S1133: The Drug Discovery Pipeline
October 28, 2019 – Exercise sheet

Instructions

- Answer the questions in a PDF file, either handwritten and scanned, or computer-generated.
- You can write in English or in French.
- The file should be named `LastName_FirstName_pipeline.pdf`.
- You are encouraged to work in groups, but submit one file per student.
- Your report is due on **October 29, 2019 at 23:59pm**.
- Please deposit your file at the following URL: https://frama.link/s1133_2019_assignments. You will not receive a confirmation message nor see your deposited file; don't worry, if there's a technical issue, we'll address it and you won't be penalized.

1. **The drug discovery pipeline** Explain in your own words (about half a page, total):
   (a) (1 point) The main paradigm explaining what a drug is and how it works;

   Solution:

   - small molecule that binds to a protein, thereby affecting its activity and therefore a biological process that is key to the disease.
   - key-lock principle.

   (b) (2 points) The steps of the drug discovery pipeline that stem from this principle;

   Solution:

   - Target definition.
   - Hits identification: finding small molecules that bind to the target.
   - Leads characterization: among the hits, find compounds with desirable ADME-Tox properties.
   - Candidates suggestion: pick the most promising leads; optimize leads.
   - Clinical trials.

   (c) (1 point) Some limitations of this principle.

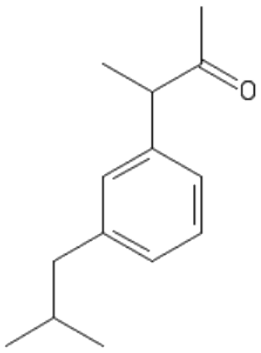
Solution:

- A small molecule may by multiple targets (drug promiscuity).
- Drug interactions are not accounted for.
- Not all drugs follow this paradigm (see e.g. monoclonal antibodies, gene therapy, cell therapy).

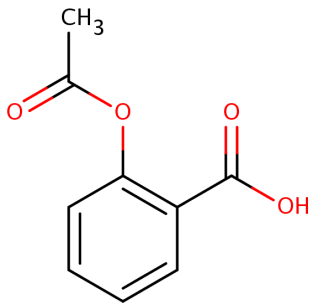
2. **Chemoinformatics**

(a) (1 point) Draw the chemical structure corresponding to the following SMILE string: `CC(C)Cc1cc(C(C)C(C)=O)ccc1`.

Solution:



(b) (1 point) Propose one SMILE string for the following chemical structure:



Solution: `CC(=O)Oc1ccccc1C(=O)O` or `O=C(C)Oc1ccccc1C(=O)O`, for example.

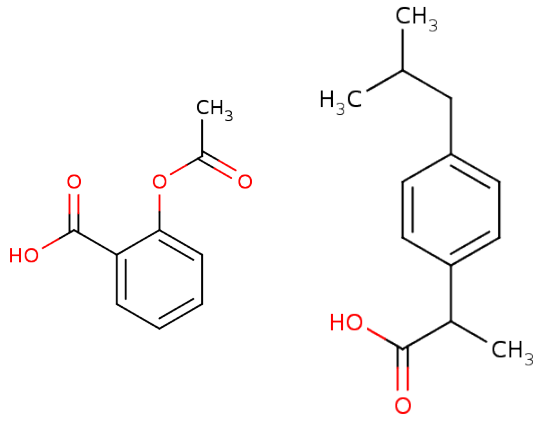
(c) (1 point) Consider two binary fingerprints A and B. Let a be the number of bits set to 1 in A, b the number of bits set to 1 in B, and c the number of bits set to 1 in both A and B. Using only a, b and c, express:

- the dot product between A and B;
- the Hamming distance between A and B;
- the Tanimoto similarity between A and B.

Solution:

- Dot product: $\langle \vec{A}, \vec{B} \rangle = c$
- Hamming/Euclidean distance: $d(\vec{A}, \vec{B}) = \sum_{i=1}^{p} |A_i - B_i| = \sqrt{\sum_{i=1}^{p} (A_i - B_i)^2} = a + b - 2c$
- Tanimoto: $k(\vec{A}, \vec{B}) = \frac{\vec{A}.\vec{B}}{||\vec{A}||^2 + ||\vec{B}||^2 - \vec{A}.\vec{B}} = \frac{c}{a+b-c}$

(d) (2 points) Here are the chemical structures of two molecules. Write out their count path fingerprints, using paths of length equal to 5 (that is to say, containing 5 *nodes*, not 5 *links*). Account for every type of atom (except H) and bound you encounter.



Solution: Paths of 5 nodes present in molecule A, starting from the "left", ignoring `H` but accounting for bound type:

- `OCccc` (x2, once with the "top" of the aromatic ring and once with the "bottom"), can also be written `cccCO`
- `O=Cccc` (x2), can also be written `cccC=O`
- `OCccO`, can also be written `OccCO`
- `O=CccO`, can also be written `OccC=O`
- `Ccccc` (x2), can also be written `ccccC`
- `CccOC`
- `ccccc` (x6, starting with each atom of the ring)
- `ccOCC` (x2, starting fromt he "left" or the "right" of the aromatic ring)
- `ccOC=O` (x2)
- `ccccO` (x2)
- `cccOC` (x2)

Paths of 5 nodes present in molecule B, starting from the "top", ignoring H but accounting for bound type:

- CCCcc (x4, starting from either of the two CH3 and going either "left" or "right" in the aromatic ring)
- CCccc (x6: x2 from the "top", and x4 at the "bottom" of the ring, starting either from the CH3 or from the C connected to two oxygens)
- Ccccc (x4: x2 from the "top", and x2 at the "bottom" of the ring)
- ccccc (x6, starting with each atom of the ring)
- ccCC=O (x2)
- ccCCO (x2)

	molecule A	molecule B
ccccc	6	6
ccccC	2	4
cccc0	2	0
cccCC	0	6
cccCO	2	0
cccOC	2	0
cccC=O	2	0
ccCCC	0	4
ccCCO	0	2
ccCC=O	0	2
ccOCC	2	0
ccOC=O	2	0
cCCCC	0	4
CccOC	1	0
OccCO	1	4
OccC=O	1	4

(e) (2 points) Here are the count fingerprints of three molecules:

molecule A	3	0	0	0	1	1	1	0	2	0	0	1
molecule B	2	1	1	0	1	2	1	0	2	0	1	1
molecule C	3	1	1	0	1	1	1	0	0	0	0	1

- Compute the Tanimoto similarity between A and B, A and C, and B and C.
- Compute the Minmax similarity between A and B, A and C, and B and C.
- Compute the Hamming distance between A and B, A and C, and B and C.
- Which pair of molecule is the closest / the most dissimilar?

Solution: In order to compute Tanimoto or Hamming, we first binarize the fingerprints:

molecule A	1	0	0	0	1	1	1	0	1	0	0	1
molecule B	1	1	1	0	1	1	1	0	1	0	1	1
molecule C	1	1	1	0	1	1	1	0	0	0	0	1

- Tanimoto(A, B) = 6/9 = 2/3 $\approx$ 0.667; Tanimoto(A, C) = 5/8 = 0.625; Tanimoto(B, C) = 7/9 $\approx$ 0.778.
- Minmax(A, B) = (2+1+1+1+2+1) / (3+1+1+1+2+1+2+1+1) = 8/13 $\approx$ 0.615; Minmax(A, C) = (3+1+1+1) / (3+1+1+1+1+1+2+1) = 7/11 $\approx$ 0.636; Minmax(B, C) = (2+1+1+1+1+1+1) / (3+1+1+1+2+1+2+1+1) = 8/13 $\approx$ 0.615.
- Hamming(A, B) = 3; Hamming(A, C) = 3; Hamming(B, C) = 2.
- The pair of closest molecules is: (B, C) according to Tanimoto and Hamming; and (A, C) according to Minmax. The pair of most dissimilar molecules is: (A, C) according to Tanimoto (it was the most similar for Minmax!); (A, B) or (B, C) according to Minmax; and (A, B) or (A, C) according to Hamming.

3. **LogP prediction** The logP coefficient of a molecule is the logarithm of its octanol-water partition coefficient. In other words, if you consider this molecule in a two-phase 1-octanol/water system, and call C_o is its concentration in the 1-octanol phase, and C_w its concentration in the water phase, its logP is given by

$$\log P = \log_{10} \frac{C_o}{C_w}.$$

A high, positive logP indicates a lipophilic molecule, much more soluble in octanol than in water, and conversely for a negative logP.

Let us consider a data set of molecules, described both by their molecular graph and their logP. We want to build a logP predictor.

(a) (1 point) Why is logP prediction interesting in the context of drug discovery? Remember that the cell membrane is composed of a bilipidic layer; its outside is hydrophobic and the inside is hydrophilic.

Solution: The distribution of a candidate molecule (that is, reaching its protein target) can only happen if the molecule is sufficiently hydrophobe to enter the bilipidic layer and hydrophile to exit it.

(b) (1 point) What large class of machine learning algorithms (clustering, dimensionality reduction, semi-supervised learning, etc) can be used to build our predictor? Give examples of such algorithms.

Solution: Any *regression* algorithm: linear regression, SVR, random forest, artificial neural network... You cannot use a classification method because the outcome (logP) is continuous, not categorical. For kernel methods (SVR, kernel ridge regression) one can use the Tanimoto kernel.

(c) (1 point) What are the pros and cons of using representations of molecules derived from their molecular graphs?

Solution:

- Pros: typically known, easy to manipulate, many tools (eg. kernels).
- Cons: less info than 3D, stereochemistry is missing, etc.

4. **Virtual screening**

(a) (2 points) Cite three different computational methods that can be used to predict the binding affinity between a small molecule and a protein, based on different levels of information or modeling. Explain the differences between them. (About half a page, total).

Solution:

- QM/MM: very local scale, uses quantum physics.
- molecular mechanics/dynamics: intermediate scale, uses classical physics (atoms modeled as balls connected by springs)
- molecular docking: global scale, uses a discretized grid and force fields.

Figure 1 shows the performance of several virtual screening methods, applied to the detection of molecules inhibiting the Respiratory Syncytial Virus (RSV). This virus is the most frequent cause of respiratory infections in children under 5. Here the molecules are separated in two classes: those with a weak or no biological activity against RSV, and those with a strong inhibitor activity.

(b) (1 point) Which of the methods performs the best?

Solution: The best ROC curve is obtained by the VS-RF (random forest) approach.

(c) (1 point) In the context of virtual screening, is it preferable to have a good *sensitivity* or a good *specificity*? Why? Which part of the ROC curve is therefore more relevant?

Solution: There will be follow up studies, so it's OK to include some false positives as long as we make sure to get as many of the true positive as possible. We therefore want to have a high sensitivity, and specificity does not matter so much. Hence we're interested in the beginning of the curve (bottom left).

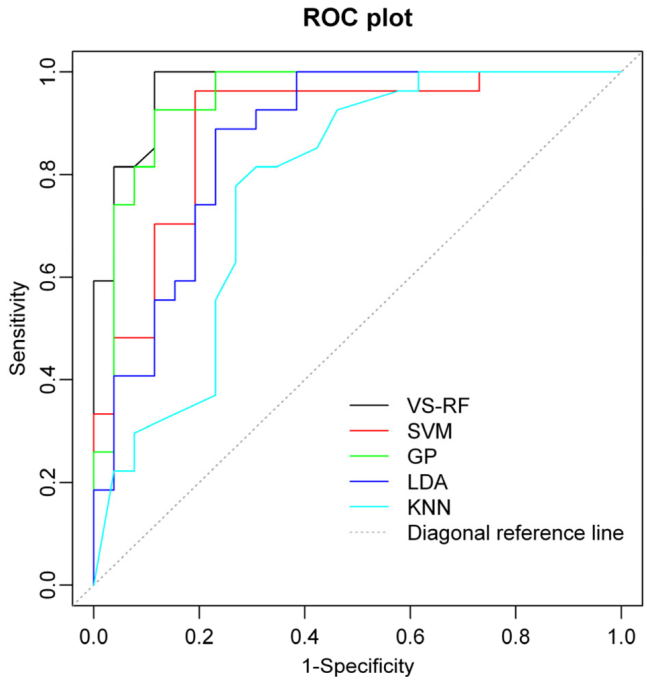


Figure 1: ROC curves for several virtual high-throughput screening methods, on the validation dataset.
VS-RF: Random Forest. SVM: Support Vector Machine. GP: Gaussian Processes. LDA: Linear Discriminant Analysis. kNN : k-Nearest Neighbors.
Source: M. Hao, Y. Li, Y. Wang, and S. Zhang, *Int. J. Mol. Sci.* 2011, 12(2), 1259-1280.