Electronic Supplementary Material (ESI) for Analyst.

Identification of Isoforms of Asp Residue in Peptides by 2D UV-MS Fingerprinting of Cold Ions

Vladimir Kopysov,[a] Mikhail V. Gorshkov,[b] and Oleg V. Boyarkin[a†]

Laboratoire de Chimie Physique Moléculaire, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

† Electronic address: oleg.boiarkin@epfl.ch

Institute of Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, Russia.

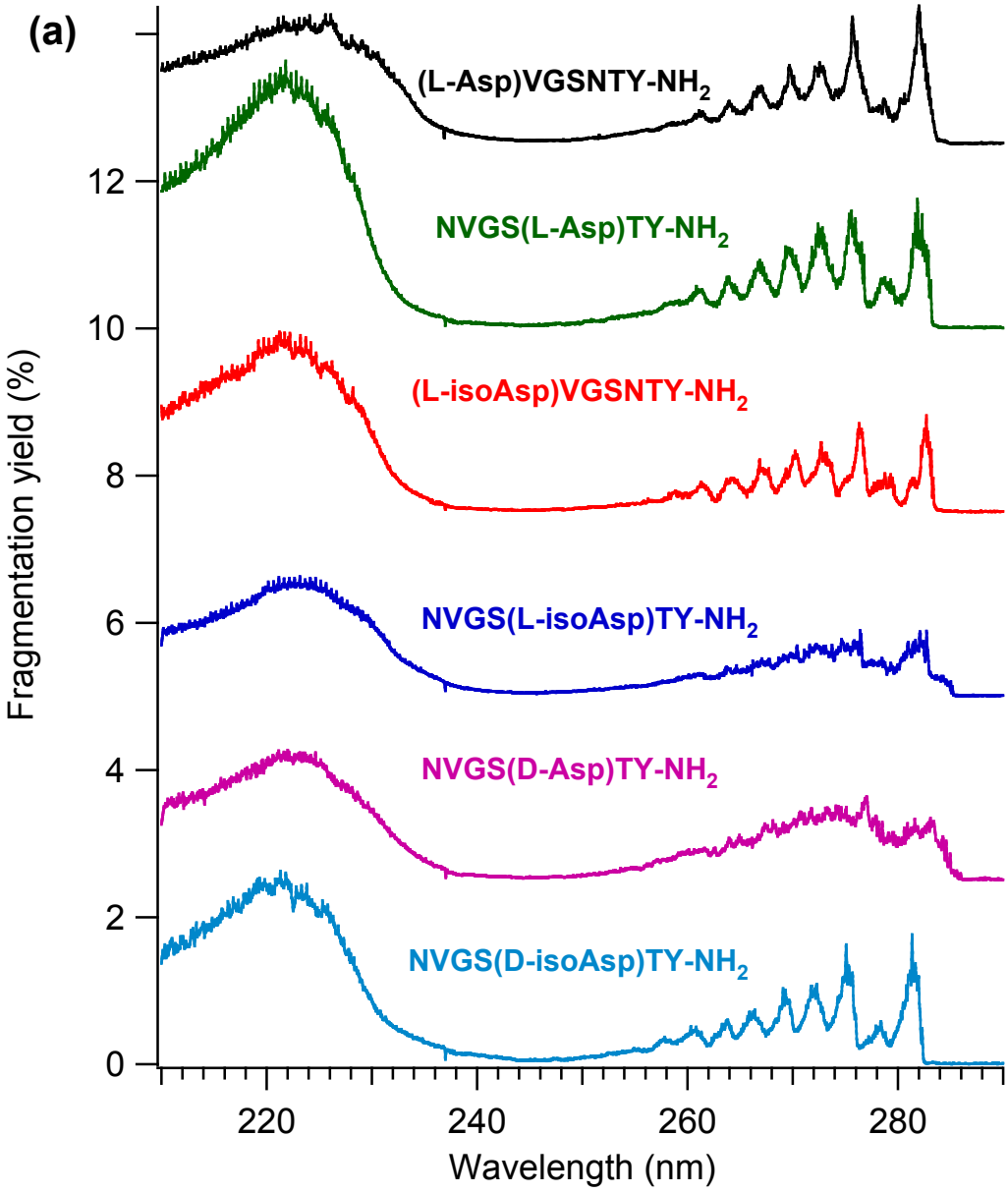


Figure S1. Photofragmentation UV spectra of singly protonated amylin(30-37) peptide with different isoforms and positions of Asp residue. The spectra have been generated by integrating the respective 2D UV-MS over m/z. The 2D UV-MS arrays are normalized to the UV OPO pulse energy; each mass spectrum is normalized to the total ion signal, including the parent ion. For graphical clarity, subsequent UV spectra are offset relative to the preceding ones by 2.5%.

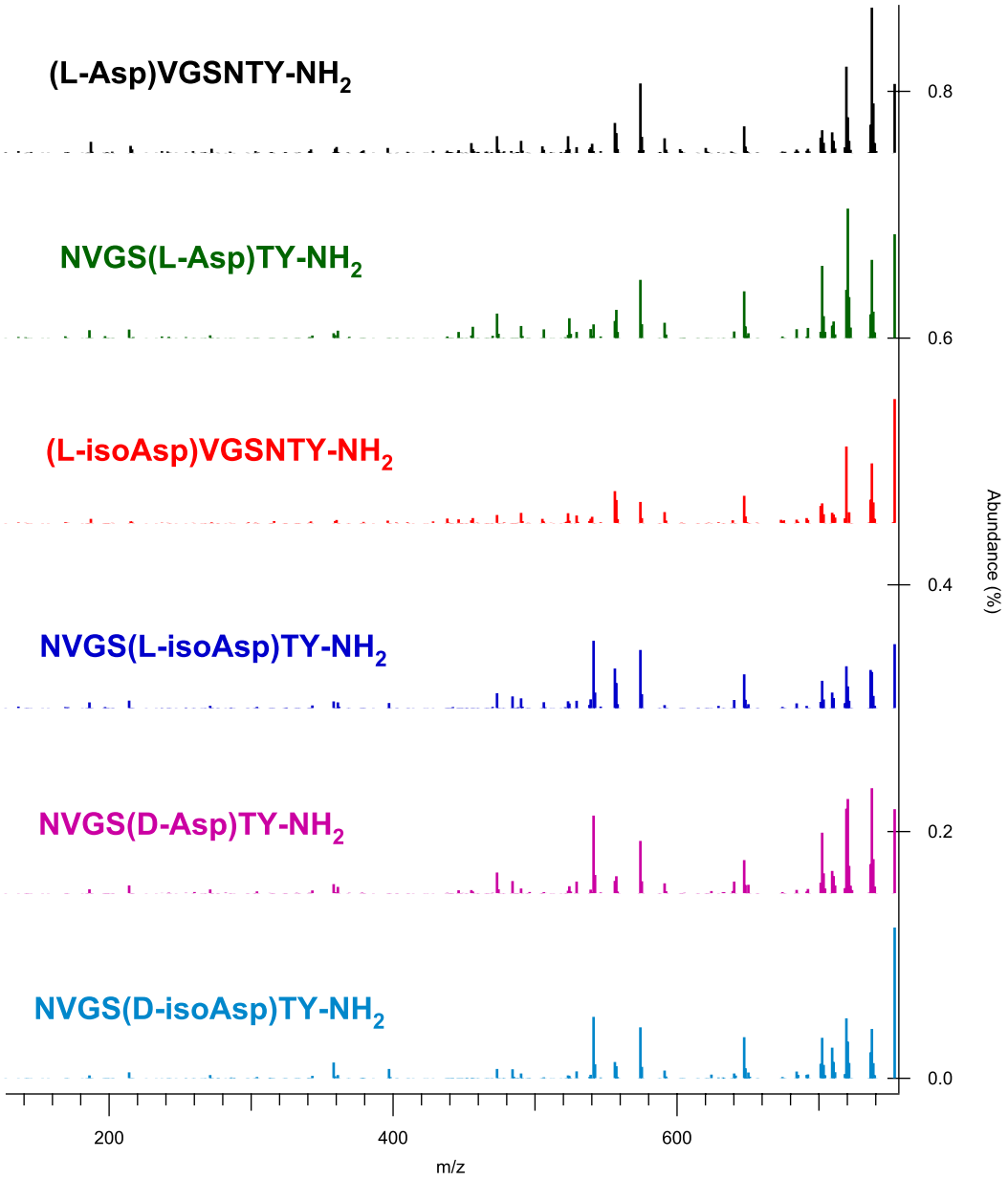


Figure S2. UVPD mass spectra of singly protonated amylin(30-37) peptide with different isoforms and positions of Asp residue. The spectra have been generated by integrating the respective 2D UV-MS over wavelength coordinate. The 2D UV-MS arrays are normalized to the UV OPO pulse energy; each mass spectrum is normalized to the total ion signal, including the parent ion. For graphical clarity, subsequent mass spectra are offset relative to the preceding ones by 0.15%, respectively.

Note S1. Library-based identification of isomers in their mixtures.

The library-based identification of isomers in their mixtures implies 1) measuring 2D UV-MS spectra of the species to be identified; 2) constructing a library of the corresponding UV-MS matrices; 3) measuring 2D UV-MS spectra of the mixtures to be analyzed; and 4) decomposing the corresponding UV-MS matrices into a linear combination of the library matrices. The last step is very similar to decomposing a vector in a basis, that is, expressing the vector as a sum of the basis vectors multiplied by the coefficients reflecting the coordinates of the vector relative to this basis.

Building a library of fingerprints for individual isomeric compounds begins with transforming 2D UV-MS spectra, which are recorded on the Exactive Orbitrap-based mass spectrometer and stored as RAW files, to UV-MS matrices stored as plain-text files. This is done by a PeakByPeak software package (Spectroswiss Sarl, Switzerland)[1] written in Python. It first calculates an integrated photofragmentation mass spectrum by summing up all single mass scans, each of which is preliminary corrected for a baseline by means of a model-free Friedrichs algorithm.[2] The mass spectrum is used to find the m/z values of all the photofragment ions that appear during the UV laser scan as well as the m/z values of parent ions. Then PeakByPeak processes all single mass scans one by one using the resulting set of m/z values as a reference peak list. For each mass scan, it evaluates the noise level (σ)[3] and identifies all the peaks, the intensity of which is higher than 5σ. Only the peaks, the m/z values of which are found in the reference peak list (with a tolerance of 10 ppm), are finally retained. The output data are stored as an n-by-m matrix D; an element D_{ij} of the matrix is the intensity of the j-th peak from the reference list in the i-th mass scan and therefore corresponds to the intensity of an ion with the mass-to-charge ratio of m/z_j at the UV laser wavelength of λ_i. Each row of the matrix D is divided by the sum of row elements, that is, normalized on the total ion signal; then, each column, except those that correspond to parent ions, are divided by the recorded OPO power curve, that is, normalized to 1 mJ energy of UV OPO. UV-MS matrices $D^{(i)}$, $i = 1...k$, where k is the number of individual components, form a library of fingerprints.

The same procedure is used to obtain a matrix $D^{(mix)}$ of a mixture from its 2D UV-MS spectrum. The UV-MS matrices of individual compounds as well as the matrix of the mixture are then imported into MATLAB and processed by an in-house developed MATLAB script. First, we define a set of wavelengths, at which the UV-MS matrix of a mixture will be analyzed, and generate a set of m/z values of all the photofragment peaks present in the UV-MS matrices of individual components. Each of the matrices is then mapped onto these new wavelength and m/z scales in the following way: if an initial matrix contains a peak at a certain m/z (from the new set of m/z values), then the UV spectrum at this m/z (i.e., a column

vector in the initial UV-MS matrix) is projected onto the new wavelength scale by means of linear interpolation; otherwise, the respective column vector in the resulting matrix is filled with zeros.

If we denote the number of individual components by *k*, the number of selected wavelengths as *n*, and the total number of mass peaks by *m*, the new matrices of individual components form a basis of a *k*-dimensional subset of a vector space of *n*-by-*m* matrices. The projected matrix of a mixture (some of *k* species may be absent in the mixture, which corresponds to zero concentration) should then be a linear combination of the library matrices:

$$D^{(mix)} = \sum_{i=1}^{k} x_i \cdot D^{(i)}$$

However, due to noise in the experimental data and possible data processing errors, the system of $n \times m$ equations becomes inconsistent. A general approach to find an approximate solution of an overdetermined system (indeed, $n \cdot m \gg k$) is to minimize the sum of squared residuals with respect to variables. More specifically, we minimize the squared Frobenius norm of the residual matrix subject to non-negativity constraints:

$$\{x_1, \ldots, x_k\} = \arg\min_{x \geq 0} \left\| D^{(mix)} - \sum_{i=1}^{k} x_i \cdot D^{(i)} \right\|_F$$

using the MATLAB built-in function *lsqnonneg*. To find the relative concentrations of the mixture components, the determined coefficients are divided by their sum:

$$\hat{c}_i = 100\% \cdot \frac{x_i}{\sum_{j=1}^{k} x_j}$$

Finally, the accuracy of the identification of individual components in a set of *s* mixtures is assessed in terms of the root-mean-square deviation (RMSD):

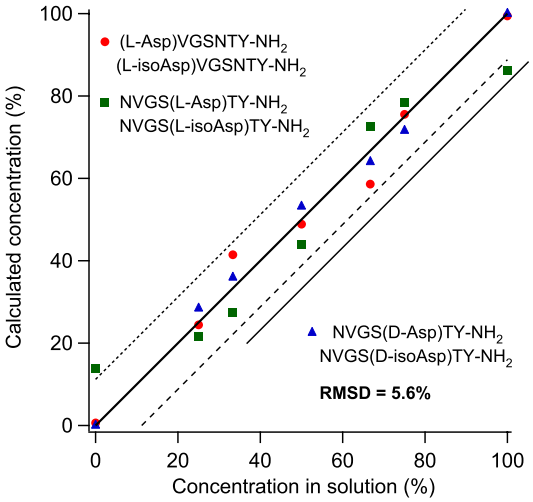
$$RMSD = \sqrt{\frac{1}{k \cdot s} \cdot \left(\sum_{i=1}^{s} \sum_{j=1}^{k} \left(c_j^{(i)} - \hat{c}_j^{(i)} \right)^2 \right)}$$

which is a measure of discrepancy between the relative concentration $c_j^{(i)}$ of the *j*-th component in the *i*-th mixture and its predicted value $\hat{c}_j^{(i)}$ averaged over all the components and all the mixtures.

References

1. Spectroswiss Sarl, http://spectroswiss.ch.
2. M. Friedrichs, *J. Biomol. NMR*, 1995, **5**, 147–153.
3. K. O. Zhurov, A. N. Kozhinov, L. Fornelli and Y. O. Tsybin, *Anal. Chem.*, 2014, **86**, 3308-3316.

(a)



(b)

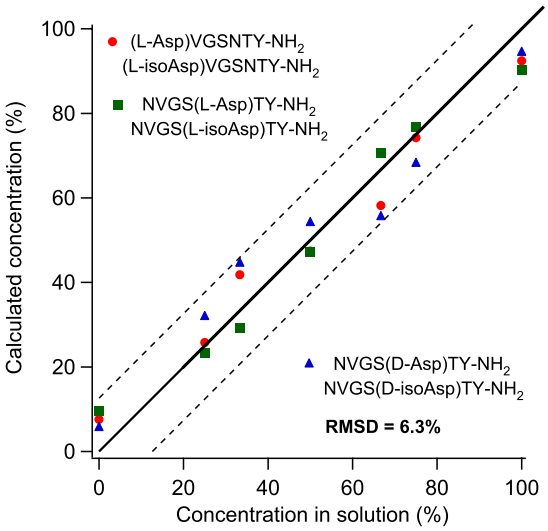


Figure S3. Calculated relative concentrations of six singly protonated amylin(30-37) peptides as a function of their relative concentrations in 12 pairwise solution mixtures for (a) the entire 290-210 nm and for 250-210 nm spectral range.

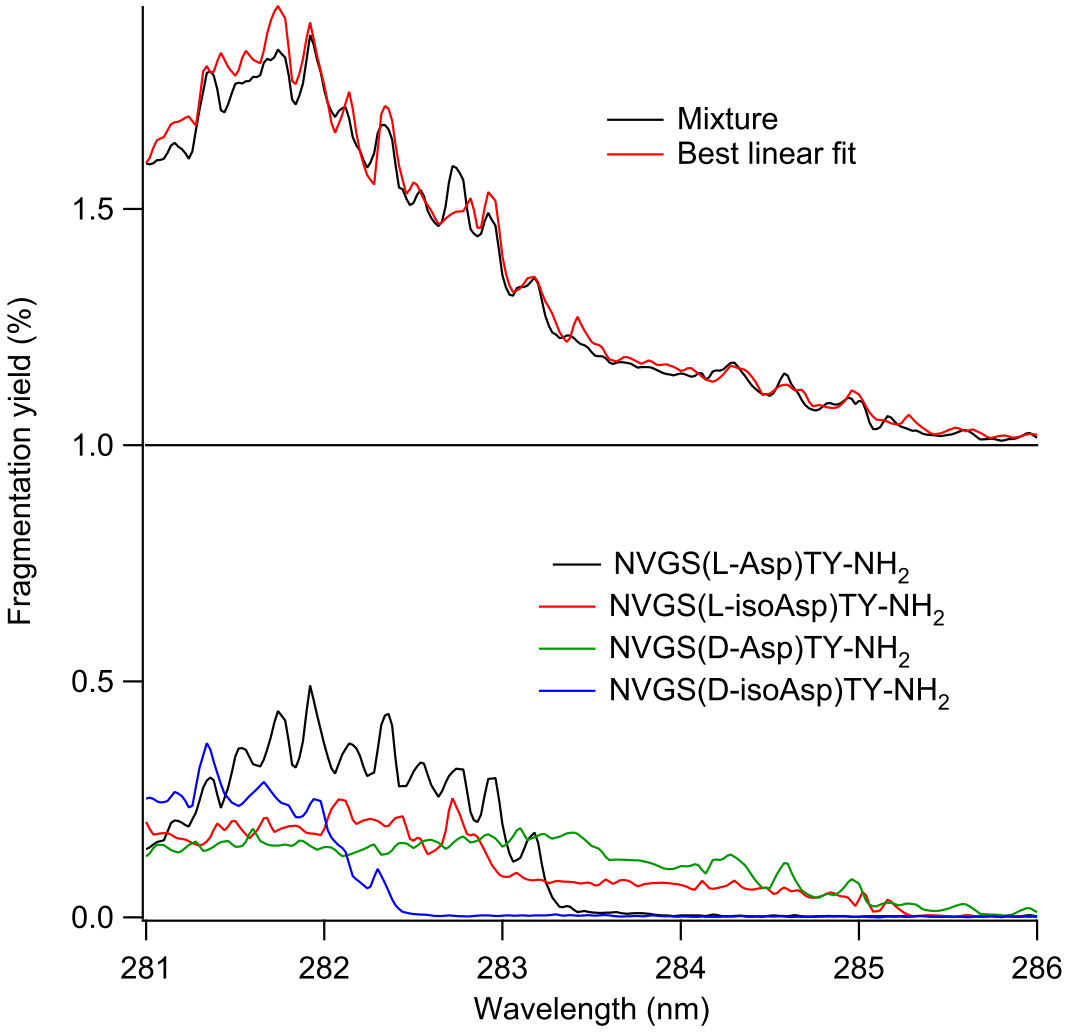


Figure S4. (a) UV spectrum of an equimolar quaternary solution mixture of singly protonated NVGS(L-Asp)TY-NH_2, NVGS(L-isoAsp)TY-NH_2, NVGS(D-Asp)TY-NH_2 and NVGS(D-isoAsp)TY-NH_2 isomeric peptides and its best linear fit by as it results from the decomposition procedure of the respective 2D UV-MS data arrays, and (b) library UV spectra of the mixed peptides, scaled by their calculated relative concentrations. All UV spectra were generated by integrating the respective 2D UV-MS data arrays over m/z.

Table S1. Wavelengths of the first maxima within 281-286 nm interval in UV spectra of protonated isomeric peptides amylin(30-37). The decompositions of 2D UV-MS data arrays with these wavelengths only result in RMSD of 2.5%.

Isomeric peptide	Wavelength, nm
NVGS(D-isoAsp)TY-NH_2	281.35
NVGS(L-Asp)TY-NH_2	281.86
(L-Asp)VGSNTY-NH_2	282.00
(L-isoAsp)VGSNTY-NH_2	282.71
NVGS(L-isoAsp)TY-NH_2	282.72
NVGS(D-Asp)TY-NH_2	283.09