Supplementary material for Norine: update of the nonribosomal peptide resource

In this supplementary material, we describe the main steps and filters used to verify data integrity and data quality before insertion in the Norine database. If quality criteria are not met, the data is not inserted.

The three databases used as external sources for new nonribosmal peptides (NRPs) are: MIBiG, BIRD and StreptomeDB. These databases tag the entries containing a NRP with the value: NRP in the general_params field (MIBiG); Y in the field _pdbx_reference_entity_sequence.NRP_flag (BIRD); and NRPs in the tag <synthesizing routes> (StreptomeDB). So only entries with these tags are processed by our script. Annotations are extracted from external entries, with correspondences established between external fields and Norine fields.

To verify if an external peptide is already in Norine, we compare its structures (SMILES and monomer graph) to Norine peptide structures. SMILES are canonized to allow quick textual comparison. Monomer graphs are generated by two processes, depending on the peptide source. BIRD (PDB) provides its own monomer graph, with its own monomer names. A connection between PDB and Norine monomers has been established. But, if at least one monomer is not already in Norine, a warning describes the external entry and the missing monomer. The other databases provide SMILES. So, the monomer graph is inferred from the SMILES using both Smiles2Monomers and rBAN. If neither software constructs a monomer graph with all atoms included in monomers, it means that the structure contains new monomers, not already in Norine. A warning describes the external entry and the results of both tools. rBAN suggests the new monomer(s) by querying PubChem. If a correct monomer graph is output, it is searched in Norine. If an identical structure is found, it means that the NRP is already in Norine, so the new annotations are inserted in the Norine entry, at least the link to the external database. If no identical structure is found, the NRP is new and a new entry is created, tagged as "unreviewed". If both structures are available, SMILES and monomer graphs of Norine and external entries are compared. Again, if differences are detected, a warning describes the problem.

In addition to the mining of external databases to find new NRPs, other scripts search for new annotations in other databases. For example, PubChem and ChEMBL are queried with the peptide names in order to find their entries and extract the SMILES. Before inserting the SMILES in the corresponding Norine entries, rBAN is run to verify that the monomer structure inferred is identical to the Norine monomer structure. If not, a warning is given with relevant information.

Each time a warning is produced, the corresponding data are not included in Norine. These warnings alert us for possible errors in Norine and the Norine team analyses them to curate existing Norine entries or to insert new peptides. Careful reading of scientific articles is often needed to determine what is the correct structure and sometimes, we observed errors in the external databases.