Multi-scale analysis of structural variability of *Caryophyllaceae* saponins by a simplex machine learning approach

Soumaya CHEIKH ALI (E-mail: cheikhalisoumaya@gmail.com)\textsuperscript{a,d}, Muhammad FARMAN (E-mail: farman@qau.edu.pk)\textsuperscript{b}, Asma HAMMAMI-SEMMAR (E-mail: asma.hamami@gmail.com)\textsuperscript{c}, Nabil SEMMAR (E-mail: nabilsemmar5@gmail.com)\textsuperscript{d,*}

\textsuperscript{a}University of Carthage, Faculty of Sciences of Bizerte, Tunisia
\textsuperscript{b}Quaid-i-Azam University, Department of Chemistry, Islamabad 45320, Pakistan
\textsuperscript{c}University of Carthage, Institut National des Sciences Appliquées et Technologies, Tunis, Tunisia
\textsuperscript{d}University of Tunis El Manar, Institut Pasteur de Tunis, Laboratory of BioInformatics, BioMathematics & BioStatistics, Tunisia

Abstract. A mass conservation law-based chemometric approach was developed to extract smoothed processes governing inter- and intra-molecular variability of structural diversity in metabolic pools. The approach consisted of a machine-learning method using simplex rule to calculate a complete set of smoothed barycentric molecules from iterated linear combinations between molecular classes (glycosylation classes). An application to four glycosylation levels (GLs) of Caryophyllaceae saponins highlighted aglycone-dependent variations of glycosylations, especially for gypsogenic acid (GA) which showed high 28-glucosylation levels. Quillaic acid (QA) and gypsogenin (Gyp) showed closer variation ranges of GLs, but differed by relationships between glycosylated carbons toward different sugars. Relative GLs of carbons C3 and C28 showed associative (positive), competitive (negative) or independent (unsensitive) trends conditioned by the aglycone type (GA, Gyp) and molecular (total) GLs (the four classes): 28-glucosylation and 28-xylosylation showed negative global trends in Gyp vs GLs-depending trends in QA. Also, relative levels of 3-galactosylation and 3-xylosylation varied by unsensitive ways in Gyp vs positive trends in QA. These preliminary...
results revealed higher metabolic tensions (competitions) between considered glycosylations in Gyp vs more associative processes in QA. In conclusion, glycosylations of GA and QA were relatively distant whereas Gyp (common precursor) occupied intermediate position.

**Introduction**

The Caryophyllaceae plant family was proved to be a wide source of saponins essentially based on three triterpenic skeleton (aglycones or sapogenins) including gypsogenin (Gyp), quillaic acid (QA) and gypsogenic acid (GA) [1]. Apart from the sapogenin type, structural variability of Caryophyllaceae saponins showed multi-factorial and multi-scale aspects due to different glycosylation levels (GLs) and glycosylation types essentially occurring at the carbons C3 and C28. By considering a wide dataset of 205 Caryophyllaceae saponins based on Gyp, QA and GA with different GL (2 to 9), a machine learning approach was applied to extract key information on inter- and intra-molecular regulatory processes governing the observed structural diversity in relation to aglycones (a), glycosylation levels and types (b, c) and substitution carbons (d) [2]. *In silico* combinations between saponin structures belonging to different molecular classes (GLs) provided a complete set of simulated theoretical molecules from which significant trends within and between glycosylated carbons were revealed to govern structural variability at inter-molecular scale. This helped to better understand hierarchical and sequential glycosylation orders responsible for diversification of saponins in Caryophyllaceae.

**Materials and Methods**

Machine learning approach was applied to the three aglycones separately (Gyp, QA, GA). It consisted in combining structural variabilities of saponins belonging to qmolecular classes (concerning one aglycone) representing increasing glycosylation ranges: for Gyp and QA, saponins were stratified into q=4 classes of glycosylation levels (GLs) (GLs = 1, 2, 3, 4) representing saponins with 3-4, 5-6, 7 and 8-9 substituted sugars, respectively; for GA, q=3 classes were considered (GLs = 1, 2, 3) corresponding to saponins with 3, 4, 5 substituted sugars, respectively. Saponins of different GL classes were initially characterized by the relative GLs of different sugars substituted at different carbons (C3, C16, C23, C28). Combinations between the qmolecular classes were applied using Scheffé’s simplex matrix (N rows x q columns) which provides a complete set of N mixtures varying gradually by different weights $w_j$ (from 0/5 to 5/5) of the q mixed GL classes $j$ (with $\Sigma w_j = 1$) [2].

In output of each combination, a barycentric molecular profile was calculated by averaging the relative levels of glycosylation (G) profiles of the n randomly sampled contributive saponins. The mixture design was iterated 30 times by bootstrap technique then the 30 resulting response matrices (containing Nelementary barycentric G-profiles) were averaged leading to a final response matrix containing Nsmoothed barycentric G-profiles and representing a deep regulatory machinery of the whole studied structural system.

The smoothed response matrix was used for graphical analysis of regulatory trends between glycosylated carbons. For two given glycosylated carbons, different regulatory trends were highlighted by considering successions of weight ellipses associated to different GL classes [2].
Results and Discussion

Data smoothing by simplex machine learning approach helped to highlight regulatory processes of glycosylation in Caryophyllaceae saponins at inter-molecular, inter-atomic and intra-atomic scales. Illustrations are provided by xylose (Xyl), glucose (Glc) and galactose (Gal) substituted at carbons C3 and/or C28.

**Inter-molecular scale variations 1 (aglycone effect).** The three simplex plots associated to the three aglycones-based saponins showed strong differentiation spaces of relationships between the same glycosylated carbons (Figure 1a, b). Illustrations are given for 28-Xyl vs 28-Glc (Figure 1) and 3-Xyl vs 3-Gal (Figure 2):

GA was markedly distant from Gyp and QA indicating some specific glycosylation orders (in GA). This could be linked to the occurrence of two carboxylic groups in GA(23- and 28-COOH) vs only one (28-COOH) in both QA and Gyp. Specific space of GA was characterized by strong relative 28-Glc levels (0.65-0.75) without competing (comparable) levels from other sugars in both C3 and C28 (Figure 1a)[1, 3, 4]. Gyp occupied intermediate position between GA and QA whereas QA was the most distant from GA (Figure 1a, 2b). Relative locations of different aglycones-based saponins in simulated graphics were compatible with the metabolism: Gyp is metabolically precursor of both QA and GA (by 16- and 23-hydroxylation, respectively); this is compatible with intermediate position of Gyp between QA and GA (competing for Gyp) [3].

Although Gyp and QA plots showed spatial neighboring (compared to GA), they differed by dispersions, inclinations and internal organization of corresponding clouds of points: relationship between 3-Xyl and 3-Gal showed significantly higher dispersion in QA than in Gyp (Figure 1b). This aspect indicated wider regulation range of 3-Gal in QA compared to Gyp[3]. However, for (28-Xyl vs 28-Glc), the two aglycones showed opposite relationship leading to inversely inclined clouds of points: (global positive trend in QA against negative trend in Gyp) (Figure 1a).

**Inter-molecular scale variation 2 (molecular glycosylation effect).** For a same aglycone, global variation trends between substituted sugars significantly varied with molecular GLs:

In Gyp, 28-Glc showed a strong peak under low molecular GL (GL=1) followed by rapid decrease to minimum in higher GLs (GL=2, 3, 4) (Figure 1b). This highlighted strong contribution of early 28-Glc in ramification process (monodesmosylation) of Gyp leading to preliminary structural diversification of saponins [3, 5, 6]. In QA, 28-Glc seemed to play intermediate modulatory role due to its low relative levels in GLs = 1 and 4 (figure 1c).
However, in Gyp, relative levels of 28-Xyl showed alternated states with molecular GLs: global minimal regulations in the less glycosylated saponins (GL=1), followed by maximal global regulations in GL=2 and 4 via slight decrease in GL=3 (Figure 1b). This highlighted key role of 28-Xyl in
elongation process for the synthesis of Gyp-based saponins with high GLs. Also, in QA, 28-Xyl showed gradual (step-by-step) increases of relative levels with molecular GLs. This was indicative of key role of xylosylation in elongation process at C28 in QA (Figure 1c).

Inter-atomic scale variation. By considering xylosylation between C3 and C28 in QA, 3-xylosylation seemed to be initially favored while 28-Xyl was at its lowest levels (in GL=1) (Figures 1c, 2b). This could be indicative of key role of 3-xylosylation in molecular ramification of QA [3]. For higher GL (2 to 4), relative level of 28-Xyl showed continuous increase vs alternated variation for 3-Xyl. This highlighted increasing affinity of C28 vs alternated affinity of C3 for xylosylation with GLs in QA.

Intra-atomic scale variation. By considering the inclinations of weight ellipses of different GLs, further regulatory processes were highlighted between 28-Xyl and 28-Glc conditionally to the aglycone type (Gyp vs QA) (Figures 1a, 2):

In Gyp, weights’ ellipses showed negative inclinations indicating systematic competitions between Xyl and Glc for substitution at C28 in all the GLs (Fig. 1b). Such a competition could suggest the implication of different glycosyl-transferases (GT) in the two glycosylations of C28 [3]. Hypothesis on specific GT in Gyp could be also locally indicated by not clearly inclined weight ellipses in the relationship 3-Xyl vs 3-Gal.

However, in QA, weights’ ellipses in (28-Xyl vs 28-Glc) showed positive inclinations indicating some associative process between these two sugars for their 28-substitution despite some global negative trends (for GLs=2, 4). Such associative substitution process could occur under the effect of a same GT having a promiscuity character and leading to sequential glycosylations by different sugars [5, 7-10]. Also, weight ellipses in 3-Xyl vs 3-Gal showed some positive inclination (except in GL=1) leading to further indication about intra-atomic associations between considered glycosyls in favor of shared (promiscuity) GT hypothesis.

Conclusions
Simplex-based machine learning applied to structural variability of Caryophyllaceae saponins highlighted strong differentiation in metabolic glycosylation governed by the aglycone type, molecular GL and substituted carbon. GA was strongly characterized by high levels of 28-Glc followed by Gyp then QA. Effect of GLs was partially associated to key role of 28-Xyl in elongation in both Gyp and QA. At inter-atomic scale, preliminary 28-Glc and 3-Xyl seemed to be responsible for molecular ramification in Gyp and QA, respectively. At intra-atomic scale, 28-Xyl and 28-Glc showed competition in Gyp and association in QA that could be indicative of different and shared GTs, respectively. Finally, metabolic tensions for glycosylation seemed to decrease from GA to QA via Gyp.

References:


