Force field generation for a bioinorganic complex: the 'Complex' mode

Tutorial by:
F. Wang,[1] J.-P. Becker, [1] P. Cieplak,[2] and F-Y. Dupradeau[1]

(1) Université de Picardie - Jules Verne, Amiens
(2) Sanford-Burnham Institute for Medical Research, La Jolla, CA

November 2014

This tutorial is related to force field generation for a bioinorganic complex.

The quantum mechanics (QM) Gaussian program is selected during the input submission procedure using the web interface: one has to register and use a private account:
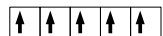See http://q4md-forcefieldtools.org/REDServer-Development/faq.php#3

1- General pieces of information
– *Spin multiplicity and total charge values*

The spin multiplicity of organic molecules/ground states generally equals one (controlled in PyRED with the `MOLECULE1-SPINMULT` keyword). On the contrary the transition metal ion within a bioinorganic complex with *d* orbitals rarely present a spin multiplicity of one [but the Zinc(II) ion]. Examples are provided below: the electronic configurations of the iron atom, the ferric ion (oxidation number of +III), the zinc atom, and of the zinc(II) ion (oxidation number of +II) are the following:[1]

```
Fe            [Ar].3d⁶.4s²
Fe(III)       [Ar].3d⁵       The spin multiplicity for the high spin system equals six.
```
$\uparrow\ \uparrow\ \uparrow\ \uparrow\ \uparrow$ *i. e.* the spin multiplicity equals 'the number of single electrons in the five *d* orbitals + 1'.
```
Zn            [Ar].3d¹⁰.4s²
Zn(II)        [Ar].3d¹⁰      The spin multiplicity equals one (no single electron).
```

The total charge (controlled with the `MOLECULE1-TOTCHARGE` keyword) of a bioinorganic complex depends on the formal charges of the metal ion and its different ligands.

– *The theory level used in QM computation*

An empirical force field is generally designed for a class of molecules, and is generated using a well-established strategy. For compatibility reason one has to follow that strategy, when one wants to extended that force field. The choice of the theory level used in QM computation is an important point for constructing a force field. The Hartree-Fock/6-31G(d) theory level is widely used for bioorganic and organic molecules.[2] On the contrary density functional theory (DFT) based methods are recommended for bioinorganic complexes,[3] because they include electron correlation (see the `RESP-X1` *versus* `RESP-A1` charge model in PyRED, as well as the `METHOD_OPTCALC`, `BASSET_OPTCALC`, `METHOD_MEPCALC` and `BASSET_MEPCALC` keywords).

– *Problems encountered during QM computation*

Obtaining an optimized geometry for a bioinorganic complex is more tedious than for an organic molecule. Indeed, the presence of single electrons leads to SCF and geometry optimization

convergence problems in unrestricted computation. Wave function instability might also be observed. To solve putative problems encountered during geometry optimization and molecular electrostatic potential computation the `Complex` mode has been developed in PyRED (see the `MOD_GAUSSIAN_JOB` keyword). It allows automatic generation of input files, and re-execution of the Gaussian program when a calculation does not correctly end. For this purpose a database of error messages produced by Gaussian has been implemented in PyRED allowing the efficient generation of inputs and the execution of Gaussian as a function of these error messages. The wave function of a minimum is also automatically tested and optimized if found instable. Gaussian keywords in input files generated by PyRED have been designed to achieve efficient SCF and geometry optimization convergence. Energy values and orbital eigenvalues obtained for the minimum are also compared to these found after molecular electrostatic computation ensuring charge derivation and force field generation for a fully characterized minimum. A drawback of this `Complex` mode is that it requires more calculation time. ***It was also implemented only for the Gaussian program.***

– *The charge fitting step and the charge model*

The mono-polar approximation or atomic charge is empirical. In the PyRED approach atomic charges are obtained by fitting them to molecular electrostatic potentials computed by QM. RESP is a popular charge model, which is widely used for bioorganic and organic molecules.[4] For bioinorganic complexes the RESP charge value observed for the metal center buried within the complex might be in some cases unrealistic. The use of a high density of points, when computing the electrostatic potential (see the `SURFMK_MEPCALC` keyword) or three stages instead of two during charge fitting (see the `RESP-X11` *versus* the `RESP-X1` or `RESP-A1` charge model) might solve the problem. ESP atomic charges (see the `ESP-X1` charge model) might also be an alternative.

– *Preparation of the input molecule for a PyRED job*

If extracted from an experimental macro-structure such as a protein, a bioinorganic complex with a metal center corresponds to a molecular fragment with different open valencies. This structure has to be converted into a whole molecule using capping groups to be used in QM computation (see http://q4md-forcefieldtools.org/Tutorial/Tutorial-4.php#2). Indeed, a molecular fragment is empirical, and is consequently not recognized in QM computation. If the metal ligands in the complex are amino acids two approaches can be followed:
- The first one involves entire amino acid residues, whose the backbone of the central fragment of each amino acid can be capped with the acetyl and the NH-methyl groups. Here *trans* peptide bonds are generally selected.
- The second one only involves amino acid side chains (*i. e.* the amino acid backbone is removed) and the α-carbon atom can be represented by a capping methyl group.

The first approach is supposed to be more representative of the bioinorganic complex within the protein than the second one, but it involves more atoms, and consequently requires more calculation time.

– *Checking the model obtained after geometry optimization*

If extracted from an experimental macro-molecule the structure obtained after geometry optimization performed by QM for the capped bioinorganic complex has to be carefully compared with that in the macro-structure. Two cases can be observed:
- In the first one the differences observed between the optimized geometry obtained by QM and the structure within the macro-molecule remains acceptable. The QM optimized geometry is thus satisfying, and can be involved in force field generation.

- In the second one the differences observed between the experimental and QM structures are too important. The QM optimized geometry is therefore not representative of the structure within the protein, and should be discarded. Two different approaches could be followed to obtain a more representative optimized geometry:
  - In the first one geometry optimization can be carried out using an implicit solvation model (instead of gas phase geometry optimization, which is the approach used by default). The last version of PyRED allows the use of implicit solvation during geometry optimization (see the definition the SCRF algorithm within the BASSET_OPTCALC and BASSET_MEPCALC keywords).
  - In the second one geometry optimization can be carried out using geometrical constraint(s), where the position of atoms, and/or bond(s), angle(s) and dihedral angle(s) can be frozen during the procedure to force the geometry of QM minimum to keep the shape of the complex within the protein (see the MOLECULE1-GEOMOPTCONST keyword).

  – *Putting the model back within the protein*

If extracted from a macro-molecule the structure of the bioinorganic complex has been capped to be involved in QM computation and empirical force field generation. However, the user is generally interested in a bioinorganic complex in the form of a molecular fragment with open valencies, which can be inserted back within the protein from which it originates. This molecular fragment is generated by PyRED by designing specific charge constraints, which are applied during the charge fitting step (*i. e.* after QM computation) on the capping groups added to construct the whole molecule (see the MOLECULE1-INTRA-MCC1, MOLECULE1-INTRA-MCC2, MOLECULE1-INTER-MCC1, MOLECULE1-INTER-MCC2 keywords), and by removing these capping groups with constrained charge values (see the principle applied on an amino acid dipeptide used to generate the corresponding central fragment http://q4md-forcefieldtools.org/Tutorial/Tutorial-4.php#16).[5]

2- Generate a bioinorganic complex in the PDB file format
Input molecules of PyRED are provided in the PDB file format:
See http://q4md-forcefieldtools.org/REDServer-Development/Documentation/readme.txt

The user has to pay particular attention to ambiguities, that arise in atom names when for instance the α carbon atom of an amino acid bears the CA atom name. Indeed the CA atom name is recognized by default as a calcium atom. Similar problems are observed for Hγ hydrogen atoms (HGx atom names) *versus* the mercury element and the Cδ carbon atom (CD atom name) *versus* cadmium. Providing chemical elements in the PDB input file allows solving these ambiguities (see the *readme.txt* file).

The formal charge of the charged groups of the bioinorganic complex can be given in the PDB input file providing an alternative way to determine the total charge of the molecule (see the *readme.txt* file).

Concerning the topology of the bioinorganic complex two approaches could be followed:
- In the first/default one the user lets the atom connectivities be determined by PyRED. Consequently the bonds between the metal ion and its ligands are considered. In this case, no charge constraint is applied on the metal ion during charge fitting letting the charge value of the metal ion be delocalized on the ligand atoms. This approach results in a metal ion with a non-integer partial charge value, that is physically bound to its ligands.

- In the second one the user wishes to control the bioinorganic complex topology to remove the bonds between the metal ion and its ligands, and provides a charge constraint for the metal ion set to a value, which equals its oxidation number. This approach results in a metal ion with an integer partial charge value, that is not physically bound to its ligands.

3- Define specific keywords in the Project.config file
See http://q4md-forcefieldtools.org/REDServer-Development/Documentation/Project.config

```
# Define the title of the molecule (optional):
MOLECULE1-TITLE     = Molecule_title
# Define the spin multiplicity of the metal center:
MOLECULE1-SPINMULT  = ..
# Define the total charge of the molecule:
#    (the total charge can also be defined in the PDB input file)
MOLECULE1-TOTCHARGE = ..
# Keeping the atom order of the PDB input file allows direct RSMD calculation
#    between the QM optimized geometry and the input structure/experimental data
MOLECULE1-ATMREORDR = OFF
# Concerning the topology of the complex two approaches can be followed:
# either set MOLECULE1-CALCONECT = On (default)
#    i. e. the atom connectivities are automatically determined.
#         Consequently these between the metal ion and its ligands are considered.
#     or set MOLECULE1-CALCONECT = OFF
#       and provide atom connectivities in the PDB input file
#           removing these between the metal ion and its ligands.
#       constrain the charge value of the metal ion to its oxidation number
#           adding below an additional MOLECULE1-INTRA-MCC1 with the 'K' flag
MOLECULE1-CALCONECT    = ON/OFF
# Define charge constraints for generation of a molecular fragment:
# Usually a constraint with a value of zero is used for a capping group.
# Different charge constraints can be defined to generate different open valencies
# Use the 'R' flag to remove the atoms from the force field library
MOLECULE1-INTRA-MCC1 = 0.0 | 1 2 3 4 5 6 | R
MOLECULE1-INTRA-MCC1 = ...
# Force the partial charge of a metal ion(III) (index in the atom order = 100)
#    to take the +3 value:
MOLECULE1-INTRA-MCC1 = 3.0 | 100 | K
# Request geometry constraints during geometry optimization (if needed):
# Different geometry constraints can be defined
# For instance the position of the α-carbon atom of each metal ligand can be fixed
MOLECULE1-GEOMOPTCONST = ...
MOLECULE1-GEOMOPTCONST = ...
```

4- Define specific keywords in the System.config file
See http://q4md-forcefieldtools.org/REDServer-Development/Documentation/System.config

```
# Request geometry optimization:
#    (using OPT_Calc = On2 can save substantial cpu time in a first approach)
OPT_Calc    = On
# Request molecular electrostatic potential computation (default)
MEPCHR_Calc = On
```

```
# Request the 'Complex' mode of PyRED (normal is the default):
MOD_GAUSSIAN_JOB = Complex
# Request frequency computation (demand more calculation time):
#    Useful to characterize the minimum (when geometrical constraints are used)
Freq_Calc = On
# Define the charge model:
#    Request two stage charge fitting i. e. RESP-X1 in a first approach
#    Compare with one and three charge fitting i. e. ESP-X1 and RESP-X11
#        To do that use the 'Re_Fit' mode:
#        See http://q4md-forcefieldtools.org/Tutorial/Tutorial-4-demo1.pdf
CHR_TYP = RESP-X1
# Use the four keywords below to overwrite default theory levels defined for
#    the CHR_TYP charge model:
# Request triple-ζ basis,[3] but cpu demanding
# Other DFT methods can be requested - instead of B3LYP[3]
# QM computation can also be carried out in presence of implicit solvent
METHOD_OPTCALC = OPBE
BASSET_OPTCALC = 6-311G(d,p)    # Triple ζ basis set; cpu demanding!
METHOD_MEPCALC = OPBE
BASSET_MEPCALC = 6-311G(d,p)
# Affect the density of points of the electrostatic potential:
# This was defined by U. Ryde; might not be useful if CHR_TYP = RESP-X11 is used
SURFMK_MEPCALC = IOp(6/33=2,6/41=10,6/42=17)
```

5- Provide a frcmod.user file

A *frcmod.user* file can be given as input file to PyRED to provide force field parameters. However it is more suitable to submit a first PyRED job, analyze the force field generated in this first job and use the 'Re_Fit' mode to provide these missing force field parameters in a second step.
See http://q4md-forcefieldtools.org/Tutorial/Tutorial-4-demo1.pdf

Completing a force field for a bioinorganic complex

– The equilibrium value of a bond or an angle can be measured from the minimum obtained after QM geometry optimization carried out with a high theory level. In this context removing the backbone as well as the α-carbon atom and its hydrogen atoms permits to reduce the number of atoms of the complex, and to use a triple-ζ basis set. Using OPT_Calc = On2, MEPCHR_Calc = Off and MOD_GAUSSIAN_JOB = Complex in the *System.config* file leads to a PDB output file, which can be directly used to measure needed equilibrium values.
– The force field constant of a bond or an angle can be deduced by analogy to existing ones (see literature as well as data in R.E.DD.B. http://q4md-forcefieldtools.org/REDDB/). It can also be computed from second derivatives (using the Gaussian fchk file).[6]
– Dihedral angle values involving a metal center could be set to zero, or better, determined from a dihedral scan, and improper dihedral angle could be discarded from the list of needed force field parameters in a first approach.
– Finally, van der Waals parameters can be obtained from literature. PyRED internally stores van der Waals parameters found in literature, and judged compatible with the generated force fields.

6- Create an archive file to upload the different PyRED inputs
```
tar -zcvf Archive.tgz  Mol_red*.pdb  Project.config  System.config
```

Or when using the 'Re_Fit' mode; see http://q4md-forcefieldtools.org/Tutorial/Tutorial-4-demo1.pdf

```
tar -jcvf Archive.tar.bz2  Mol_red*.pdb  *.config  frcmod.user  Data-R.E.D.Server/*
```

7- Using the force field generated by PyRED within the AmberTools

The force field generated by PyRED can be found in the following sub-directories:
- The *force field libraries* are available in the `Mol_m1` or in the `Mol_MM` sub-directory (the mol3 file format is the default; see http://q4md-forcefieldtools.org/Tutorial/leap-mol3.php) if the number of PDB input file(s) equals one, or is strictly greater than one, respectively.

- The *force field parameter files* are available in the *Data-Default-Proj* sub-directory.
  Three files are generated:
  - the *frcmod.known* file lists the known force field parameters determined by PyRED based on the atom types assigned to the input molecules,
  - the *frcmod.corespondence* file lists the force field parameters, which were determined by analogy to the existing ones,
  - the *frcmod.unknown* file lists the force field parameters, which remains unknown.
    The 'Re_Fit' mode can be used to provide a *frcmod.user* input file in agreement with the atom types assigned by PyRED, or after update of these atom types (see the `MOLECULE1-ATMTYPE` keyword):
    See http://q4md-forcefieldtools.org/Tutorial/Tutorial-4-demo1.pdf

- The *leaprc.q4mdfft* script available in the *Data-Default-Proj* sub-directory allows the direct use of the force field generated by PyRED within the LEaP program (see the Amber tools at http://ambermd.org/#AmberTools):

```
# Extract the PXXXX.tar.bz2 R.E.D. Server Dev./PyRED job after download
tar -jxvf PXXXX.tar.bz2
cd PXXXX/Data-R.E.D.Server/Data-Default-Proj
# Edit to update the LEaP script and extend its use
gedit leaprc.q4mdfft &
# Run the LEaP program:
# see http://q4md-forcefieldtools.org/REDServer-Development/images/PyRED.gif
xleap -f leaprc.q4mdfft
```

References
[1] E.I. Solomon, R.A. Scott, R.B. King, Computational Inorganic and Bioinorganic Chemistry, John Wiley & Sons, 2013.
[2] F. Jensen, Introduction to Computational Chemistry, John Wiley & Sons, 1999.
[3] C.J. Cramer, D. G. Truhlar, *Phys. Chem. Chem. Phys.* 2009, 11, 10757.
[4] C.I. Bayly, P. Cieplak, W.D. Cornell, P.A. Kollman *J. Phys. Chem.* 1993, 97, 10269.
[5] P. Cieplak, W.D. Cornell, C.I. Bayly, P.A. Kollman *J. Comput. Chem.* 1995, 16, 1357.
[6] J. M. Seminario, *Int. J. Quantum Chem.* 1996, 60, 1271.

Last update of this tutorial: November 30th, 2014.