NUCS: Global scaling protocol

Sonja M. Schwarzl and Stefan Fischer

June 17-2008, version 1.2

As described in: J. Comput. Chem. 26, 1359-1371 (2005). *Non-Uniform Charge Scaling (NUCS): a practical approximation of solvent electrostatic screening in proteins.* S.M. Schwarzl, D. Huang, J.C. Smith and S. Fischer.

The present CHARMM scripts require CHARMM version 28 or higher (as the scripts make use of the PBEQ module).

1 Theoretical summary

The protein is split into groups, for which scaling factors are to be determined. Each side chain is a separate group. Back bone groups are defined over two residues in order not to split the amide bond dipole. Thus, the carbonyl carbon and oxygen of a given residue and the C_{alpha}, nitrogen and hydrogen of the next residue form a back bone group. An exception is proline, where the side chain atoms are also included in the back bone group because of the rigidity of the ring system.

First, a group I is selected, for which the scaling factor is to be determined. The charges on this group are set to their respective values. The charges of all other groups are set to zero. Then the electrostatic potential ϕ_I due to the charges of group I is calculated using a finite difference scheme with a focusing approach. It is then interpolated to all atom positions $\phi_I(\vec{r}_j)$.

Next, the electrostatic interaction energies E_{IJ}^{vac} and E_{IJ}^{solv} between the source group I and all other groups J are calculated. From this the pairwise dielectric constants ϵ_{IJ} are determined. If $\epsilon_{IJ} < 0$, the energies are zeroed, otherwise they are put to absolute values. Finally, the dielectric constant ϵ_I and the scaling factor λ_I are determined.

2 Step 1: Getting the box dimensions

The CHARMM input file *get_box.inp* can be used to derive the correct box dimensions. It streams *generate.str* and *settings1/2.str*. In *settings2.str* (see next section) the PBEQ settings do not need to be correctly set yet.

In *generate.str*, the user must set up the system (i.e. read in topology and parameter files, sequence information, generate the protein structure file with all segments) and read in the coordinates for which the NUCS is to be performed (NUCS factors are conformation dependent).

The *get_box.inp* script orients the coordinate (with COOR ORIENT) and writes them to a new coordinate file, whose name must be given in *get_box.inp*. ATTENTION: !!! Make sure that you use these oriented coordinates for the whole subsequent NUCS procedure (by changing the name of the coordinate-file in *generate.str*) !!!

The script also performs a "COOR STAT", which gives the dimensions of the system. It gives the minimum and maximum X-, Y-, and Z-coordinates of the centers of the atoms. This will be needed in the next section.

3 Step 2: Computing the raw scaling-factors λ_I'

The user has to set the variables in *settings1/2.str*, see section **??**. Use the info resulting from the previous step to edit the PBEQ sectionn in *settings2.str* as follows:

First, to the COOR STAT maximum distances, add 6 Å (i.e., 1.5 Å for the atom radius plus 1.5 Å for the water radius on each side).

Then, for the first level of PBEQ focusing, add another 25 Å on each side of the molecule. This gives the size of the box. Use a grid spacing of 4.0 Å to obtain the number of grid points with $boxsize/gridspacing = (Xmax - Xmin + 6 + 2 \cdot 25)/4.0$. In order not to get any box size problems, the values should always be rounded up. CHARMM only takes odd numbers, if an even number of grid points is specified, it is automatically incremented by one.

For the second level of PBEQ focusing, add 10 Å (instead of 25 Å) on each side and use a grid spacing of 2.0 Å .

For the third level of PBEQ focusing, add only 5 Å and use a grid spacing of 1.0 Å.

generate.str should remain unchanged throughout the subsequent NUCS procedure.

3.1 User defined settings

The user must also specify the following variables in *settings.str*:

- General settings. The variable $FIRSTRESIDUE$ should be set to the first residue for which the scaling factor should be determined (if the residue $FIRSTRESIDUE - 1$ has a side chain, this scaling factor is also determined). $LASTRESIDUE$ is the last residue, for which λ is determined ($BACK1$ of $LASTRESIDUE + 1$ is also included).

Specifying $FIRSTRESIDUE$ and $LASTRESIDUE$ allows to plit the NUCS procedure for a large protein into several jobs, each getting the scaling factors for a sub-set of residues. These can then be re-assembled into a single file with *reassemble_large_protein.inp*

The names of the following output files must be given:

1. Scaling factors. The final scaling factors are written into the WMAIN column of a .crd coordinate file. The name is specified in $OUTPUT$
2. Interaction energies E_I^{solv}. The reference interaction energies between each group and the rest of the protein calculated from the Poisson-Boltzmann calculation are written into the WMAIN column of a .crd coordinate file. The name is specified in $OUTPUT1$
3. Paiwise interaction energies. The pairwise interaction energies E_{IJ}^{vac} and $E_I^{solv}J$ that are evaluated during the calculation are printed into an output file specified in $OUTPUT2$.

- Energy function settings. The parameter file version has to be given in $PARAM$(19 extended carbon model or 22 for an explicit hydrogen representation), the cutoff type in $ELECTR$ (cdiel for a constant dielectric permeability and rdiel for a distance-dependent dielectric permeability), and cutoff values in the variables $UCTONNB, UCTOFNB, UCUTNB, UWMIN$ (U for user defined).

- PBEQ settings. These settings are needed for the calculation of E_I^{solv} with the Poisson-Boltzmann module in CHARMM (PBEQ). $offset$ is a value added to the van der Waals radii used for the determination of the dielectric boundary. sw gives the value of the smoothing window to be used for smoothing the transition of the dielectric constant from $EPSWAT$ to $EPSPROT$ at the dielectric boundary between high and low dielectricum. A combination of $sw = 1.5$ and $offset = 0.7$ has been found to reproduce accurate electrostatic calculations done with UHBD. The ion concentration in the solvent is given in CON and the temperature in TEM. The grid spacings and number of grid points in all direction are given in $DCELn$ and $NCLdn$, where $n = 1, 2, 3$ and $d = X, Y, Z$. To determine these values see Section 2.

- USERATOMS. Each non-standard residue (i.e., other than protein residues or waters), for example ATP, Mg, Retinal, etc., must have ONE (and only one) of its atoms (it does not matter which one) included in the selection called UATOMS.

4 Running the main script

The script *scaling.inp* is piped into CHARMM. The print and warning levels are set so as to suppress most of the output. Otherwise, the output files will be very large.

The structure of this script is shown in Figure 1, and it is described in more detail in the last sections.

5 Counterscaling of scaling factors

The scaling factors as determined by the procedure described above are only raw factors. They will be denoted with λ'_I in this section. Having calculated the λ'_I it is possible to calculate the electrostatic interaction energy between a group I and the rest of the protein in vacuum $E_I^{shield'}$ using the raw scaling factors. The $E_I^{shield'}$ are an estimate for the target E_I^{solv}.

$$E_I^{shield'} = \sum_{i \in I} \sum_{j \in J \neq I}^{groups} \frac{q_i}{\lambda'_I} \frac{q_j}{\lambda'_J} \frac{1}{r_{ij}} \tag{1}$$

The λ'_I may be counterscaled to give the scaling factors λ_I. The E_I^{NUCS} calculated with the λ_I give a more accurate correlation with the E_I^{solv}. Counterscaling is achieved by dividing the raw scaling factors λ'_I by a factor γ,

$$E_I^{NUCS} = \sum_{i \in I} \sum_{j \in J \neq I}^{groups} \frac{q_i}{\lambda_I} \frac{q_j}{\lambda_J} \frac{1}{r_{ij}} = \gamma^2 \sum_{i \in I} \sum_{j \in J \neq I}^{groups} \frac{q_i}{\lambda'_I} \frac{q_j}{\lambda'_J} \frac{1}{r_{ij}} = \gamma^2 E^{shield'} \tag{2}$$

The E_I^{NUCS} are a better estimate of E_I^{solv}, so γ^2 can be determined by doing a least square fit of the E_I^{solv} against the $E_I^{shield'}$ according to Equation 2. For better accuracy, different γ coefficients are determined for backbone (γ_{BACK}) and side-chain (γ_{SIDE}) groups.

5.1 Calculation of the counterscaling coefficients

To get the data necessary for the least squares fit, the E_I^{solv} are taken from the previous output file (which had been specified as $OUTPUT1$) and the $E_I^{shield'}$ are recalculated. Both are then written out in columnar fashion. The CHARMM script that does this is :

./scripts/global/counterscale/verify_factors.inp

The structure of this script is detailed at the end of this document and shown in Figure 2.

./scripts/global/counterscale/verify_factors.inp streams *generate.str* and *settings2.str*, which should be exactly the same as used before to calculate the raw scaling factors.

The user must set variables *settings1.str*. Warning: The settings here are slightly different from the previous *settings1.str* file:

The two output coordinate files (from the previous section), which contain 1) the raw scaling factors λ'_I in the WMAIN column and 2) the interaction energies in solution (E_I^{solv}) need to be specified.

Two output files are generated, in which the calculated values for backbone groups and side chain groups are tabulated. These values must then be used in another program to determine γ_{BACK} and γ_{SIDE} by performing a linear regression fit.

This can for instance be done with the programs provided in the directory *./tools/* . For example:

./select_columns -s1 "5,6" < back_cdiel.dat | ./linear-regression

The square root of the slope of the resulting linear regression through the origin then gives the counterscaling coefficient γ_{BACK} (and γ_{SIDE} respectively).

5.2 Getting the final (counterscaled) scaling factors

Once γ_{BACK} and γ_{SIDE} have been determined, setting BACKFAC and SIDEFAC (in *settings1.str*) to γ_{BACK} and γ_{SIDE} respectively, and running *verify_factors.inp* again will yield the final (counterscaled) scaling factors. They are written to the file specified by the variable OUTPUT in *settings1.str*.

This also yields the final values for E_I^{NUCS} for the bacbone or the sice-chain groups, so that the root-mean-square error of E_I^{NUCS} with respect to the ideal PB interaction energies can then be determined. For example :

./select_columns -s1 "5,6" < back_cdiel.dat | ./rmsd

5.3 Other settings for *verify_factors.inp*

The user must specify the following variables in *settings1.str*

- Scaling factors and energies. The variable $INPUTFAC$ must be set to the path and file name of the .crd coordinate file with scaling factors in the WMAIN column. The variable $INPUTENER$ must be set to the path and file name of the coordinate file with the PB interaction energies E_I^{solv} in the WMAIN column.

- Charge scaling settings. The variables $BACKFAC$ and $SIDEFAC$ should be set to 1.0, if E_I^{vac}, E_I^{solv}, and $E_I^{shield'}$ are to be calculated.

  They should be set to γ_{BACK} and γ_{SIDE}, respectively, if the E_I^{NUCS} are to be calculated.

  The ouput file names must be given in the variables $OUTPUTBACK$ and $OUTPUTSIDE$, respectively.

6 Details about the script *scaling.inp*

The scripts rely on an atom-selection called GROUPATOMS, in which each group must have ONE (and only one) of its atoms included. For standard groups such as protein side chains, protein back bones, or water this is automatically done. For non-standard groups, the user must add (in *settings2.str*) this definition in the selection called UATOMS.

In *settings1/2.str*, the user must specify several variables needed for the calculation. In *initialize.str* several variables are initialized (see below). After an energy call to initialize the energy function, *loop.str* is streamed that loops from the first to the last residue, for which scaling factors shall be determined. After definition of the source group I, $calcfactors.str$ is streamed. There, *pbeq.str* is streamed to determine the electrostatic potential $\phi_I(\vec{r}_j)$. The scaling factor λ_I is then calculated in *scaling.str*, which streams *loop2.str* and determines the actual λ_I. *loop2.str* loops over all residues within the cutoff distance around group I and streams *calculation.str*. In this stream file, the pairwise interaction energies in vacuum E_{IJ}^{vac} and in solution E_{IJ}^{solv} are calculated (in the stream files *vacuum.str* and *solution.str*, respectively) and the check for change in sign is performed.

The stream files *terminal.str*, *terminal2.str*, *terminal3.str*, and *terminal4.str* are used in *loop.str* and *loop2.str*, respectively, to ensure adequate group definition at segment borders in CHARMM.

6.1 Initialization

The initialization of storage vectors is done in *initialize.str*. The following storage vectors are used:

- SCA1 storage of radii
- SCA2 storage of charges
- SCA3 intermediate storage of pairwise interaction energies in vacuum
- SCA4 intermediate storage of pairwise interaction energies in solution
- SCA5 intermediate storage of pairwise scaling factors
- SCA6 storage vector for scaling factors
- SCA7 intermediate storage of pairwise interaction energies in solution within cutoff distance E_{IJ}^{solv}
- SCA8 storage of interaction energies in solution between each group and the rest of the protein within cutoff distance E_{I}^{solv}
- SCA9 intermediate storage of electrostatic potential

initialize.str also streams *defi19.str* or *defi22.str*, where the definition of backbone and side chain atoms is done. The backbone groups are defined over two residues: BACK1 includes the atom types C_α, H_α, N, HN; BACK2 includes C and O for param22. For param19 the definitions are accordingly. The termini and capping groups ACE and CBX (param19 only) are treated correctly. All atoms not belongig to BACK1 or BACK2 are supposed to be side chain atoms. Thus, each protein residue is split into two groups, a backbone group (spanning two amino acids) and a side chain group.

All non-protein moieties, such as water, ions, or ligands, are treated as protein residues without backbones. A scaling factor is determined for each group. The prolines are treated specially in that only one scaling factor is determined for both backbone and side chain atoms.

Finally, a selection is defined that contains exactly one atom per group.

7 Details about the script *verify_factors.inp*

In *initialize.str* several variables are initialized. After an energy call to initialize the energy function, *loop.str* is streamed. It loop over all residues for which the scaling factors shall be tested. After definition of the source group I, *loop.str* streams *calenergy.str*, which in turn streams *vacuum.str* to calculate E_I^{vac} with unscaled charges, extracts E_I^{solv}, and streams *shield.str* to calculate $E_I^{shield'}$ with scaled charges using the raw scaling factors.

7.1 Initialization

initialize.str streams *defi19.str* or *defi22.str*, calculates scaled charges and initializes the following storage vectors:

- SCA1 storage of scaling factors
- SCA2 storage of scaled charges
- SCA3 storage of original partial atomic charges
- SCA4 storage of E_I^{solv}
- SCA5 storage of E_I^{vac} (original charges)
- SCA6 storage of $E_I^{shield'}$ or E_I^{shield} (from scaled charges)

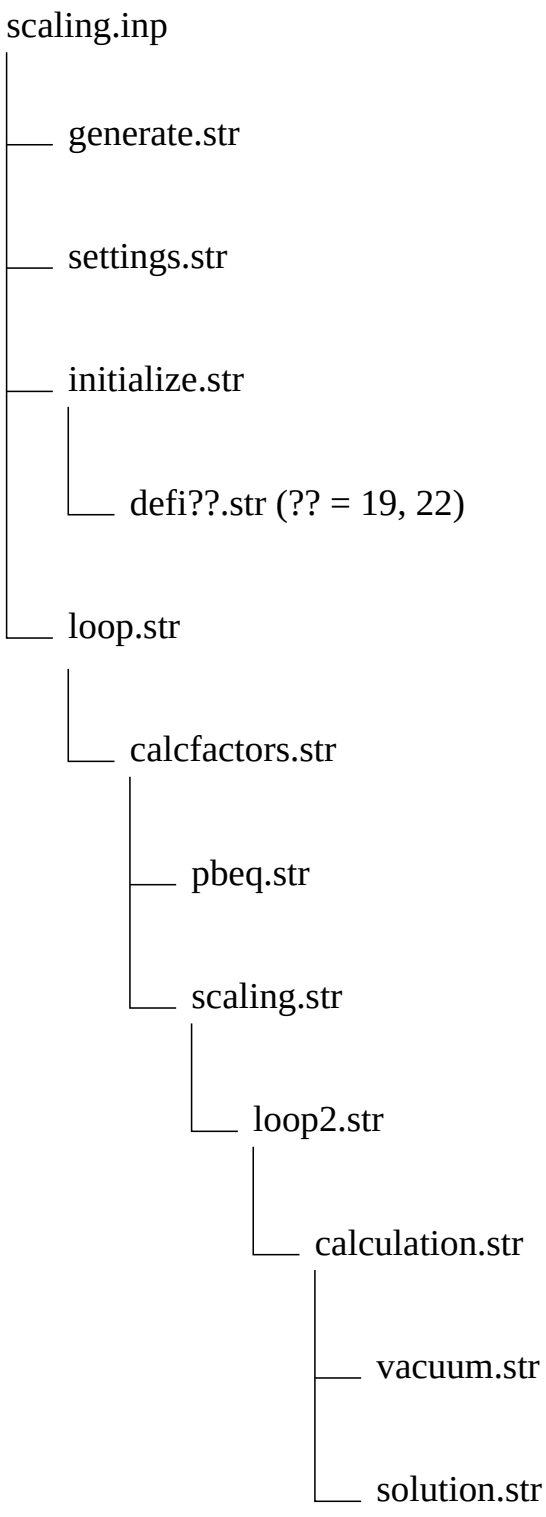

Figure 1: Flow of the charge-scaling script *scaling.inp*

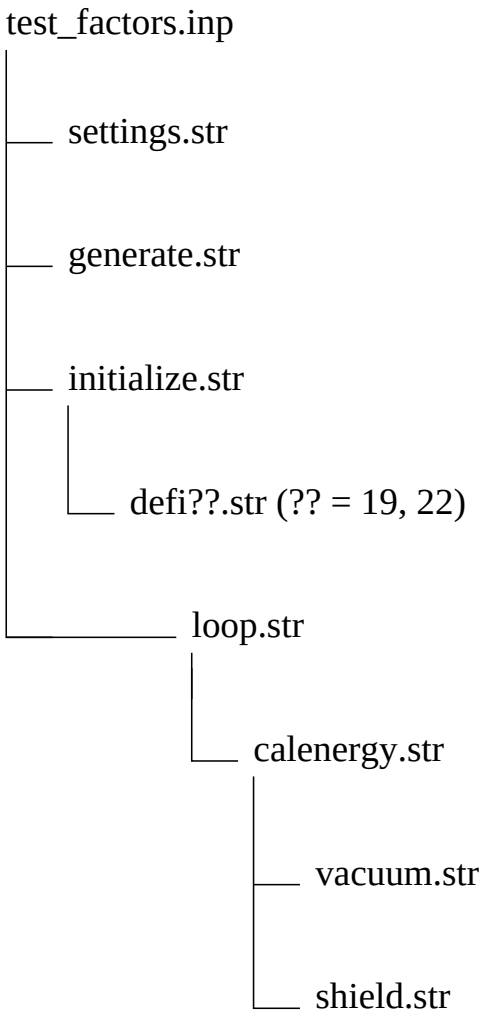

Figure 2: Flow of the counterscaling script *verify_factors.inp*.