A New Approach to Characterization of Insulin Derived from Different Species Using $^1$H-NMR Coupled with Multivariate Analysis

Akiko Ohno,*a Toru Kawanishi,b Haruhiro Okuda,a and Kiyoshi Fukuharaa

aDivision of Organic Chemistry, National Institute of Health Sciences; and bDivision of Drugs, National Institute of Health Sciences; Setagaya-ku, Tokyo 158–8501, Japan.

Received October 7, 2011; accepted December 8, 2011; published online December 27, 2011

Most of the active components of polypeptides have a complex molecular structure, large molecular size. Such components may also be structurally heterogeneous. Therefore, development of a method that can confirm the consistency of polypeptides amino-acid sequences for product characterization is desirable. In general, it is extremely difficult to distinguish differences of a few amino acid residues in the $^1$H-NMR spectrum of polypeptides with molecular weights greater than several thousand. However, we have been able to distinguish between three insulin species differing in one to three amino acid residues using a combination of multivariate statistics and $^1$H-NMR spectra. These results demonstrate that this methodology could be useful for characterization of polypeptides.

Key words insulin; polypeptide; characterization; principal component analysis; $^1$H-NMR

Nuclear magnetic resonance (NMR), which is frequently used for structure identification of unknown chemical substances in the fields of organic and natural products chemistry, is the only technique that can provide structural information on all of the components of a chemical substances. Although NMR techniques uniquely provide spectral information on primary and higher-order structure of large polymeric compounds such as polypeptides, such spectra are generally difficult to analyze in detail because of their complexity. NMR measurement has usually been applied to the structural analysis of single chemical substances, and to date, been not suitable for analyzing samples that contain multiple compounds because of the problem of overlapping peaks in the $^1$H-NMR spectrum. However, NMR techniques have come to be used recently to analyze biogenic substances, which have made it possible to discriminate between small spectral differences by performing statistical analysis of the $^1$H-NMR spectral data. Multivariate statistical methods, such as principal component analysis (PCA) and partial least-squares discriminate analysis (PLS-DA), are often useful for profiling and classifying sample groups and for characterizing the most effective variables of separated compounds.1,2) Currently, multivariate statistical methods, which combine various analytical methods, have been widely used to evaluate the quality of drugs and foods, quantitatively or qualitatively, in addition to enabling predictions of drug metabolism, and toxicity.3–10) Compared with small-molecule drugs, polypeptides are heterogeneous and are more complex in their makeup for a number of possible reasons including mutations in the amino acid sequence of the protein, different posttranslational modifications, or by being a mixture of molecules with different terminal structures due to degradation by contaminating proteases.11) The structural heterogeneity caused by these factors may affect the physiological activity and pharmacokinetics of polypeptides, resulting in changes in drug efficacy and safety.11) Therefore, the development of analytical procedures that can confirm the constancy of multiple amino-acid sequences of a polypeptide for product characterization is desirable. However, characterizing the full complexity of polypeptides by presently available analytical methods is still difficult. Therefore, it is necessary to provide a more detailed evaluation of special characteristics of polypeptides by a new physicochemical index or new analytical techniques. In our recent study, it was reported that peak changes in the determination of characteristic spectral changes associated with time-dependent alterations of oxytocin (OXT) were also observed in the PCA loading plot.8,12) Thus, the possibility of evaluating slight differences in the quality of a polypeptide was demonstrated.

In this study, we examined multivariate statistics coupled with $^1$H-NMR to analyze the difference of amino acid sequences in three species of insulin and to show this method to be effective for the characterization of the polypeptides. As a result, we have succeeded in precisely characterizing human, bovine, and porcine insulins with sequence differences of one or three amino acids, by performing $^1$H-NMR measurements of the individual insulins and their mixtures followed by PCA of the spectra. The results suggest that this methodology could be useful for the characterization of species-related sequence differences in polypeptides.

Experimental

Chemicals and Reagents All reagents used for $^1$H-NMR experiments were of analytical grade (purity >99%) from Wako Chemicals and were used without further purification. Human recombinant insulin expressed in yeast (CAS# 11061-68-0) and insulin from bovine pancreas (CAS# 11070-73-8) were purchased from Aldrich (St. Louis, MO, U.S.A.); porcine insulin (CAS# 12584-58-6) was purchased from MP Biomedicals. Acetonitrile-$d_3$ for NMR was purchased from Acros Organics (CN). Deuterium oxide (D$_2$O, isotopic purity 99.9%) and 3-(trimethylsilyl) propionic-2,2,3,3-$d_4$ acid, sodium salt (TSP) were purchased from Aldrich (St. Louis, MO, U.S.A.). TSP was used as an internal standard with a chemical shift (δ) of 0.0 ppm in $^1$H-NMR measurements.

Sample Preparation and $^1$H-NMR Spectroscopic Analysis Each insulin (14.7 mg) was dissolved in 140 μL of 0.1 M HCl, 70 μL of 0.1 M NaOH, 200 μL of Milli-Q water, 70 μL of 5 mM TSP/D$_2$O, and the pH was adjusted to 3.6 by adding aliquots of 0.1 M NaOH or HCl. Milli-Q water was added to give a total volume of 910 μL, and this mixture solution was diluted with 490 μL of CD$_3$CN. The solvent used in the present work was H$_2$O/D$_2$O : CD$_3$CN (65/35 vol%). For $^1$H-NMR

*To whom correspondence should be addressed. e-mail: ako-ohno@nihs.go.jp © 2012 The Pharmaceutical Society of Japan
measurements, a total sample volume of 700μL with at least a 1.7 mM concentration (pH 3.6) was used. Samples of each insulin mixture ratio are shown in Table 1.

The sample was introduced into an NMR test tube, and nuclear Overhauser effect spectroscopy (1H-NOESY) spectra were recorded at 25°C using a Varian 600 MHz NMR spectrometer equipped with a cold probe. Thirty-two free induction decays (FIDs) with 75 K data points per FID were collected using a spectral width of 9615.4 Hz, an acquisition time of 4.00 s, and a total pulse recycle delay of 2.02 s. The water resonance was suppressed using presaturation during the first increment of the NOESY pulse sequence, with irradiation occurring during the 2.0 s relaxation delay and also during the 200 ms mixing time. Prior to Fourier transformation, the FIDs were zero-filled to 128 K and an exponential line broadening factor of 0.5 Hz was applied. Spectral 1H-NMR assignments were achieved according to the literature values of chemical shifts in various media.

**NMR Data Reduction and Preprocessing** All 1H-NMR spectra were phased and baseline corrected by Chenomx NMR Suite 6.0 software, professional edition (Chenomx Inc., Canada). 1H-NMR spectra were subdivided into regions having an equal bin size of 0.04 ppm over a chemical shift range of 0.04—10.0 ppm (excluding the region around the water signal; 4.2—4.6 ppm), and the regions within each bin were integrated. The integrated intensities were then normalized considering the intrinsic variation in the data sets. The quality of the models was described by the total spectral area, and the data were converted from the Chenomx software format into Microsoft Excel format (*.xls). The resulting data sets were then imported into SIMCA-P version 12.0 (Umetrics AB, Umeå, Sweden) for multivariate statistical analysis.

**Multivariate Data Analysis** PCA was performed to examine the intrinsic variation in the data sets. The quality of the models was described by the $R^2_x$ and $Q^2$ parameters, which indicate the proportion of variance in the data explained by the models and the goodness of fit. $R^2_x$ represents the goodness of fit of the PCA model, and $Q^2$ expresses the predictability of the PCA model. The quality of the PCA models was described by the total variance of principal component 1 (PC1) and principal component 2 (PC2) at a confidence level of 95%.

### Results

The amino-acid sequences of human, bovine, and porcine insulins are shown in Fig. 1. Bovine insulin differs from human insulin at the following positions: alanine (Ala) for threonine (Thr) at position A8 (8th position on the A Chain), valine (Val) for isoleucine (Ile) at A10, and Ala for Thr at the carboxyl terminus of the B-chain. Porcine insulin differs from human insulin with an Ala substituted for Thr at the carboxy terminus of the B-chain. 1H-NMR spectra of the three types of insulin are shown in Fig. 2. While a simple visual inspection suggests that the three spectra might be indistinguishable, actual spectral differences may be detected if changes can be represented as points in a multidimensional space and examined using PCA. PCA of each insulin spectrum was performed, and distinct differences among the 1H-NMR spectra at each sample mixture ratio were readily detected by the scores of both PC1 and PC2, which could be clearly depicted as points on the lines of the triangular phase diagram as shown in Fig. 3. In the PC1-PC2 plane, all samples were displayed in a triangular phase diagram bearing the three types of single composition insulin at each vertex. The cumulative contribution rate by PCA of the first two principal components, PC1 and PC2, was 63.8 and 83.5%, respectively. Thus, a spectrum change was characterized by PC1 and PC2 with species-related differences of insulin at a high contribution ratio. This result suggests a large contribution of human and bovine insulin to PC1, indicating that the positive direction from the center of the PC1 coordinate on the horizontal axis corresponded to bovine insulin and the negative direction corresponded to human insulin. Meanwhile, a high contribution to PC2 indicated that the positive direction from the center of the PC2 coordinate on the vertical axis corresponded to porcine insulin and the negative direction corresponded to human and bovine insulin. In addition, the mixed sample of all three types of insulin in a 1 : 1 : 1 ratio is in the center of the triangular phase diagram, and samples of each insulin mixture ratio radiate from the center toward each vertex.

The loading plot of all the evaluated 1H-NMR signals is shown in Fig. 4. Each variable represents a peak at a particular chemical shift in the 1H-NMR spectra shown in Fig.

### Table 1. Each Insulin Mixture of the Sample Used in This Experiment: Human Insulin (A), Bovine Insulin (B), and Porcine Insulin (C)

<table>
<thead>
<tr>
<th>Sample</th>
<th>A :</th>
<th>B :</th>
<th>C :</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A4B1</td>
<td>4</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>A1B1</td>
<td>1</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>A1B4</td>
<td>1</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td>B1</td>
<td>—</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>B1C1</td>
<td>—</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>B1C4</td>
<td>—</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C1</td>
<td>—</td>
<td>—</td>
<td>1</td>
</tr>
<tr>
<td>A1C4</td>
<td>1</td>
<td>—</td>
<td>4</td>
</tr>
<tr>
<td>A1C1</td>
<td>1</td>
<td>—</td>
<td>1</td>
</tr>
<tr>
<td>A1C4</td>
<td>1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A1B1C1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A2B1C1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A1B2C1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>A1B1C2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
This loading plot reveals the contributions of particular variables towards either an increase or decrease in the integrated intensities of the $^1$H-NMR spectra. Six variables showing typical fluctuations of the different insulin species were identified (Fig. 4), and the integrated intensities of these variables corresponding to the insulin ratios are shown in Fig. 5. The tendency of the intensity of each of these variables to either increase or decrease is evident in the differences of the mixed insulin ratios and suggests specificity of the variables for the different insulin species. The corresponding characteristic variables of human, bovine, and porcine insulin were $\delta$ 1.1, 0.86 and 4.06 ppm, respectively. In addition, a common characteristic variable of bovine and porcine insulin occurred at $\delta$ 1.3 ppm and that of human and porcine insulin occurred at $\delta$ 1.18 ppm. The typical chemical shifts responsible for these variables were shown in Fig. 6. Therefore, these variables were indicative of the species specificity of insulin since the insulin species characterized by each variable tended to have a high intensity either for a single insulin species or for a species making up a high proportion of the mixture as shown in Fig. 5. Next, we examined whether the larger variables characterizing each species were derived from the type of amino acid residue by analysis of the loading and trend plots.

First, determination of the variables contributing significantly to the variation of the PC1 coordinate axis for the human and bovine insulin groups was performed by analysis of the amino acid residue peaks corresponding to these particular variables. Peaks for each variable could be identified: the variable at 1.1 ppm was attributed to both the $\gamma$H of Ile at A10 and to the $\gamma$H of Thr at B30; the variable of the amide group region at 8.06 ppm was attributed to the NH of Thr at A8 of the characteristic amino-acid sequences of human insulin. On the other hand, the variable at 0.86 ppm was attributed to the $\gamma$H of Val at A10; and the variable at 1.5 ppm was attributed to the $\beta$H of Ala at A8 and/or B30; the variable of the amide group region at 8.46 ppm was attributed to the NH of Val at A10 of the characteristic amino-acid sequences of bovine insulin.

Next, determination of the variables contributing significantly to the variation of the PC2 coordinate axis for the two groups, porcine, and both human and bovine was performed by the analysis of the amino acid residue peaks corresponding to these particular variables. Peaks for each variable could be identified: the variable at 4.06 ppm was attributed to the $\alpha$H of Ala at B30; the variable of the amide group region at 7.78 ppm was attributed to the NH of Ala at B30 of the characteristic amino-acid sequences of porcine insulin. Therefore, identified amino acids for each characteristic variable were by inference associated with the primary structure of the three types of insulin.
The three insulin species used in this study each consist of 51 amino acid residues; bovine insulin has three different positions and porcine insulin has one different position from human insulin. In general, it is extremely difficult to distinguish differences of one to three amino acid residues in the $^1$H-NMR spectrum of polypeptides of approximately MW 6000. However, we succeeded in distinguishing between the insulin species using a combination of multivariate statistics and $^1$H-NMR spectra. PCA of the insulin spectrum for each mixture ratio was performed, and variables due to several different amino acid residues were detected from both the scores of PC1 and PC2, clearly depicting three separate groups, as shown in Figs. 3 and 4. In addition, from the scores plot, the components of the single insulin species and of mixtures consisting of two-three species were statistically distributed in the triangular phase diagram according to differences in the species composition ratio (Fig. 3, Table 1). This result shows that it is possible to analyze the composition of a mixture of species with a small number of amino acid sequence differences by taking advantage of the scores plot.

We also examined by PCA the partial aliphatic and amide regions of each insulin spectrum (data not shown). The determination of the variables contributing significantly to the variation of the PC1 coordinate axis for the human and bovine insulin groups and to the variation of the PC2 coordinate axis for the two groups, porcine, and both human and bovine insulin groups of the loading plots was performed by analysis of the amino acid residue peaks corresponding to these particular variables. Differences of the characteristic amino acid residues for each insulin species as well as the results from analyzing entire spectra were reflected in the scores plot. It is noteworthy that a similar tendency from analysis of the complete spectral region was observed with the scores plot of the amide region. These results suggest that it may be possible to distinguish slight sequence differences of polypeptides by PCA analysis of the amide region of the NMR spectrum.

On the other hand, the biological effect of polypeptides is also influenced by differences in their higher-order structure. It is difficult to evaluate differences in the higher-order structure of polypeptides by conventional analytical methods. In this PCA analysis study, it was found that the first and second principal components accounted for a majority of the vari-
ability differences of the primary insulin species structures. Interestingly, there were some variables that were not derived from the differences of amino acids among the three insulin types. Through a comparison of the NOESY spectra of the insulin species, it was proposed that these variables were associated with differences in the higher-order structures (data not shown). Therefore, the NMR technique coupled with PCA might also be useful as a tool for analyzing the higher-order structure, which is associated with the quality of a polypeptide, because in addition to primary structure information of the amino acid residues, the $^1$H-NMR spectrum also contains information related to the three-dimensional structure of a molecule that is dependent on the solution conformation.

In conclusion, we have succeeded in precisely characterizing samples of human, bovine, and porcine insulin, molecules that differ in amino acid sequence from one to three amino acid residues, by $^1$H-NMR spectroscopy coupled with PCA. Currently, assessment of the differences of higher-order structure using PCA analysis of NOESY spectral data are underway.

Acknowledgements We thank Dr. J. Kurita and Mr. K. Kushida (Agilent Technologies Japan Ltd.) for their technical assistance in measuring the NMR spectra. This work was supported by a Health Labour Sciences Research Grant from the Ministry of Health, Labour, and Welfare, Japan and by a ‘Grant-in-Aid for Young Scientists (B)’ (No. 22790126) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References