Sidechain prediction validation set 1: 4BVG

Residues included in the consideration (for sidechain prediction and Prime minimization) are 144-180.195,199,204,207,210,227-234,248,251,291,294,324 plus intermediate in minimization.

The choice of the residues is based on the NAM binding pocket (C pocket) residues, which are identified by selecting residues within 7.5 Angstroms of NAM in SIRT3/INT/NAM complex converted from 4FVT or by docking NAM into 4BVG. The flexible loop was also included to make up the selection.

Two starting structures were used.

1. One used the 4BVG prepared using Protein Preparation Wizard and OPLS minimization on hydrogen only (without relaxation of heavy atoms).

|  |  |
| --- | --- |
| **Structures** | **Prime Energy** |
| Struct 1: 4BVG as prepared w/ h-opt | -9614.9 |
| Struct 2: Struct 1 w/ sidechain opt w/ backbone within 1 residue | -9871.5 |
| Struct 3: Struct 1 w/ sidechain opt w/ default option | -9896.1 |
| Struct 4: Struct 1 w/ sidechain opt w/ Monte Carlo, **rank2** | -10174.6 |
| Struct 5: Struct 1 w/ sidechain opt w/ Monte Carlo, rank1 | -10175.5 |
| Struct 6: Struct 1 w/ sidechain opt w/ CA-CB vector sampling | -10179.1 |
| Struct1 with prime minimization on selected residues | -10301.2 |
| Struct2 with prime minimization on selected residues | -10316.1 |
| Struct3 with prime minimization on selected residues | -10316.4 |
| Struct4 with prime minimization on selected residues | -10321.0 |
| Struct5 with prime minimization on selected residues | -10321.1 |
| The above structure with prime re-minimization on selected residues | -10320.9 |
| Struct6 with prime minimization on selected residues | -10320.7 |
| Struct 7: Struct 1 w/ sidechain opt w/ backbone within 3 residue | -10268.2 |
| Struct7 with prime minimization on selected residues | -10463.6 |
| The above structure with prime re-minimization on selected residues | -10467.1 |
| The above structure with prime re-minimization on selected residues | -10467.5 |

There are four options available in Prime Sidechain Prediction: Default, **Monte Carlo**, w/ **CA-CB vector sampling**, and w/ **backbone sampling** (by default with 3 residues each time, leading to extra residues included in optimization.) Reducing backbone sampling to 1 actually remove the backbone sampling, and has to be manually edited in the input file to carry out the calculation.

The results also show extra steps of minimization in testing the convergence.

The RMSDs for each residue are included in EXCEL file 4BVG\_h\_minimized\_as\_Starting\_for\_sidechain\_prediction\_1.xlsx.

Some convergence on structures can be identified. **Two minimums identified at around -10316 and -10321** with only the selected residues included. And the default sidechain prediction didn’t locate the better minimum. Residues with significant change from crystal structure were highlight in EXCEL file.

1. The structure prepared using Protein Preparation Wizard and OPLS minimization on hydrogen on 4BVG were **further minimized using Prime** and serve as a starting structure for further sidechain prediction investigation.

|  |  |
| --- | --- |
| **Structures** | **Prime Energy** |
| Struct 1a: 4BVG as prepared w/ h-opt followed by Prime minimization of selected residues | -10301.2 |
| Struct 2a: Struct 1a w/ sidechain opt w/ default option | -10261.6 |
| Struct 3a: Struct 1a w/ sidechain opt w/ CA-CB vector sampling | -10301.9 |
| Struct 4a: Struct 1a w/ sidechain opt w/ backbone within 1 residue | -10304.7 |
| Struct 5a: Struct 1a w/ sidechain opt w/ Monte Carlo, rank1 | -10322.7 |
| Struct 6a: Struct 1a w/ loop refine with ultra extended sampling (res. 156-169) rank 1 | -10334.8 |
| Struct2a with prime minimization on selected residues | -10289.7 |
| Struct3a with prime minimization on selected residues | -10304.3 |
| Struct4a with prime minimization on selected residues | -10320.4 |
| Struct5a with prime minimization on selected residues | -10324.8 |
| Struct6a with prime minimization on selected residues | -10374.4 |
| Struct 7a: Struct 1a w/ sidechain opt w/ backbone within 3 residue | -10464.5 |
| Struct7a with prime minimization on selected residues | -10470.2 |

The results above show that sampling error does exist as least for the **sidechain prediction with default option, as it located a structure with higher energy**. And various sidechain predictions also gave different minimums. The fact that loop refinement on only a subset of selected residues (but it includes extra residues within 7.5 A in optimization) reduced structure significantly suggest there may be potential energy errors as well.

The above two sidechain prediction run results point to a limited set of residues that contribute to the **change of energy** in sidechain prediction (as seem from RMSD data), which can be used in the future in making a small set of residues for optimization.

We carried out something similar for 4FVT using the carba-NAD+ as in the crystal structure, and almost the same choice of residues except carbaNAD+ and ac-LYS in the place of intermediate in 4BVG.

|  |  |
| --- | --- |
| **Structures** | **Prime Energy** |
| Struct 1: 4FVT prepared w/ Protein Preparation Wizard w/ h-opt | -10055.6 |
| Struct 2: Struct 1 w/ sidechain opt w/ backbone sampling of 1 residue | -10339.6 |
| Struct 3: Struct 1 w/ sidechain opt w/ default option | -10342.2 |
| Struct 4: Struct 1 w/ sidechain opt w/ CA-CB vector sampling | -10601.9 |
| Struct 5: Struct 1 w/ sidechain opt w/ Monte Carlo, rank1 | -10605.9 |
| Struct 6: Struct 1 w/ sidechain opt w/ Monte Carlo, rank2 | -10605.1 |
| Struct 7: Struct 1 w/ loop refine with ultra extended sampling (res. 156-169) rank 1 | -10382.2 |
| Struct1 with prime minimization on selected residues | -10667.1 |
| Struct3 with prime minimization on selected residues | -10715.9 |
| The above structure with prime re-minimization on selected residues | -10716.2 |
| Struct4 with prime minimization on selected residues | -10725.0 |
| Struct5 with prime minimization on selected residues | -10728.2 |
| Struct6 with prime minimization on selected residues | -10726.2 |
| Struct7 with prime minimization on selected residues | -10742.8 |
| Struct 8: Struct 1 w/ sidechain opt w/ backbone sampling of 3 residue | -10690.8 |
| Struct8 with prime minimization on selected residues | -10894.0 |

Further analysis on 4FVT will be provided on next work day.

Sidechain prediction methods:

The Predict Side Chains task predicts the conformation of side chains in one or more proteins by sampling orientations to minimize the energy. The sampling can include the orientation with respect to the backbone, and sampling of the backbone as well.

You can select residues for side-chain prediction with the Residues for side chain refinement atom selection tools.

The Sampling algorithm option menu allows you to choose a sampling algorithm for the prediction of side chains. These are primarily aimed at cases where some degree of backbone sampling is likely to be required, such as in homology modeling or cross-docking, where the backbone conformation of the initial structure might not be entirely correct. These methods allow for progressively more backbone movement during the side-chain prediction calculations.

There are four choices on the Sampling algorithm option menu:

• Default—No backbone sampling or reorientation of the CA – CB bond is performed.

• Monte Carlo—Perform Monte-Carlo sampling of side-chain conformations. When this algorithm is selected, you can select Maximum number of structures to return to return more than one side-chain conformation, and enter the desired number in the text box.

• With CA-CB vector sampling—Vary the orientation of the CA – CB bond by up to 30° from the initial direction.

• With backbone sampling—Sample the backbone by running a loop prediction on a set of 3 residues centered on the residue for which the side chain is being refined.

The structure refinement program uses the following procedure to re-predict conformations for the side chain set that you select.

* The default algorithm is as follows:

1. Side-chain rotamers are randomized for nonconserved residues (the default) or for all residues.

2. Beginning with the first residue to be predicted, the side-chain rotamer library is used to find the rotamer with the lowest energy while keeping all other side chains fixed. Once the process is complete for the first residue, the next residue is treated, and so forth until all have been done once (a single pass has been completed).

3. Once the pass is complete, Step 2 is repeated from the beginning, and then repeated again until the side-chain rotamers appear to be converged (no more changes are occurring).

4. Minimization is run on all of the side-chain atoms (but not backbone atoms) of the residues being treated.

Two methods are available for situations where some degree of backbone movement is likely to be required to get the correct side-chain conformation. This typically happens when working with homology models or when performing cross-docking studies, where the backbone conformation of the initial structure might not be entirely correct. These two methods allow for progressively more backbone movement during the side chain prediction calculations, and are described below. They are intended for use on a small set of residues, not the entire structure.

* With CA-CB vector sampling

Even minor discrepancies in backbone conformation can result in a slight shift in the location of the CB atom. This slight shift can often cause subsequent side-chain predictions to fail in placing the side chain optimally, particularly on large rigid aromatic residues. Allowing for sampling of the CA-CB vector in conjunction with the side-chain prediction can alleviate this problem. The method is turned on in the input file with the keyword SAMPLE\_CBETA. The algorithm is as follows:

1. CB positions in a conical region are systematically sampled around the initial position with a maximum displacement in the CA-CB vector of 30.0 degrees (settable with input file keyword MAXCONEANG).

2. At each position, the optimal side chain rotamer is selected.

3. The best overall combination of CB position and rotamer for the residue is selected from the results for all positions.

4. The sampling iterates over all requested residues until convergence is achieved in the same way as with conventional side chain prediction (see above).

5. Once all side-chain sampling has completed, the side chains and backbone atoms of the predicted residues are minimized.

* With backbone sampling

This method is intended for situations where even larger backbone movements may be required, but where a full loop prediction on the entire region may produce more movement than desired. It is turned on in the input file with the keyword SAMPLE\_BACKBONE.

For each residue for which side chains are predicted, a backbone loop segment of 3 residues (settable with the input file keyword BACKBONE\_LEN), centered about that residue is reconstructed using the normal loop prediction algorithm. Because this loop prediction includes the additional residues on either side of each requested residue, you should expect to see the conformations of these residues optimized as well (i.e. the region that is optimized extends somewhat beyond the residues actually selected for side-chain optimization).

Side-chain prediction is not performed for GLY and ALA residues with any of these methods.

Table 10.8. Keywords for sidechain prediction (SITE\_OPT) jobs

Keyword syntax Description

1. BACKBONE\_LEN n Number of residues to include in the loop defined for backbone sampling in side-chain prediction. Only used with SAMPLE\_BACKBONE. Default: 3.
2. MAXCONEANG value Maximum angular displacement of the CA-CB vector from the initial position when sampling CB positions. Only used with
3. SAMPLE\_CBETA. Default: 30°.
4. NITER\_SIDE n Number of iterations of side-chain prediction to perform. Larger numbers generally give better results. If set to zero, only the initial selection of rotamers without clashes is performed. Default: 1.
5. NPASSES n Number of passes through side chain optimization of the protein residues followed by minimization of the protein residues (including the backbone). These two steps are repeated the specified number of times before proceeding to optimization of the protein and the ligand. Default: 2. Files written from Maestro have NPASSES set to 1.
6. SAMPLE\_BACKBONE {yes | no} Sample the backbone by running a loop prediction on a set of residues centered on the residue for which the side chain is being refined. The number of residues in the loop is defined by the BACKBONE\_LEN keyword. Default: no.
7. SAMPLE\_CBETA {yes | no} Sample CB positions in a conical region around the initial position. The maximum displacement of the CA-CB vector in this region is set by the MAXCONEANG keyword. Default: no.