Mass Distribution Calculation for Isotope-Enriched Macromolecules

John Kuo, Michael May, Michael Gray, C.T. Tan
Quality Control Laboratory
Isotec division, Sigma-Aldrich
Miamisburg, Ohio 45342

Introduction

Molecular mass distribution calculations are well developed for natural abundance compounds. Detailed isotope ratio tables have been compiled and published. However, the published isotope ratio tables are mostly limited to relatively small molecules of natural abundance. In addition to this, the reported molecular ion mass distributions often span only a few atomic mass units, such as M, M+1, and M+2. For highly isotope-enriched compounds, especially biological macromolecules, the complete (molecular ion) mass distribution calculation remains challenging.

It is important for chemists producing isotope-enriched compounds to know their molecular mass distribution, provided the enrichment for each atom is known. If the molecule is relatively small, the calculation is correspondingly simple. For the case of macromolecules such as Insulin, the calculation becomes quite involved. Here, a computer program to estimate molecular mass distributions was developed that utilizes matrix iteration. This program can handle differentially enriched atoms (at different structural positions) and relatively large molecules.

Math Model for Mass Profile Calculation of Natural Abundance Bio-Macromolecules

When molecule size becomes relatively large, such as for bio-macromolecules, the mass distribution profile widens (as observed by mass spectrometry). With present advances in isotope chemistry, isotope-enriched proteins and nucleic acids are becoming possible. To observe the isotope-enriched macromolecule molecular ion and calculate its mass distribution profile present on-going challenges in scientific research.

Assume the empirical formula of protein molecule is given as $C_mH_nN_oO_pS_q$; the molecular mass distribution calculation is then given by polynomial expansion:

$(^{12}C\% + {}^{13}C\%)^m * (^{1}H\% + {}^{2}H\%)^n * (^{14}N\% + {}^{15}N\%)^o * (^{16}O\% + {}^{17}O\% + {}^{18}O\%)^p * (^{32}S\% + {}^{33}S\% + {}^{34}S\% + {}^{36}S\%)^q$

Polynomial expansion yields a total of $(m+1)(n+1)(o+1)(\frac{p^2+3p+2}{2})(\frac{q^3+6q^2+11q+6}{6})$ terms, each of which represents a unique combination of nuclides with unique mass.

Math Model for Mass Profile Calculation of Isotope Enriched Bio-Macromolecules

Isotope-enriched, uniformly labeled bio-macromolecules are becoming technically accessible. Calculation of the molecular ion mass profile for uniformly labeled bio-molecules is similar to that for natural abundance. Again using $C_mH_nN_oO_pS_q$ as an example:

$(^{12}C\% + ^{13}C\%)^m * (^{1}H\% + ^{2}H\%)^n * (^{14}N\% + ^{15}N\%)^o * (^{16}O\% + ^{17}O\% + ^{18}O\%)^p * (^{32}S\% + ^{33}S\%+^{34}S\% + ^{36}S\%)^q$

where each nuclidic atom% enrichment is input. This calculation can equally be applied to isotope-depleted macromolecules, which are also of interest in mass spectrometry.

Consider a bio-molecule with natural abundance C/H/N/O/S at some structural positions, and isotope-enriched $^{13}C/ ^{2}H/ ^{15}N/ ^{18}O$ at other structural positions (we exclude all radioactive labels for simplicity), then the empirical formula can be stated as $C_mH_nN_oO_pS_q{}^*C_{m'}{}^*H_{n'}{}^*N_{o'}{}^*O_{p'}$ and the molecular ion mass distribution can be expressed:

$(^{12}C\% + ^{13}C\%)^m * (^{1}H\% + ^{2}H\%)^n * (^{14}N\% + ^{15}N\%)^o * (^{16}O\% + ^{17}O\% + ^{18}O\%)^p * (^{32}S\% + ^{33}S\%+^{34}S\% + ^{36}S\%)^q *(^{12}\boldsymbol{C}\% + ^{13}\boldsymbol{C}\%)^{m'} * (^{1}\boldsymbol{H}\% + ^{2}\boldsymbol{H}\%)^{n'} * (^{14}\boldsymbol{N}\% + ^{15}\boldsymbol{N}\%)^{o'} * (^{16}\boldsymbol{O}\% + ^{17}\boldsymbol{O}\% + ^{18}\boldsymbol{O}\%)^{p'}$

Polynomial expansion now yields the following number of simplified terms:

$$(m+1)(n+1)(o+1)\left(\frac{p^2+3p+2}{2}\right)\left(\frac{q^3+6q^2+11q+6}{6}\right) * (m'+1)(n'+1)(o'+1)\left(\frac{p'^2+3p'+2}{2}\right)$$

Matrix Bin approach versus Polynomial Expansion

Given a mass spectrometer with unit-amu mass resolution, most expanded terms would effectively become "degenerate", meaning many unique terms share the same nominal mass. One computational approach for simulating the molecular ion mass profile is to "bin" terms of equal nominal mass together. This becomes more involved as exponent indices (m/ n/ o/ p/ q) become large. For example, Bovine Insulin ($C_{254}H_{377}N_{65}O_{75}S_6$) has about 10^{+12} unique nuclidic terms, each with unique mass. In practice, the observable Insulin molecular ion would span about 15 amu. Unless ultrahigh-resolution mass spectrometry is applied (such as FT-ICR-MS), these numerous terms (ie, m/z signals) would be observed as one mass envelope.

One might notice that rigorous polynomial expansion leads to factorial coefficients, which can yield rather large numbers. Approximation techniques to estimate factorial magnitudes (such as those of Stirling and Lanczos) have been utilized in prior algorithms of researchers to compute mass spectrometric profiles. However, the use of such approximations introduces error into simulated molecular ion profiles, particularly at low molecular mass.

Here, we propose an alternate technique in which iterative matrix operation is used instead of factorial calculation. This "matrix bin" approach has been tested on Wintel-class personal computers and found to be adequately fast, even for C_{10000}.

Mass Profile Calculation using Matrix approach

As an initial test case, consider the hydrocarbon C_mH_n. For purposes of discussion, assume that carbon isotope enrichment is 1-atom%-^{12}C and 99-atom%-^{13}C. The algorithm of this (molecular ion) mass distribution calculation involves distribution matrix **A** (nx1), enrichment matrix **B** (1x2), and 2-D probability matrix **C** (nx2). These three matrices are relatively small in size. In this approach, the maximum value of "n" is set to 500.

The core calculation for C_mH_n includes (m+n) iterations, each iteration being matrix multiplication [n x 1] x [1 x 2] followed by a sorted summation step. It has been found that computation is reasonably fast even when (m+n) approaches 10^{+5}.

The first iteration loop starts with multiplication of matrix **A** (nx1) and matrix **B** (1x2); the resultant matrix **C** is then sorted, summed, and written to matrix **A**:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \dots \end{bmatrix} x \begin{bmatrix} 0.01 & 0.99 \end{bmatrix} = \begin{bmatrix} 0.01 & 0.99 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \dots & \dots \end{bmatrix} \rightarrow \begin{bmatrix} 0.01 \\ 0.99 \\ 0 \\ 0 \\ 0 \\ \dots \end{bmatrix}$$

A x **B** = **C** → **A**

Mass Profile Calculation using Matrix approach

The second iteration loop proceeds as follows:

$$\begin{bmatrix} 0.01 \\ 0.99 \\ 0 \\ 0 \\ 0 \\ \dots \end{bmatrix} \quad x \begin{bmatrix} 0.01 & 0.99 \end{bmatrix} \quad = \quad \begin{bmatrix} 0.0001 & 0.0099 \\ 0.0099 & 0.9801 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \dots & \dots \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} 0.0001 \\ 0.0198 \\ 0.9801 \\ 0 \\ 0 \\ \dots \end{bmatrix}$$

A x **B** = **C** → **A**

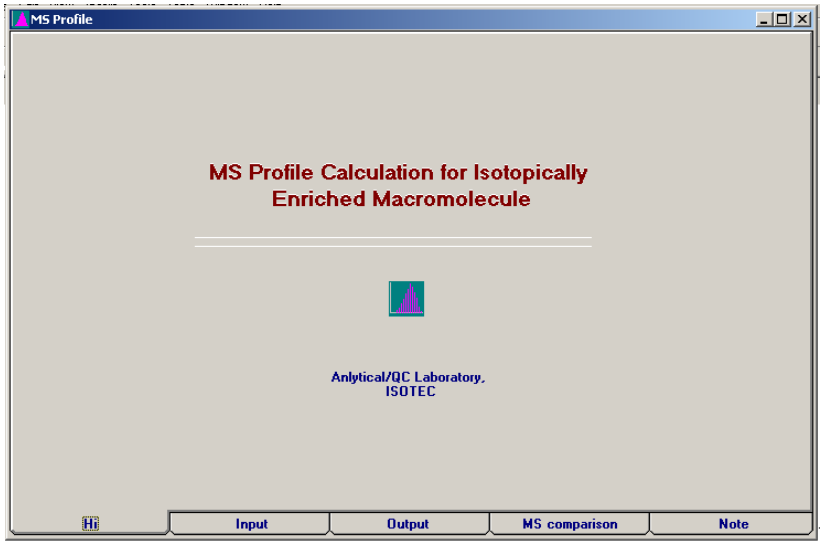
Resultant matrix **C** represents combinatorial probabilities for two carbon positions in C_mH_n, which can be symbolized $^{12}C_{\#1}$-$^{12}C_{\#2}$, $^{12}C_{\#1}$-$^{13}C_{\#2}$, $^{13}C_{\#1}$-$^{12}C_{\#2}$, and $^{13}C_{\#1}$-$^{13}C_{\#2}$, with respective probabilities 0.0001, 0.0099, 0.0099, and 0.9801. Consideration of the resultant matrix leads one to realize that matrix elements linked by back-diagonal lines (top-right to left-bottom) share equal nominal mass. Sorting and summing the simplified terms (of equal nominal mass) yields in-process probabilities for 24-amu, 25-amu, and 26-amu, which respectively are 0.0001, 0.0198, and 0.9801. New matrix **A** (n x 1) is then used in the next loop for carbon atom $C_{\#3}$. This iterative process continues until the loop number equals the number of carbons. Next the analogous technique is applied to Hydrogen H_n.

The matrix bin approach is also useful for bio-macromolecules, which often contain Carbon, Hydrogen, Nitrogen, Oxygen, and Sulfur. Indeed, the computation can become a bit tricky when oxygen, sulfur, chlorine, or bromine are involved. Oxygen has three isotopes (^{16}O/ ^{17}O/ ^{18}O) spanning three amu. Thus oxygen matrix **B** requires dimension (1x3) and matrix **C** (n x 3). Sulfur matrix **B** should have dimension (1 x 5) and matrix **C** (n x 5). Although Chlorine

and Bromine have two major isotopes apiece, they each span 3 amu; therefore, either halogen atom calls for the matrix **B** (1 x 3) and matrix **C** (n x 3). Throughout these calculations the computer power is directed toward matrix iteration (multiply/ sort/ sum/ transfer), while factorial approximations are avoided.

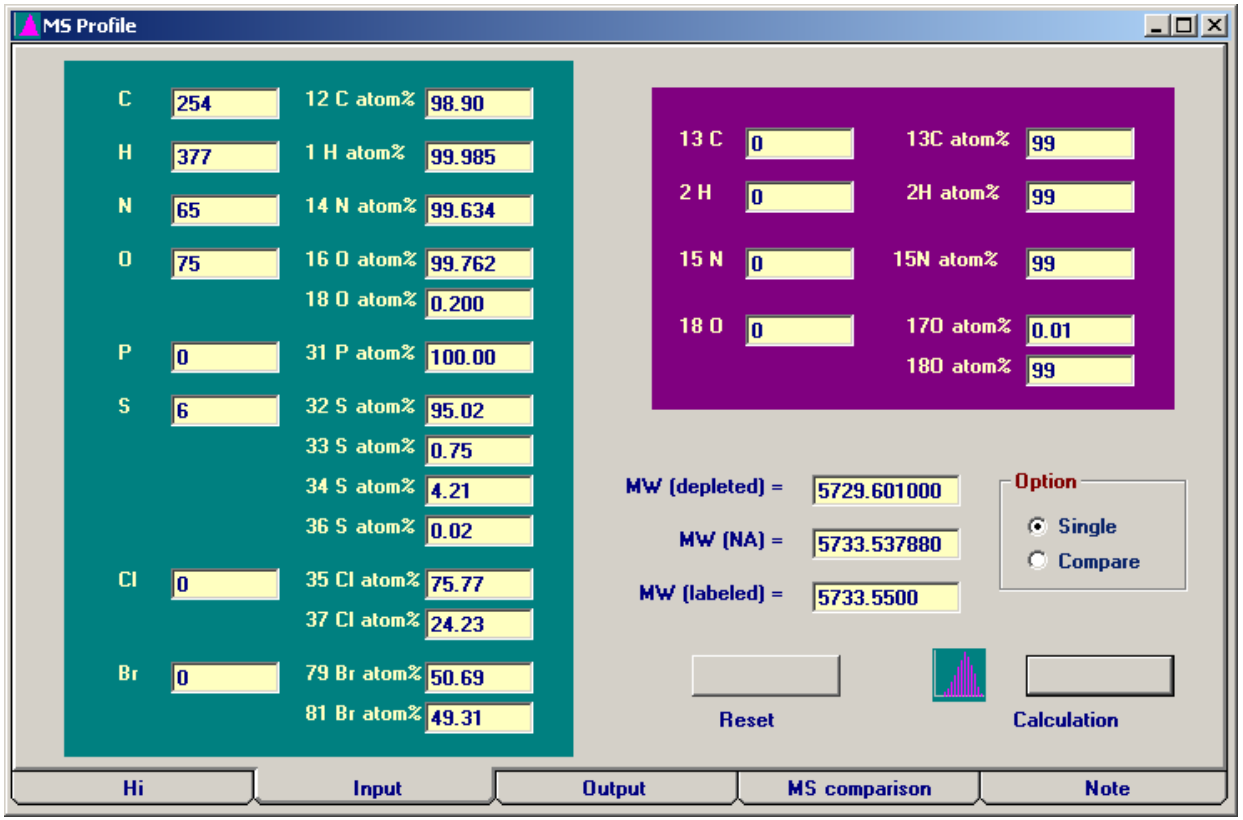
For macromolecules having atoms of the same type (such as Carbon) that are differentially enriched as a function of structural position, the molecular ion mass profile can still be calculated. The main criterion is that isotope enrichment values would need to be known at each structural position.

MS Profile Calculation Program (I)



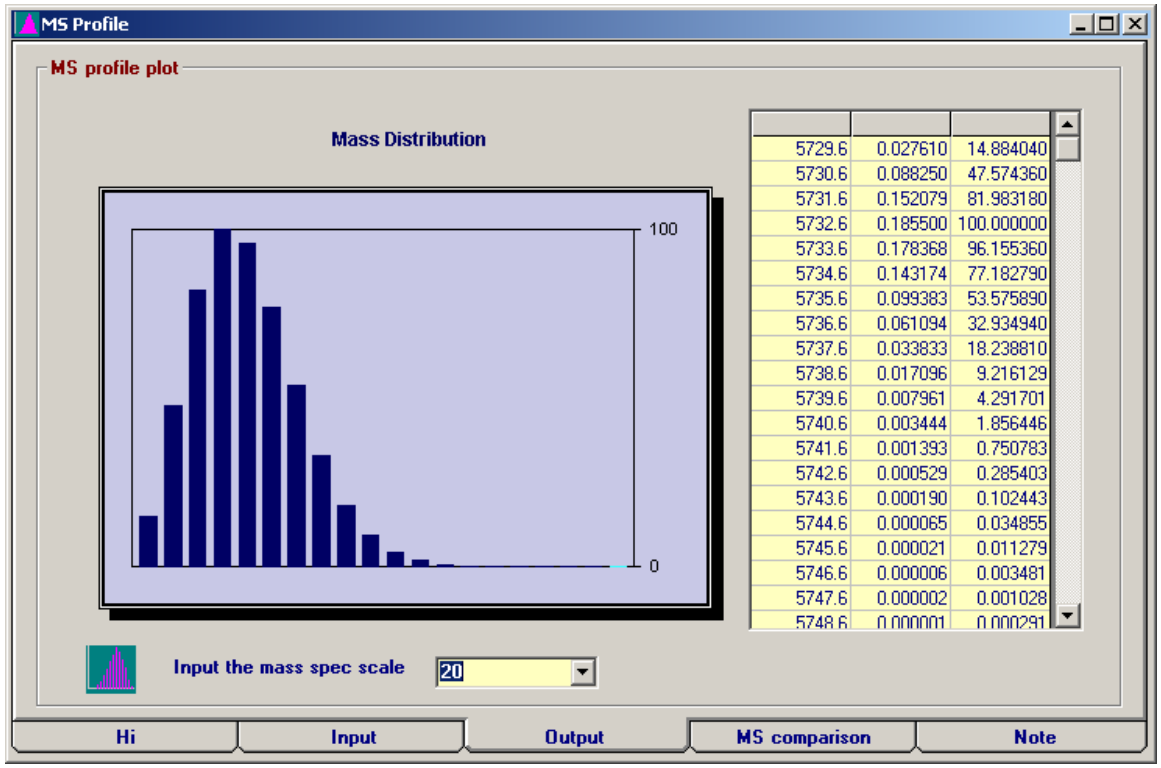
Mass profile program was developed in Visual Basic (v6) using a Wintel-PC.

MS Profile Calculation Program(II)



User input includes stable nuclides of these elements: C/ H/ N/ O/ P/ S/ Cl/ Br.

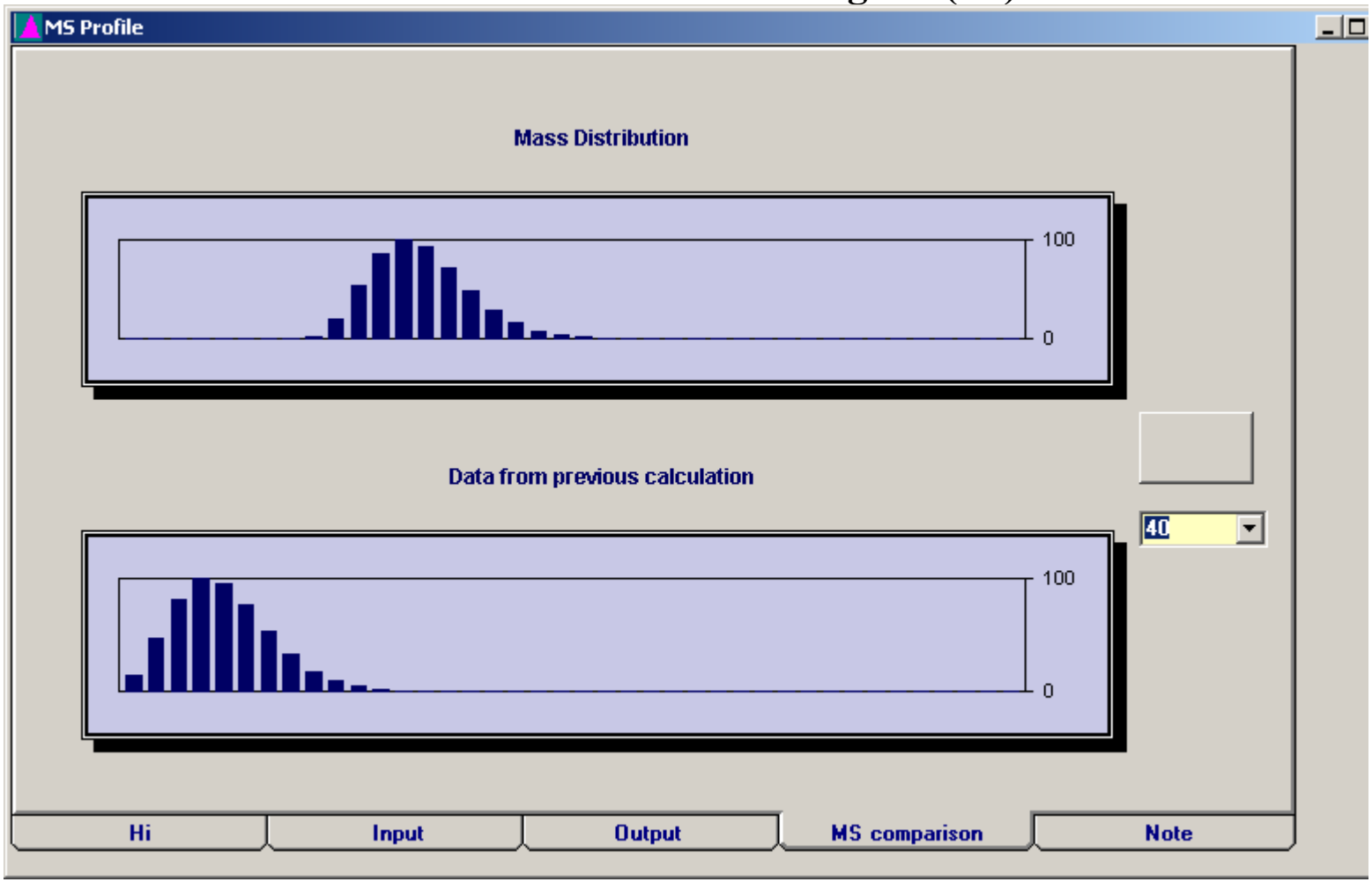
MS Profile Calculation Program(III)



5729.6	0.027610	14.884040
5730.6	0.088250	47.574360
5731.6	0.152079	81.983180
5732.6	0.185500	100.000000
5733.6	0.178368	96.155360
5734.6	0.143174	77.182790
5735.6	0.099383	53.575890
5736.6	0.061094	32.934940
5737.6	0.033833	18.238810
5738.6	0.017096	9.216129
5739.6	0.007961	4.291701
5740.6	0.003444	1.856446
5741.6	0.001393	0.750783
5742.6	0.000529	0.285403
5743.6	0.000190	0.102443
5744.6	0.000065	0.034855
5745.6	0.000021	0.011279
5746.6	0.000006	0.003481
5747.6	0.000002	0.001028
5748.6	0.000001	0.000291

Graphic display of molecular ion profile for Bovine Insulin.

MS Profile Calculation Program(IV)



Comparison of Insulin molecular ion profiles: natural abundance and isotope-enriched.

Results and Discussion

1/ Collecting Simplified-terms from Permutations

Based on the Pascal triangle, expansion of $(a + b)^n$ yields:

$$(a+b)^n = \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} a^i b^{n-i}$$

For binomial expansion, there exist 2^n permutations, which for the case of mass spectrometry can be combined into n+1 simple terms (since permutations such as $^{12}C_{\#1}$-$^{13}C_{\#2}$ and $^{13}C_{\#1}$-$^{12}C_{\#2}$ are mass equivalent).
Trinomial expansion of $(a + b + c)^n$ yields:

$$(a+b+c)^n = \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} a^i \sum_{j=0}^{n-i} \frac{(n-i)!}{j!(n-i-j)!} b^j c^{n-i-j}$$

For this case, there exist 3^n possible permutations. Using combinatorial algebra, it can be shown that the expression above gives $(\frac{n^2+3n+2}{2})$ simplified terms.

Polynomial expansion of $(a + b + c + d)^n$ generates 4^n permutations and $\left(\frac{n^3 + 6n^2 + 11n + 6}{6}\right)$ simplified terms:

$$(a+b+c+d)^n = \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} a^i \sum_{j=0}^{n-i} \frac{(n-i)!}{j!(n-i-j)!} b^j \sum_{k=0}^{n-i-j} \frac{(n-i-j)!}{k!(n-i-j-k)!} c^i d^{n-i-j-k}$$

For Glucagon ($C_{153}H_{224}N_{42}O_{50}S$), the number of simplified terms can thus be calculated:

$$(153+1)*(224+1)*(42+1)*[0.5*(50^2 + 3*50 + 2)]*(4) = 7.9 \text{ x } 10^{+9}$$

For Bovine Insulin ($C_{254}H_{377}N_{65}O_{75}S_6$) the number of simplified terms is given below:

$$(254+1)*(377+1)*(65+1)*[0.5*(75^2 + 3*75 + 2)]*\left(\frac{6^3 + 6*6^2 + 11*6 + 6}{6}\right) = 1.6 \text{ x } 10^{+12}$$

As macro-molecules become large like Insulin, it becomes challenging to store all permutations and simplified-terms. To simulate molecular ion profiles with 1-amu resolution, the present matrix approach is advantageous.

2/ Insulin result from Matrix Iteration program versus prior Computer models

Mass	Literature 1	Literature 2	Matrix iteration
5729.6	14.84	14.56	14.88
5730.6	47.51	46.98	47.57
5731.6	81.91	81.47	81.98
5732.6	100	100	100
5733.6	96.15	96.48	96.16
5734.6	76.97	77.58	77.18
5735.6	53.03	53.59	53.58
5736.6	32.09	32.38	32.93
5737.6	17.33	17.37	18.24
5738.6	8.43	8.00	9.21
5739.6	3.56	3.04	4.29
5740.6	1.30	0.89	1.86
5741.6	0.15		0.75
5742.6			0.28

The computer results are comparable to each other. Minor peak intensity differences can be attributed to the specifics of each algorithm. For Matrix Iteration, any probability term $<10^{-20}$ is treated as zero, which might lead to minor errors in the overall computation.

References

1/ D.B. Hibbert, *Chemometrics and Intelligent Laboratory Systems*, 6, 203-212 (1989).
2/ J.A. Yergey, *International Journal of Mass Spectrometry and Ion Physics*, 52, 337-349 (1983).
3/ M.L. Brownawell, J. San Filippo, *Journal of Chemical Education*, 59, 663-665 (1982).
4/ J.L. Margrave, R.B. Polansky, *Journal of Chemical Education*, 39, 335-337 (1962).
5/ Nuclides and Isotopes, 15th edition, General Electric and KAPL, 1996.
6/ Professor Steen Pedersen, Wright State University, private communication (2003).