Oral Contributions

[MS4-03] Representing φ/ψ-Dependency of a Variable as a Continuous Function Dale E. Tronrud, P. Andrew Karplus

Department of Biochemistry and Biophysics, Oregon State University, Corvallis OR, 97331, USA.
E-mail: tronrudd@science.oregonstate.edu

In work to develop conformation dependant libraries (CDL) [1] for peptide geometry it is desirable to describe the variation of target values for main chain bond angles as a continuous function of φ and ψ. In our previous work [2, 3] the Ramachandran plot was broken into 10°x10° tiles and the bond angle values for real structures assigned to the appropriate tiles.
The variation in average values across the tiles was smoothed with kernel density averaging to generate a table of target values for that bond angle. In protein structures the conformations of residues are not uniformly distributed, so some tiles may contain a great many samples while others contain none. The "unpopular" tiles must be treated as special cases and the transition between the populated regions and the unpopulated creates difficulties. In addition, any target value for a bond angle that changes quickly relative to the 10° spacing of the tiles will become a "stepped pyramid" and not a smooth hillside. Specifically, with sparse datasets, such as occur in our current work to develop CDLs for *cis*-peptide units, the number of sample points in a residue category range from 796 to 6 and their density in φ/ψ space also varies. We believed that the level of detail of the CDL should be contingent on the nature of the sample sets, but optimizing the predictive power of the library by changing the size of the tiles, the kernel density averaging parameter κ, and the significance cutoff proved impossible due to the high variability of the target function caused by the discontinuities of the form of the library.
We have explored the potential to represent the variation of geometric target values as Fourier summations. The detail of the CDL can be limited by the number of terms in the summation, i.e. the highest index allowed for an amplitude. A function with a larger limiting index will contain greater detail than one with a smaller limit. Since the structures upon which the CDL is based do not cover Ramachandran space uniformly the Fourier coefficients cannot be calculated by a simple FFT. The estimation of the Fourier coefficients was stabilized by a restraint which minimized their amplitudes and biased the result toward lower resolution components. The actual resolution limit was determined on a case-by-case basis using complete cross-validation. It was found that remarkably few Fourier coefficients were required to fit the observed distributions.

The resulting CDLs have the expected properties. The optimal predictive fits for residue classes with a small number of sample points scattered widely in conformation space have very small maximal indices, as small as zero. CDLs with more densely sampled regions, even when most of conformational space is unsampled, contain greater detail. Surprisingly, for the sequence category with the largest number of sample points, 796, a maximal index of only 7 results in the optimal complete crossvalidation. This CDL is both continuous and simpler, being described by only 225 parameters where the original "tile" parameterization required 36x36 or 1296 parameters.

[1] Berkholz, D.S., Shapovalov, M.V., Dunbrack, R.L. Jr & Karplus, P.A. (2009). *Structure*, **17**, 1316-1325.
[2] Tronrud, D.E., Berkholz, D.S. & Karplus, P.A. (2010). *Acta Cryst.* **D66**, 834-842.
[3] Tronrud, D.E., & Karplus, P.A. (2011). *Acta Cryst.* **D67**, 699-706.

Keywords: conformation dependant library; cross-validation; Fourier transform