FragBuilder: An efficient Python library to setup quantum chemistry calculations on peptides models

Anders S. Christensen[1], Thomas Hamelryck[2], and Jan H. Jensen[1]

[1]Department of Chemistry, University of Copenhagen, Copenhagen, Denmark
[2]Bioinformatics Centre, Department of Biology, Bioinformatics group, University of Copenhagen, Copenhagen, Denmark

December 2013

We present a powerful Python library to quickly and efficiently generate realistic peptide model structures. The library makes it possible to quickly set up quantum mechanical calculations on model peptide structures. It is possible to manually specify a specific conformation of the peptide. Additionally the library also offers sampling of backbone conformations and side chain rotamer conformations from continuous distributions. The generated peptides can then be geometry optimized by the MMFF94 molecular mechanics force field via convenient functions inside the library. Finally, it is possible to output the resulting structures directly to files in XYZ and PDB formats, or optionally directly as input files for a quantum chemistry program. FragBuilder is freely available at https://github.com/jensengroup/fragbuilder/ under the terms of the BSD open source license.

1

1 Introduction

Modeling of chemical properties of proteins is a challenging task in modern computational biochemistry, mainly due to the large number of atoms that need to be treated computationally, compared to the computational speed of modern computers. Although theoretical methods to treat large systems are being developed, it is computationally more feasible to investigate properties of small, representative, protein-like structures. For example, calculations on peptides have been used to parametrize protein-specific molecular mechanics force fields, and NMR properties of proteins such as chemical shifts and spin-spin coupling constants [Mackerell, 2004, Vila et al., 2009, Case et al., 2000].

Recently, we have used the presented Python library to carry out calculation on peptides modeling the backbone of a protein in the parametrization of amide proton chemical shifts [Christensen et al., 2013]. Since this study, we have carried out more than 1.5 million quantum mechanical geometry optimization and NMR shielding calculations on protein-like peptides in order to extend our model of protein chemical shifts. Naturally, an efficient and stable method is needed in order to generate such a number of peptide models. Two recent programs that can generate peptide structures are the Ribosome program[Srinivasan, 2013] and the PeptideBuilder library [Tien et al., 2013]. The Ribosome program is written in FORTRAN and thus difficult to extend and therefore not ideal for use in an automated, scripting fashion. The PeptideBuilder library is written in Python and is therefore very attractive for this purpose. Our library which is presented here is very similar to PeptideBuilder, but offers a number of additional features which we found necessary for our purpose. Most importantly, our library includes methods for geometry optimization with a molecular mechanics force field, efficient conformational sampling from continuous probability distributions and lastly output to a variety of output formats or, optionally, directly as input file for a quantum chemistry program. Currently Gaussian 09 is supported [Frisch et al., 2009].

2 Methods

FragBuilder is implemented in Python and is a library that can be imported and used in simple Python scripting style. Python is attractive, since a very large number of scientific libraries are already available in Python, and thus easy to extend and combine with new code. FragBuilder is implemented using the Open Babel library as back-end for handling the molecular structure of the peptide via existing classes and methods [O'Boyle et al., 2011]. The methods present in FragBuilder thus have access to a multitude of existing chemistry and cheminformatics related library routines which are maintained separately by Open Babel. Especially, the code for manipulating a molecular structure, molecular mechanics and file writers from Open Babel are used in FragBuilder. FragBuilder also comes with the BASILISK library which can sample protein backbone and side chain conformations from a joint probability distribution [Harder et al., 2010].

The only dependencies for running FragBuilder are the NumPy mathematics library [Oliphant, 2006] and Open Babel with Python bindings. These packages are already available through package managers on virtually every recent Linux distribution, or otherwise freely available and open source.

2

3 Functionality and usage

The functionality to create a peptide is implemented in the `Peptide` class which is imported from `fragbuilder` module. A typical work flow creates a peptide, defines torsion angles, performs a constrained geometry optimization and finally writes the resulting structure to a file. A chart describing a typical use case is displayed in Fig. 1, and detailed examples of the functionality of FragBuilder are given below.

3.1 Creating peptides

The structure of a peptide molecule is generated as a Python object by using the `Peptide` class instantiated with the sequence as argument. The `Peptide` class has access to classes for each type of residues which each contain a structure for that residue in XYZ format. Routines from Open Babel are then used to automatically rotate, translate, and connect the residues. Finally the structure is stored in the `Peptide.molecule` class variable as an Open Babel `OBMol`.

The sequence interpreted uses the single letter abbreviation for each amino acid. E.g. `Peptide("GLG")` will create a glycine-leucine-glycine tripeptide molecule which can then be manipulated through the interface. The minimal code to achieve this could be:

```
1 from fragbuilder import Peptide
2 pep = Peptide("GLG")
```

As default values, the ϕ, ψ and ω backbone torsion angles are set to $-120°, 140°$ and $-180°$, which corresponds to a typical extended β-strand. The side chain torsion χ angles are set so two neighboring side chains will not have steric clashes when side chain torsion angle input is given. After the peptide has been instantiated, the structure can be manipulated through built-in methods. Several convenient methods of the `Peptide` class are presented in the next sections.

3.2 Setting dihedral angles

The `Peptide` class allows for dihedral angles to be manually specified through setter and getter type functions that set or read backbone and side chain torsion angles. Examples of torsion angles that can be set in FragBuilder are shown in Fig. 3.

Alternatively, it is possible set torsion angles to values from predefined distributions, such as the Ramachandran-plot for backbone angles or rotamer distributions for side chain angles. This allows for fast and efficient sampling of realistic peptides conformations and rotamer distribution without the need for a molecular dynamics or Monte Carlo simulation. For this purpose FragBuilder includes the BASILISK library and convenient methods to access BASILISK from the `Peptide` class.

3.2.1 Manual input of dihedral angles

For example, making a glycine-leucine-glycine peptide and setting the backbone angles to $\phi = -60.0°$ and $\psi = -30.0°$, and side chain angles to $\chi_1 = 180°$ and $\chi_2 = 60°$ of the leucine (residue 2) can be done through the following code:

3

```
1 pep = Peptide("GLG")
2 pep.set_bb_angles(2, [-60.0, -30.0])
3 pep.set_chi_angles(2, [180.0, 60.0])
```

This way it is possible to precisely specify dihedral angles manually. This code can be used, for instance, to set up a scan of torsion angles or making peptides with geometries extracted from experimental structures. An example of a scan is shown in Fig. 4. This scan was created in the following manner:

```
1 pep = Peptide("GLG")
2 for i in range(10):
3     pep.set_bb_angles(2, [-120.0, 100.0+20.0*i])
4     pep.write_xyz("pep_%i.xyz" % (i))
```

The method `Peptide.write_xyz()` writes the structure to a file in XYZ format and is described later in this section.

3.2.2 Sampling dihedral angles from BASILISK

BASILISK is a dynamic Bayesian network trained on a large set of representative structures from the Protein Data Bank[Berman et al., 2000] and is able to sample backbone angles and side chain angles. BASILISK makes use of directional statistics - the statistics of angles, orientations and directions - to formulate a well-defined joint probability distribution over side and main chain angles. Backbone angles are essentially sampled from the Ramachandran-plot via BASILISK. Similarly, side chain angles are sampled from corresponding rotamer distributions. The distributions offered by the BASILISK library are continuous, in contrast to most approaches based on discrete rotamer libraries. BASILISK can sample side chain angles either in a backbone conformation-dependent mode or -independent mode (where backbone dependency is the default behavior).

The following code will sample sets of backbone and side chain torsion angles in the second residue (leucine) of a glycine-leucine-glycine.

```
1 pep = Peptide("GLG")
2 new_bb = pep.sample_bb_angles(2)
3 new_chi = pep.sample_chi_angles(2)
```

Using the sampler functions in the `Peptide` will set the torsion angles of the residue and return the new set of angles. It is also possible to get samples from BASILISK via Frag-Builder by using the `fragbuilder.BasiliskDBN` class which provides direct access to the BASILISK library.

3.3 Capping peptides

One aspect of carrying out quantum mechanical calculations on peptide fragments is the way the peptide strands are terminated or capped. This can be important, since the properties calculated from a quantum mechanical calculation may be affected by how the protein is truncated to a model peptide.

By default, FragBuilder generates methyl caps by adding a CH_3-C(=O)- group to the N-terminus and an -NH-CH_3 group to the C-terminus. Additionally, it is possible to cap the

4

ends of the peptide as normal N- and C-termini (amine or carboxyl groups, respectively). These can be set to either a charged or a neutral state. See Fig. 5 for a schematic of the possible caps.

The type of cap can be specified through a keyword argument when the Peptide object is instantiated. If no cap is specified, methyl caps are generated by default. A glycine-leucine-glycine residue with a positively charged N-terminus and a neutral C-terminus is generated by the following code:

```
1 pep = Peptide("GLG", nterm="charged", cterm="neutral")
```

3.4 Optimization

An occurring problem when generating peptides with a specific set of dihedral angles is that the structure may in some cases contain steric clashes. We found this prevented us from starting quantum mechanical geometry optimization on the structures, even when these were generated to match angles from experimental structures. Typical problems with these structures were SCF convergence issues and very large molecular gradients which cause the program to fail.

To circumvent this issue, FragBuilder offers specialized optimization routines, specifically designed to constrain the dihedral angles of peptides while removing steric clashes. Optimization is performed through the Open Babel which provides access to several force fields and a number of optimizers. The MMFF94 force field [Halgren, 1996] is arguably the most advanced force field for biomolecules in Open Babel and is used exclusively in FragBuilder along with the conjugate gradient method. FragBuilder offer three kinds of optimization methods in the Peptide class.

The method Peptide.optimize() will perform a conjugate gradient optimization of the peptide with no restraints, until the default convergence criterion of Open Babel is reached ($\Delta E < 1.0 \ 10^{-6}$ kcal/mol or a max of 500 steps). Another option is to impose harmonic constraints on all dihedral angles. This is achieved through an extra keyword, i.e. Peptide.optimize(constraint=True). This will perform a conjugate gradient minimization through Open Babel with harmonic potentials on ϕ, ψ and ω backbone angles as well as all side chain χ angles.

A harmonic potential does not keep torsion angles completely fixed during optimization, and after convergence they deviate slightly from the starting values. For situations where this is problematic, FragBuilder is offering a routine termed "regularizing" which is accessed via the Peptide.regularize() method.

Regularizing cycles between a few constrained geometry optimization steps and resetting the dihedral angles to the initially specified angles, until self consistency is reached. A default regularization cycles 10 times between 50 conjugate gradient steps and angle resets. In most cases this converges the constrained optimization to less than $0.002°$ from the specified dihedral angles, which are then set to the specified values.

We found our regularization procedure with flexible bond lengths and angles through the MMFF94 force field to allowing convergence of QM calculations in many cases, which would have been hindered by steric clashes due to fixed bond length and angles.

A similar approach to avoid spurious conformations has been adopted by Vila *et al.* in the creation of the CheShift chemical shifts predictor, which is parametrized from quantum

5

mechanical calculations on model peptides [Vila et al., 2009]. Here bond angles and lengths are simply set to the standard values of the ECEPP/3 force field [Nemethy et al., 1992]. Subsequently the internal energy of the peptide is calculated with the ECEPP-05 force field and any conformation with an internal energy > 30 kcal/mol is rejected as being unphysical.

Fig. 6 shows an example of a tryptophan-aspartate-glycine peptide with methyl caps in which the backbone torsion angles are taken from the experimental structure of xylanase (PDB-code: 1XNB), residues 99-101. This choice of angles causes a clash between a hydrogen bonding O...H pair, and a geometry optimization at the B3LYP/6-31+G(d,p) level in Gaussian 09 could not start due to a large gradient in the initial geometry. Regularization removes the clash, while retaining the specified dihedral angles.

A peptide can be created and regularized using the following code, which also prints the MMFF94 force field energy in units of kcal/mol:

```
1 pep = Peptide(sequence)
2 # The user can manipulate the structure here angles here
3
4 pep.regularize()
5 print pep.get_energy()
```

3.5 File output and interface to QM programs

Open Babel provides very flexible file readers and writers. The `Peptide` class wraps Open Babel with functions to directly write the geometry of a `Peptide` object to a file in XYZ or PDB format. This can be done simply as:

```
1 pep = Peptide(sequence)
2 pep.write_xyz("pep.xyz")
3 pep.write_pdb("pep.pdb")
```

It is also possible to write to any of the nearly 100 formats supported in Open Babel by using the method `Peptide.write_file(filetype, filename)` which offers direct access to Open Babel's `OBConversion.WriteFile()` method.

FragBuilder additionally offers an interface to write input-files for Gaussian 09, beyond the capabilities of Open Babel. Currently, it is possible to set up geometry optimization, single-point energy calculations and calculation of NMR shielding.

An example for a simple work flow that will generate a file for geometry optimization of a peptide in Gaussian 09 at the B3LYP/6-31G(d) level (using the `fragbulder.G09_opt` class) is as follows:

```
1 from fragbuilder import Peptide, G09_opt
2
3 pep = Peptide(sequence)
4 # The user can manipulate the structure here
5
6 opt = G09_opt(pep)
7 opt.set_method("B3LYP/6-31G(d)")
8 opt.write_com("pep.com")
```

6

If no method or basis set is specified, the file writer defaults to PM6[Stewart, 2007] for geometry optimization. Other classes that interface to Gaussian 09 are the `fragbuilder.G09_NMR` and `fragbuilder.G09_energy` classes, which are imported and instantiated similarly.

4   Conclusion

We have implemented routines to generate peptide models, from either specific geometries or efficient conformational sampling through the BASILISK library. We have furthermore implemented necessary code to perform constrained geometry optimizations of the peptide models, remove steric clashes and prepare the structure for use in a quantum chemistry program . In addition, we offer file writers to useful formats.

The `Peptide` class wraps functionality from Open Babel offered through its Python interface. The molecular structure is stored as an Open Babel `openbabel.OBMol` object in the `Peptide.molecule` class variable. This means that developers and users effectively have access to all the tools present in Open Babel to further manipulate the structure, or extend FragBuilder by wrapping and combining functionality from Open Babel.

FragBuilder is open source and published under the BSD 2-Clause license. Note that the packaged BASILISK library is published under the GNU General Public License version 3. FragBuilder is freely available at `https://github.com/jensengroup/fragbuilder/` where additional examples and full documentation can be found.

5   Acknowledgements

References

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28:235–242.

[Case et al., 2000] Case, D. A., Scheurer, C., and Brüschweiler, R. (2000). Static and dynamic effects on vicinal scalar j couplings in proteins and peptides: A md/dft study. *Journal of the American Chemical Society*, 122:10390–10397.

[Christensen et al., 2013] Christensen, A. S., Linnet, T. E., Borg, M., Boomsma, W., Lindorff-Larsen, K., Hamelryck, T., and Jensen, J. H. (2013). Protein structure validation and refinement using amide proton chemical shifts derived from quantum mechanics. *PLOS ONE*, page (In press).

[Frisch et al., 2009] Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, Jr.,

7

J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J., and Fox, D. J. (2009). Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT 2009.

[Halgren, 1996] Halgren, T. A. (1996). Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17:490–519.

[Harder et al., 2010] Harder, T., Boomsma, W., Paluszewski, M., Frellesen, J., Johansson, K. E., and Hamelryck, T. (2010). Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, 11:306–318.

[Mackerell, 2004] Mackerell, A. D. (2004). Empirical force fields for biological macro-molecules: Overview and issues. *Journal of Computational Chemistry*, 25:1584–1604.

[Nemethy et al., 1992] Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S., and Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. improved geometrical parameters and nonbonded interactions for use in the ecepp/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry*, 96:6472–6484.

[O'Boyle et al., 2011] O'Boyle, N. M., Banck, M., a C Morley, C. A. J., Vandermeersch, T., and Hutchinson, G. R. (2011). Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33–46.

[Oliphant, 2006] Oliphant, T. (2006). NumPy. `http://www.numpy.org/` (Accssed 10 December 2013).

[Srinivasan, 2013] Srinivasan, R. (2013). Ribosome - program to build coordinates for peptides from sequence. Available at `http://roselab.jhu.edu/~raj/Manuals/ribosome.html` (Accssed 10 December 2013).

[Stewart, 2007] Stewart, J. (2007). Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *Journal of Molecular Modeling*, 13:1173–1213.

[Tien et al., 2013] Tien, M. Z., Sydykova, D. K., Meyer, A. G., and Wilke, C. O. (2013). Peptidebuilder: A simple python library to generate model peptides. *PeerJ*, page 1:e80.

[Vila et al., 2009] Vila, J. A., Arnautova, Y. A., Martin, O. A., and Scheraga, H. A. (2009). Quantum-mechanics-derived 13Ca chemical shift server (cheshift) for protein structure validation. *Proceedings of the National Academy of Science*, 106:16972–16977.
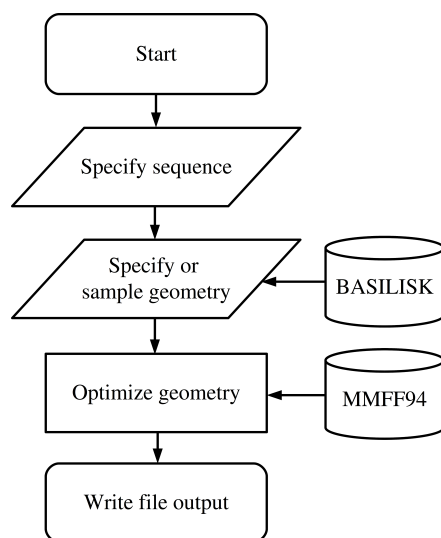
Figure 1: Simple chart of a common work flow using FragBuilder. First a peptide is generated from the sequence. Then torsion angles are set - either specified manually or sampled through BASILISK and a quick geometry optimization is performed using the MMFF94 force field. Finally, the structure is written to a file.
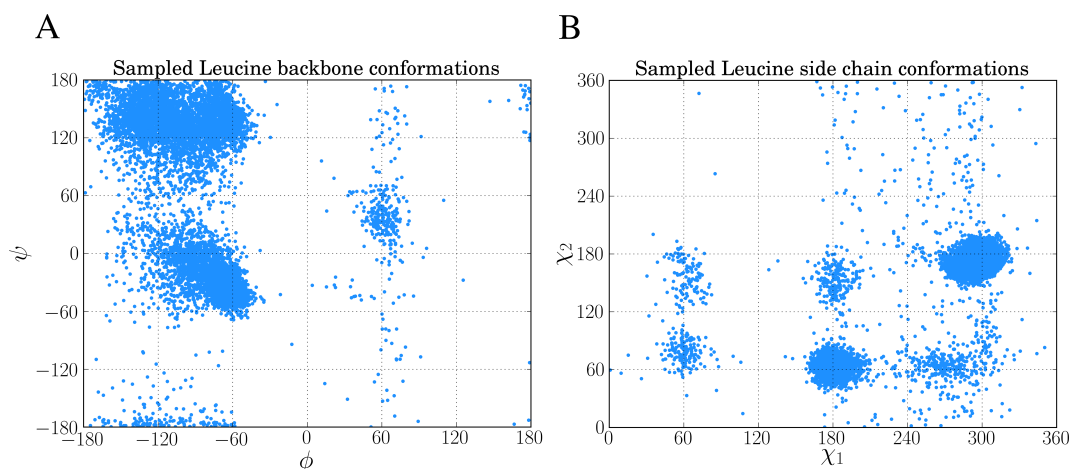

Figure 2: Examples of sampling dihedral angles through BASILISK in fragbuilder. 10,000 samples from BASILISK are shown for a leucine residue. (ϕ, ψ) backbone torsion angles pairs are shown in A and (χ_1, χ_2) side chain torsion angles pairs are shown in B.
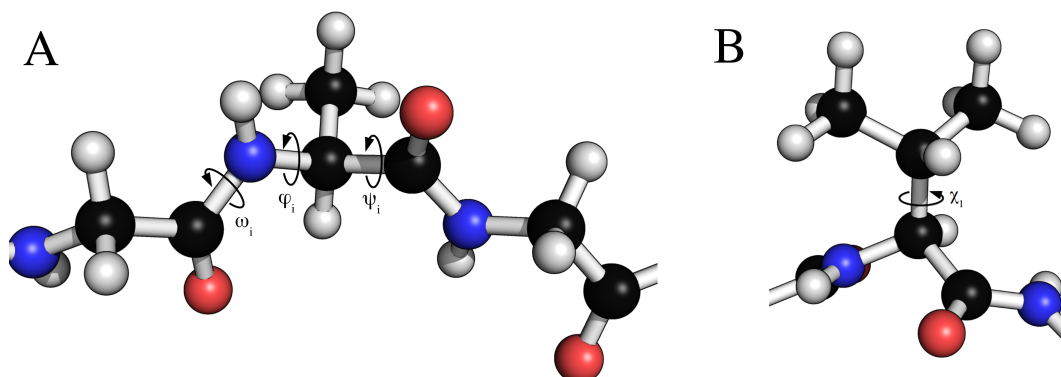
9

A   B

Figure 3: Examples of dihedral angles that can be set via FragBuilder. In A the backbone ω, ϕ and ψ torsion angles are shown for the i'th alanine residue of a peptide strand. In B, the χ_1 torsion angle is shown for a valine side chain.
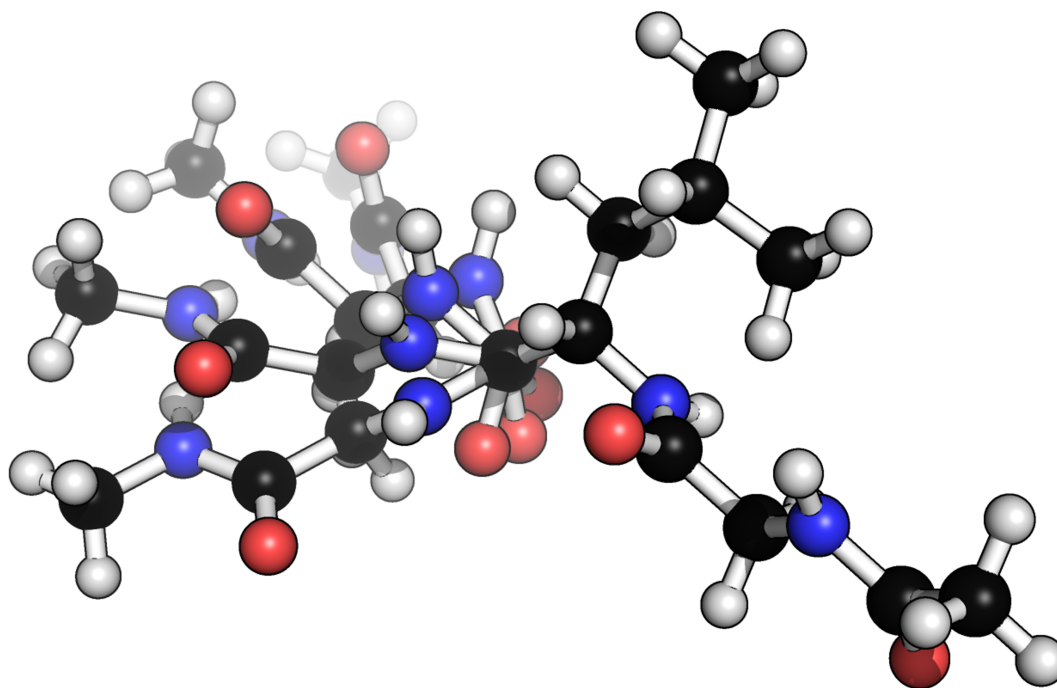
Figure 4: Example of four different conformers of a glycine-alanine-glycine tri-peptide, generated from a scan over the ψ backbone torsion angle of the alanine residue.
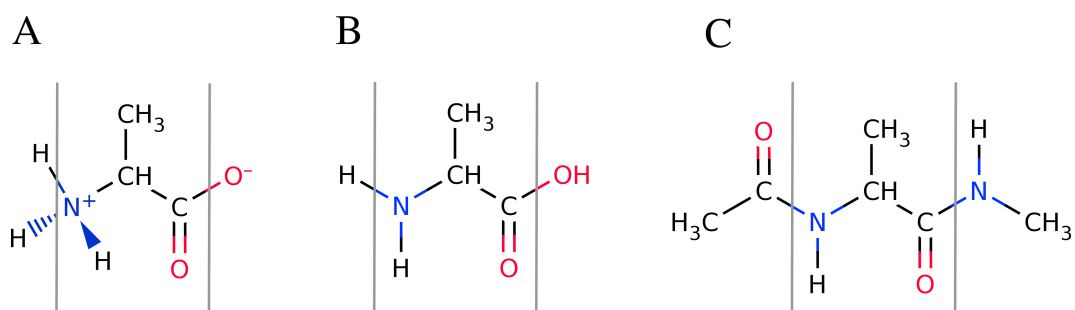
10

A B C

Figure 5: Overview of the available peptide-capping schemes available in fragbuilder. All three example shows an alanine residue (show between a set of gray lines). In A, the caps are the N- and C-termini in their charged states. In B the the caps are the N- and C-termini in their neutral states. In C the caps are methyl groups. Caps can be mixed and matched according to the users specifications.
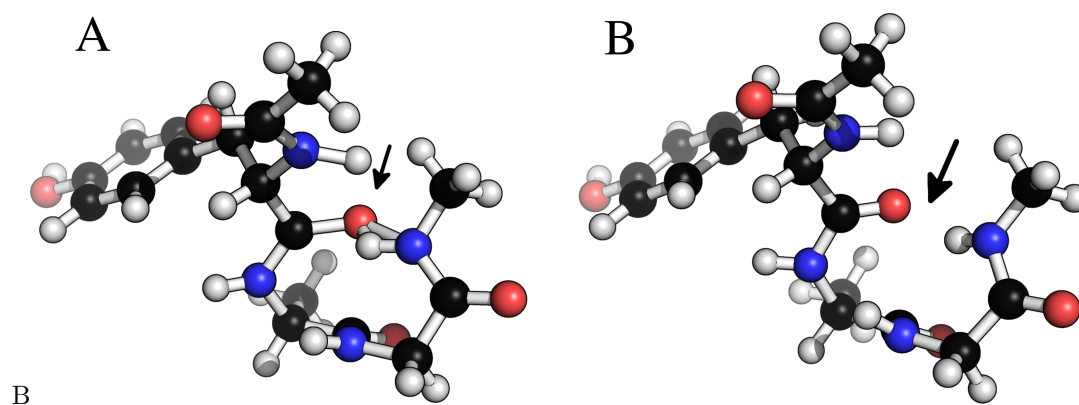
A B

B

Figure 6: Removing clashes by regularization in a tryptophan-apartate-glycine peptide. In A the peptide clashes between the amide proton on the C-terminal methyl cap and the amide oxygen in residue 1. in B this clash has been removed by constrained relaxation during the regularization procedure. Both structures have identical ϕ, ψ and ω backbone torsion angles.

11