Symmetry-aware placement of hydrogens in molecules: Reduce and cctbx*

Jack Snoeyink† Auston Sterling† Vishal Verma†

1 Introduction

We describe, from a geometric algorithms perspective, the role of symmetry in molecular structure determination by crystallography, and how that affects algorithms for structure validation. We focus on our work with the Richardson Lab at Duke University on Reduce, part of the Molprobity suite of validation tools [2]

The Protein Data Bank (PDB) [1] records the atomic coordinate of model molecular structures that are published in key journals or funded by agencies like the US' NIH. As of 10 April 2014, there are 99,293 PDB files, almost 90% of which are determined by X-ray crystallography, a process that makes a pure crystal of a protein sample, then destroys it with X-rays to observe refractions off electron shells (structure factors), which give data about atom positions. I suppress details, including the *phase problem* [5], to point out the role of symmetry: only by having many copies of an atom, packed with known symmetry, can the weak signal from X-ray refractions be amplified to obtain coordinates. Hydrogen atoms, despite making up nearly half of the atoms of an organic molecule, are not given coordinates because refractions from their single electron (often delocalized) usually cannot be resolved.

Molprobity validates model structures by comparing sets of interatomic distances and angles to statistics of a hand-curated set of 8,000 high resolution protein chains. Before it can do so, Reduce, originally written by Mike Word [6], must place hydrogens. Many can be placed by geometric rules, but some must be optimized (MET rotations, ASN and GLN flips), and the optimization function used is not pairwise. We showed in 2005–07 that the interaction (hyper)graph typically has small treewidth, so these problems are solved effieciently by dynamic programming [4].

Because structures actually interact with symmetric copies in the crystal, it is no longer enough to analyze isolated model structures for validation, and especially for determining statistics. In this talk we present the key concepts (model structure, unit cell, asymmetric unit) of working with the crystallographic symmetry groups to extend Reduce, some lessons learned, and some open problems.

One lesson is to begin with a good crystallographic library, even if you will later implement your own. Of the 230 possible crystallographic space groups, 91 appear in the Protein Data Bank, with 21 accounting for 90% of all structures with recorded symmetry, and 53 accounting for 99%. Thus, we used the open-source Computational Crystallography Toolbox (cctbx), a component of the PHENIX crystallography suite [3].

References

[1] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112(3):535–542, May 1977.

[2] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, 35(Web Server issue):W375–383, Jul 2007.

[3] Ralf W. Grosse-Kunstleve, Nicholas K. Sauter, Nigel W. Moriarty, and Paul D. Adams. The *Computational Crystallography Toolbox*: crystallographic algorithms in a reusable software framework. *Journal of Applied Crystallography*, 35(1):126–136, Feb 2002.

[4] Andrew Leaver-Fay, Yuanxin Liu, Jack Snoeyink, and Xueyi Wang. Faster placement of hydrogens in protein structures by dynamic programming. *ACM Journal of Experimental Algorithmics*, 12, 2007.

[5] Gale Rhodes. Crystallography made crystal clear. Elsevier, 2002.

[6] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285(4):1735–1747, Jan 1999.

*Research supported by NIH Molprobity grant and NSF

†Dept. of Computer Science, UNC Chapel Hill