Early Life Relict Feature in Peptide Mass Distribution

Roman A. Zubarev*, Konstantin A. Artemenko, Alexander R. Zubarev,

Corina Mayrhofer & Y. M. Eva Fung

Molecular Biometry, Department for Cell and Molecular Biology, Uppsala University, Box 596, SE-75 124 Uppsala, Sweden

ABSTRACT

Molecular mass of a biomolecule is characterized in mass spectroscopy by the monoisitopic mass M_{mono} and the average isotopic mass M_{av}. We found that peptide masses mapped on a plane made by two parameters derived from M_{mono} and M_{av} form a peculiar global feature in form of a 'band gap' 5-7 ppm wide stretching across the whole 'peptide galaxy', with a narrow (FWHM ≈0.2 ppm) 'line' in the centre. The *a priori* probability of such a feature to emerge by chance is less than 1:100. Peptides contributing to the central line have elemental compositions following the rules S=0; Z = (2C - 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. Thus the central line is a relic of ancient life, and likely a signature of its emergence in abiotic synthesis. The linear correlation between M_{av} and M_{mono} reduces the complexity of polypeptide molecules, which may have increased the rate of their abiotic production. This, in turn may have influenced the selection of these amino acid residues for terrestrial life. Assuming the line feature is not spurious, life has emerged from elements with isotopic abundances very close to terrestrial levels, which rules out most of the Galaxy.

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 also variable. In biomolecular mass spectroscopy, the traditional approach is the assumption of constant isotopic compositions. MM of a biopolymer is then considered to have two constant values, the monoisotopic mass M_{mono} and the average isotopic mass M_{av}. Reduction of dimensionality to two affords easy visualization on a two dimensional (2D) plot using derivatives of the two mass values as axes. One convenient mass derivative is the monoisotopic mass defect,

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

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. The other derivative is the isotopic mass shift

$$\Delta M_{is} = M_{av} - M_{nom} \qquad (1b)$$

determined by the relative abundances of the less abundant isotopes. Since heavier isotopes can be enriched or depleted in physico-chemical process, ΔM_{is} is a constant only approximately. ΔM_{is} values are tabulated, and for biogenic elements C, H, N and O are constant within 1-3% (with a bigger variation for H) for most terrestrial organic molecules. To eliminate the mass dependence, introduce

$$NMD = 1000 \bullet \Delta M_m / M_{nom} \qquad (2a)$$

$$NIS = 1000 \bullet \Delta M_{is} / M_{nom} \qquad (2b)$$

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.

Mapping theoretical NMD and NIS values of 3600 tryptic peptides revealed an unexpected global feature in form of a 'band gap' across the 'peptide galaxy', with a narrow line in the centre (Figure 1). The best fit to the Line of a linear equation:

$$NIS = a*RMD + b \quad (3)$$

gave a = 0.350574 and b = 0.395622 ‰ with R^2=0.99999. The distance d from the dots to the Line is shown in Figure 2, with the Line now represented by a peak at d = 0 ppm. No dots outside the central peak are found in the region between -3.0 ppm and +2.5 ppm. The molecules on the Line have elemental compositions obeying the rule:

$$s = 0; \quad h = 2c - n. \quad (4)$$

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

$$z = (2c - n - h)/2 \quad (5)$$

reflects the deviation of the elemental composition from the rule (4). 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 with $z<0$ lie under the Line, while those with $z>0$ occupy the region above the Line.

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), with the peptides with z=±2 forming even broader distributions farther away. 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 focussed into the Line by adding/subtracting nH_2O_x to the elemental formula, 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 $n\mathrm{H_2O}_x$ does not alter the Line position, and only moves the peptide masses on the 2D plot.

According to Table 1, nine out of 18 sulfur-free residues have z =0, i.e. obey the rule (4). Acidic residues Glu and Asp have a z = 1. Since only basic Lys and Arg have negative z-values, peptides with $z<0$ *must* contain more than one basic residue. Thus, if these peptides are produced by trypsinolysis, $z<0$ means at least one missed cleavage. Therefore, z-values contain important analytical information. The z-value of any peptide (or indeed, any molecule containing only C, H, N and O) can be estimated (e.g. by means of mass spectrometry) its distance from the Line, and determined *exactly* by adding/subtracting the masses of $z\mathrm{H_2O}_x$ until the summed mass fits the Line. The z value represents a *class* of elemental compositions and thus it is more easily determined than the elemental composition itself. Such determination requires 1-2 ppm mass accuracy in both $\mathrm{M_{av}}$ and $\mathrm{M_{mono}}$ irregardless of the molecular mass.

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,

$$\mathrm{NMD} = 1000\cdot(D_c\cdot c + D_h\cdot h + D_n\cdot n + D_o\cdot o)/(12\cdot c + h + 14\cdot n + 16\cdot o) \quad (6a)$$

and

$$\mathrm{NIS} = 1000\cdot(I_c\cdot c + I_h\cdot h + I_n\cdot n + I_o\cdot o)/(12\cdot c + h + 14\cdot n + 16\cdot o), \quad (6b)$$

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

Combination of (3) and (4) is equivalent to (D_c=0 in the scale of ^{12}C=12.000):

$$c(I_c + 2I_h - 2aD_h - 14b^*) + n(I_n - 13b^* - I_h - aD_n + aD_h) + o(I_o - aD_o - 16b^*) = 0, \quad (7)$$

where $b^* = b/1000$.

Since c, o and n are arbitrary numbers, (7) is equivalent to the system of three equations:

$$I_c + 2I_h - 2aD_h - 14b^* = 0 \quad (8a)$$

$$I_n - 13b^* - I_h - aD_n + aD_h = 0 \quad (8b)$$

$$I_o - aD_o - 16b^* = 0 \quad (8c)$$

Using the NIS values[3] D_h = 0.007825032, D_n = 0.0030740053, I_c = 0.001078, I_h = 0.00012197, and I_n = 0.003646, obtain from (8a-b): a = 0.3481012, b = 0.3982968 ‰. Note that (8c) is independent from (8a-b) and contains only values related to oxygen. Knowing that D_o = -0.005085378 and using the values of a and b from (8a-b), find 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.

The table value of I_o is 0.004515, i.e. just 19.5‰ 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‰.[4] 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%.

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.

Not all amino acids are proportionally represented in on-Line peptides. As expected, amino acids with $z = 0$ (Table 1) are overrepresented, while sulfur-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 hypothesis and determined the consensus order of amino acid emergence.[2] We found an overall correlation (p=0.05) between the consensus ranking of a residue and the probability to be found in on-Line peptides (Figure 4). The correlation means that the more ancient an amino acid is, the more likely it is found in on-Line peptides.

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. Even today, the z=0 residues account for 57% of the amino acid abundance in natural proteins. When life first emerged, the relative abundance of these and other low-z amino acids was certainly higher, possibly as high as 100%. In the course of evolution, as was discovered by Zuckerkandl[5] et al. and recently by Jordan et al.[6], the 'ancient' amino acids which tend to have low z values were consistently lost, being replaced by 'novel' amino acids with high z-values. 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.

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,[7] 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.[2] Note that 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). Ribose $C_5H_{10}O_5$ that is the basis of RNA *and* has plausible prebiotic syntheses, also has z=0. Thus there may also 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 explanation of the benefit of a linear correlation between the mass defects and isotopic shifts of the 'lucky' molecules.

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 sulfur) are independent upon the isotopic masses.[8] 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 a reduction in the number of quantum-mechanical states due to the symmetry of some isotopomers.[9] Similarly, in our case the peptides with z=0 are characterized by a 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 linear correlation 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 must accelerate certain reactions involving molecules with z = 0. MIF is usually observed in non-equilibrium processes involving photo-, electronic excitation or high

temperatures;[8] abiotic synthesis of amino acids and other building bocks of life is thought to be involving similar mechanisms.[10]

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 abundance 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 and more.[11] For instance, the ratio $^{12}C/^{13}C$ is 92 on Earth and ≈20 in the Galactic centre.[12] In general, terrestrial isotopic abundances are atypical for our Galaxy.[12] Thus if the feature in Figure 1 is not spurious, life has emerged either in the Solar system or in an environment with very similar to terrestrial isotopic abundances of biogenic elements.

References

1. Demirev, P. A. & Zubarev, R. A. Probing Combinatorial Library Diversity by Mass Spectrometry. *Anal. Chem.* **69**, 2893-2900 (1997).

2. Trifonov E. N. The Triplet Code from First Principles, J. Biomolec. Struct. Dynamics, 22, 1-11 (2004).

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

4. Kroopnick, P. & Craig, H. Atmospheric Oxygen; Isotopic Composition and Solubility Fractionation. Science 175, 54–55 (1972).

5. Zuckerkandl, E., Derancourt, J., Vogel, H. Mutational trends and randomprocesses in the evolution of informational macromolecules. J. Mol. Biol. 59, 473-490 (1971).

6. Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., Wolf ,Y. I., Koonin, E. V., Kondrashov, A. S., Sunyaev, S. A universal trend of amino acid gain and lossin protein evolution. Nature 433, 633-638 (2005).

7. Miller, S. L. A Production of Amino Acids Under Possible Primitive Earth Conditions, Science 117, 528 - 529 (1953).

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

9. Gao, Y. Q. & Marcus, R. A. Strange and Unconventional Isotope Effects in Ozone Formation, Science, 293, 259 – 263 (2001).

10. Bada, J. L. How life began on Earth: a status report. Earth and Planetary Sci. Lett. 226, 1-15 (2004).

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

12. 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. 326, 139–142 (1997).

Figure legends

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. The distances are calculated as $d = (NIS - a \cdot NMD - b)/(1 + a^2)^{0.5}$.

Figure 3. Distances in ‰ from the line feature for peptide brutto-formulae, and with added or subtracted water molecule.

Figure 4. Correlation between the degree of enrichment of amino acids in on-Line peptides and their consensus average rank of emergence in ref. 2. The enrichment degree is calculated as $(N_L/N_T)\cdot(P_T/P_L)$, where N_L is the number of amino acids of a particular type in on-Line peptides, N_T is the number of these amino acids in all peptides, P_T is the total number of amino acids in all peptides and P_L is the total number of amino acids in on-Line peptides.

Table 1. Sulfur-free amino acid residues, their *z* values $z = c - (n + h)/2$, and the consensus order of emergence[2]

Alanine, C_3H_5NO	0	4.0
Arginine, $C_6H_{12}N_4O$	-2	11.0
Asparagine, $C_4H_6N_2O_2$	1	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

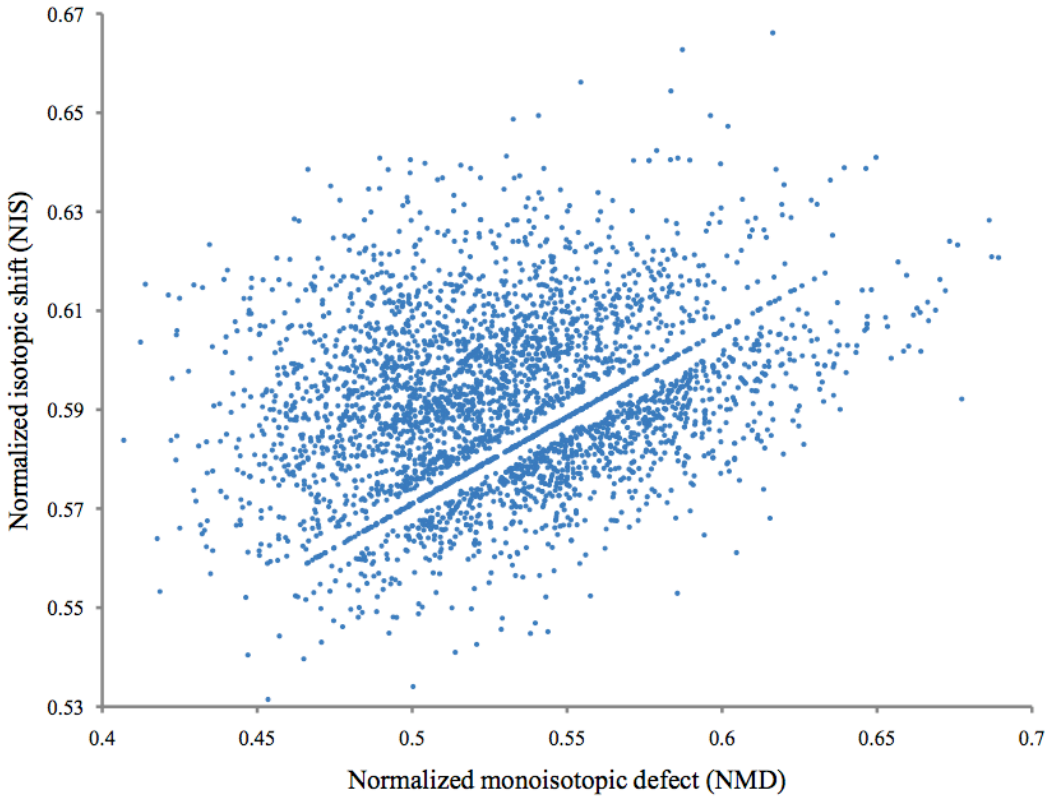
0.67
0.65
0.63
0.61
0.59
0.57
0.55
0.53
Normalized isotopic shift (NIS)
0.4
0.45
0.5
0.55
0.6
0.65
0.7
Normalized monoisotopic defect (NMD)

Figure 1.

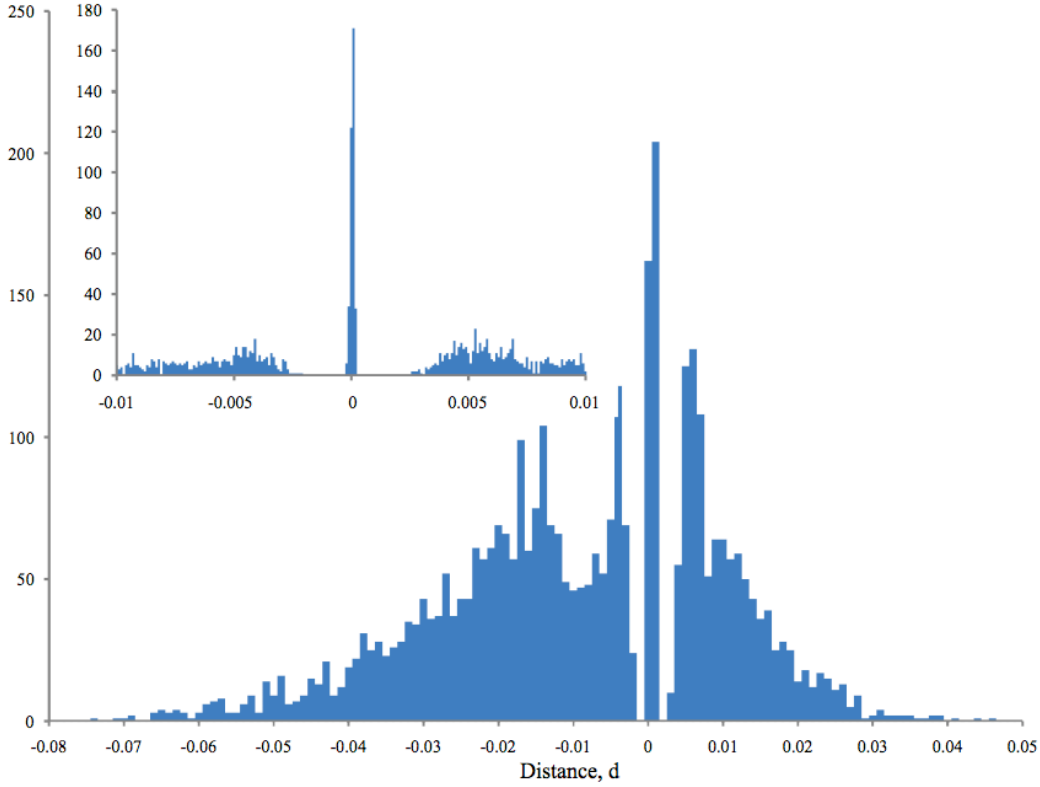
250
200
150
100
50
0
-0.08
-0.07
-0.06
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
0.05
Distance, d
180
160
140
120
100
80
60
40
20
0
-0.01
-0.005
0
0.005
0.01

Figure 2.

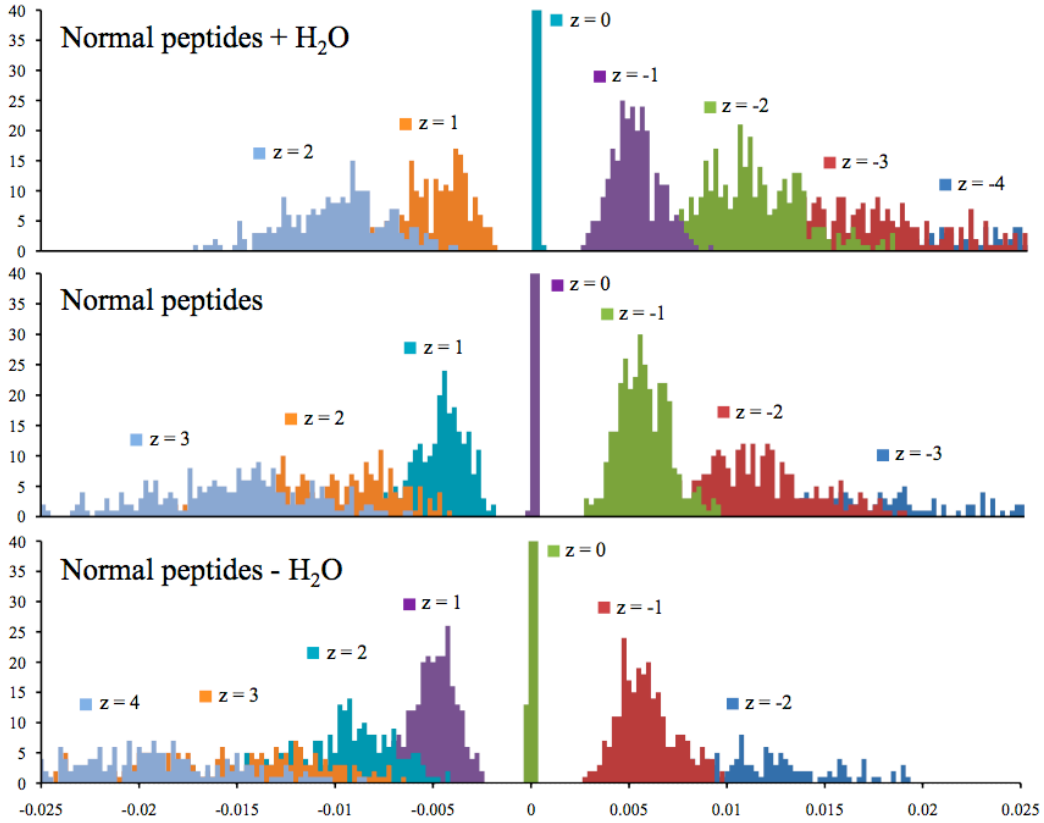
Normal peptides + H2O
z = 0
z = -1
z = -2
z = -3
z = -4
z = 1
z = 2
Normal peptides
z = 0
z = -1
z = 1
z = -2
z = 2
z = -3
z = 3
Normal peptides - H2O
z = 0
z = -1
z = 1
z = -2
z = 2
z = 4
z = 3
0 5 10 15 20 25 30 35 40
-0.025 -0.02 -0.015 -0.01 -0.005 0 0.005 0.01 0.015 0.02 0.025

Figure 3.

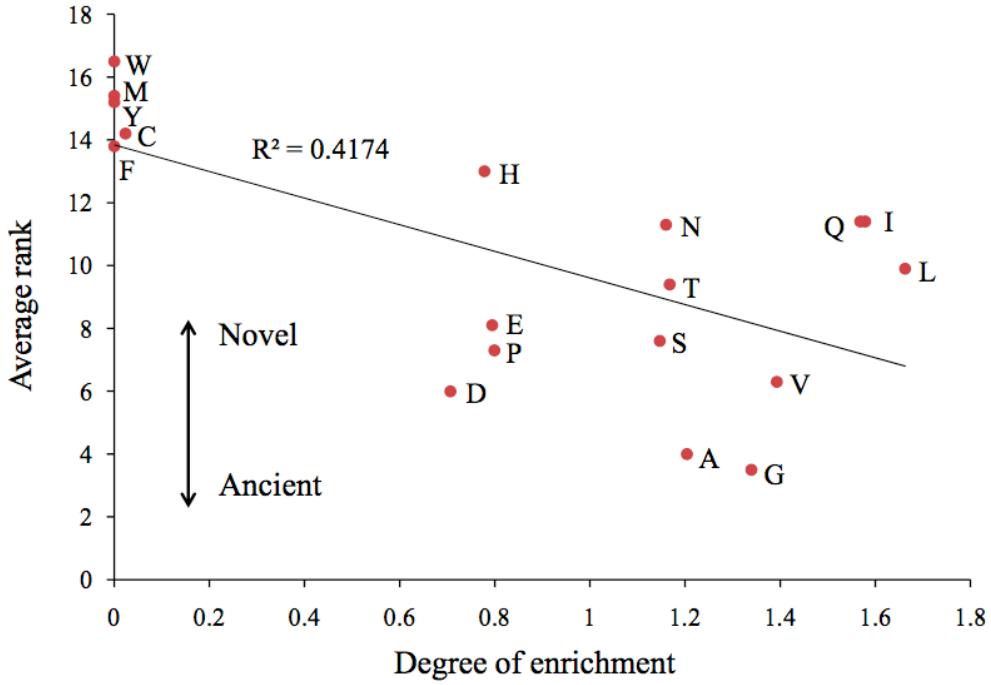
18
16
14
12
10
8
6
4
2
0
Average rank
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Degree of enrichment
R² = 0.4174
W
M
Y
C
F
H
N
Q
I
L
T
E
P
S
V
D
A
G
Novel
Ancient

Figure 4.