Building Molecular Structures for NAMD

Justin Gullingsrud
Theoretical Biophysics Group
Beckman Institute, UIUC

Why do we need psfgen?

- "I thought PDB files contained structure information already."
- Biomolecules can be represented in a variety of ways; many different force fields can be used to describe their interactions.
- Psfgen maps the *abstract representation* of a molecule in a PDB file to a *concrete representation* needed for an MD simulation.

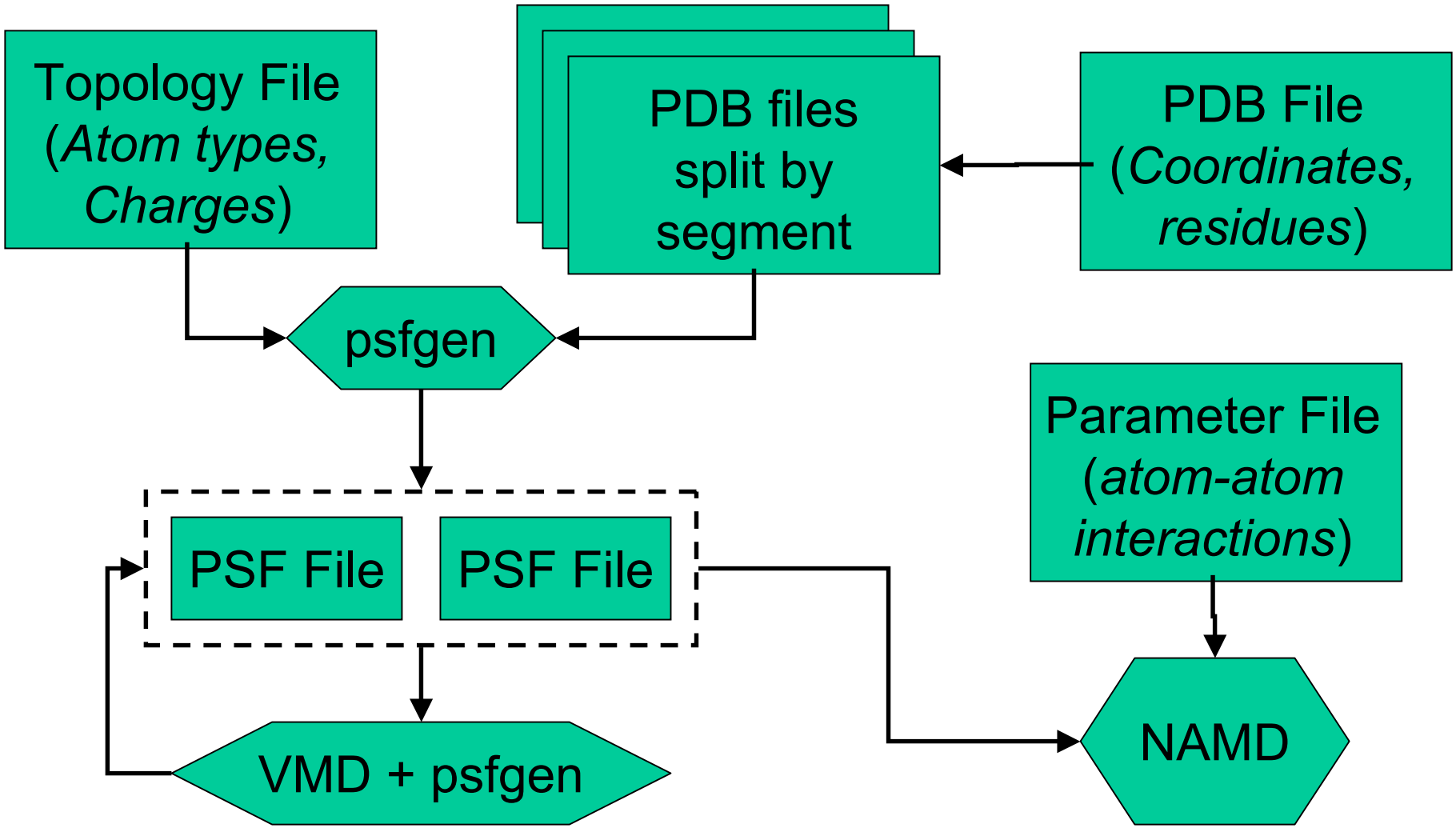
What does psfgen do?

- Maps residues to entries in a Charmm topology file.
- Links residues to form connected segments.
- Combines segments to form a complete structure file.
- Patches residues to form new covalent bonds or modify charge states.
- Guesses coordinates for missing atoms.
- Writes PSF and PDB files for NAMD.

What does psfgen not do?

- Arbitrary manipulation of structure, including mutating side chains.
- Translation or rotation of coordinates.
- Automatic hydration of molecules.
- Determination of protonation states.
- Force and energy evaluation.
- **However:** the first three things can all be done in combination with VMD.

Structure building flowchart



Data files for psfgen

- Topology files:
  - Atom definitions (just the mass)
  - Residue definitions:
    - atom names, types, and charges;
    - bonds and impropers (but not angles and dihedrals)
  - Patches for initial, terminal and other residues
- PDB file: sequence and coordinate data
- PSF file: Every interaction in the simulation (bonds, angles, dihedrals, etc.)

Running psfgen

- Psfgen is typically run in batch mode:

```
psfgen < mkmol.inp >& mkmol.log
```

- When running within VMD, psfgen commands can be freely intermingled with VMD commands:

```
vmd> set badwat [atomselect top "water
and within 2.4 of protein"]
vmd> foreach segid [$badwat get segid]
resid [$badwat get resid] {
?  delatom $segid $resid]
}
```

Tcl, VMD and psfgen

- Tcl is a full-featured scripting language, and psfgen extends Tcl with structure-building commands.
- Running psfgen from within VMD gives you access to VMD's powerful atom selection capabilities.
- You can write Tcl scripts that generate lipid bilayers or automatically solvate proteins.

BPTI walkthrough

- Builds BPTI from Protein Data Bank files.
- Illustrates multiple segments and patches.
- Split the PDB file 6PTI.pdb into two pieces, one for each segment:

```
grep -v '^HETATM' 6PTI.pdb >
  6PTI_protein.pdb
grep 'HOH' 6PTI.pdb > 6PTI_water.pdb
```

Topology files

- Input:

```
topology toppar/top_all22_prot.inp
```

- Output:

```
reading topology file toppar/top_all22_prot.inp

>>>>>>>>CHARMM22 All-Hydrogen Topology File for Proteins
   <<<<<
>>>>>>>>>>>>>>>>>>>>>>>> July 1997
   <<<<<<<<<<<<<<<<<<<<<<<<<<<<
>>>>>>> Direct comments to Alexander D. MacKerell Jr.
   <<<<<<<<<
>>>>>> 410-706-7442 or email: alex,mmiris.ab.umd.edu
   <<<<<<<<<

Created by CHARMM version 22 1
```

Reading sequences

- For structure-building purposes, the PDB file is just a source of sequence information:

```
segment BPTI {
  pdb output/6PTI_protein.pdb
}
reading residues from pdb file output/6PTI_protein.pdb
extracted 57 residues from pdb file
generating structure at end of segment
no residue 1 before ARG:1 of segment BPTI
add improper failed in residue ARG:1
no residue 1 past GLY:57 of segment BPTI
add bond C(0) N(1) failed in residue GLY:57
```

Applying patches

- Create three disulfide bridges using the patch residue (PRES) defined in the topology file:

```
patch DISU BPTI:5 BPTI:55
patch DISU BPTI:14 BPTI:38
patch DISU BPTI:30 BPTI:51
```

- Output:

```
applying patch DISU to 2 residues
applying patch DISU to 2 residues
applying patch DISU to 2 residues
```

Reading coordinates

- Read the PDB file again to get coordinates. Names in the PDB file don't always match names in the topology file, so we have to *alias* them:

```
alias atom ILE CD1 CD
alias atom LEU CD1 CD2
alias atom LEU CD2 CD1
coordpdb output/6PTI_protein.pdb BPTI
```

- Output:

```
aliasing residue ILE atom CD1 to CD
aliasing residue LEU atom CD1 to CD2
aliasing residue LEU atom CD2 to CD1
reading coordinates from pdb file
  output/6PTI_protein.pdb for segment BPTI
```

More on using aliases

- Alternative to editing the input files.
  - Only affects reading from input files.
  - Output names match topology files.
- Aliasing residue names (for sequence):

  alias residue HIS HSD
- Aliasing atom names (for coordinates):

  alias atom ILE CD1 CD

A segment of water

- Build the water ‘segment’ from the PDB file:

```
alias residue HOH TIP3
segment SOLV {
  auto none
  pdb output/6PTI_water.pdb
}
```

- Output:

```
aliasing residue HOH to TIP3
building segment SOLV
disabling angle autogeneration
disabling dihedral autogeneration
reading residues from pdb file output/6PTI_water.pdb
extracted 73 residues from pdb file
generating structure at end of segment
```

Water coordinates

- Input:

```
alias atom HOH O OH2
coordpdb output/6PTI_water.pdb SOLV
```

- Output:

```
aliasing residue HOH atom O to OH2
reading coordinates from pdb file
  output/6PTI_water.pdb for segment
  SOLV
```

Guessing coordinates

- Psfgen can build missing atoms provided:
  - IC records are available; and
  - Enough atom coordinates have been specified.
- Input:

```
guesscoord
```

- Output:

```
guessing coordinates based on topology file
Warning: guessing coordinates for 583 atoms
Warning: poorly guessed coordinates for 151
  atoms
```

Writing output

- Input:

```
writepsf output/bpti.psf
writepdb output/bpti.pdb
```

- Output:

```
writing psf file output/bpti.psf
total of 1101 atoms
total of 1115 bonds
total of 1681 angles
total of 2366 dihedrals
total of 139 impropers
writing pdb file output/bpti.pdb
```

Minimizing guesses

- If you see more guessed coordinates than expected, residue or atom aliases might be indicated.
- Atomic coordinates should always be minimized before starting a simulation:

  "minimization on"
- For initial minimization, one can minimize only atoms with guessed coordinates:

  "fixedAtoms on"

Checking results

- Minimize guessed atoms:
  - Large motions indicate bad guesses.
  - May indicate indicate switched atom names.
- Minimize entire system:
  - Look for strange conformations.
  - May indicate errors in topology file.
- Bad IC records for can often be removed.

Correcting atom names

- If errors occur when reading coordinates:
  - Look at source pdb in VMD w/o psf file.
  - Compare guessed structure to topology file.
  - Alias atom names to match.
- Reversed atom names will slip through:
  - Look for strange guessed coordinates.
  - Use two atom aliases to reverse this.

What is the future of psfgen?

- Better integration with Tcl.
- Automated methods for solvation.
- Incorporation into NAMD front end.
- Incorporation into VMD.

Understanding IC records

```
IC -C   CA   *N   HN    1.3551 126.4900  180.0000 115.4200  0.9996
IC -C   N    CA   C     1.3551 126.4900  180.0000 114.4400  1.5390
```

- Four atom names (A,B,C,D):
  - optional -/+/# for prev, next, next-of-next
  - number of residue (1,2,3,4) for patches
  - * on third indicates improper version
- Five numbers:

  d(AB), ang(ABC), dihe(ABCD), ang(BCD), d(CD), or
  d(AC), ang(ACB), impr(ABCD), ang(BCD), d(CD)