
SciForum
MOL2NET

# Chemometric highlighting of metabolic diversification factors at inter- and intra-molecular scales by a new simplex-based training approach: application to *Astragalus* saponins

**Abir SARRAJ-LAABIDI [1, 2], Habib MESSAI [3], Asma HAMMAMI-SEMMAR [4] and Nabil SEMMAR [1, 5, *]**

[1] Université de Tunis El Manar, Institut Pasteur de Tunis, Laboratory of Bioinformatics, Biomathematics and Biostatistics (BIMS), 1002, Tunis, Tunisia; sarraj.abir@live.fr; nabilsemmar@yahoo.fr

[2] Université de Tunis El Manar, Faculté des Sciences de Tunis, Campus Universitaire, 2092 Tunis, Tunisia; sarraj.abir@live.fr

[3] Université de Tunis El Manar, Institut Pasteur de Tunis, Laboratory of Biomedical Genomics and Oncogenetics, 1002, Tunis, Tunisia; habib.messai@gmail.com

[4] Université de Carthage, National Institute of Applied Sciences and Technology (INSAT), 1080, Tunis, Tunisia; asma.hamami@gmail.com

[5] Aix-Marseille Université, Faculté des Sciences de Saint-Jérôme, 13397 Marseilles, France; nabilsemmar@yahoo.fr

***** Author to whom correspondence should be addressed; nabilsemmar@yahoo.fr ;

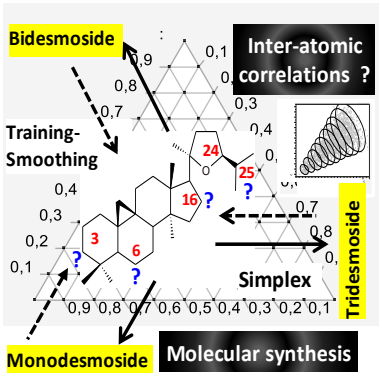
**Abstract:** *Metabolisms represent highly organized systems characterized by strong regulations obeying to mass conservation law. This makes a whole chemical resource to be competitively shared between several biosynthesis components (ways) at both intra-and inter-molecular scales. Statistically, the whole shared-resource principle can be considered under a constant or unit sum constraint which represents the basis of simplex mixture rule. In this work, a new simplex-based chemometrics approach was developed to extract scaffold information on different biosynthesis regulation factors responsible for the chemical structural diversity at atomic and molecular scales within a large metabolic system. This approach consisted in linearly combining different (q) molecular clusters into a complete set of N mixtures by gradually varying their relative weights. The complete set of the N combinations was given by Scheffé's mixture matrix. In output of Scheffé's design, each molecular combination was represented by a theoretical average (barycentric) molecule which was trained by the characteristics of the different weighted clusters. The mixture design was iterated several (K) times by bootstrap technique to explore chemical variability between and within clusters. Finally, the K response matrices issued from the K iterations were averaged to obtain a smoothed matrix containing scaffold information on different*

*regulation factors responsible for molecular diversification at inter- and intra- (atomic) molecular scales. This matrix was used as a backbone for graphical analysis of positive and negative trends between atomic characteristics: the chemical substitutions levels of carbons. This new simplex approach was illustrated by cycloartane-based saponins of Astragalus genus by considering three desmosylation clusters (mono-, bi- and tridesmoside saponins) characterized by relative glycosylation levels of different aglycones' carbons.*

---

**Keywords:** *Simulation, molecular diversification, glycosylation, desmosylation, simplex mixture design*

**Graphical Abstract:**



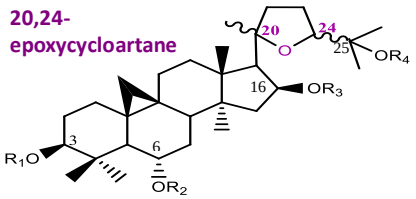

.

.

**Introduction:** This work presents a new chemometric approach combining structural information contained in a large set of molecular structures to highlight mechanistic order governing molecular diversification via sequential ramifications and elongations occurring at different carbons. The approach was based on Scheffé's mixture design [1].

Simulated results are initially trained by inter- and intra-molecular variability contained in a large set of chemical structures available in literature. The simplex approach was applied to saponins of *Astragalus* genus which is the largest taxon of terrestrial plants with more than 2200 species [2]. Saponins are essentially based on cycloartane aglycone which has several forms including structures with either aliphatic lateral chain or epoxylated cycle(s) [3]. Epoxylated forms consists mostly of 20,24-epoxycycloartane (**Figure 1**).

.




.

**Figure 1**. Chemical structure of 20,24-cycloartane with different substitutions implied in saponins synthesis. Xyl, xylose; Rha, rhamnose; Glc, glucose; Ara, arabinose; Api, apiose; Ac, acetyl.

.

.

**Materials and Methods:** Molecular formations obey to mass conservation principle under which the whole mass of a system can be shared between $q$ components ($A$, $B$, $C$, etc.) according to many ways. This leads to many possible multiplets where weights or proportions $w_j$ vary the ones at the expense of the others under constant or unit

sum constraint (**Eq. 1**), i.e. the sum of *q* mass parts equal to the whole initial mass (**Figure 2**):

.

$$\sum_{j=1}^{q} weights \quad w_j = w = Cst \qquad (1)$$

Sharing processes between exclusive and complementary system's components is statistically governed by simplex rule where components' parts vary the one relatively to the other under the constraint of limited total resource available in the whole system.

In molecular systems, the simplex rule find application both between and within molecules (i.e. between carbons of a same molecule) (**Figure 2**):

At inter-molecular level, molecular clusters exclude the one the others by simple biosynthesis (**Figure 2a**). At intra-molecular scale, a molecule can be conceived as a set of carbons competing for chemical substitutions leading the whole molecular substitution level to be shared between carbons (**Figure 2b**).

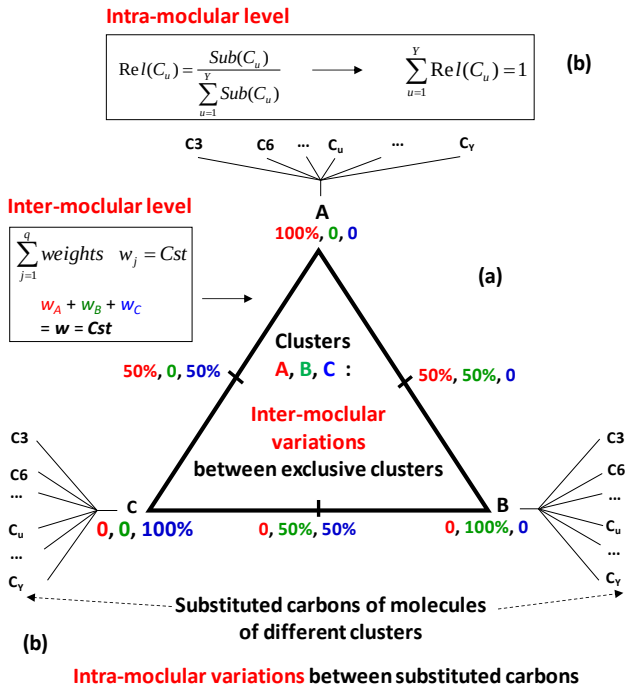



**Figure 2**. Geometric representation and numeric formulation of simplex rule governing mass conservation between several components at inter- (a) and intra- (b) molecular scales.

In this work, molecular clusters consisted of three desmosylation levels represented by 108 *Astragalus* saponins based on 20,24-epoxycycloartane and characterized by different relative glycosylation levels of carbons. Desmosylation (*D*) and glycosylation (*G*) are responsible for saponins' ramification and elongation, respectively (**Figure 3**). In the current work conceptualization, *D* and *G* represented two metabolic variability factors at inter-molecular and intra-molecular scales, respectively.

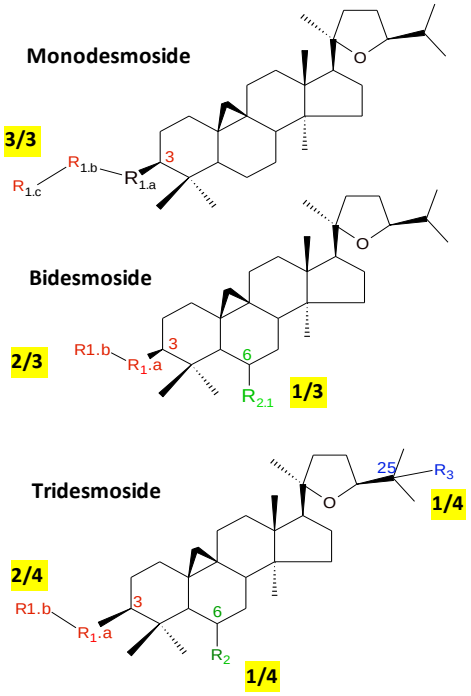



**Figure 3**. Illustration of three saponin clusters associated to three desmosylation levels and characterized by intra-molecular variation of relative glycosylation levels of different carbons

Molecular diversification mechanisms of *Astragalus* saponins were computztionally approached by a training process based on iterative combinations of structural information characterizing the three desmosylation clusters. The complete combinatorial process between the three clusters was based on Scheffé's mixture design (**Figure 4a**) [1]. With $q = 3$ clusters and by fixing the total weight $w$ to 10 molecules per mixture, combinatorial formula gives a total $N$ of 66 combinations (**Eq. 2**) (**Figure 4b**):

$$N = \frac{(w+q-1)!}{(q-1)!w!} \qquad (2)$$

Application of Scheffé's mixture design consisted in linearly combining the three desmosylation clusters $j$ by varying their weights $w_j$ the ones at

the expense of the others (**Figure 4c**). Each linear combination was applied by randomly sampling a total of w molecules from the three clusters by respecting the three variable weights $w_j$ given by Scheffé's mixture design. In output, the *w* sampled molecules were summarized by a theoretical barycentric molecule calculated by averaging the w glycosylation profiles representing mono-, bi- and tri-desmoside clusters (**Figure 4d**). With 66 combinations, we obtained 66 elementary responses consisting of 66 barycentric molecules characterized by 66 average relative glycosylation profiles (**Figure 4e**).

Scheffé's mixture design was iterated several times in order to explore molecular variability between and within clusters (**Figure 5**): a single mixture design combining $w$=10 molecules per mixture is insufficient for good training from the whole available molecular data. For that, Scheffé's matrix was iterated several times, i.e. *K* times, leading to *K* response matrices (**Figure 5a**). After $K$ = 30 iterations of mixture design, 30 elementary response matrices were obtained, each one containing $N$=66 barycentric molecular profiles (**Figure 5b**). Finally, the 30 elementary response matrices were averaged leading to a final matrix containing 66 smoothed molecular profiles integrating high inter- and intra-molecular variability (**Figure 5c**). This final matrix was used for graphical analysis of relationships between substituted carbons (**Figure 5d**).

Relationships between carbons were conditional to each desmosylation clusters. For that, in each plot, the three desmosylation clusters *j* were separately considered by projecting their 11 weights $w_j$ ($w_j$ = 0 to 10) on corresponding points (**Figure 6**). Then, equal weights were statistically grouped by a confidence ellipse. The succession of the eleven ellipses from 0 to $w$=10 resulted in a trajectory highlighting how the two considered carbons varied the one in relation to the other for the formation of considered desmosylation level.

**Results and Discussion:** Smoothed relationships highlighted several relationships between carbons depending on the desmosylation level (molecular cluster) (**Figure 6**):

.

Global trajectories for monodesmoside formation initially implied increase in glycosylation level of C3 at the expense of C6, C25 (**Figure 6a**). Moreover, at local (intra-molecular) scale, negatively inclined ellipses in the three plots indicated systematic tension between C3, C6, C25 for glycosylation in favor of C3.

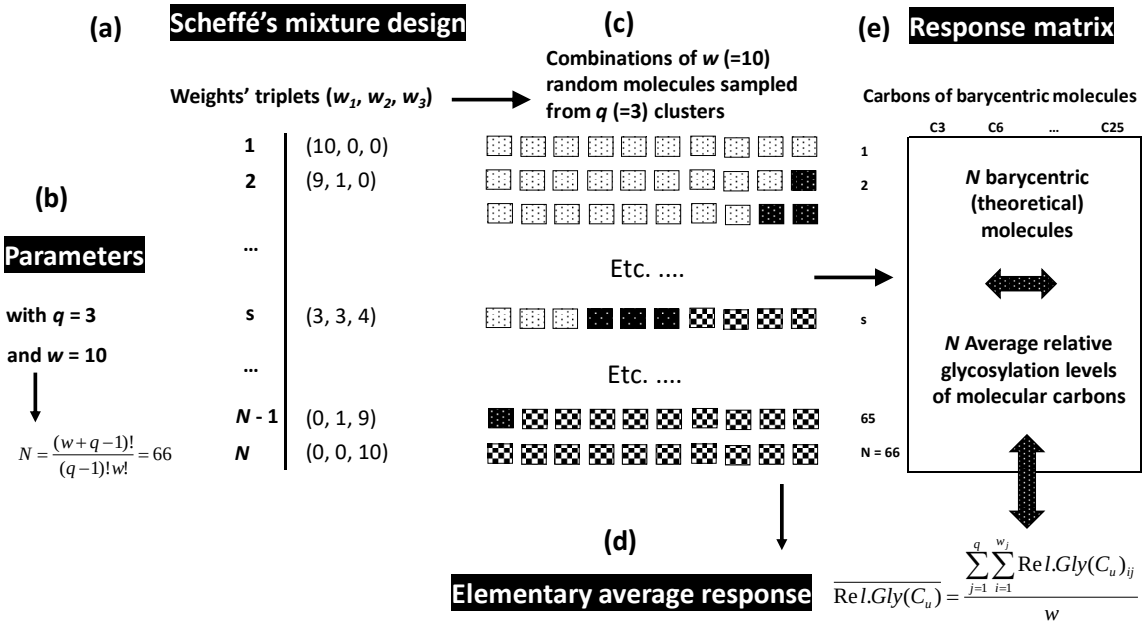
For bidesmosylation formation, global relationships between carbons highlighted significant increase of 6-glycosylation at the expense of C3 which stabilized at intermediate relative glycosylation level (**Figure 6b1**). However, C25 was not favored in bidesmosylation system (**Figures 6b2, 6b3**). However, weights' ellipses in C25 vs C6 showed positive inclination indicating that 6-glycosylation could be favorable factor for 25-glycosylation. Such hypothesis found checking in tridesmosylation:

Tridesmosylation implied further decrease of relative 3-glycosylation in favor of C25-glycosylation (**Figure 6c2**). However, C6-glycosylation slightly decreased by maintaining relatively high level indicating its open (alternative) role for glycosylation for tridesmoside formation (**Figure 6c3**). Systematic intra-molecular positive trend between C6- and C25-glycosylations was highlighted by positively inclined weights' ellipses contrary to negative states in C6 and C25 vs C3 plots (**Figures 6c1-6c3**).

Highlighted sequential glycosylation mechanisms revealed in agreement with [4-8]:

- The promiscuity characterizing glycosyl-transferases of saponins.
- The initial hydroxylation occurring at C3 during aglycone formation from 2,3-oxidosqualene.

**Conclusions:** Simplex simulation approach provides useful chemometrics tool for extraction and visualization of inter-atomic factors governing molecular diversity in a large metabolic dataset (population). It can be applied on other structural criteria than desmosylation leading to multiple analysis ways of molecular diversification factors. In consequence, such a computational tool has wide perspective applications in metabolomics.




.

**Figure 4**. Principle of the simplex approach based on Scheffé's matrix (**a**) and combining structural information of three molecular clusters (**c**) to simulate a response matrix of barycentric molecules (**e**) by averaging the contributive molecules to combinations (**d**). Required number of combinations is calculated from the number *q* of clusters and the whole weight *w* of mixture (**b**).

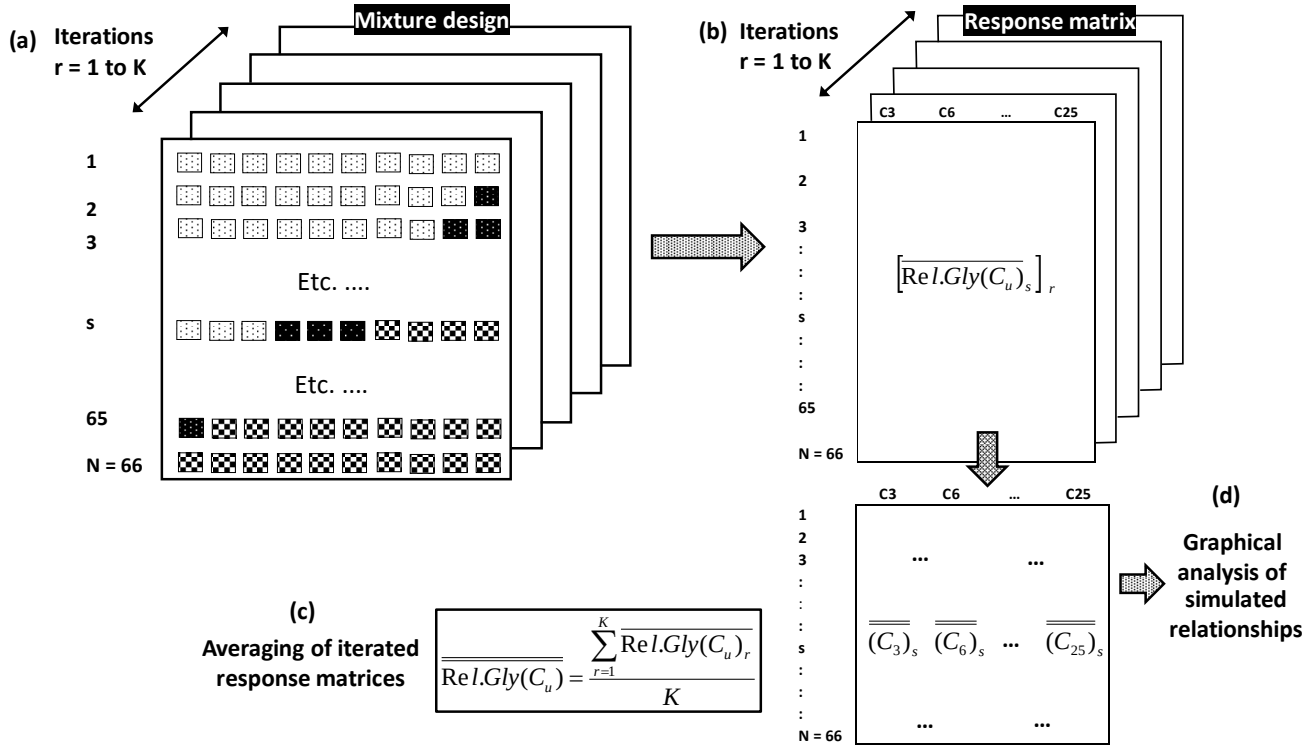



**Figure 5**. Iteration of Scheffé's mixture design (**a**) and its response matrix of barycentric molecules (**b**) to calculate a final smoothed matrix (**c**) used for graphical analysis of relationships between carbons leading to highlight molecular diversification mechanisms (**d**).

.

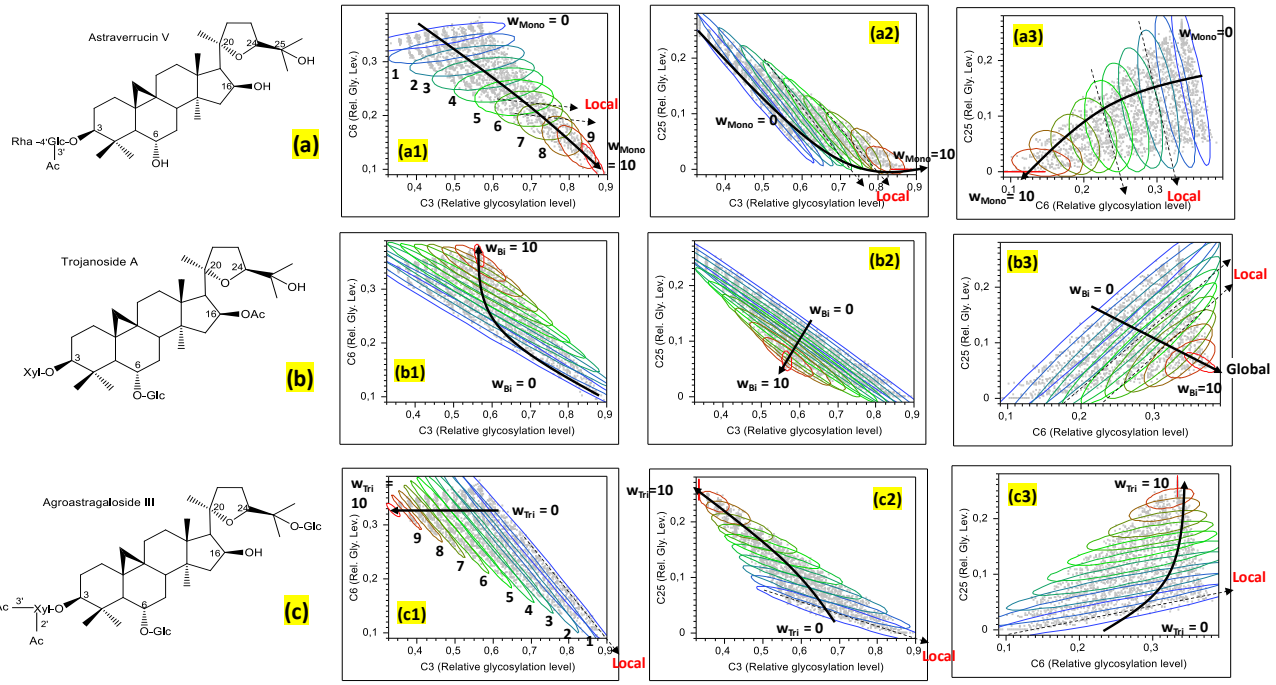



**Figure 6**. Smoothed relationships between three glycosylated carbons (C3, C6, C25) of *Astragalus* saponins showing global trajectories and local variations associated to metabolic diversification factors at inter- and intra-molecular scales, respectively.

**References:**


1. Scheffe, H. The simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)* **1963**, 235-263.
2. Ionkova, I.; Shkondrov, A.; Krasteva, I.; Ionkov, T. Recent progress in phytochemistry, pharmacology and biotechnology of *Astragalus* saponins. *Phytochemistry Rev.* **2014,** *13* (2), 343-374.
3. Sarraj-Laabidi, A.; Semmar, N.. Interspecific Chemical Differentiation within the Genus *Astragalus* (Fabaceae) Based on Sequential Variability of Saponin Structures. In: *Plant Biodiversity*, Ansari, A., Gill, S.S., Abbas, Z.K., Naeem, M. (Eds). CABI, London, UK, **2016**; Chapter 27.
4. Augustin, J. M.; Kuzina, V.; Andersen, S. B.; Bak, S. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry* **2011,** *72* (6), 435-457.
5. Hansen, K. S.; Kristensen, C.; Tattersall, D. B.; Jones, P. R.; Olsen, C. E.; Bak, S.; Møller, B. L. The in vitro substrate regiospecificity of recombinant UGT85B1, the cyanohydrin glucosyltransferase from *Sorghum bicolor*. *Phytochemistry* **2003**, 64 (1), 143-151.
6. Kramer, C. M.; Prata, R. T. N.; Willits, M. G.; De Luca, V.; Steffens, J. C.; Graser, G. Cloning and regiospecificity studies of two flavonoid glucosyltransferases from *Allium cepa*. *Phytochemistry* **2003**, 64 (6), 1069-1076.
7. Modolo, L. V.; Blount, J. W.; Achnine, L.; Naoumkina, M. A.; Wang, X.; Dixon, R. A. A functional genomics approach to (iso) flavonoid glycosylation in the model legume *Medicago truncatula*. *Plant Molecular Biology* **2007**, 64 (5), 499-518.
8. Vogt, T.; Jones, P. Glycosyltransferases in plant natural product synthesis: characterization of a supergene family. *Trends Plant Sci*. **2000**, 5 (9), 380-386.