Near Infrared Spectroscopy As High-Throughput Technology for Screening of Xylose-Fermenting Recombinant Saccharomyces cerevisiae Strains

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2011

Impact Factor: 5.64 · DOI: 10.1021/ac103128p · Source: PubMed

analytical chemistry

ARTICLE

pubs.acs.org/ac

Near Infrared Spectroscopy As High-Throughput Technology for Screening of Xylose-Fermenting Recombinant *Saccharomyces cerevisiae* Strains

Hiroyuki Morita,[†] Tomohisa Hasunuma,[‡] Maria Vassileva,[§] Roumiana Tsenkova,[*,§] and Akihiko Kondo[*,†]

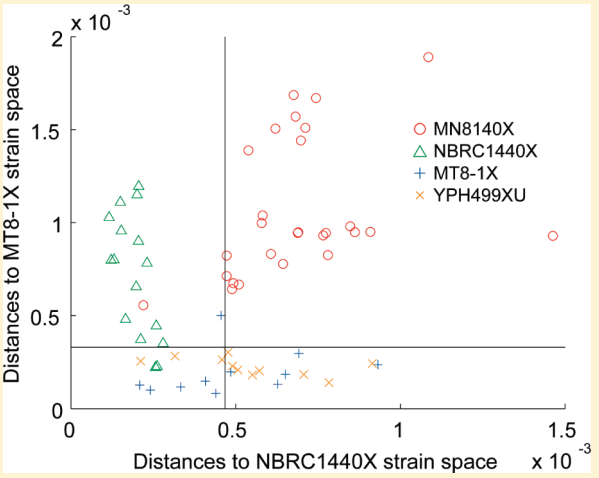
[†]Department of Chemical Science and Engineering, Graduate School of Engineering, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan

[‡]Organization of Advanced Science and Technology, Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, Japan

[§]Department of Environmental Information and Bioproduction Engineering, Graduate School of Agricultural Science, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan

Supporting Information

ABSTRACT: Recently, genetic engineering efforts have been made to develop recombinant *Saccharomyces cerevisiae* strains able to utilize xylose, an inexpensive and abundant carbon source. However, their construction and selection processes are limited by the speed and expenses of the existing testing methods, thus a rapid and equally precise method will significantly increase the number of tested strains. Here, near infrared (NIR) spectroscopy is proposed as a successful alternative method for screening recombinant xylose-fermenting *S. cerevisiae*. Supernatant samples of fermentation solutions from one diploid and three haploid recombinant strains were collected along the fermentation process. NIR spectra of the diluted supernatant provided effective differentiation of strains consistent with their phenotypic and genotypic features. This result could be used as a feedback for multicomponent analysis, in order to develop regression model for quantification of consumed glucose and xylose, produced ethanol, glycerol, and xylitol. Robust partial least-squares regression models with high prediction accuracy that are effective with any strain were achieved for all components when the modeling was performed with combined data of all strains, achieving 0.21−1.49 g/L of standard error of prediction with calibration, prediction, limit of detection and limit of quantification in the range of 1.0−4.5 and 3.0−13.4 g/L, respectively.



Numerous environmental and social benefits could result from the replacement of petroleum-based transport fuels with bioethanol converted from cellulosic materials, an inexpensive and abundant source of sugar for fermentation.[1,2] The commonly used *Saccharomyces cerevisiae* has many advantages as an ethanol producer, such as fast sugar consumption, high glucose to ethanol conversion rate, and high resistance to ethanol. However, a major drawback is that *S. cerevisiae* cannot utilize xylose, the most common pentose sugar in the hemicellulose that makes up a sizable fraction of lignocellulosic hydrolysates. Xylose is one of the most abundant carbohydrates in lignocellulosic materials and the construction of recombinant yeast strains that simultaneously and efficiently ferment both glucose and xylose would enormously increase the efficiency of bioethanol production.[3] Most efforts in the engineering of *S. cerevisiae* for xylose fermentation have focused on gene expression analysis and manipulation of the initial xylose metabolic pathway.[4]

Various reports have been published on improvements of recombinant *S. cerevisiae* fermentation from lignocellulosic materials by genetic manipulation.[5−9] Each genetic modification process requires a fast screening procedure for the achievement of improved fermentation abilities such as ethanol production, substrate consumption and reduction in byproduct formation. In the above-mentioned reports, the comparison of fermentation abilities among genetically modified strains were performed with standard analytical methods as high-performance liquid chromatography (HPLC)[6−8] and enzyme assays.[5,9] Recombinant strains' construction and selection processes are limited by the speed and expenses of the existing testing methods, thus a simple, rapid, and equally precise method will significantly increase the number of tested recombinant strains.

Received: November 29, 2010
Accepted: April 26, 2011
Published: April 26, 2011

This work presents the availability of near infrared (NIR) spectroscopy to achieve high throughput screening for recombinant strains. Electromagnetic waves in the NIR region provide deeper penetration into aqueous samples due to reduced absorption relative to infrared (IR) region. NIR spectral analysis for fermentation process can provide detection of constituents of interest with accuracy close to conventional time-consuming and/or expensive reference methods. Recently, NIR spectroscopy has been applied for monitoring of fermentation processes as it provides simultaneous quantification of components and does not require sample preparation. NIR technology use for monitoring glucose fermentation by yeast strains, such as *S. cerevisiae*, to produce ethanol,[10−12] glycerol and acetic acid[13] have been reported. *S. cerevisiae* fermentation in onion juice containing fructose, glucose and sucrose,[14] or wheat and rye degraded to glucose by enzymes[15] has been also monitored by NIR spectroscopy. NIR analysis has been applied in monitoring other microbes, such as *Rhodotorula mucilaginosa* to remove glycerol as byproduct from acidified biodiesel fuel wastewater,[16] recombinant *Pichia pastoris* for producing a therapeutic protein from glycerol and methanol,[17] and *Streptomyces fradiae* for producing antibiotic tylosin from methyl oleate, glucose, glutamate and ammonium.[18] However, the identification of xylose-fermenting strains with NIR spectroscopy and model development for xylose quantification based on the spectra of various *S. cerevisiae* strains' supernatant has never been reported.

In this study, we have developed a quantitative analytical method using NIR spectroscopy as alternative tool for evaluating fermentation performance of recombinant *S. cerevisiae* that could ferment mixed sugars, glucose and xylose. Supernatant samples of fermentation solutions were investigated, as half of recombinant strains have flocculation properties that causes light scattering from cell gathering and also to duplicate the analytical conditions of the reference HPLC method. As a result of qualitative analysis, NIR spectra acquired from diluted supernatants of fermentation solutions facilitated the differentiation of strains by their phenotypic and genotypic features. Based on the spectral variations of each strain, robust regression model was developed to include various strains' features and fermentation conditions. The PLS regression model for components of interest constructed from NIR spectral data of all strains could determine simultaneously consumption of glucose and xylose substrates, yield of ethanol product, and glycerol and xylitol byproducts.

MATERIALS AND METHODS

Yeast Strains. Recombinant xylose-fermenting *S. cerevisiae* strains, MN8140X, MT8-1X,[7] NBRC1440X, and YPH499XU were constructed in the following way. Plasmid pIUX1X2XK[6] was digested with *Pst*I and the linear product was transformed into haploid strains NBRC1440Δ*H*UWL[19] and YPH499,[20] to yield NBRC1440X and YPH499XU, respectively. Plasmids pRS403 (Stratagene, La Jolla, CA) was digested with *Msc*I and the linearized product was transformed into strain MT8−1X to yield MT8−1X/pRS403. Diploid strain MN8140X was constructed by mating haploid strains MT8−1X/pRS403 and NBRC1440X. The strains were incubated in YPD medium (10 g/L of yeast extract, 20 g/L of Bacto-peptone (Difco Laboratories, Detroit, MI), and 20 g/L of glucose). After preincubation for 24 h, the two haploid strains were harvested and spread together on YPD medium plates containing 20 g/L of agar. After incubation for 72 h at 30 °C, the plates were replicated onto SD medium plates (6.7 g/L of yeast nitrogen base without amino acids (Difco Laboratories) and 20 g/L of glucose) and subsequently incubated for 3 days at 30 °C to yield a diploid strain. The resulting diploid MN8140X strain formed a single colony on the SD plate.

Fermentation. The recombinant *S. cerevisiae* strains were cultivated aerobically in YPD medium for 48 h at 30 °C. The cells were collected by centrifugation at 3000*g* for 10 min at 4 °C and washed with sterilized distilled water two times. Fermentation was performed at 30 °C with mild agitation (at 500 rpm) in 100 mL bottles exhausting CO_2 in YP medium (10 g/L of yeast extract, 20 g/L of Bacto-peptone) containing reciprocally 10−90 g/L glucose and 90−10 g/L xylose to a total volume of 100 g/L. Initial cell concentration was prepared at 50 g wet weight/L.

Yeast cells suspension was collected by regular samplings from the fermentation. The collected samples were centrifuged at 6000 rpm at 4 °C for 5 min to separate the supernatants. Supernatant samples, a total of 266, were collected from MN8140X (9 clones), NBRC1440X (6 clones), MT8-1X (5 clones), and YPH499XU (5 clones) at the 72-h, 48 h, 48 h, and 48 h period, respectively. After 2-fold dilution with Milli-Q water (Milli-pore, Bedford, MS), the supernatants were subjected to HPLC and NIR spectroscopy analyses.

HPLC Analysis. HPLC system equipped solvent delivery module LC-20AB (Prominence, Shimadzu, Kyoto, Japan) was operated with Milli-Q water (0.6 mL/min of flow rate) as mobile phase. A Shim-pack SPR-Pb column with 250 mm length and 7.8 mm inner diameter (Shimadzu, Kyoto, Japan) was used at 80 °C with a refractive index detector (model RID-10A, Shimadzu) kept at 40 °C. The instrument manufacturer software LC solution (version 1.25, Shimadzu) drove the injection of 10 μL by the sampling module (SIL-20AC, Shimazu) and recorded chromatogram of each sample. The calibration model was constructed using authentic standards for glucose, xylose, ethanol, glycerol, and xylitol expressed in g/L with the provided by the software LC solution.

NIR Spectroscopy Measurements. Transmittance spectra of diluted supernatant samples were acquired with NIR Systems 6500 spectrophotometer (Foss NIR Systems, Laurel, MD) using quartz cuvette cell with 1 mm path length after warming the samples at 30 °C for 1.5 min with the instrument heater inbuilt in the sample transport module. Each spectrum was recorded from the average of 32 scans for 30 s built up at 2 nm intervals in a wavelength region from 1100 to 2498 nm. The spectrum for each sample was consecutively recorded three times, and the three spectra were averaged for further analysis.

Model Development and Validation. To assess the prediction ability of analytical models, external validation was performed using independent prediction set. The whole data was systematically split into two groups sorted by fermentation time: for internal validation as a calibration set and for external validation as a prediction set. The samples number in the calibration and prediction set was 81 and 27 spectra for MN8140X, 51 and 16 spectra for NBRC1440X, 34 and 11 spectra for MT8-1X, and 35 and 11 spectra for YPH499X, respectively.

Qualitative analysis with soft independent modeling of class analogy (SIMCA) method was performed to investigate the difference of each spectral data obtained from each strain. SIMCA method is based on principal component analysis (PCA),[21−23] where PCA is applied to classify each group, that is, spectral data of each strain in this study, as scores in principle component (PC) space. Each sample is evaluated whether it belongs to certain PC space or not. In order to visualize groups' classification, the

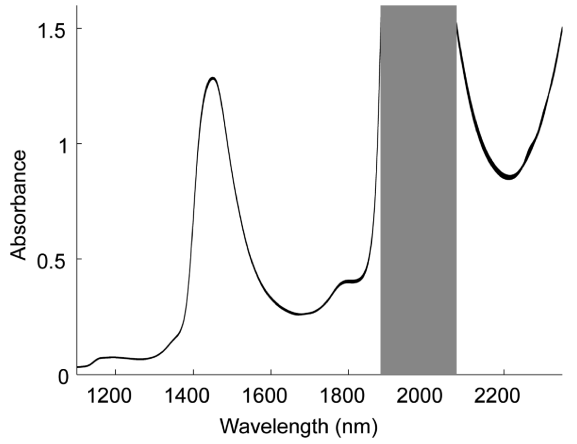


Figure 1. Raw NIR spectra acquired from 2-fold diluted fermentation supernatant samples. Spectra were acquired in transmittance mode with 1 mm path length at 30 °C and analyzed over 1100−2350 nm region with 2 nm step.

residual standard deviation of each sample calculated from PC space is presented for two groups in traditional Coomans plot. The residual standard deviation of each group is converted to interclass distance using PC space of the other group, and can be also used as a criterion to investigate how each group separates from the others. In this study, 5% level of significance threshold was used for *F* test in SIMCA method. As data pretreatment, mean centering was applied.

All the regression models were developed using partial least-squares (PLS) regression method.[24] Development of all PLS regression models was carried out by Pirouette software (version 4.0, Infometrix, Bothell, WA). For the determination of optimal number of latent variables (LVs) to construct robust regression model, a calculation tool using *F* test, that is included in the software, was applied to predict the residual error sum of squares (PRESS) calculated from leave-one-out cross-validation. The coefficient of determination (R^2) or the standard error of cross-validation (SECV) defined by leave-one-out method and standard error of prediction (SEP) were used to assess the prediction quality of regression models. Additionally, bias, which was calculated from average of differences between reference concentration data acquired with HPLC system and predicted data by PLS regression, was used for assessment of regression performance. Overfitting of regression models by the use of a large number of LVs was prevented by direct orthogonal signal correction (DOSC).[25] Only one DOSC component was applied to the data set after mean centering.

RESULTS AND DISCUSSION

Evaluation of NIR Spectroscopy and HPLC Method for Fermentation Samples. Raw NIR spectra of diluted supernatants collected from fermentation solutions are shown in Figure 1. The peaks at around 1450 and 1940 nm are strong water absorbance bands.[26] The absorption band in 1880−2080 nm region (gray box) was excluded from further analysis because of the strong saturated absorption of water considered as noise. The region 2350−2498 nm was also excluded because of the high detector sensitivity to strong absorption in IR neighboring region. Therefore, the regions 1100−1880 and 2080−2350 nm were used for analysis.

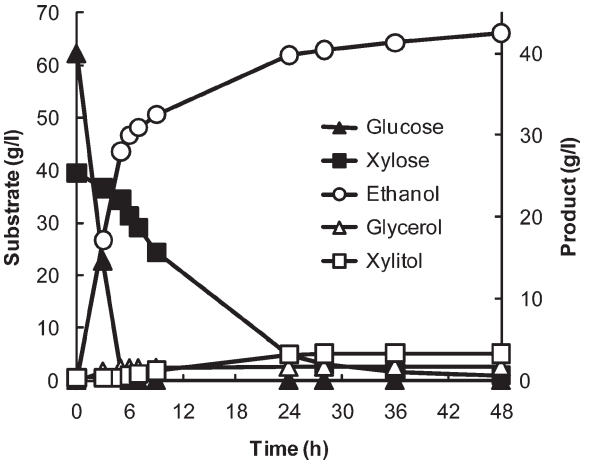


Figure 2. Time course of fermentation process from mixed sugar source, glucose, and xylose, by MN8140X strain. The product of interest, ethanol (open circle) and side products glycerol (open triangle) and xylitol (open square) were produced from glucose (close triangle) and xylose (close square). Note the steep decrease of glucose (preferred energy source) and much slower consumption rate of xylose, as well as sharp increase of ethanol production and relatively stable low concentration of byproducts glycerol and xylitol.

Table 1. Statistics for the Rate of Substrate Consumption and Product Formation in Glucose and Xylose Fermentation During 3 h

	MN8140X		NBRC1440X		MT8−1X		YPH499XU	
components	mean	SD	mean	SD	mean	SD	mean	SD
glucose (g/l/h)	11.34	3.06	8.56	2.15	7.34	2.43	7.19	1.62
xylose (g/l/h)	1.94	1.07	1.30	0.71	1.39	1.61	1.10	0.65
ethanol (g/l/h)	5.42	1.90	4.10	0.98	3.56	0.77	3.45	1.01
glycerol (g/l/h)	1.03	0.36	0.61	0.34	1.14	0.27	0.95	0.31
xylitol (g/l/h)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

An example of fermentation process by MN8140X strain in YP medium containing 60 g/L of glucose and 40 g/L of xylose is shown in Figure 2. Consumption rate of glucose was greater than that of xylose, and production yield of xylitol was greater than that of glycerol. However, the fermentation ability such as consumption of substrates (glucose and xylose) and formation of products (ethanol, glycerol and xylitol) was different among strains. The statistics, mean and standard deviation (SD), for the rate of substrate consumption and product formation by each strain during 3 h of fermentation is shown in Table 1. No xylitol production was detected in all strains for 3 h, MN8140X strain showed superior fermentation ability in glucose and xylose consumption, and ethanol production compared with other strains.

Analysis of the Differences among Strains. In previous reports, it was shown that the robustness of the regression model was affected by various external parameters, such as the different strain data.[27,28] In other words, if strain difference could be detected by NIR analysis, it could be further used for feedback to speed up the strain construction and selection process. In order to investigate the differences among each strain samples and reflect them in regression analysis, SIMCA was performed for NIR spectral data as qualitative analysis. SIMCA classification

Table 2. Interclass Distances by SIMCA Classification for Calibration Sets of Four Strains

actual strain	MN8140X	NBRC1440X	MT8-1X	YPH499XU
MN8140X	0			
NBRC1440X	1.19	0		
MT8−1X	2.34	2.07	0	
YPH499XU	1.53	1.15	0.51	0

was applied to the spectral data split into individual calibration and prediction sets after mean centering. The wavelength region applied for SIMCA was 1100−1880 nm and 2080−2350 nm combined. The number of PCs used for each strain's data space was determined by the cumulative proportion of variance in spectral data greater than 99.5%, in which the cumulative proportion variance among each strain was less than 95%, in order to give even information on each strain to each PC space. Thus, the number of applied PCs was four, four, five, and five for strain MN8140X, NBRC1440X, MT8-1X, and YPH499XU spectral data, respectively.

In calibration results, interclass distances were calculated by PC model in SIMCA method between two strains for each strain's spectral data. The interclass distance between diploid MN8140X and haploid MT8−1X strain was the greatest value of 2.34 (Table 2). According to previous reports,[21,23] the interclass distance of less than 0.8 indicates a small difference between two groups while the distance of more than 3 supports a reliable separation. Therefore, these two strains were considered as well separated among four strains. The closest haploid strain to the diploid strain MN8140X was NBRC1440X in terms of interclass distances. Furthermore, SIMCA prediction results gave 96.3% (26/27 samples) of correct classifications for diploid strain MN8140X spectral data. Haploid strains were identified with accuracy of 75.0% (12/16 samples) for NBRC1440X and 63.6% (7/11 samples) for MT8−1X and YPH499XU. Samples of haploid strains were misclassified as other haploid strains but not as a diploid strain (Table S-1, Supporting Information). These results indicate that the difference of fermentation ability among individual strains (shown in Table 1) could be reflected in the spectral data of each strain. The diploid strain MN8140X that has superior fermentation ability was significantly distinguished from other haploid strains by SIMCA classification for spectral data. A Coomans like plot for prediction set (shown in Figure 3) was constructed by SIMCA classification with distances to MT8-1X versus distance to NBRC1440X, which enabled to clearly visualize the classification of four strains. The plot indicates that some diploid strain MN8140X samples were closer to the haploid strain NBRC1440X. It is noteworthy to mention that MN8140X resembles NBRC1440X as they both exhibit flocculation properties. On the other hand, MT8-1X and YPH499XU are haploid laboratory strains without flocculation properties. Expectedly, the closest interclass distance in the calibration set was calculated between these two strains (Table 2), and on the Coomans like plot for prediction set they overlapped (Figure 3). These two haploid strains can be considered most similar by the evaluation of SIMCA analysis using spectral data.

The ploidy of strain plays role in the fermentation ability based on the efficiency of metabolism. Therefore, spectral data can be influenced by different amount of exocellular intermediate product. Also, because the flocculation of yeast strains is caused by the interaction between a lectine-like protein and a mannose

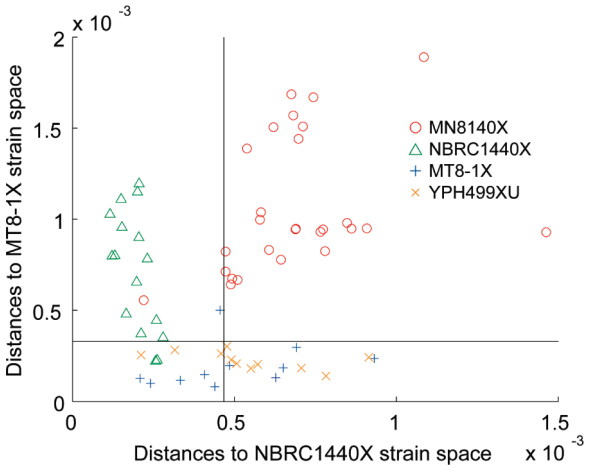


Figure 3. SIMCA classification results visualized by Coomans like plot showing computed distances between individual strains' data sets in prediction set. MN8140X (circle), NBRC1440X (triangle), MT8-1X (cross), and YPH499XU (diagonal cross). The diploid strain MN8140X is with highest distance, while haploid and nonadhesive MT8-1X and YPH499XU strains have the lowest distance and appear in the same cluster.

chain located on the yeast cell surface,[29] flocculation property influences fermentation supernatant composition and thus the spectral information. Reliable classification results representing those features of recombinant yeasts were obtained by SIMCA method applied to spectral data, since water configurations changes due to interaction with intermediate products in aqueous media appear in NIR spectral data.[30] Furthermore, the strains classification rates in the prediction set were consistent with the evaluation of interclass distances calculated from the calibration spectral data. Therefore, it can be concluded that genotypic and phenotypic information as that for ploidy and flocculation properties can be expressed by the strains' NIR spectral data.

PLS Model for Components Quantification in Mixed Sugars Fermentation. As a first step, fermentation of each strain was separately analyzed and glucose, xylose, ethanol, xylitol, and glycerol were measured by individual PLS regression models. To prevent possible overfitting when developing the models, calibration set was designed to cover wider components range in prediction set (Table S-2, Supporting Information), since extrapolation of reference data in the prediction domain can lower the prediction accuracy due to overfitting in the regression model.[31] In addition, DOSC with one component was applied as data treatment after mean centering to prevent overfitting by a lot of LVs. The most accurate prediction results with least SEP are shown in Table 3. Three spectral regions, 1100−1880 nm and/or 2080−2350 nm, were evaluated and the best region for each component in each individual model is presented. The number of selected LVs was different for each component among different strains. Optimal wavelength region was not unified for each component and different for individual models (Figure S-1A to S-1E, Supporting Information).

The difference in strains' spectral data can affect the regression performance when using model constructed on other strain's data. In order to investigate the differences in models' performance among different strains and see if the models can be interchangeable, regression model constructed using one strain's data set was applied to prediction of other strains' components based on their

Table 3. Prediction Results for Each Strain Data Set by PLS Regression Model Constructed from the Same Strain's Data[a]

	MN8140X				NBRC1440X				MT8−1X				YPH499XU			
components	no. of LVs	R^2	SEP (g/L)	bias (g/L)	Lno. of Vs	R^2	SEP (g/L)	bias (g/L)	no. of LVs	R^2	SEP (g/L)	bias (g/L)	no. of LVs	R^2	SEP (g/L)	bias (g/L)
glucose	3[c]	0.998	1.25	0.38	4[c]	0.999	0.98	0.13	2[d]	0.998	0.73	0.00	4[d]	0.995	1.10	0.03
xylose	2[d]	0.996	1.12	−0.51	4[b]	0.994	1.31	0.37	9[c]	0.997	1.25	0.02	5[b]	0.988	1.70	−0.59
ethanol	6[b]	0.988	1.64	−0.45	2[d]	0.994	1.33	−0.15	5[b]	0.998	0.55	−0.11	4[d]	0.994	1.48	1.18
glycerol	2[c]	0.976	0.39	−0.06	2[c]	0.989	0.25	−0.03	4[d]	0.994	0.23	−0.11	4[d]	0.995	0.20	0.15
xylitol	2[c]	0.963	0.41	0.0	2[c]	0.989	0.39	0.08	10[c]	0.995	0.36	−0.07	9[c]	0.993	0.36	−0.01

[a] Average of glucose concentrations in prediction set samples: 14.19 g/L. Average of xylose concentrations: 27.97 g/L. Average of ethanol concentrations: 28.17 g/L. Average of glycerol concentrations: 4.48 g/L. Average of xylitol concentrations: 3.79 g/L. [b] Applied wavelength region is 1100−1880 nm. [c] Applied wavelength region: 1100−1880 and 2080−2350 nm. [d] Applied wavelength region: 2080−2350 nm.

Table 4. Prediction Results for Other Strains Data Sets Based on PLS Regression Model of NBRC1440X Strain (Calibration Model—NBRC1440X, Prediction Set—Individual Strain)[a]

		MN8140X			MT8−1X			YPH499XU		
components	no. of LVs	R^2	SEP(g/L)	bias (g/L)	R^2	SEP (g/L)	bias (g/L)	R^2	SEP (g/L)	bias (g/L)
glucose	4[c]	0.995	2.43	1.68	0.997	2.17	−0.99	0.993	1.54	0.89
xylose	4[b]	0.951	5.77	−4.41	0.993	3.76	−3.43	0.981	5.05	−4.47
ethanol	2[d]	0.984	1.81	−0.02	0.989	1.85	−2.23	0.983	1.70	−2.46
glycerol	2[c]	0.670	1.42	−0.41	0.952	0.78	0.54	0.960	0.55	0.28
xylitol	2[c]	0.742	1.32	0.49	0.987	1.26	−0.98	0.987	1.10	−0.90

[a] Average of glucose concentrations in prediction set samples: 12.67 g/L. Average of xylose concentrations: 32.48 g/L. Average of ethanol concentrations: 21.17 g/L. Average of glycerol concentrations: 4.54 g/L. Average of xylitol concentrations: 3.86 g/L. [b] Applied wavelength region is 1100−1880 nm. [c] Applied wavelength region: 1100−1880 and 2080−2350 nm. [d] Applied wavelength region: 2080−2350 nm.

Table 5. Prediction Results of PLS Regression Constructed from All Strains' Data[a]

		MN8140X			NBRC1440X			MT8−1X			YPH499XU		
components	no. of LVs	R^2	SEP (g/L)	bias (g/L)	R^2	SEP (g/L)	bias (g/L)	R^2	SEP (g/l)	bias (g/L)	R^2	SEP (g/L)	bias (g/L)
glucose	4[b]	0.998	1.08	0.03	0.999	0.82	−0.09	0.996	1.17	0.50	0.998	0.73	0.17
xylose	4[b]	0.996	1.19	−0.17	0.994	1.28	0.06	0.994	1.37	−0.19	0.994	1.18	−0.29
ethanol	4[c]	0.991	1.47	−0.51	0.992	1.49	−0.24	0.994	1.02	0.11	0.993	1.07	0.54
glycerol	2[c]	0.977	0.42	0.17	0.980	0.37	−0.04	0.991	0.26	0.04	0.991	0.21	0.04
xylitol	3[c]	0.976	0.34	0.03	0.991	0.33	−0.06	0.993	0.50	0.03	0.998	0.23	−0.10

[a] Average of glucose concentrations in prediction set samples: 14.19 g/L. Average of xylose concentrations: 27.97 g/L. Average of ethanol concentrations: 28.17 g/L. Average of glycerol concentrations: 4.48 g/L. Average of xylitol concentrations: 3.79 g/L. [b] Applied wavelength region is 1100−1880 nm. [c] Applied wavelength region: 1100−1880 and 2080−2350 nm.

respective spectra. The model with the widest concentration range of almost all components, that for NBRC1440X strain, was selected as a representative regression model to prevent overffiting by extrapolation of reference data in the prediction set.[31] The prediction results for other strains by the NBRC1440X regression model are shown in Table 4. Prediction accuracy has decreased for some components (particularly glycerol and xylitol for MN8140X strain) and SEP has increased in all cases compared with the prediction results of individual models.

When using quantification by regression model built on one strain for other strains, the limitations of regression performance for other strains have become evident. As mentioned above (in Analysis of the Differences among Strains section), the difference in fermentation components of each developed recombinant strain is reflected in the NIR spectral data.[30] Therefore, a new PLS model was constructed using all four strains spectral data split into calibration set (201 spectra), validated by leave-one-out method as internal validation, and evaluated by independent prediction set (65 spectra). DOSC with one component was applied to the spectral data to prevent overfitting after mean centering. The most accurate results for each component achieved by the new model are shown in Table 5. Similarly high or even higher accuracy results were obtained compared with results based on individual strains data (R^2 from 0.976 to 0.998, SEP from 0.21 to 1.49 g/L and bias from −0.51 to 0.54 g/L, Table 5), and the optimal wavelength region for each component was unified at 1100−1880 nm for substrates (glucose and xylose), and 1100−1880 and 2080−2350 nm for product (ethanol) and byproducts (glycerol and xylitol), respectively. The average concentration of each component in the prediction set samples was

dx.doi.org/10.1021/ac103128p |Anal. Chem. 2011, 83, 4023–4029

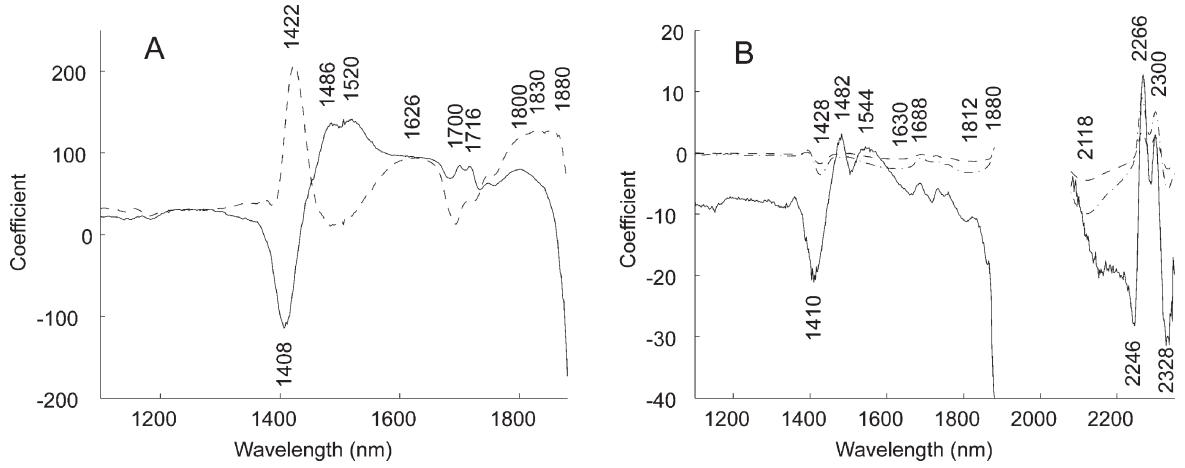


Figure 4. PLS regression vectors for the prediction of glucose (dash line) and xylose (solid line) in the region 1100−1880 nm (A), and ethanol (solid line), glycerol (dash line) and xylitol (dash-dot line) in the region 1100−1880 and 2080−2350 nm (B). The regression vectors are based on the all strains data sets. Note the distinct ethanol pattern and overall similar regression vectors for glycerol and xylitol. Note also the inverse pattern in glucose/xylose regression vectors in the region of 1350−1760 nm. The informative absorbance bands (high intensity peaks) for each regression vector are identified, in nanometers. The most prominent absorbance band common for all vectors is that at 1410 nm, attributed to water.

14.19 g/L for glucose, 27.97 g/L for xylose, 23.17 g/L for ethanol, 4.48 g/L for glycerol and 3.79 g/L for xylitol. Thus the SEP values in the presented combined model of all strains are suitable for practical application.

The figures of merit were determined for calibration and prediction performance of PLS model for each fermentation component by limit of detection (LOD) and quantification (LOQ). These parameters were estimated using the following equations.[32]

$$\mathrm{LOD} = 2t_{1-\alpha,\nu} \times \mathrm{SECV(or\ SEP)}$$

$$\mathrm{LOQ} = 10 \times \mathrm{SECV(or\ SEP)}$$

Where $t_{1-\alpha,\nu}$ represents a statistical parameter for the Student t test with one-sided tails of the distribution in false positive error of α% probability with ν degrees of freedom. The estimated LOD was calculated as α = 5% and ν = 201 degree of freedom for calibration set with SECV or ν = 65 for prediction set with SEP. The regression model performance was evaluated as LOD of 1.0−4.4 g/L and LOQ of 3.0−13.3 g/L for calibration, and LOD of 1.2−4.5 g/L and LOQ for 3.5−13.4 g/L for prediction (Table S-3, Supporting Information).

NIR analysis allowed simultaneous precise quantification of all components from diluted supernatant samples without any further sample preparation. The validity of the figures of merit for regression performance was confirmed in the correlation plots (Figure S-2, Supporting Information).

The regression vectors for substrates and products, based on the unified new model are shown in Figure 4. The bands at 1410 and 1884 nm are considered as related to O−H overtones due to water absorbance, as described in wine analysis study,[33] also 1410 nm is described as the band of free water molecular species well scattered within the range.[30] The highest coefficients in 1100−1880 nm region for the constructed PLS models were obtained around these wavelengths for all components.

In a previous report, the regions 1480−1580 nm and 2030−2130 nm were selected for glucose detection based on wavelength selection.[13] On the other hand, Yano et al. have reported that for glucose prediction in aqueous solution of blood anticoagulant the use of 1674 and 2274 nm led to higher regression accuracy compared with model based only on 2274 nm.[34] In this study using mixed sugars, a model based on 1100−1880 nm was found optimal for glucose and xylose prediction. Regression coefficients around 1626 and 1830 nm were higher for glucose. The regions 1486−1520 nm and 1700−1716 nm were calculated to be highly informative for xylose quantification (Figure 4A). Thus it can be concluded that even though the spectral range in each study is slightly different, the middle NIR range appears as highly informative for sugars identification.

In previous reports, the ranges 1660−1780 nm and 2200−2360 nm were regarded as the regions for absorption bands of ethanol,[13] and the bands at 2266 and 2300 nm were associated with the CH_2 group of ethanol in wine analyses.[33] For quantification of glycerol, the combined region of 1696−1706 nm, 2072−2098 nm and 2264−2284 nm was chosen in a glycerol and methanol fermentation process.[17] Moreover, 1688 or 2264 nm were selected for glycerol calibration in acidified biodiesel wastewater.[16] As shown in Figure 4B, intensive peaks similarly appeared around 2266 and 2300 nm in the regression coefficient plot for products. The former peak was shifted onto 2268 nm for ethanol, 2264 nm for glycerol, and 2266 nm for xylitol. In this study, the effective wavelength for glycerol corresponded to previous reports.[16,17] The region around 1688 nm was effective for ethanol. The intensive peaks for xylitol except the peaks mentioned above appeared at 1812 nm, 2118 and 2334 nm. The other peaks for each product were found in the region of 1482−1880 nm. Therefore, it can be assumed that the widely applied region combining 1100−1880 nm and 2080−2350 nm could be a useful one for the measurement of ethanol, glycerol, and xylitol.

CONCLUSION

In this paper, we demonstrate the use of NIR spectroscopy for successful quantification of multiple components in fermentation process consumed and produced by a number of recombinant *S. cerevisiae* strains from mixed-sugar environment of glucose and xylose. For the first time, phenotypic and genotypic differences among genetically modified strains were successfully extracted and visualized by qualitative analysis based only on NIR spectral data of diluted supernatant samples acquired from fermentation

solution along the process of fermentation. When individual strains were used for quantification of fermentation components, we pointed out the vulnerability of the regression models constructed from such single-strain data sets when applied to other strains data. Therefore, it is suggested that accurate regression model should be constructed from plural number of strains that have different genotypic and phenotypic features. Accordingly, four kinds of strain data lead to a robust regression model. Thus, NIR spectral data provided multifactor information that allows not only qualitative but also quantitative analysis. These results prove that NIR spectroscopy is an effective and highly accurate screening method, suitable as high-throughput technology for decision making in biotechnological development of new industrial recombinant strains able to efficiently ferment glucose and xylose into ethanol from cellulosic biomass.

■ ASSOCIATED CONTENT

Supporting Information. Figures S-1 and S-2 and Tables S-1−S4. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone/Fax: +81-78-803-5911 (R. T.); +81-78-803-6196 (A. K.). E-mail: rtsen@kobe-u.ac.jp (R. T.); akondo@kobe-u.ac.jp (A. K.).

■ ACKNOWLEDGMENT

This work was supported by project P07015 from the New Energy and Industrial Technology Development Organization (NEDO), under the sponsorship of the Ministry of Economy, Trade, and Industry (METI) of Japan.

■ REFERENCES

(1) Hahn-Hägerdal, B; Galbe, M; Gorwa-Grauslund, M. F.; Lidén, G.; Zacchi, G. *Trends Biotechnol.* **2006**, *24*, 549–556.
(2) Ragauskas, A. J.; Williams, C. K.; Davison, B. H.; Britovsek, G.; Cairney, J.; Eckert, C. A.; Frederick, W. J., Jr.; Hallett, J. P.; Leak, D. J.; Liotta, C. L.; Mielenz, J. R.; Murphy, R.; Templer, R.; Tschaplinski, T. *Science* **2006**, *311*, 484–489.
(3) Jeffries, T. W. *Curr. Opin. Biotechnol.* **2006**, *17*, 320–326.
(4) Nevoigt, E. *Microbiol. Mol. Biol. Rev.* **2008**, *72*, 379–412.
(5) Katahira, S.; Fujita, Y.; Mizuike, A.; Fukuda, H.; Kondo, A. *Appl. Environ. Microbiol.* **2004**, *70*, 5407–5414.
(6) Katahira, S.; Mizuike, A.; Fukuda, H.; Kondo, A. *Appl. Microbiol. Biotechnol.* **2006**, *72*, 1136–1143.
(7) Katahira, S.; Ito, M.; Takema, H.; Fujita, Y.; Tanino, T.; Tanaka, T.; Fukuda, H.; Kondo, A. *Enzyme Microb. Technol.* **2008**, *43*, 115–119.
(8) Saitoh, S.; Hasunuma, T.; Tanaka, T.; Kondo, A. *Appl. Microbiol. Biotechnol.* **2010**, *87*, 1975–1982.
(9) Tanino, T.; Hotta, A.; Ito, T.; Ishii, J.; Yamada, R.; Hasunuma, T.; Ogino, C.; Ohmura, N.; Ohshima, T.; Kondo, A. *Appl. Microbiol. Biotechnol.* **2010**, *88*, 1215–1221.
(10) Cavinato, A. G.; Mayes, D. M.; Ge, Z.; Callis, J. B. *Anal. Chem.* **1990**, *62*, 1977–1982.
(11) Blanco, M.; Peinado, A. C.; Mas, J. *Anal. Chim. Acta* **2006**, *556*, 364–373.
(12) Finn, B.; Harvey, L. M.; McNeil, B. *Yeast* **2006**, *23*, 507–517.
(13) Blanco, M.; Peinado, A. C.; Mas, J. *Biotechnol. bioeng.* **2004**, *88*, 536–542.
(14) González-Sáiz, J. M.; Pizarro, C.; Esteban-Díez, I.; Ramírez, O.; González-Navarro, C. J.; Sáize-Abajo, M. J.; Itoiz, R. *J. Agric. Food Chem.* **2007**, *55*, 2930–2936.
(15) Liebmann, B.; Friedl, A; Varmuza, K. *Anal. Chim. Acta* **2009**, *642*, 171–178.
(16) Kohda, J.; Ooshita, K.; Nakano, Y.; Yano, T. *J. Near Infrared Spectrosc.* **2008**, *16*, 199–204.
(17) Crowley, J.; Arnold, S. A.; Wood, N.; Harvey, L. M.; McNeil, B. *Enzyme Microb. Technol.* **2005**, *36*, 621–628.
(18) Arnold, S. A.; Crowley, J.; Vaidyanathan, S.; Matheson, L.; Mohan, P.; Hall, J. W.; Harvey, L. M.; McNeil, B. *Enzyme Microb. Technol.* **2000**, *27*, 691–697.
(19) Yamada, R.; Bito, Y.; Adachi, T.; Tanaka, T.; Ogino, C.; Fukuda, H.; Kondo, A. *Enzyme Microb. Technol.* **2009**, *44*, 344–349.
(20) Sikorski, R. S.; Hieter, P. *Genetics* **1989**, *122*, 19–27.
(21) Blomquist, G.; Johansson, E.; Söderström, B.; Wold, S. *J. Anal. Appl. Pyrolysis* **1979**, *1*, 53–65.
(22) Derde, M. P.; Coomans, D.; Massart, D. L. *Anal. Chim. Acta* **1982**, *141*, 187–192.
(23) Kvalheim, O. M.; Karstang, T. V. In *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*; Brereton, R. G., Ed.; Elsevier: Amsterdam, The Netherlands, 1992; pp 209−248.
(24) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley and Sons: Chichester, 1989; pp 116−119.
(25) Westerhuis, J. A.; de Jong, S.; Smilde, A. K. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 13–25.
(26) Williams, P. C. In *Near-Infrared Technology in the Agricultural and Food Industries*, 2nd ed.; Williams, P., Norris, K., Eds.; American Association of Cereal Chemists: St. Paul, MN, 2001; pp 145−169.
(27) Peirs, A.; Tirry, J.; Verlinden, B.; Darius, P.; Nicolaï, B. M. *Postharvest Biol. Technol.* **2003**, *28*, 269–280.
(28) Roger, J.-M.; Chauchard, F.; Williams, P. *J. Near Infrared Spectrosc.* **2008**, *16*, 311–315.
(29) Miki, B. L. A.; Poon, N. H.; James, A. P.; Seligy, V. L. *J. Bacteriol.* **1982**, *150*, 878–889.
(30) Tsenkova, R. *J. Near Infrared Spectrosc.* **2009**, *17*, 303–314.
(31) Estienne, F.; Pasti, L.; Centner, V.; Walczak, B.; Despagne, F.; Jouan Rimbaud, D.; de Noord, O. E.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 195–211.
(32) Currie, L. A. *Pure Appl. Chem.* **1995**, *67*, 1699–1723.
(33) Yu, H.; Lin, H.; Xu, H.; Ying, Y.; Li, B.; Pan, X. *J. Agric. Food Chem.* **2008**, *56*, 307–313.
(34) Yano, T.; Funatsu, T.; Suehara, K.; Nakano, Y. *J. Near Infrared Spectrosc.* **2001**, *9*, 43–48.