Table 18.3.2.1 (continued)

Phenylalanine, 1076/1616, $C_6H_5CO—NH—CH_2—C\_phenyl$

| Phenyalanine | 1076/1616 | $C_6H_5CO—NH—CH_2—C\_phenyl$
|--------------|-----------|--------------------------------------------------|

Proline, 262/255, $trans—C—CO—pyrrolidine—CO—N$

| Proline | 262/255 | $trans—C—CO—pyrrolidine—CO—N$
|---------|---------|--------------------------------------------------|

Serine, 33/39, $NH—CH(CO)—CH_2—OH$

| Serine | 33/39 | $NH—CH(CO)—CH_2—OH$
|--------|------|--------------------------------------------------|

Threonine, 20/25, $NH—CH(CO)(CH_2)OH—CH_3$

| Threonine | 20/25 | $NH—CH(CO)(CH_2)OH—CH_3$
|-----------|-------|--------------------------------------------------|

Tryptophan, 123/135, $CH_2—indole$

| Tryptophan | 123/135 | $CH_2—indole$
|------------|--------|--------------------------------------------------|

Tyrosine, 124/161, $para—(—C—CH_3)—phenol$

| Tyrosine | 124/161 | $para—(—C—CH_3)—phenol$
|----------|--------|--------------------------------------------------|

Valine, 198/313, $N—CH(CO)—CH—(CH_3)$

| Valine | 198/313 | $N—CH(CO)—CH—(CH_3)$
|--------|--------|--------------------------------------------------|

---

18.3. STRUCTURE QUALITY AND TARGET PARAMETERS

18.3.2.2. Treatment of outliers

A truly Gaussian distribution should include outliers at high $\sigma$ values (about 0.01% for 4$\sigma$). We should expect, however, that the width of the distribution is affected not only by inherent variation in the variables to be parameterized, but also by variability in the experimental conditions (e.g. resolution) and by erroneous structures. This weakens a strategy of automatic rejection of outliers beyond a specific cutoff value. The possibility of visualizing the distributions with CSD software allows refinement of this rejection strategy, with, however, the introduction of considerable subjectivity in the criteria. For this work, a 4$\sigma$ cutoff was generally considered a flag for erroneous outliers. However, broad and flat tails in the distribution were relatively frequent and often asymmetric. These deviations from Gaussian behaviour ‘artificially’ increased $\sigma$ values. In these cases, the 4$\sigma$ cutoff rule was not applied automatically, but was applied after examination and rejection of conspicuous outliers. From an algorithmic viewpoint, this was the additional use of skew and kurtosis (third and fourth moments of the distribution) for rejection criteria. In most cases, uncertainty in rejection criteria affected the average values little, but could significantly alter standard deviations.

18.3.2.3. Bonds and angles

18.3.2.3.1. Peptide parameters: proline, glycine, alanine and CB substitution

Fragments representing five-atom lengths of the backbone currently provide adequate statistics for peptide compositions of varieties including glycine, proline and side chains branched at CB. Peptide cyclicity was generally allowed on the assumption that this does not introduce distortions greater than typical protein secondary-structure interactions. The results are presented in Table 18.3.2.3. With one exception, none of the values deviates from those of 1991 by more than one sample standard deviation. However, the very large $\sigma$ values for the proline $C—N—CA$ and $C—N—CD$ angles (Table 18.3.2.1) are conspicuous. Using high-resolution protein structures, Lamzin et al. (1995) identified geometries of proline that were inconsistent with high-resolution protein structures and also noted inconsistencies in $C—CA—CB$ angle parameters (see also the sections on individual amino acids below). In the case of proline, a bimodal distribution of these parameters could be resolved with the discrimination between cis and trans forms (Fig. 18.3.2.1). A scatter plot of the angles against $\omega$ torsion angle resolves the averages (and $\sigma$’s) of 122.6 (50) and 125.4 (44$\sigma$) for $C—N—CA$ and $C—N—CD$, respectively, into cis- and trans-dependent values with much smaller sample deviations (see Table 18.3.2.2). The large $\sigma$ value for CB—CG remains, however, particularly for trans-proline. Its origin is unknown, but proline pucker may play a role.

Glycine, with its unique CH$_3$ as CA, required new atom-type definitions for Engh & Huber (EH) (1991) parameterization to account for parameter-average differences of about one-half of a sample standard deviation. These also included $C—N—CA$ for which the average angles were 120.6$^\circ$ for glycine and 121.7$^\circ$ for the rest. The new statistics with 83 $C—CO—NH—CH_2—C$ fragments estimate a larger value of 122.3$^\circ$ for the glycine $C—N—CA$ angle.