Investigation of the Similarity of Allergens within the FARRP5 Dataset: Implications for the Current FAO/WHO Similarity Threshold

Robert Cressman, Mark Whitsitt, and Greg Ladic. Dupont/Pioneer Crop Genetics Regulatory Science Department, Dupont Experimental Station, Wilmington, DE.

Abstract

The current amino acid identity threshold for similarity of a transgenic protein product to known allergenic proteins is a FASTA identity of 35% over an 80-residue window using the (FAO/WHO 2001, CODEX 2003). We desired to investigate the residue identity between the allergens comprising the FARRP5 dataset (www.allergenonline.com) and compare the similarity thresholds observed to those currently employed. All manipulations were performed using the Biofacet software package (Gene-IT, Worcester, MA). In order to reduce the number of highly similar sequences, which could skew results in subsequent analyses, 627 of the proteins were clustered into 175 groups and the longest member of each cluster retained as representative of that cluster. These representative sequences and the remaining un-clustered sequences were combined into a new dataset containing 739 proteins. This dataset was then subjected to single linkage cluster analysis using pair-wise alignments of these sequences with varying match criteria from 35% to 50% identity over a minimum alignment length of 80 amino acids in 5% intervals (e.g. 35% to 40%, 40% to 45%, etc.). Multiple sequence alignment of the clusters generated showed that they were biologically relevant and representative of several large families of known allergenic proteins such as the Bet v 1 pollen allergens, Trichicum gliadins/glutenins, and the Lol pollen allergen family. Because of the disparate nature of its members, the Bet v 1 cluster, which included the allergens Mal d1, Api g 1, Pru AV 1, Cor a 1, and Dau c 1, was chosen for further study. Within this family, Dau c 1 and Api g 1, while known to cross-react with Bet v 1, were among the least similar to Bet v 1 (~38% and 41% identical, respectively, across the best local alignment greater than or equal to 80 residues). As clustering stringency increased to 50% or greater, the 6 Bet v 1 cluster members most dissimilar to Bet v 1, including, Dau c 1 and Api g 1, were excluded from the cluster. This suggests that 50% identity across the best local alignment greater than or equal to 80 residues may be an upper bound for assessing cross-reactivity risk using sequence analysis. The ~40% identity between Bet v 1 and Dau c 1, for example, may be a reasonable lower bound to this criterion. In addition, when the Bet v 1 cluster was used in the standard 35% identity over 80 amino acid fragment assay, neither Dau c 1 nor Api g 1 were detected as potential matches, suggesting that limiting the window of similarity may negatively affect sensitivity. While these results provide tentative upper and lower bounds to similarity criteria that may be used to assess cross-reactivity risk, they can currently only be applied to the Bet v 1 family of allergens; these analyses must be extended to other allergen protein family clusters before the results can be generalized.