Supplementary data :
Study of groups of variants of nonribosomal peptides

As mentioned in the paper, no universal definition of NRP variants has been proposed so far. Some peptides sharing a common name show similar monomeric composition, while others have very diverse ones. To limit the bias in our analyses, we decided to select only one representative variant for groups composed of similar peptides. The definition of those groups is based on a distance between peptides.

Peptides are encoded by non-oriented labeled graphs. The distance d between two graphs (representing two peptides) $G1$ and $G2$ is computed as follows :

$$d(G1, G2) = 1 - \frac{|mcs(G1, G2)|}{Max(|G1|, |G2|)} \cdot \delta(G1[V_{mcs}], G2[V_{mcs}])$$

, where

- $|mcs(G1, G2)|$ is the number of nodes in the maximal common subgraph (MCS), i.e. the number of common monomers in the common substructure between the two peptides.
- $|Gi|$ is the number of nodes in a graph Gi, i.e. the number of monomers in peptide i encoded by a graph Gi
- $Gi[V_{mcs}]$ is the subgraph of Gi induced by V_{mcs}. $G1[V_{mcs}]$ and $G2[V_{mcs}]$ have the same node set (V_{mcs}), they contain the edge set E_{mcs} but each one can contain supplementary edges.
- coefficient δ is the number of different edges between $G1[V_{mcs}]$ and $G2[V_{mcs}]$, divided by the number of possibles edges on V_{mcs}. It is computed as follows :

$$\delta(G1[V_{mcs}], G2[V_{mcs}]) = 1 - \frac{|E(G1[V_{mcs}])| + |E(G2[V_{mcs}])| - 2|E_{mcs}|}{\frac{|V_{mcs}|(|V_{mcs}|-1)}{2}}$$

  where $|E(G)|$ is the number of edges in graph G.

With this formula, two identical peptides have distance 0 and two peptides without common monomers have distance 1. In order to reduce the influence of substitutions between similar monomers, we defined clusters of monomers (corresponding to a simplified amino acid substitution matrix). First, we made a cluster with all the lipids, another with all the glucids, another with all the chromophores and a cluster for each amino acid and its derivatives. Then, we merged together amino acids clusters sharing similar physico-chemical properties, taking inspiration from the small clusters of Rausch *et al* (NAR, 2005). Monomers of the same cluster are considered to be identical if we compute a distance with clustering.

To obtain the average distance in a group of n peptides, we computed the pairwise distance between all the peptides of the group and divided the sum by n. The 1071 NRPs of the study are divided into 183 groups of which 62 groups contain only one peptide. For the 121 groups containing at least two variants, we have computed the average distance with the second level of clustering. The histogram of average distance, ordered by increasing order, for the 80 groups with a value not equal to 0 in Figure 1 shows that the threshold 0.4 can be used for identification of groups with non-similar variants.

Ten groups have an average distance greater than 0.4 : dolastatines, guineamides, hymenamides, kahalalides, kapakahines, phakellistatins, pyoverdins, serrawettins, stylopeptides and peptaibols. For the nine first groups, we keep all the variants in all studies. The last one, the peptaibol group, is not a real group but a big class of nonribosomal antibiotics that share common features. The 130 peptaibols are divided into 20 groups of variants, in the 1071 peptide set. For each of the 20 groups, we keep only one variant.

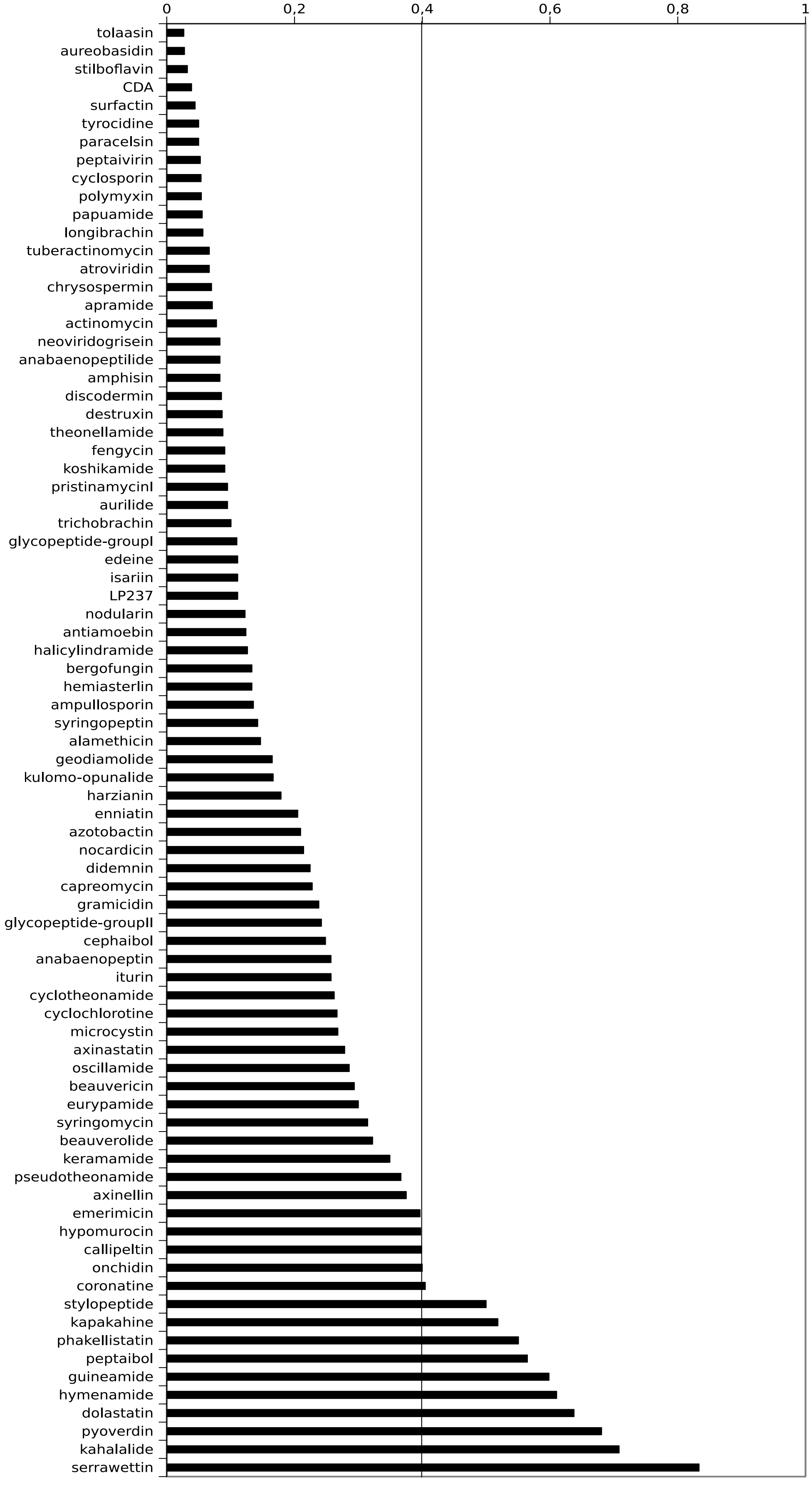

Fig. 1 – Average distance by increasing order in the 80 NRP groups in which there are at least two peptides and the average distance is different of 0.