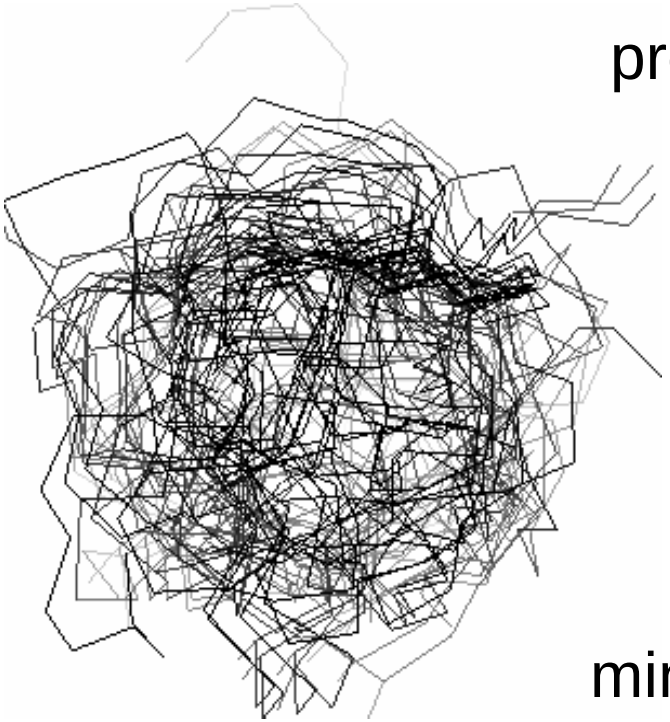
ANALYZE

A Program for Cluster Analysis
and
Characterization of
Conformational Ensembles of
Polypeptides

- Calculations of conformational characteristics, such as hydrogen bonds, turn position and types, RMS deviation from a reference conformation, interchromophore distances, interproton distances, etc.
- Calculation of Boltzmann-averaged properties of the conformational ensemble.
- Calculate the dihedral angles from supplied Cartesian coordinates.
- Cluster analysis of the conformational ensemble by the minima spanning tree or minimum-variance method.
- Fitting the statistical weights of the conformations so as to achieve the best agreement between the calculated average and experimental NOE spectra and coupling constants.

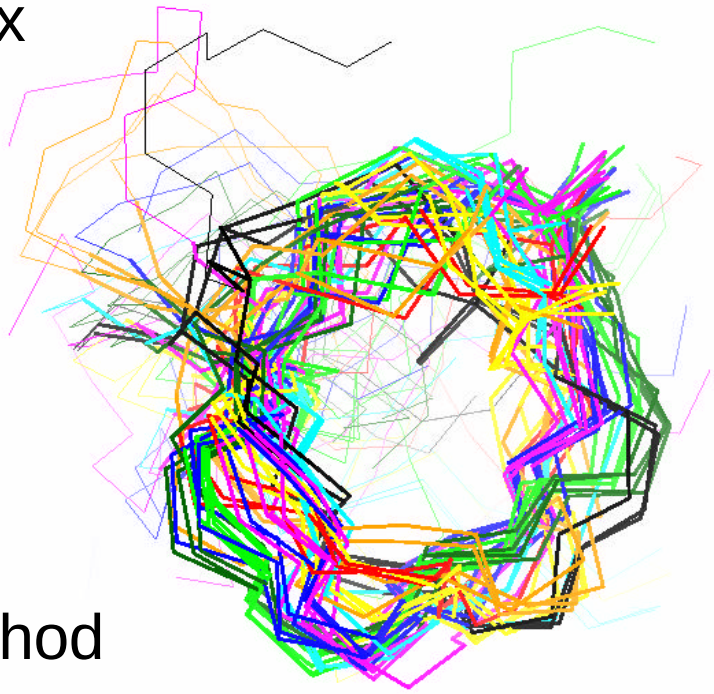
Cluster analysis



data

proximity (RMS) matrix

energy

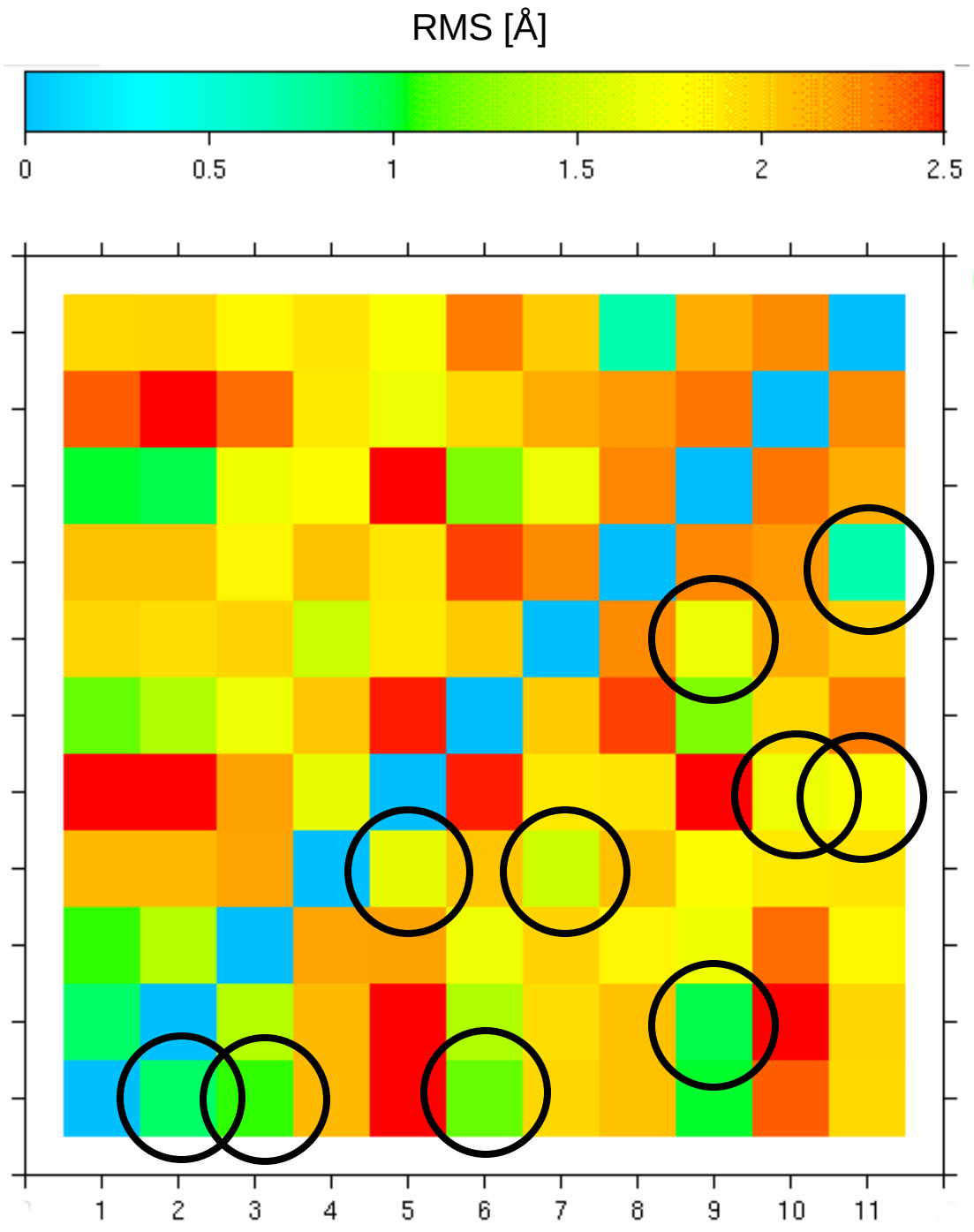


algorithms

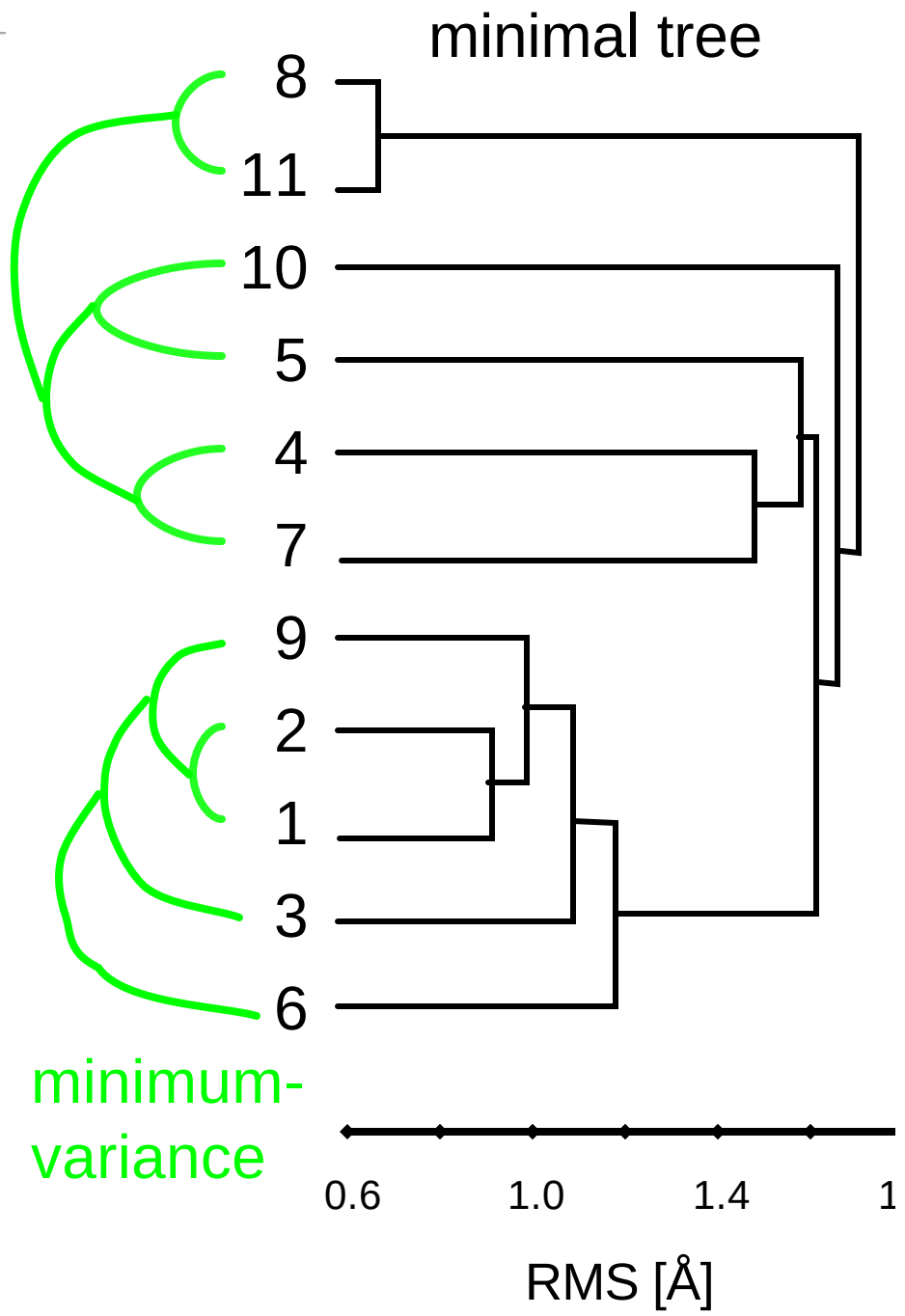
minimum-variance method

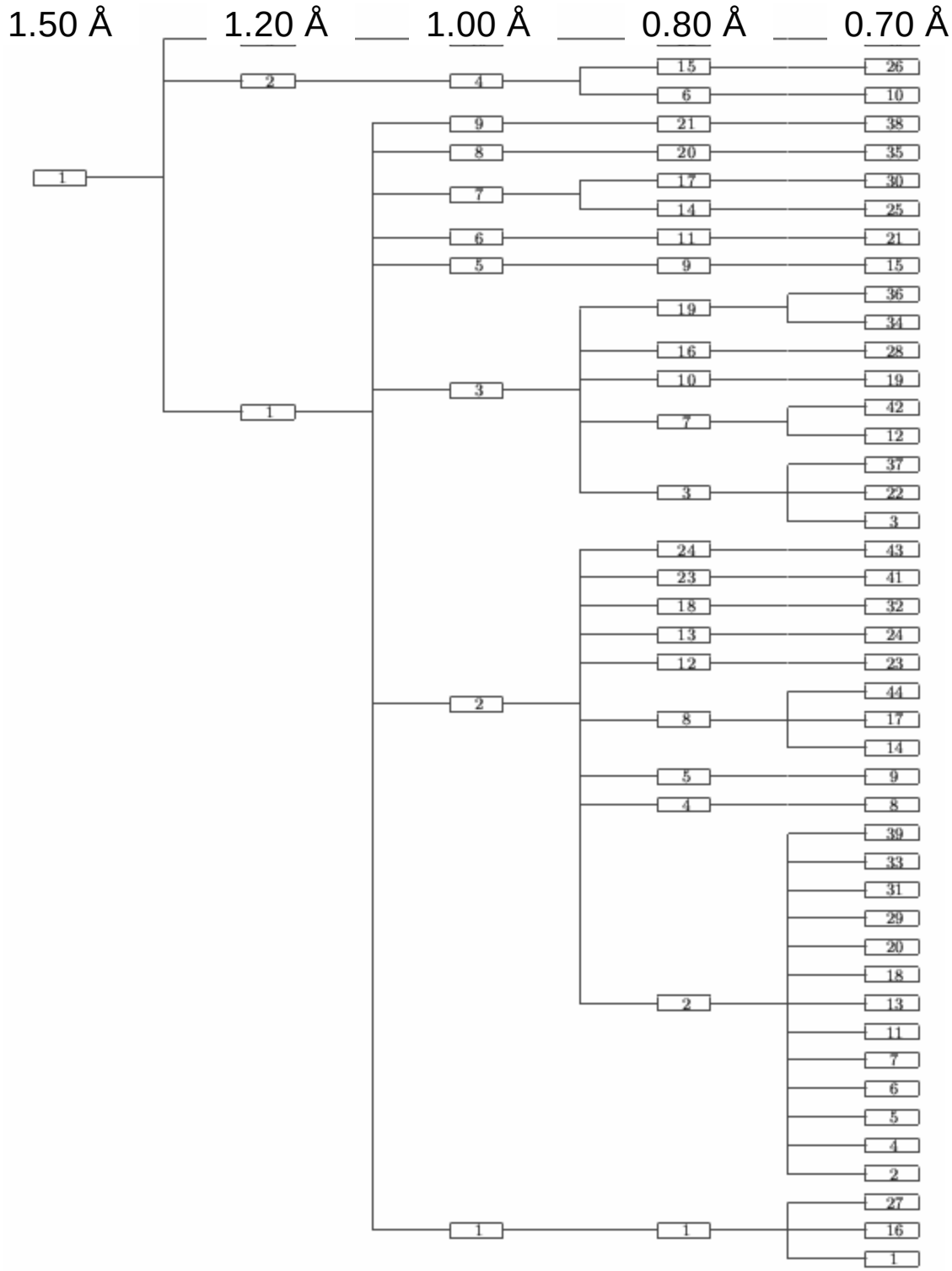
minimum spanning tree

with and without explicit generation of minimal tree



proximity matrix

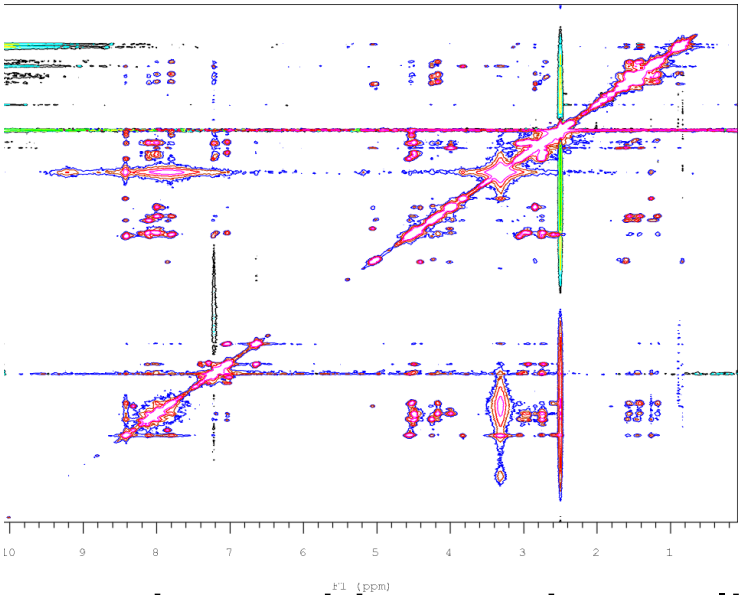



1.50 Å
1.20 Å
1.00 Å
0.80 Å
0.70 Å
1
2
1
4
9
8
7
6
5
3
2
1
15
6
21
20
17
14
11
9
19
16
10
7
3
24
23
18
13
12
8
5
4
2
1
26
10
38
35
30
25
21
15
36
34
28
19
42
12
37
22
3
43
41
32
24
23
44
17
14
9
8
39
33
31
29
20
18
13
11
7
6
5
4
2
27
16
1

Determination of conformational ensemble of flexible polypeptide in solution

low energy conformations

+



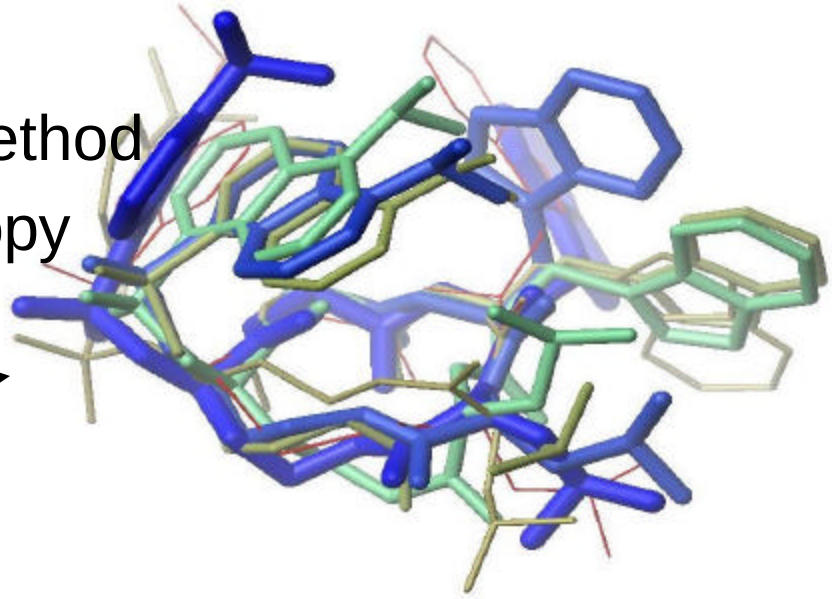
NOE intensities and coupling constants from NMR experiment

algorithms

least-square method

maximum-entropy approach

→



i. simulate the NOE spectra and constant for each conformation

ii. determine statistical weight of conformation in order to obtain the best fit of averaged values t the experimental quantities

The theoretical NOE intensities are **averages** over all conformations of the ensemble:

$$\overline{v_{kl}} = V_\circ \sum_{i=1}^{NC} x_i v_{ikl} \quad k,l = 1,2,\ldots NP$$

$$x_i \geq 0, \quad i = 1,2,\ldots NC$$

$$\sum_{i=1}^{NC} x_i = 1$$

$\overline{v}_{kl}$ – averaged integral intensity of the NOE peak

v_{ikl} – NOE intensity for conformation i

x_i – the statistical weight (fraction)

$V_\circ$ – scaling factor

NP – the number of protons

NC – the number of conformations

The empirical Karplus relationship for NH–C$^{\alpha}$H coupling constants:

$$J_{ik} = a_{\circ k} + a_{1k} \cos\theta_{ik} + a_{2k} \cos^2\theta_{ik}$$

J_{ik} – the coupling constant of kth angle and ith conformation
θ_{ik} – NH–C$^{\alpha}$H angle

And as in the case of NOE intensities, they must be averaged:

$$\overline{J_k} = \sum_{i=1}^{NC} x_i J_{ik}$$

The weights are determined by least square fitting:

$$\min \Phi(V_\circ, x_1, x_2, \ldots, x_{NC}, a_{\circ 1}, a_{11}, \ldots, a_{NJ}) =$$

$$\sum_{(k,l)\in\mathcal{K}} w_{kl}\left[v_{kl}^{exp} - \overline{v_{kl}}(V_\circ, x_1, x_2, \ldots x_{NC})\right]^2$$

$$+w_J \sum_{k=1}^{N\theta}\left[J_k^{exp} - \overline{J_k}(x_1, x_2, \ldots x_{NC})\right]^2$$

$$+\sum_{I=1}^{NJ}\frac{1}{\sigma_{a_{\circ I}}^2}(a_{\circ I} - a_{\circ I}^\circ)^2 + \frac{1}{\sigma_{a_{1I}}^2}(a_{1I} - a_{1I}^\circ)^2 \quad + \frac{1}{\sigma_{a_{2I}}^2}(a_{2I} - a_{2I}^\circ)^2$$

$\mathcal{K}$ – the set of all signals

w_{kl} – the weight of the the NOE peak $w_{kl} = \frac{1}{v_{kl}^{exp}+a}$

w_J – the weight of the coupling-constant term

$N\theta$ – the number of angles with determined J

NJ – the number of the sets of the constants in the Karplus equation

a_{kI}° – the "standard" value of a_{kI} in the Karplus equation

$\sigma_{a_{kI}}$ – estimated standard deviation of a_{kI}°

The maximum-entropy algorithm

$$\Psi(V_\circ, x_1, x_2, \ldots, x_{NC}) = \Phi(V_\circ, x_1, x_2, \ldots, x_{NC}) + \alpha \sum_{i=1}^{NC} x_i \log x_i$$

the "entropy" term

$$-\alpha \sum_{i=1}^{NC} x_i \log x_i$$

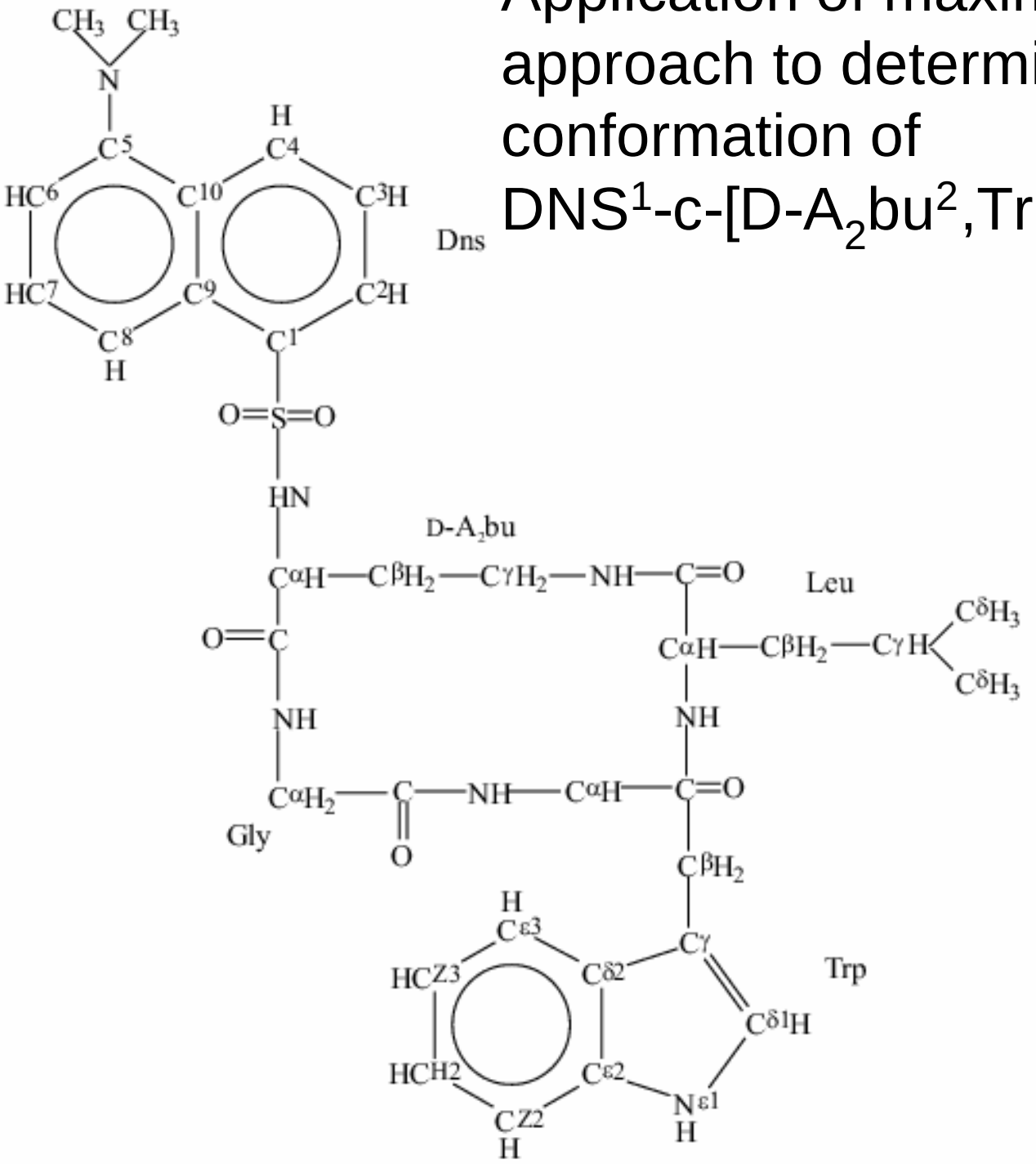
is subtracted from the minimized sum of squares.

The maximum-entropy approach prevents overfitting.

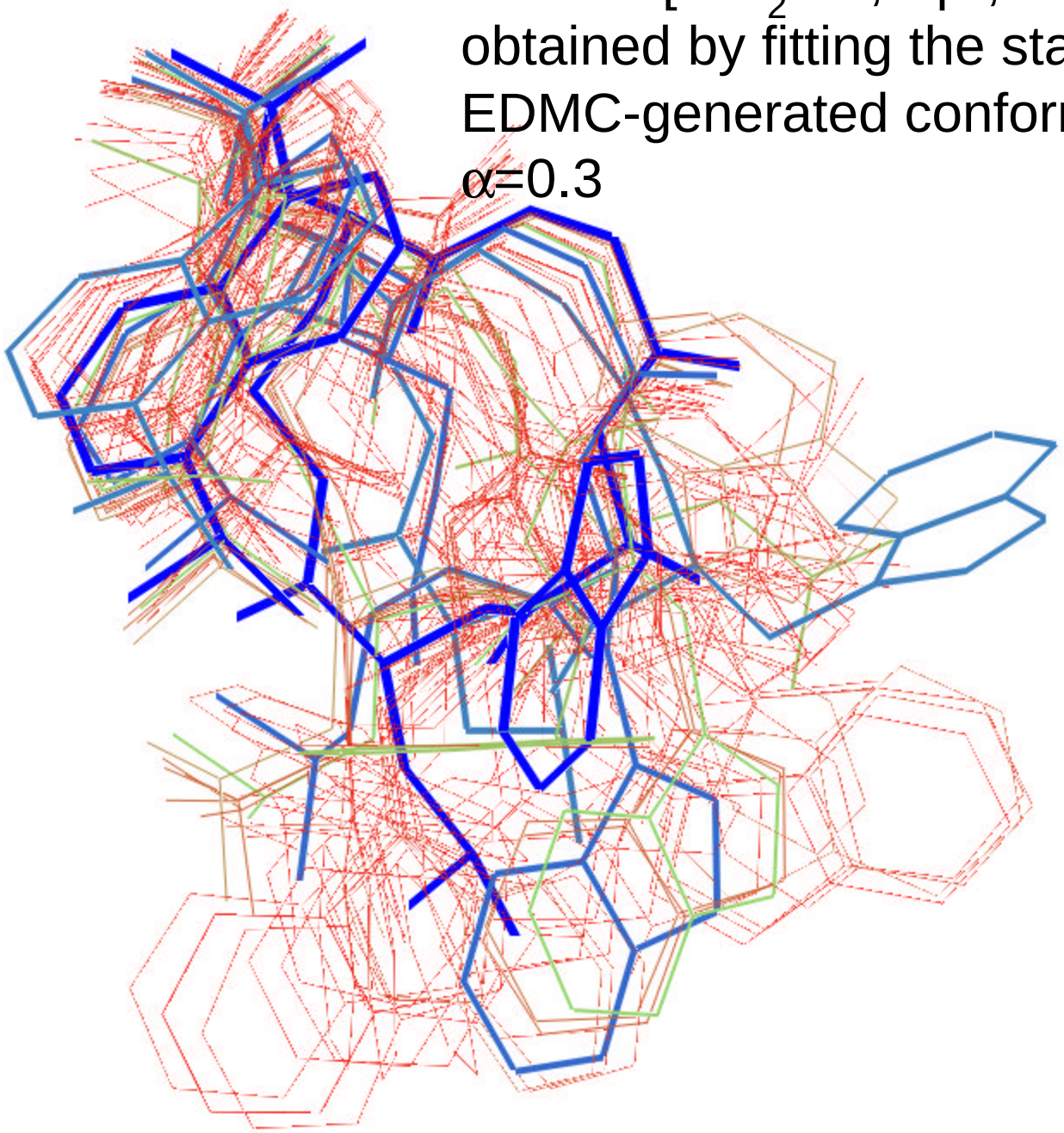
The entropy term reaches its global minimum, if the statistical weights of all conformations are equal. Weight differentiating comes only from the Φ term that includes experimental information.

Therefore a common procedure is to choose the coefficient at the entropy term, α, so that the weighted χ^2 value be equal to the number of observations, which is equivalent to the requirement that the mean errors in the fitted quantities be comparable with the error estimates.

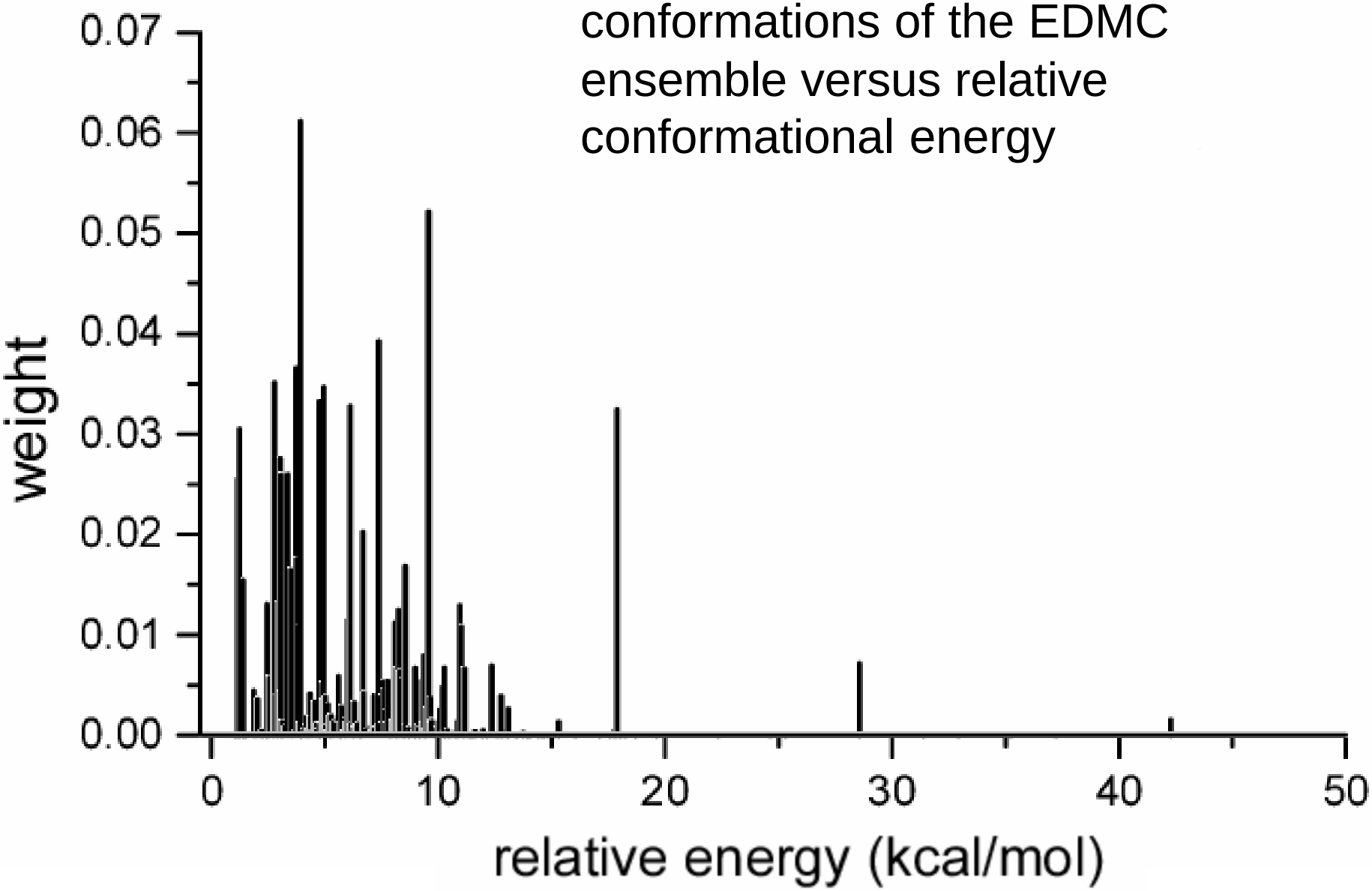
Application of maximum entropy approach to determination of solution conformation of DNS^{1}-c-[D-A_2bu^{2},Trp^{4},Leu^{5}]-enkephalin



The conformational ensemble of DNS^{1}-c-[D-A$_2$bu^{2},Trp^{4},Leu^{5}]-enkephalin obtained by fitting the statistical weights c EDMC-generated conformations with α=0.3



Impulse plot of statistical weights of conformations of the EDMC ensemble versus relative conformational energy



Determination of conformational equilibrium of peptides in solution by NMR spectroscopy and theoretical conformational analysis: application to the calibration of mean-field solvation models.

- If an adequate force field is used, conformations with large statistical weights obtained from the weight-fitting procedure should also have low energies
- The solvation parameters of simple mean-field models can be adjusted to achieve this
- Force field calibration based on such a procedure is particularly attractive regarding the parameterization of the solvation energy in non-aqueous solvents e.g.,dimethyl sulfoxide, for which thermodynamic solvation data are scarce.

Atom[a] type	Atom name	σ		
		water	DMSO	
			iteration 1	iteration 2
1	H (hydroxyl and amine)	0.050	-0.066	-0.094
2	H (amide)	-0.008	-0.040	-0.027
5	$CH_3(sp^3)$	0.001	-0.012	-0.017
6	$CH_2(sp^3)$	-0.002	-0.030	-0.044
7	CH (sp^3)	0.032	-0.083	-0.139
8	$CH_2(sp^3,5)$[c]	0.000	-0.034	0.024
9	CH (sp^3,5)[c]	-0.031	-0.013	-0.128
10	CH (ar)[d]	-0.005	-0.011	-0.014
11	C (ar)[d]	-0.099	0.808	0.358
13	C (>C-OH)[e]	0.022	0.156	0.023
14	C (C=O)	0.162	0.243	0.106
15	N (sp^3)	-0.105	-0.031	-0.042
22	N (amide)	-0.142	-0.649	-0.871
23	O (sp^3)	-0.125	-0.041	-0.064
26	O (C=O)	-0.138	-0.018	-0.037

[a]Atom solvation types are same as in SRFOPT.
[b]Original SFROPT hydration parameters.
[c]Five-membered proline ring.
[d]Aromatic carbon or CH group.
[e]Tyrosine ring carbon atom connected to the hydroxyl.

The package includes two versions of ANALYZE: one for fitting NMR spectra and one for cluster analysis and other purposes. The source files are contained in source_clust and source_nmr, respectively. This division is caused by practical reasons: the NMR and clustering parts are very memory consuming, which practically excludes their incorporation into one program.

The main input file for the ANALYZE program is organized in the same way as the input file for the ECEPP program.

Instructions are collected into "Data Groups" identified by a opening keyword which contains symbol '$' as the first character, i.e. $CLUSTER, $NOES, and closed by keyword $END.

Cluster analysis of the conformational ensemble

A. minimal-tree clustering

0.- go to directory cluster/MINTREE

1.- Prepare input file with suffix "inp" (e.g. tree.inp) with instructions for ANALYZE, include the following data groups (look into User Manual for detailed description of each group): $title $cntrl $seq $bridge $cluster $supat

2.- in $cntrl group use the following keywords:

$cntrl
runtyp=cluster nrclus1=1 nrclus2=6 res_code=one_letter
verbose print_pdb=1 tree
$end

short description of keywords

runtyp=cluster	defines the type of the run
tree	defines the algorithm for clustering
nrclus1=1 nrclus2=6	residues 1-6 will be superposed
res_code=one_letter	$seq will be defined by a one-letter code
print_pdb=1	write coordinates of only leading

3.- for this example data from EDMC run of oxytocin peptide will be used so the $seq should look like (see table in ECEPP manual for one-letter code residue names)

```
$seq
  HC_YIQNC_PLGN
$end
```

and the definition of disulfide bond is necessary in data group $bridge

```
$bridge
2 7
$end
```

4.- finally in data groups $cluster and $supat we should include RMS cut-offs for clustering and atoms to be superposed

```
$cluster
5 1.5 1.2 1.0 0.8 -0.7
5.0
$end
```

the -0.7 means that dihedral angles and Cartesian coordinates will

$supat
4
CA C N SG
$end

atoms CA,C,N,SG from residues 1-6 (nrclus1=1 nrclus2=6 in $cntrl) will be
used for superposing structure

5.- Save the file and run ANALYZE

In the command line type

analyze-clust tree otv16cl otv_tree

tree.inp is just prepared input file

outo.otv16cl contains the results of EDMC run for oxytocin used as input in this example

6.- Analyze output files

otv_tree.out contains output list file

otv_tree.ang contains the dihedral angles of the leading families

otv16cl.tex contains LaTeX picture output of the graph representing the minimal spanning tree, to prepare postscript file tree.ps use tree.tex template:

use PCTex32 to load tree.tex, and click *typeset* to see the picture

files otv16cl0001.pdb - otv16cl0009.pdb contain Cartesian coordinates of leading members of each family within energy within defined 5.0 kcal cutoff, you can load them to molmol program using script molmol.csh

to run script with Cygwin tcsh shell type in command prompt

tcsh molmol.csh

3. clustering with hydrogen bond and turn analysis without explicit generation of minimal-tree

).- go to directory cluster/NOTREE

..- Copy tree.inp from previous example to current directory with name notree.inp

2.- Make the following changes

remove keyword 'tree' and add keywords 'beta_turns' and 'h_bonds' in $cntrl data group

3.- Save the file and run ANALYZE

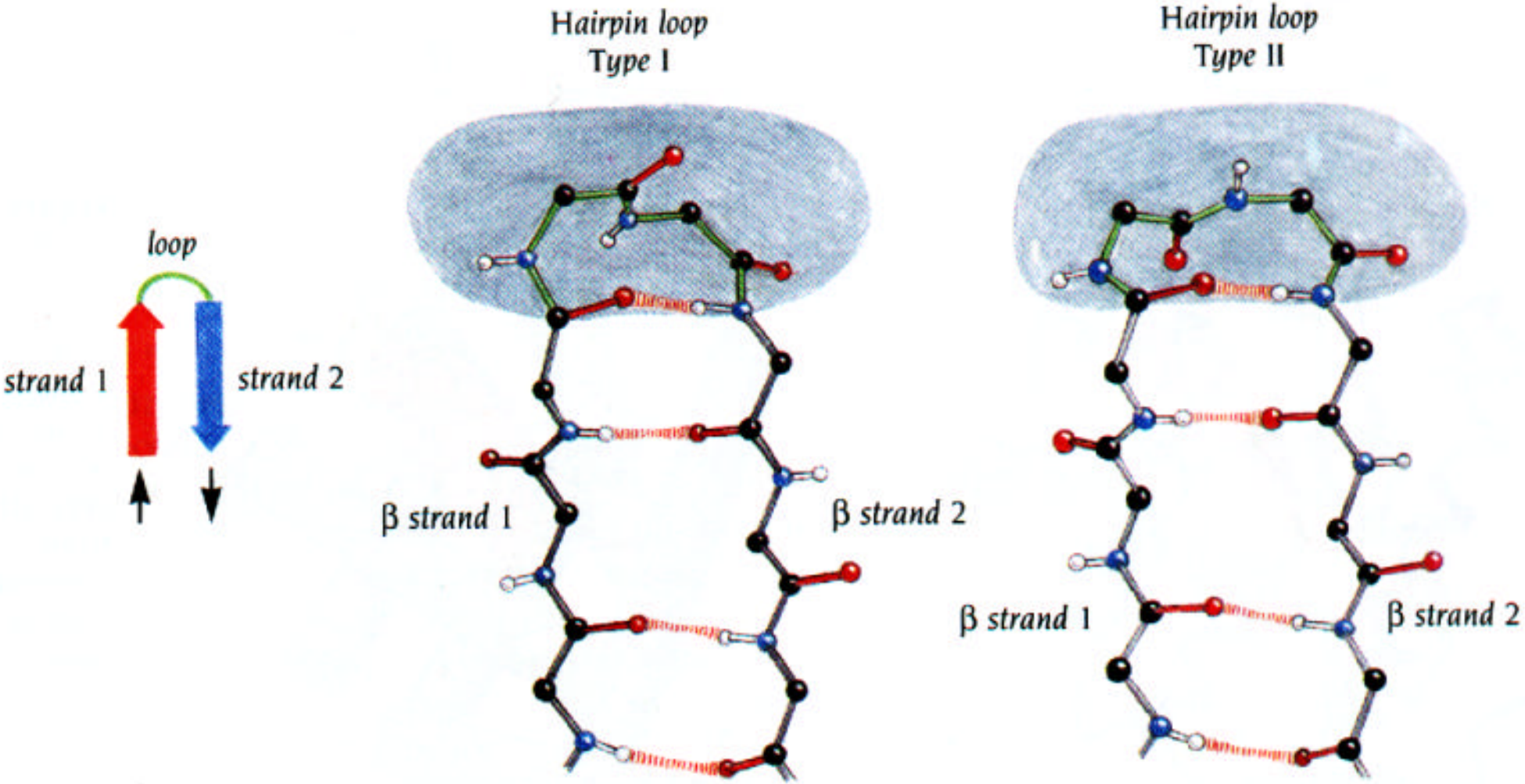
n the command line type

analyze-clust notree otv16cl otv_notree

4.- Analyze output files

look at additional information in otv_tree.out file

For real situation, clustering of several thousands of conformations, algorithm working without explicit construction of minimal-tree is much faster. The obvious disadvantage is that minimal-tree, which gives some idea about how to best partition the set of the conformations is not printed.



Lewis, P.N., Momany, F.A. and Scheraga, H.A., Chain reversals in proteins, *Biochim. Biophys. Acta*, **303**, 211-229 (1973)

C. clustering using minimum-variance method

0.- go to directory cluster/MINVAR

1.- Copy tree.inp from first example to current directory with name minvar.inp

2.- Make the following changes

replace keyword 'tree' with keywords 'min_var' , increase print_pdb=10 (so coordinates of up to 10 structures from each family will be written) and change the $cluster group into

```
$cluster
6 -2.0 1.5 1.2 1.0 0.8 0.7
5.0
$end
```

.- Save the file and run ANALYZE

n the command line type

analyze-clust minvar otv16cl otv_minvar

.- Analyze output files

heck in otv_tree.out how the results of cluster analysis using
inimum-variance method differs from minimum-tree

db files contain Cartesian coordinates of up to 10 members of each
amily
ith energy within the 5.0 kcal cutoff, you can load them to molmol
rogram using script molmol.csh

to run script with Cygwin tcsh shell type in command prompt

tcsh molmol.csh

Fitting the statistical weights of the conformations so as to achieve the best agreement between the calculated average and experimental NOE spectra and coupling constants.

A. Pure least-squares fitting with no entropy term.

0.- go to directory morass/least_squares

1.- Prepare input file least_square.inp with instructions for ANALYZE, include the following data groups (look into User Manual for detailed description of each group): $title $cntrl $seq $bridge $noes $morass $marquardt $coupling

2.- in $cntrl group use the following keywords:

$cntrl
untyp=morass
es_code=three_letter
$end

3.- for this example data from EDMC run of DNS[1]-c-[D-A_2bu[2],Trp[4],D-Leu[5]] enkephalin will be used so include following $seq $bridge data groups

```
$seq
DAN
lab Gly Trp lep
CXX
end
$bridges
: 5
end
```

4.- data groups $noes and $marquardt controls the fitting procedure

```
$noes
mode=fitting conf=all bystrov=yes antinoe=long
geminal=no vicinal=yes rigid=no
wbase=0.01 wei_coupl=0.1
alpha_ent=0.0
end
$marquardt
olf=0.00001
```

5.- data group $morass includes parameters necessary for simulation of NOE
spectra by MORASS program

$morass
auc=0.45 time=0.300 vol0=100 sfrq=500 cutt=6.0
$end

6.- finally in data group $coupling there are experimental coupling constants

$coupling
1 0
2 2 1 5 # dab 1 phi, chi 2
1 9.28 -19.5 2 11.23 0 #J,phase angle,J,phase angle
3 1 1 # Gly 2 phi
2 12.21 0
4 1 1 # Trp 3 phi
1 7.81 60
5 1 1 # leu 4 phi
1 9.77 -60

'.- Save the file and run ANALYZE

n the command line type

.nalyze-morass least_square dansylD_clust least_square dansylD

east_square.inp is the prepared input file,

uto.dansylD_clust contains the results of cluster analysis of EDMC run or DNS^{1}-c-[D-A$_{2}$bu^{2},Trp^{4},D-Leu^{5}] enkephalin used as input in this example,

lansylD.noe contains experimental NOE intensities.

3.- Analyze output files

east_square.out contains output list file

n subdirectory ..\PDB there are files with Cartesian coordinates for all conformations used in fitting, they were produced together with uto.dansylD_clust by cluster analysis. You can visualize their weights with the molmol program using molmol_nmr.csh script with the number of conformation to be shown as argument. Only 6 conformation have weight > 0.0 so type

3. Maximum-entropy fitting example

.- go to directory morass/FITTING/maxent

.- Copy least_square.inp from previous example to current directory
/ith name max_entropy.inp

.- Set parameter alpha_ent to 0.3

.- Save the file and run ANALYZE

ı the command line type

analyze-morass max_entropy dansylD_clust max_entropy dansylD

.- Analyze output files, compare weights with results of previous
xample

he entropy term forces the weights to be equal to each other, while the
sum of error" term Φ picks up the conformations that best fit to the
xperimental observables; the latter usually results in the selection of
nly a few out of several hundred, which is regarded rather strange by
ıe authors of the program. Just a little admixture of the "disorder" term

You can visualize their weights in molmol program using molmol_nmr.csh
script with the number of conformation to be shown as argument. For instance
weights of 10 conformation sum up to 0.75

```
tcsh molmol_nmr.csh 10
```