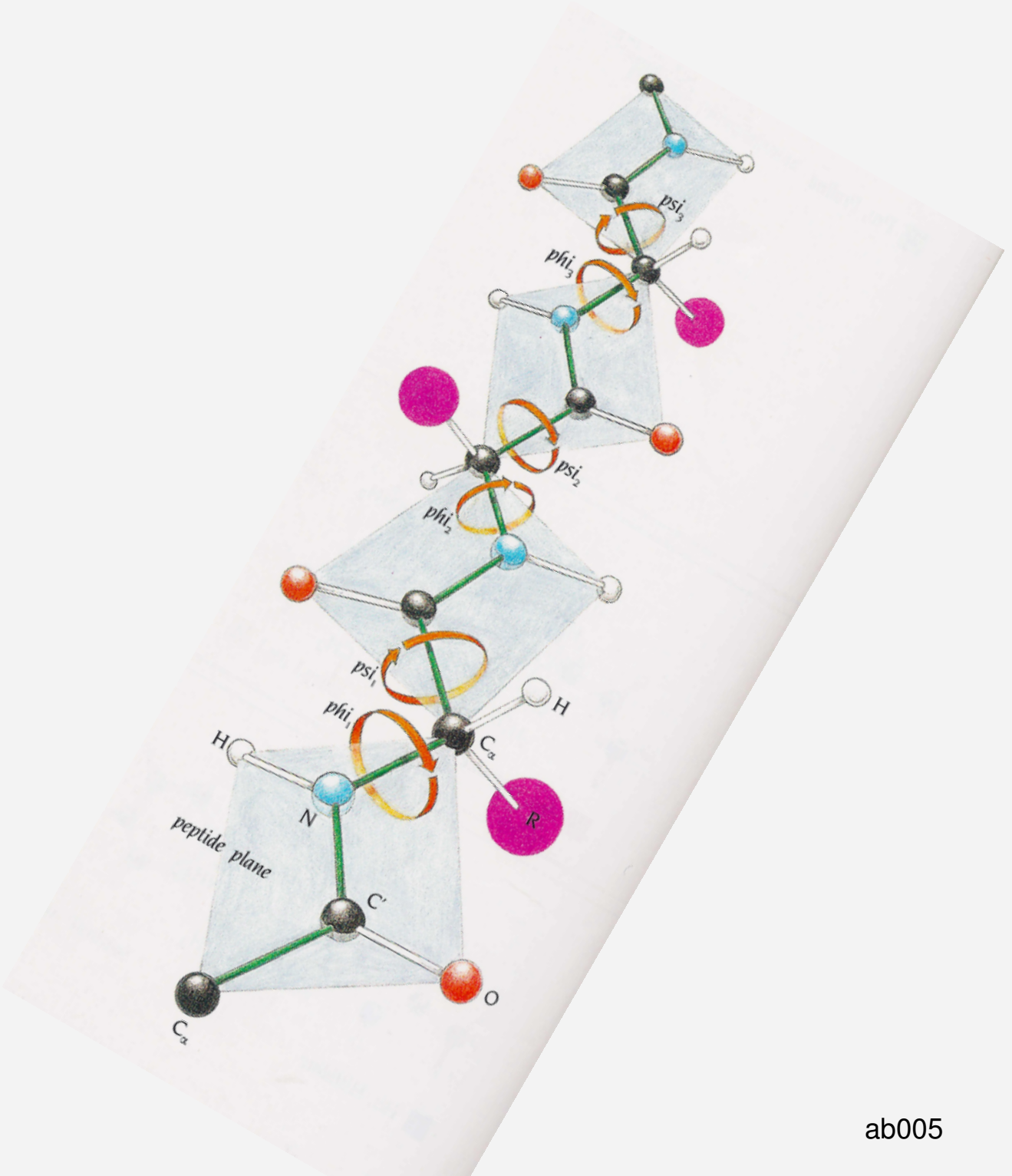
Use a simplified representation of amino acids:

a side chain is represented by a single atom with variable distance to C_α.



The polypeptide backbone

ϕ and ψ angles are the main degrees of freedom



ab005

New problem:

Potentials even harder to calculate!

Therefore various “data driven” and machine learning approaches have been used to estimate pair potentials from known 3D structures.

Best known are “statistical potentials”, which derives an amino acid pair potential from the length distribution in PDB for the two amino acids.

We propose a maximum likelihood approach to the problem.

Remember physics?

Protein seq with atomic coordinates $\mathbf{R}$.

Energy function $E(\mathbf{R}, \text{seq})$

Probability that the protein is in a particular conformation $\mathbf{R}_0$ (Boltzmann-Gibbs distribution):

$$P(\mathbf{R}_0) = \frac{\exp(-\beta E(\mathbf{R}_0, \text{seq}))}{Z}$$

Where $\beta = (k_{\text{Boltzmann}} T)^{-1}$ and the partition function Z is a normalizing constant:

$$Z = \int \exp(-\beta E(\mathbf{R}, \text{seq})) \, d\mathbf{R}$$

Maximum Likelihood

Problem: We don't know the energy.
Assume we have a parametrized energy $E_\theta(\mathbf{R}, \text{seq})$.
We need to find the correct parameters, θ.

Training set with known structure: $(\text{seq}_1, \text{seq}_2, \ldots)$.

Maximize the the probability of the native structures w.r.t. the parameters

$$\max_\theta \prod_i P(\text{nat}_i|\text{seq}_i, \theta)$$

with

$$P(\text{nat}_i|\text{seq}_i, \theta) = \frac{\int_{\text{nat}_i} \exp(-\beta E_\theta(\mathbf{R}, \text{seq}_i))\ d\mathbf{R}}{\int \exp(-\beta E_\theta(\mathbf{R}, \text{seq}_i))\ d\mathbf{R}}$$

nat_i means a volume around the native structure (hypersphere).

Maximize by gradient ascent

Use gradient ascent on $\log(P)$:

$$\theta^{\text{new}} := \theta^{\text{old}} + \eta \nabla_\theta \sum_i \log P(\text{nat}_i|\text{seq}_i, \theta),$$

where η is the "learning rate".

$$\nabla_\theta \sum_i \log P(\text{nat}_i|\text{seq}_i, \theta) =$$
$$\beta \sum_i [\langle \nabla_\theta E_\theta(\mathbf{R}, \text{seq}_i) \rangle - \langle \nabla_\theta E_\theta(\mathbf{R}, \text{seq}_i) \rangle_{\text{nat}_i}]$$

The $\langle \ldots \rangle_{\text{nat}_i}$ is an average over the volume around the native structure.

Learning Algorithm

1. Simulate folding of the whole training set and sample the gradient (w.r.t. the parameters).
2. Simulate folding of the whole training set *around the native structure* and sample the gradient.
3. Average the gradients. Update the parameters of the potential.
4. Stop if you are happy (or ran out of time).
   Otherwise go to 1.

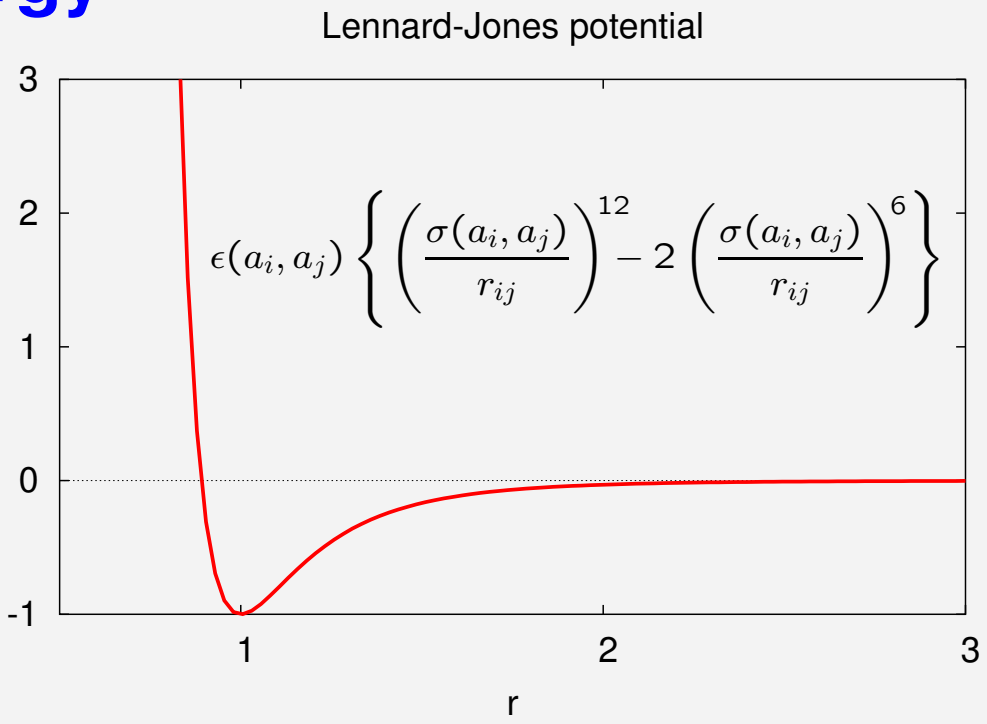
This is similar to Boltzman learning:

Learn the native structure and unlearn false maxima.

The Parametrized Energy

- Lennard-Jones potential between side chains a_i and a_j

  Free parameters σ (radius) and ϵ (depth) depends on side chains.



- Similar potential for $C_\alpha - C_\alpha$ interactions.
- 12-10 Lennard-Jones potential + angular term for hydrogen bonds (between $N - H$ and $C' - O$)
- Steric constraints: local interactions between neighboring amino acids
- Surface energy term: “counts” the number of atoms in a neighborhood

The Gradient – an example

The term in the gradient from $\epsilon(a_i, a_j)$ is easy to calculate.

It results in this parameter update:

$$\epsilon^{\text{new}}(a_i, a_j) = \epsilon(a_i, a_j) + \frac{\eta}{\epsilon(a_i, a_j)} \sum_n \left(\langle E_p^n(a_i, a_j) \rangle - \langle E_p^n(a_i, a_j) \rangle_{\text{nat}_n} \right)$$

(The sum is over the training set.)

If the native energy on average is lower (=better) than the free-running, $\epsilon(a_i, a_j)$ is increased, and vice versa.

Simulations

We use Monte Carlo simulations:

1. the protein is perturbed (changing only ϕ, ψ angles)
2. energy change $\triangle E$ is calculated
3. the new conformation is accepted if $\triangle E <= 0$
   or with probability $exp(-\beta \triangle E)$ otherwise.

Parallel tempering: the system is simulated independently at a number of different temperatures. Sometimes the temperature of two random systems are exchanged with the Metropolis probability.

This gives a much better sampling of the system.

Test

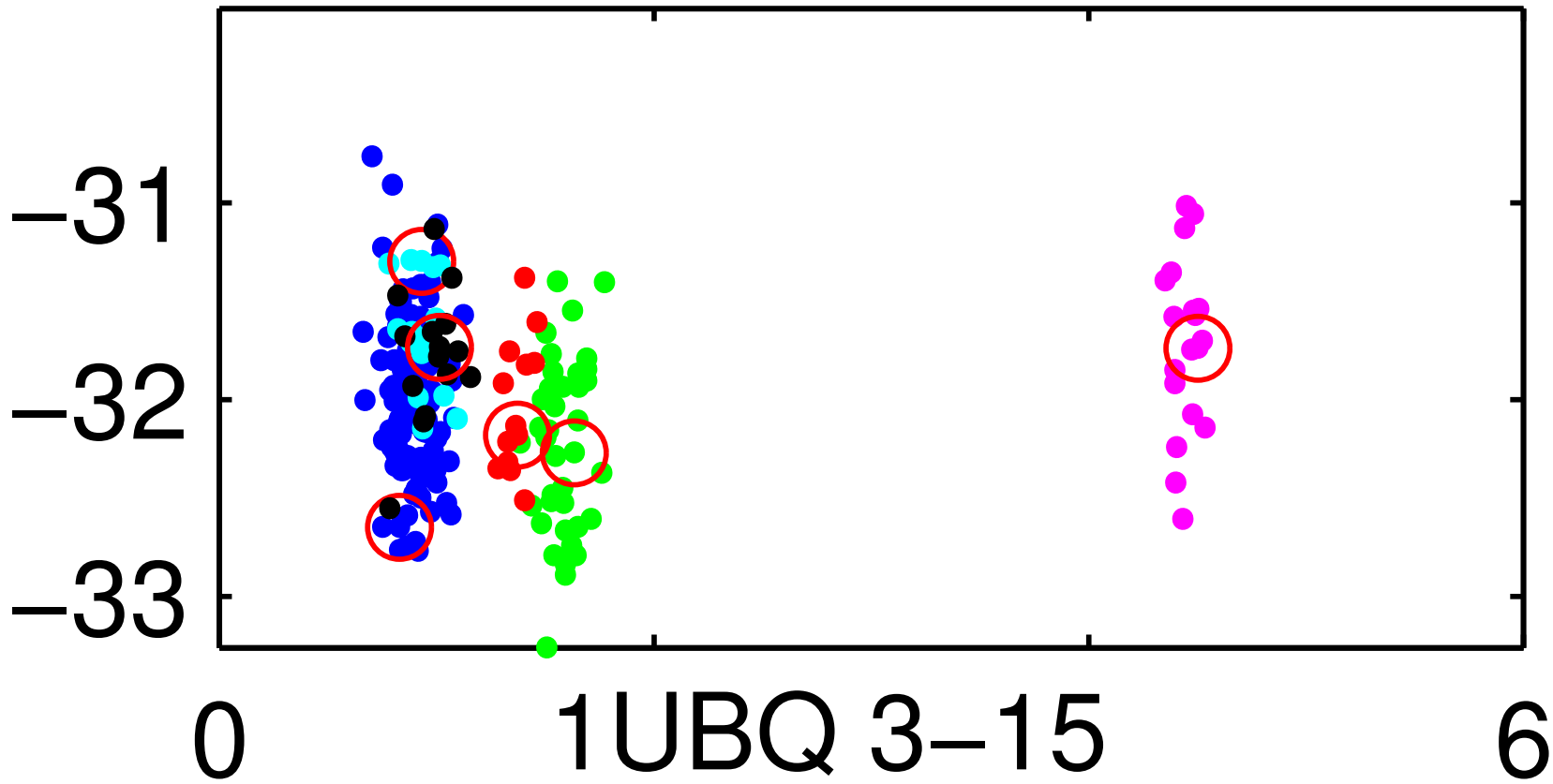
To see if the procedure works, it was initially tested on a small training set of 24 peptides of length 11–24.

Distributed processing: easy to parallelize for multiple processors (8 processor SGI, Origin 3000)

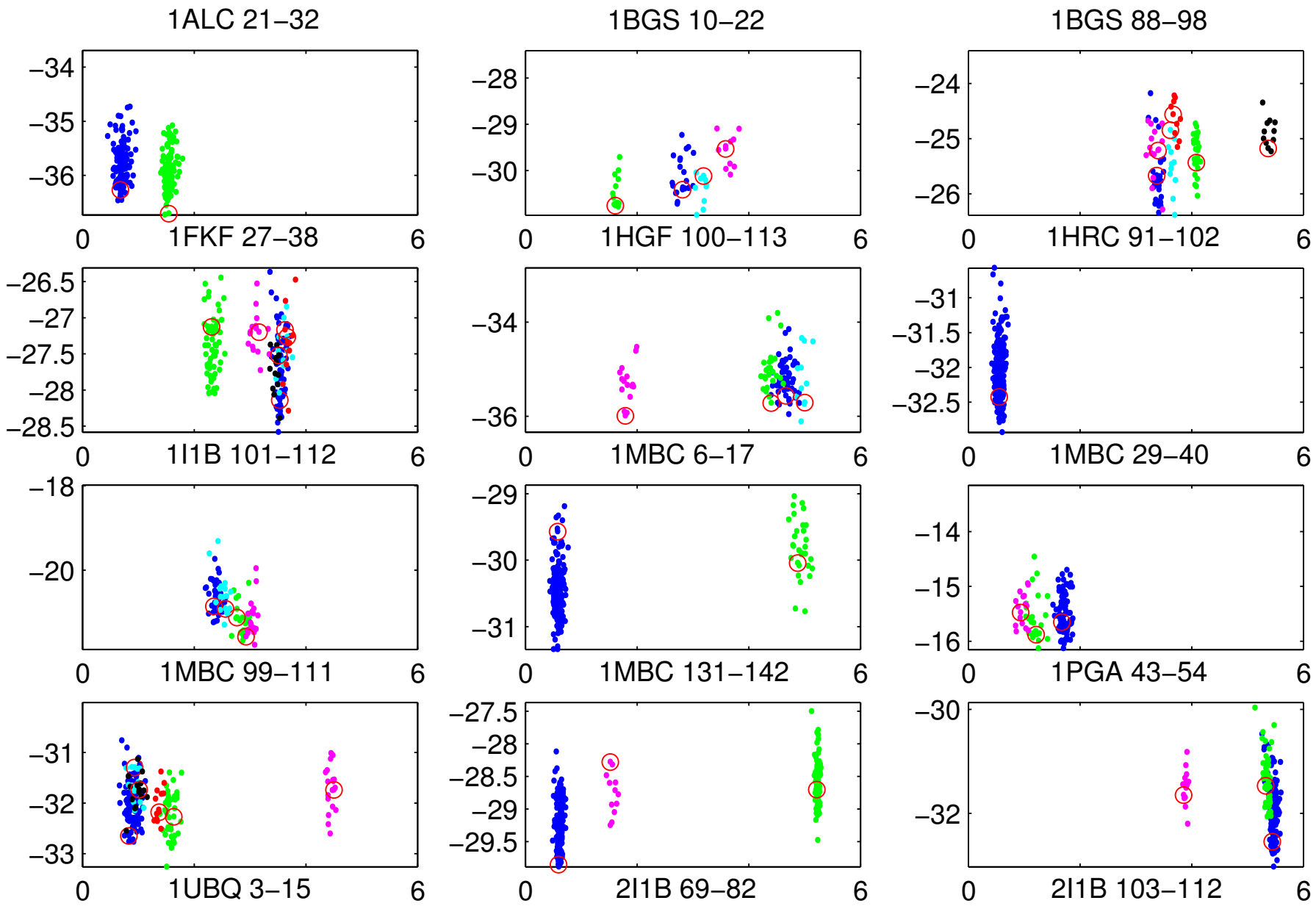
We only tested if the folding of the training set improved. (The data set is too small to test generalization.)

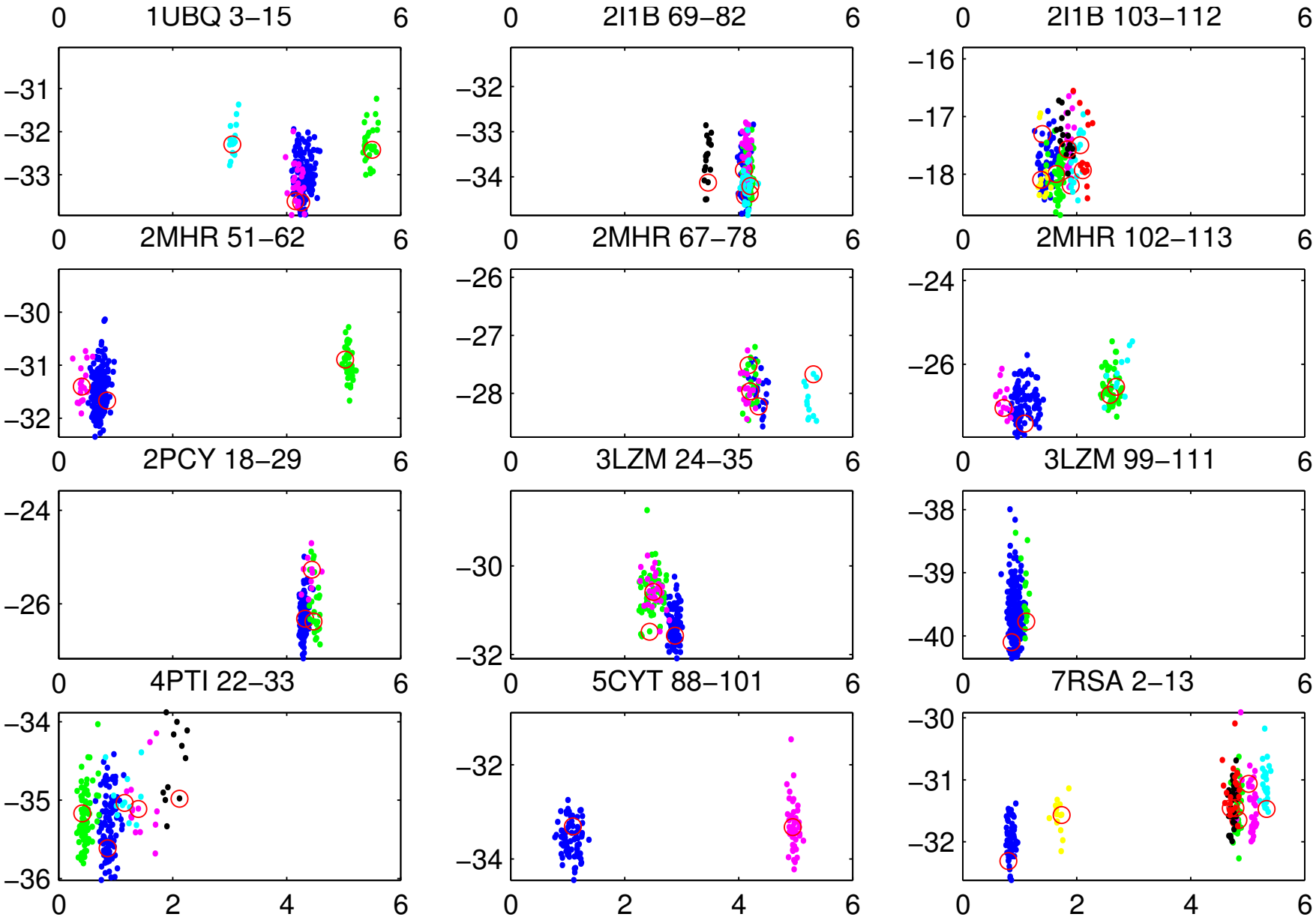
After a number of learning steps (limited by CPU time) conformations where sampled in a long simulation with the final potentials.

Conformations where clustered: the conformation with most neighbors within 0.5 Å (RMSD) defines the first cluster, etc.



ab020


1ALC 21–32
1BGS 10–22
1BGS 88–98
1FKF 27–38
1HGF 100–113
1HRC 91–102
1I1B 101–112
1MBC 6–17
1MBC 29–40
1MBC 99–111
1MBC 131–142
1PGA 43–54
1UBQ 3–15
2I1B 69–82
2I1B 103–112
ab020


1UBQ 3–15
2I1B 69–82
2I1B 103–112
2MHR 51–62
2MHR 67–78
2MHR 102–113
2PCY 18–29
3LZM 24–35
3LZM 99–111
4PTI 22–33
5CYT 88–101
7RSA 2–13

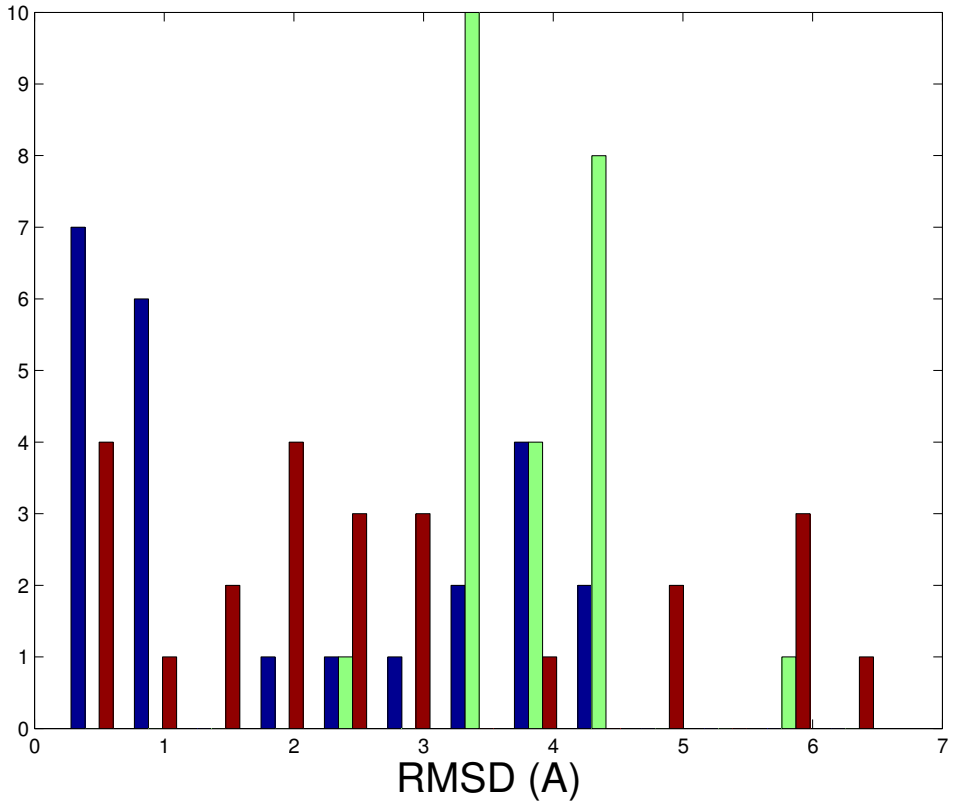
ab020

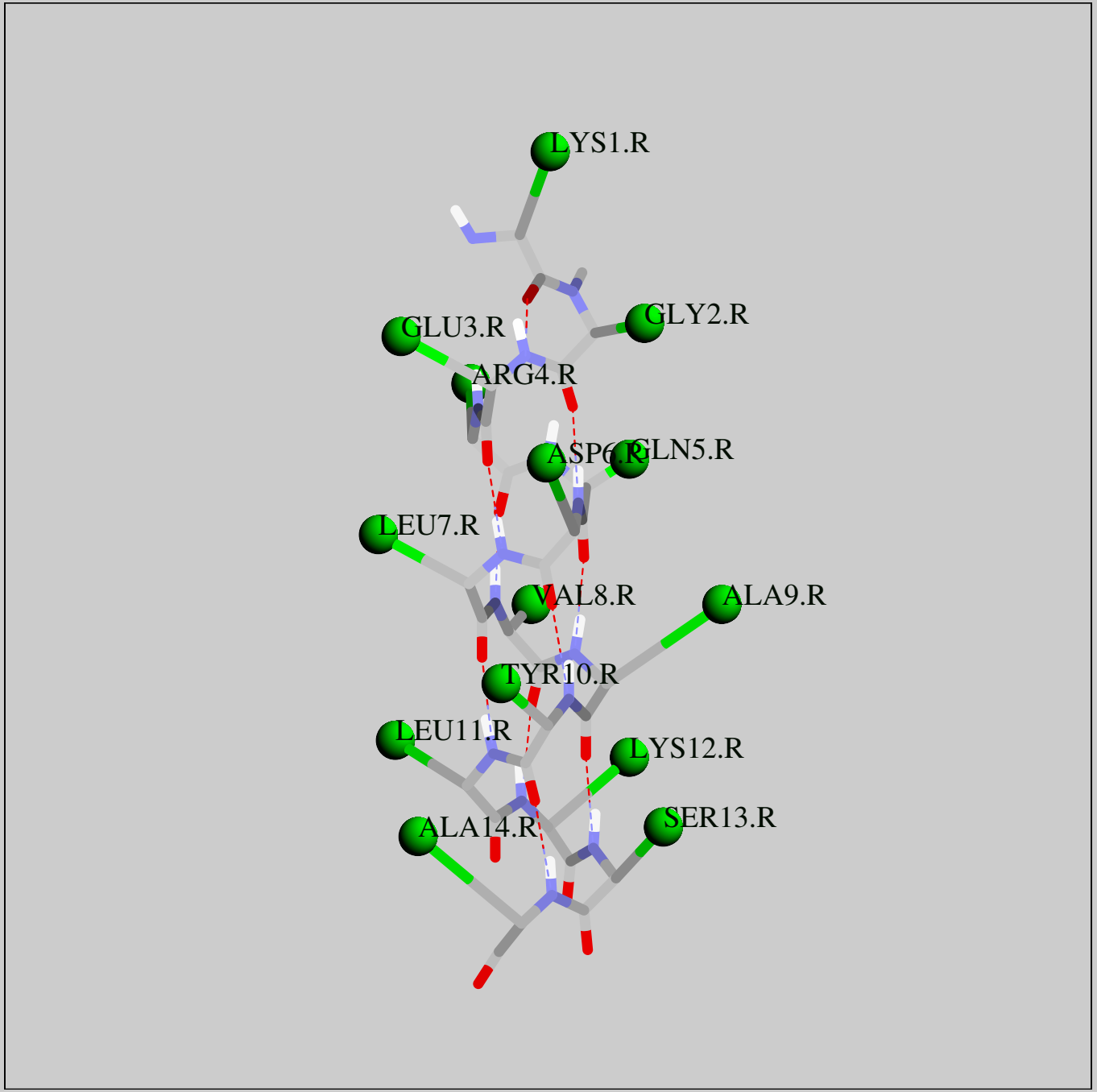
RMSD to the native structure

Predicted structure: the center of the largest cluster

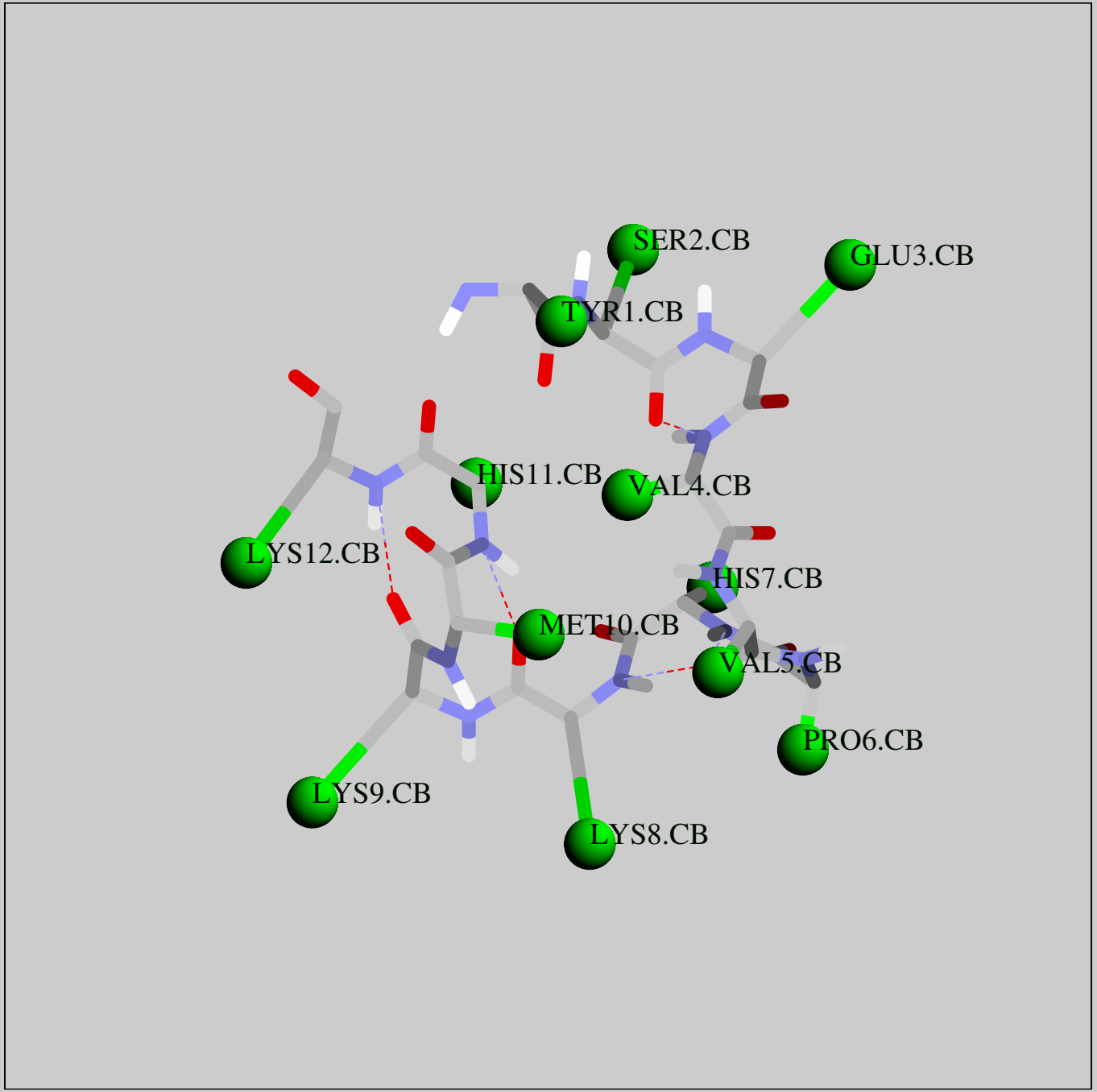
Green: Initial potentials — Blue: After training

Red:Pedersen and Moult, J. Mol. Biol. **269**, 240 (1997).

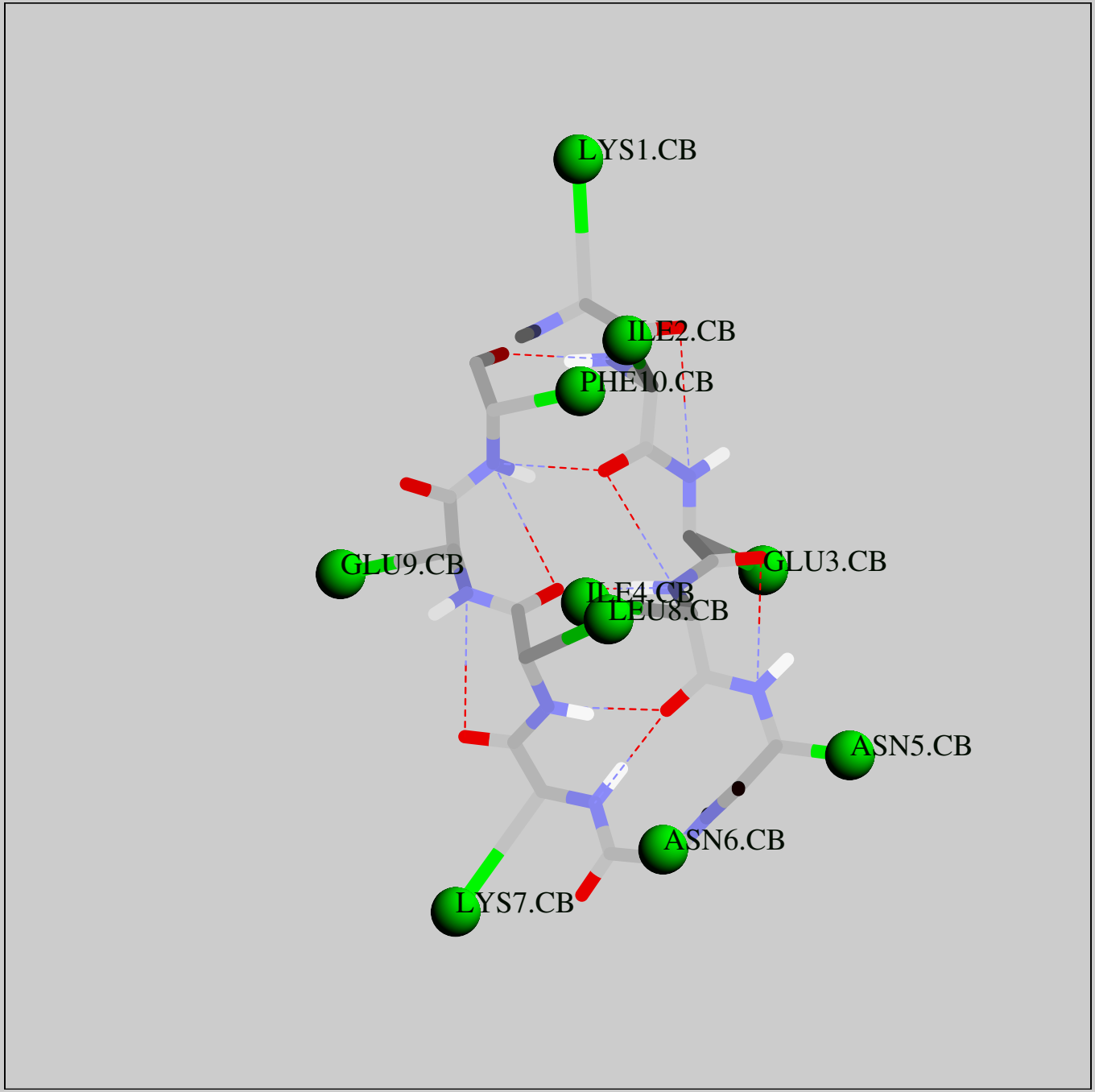



LYS1.R
GLY2.R
GLU3.R
ARG4.R
ASP6.R
GLN5.R
LEU7.R
VAL8.R
ALA9.R
TYR10.R
LEU11.R
LYS12.R
SER13.R
ALA14.R

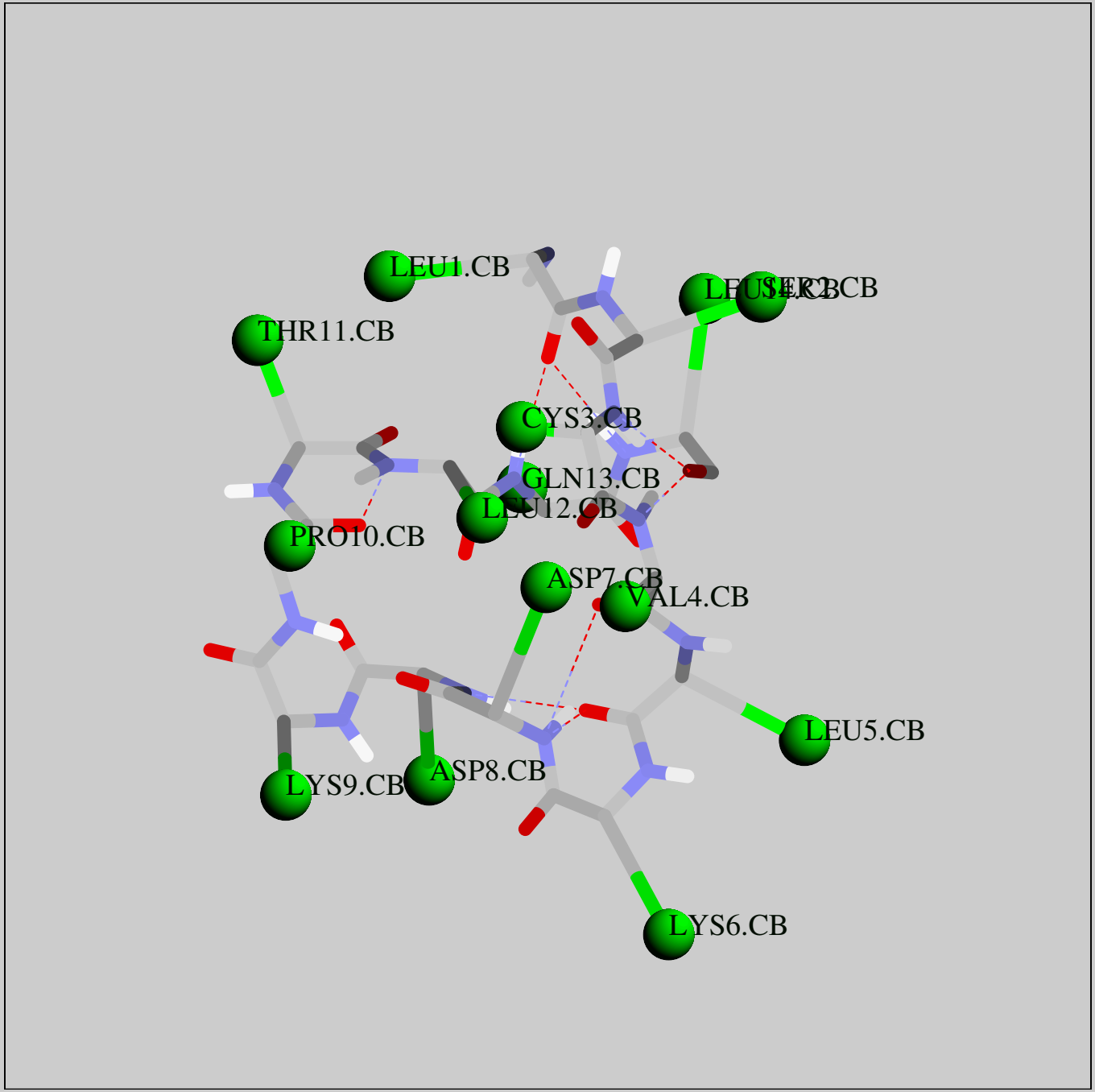
ab022


SER2.CB
GLU3.CB
TYR1.CB
HIS11.CB
VAL4.CB
LYS12.CB
HIS7.CB
MET10.CB
VAL5.CB
PRO6.CB
LYS9.CB
LYS8.CB

ab022


LYS1.CB
ILE2.CB
PHE10.CB
GLU9.CB
GLU3.CB
ILE4.CB
LEU8.CB
ASN5.CB
ASN6.CB
LYS7.CB

ab022


LEU1.CB
THR11.CB
CYS3.CB
GLN13.CB
PRO10.CB
ASP7.CB
VAL4.CB
LEU5.CB
LYS9.CB
ASP8.CB
LYS6.CB

ab022

Conclusions

1. The minimization procedure works
2. The results are as good as an all-atom potential
3. Main problem: Side chains are too simple.

Future

- Scale up to larger data set, real proteins
- Improve model of amino acids
- Test generalization
- Apply same technique on RNA structure