Supplementary Material

Which intermolecular interactions have a significant influence on crystal packing?

Robin Taylor, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK. Correspondence e-mail: robin@justmagnolia.co.uk

Validation of the R_F confidence intervals

Consider a structure containing N base atoms of type T_B, of which n form their primary interactions to atoms of type T_A. Under the null hypothesis, the probability of this result or one more extreme (*i.e.* >n) is given by the binomial formula:

probability = $\Sigma\ P^i Q^{N-I} N! / i!\ (N-i)!$

where $P = S(T_A)/S(total)$, $Q = 1-P$, and the summation is from i=n to i=N. The smaller the probability, the more likely it is that the null hypothesis is false and the alternative hypothesis is true. A separate binomial probability can be obtained from every structure in the data set containing atoms of types T_B and T_A. The question then arises how they may be combined into a single, overall test of significance. This would be straightforward if the probabilities were continuously distributed.[1,2] However, binomial probabilities are discontinuous and there is no analytical method for determining the expected distribution of the probabilities if the null hypothesis is true.[3] Fortunately, it is easy to estimate by simulation. The procedure used was as follows. For each structure from which a binomial probability was obtained, the primary interactions in 10 "pseudo-structures" were generated (10 was chosen arbitrarily but is large enough to produce well-converged results). In a given pseudo-structure, the atom to which each base atom formed its primary interaction was chosen by a random number generator, with the probability of choosing an atom of any given type T_X equal to $S(T_X)/S(total)$. The distribution of binomial probabilities calculated from the pseudo-structures was therefore the distribution expected under the null hypothesis.

The statistic used to determine whether the distribution of binomial probabilities from the real structures (the "true" probabilities) differed significantly from the simulated null distribution was:

Δ_{mean} = mean(simulated) – mean(true)

where mean(simulated) and mean(true) are the means of the simulated and true binomial probabilities, respectively. A positive (negative) value of Δ_{mean} indicates that the true probabilities tend to be smaller (larger) than those in the simulated null distribution, implying

that the interaction occurs more (less) often than would be expected by chance. The standard error of Δ_{mean} can be calculated in the usual way:

$$SE(\Delta_{mean}) = \sqrt{[\sigma^2(simulated)/n(simulated) + \sigma^2(true)/n(true)]}$$

where σ^2(simulated) and σ^2(true) are, respectively, the sample variances of the simulated and true binomial probabilities, and n(simulated) and n(true) are the sample sizes. Since the sample sizes in this study are almost always large, Δ_{mean} can be assumed normally distributed, allowing us to conclude that a Δ_{mean} is significantly different from zero at the 95% confidence level if:

$$|\Delta mean| > 1.96.SE(\Delta_{mean}).$$

The reliability of the bootstrapped 95% confidence intervals of R_F may therefore be checked by comparing the interactions whose confidence interval does not span the value 1 (implying R_F significantly different from 1 at the 95% confidence level) with those whose Δ_{mean} satisfies the above inequality (implying Δ_{mean} significantly different from 0 at the same confidence level). A high degree of concordance will suggest that the bootstrapped confidence intervals are reliable.

The degree of concordance was established as follows. The coarse atom-typing scheme (Table 1 of main paper) was used to assign atom types to both base atoms and the atoms to which they form their primary interactions. Each combination of atom types was considered, giving a total of $17^2 = 289$ different types of interaction. Of these, 24 were eliminated because there were fewer than 10 structures in which they could possibly occur. Δ_{mean}, $SE(\Delta_{mean})$, and R_F and its 95% confidence interval were computed for the remaining interactions using the Bondi radii. The results from the two statistical methods were compared. There was disagreement on two of the interactions, F…F and Br…Br. The Δ_{mean} statistic indicated with >95% confidence that these interactions occur less often than expected under the null hypothesis (Δ_{mean} = -0.020(4) and -0.009(3), respectively), while R_F suggested at the same confidence level that they occur more often than expected (R_F = 1.03 and 1.05, respectively, with the lower boundary of the confidence intervals > 1). There were a further 9 interactions for which one of the methods gave a significant result (*i.e.* rejected the null hypothesis) and the other did not. However, the methods agreed on the remaining 254 interactions. Of these, there were 25 for which the null hypothesis could not be rejected at the 95% confidence level, *i.e.* they occur about as often as would be expected by chance. Some 151 occur significantly less often than would be expected by chance, and the other 78 significantly more often.

The statistical methods therefore showed excellent agreement, which strongly suggests that the confidence intervals derived for R_F by bootstrapping are reliable.

References

1. R. A. Fisher, *Statistical Methods for Research Workers*, 4th ed., Oliver & Boyd, Edinburgh, 1932.
2. M. C. Whitlock, *J. Evolutionary Biol*., 2005, **18**, 1368-1373.
3. W. A. Wallis, *Econometrica*, 1942, **10**, 229-248.