Use of Learning Vector Quantization and BCI fingerprints for the Predictive Toxicological Challenge 2000–2001.

Baurin, N *.; Marot, C.; Mozziconacci, J.C.; Morin–Allory, L.
ICOA, UMR 6005, Université d'Orléans, BP 6759, 45067 Orléans Cedex 2, France.

For the submission to the PTC 2000–2001, *i.e* the prediction of carcinogenicity for 185 molecules part of a Test set, we used LVQ models. A Learning Vector Quantization model is built during a learning phase in which the model learns to discriminate the "Positive" from the "Negative" molecules based upon the structural features and carcinogenicity classifications of the molecules part of a Training set. The data needed for this Learning phase were provided during the stage I of the PTC 2000–2001 :
the structural features of the molecules are the BCI fingerprints; each molecules is described by 6150 fingerprints, *i.e* the presence or absence within a molecule of various structural features:

- augmented atoms
- atom pairs with a distance between atoms of any length desired
- atom sequences of any length desired
- ring composition sequences for any size rings desired
- ring fusions ? sequences of ring connectivities

the carcinogenicity classifications are the National Toxicology Program carcinogenicity classifications (a molecule is classified "CE", "SE", "EE", "NE", "IS" , "P", "E" or "N") on 4 sex/species combinations (Male Mouse, Female Mouse, Male Rat, Female Rat).
The algorithm used so that the model learns to discriminate the "Positive" from the "Negative" molecules is a LVQ algorithm.

That way, 4 LVQ models were retained to be sub–optimal models according to the ROC convex hull criteria: **FM–EN**, **MM–EN**, **MR–3class**, **MR–EH**.

Handling of the carcinogenicity data. The heterogeneity of the initial carcinogenicity data for the training set was handled in order to get 3 types of carcinogenic classification ("Positive", "Negative" and "Equivocal"):
if the molecules of the training set were initially marked "CE", "SE" or "P", they were considered as "Positive" molecules.
if the molecules of the training set were initially marked "NE" or "N", they were considered as "Negative" molecules.
if the molecules of the training set were initially marked "EE" or "E", they were considered as "Equivocal" molecules.
if the structures were initially marked "IS", they were excluded from the training set.

Since the predictions should be whether "Positive=Carcinogenic" or "Negative=Non–Carcinogenic", various strategies were used in the various models selected by the PTC/2000–2001:
FM–EN, this model is a 2 classes model for the Female Mouse carcinogenicity, where we have considered the Equivocal molecules as "Negative" before the building of the model (143 "Positive" structures and 228 "Negative" structures were part of the training set).

MM–EN, this model is a 2 classes model for the Male Mouse carcinogenicity, where we have considered the Equivocal molecules as "Negative" before the building of the model (129 "Positive" structures and 248 "Negative" structures were part of the training set).
MR–3class, this model is a 3 classes model for the Male Rat carcinogenicity, where we have considered the Equivocal molecules as "Positive" after the building of the model (152 "Positive" structures, 44 "Equivocal" structures and 192 "Negative" structures were part of the training set). That mean that the structures of the test set that were predicted "Equivocal" by this model were proposed as "Positive=carcinogenic" structures.
MR–EP, this model is a 2 classes model for the Male Rat carcinogenicity where we have considered the Equivocal molecules as "Positive" before the building of the model (196 "Positive" structures and 192 "Negative" structures were part of the training set).

Handling of the descriptors. From the 6150 BCI initial descriptors of the training set, we excluded the non–variant descriptors (*i.e* a descriptor that is constant for all the structures of the training set is not able to explain the structural differences between the "Positive" and "Negative" structures, so it is useless). That way the **FM–EN** model was built using 5456 BCI descriptors, the **MM–EN** model was built using 5467 BCI descriptors , the **MR–3class** model was built using 5905 BCI descriptors and the **MR–EP** model was built using 5905 BCI descriptors. We transformed (centering and autoscaling) the descriptors of the training set descriptor matrix (this operation is usually done and enable to consider all descriptors on the same basis (mean=0, standard deviation=1)) and we transformed in the same way the descriptors of the test set descriptor matrix using the initial means and standard deviations of the training set descriptors.

Building of the model. We used 3 types of LVQ algorithms to build a model: all those algorithms are supervised classification techniques which handle several **codevectors** during the learning phase. Basically, in that case, a codevector can be considered as a virtual molecule; the LVQ strategy used, for example to discriminate "Positive" from "Negative" molecules, is starting creating 400 codevectors (if the molecules studied are described by 5456 descriptors, each of our codevector is made of 5456 values that are initially randomly assigned); 200 of those codevectors will be labelled "Positive" and 200 codevectors will be labelled "Negative". Then, by comparing those labelled codevectors to the molecules of the training set which are either "Positive " or "Negative", the codevectors are slightly modified so that the "Positive" molecules will be more similar to the "Positive" codevectors compared to the "Negative codevectors, and the "Negative" molecules will be more similar to the "Negative" codevectors compared to the "Positive" codevectors. After several iterations during which all the molecules of the training set will be used several times to discriminate the "Positive" codevectors from the "Negative" codevectors, we got a LVQ model, *i.e* 2 populations of labelled codevectors that can be used to make predictions.
Algorithmic Strategy: We first used the OLVQ1 algorithm (3 times to balance) then we used several combinations of LVQ3+LVQ2.1 algorithms till we reach a maximum in the ability of the model to efficiently discriminate the classes within the training set.
Details: we initially used 400 codevectors, 16000 iterations for the OLVQ1 algorithm, 1330 iterations for the LVQ3 algorithm [°(0)=0.03, °=0.1, °=0.3] and 1330 iterations for the LVQ2.1 algorithm [°(0)=0.03, °=0.3]).

Prediction of carcinogenicity for the 185 ligands test set. Using a LVQ model, *i.e* a population of labelled codevectors, we predict the carcinogenicity of a ligand by looking at the most similar codevector in a LVQ model (euclidian distance) : for example, if the ligand X is closest (similarity) to a codevector labelled as "Positive", we will propose that the ligand X is "Positive=Carcinogenic". This method of prediction is a similarity–based method.

Concerning the interpretation of which structural features are responsible of the Carcinogenic or Non–Carcinogenic property of a compound, this is not easily discernible since all the BCI descriptors are used (more than 5000 fragments in all models retained) to discriminate the "Positive" from the "Negative" molecules.

Materials: In–house VC++ routines were used for the preparation of the data. The LVQ_PAK 3.1 (http://www.cis.hut.fi/research/som–research/nnrc–programs.shtml) developed by the Neural Network Research Center, Helsinky University of Technology in Finland, was used for the building of the LVQ models.

* Corresponding author. E–mail: nicolas.baurin@univ–orleans.fr