Description of the software

C13dist

Author: Naoki Sato (University of Tokyo)

E-mail: naokisat@bio.c.u-tokyo.ac.jp

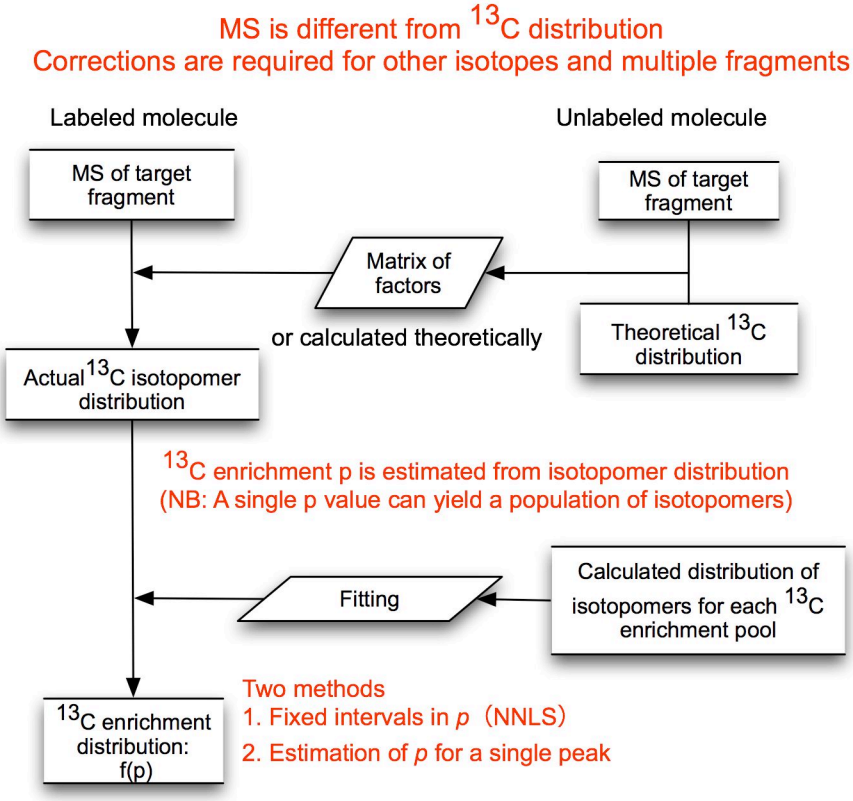
Overview

This document describes the computational methods for the analysis of ^{13}C content in ^{13}C-labeled compounds. A ^{13}C-labeled compound contains not only the ^{13}C incorporated as a result of labeling but also naturally occurring ^{2}H, ^{18}O and ^{13}C (and ^{28}Si and ^{30}Si for trimethylsililated (TMS) compounds). ^{13}C, which is not related to the labeling, is also present in the methoxy group in the fatty acid methyl esters or methyl groups of TMS-derivatives. In addition, an apparently single peak of molecular ion (M^+) accompanies small peaks of $(M-1)^+$ and $(M-2)^+$. In some cases in which M^+ is weak, relevant fragment peaks are used for obtaining information on labeling.

To remove these effects, a raw mass spectrum should be converted to a pure isotopomer distribution, namely, a composition of various molecular species containing different numbers of ^{13}C labels. The method is now quite common for simple calculation involving natural abundance of isotopes (Wahl et al. 2004, Hellerstein and Neese 1999), but a more universal methodology involving the use of fragment ions is described here. The method was first described in an early paper on the direct desaturation of fatty acids (Sato et al. 1986), but was not recognized by later papers specialized in mass spectrometric methods (Lee et al. 1991, Fernandez et al. 1996). Now a new version of software was developed using the GNU Scientific Library (GSL). In addition, the current software uses the NNLS (non-negative least square) program (Lawson and Hanson 1974). The source code “nnls.c” was obtained from http://hesperia.gsfc.nasa.gov/~schmahl/nnls/.

This software is also useful in simulating ^{13}C isotopomer distribution with different concentrations of the isotope, and deconvoluting actual isotopomer distribution into various different pools labeled with different abundance. In contrast with the published methods (Lee et al. 1991, Fernandez et al. 1996, Wahl et al. 2004), the present method is an empirical one that uses the actual mass spectrum of an unlabeled compound (labeled at the natural abundance) for calculating the conversion factors, which is then used for calculating the isotopomer distribution from the mass spectrum of labeled compound. In this way, isotopomer distribution can be obtained from the composite mass spectrum of $(M-31)^+$ and $(M-32)^+$ ions of methyl oleate, in which M^+ ion is rather weak. This is also useful in deducing isotopomer distribution of glycerol from the complex spectrum around the $(M-90)^+$ and $(M-91)^+$ ions of TMS-glycerol.

The usage of the software is schematically shown in the figure below.



References

Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. and Brunengraber, H. (1996) Correction of ^{13}C mass isotopomer distributions for natural stable isotope abundance. J. Mass Spectrom. 31, 255-262.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M. and Rossi, F. (2009) GNU Scientific Library Reference Manual - Third Edition. Network Theory Ltd. http://www.gnu.org/software/gsl/

Hellerstein, M. K. and Neese, R. A. (1999) Mass isotopomer distribution analysis at eight years: theoretical, analytic, and experimental considerations. Amer. J. Physiol. Endocrinol. Metab. 276, E1146-E1170.

Lawson, C. L. and Hanson, R. J. (1974). Solving Least Squares Problems. Prentice Hall, Englewood Cliffs NJ, USA.

Lee, W.-N. P., Byerley, L. O., Bergner, E. A. and Edmond, J. (1991) Mass isotopomer analysis: Theoretical and practical considerations. Biol. Mass Spectrom. 20, 451-458.

Sato, N., Seyama, Y. and Murata, N. (1986) Lipid-linked desaturation of palmitic acid in monogalactosyl diacylglycerol in the blue-green alga (cyanobacterium) *Anabaena variabilis* studied in vivo. Plant Cell Physiol. 27, 819-835.

Wahl, S. A., Dauner, M. and Wiechert, W. (2004) New tools for mass isotopomer data evaluation in ^{13}C flux analysis: mass isotope correction, data consistency checking, and precursor relationships. Biotechnol Bioeng. 85, 259-268.

1. Basic explanation of the calculation

1.1 Definition

1.1.1 Vectors (dimension is n or $n+a$)

The following vectors are defined:

$\boldsymbol{f}$: fragment pattern vector (components describe the intensities of fragments)

$\boldsymbol{h}$: non-C isotopomer distribution (components describe the distribution of isotopomers of H, O, Si, or C in derivative groups)

$\boldsymbol{c}$: C13 isotopomer distribution (components describe the distribution of ^{13}C isotopomers)

$\boldsymbol{s}$: mass spectrum (components describe the intensities of mass peaks starting from the base mass)

In all vectors, the dimension is n (the number of carbon atoms in the target molecule) or $n+a$, where a is a small positive integer, covering all the ions (completely labeled ions contains additional peaks due to labels in H etc.).

1.1.2 Matrices (dimension is n or $n+a$)

The following matrices that correspond to the vectors are defined:

$\boldsymbol{F}$: lower triangular matrix for $\boldsymbol{f}$

$\boldsymbol{H}$: lower triangular matrix for $\boldsymbol{h}$

$\boldsymbol{C}$: lower triangular matrix for $\boldsymbol{c}$

1.1.3 Conversion of matrix and vector

The vector representation and the matrix representation are equivalent as shown below.

$$\boldsymbol{C} = \begin{bmatrix} c_0 & 0 & 0 & 0 \\ c_1 & c_0 & 0 & 0 \\ c_2 & c_1 & c_0 & 0 \\ c_3 & c_2 & c_1 & c_0 \end{bmatrix} \Leftrightarrow \boldsymbol{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

1.2 Master equation

The vectors and matrices are related by the following equations:

$\boldsymbol{C}\,\boldsymbol{H}\,\boldsymbol{f} = \boldsymbol{H}\,\boldsymbol{C}\,\boldsymbol{f} = \boldsymbol{s}$

If $\boldsymbol{f}' \equiv \boldsymbol{H}\,\boldsymbol{f}$, then

$\boldsymbol{C}\,\boldsymbol{f}' = \boldsymbol{s}$

or

$\boldsymbol{F}'\,\boldsymbol{c} = \boldsymbol{s}$

where $\boldsymbol{F}'$ is a lower triangular matrix for $\boldsymbol{f}'$, as defined like $\boldsymbol{C}$ and $\boldsymbol{c}$ (see 1.1.3).

$$\boldsymbol{F}' = \begin{bmatrix} f'_0 & 0 & 0 & 0 \\ f'_1 & f'_0 & 0 & 0 \\ f'_2 & f'_1 & f'_0 & 0 \\ f'_3 & f'_2 & f'_1 & f'_0 \end{bmatrix} \Leftrightarrow \boldsymbol{f}' = \begin{bmatrix} f'_0 \\ f'_1 \\ f'_2 \\ f'_3 \end{bmatrix}$$

1.3 Factor vector

The factor vector $\boldsymbol{f}'$ is obtained by using the mass spectrum $\boldsymbol{s}_0$ of unlabeled compound and theoretical abundance of isotopomers $\boldsymbol{C}_0$. In the software, this is performed by the mode 3.

Observed: $\boldsymbol{s}_0$

Theoretical: $\boldsymbol{C}_0$

$\boldsymbol{f}' = \boldsymbol{C}_0^{-1}\, \boldsymbol{s}_0$

1.4 Isotopomer distribution

The isotopomer distribution of an actual labeled compound $\boldsymbol{c}_1$ is calculated from observed mass spectrum $\boldsymbol{s}_1$ and the factor vector (matrix) $\boldsymbol{F}'$. In the software, this is performed by the mode 4.

Observed: $\boldsymbol{s}_1$

Deduced: $\boldsymbol{F}'$

$\boldsymbol{c}_1 = \boldsymbol{F}'^{-1}\, \boldsymbol{s}_1$

1.5 Multi-fit linear regression

If the dimension of $\boldsymbol{s}$ is larger than the number of C atoms, multi-fit regression analysis is used for solving the equations in the steps 1.3 and 1.4.

1.6 Distribution of isotopic abundance

Photosynthetically incorporated ^{13}C shows ideally an isotopomer distribution, which is determined by a binomial distribution $B(n, p)$ depending on the isotopic abundance p, and the distribution of isotopic abundance $D(p)$, where n is the number of carbon atoms of the compound. It is essential to distinguish the isotopomer distribution and the distribution of isotopic abundance. Imagine a situation in which a compound is labeled with ^{13}C at an abundance of $p = 0.5$, namely, a mixture of 50% ^{12}C and 50% ^{13}C. Then, various isotopomers are observed with a distribution $B(n, 0.5)$, having a peak at $0.5n$. This distribution, however, represents the simple fact that the compound is labeled by the pool of carbon containing 50% ^{13}C. This intrinsic value representing the ^{13}C abundance in the labeled molecules is called “isotopic abundance” in the present article. If a model of isotopic abundance distribution $D(p)$ is defined, the distribution of isotopic abundance can be calculated by fitting. The model may be (1) discrete model consisting of various discrete values of p, (2) continuous model consisting of continuous values of p from a lower limit to an upper limit, (3)

fixed-level average model, or any model described by a mathematically defined function. The current software implements the two models, (1), (2) and (3). A general model may be useful in a general sense, but in the actual analysis, the above-mentioned models may be practical, since the dimension of variables is limited by n. In the software, this is performed by the mode 5.

2. Downloading and compiling of the software

The software is written in the C language, and can be compiled on any UNIX systems, such as Linux and MacOS X. It uses the GNU Scientific Library (GSL) for matrix calculation, and this library (versions 1 or 2) should be installed on the system before installing the C13dist software. In addition, the current software uses the NNLS (non-negative least square) program. This algorithm was originally developed by Lawson and Hanson (1974). The source code "nnls.c" was obtained from http://hesperia.gsfc.nasa.gov/~schmahl/nnls/. The source code for the C13dist program is available at the URL: http://nsato4.c.u-tokyo.ac.jp/old/C13dist.html (note that the 'old' directory is not old). After decompressing the archive, the software is compiled by the command 'make'. The contents of the 'makefile' should be modified to fit the system. Sample files are available in the archive. The software does not use a graphical interface. It is used only on a command-line interface (such as 'Terminal' in MacOS X).

3. Basic operation

All files are text files. If the input file (mass list) is an output of Windows software (usually used for the operation of mass spectrometer), the end-of-line (return) code should be adjusted to fit to UNIX system by using either a perl script or some software. We use the SISEQ tools to change the return code. Input file 'file1.txt' is converted to 'a.txt' by the txtr command:

txtr file1.txt a.txt cr

SISEQ software is available at the URL: http://nsato4.c.u-tokyo.ac.jp/old/Siseq.html.

Basic commands of the software C13dist are as follows (examples are included in the software package):

Mass spectrum analysis

-stdev is used for data with standard deviation

Error calculation is not completely implemented in the current software.

Mode 3: Reads mass spectrum of a compound with natural abundance and prints a factor table

Input: intensity distribution table of a single fragment

Example 8:

C13dist -mode=3 -stdev -C=37 -dim=30 18_1_16_0_MADG_std output8

Output: **stdout** (input file and estimated C13 distribution at natural abundance), file **output8** (factor table)

Mode 4: Converts mass spectrum to C13 isotopomer histogram

Input: factor table (result of 4.1) and intensity distribution table of a single fragment (after labeling)

Example 9:

C13dist -mode=4 -stdev -C=37 -dim=30 output8 18_1_16_0_MADG_0h output9

Output: **stdout** (input files), file **output9** (C13 isotopomer histogram)

Mode 5: Fitting C13 content distribution

Input: model number, C13 isotopomer histogram (result of 4.2)

Example 10: (discrete model: isotopomer distribution is fitted with <dim>-step histograms)

C13dist -mode=5 -model=1 -stdev -C=37 -dim=10 output9 output10

Example 11: (average model: fitting with the minimum (natural abundance) and maximum level (estimated) of enrichment and continuous flat distribution between them. p is the lower boundary of the search for the maximum level. p must be less than the estimated maximum level. This should be re-entered after viewing the output of fitting error.)

C13dist -mode=5 -model=2 -stdev -C=37 -p=0.4 output9 output11

The **output11** file indicates that the observed C13 isotopomer

distribution is fitted by two distributions, each having the natural abundance and p=0.51, respectively, and the populations of them are 83 and 17%, respectively.
Output: **stdout** (intermediate data), file **output11** (finally fitted C13 enrichment pattern)

4. Preparation of a matrix for unlabeled molecule

A text file 'A. txt' describing the intensities of a mass spectrum of unlabeled molecule is prepared from the output of MALDI-TOF MS or GC/MS. It is desirable to use an average of several measurements. An example of methyl palmitate (16:0-Me) is like this:

```
16:0me
268
8
10
20
15269
6001
205
10
5
0
```

In this file, all the values are described without spaces or indentation. The first line is the name of compound (or any word). The second line is the mass of the base peak (this is NOT necessarily the strongest peak. Base peak means just the "starting peak" in the list). The third line is the number of peaks described. The following lines describe the peak intensities. The main peak of 16:0-Me appears at m/z=270, and the peaks at 271, 272, etc. represent peaks of ^{13}C-containing isotopomers. The small peaks usually appear at 268 and 269, too. These might be M-1 and M-2 peaks lacking H atom(s). Theoretically, there must be ^{13}C-isotopomers of such peaks, which complicates the spectrum. In the calculations, all these peaks can be treated just as satellite peaks accompanying the main peak. A return code should be added at the last line of the file.

5. Calculation of a factor table

Now this file is converted to a factor table. A factor table is a hypothetical mass distribution by removing the effect of natural abundance level of ^{13}C. The table includes the effects of multiple ions (such as M-1 and M-2 as described above) and the effects of isotopes of H and O, as well as C in the methyl group, which is not the subject of ^{13}C labeling. The command is like this:

```
C13dist 3 –C=16 –dim=8 a.txt 160me.f0
```

Here, -C=16 defines that the number of carbon atoms to be analyzed in this experiment is 16 (but not 17 which is included in the entire molecule of 16:0-Me). The –dim=8 message defines the size of internal matrix, which is suitable for analyzing 8 peaks in the a.txt file. The output file is 160me.f0. The extension 'f0' is used for the factor table. The contents of the f0 file are scrolled on the screen. Note that the **–mode=3** option can be replaced by just **3**.

In the case of analyzing 18:1-Me by GC/MS (EI mode), in which M^+ peak is rather weak but $(M-31)^+$ and $(M-32)^+$ peaks are intense, the isotope peaks due to these ions are useful. Then, the factor table is calculated using a mass table describing from M-33. For 18:2-Me and 18:3-Me, M^+ is useful without problems. For MALDI-TOF-MS, $(M+Na)^+$ or $(M+1)^+$ ions (or $M+NH_4)^+$ are useful.

6. Conversion of a mass spectrum of labeled molecule into the isotopomer distribution

An intensity list of the mass spectrum of a labeled molecule is then prepared as described in Section 4. Normally, the mass range is wider now from *M*-2 until *M*+*N*+2, where *N* is the number of carbon atoms to be labeled. The dimension should be larger than the value of *N*: 20 for 16:0-Me, for example. The conversion is effected by the following command:

```
C13dist 4 -C=16 -dim=20 160me.f0 datafile.txt datafile.c1
```

The output file 'datafile.c1' in this case, using the extension 'c1', describes the distribution of isotopomers.

7. Mathematical analysis of the isotopomer distribution: Estimation of the distribution of isotopic abundance

Now we can estimate the isotopic abundance and the percentage of labeled molecules, using the 'fit' mode. The following calculations are used to estimate the distribution of isotopic abundance from the isotopomer distribution.

The following examples are used for analyzing palmitic acid (having 16 carbons).

```
C13dist -fit -model=2 -C=16 -p=1.0 -q=0.7 datafile.c1 datafile.d
```

The file 'datafile.d' reports the results of fitting, or estimated distribution of the isotopic abundance, from 0.7 to 1.0.

```
C13dist -range -C=16 -low_lim=6 -up_lim=16 datafile.c1 datafile.cL
C13dist -fit -model=2 -C=16 -p=1.0 -q=0.7 datafile.cL datafile.dL
```

The first command prepares the datafile for the isotopomers containing 6 to 16 ^{13}C. The second command uses this datafile for fitting the distribution to estimate the distribution of isotopic abundance in the highly labeled molecules.

```
C13dist -fit -model=4 -C=16 -dim=14 datafile.c1 datafile.d1
```

This command prepares a histogram of isotopic abundance, datafile.d1, using 14 divisions (discrete model). The number of divisions must be smaller than the value of *n*. Model 1 is outdated.

8. Analysis of combinatomers of GlcDG/MGDG

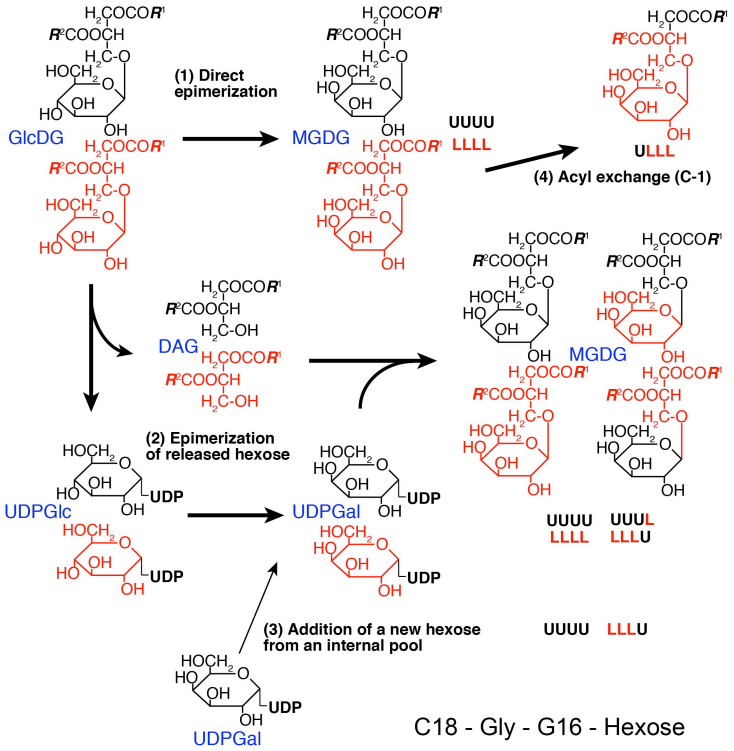
8.1 Notion of combinatomer

GlcDG and MGDG are composed of various parts, such as glycerol, two acyl groups (C18 and C16 at the *sn*-1 and *sn*-2 positions, respectively, in the cyanobacteria like *Anabaena*), and a hexose (glucose and galactose in GlcDG and MGDG, respectively). During the photosynthetic labeling with ^{13}C, each of the parts are labeled at the isotopic abundance p (labeled parts). The preexisting lipids are entirely composed of parts in which p = natural abundance (unlabeled parts). Within the cells, some lipid molecules may be composed entirely of labeled parts. In addition, some molecules are chimeric, in that some parts are labeled but some other parts are unlabeled. Theoretically, there are $2^4 = 16$ types of molecules, according to the combinations of labeled and unlabeled parts. We define these different types of molecules formed by combinations of labeled and unlabeled parts 'isotopic combinatomers'. Note that each part is characterized by its own distribution of isotomopers. We use this notion to describe 'partially labeled' molecules. In the current study, we use four-character representation, such as UUUU for completely unlabeled combinatomer. There four characters indicate, from the left to the right, C18, glycerol, C16, and hexose, respectively.

Assume that the original pool of 34:1-GlcDG contains only the completely labeled combinatomer and the completely unlabeled combinatomer, as a result of photosynthetic labeling with ^{13}C. If the molecule is converted to MGDG without replacing the hexose moiety during the chase period, then the resulting MGDG will contain only completely labeled (LLLL) and completely unlabeled (UUUU) combinatomers. Note that the unlabeled combinatomer is also present before the start of the chase. If, however, the hexose is replaced during the conversion of GlcDG to MGDG by exchange of glucose with galactose, then the incoming galactose might be unlabeled. The removed glucose could be converted to galactose and re-used for incorporation into MGDG, but such sugar must be diluted with large amount of unlabeled hexose pool. As a result, the resultant MGDG pool will contain a significant amount of LLLU combinatomer. If the levels of LLLL and LLLU are compared, we can determine whether the conversion accompanies replacement of the hexose or not. In other words, we can determine whether the conversion of GlcDG to MGDG is a direct epimerization occurring on the lipid.

A similar logic may be used for analyzing the mechanism of fatty acid desaturation during the conversion of 34:1-GlcDG to 34:2-GlcDG or 34:2-MGDG. In this case, the oleic acid at the sn-1 position is converted to linoleic acid. Oleic acid might be removed from the lipid, then desaturated to linoleic acid as a CoA ester, and then re-acylated to form MGDG. Another possibility is the direct desaturation, in which the oleoyl group is desaturated to linoleoyl group on the lipid without deacylation-reacylation. We already demonstrated the direct desaturation of the palmitoyl group at the sn-2 position (Sato et al. 1986), but no definite data are available for the C18 acids at the *sn*-2

position. In the current study, production of a ULLL combinatomer will be a sign of deacylation and reacylation.



8.2 Practical analysis of combinatomers

We developed the **–mgdg** option of the C13dist software. In this mode, two models are available: model 1 uses the isotopomer distribution of GlcDG/MGDG, whereas model 2 uses the MS/MS data of GlcDG/MGDG. In the latter, there types of data are used. In the MS/MS analysis of GlcDG/MGDG, each of the MS peak (corresponding to 44 different isotopomers) was subject to isonization (electron ionization), and the resultant 44 MS spectra were recorded. After correction by the mode 4, all these spectra were combined to form a single 2D spectrum. We focus on the diacylglyceryl peaks (DAG), 1-acylglyceryl peaks (C18Gly) and 2-acylglyceryl peaks (C16Gly), so that we obtain three 2D spectra.

In the **–mgdg** mode, eight files describing the isotopomer distributions of the labeled and unlabeled parts (C18, glycerol, C16, and hexose) are also used to construct isotopomer distributions (1D or 2D) of all possible combinatomers, which are also output to files. These 16 distributions were used to reproduce the actual MS or MS/MS spectra by fitting with the NNLS method.

Factor tables

Fatty acid methyl esters (GC-MS)

16:0-me (base=268)

i	value	*=0.011980
0	0.000000e+00	|
1	1.191232e-02	**
2	9.583678e-01	**********
3	2.557632e-02	***
4	4.143542e-03	|
Total	1.000000	

18:0-me (base=296)

i	value	*=0.012171
0	0.000000e+00	|
1	7.973800e-03	**
2	9.737052e-01	**********
3	1.395825e-02	**
4	4.362728e-03	|
Total	1.000000	

18:1-me (base=264 for M-32 ions)

i	value	*=0.008703
0	0.000000e+00	|
1	9.146565e-03	**
2	6.962625e-01	**********
3	2.860231e-01	**********
4	2.187537e-03	|
5	6.380252e-03	**
Total	1.000000	

18:2-me (base=294)

i	value	+- stdev	*=0.012212	use_std=yes
0	1.308434e-03	1.456610e-03	|	
1	6.734961e-03	4.028627e-03	**	
2	9.769249e-01	5.322489e-03	**********	
3	1.084257e-02	2.319329e-03	**	
4	4.189110e-03	6.427050e-04	|	
5	0.000000e+00	2.524796e-04	|	
Total	1.000000			

16:1-me (base=232 for M-32 ions)

i	value	+- stdev	*=0.008091	use_std=yes
0	0.000000e+00	1.487279e-04	|	
1	9.292414e-03	1.684990e-03	**	
2	6.472810e-01	6.398504e-03	**********	|*|
3	3.345610e-01	6.860431e-03	**********	$|
4	0.000000e+00	2.023168e-03	|	
5	8.865542e-03	4.432158e-04	**	
Total	1.000000			