Filip Jagodzinski
Overview of Errors in the PDB
October 21, 2008

I describe here several of the "errors" that are contained in the individual files in the PDB. References are given for several research articles that address the errors in the PDB.

1 Annotation Errors

PDB files with annotation errors have erroneous declarations of secondary structures, hydrogen bonds, sulfide bridges, etc. Although such errors are not necessarily "crucial", in that a scientists who is only concerned with the coordinates of the atoms will be undeterred by the annotation errors, they are errors nonetheless.

For example, the following fragment, from PDB file 1AMR:

```
HELIX   13  HM PHE A  352  LYS A  355  1
TURN    16 T16 ILE A  353  GLN A  356  TYPE I
```

declares that atoms 353 through 355 are part of an alpha helix, but that likewise atoms 353 through 356 are part of a turn; an obvious overlap.

As found by the Leibniz Institute for Age Research, there are approximately 2,000 such PDB files that have annotation errors, including 1ARC, 1ACF, 1AKA, 1AY5, and 1BRS. Source:
http://www.fli-leibniz.de/ rhuehne/jmol/analyze_sec_struct-2008_02_26b-overlap.txt

Possible Student Projects The level of difficulty for this project is easy. It simply involves parsing the header of a PDB file to determine where overlaps exist. If a single atom number is included in more than one secondary structure, then an errors exists.

2 Residue out of Sequence Errors

Scientists who submit data files for inclusion in the PDB often-times "clean up" the PDB file prior to submission. This is done to remedy any spurious errors that any proprietary fitting program might have introduced into the PDB. Unfortunately, in doing so, often-times additional errors are introduced. For example, the residues numbers in the PDB should be sequential, and thus the following snipped is incorrect, because residue 5 follows residue 1, and residue 3 follows residue 5:

```
...
ATOM      7  ND1 HIS A   1      49.636  26.144   7.860  1.00 16.00           N
ATOM      8  CD2 HIS A   1      51.797  26.043   7.286  1.00 16.00           C
ATOM      9  CE1 HIS A   1      49.691  26.152   6.454  1.00 17.00           C
ATOM     10  NE2 HIS A   1      51.046  26.090   6.098  1.00 17.00           N
ATOM     11  N   SER A   5      49.788  27.850  10.784  1.00 16.00           N
ATOM     12  CA  SER A   5      49.138  29.147  10.620  1.00 15.00           C
ATOM     13  C   SER A   5      47.713  29.006  10.110  1.00 15.00           C
ATOM     14  O   SER A   5      46.740  29.251  10.864  1.00 15.00           O
ATOM     15  CB  SER A   5      49.875  29.930   9.569  1.00 16.00           C
ATOM     16  OG  SER A   5      49.145  31.057   9.176  1.00 19.00           O
ATOM     17  N   GLN A   3      47.620  28.367   8.973  1.00 15.00           N
ATOM     18  CA  GLN A   3      46.287  28.193   8.308  1.00 14.00           C
...
```

The above snipped is for the original submission 1GCN, which has since been fixed, but surely other such errors in the PDB exist.

Possible Student Projects The level of difficulty for this project is easy. It merely involves the parsing of the PDB file to determine if the amino acid residue number numbering sequence is out of order.

3 Missing residues

Often-times, there are residues missing from the PDB data file. This occurs most frequently near the N and C-termini, where there is a unproportionally large amount of motion, and hence those regions exhibit a high B-factor, and so many programs do not include such atoms in their output because to indicate a coordinate for such atoms would be a mere guess, at best. For example, PDB file 1A33 begins with the following:

```
ATOM      1  N   LYS A   4      56.486  36.933  31.281  1.00 48.60           N
ATOM      2  CA  LYS A   4      57.521  37.793  30.641  1.00 50.48           C
ATOM      3  C   LYS A   4      56.855  38.698  29.594  1.00 51.34           C
ATOM      4  O   LYS A   4      56.297  39.760  29.915  1.00 51.34           O
ATOM      5  CB  LYS A   4      58.600  36.900  29.989  1.00 48.28           C
ATOM      6  N   ASP A   5      56.923  38.257  28.339  1.00 52.29           N
ATOM      7  CA  ASP A   5      56.324  38.958  27.210  1.00 49.77           C
ATOM      8  C   ASP A   5      55.119  38.105  26.808  1.00 45.76           C
ATOM      9  O   ASP A   5      54.901  37.823  25.629  1.00 43.73           O
ATOM     10  CB  ASP A   5      57.320  39.028  26.045  1.00 54.77           C
```

Note that missing residues do not necessarily have to occur at the front or rear portions of the amino acid sequence; often-times, missing residues are interior residues.

Note that the sequence of the residue chain IS most often known and indicated in the header of the PDB file, as for example for 1A33:

```
SEQRES   1 A  177  MET SER LYS LYS ASP ARG ARG ARG VAL PHE LEU ASP VAL
SEQRES   2 A  177  THR ILE ASP GLY ASN LEU ALA GLY ARG ILE VAL MET GLU
SEQRES   3 A  177  LEU TYR ASN ASP ILE ALA PRO ARG THR CYS ASN ASN PHE
SEQRES   4 A  177  LEU MET LEU CYS THR GLY MET ALA GLY THR GLY LYS ILE
SEQRES   5 A  177  SER GLY LYS PRO LEU HIS TYR LYS GLY SER THR PHE HIS
SEQRES   6 A  177  ARG VAL ILE LYS ASN PHE MET ILE GLN GLY GLY ASP PHE
SEQRES   7 A  177  THR LYS GLY ASP GLY THR GLY GLY GLU SER ILE TYR GLY
SEQRES   8 A  177  GLY MET PHE ASP ASP GLU GLU PHE VAL MET LYS HIS ASP
SEQRES   9 A  177  GLU PRO PHE VAL VAL SER MET ALA ASN LYS GLY PRO ASN
SEQRES  10 A  177  THR ASN GLY SER GLN PHE PHE ILE THR THR THR PRO ALA
SEQRES  11 A  177  PRO HIS LEU ASN ASN ILE HIS VAL VAL PHE GLY LYS VAL
SEQRES  12 A  177  VAL SER GLY GLN GLU VAL VAL THR LYS ILE GLU TYR LEU
SEQRES  13 A  177  LYS THR ASN SER LYS ASN ARG PRO LEU ALA ASP VAL VAL
SEQRES  14 A  177  ILE LEU ASN CYS GLY GLU LEU VAL
```

4 Missing atoms

Side-chain atoms are often missing. This occurs quite frequently. For example, the protein 1A2J is missing all of the side-chain atoms for the first residue, Alanine, which has a chemical formula HO2CCH(NH2)CH3; only the backbone atoms are provided:

```
ATOM      1  N   ALA A   1      24.782   1.767  29.597  1.00 50.96           N
ATOM      2  CA  ALA A   1      25.511   0.468  29.633  1.00 50.43           C
ATOM      3  C   ALA A   1      25.115  -0.295  30.894  1.00 48.83           C
ATOM      4  O   ALA A   1      24.376   0.229  31.731  1.00 49.41           O
ATOM      5  CB  ALA A   1      27.019   0.711  29.608  1.00 47.80           C
ATOM      6  N   GLN A   2      25.588  -1.534  31.008  1.00 45.37           N
```

Possible Student Projects The level of difficulty for this project is medium or difficult, depending on the actual project that is chosen. First, the PDB file may be parsed, and each residue can be checked to make sure that all of the necessary side chain atoms exist. This requires a knowledge of the 20 amino acids and the nomenclature that is used to describe the relative positions of the atoms. Second (and which is the harder project), a student may want to re-insert into the PDB file those missing atoms so that the final PDB file contains all of the atoms located at the proper location (i.e., the “missing” atoms cannot just be placed at the end of the PDB file). Of course, the truly difficult problem lies in determining the PROPER, i.e., non-steric, location of the missing atoms. This is a research topic, and so there are many viable approaches.

5 References

1. John Westbrook, Zukang Feng, Shri Jain, T. N. Bhat, Narmada Thanki, Veerasamy Ravichandran, Gary L. Gilliland, Wolfgang Bluhm, Helge Weissig, Douglas S. Greer, Philip E. Bourne and Helen M. Berman, “The Protein Data Bank: unifying the archive”, *Nucleic Acids Research*, 2002, Vol. 30, No. 1, 245-248

2. T. N. Bhat, Philip Bourne, Zukang Feng, Gary Gilliland, Shri Jain, Veerasamy Ravichandran, Bohdan Schneider, Kata Schneider, Narmada Thanki, Helge Weissig, John Westbrook and Helen M. Berman, “The PDB data uniformity project”, *Nucleic Acids Research*, 2001, Vol. 29, No. 1, 214-218.

3. A Perrakis, R Morris, VS Lamzin, “Automated protein model building combined with iterative structure refinement“ *Nature Structural Biology*, 1999, Vol. 6 No. 5, 458-463.

4. Miller, R., Gallo, S.M., DeTitta, G.T., Khalak, H.G. and Weeks, C.M. SnB, “Crystal Structure Determination Via Shake-and-bake”. *Journal of Applied Crystallography*, 1994, Vol. 27, 613-621.