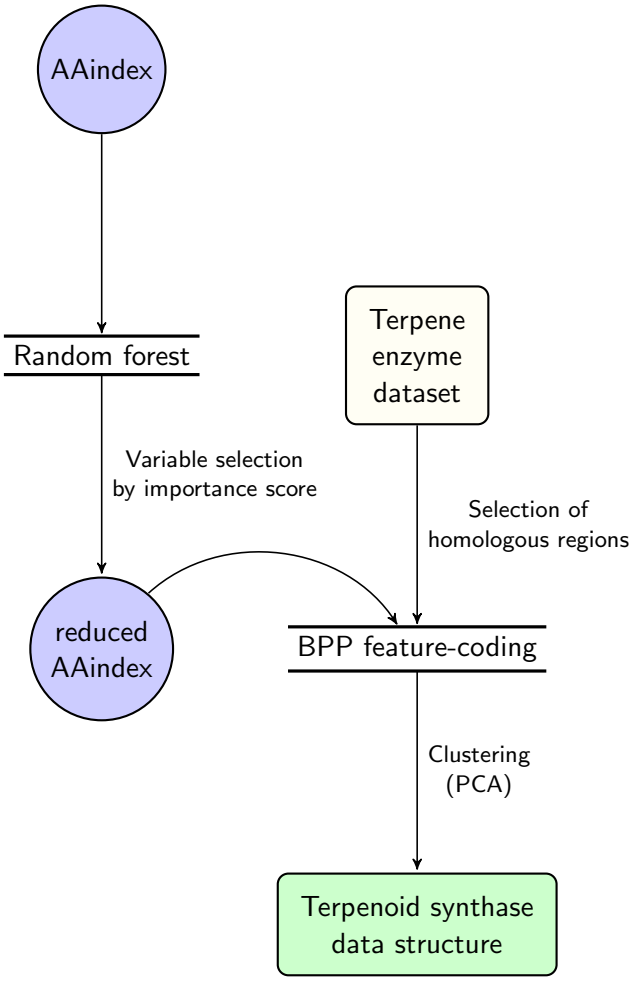

Figure 1: Materials and methods: depicts the procedure carried out on this report. The starting point is the amino acid index database which is reduced to a small subset based on variable importance scores derived from random forest algorithm. The reduced set (rAAindex) is used for to encode biochemical and physical properties into protein sequences for examination of the data structure of the sub-families terpene synthases.