VERSITA

Cent. Eur. J. Biol. • 5(2) • 2010 • 190–196
DOI: 10.2478/s11535-009-0069-2

Central European Journal of **Biology**

Early life relict feature in peptide mass distribution

Research Article

Roman A. Zubarev*, Konstantin A. Artemenko, Alexander R. Zubarev, Corina Mayrhofer, Hongqian Yang, Y.M. Eva Fung

Division of Molecular Biometry, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, S-171 77 Stockholm, Sweden

Received 16 June 2009; Accepted 6 October 2009

Abstract: The molecular mass of a biomolecule is characterized by the monoisitopic mass M_{mono} and the average isotopic mass M_{av}. We found that tryptic peptide masses mapped on a plane made by two parameters derived from M_{mono} and M_{av} form a peculiar feature in the form of a 'band gap' stretching across the whole 'peptide galaxy', with a narrow line in the centre. The purpose of this study was to investigate possible reasons for the emergence of such a feature, provided it is not a random occurrence. The *a priori* probability of such a feature to emerge by chance was found to be less than 1:100. Peptides contributing to the central line have elemental compositions following the rules $S = 0$; $Z = C - (N + H)/2 = 0$, which nine out of 20 amino acid residues satisfy. The relative abundances of amino acids in the peptides contributing to the central line correlate with the consensus order of emergence of these amino acids, with ancient amino acids being overrepresented in on-Line peptides. Since linear correlation between M_{av} and M_{mono} reduces the complexity of polypeptide molecules, and the turnover rate of less complex molecules should be faster in non-equilibrium abiotic synthesis, we hypothesize that the line could be a signature of abiotic production of primordial biopolymers. The linear dependence between the average isotopic masses and monoisotopic masses may have influenced the selection of amino acid residues for terrestrial life.

Keywords: *Molecular mass • Mass spectrometry • Origin of life • Abiotic production*

1. Introduction

Molecular mass (MM) of a biopolymer is a distribution containing information on the exact masses and abundances of all stable isotopes of the constituent elements. For polypeptides, the five constituent elements C, H, O, N and S encompass 12 stable isotopes, the five lightest isotopes being dominant. The full description of the peptide MM is thus a combination of the five variable coefficients of a brutto-formula (*c, h, o, n* and *s*), 12 constant isotope masses and seven relative abundances of less abundant isotopes, which are variable to some extent. In biomolecular mass spectroscopy, the traditional approach is the assumption of constant isotopic compositions. MM of a biopolymer is then characterized by two constant values, the monoisotopic mass M_{mono} and the average isotopic mass M_{av}. Reduction of the dimensionality to two affords easy visualization on a two dimensional (2D) plot using derivatives of the two mass values as axes [1]. One convenient mass derivative is the monoisotopic mass defect,

$$\Delta M_m = M_{mono} - M_{nom}, \qquad \text{(Eq. 1a)}$$

where M_{nom} is the nominal (integer) mass (e.g. $^{14}N = 14$, $^{32}S = 32$, *etc.*). ΔM_m is related to the binding energy of the nucleons in the atomic nucleus, and for each element it is a strict constant in the whole universe. The other derivative is the isotopic mass shift [2]

$$\Delta M_{is} = M_{av} - M_{mono} \qquad \text{(Eq. 1b)}$$

determined by the relative abundances of the less abundant isotopes. Since heavier isotopes can be enriched or depleted in physico-chemical processes,

* *E-mail: roman.zubarev@ki.se*

ΔM_{is} is a constant only approximately. ΔM_{is} values are tabulated, and biogenic elements C, H, N and O are constant within 1-3% (with bigger variations for H) for most terrestrial organic molecules. To eliminate the mass dependence, introduce

$$NMD = 1000 \cdot \Delta M_m / M_{nom} \quad \text{(Eq. 2a)}$$

$$NIS = 1000 \cdot \Delta M_{is} / M_{nom} \quad \text{(Eq. 2b)}$$

where NMD [‰] is the normalized monoisotopic defect and NIS [‰] is the normalized isotopic shift, respectively. NMD and NIS can be considered independent mass parameters for peptides with arbitrary amino acid compositions.

In our previous publication [1], we introduced 2D mass mapping as a visualization method for moderately complex peptide mixtures that retains and reveals information on the peptide families, physico-chemical properties and even their biological function. However, mapping theoretical NMD and NIS values of large number of tryptic peptides from a typical proteome-wide analysis was expected to produce a more or less homogeneous distribution modulated by CH_2 and other groups common for peptides [3]. Instead, the map contained an unexpected global feature in the form of a 'band gap' across the 'peptide galaxy', with a narrow line in the centre (Figure 1). Below we report on the investigation of the origin of this feature and its possible causes.

2. Results and Discussion

2.1 Feature appearance

The 2D map of theoretical masses of 3600 tryptic peptides detected in a proteomics analysis of a mouse kidney is shown in Figure 1. The best fit to the central Line of a linear equation:

$$NIS = a*NMD + b \quad \text{(Eq. 3)}$$

gave a=0.350574 and b=0.395622‰ with R^2 =0.99999. The distance d from the dots to the Line,

$$d = (NIS - a \cdot NMD - b)/(1 + a^2)^{0.5}, \quad \text{(Eq. 4)}$$

is shown in Figure 2, with the Line now represented by a peak at d=0 ppm. No dots outside the central peak were found in the region between -3.0 ppm and +2.5 ppm.

2.2 Origin of the line

The molecules on the Line have elemental compositions obeying the rule:

$$s = 0; \quad h = 2c - n. \quad \text{(Eq. 5)}$$

For sulphur-free peptides (*i.e.* no Met or Cys), the factor

$$z = c - (n + h)/2 \quad \text{(Eq. 6)}$$

reflects the deviation of the elemental composition from the rule (Eq. 5). Unlike both NMD and NIS, z is an additive parameter, z[A+B] = z[A] + z[B]. The position on the 2D mass map depends upon the z-value: molecules

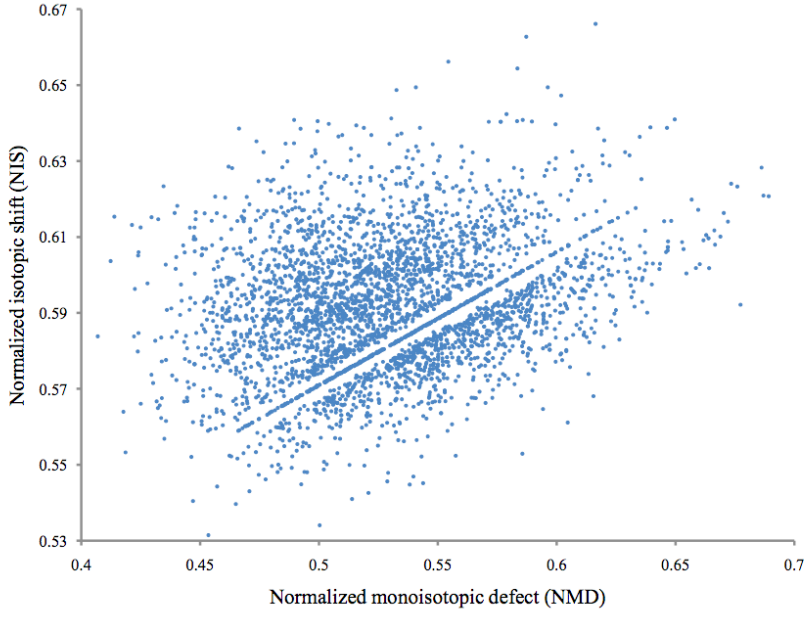

Figure 1. Map of theoretical mass values of 3600 tryptic peptides from mouse kidney identified in a shotgun proteomics experiment.

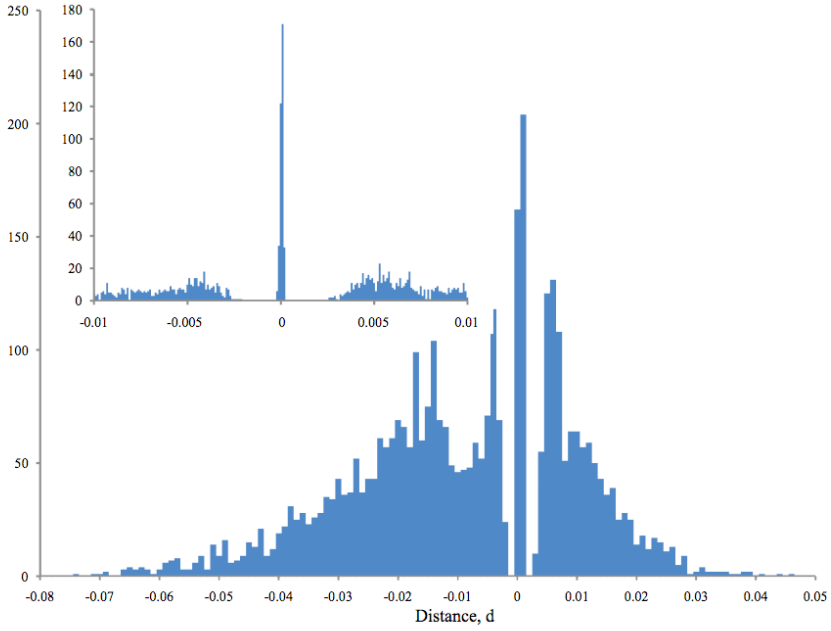

Figure 2. Distances in ‰ of mass dots in Figure 1 to the line feature.

with *z<0* lie under the Line, while those with *z>0* occupy the region above the Line. Note that the *z*-value is related to the well-known Ring plus Double Bond index (RDB) as

$$RDB = z + n. \quad \text{(Eq. 7)}$$

2.3 Discrete mass distributions

While masses of peptides with *z=0* are mapped onto the Line, those with *z=±1* contribute to broader distributions on both sides of it (Figure 3), and the peptides with *z=±2* form even broader distributions farther away. The broadness of distributions with *z≠0* is due to the different masses of the contributing peptides: for the same *z*, molecules with lower masses are found farther away from the Line.

2.4 Origin of the gap

The discrete nature of *z* is the reason for the gap between the Line and neighbouring points with *z=±1*. The peptide masses with *z=n* can be focused into the Line by adding/subtracting to/from the elemental formula nH_2O_x, as shown in Figure 3. The number of oxygens *x* can be arbitrary, and will only affect the position on the Line. Note that the addition/subtraction of nH_2O_x does not alter the Line position, and only moves the peptide mass distribution on the 2D plot.

2.5 Analytical content of z

According to Table 1, nine out of eighteen sulphur-free residues have *z=0, i.e.* obey the rule [5]. Acidic residues Glu and Asp have a *z=1*. Since only basic Lys and Arg have negative *z*-values (*z≥-2*), peptides with *z<-3* must contain more than one basic residue. Thus, if these peptides are produced by trypsinolysis, *z<-3* means at least one missed cleavage. Therefore, *z*-values contain important analytical information.

Alanine, C_3H_5NO	0	4.0
Arginine, $C_6H_{12}N_4O$	-2	11.0
Asparagine, $C_4H_6N_2O_2$	0	11.3
Aspartic Acid, $C_4H_5NO_3$	1	6.0
Glutamic Acid, $C_5H_7NO_3$	1	8.1
Glutamine, $C_5H_8N_2O_2$	0	11.4
Glycine, C_2H_3NO	0	3.5
Histidine, $C_6H_7N_3O$	1	13.0
Isoleucine, Leucine $C_6H_{11}NO$	0	11.4, 9.9
Lysine, $C_6H_{12}N_2O$	-1	13.3
Phenylalanine, C_9H_9NO	4	14.2
Proline, C_5H_7NO	1	7.3
Serine, $C_3H_5NO_2$	0	7.6
Threonine, $C_4H_7NO_2$	0	9.4
Tryptophan, $C_{11}H_{10}N_2O$	5	16.5
Tyrosine, C_9H_9NO	4	15.2
Valine, C_5H_9NO	0	6.3

Table 1. Sulphur-free amino acid residues, their z values and the consensus order of emergence [6].

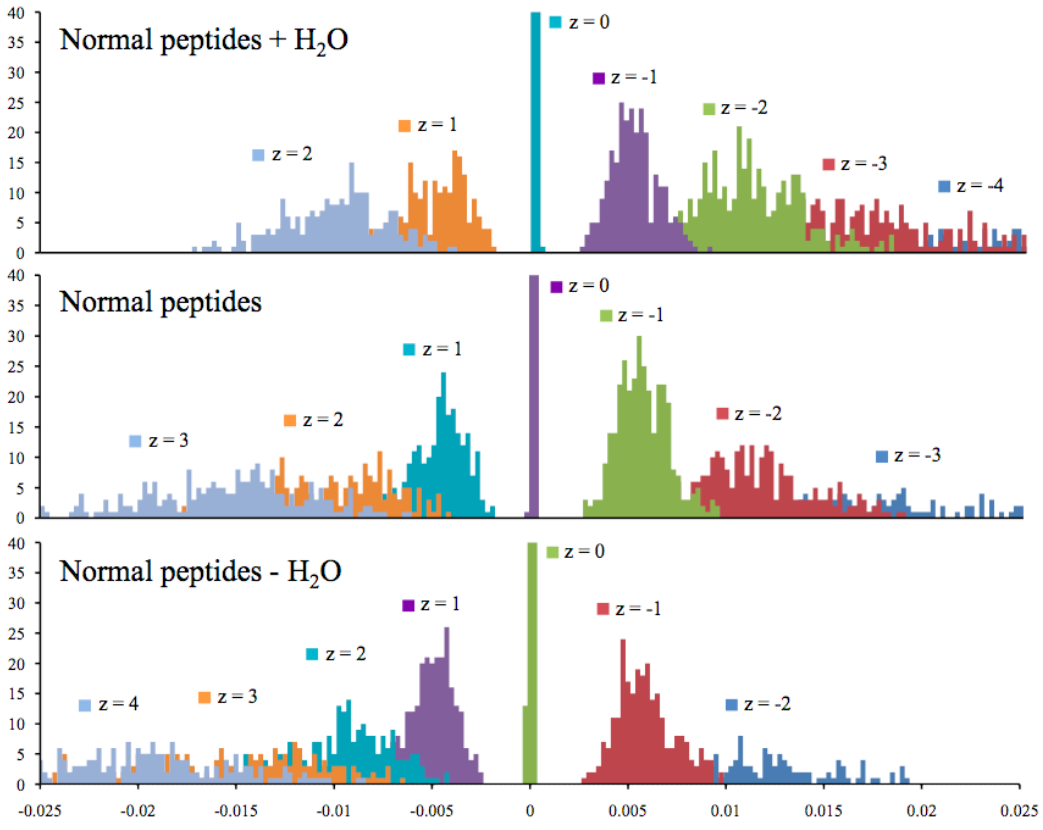

Figure 3. Distances in ‰ from the line feature.

The *z*-value of any peptide (or indeed, any molecule containing only C, H, N and O) can be determined by means of high-accuracy mass spectrometry by adding/subtracting the masses of zH_2O_x until the summed masses approach the Line, *i.e.* by minimization of the distance *d*. The *z* value represents a *class* of elemental compositions and thus it is in principle more easily determined than the elemental composition itself. Such determination requires <5 ppm mass accuracy in both M_{av} and M_{mono}, which is within the reach of modern mass spectrometry.

2.6 Linear relationship between M_{av} and M_{mono}

The reason for the focusing of the molecules with *z=0* into the Line is the peculiar relationship between the isotopic shifts and mass defects of the elements C, H, N and O. Indeed,

$$NMD = 1000\cdot(D_c\cdot c + D_h\cdot h + D_n\cdot n + D_o\cdot o)/ /(12\cdot c + h + 14\cdot n + 16\cdot o) \quad \text{(Eq. 8a)}$$

and

$$NIS = 1000\cdot(I_c\cdot c + I_h\cdot h + I_n\cdot n + I_o\cdot o)/ /(12\cdot c + h + 14\cdot n + 16\cdot o), \quad \text{(Eq. 8b)}$$

where D_x and I_x are atomic mass defects and elements' isotopic shifts, respectively.

Combination of (Eq. 3) and (Eq. 5) is equivalent to (D_c=0 in the scale of ^{12}C=12.000):

$$c(I_c + 2I_h - 2aD_h - 14b^*) + n(I_n - 13b^* - I_h - aD_n + aD_h) + o(I_o - aD_o - 16b^*) = 0, \quad \text{(Eq. 9)}$$

where $b^* = b/1000$.

Since *c, o* and *n* are arbitrary numbers, (Eq. 9) is equivalent to the system of three equations:

$$I_c + 2I_h - 2aD_h - 14b^* = 0 \quad \text{(Eq. 10a)}$$

$$I_n - 13b^* - I_h - aD_n + aD_h = 0 \quad \text{(Eq. 10b)}$$

$$I_o - aD_o - 16b^* = 0 \quad \text{(Eq. 10c)}$$

Using the NIS values [4] D_h = 0.007825032, D_n = 0.0030740053, I_c = 0.001078, I_h = 0.00012197,

and I_n = 0.003646, obtain from (Eq. 10a-b): a = 0.3481012, b = 0.3982968‰. Note that (Eq. 10c) is independent from (Eq. 10a-b) and contains only values related to oxygen. Knowing that D_o = -0.005085378 and using the values of a and b from (Eq. 10a-b), finds the 'resonance' value of $^{\#}I_o = aD_o + 16b^* = 0.004603$. When this value is used, the Line becomes a mathematically ideal line, and the central peak in Figure 2 becomes infinitely thin.

2.7 Terrestrial isotopic abundances are close to resonance values

The table value of I_o is 0.004515, *i.e.* just 19.5‰ (≈2%) off the resonance value. Since most of the isotopic shift of oxygen comes from ^{18}O, the resonance and table abundances of this isotope differ by less than 10‰. Such a difference is within the range of natural variations: e.g. oxygen in ambient air is enriched by 23.5‰ [5]. Instead of I_o, the value of I_n could be adjusted downwards by 2% to achieve the resonance, which is also within the natural isotopic abundance variation. Alternatively, I_c could be lowered by 22% or I_h increased by 60%.

2.8 p-value of the Line

The linear correlation between the isotopic abundances of C, H, N and O and their respective monoisotopic mass defects for molecules with z=0 holds true irrespective of the mass standard. Changing the I_o value in the range from 0.001000 to 0.010000 produces a plurality of resonance I_o values that create linear features on a 2D mass map, but no feature is comparable in size and prominence (number of participating peptides) with the one in Figure 1. Thus the *a priori* probability for the Line to emerge by pure chance is less than (0.004603 - 0.004515)/(0.010000 - 0.001000), *i.e.*<0.01.

2.9 Ancient amino acids in on-Line peptides

Not all amino acids are proportionally represented in on-Line peptides. As expected, amino acids with z=0 (Table 1) are overrepresented, while sulphur-containing residues and aromatic residues Tyr, Phe and Trp that have large positive z-values are practically absent. Trifonov has reviewed a bulk of origin-of-life hypotheses and determined the consensus order of amino acid emergence [6]. We found an overall correlation (P<0.06) between the consensus ranking of a residue and the probability to be found in on-Line peptides (Figure 4). This finding means that the more ancient an amino acid is, the more likely it is to be found in on-Line peptides.

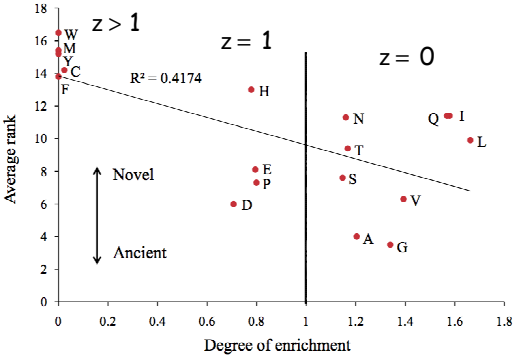

Figure 4. Correlation between the degree of enrichment of amino acids in on-Line peptides and their consensus average rank of emergence in [6]. The enrichment degree is calculated as (NL/NT)·(PT/PL), where NL is the number of amino acids of a particular type in on-Line peptides, NT is the number of these amino acids in all peptides, PT is the total number of amino acids in all peptides and PL is the total number of amino acids in on-Line peptides.

2.10 Back in time the Line was more prominent

On average, the nine amino acids with z=0 (Table 1) are overrepresented in on-Line peptides by 23%. Their average consensus ranking is 8.3, as opposed to 11.6 of the other nine non-sulphur residues. Today, the z=0 residues account for 57% of the amino acid abundance in natural proteins. In the course of evolution, as was discovered by Zuckerkandl *et al.* [7] and recently by Jordan *et al.* [8], the 'ancient' amino acids which tend to have low z values were consistently lost, being replaced by 'novel' amino acids with high z-values. Thus when life first emerged, the relative abundance of low-z amino acids was much higher than today, possibly as high as 100%. But even today peptides with small z comprise the majority of tryptic sequences (Figure 3). The ancient polypeptides must be much more concentrated on and around the Line. Therefore, the Line is a relic of the time when life was emerging on Earth and was employing a limited set of ancient amino acids. Note that this conclusion is valid *regardless* of whether or not the linear correlation between M_{av} and M_{mono} is spurious.

2.11 Low-z amino acid were selected for terrestrial life?

One of the questions intimately related to the origin of life is why out of the 500 different amino acids abiotically produced in the seminal experimental series started by Miller [9], only ten (G, A, V, L, I, S, T, P, D, and E) are found among the 20 amino acids common in terrestrial biological systems [6]. Out of these ten primary amino acids, the first seven have z=0 as residues, and the remaining three have z=0 in the free form (*i.e.* upon

water addition). Interestingly, ribose $C_5H_{10}O_5$ which is the basis of RNA *and* has plausible prebiotic syntheses, also has z=0. Thus there may be, besides empirical, also a causal link between having z=0 and being selected as a basis of terrestrial life. To reveal this link, one has to find a plausible basis for the selection rule.

2.12 Complexity reduction may be a crucial factor in life emergence

A useful analogue in this case is the 'slope = 1' line in mass-independent fractionation (MIF) reactions, in which the degrees of enrichment/depletion of several isotopes of the same element (e.g. oxygen or sulphur) are independent upon the isotopic masses [10]. The 'slope =1' means a significant complexity reduction in the reacting system, as the isotopic masses disappear from the kinetic equation. Gao and Marcus have explained MIF by quantum-mechanical effects involving reduction in the number of quantum-mechanical states due to the symmetry of some isotopomers [11]. Similarly, in our case the peptides with z=0 are characterized by significant complexity reduction and thus by a decreased number of quantum-mechanical states. Indeed, while in general M_{av} is defined by 14 parameters (four monoisotopic masses and five masses of isotopes and their five relative abundances), the presence of a strict linear relationship between M_{av} and M_{mono} reduces the number of parameters to just six (four monoisotopic masses and two coefficients of the linear equation). Therefore, in analogy with MIF phenomena where 'slope=1' reactions have much higher rates than conventional equilibrium fractionation reactions, the presence of the Line could accelerate certain reactions involving molecules with z=0. MIF is usually observed in non-equilibrium processes involving photo-, electronic excitation or high temperatures [10]; abiotic synthesis of amino acids and other building bocks of life is thought to involve similar mechanisms [12].

2.13 Life emergence hypothesis

Thus one can hypothesize that the choice of amino acids for terrestrial life has been affected by the isotopic abundances of biogenic elements C, H, N, and O: amino acids and peptides with z=0 where preferred. Note that within the solar system, isotopic abundances of biogenic elements are similar to terrestrial values, while for objects originating from the outside space (e.g. some comets), these values may differ by a factor of two or more [13]. For instance, the ratio $^{12}C/^{13}C$ is 92 on Earth and ≈20 in the Galactic centre [14]. In general, terrestrial isotopic abundances are atypical for most of our Galaxy [14]. In view of the above hypothesis, life should preferentially emerge at conditions when abiotic production of amino acids is facilitated by the presence of a resonance between monoisotopic and average isotopic masses, *i.e.* Line on a mass plane. We found that a deviation of isotopic compositions of either nitrogen or oxygen (but not carbon) by a few percent compared to the terrestrial levels destroys the Line feature. Therefore, in solar systems of our Galaxy where isotopic compositions of these elements seem to differ significantly from terrestrial levels, spontaneous emergence of life should be unlikely.

3. Conclusions

Irrespective of the actual reasons for the empirically observed strong linear correlation between the average and monoisotopic masses of terrestrial molecules composed on the biogenic elements C, H, N and O and following the rule $z=c-(n+h)/2=0$, early life was based on polypeptides with masses concentrated on and around the central Line. Thus the linear feature can be viewed as a relic of early life. If the reason behind the linear relationship between the monoisotopic and average masses is not spurious, but relates to the complexity reduction for efficient abiotic material production, life in its known form has probably emerged either in the Solar system or in an environment with very similar to terrestrial isotopic abundances of biogenic elements, which rules out most of the Galaxy. The increased yield of abiotically produced amino acids in the presence of a Line-type relationship between masses of elements represents a falsifiable hypothesis that can be tested in a Miller type of experiment.

Acknowledgements

This work was supported by the Knut and Alice Wallenberg Foundation, European Union (consortium EPITOPES) as well as the Swedish research council (grant 621-2007-4410).

References

[1] Artemenko K.A., Zubarev A.R., Samgina T.Yu., Lebedev A.T., Savitski, M.M., Zubarev R.A., 2D Mass Mapping as a General Method of Data Representation in Comprehensive Analysis of Complex Molecular Mixtures, Anal. Chem., 2009, 81, 3738-3745

[2] Zubarev R.A., Some Second-order Errors in Average Mass Measurements of Complex Biological Molecules in Time-of-flight Particle Desorption Mass Spectrometry, Int. J. Mass Spectrom. Ion Processes, 1991, 107, 17-27

[3] Demirev P.A., Zubarev R.A., Probing Combinatorial Library Diversity by Mass Spectrometry, Anal. Chem., 1997, 69, 2893-2900

[4] http://physics.nist.gov/cgi-bin/Compositions/stand_alone.pl?ele=&ascii=html&isotype=some, downloaded 07/29/2008.

[5] Kroopnick P., Craig H., Atmospheric Oxygen; Isotopic Composition and Solubility Fractionation, Science, 1972, 175, 54-55

[6] Trifonov E.N., The Triplet Code from First Principles, J. Biomol. Struct. Dynamics, 2004, 22, 1-11

[7] Zuckerkandl E., Derancourt J., Vogel H., Mutational trends and random processes in the evolution of informational macromolecules, J. Mol. Biol., 1971, 59, 473-490

[8] Jordan I.K., Kondrashov F.A., Adzhubei I.A., Wolf Y.I., Koonin E.V., Kondrashov A.S., et al., A universal trend of amino acid gain and loss in protein evolution, Nature, 2005, 433, 633-638

[9] Miller S.L., A Production of Amino Acids Under Possible Primitive Earth Conditions, Science, 1953, 117, 528-529

[10] Thiemens M.H., Heidenreich J.E. III, The mass-independent fractionation of oxygen: a novel isotope effect and its possible cosmochemical implications, Science, 1983, 219, 1073-1075

[11] Gao Y.Q., Marcus R.A., Strange and Unconventional Isotope Effects in Ozone Formation, Science, 2001, 293, 259-263

[12] Bada J.L., How life began on Earth: a status report, Earth Planet. Sci. Lett., 2004, 226, 1-15

[13] Wasserburg G.J., Busso M., Gallino R., Nollett K.M., Short-lived nuclei in the early Solar System: Possible AGB sources, Nucl. Phys. A, 2006, 777, 5-69

[14] Wielen R., Wilson T.L., The evolution of the C, N, and O isotope ratios from an improved comparison of the interstellar medium with the Sun, Astron. Astrophys., 1997, 326, 139-142