근적외선분광분석에 의한 육성계통 벼 유전자원의 아밀로스 및 단백질 성분함량에 관한 통계분석

오세종* · 최유미 · 윤혜명 · 이명철 · 신명재 · 유은애 · 현도윤 · 채병수

농촌진흥청 국립농업과학원 농업유전자원센터

Statistical Analysis of Amylose and Protein Content in Breeding Line Rice Germplasm Collected from East Asian Countries Based on Near-infrared reflectance spectroscopy

Sejong Oh*, Yu Mi Choi, Hyemyeong Yoon, Sukyeung Lee, Myung Chul Lee, Myoung-Jae Shin, Eunae Yoo, Do Yoon Hyun, and Byungsoo Chae

National Agrobiodiversity Center, NAS, RDA, Jeonju 54874, Republic of Korea

Abstract A statistical analysis of 9,771 non-glutinous rice in breeding line germplasm collected from Korea (2,836), China (2,136), Japan (1,219), and the Philippines (1,213) was conducted using normal distribution, variability index value (VIV), analysis of variation (ANOVA) and Duncan's multiple range test (DMRT) based on the data obtained from NIRS analysis. According to the normal distribution, the average protein content was 7.9%, and non-glutinous rice ranging over 10% amylose had 23.6% average content. Most resources were between 5.3 and 10.5% in protein content, and 15.7 and 31.5% in amylose content. The VIV was 0.54 for protein, and 0.83 for amylose. The average amylose content was 25.18%, 24.54%, 22.08%, and 21.47% in Filipino, Chinese, Korean, and Japanese resources, respectively, whereas the average protein content was found to be 8.19%, 7.79%, 7.58%, and 7.42% in Filipino, Chinese, Korean, and Japanese resources, respectively. The ANOVA of amylose and protein content showed significant differences at the level of 0.01. The F-test value was 412.2 for amylose content, and 108.4 for protein when compared with the critical value of 3.78. The DMRT of amylose and protein content showed significant differences ($p < 0.01$) among resources from different countries. The Filipino resources had the highest level of amylose and protein content, whereas; the lowest level of amylose and protein content were found in Japanese when compared with resources of other origins. These results are recommended as helpful materials in the field of breeding.

Keywords Amylose, ANOVA, DMRT, Protein, NIRS, Non-glutinous, Rice germplasm

Received on July 9, 2019. Revised on July 22, 2019. Accepted on August 21, 2019.

* Corresponding Author (E-mail: pleurotus@korea.kr Tel: +82-63-238-4910, Fax: +82-63-238-4909)

서 언

쌀(*Oryza sativa* L.)은 세계 3대 곡물중의 하나이며 우리나라를 포함한 아시아 지역에서 전 세계 쌀생산량의 90% 이상이 생산되는 중요한 식량원이다(Juliano 2003, Kim et al. 2012). 쌀의 주요성분은 아밀로스와 단백질로 구성되며 품종에 따라 성분 함량이 다르다. 쌀을 주식으로 하는 나라에서는 이들 성분의 특정 함량을 가진 벼 품종을 선호하는 경향이 있으며, 따라서 벼 품종육성분야에 있어 아밀로스 성분이 중요 지표 자료로 활용되기도 한다(Kim 2004). 아밀로스와 단백질은 쌀의 미질을 결정하는 중요한 이화학적 특성이고 품질관련변수이며 취반특

성 뿐만 아니라 물리적 특성에도 영향을 준다(Champagne et al. 1997, Son et al. 2002).

Near-Infrared Reflectance Spectroscopy (NIRS)는 시료 내의 특정 작용기가 근적외선 영역의 특정 파장 의 빛 에너지를 흡수하여 분자진동을 일으키며 이러한 흡광현상의 크기는 작용기들의 함량에 비례한다는 램버트 비어 법칙에 근거한 정량분석방법이다. 농산품의 경우 주요구성 성분인 지방(C-H), 수분(O-H), 단백질(N-H, S-H)이 근적외선을 흡수하므로 근적외선 분광분석법을 이용해 이들 성분의 동시 분석이 가능하다(Williams et al. 1987). 시료 내 다양한 성분들의 근적외선 흡수과장대가 서로 중첩되는 현상 또는 시료 입자 크기, 밀도 등과 같은 물리적 요인들에 의해

재료 및 방법

시험자원 분석 및 처리

본 실험에 사용한 국내외 육성계통 벼 유전자원은 농업유전자원센터 GMS 프로그램을 이용하여 한·중·일 등 57개국에서 수집된 자원을 대상으로 IT Number 기준으로 정렬하여 9,998자원을 선발하였고, 선발자원 각각의 종자 50립을 실험실용 현미기로 탈영시킨 후 사미, 피해립을 제거하여 현미립을 준비하였다. 이후 분쇄기(Bistro electric coffee grinder, Bodum®)를 사용하여 26,000 rpm 조건에서 약 2분간 분쇄한 후 300 µm sieve를 통과하는 균일한 현미가루 상태로 만든 후 NIRS를 이용하여 아밀로스 함량과 단백질 함량을 측정하였다. 벼 유전자원에 대한 고유특성을 최대한 반영하기 위하여 현미를 대상으로 분석을 실시하였다. 선발자원 중 아밀로스 함량 10% 이상의 메벼 9,771자원만을 택하여 아밀로스 및 단백질 함량분포 분석과 통계분석을 실시하였다(Table 1). 수분함량 분석은 곡류 및 곡류제품-수분함량측정(산업통상자원부 국가기술표준원, 표준번호: KS H ISO 712)을 참조하여 수분측정기(MOC63u, Shimadzu, Japan)를 사용하여 130°C 온도조건에서 측정하였다. 현미시료들은 13°C 저온저장고에 보관하였으며 수분측정기로 측정된 시료의 수분함량은 14% 이하였다.

NIRS 검량식 작성 및 검증

벼 유전자원의 아밀로스 및 단백질 성분 분석용 NIRS 예측모델을 선행연구에서 개발하고 미지시료 적용가능성을 검증하였다(Oh et al. 2017a). 검량식 작성에 사용되는 벼 자원은 제현, 마쇄와 같은 전처리과정을 거쳤고, NIRS를 이용하여 근적외선

스펙트럼의 변화가 일어난다. 이러한 변화에 의한 분석 오차를 줄이고 상호 중첩된 파장들을 분리하기 위해 NIRS 전용 프로그램을 사용하여 스펙트럼을 보정하는 등의 수학적 전처리 과정을 수행하게 된다(Ahn & Kim 2012). 이후 습식분석결과와 근적외선 스펙트럼간의 상관관계를 회귀분석을 통해 분석하여 검량식 작성 등을 통해 최종적으로 NIRS 예측모델을 개발한다. 개발된 모델이 적용된 NIRS를 이용하여 유전자원의 성분 함량을 측정하고 얻어진 data는 통계 처리 과정을 거쳐 함량 특성 분석에 활용된다.

농업분야에서 많이 사용되는 통계처리방법에는 분산분석(ANOVA)과 Duncan's 다중검정(DMRT)이 있다. 분산분석은 세 개 이상 집단들 간의 평균 차이를 비교하기 위하여 집단 간 분산과 집단 내 분산을 계산하고 이들 상호간의 비를 이용한다. 집단의 표본평균들이 동일한 모집단에서 나온 것인지를 검정하며, 실험에 관련된 요인들 가운데 가장 큰 영향을 끼치는 것을 찾아내는 분석법이다(Pak & Oh 2010, Kim 2016, Park et al. 2016). ANOVA를 통해 집단들 간의 평균들이 유의미하게 차이가 있다는 결과를 얻게 되면 이후 DMRT를 사용하여 각 집단들 간의 평균차이를 상호 비교하여 구체적으로 어느 집단 간에 차이가 있는지를 검정한다. 이러한 통계처리는 연구 결과의 신뢰도를 높일 수 있으므로 다양한 연구 분야에 적용되고 있다. 특정 식물에 존재하는 항산화, 항염증, 항혈전 물질 함량과 활성간의 유의성 검정(Kim & Cha 2017, Sim et al. 2017), 작물의 생장과 환경요인 상호간 영향 평가(Yang et al. 2001, Kim et al. 2010, Sung et al. 2014, Nam et al. 2015, Lee et al. 2017), 작물 생장환경의 인위적 변화 또는 화학물질 처리가 작물 생장에 미치는 영향 평가(Lee et al. 2010, Song 2015, Lim et al. 2015, Choi & Jung 2017), 작물의 잔류 농약 함량 변화 조사(Jang et al. 2011, Son et al. 2013) 등의 연구가 보고되었다.

통계처리방법은 그 의미와 해석에 관한 이해를 돕고 효율적인 적용을 통해 연구결과와 신뢰성을 높이고 객관성을 확보하고자 기초이론을 보완함과 동시에 농업유전자원센터에 보존되어있는 육성계통 벼 유전자원의 아밀로스 및 단백질 함량에 관한 NIRS 성분 분석 자료를 바탕으로 Normal Distribution, Variability Index Value (VIV), ANOVA, DMRT를 수행하였다. 보존자원의 다양성과 성분함량에 따라 자원 집단을 구분하고, 국가별 육성계통 자원들에 대하여 통계분석을 하였으며, NIRS 대량평가체계를 기반으로 자원들 간의 집단분석 결과를 기초자료로 제시하고자 이 연구를 수행하였다.

Table 1. Distribution of breeding line germplasm based on geographical origin.

Origin	Number of accessions	Ratio of accessions (%)
Korea	2,386	24.42
China	2,136	21.86
Japan	1,219	12.48
Philippines	1,213	12.41
Taiwan	238	2.44
India	179	1.83
Thailand	135	1.38
Unknown	1,580	16.17
The others	685	7.01
Total	9,771	100.00

영역의 스펙트럼을 측정하였다. 이후 각 자원의 아밀로스 함량은 요오드 분석법을, 단백질 함량은 켈달 분석법을 사용하여 결정하였다. 측정된 스펙트럼 특성과 분석된 성분 함량 간의 회귀분석을 통해 검량식을 작성하였고 검정과정을 통해 작성된 검량식의 정확성과 대량분석 적용가능성을 평가하였다. NIRS 예측모델의 검량식 작성 과정을 Fig. 1에 나타냈다.

현미가루 상태의 벼 자원에 대한 근적외선 스펙트럼은 실온 조건하의 가시광선 및 근적외선 대역(400-2500 nm)에서 한 자원 당 3반복 scanning하여 얻었다. 이후 검량식 작성을 위하여 벼 자원 별 아밀로스함량 및 조단백질 함량을 입력하여 습식분석값과 근적외선 스펙트럼을 파일형태로 저장하였다. 습식분석값과 NIRS 스펙트럼 간의 통계적 해석과 상관관계분석을 통해 검량식을 작성하기 위해 WINISI III project manager의 global equation 프로그램을 사용하였다. 검량식 작성과 RSQ (R square, R^2)의

이론적 배경은 아래와 같으며 이를 바탕으로 수동으로 검량식을 작성하였고, WINISI III project manager 프로그램을 사용하여 작성된 검량식과 비교하였다.

x, y 좌표상의 실측점 (x_i, y_i) 과 임의의 직선 $y = a + bx_i$ 로부터 설정된 좌표 $(x_i, a + bx_i)$ 와의 거리제곱 합을 S 라 하면

$$S_i = \{y_i - (a + bx_i)\}^2$$

$S = \sum_{i=1}^n \{y_i - (a + bx_i)\}^2$ 이 최소가 되는 a, b 를 구함 (편차자승의 합이 최소가 되도록 함)

따라서 $S = \sum_{i=1}^n \{y_i - (a + bx_i)\}^2$ 이 최소가 되도록 하기 위해서는 이차함수 $f(x) = \{y_i - (a + bx_i)\}^2$ 의 도함수(기울기)가 $f'(x) = 0$ 되도록 설정함 즉, $*S = f(U)^2$ 형태는 합성함수이고 다르게 표시하면 $Y = f(g(x))$ 형태이므로 미분하면 $Y' = f'(g'(x))$ 이와 같은 형식으로 미분된다.

또한 미지수가 a, b 두 개이므로 각각에 대해 미분해야 하기 때문에 편미분을 적용한다.

$$S \text{를 } a \text{에 대해서 편미분하고 } \frac{\partial S}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$b \text{에 대해서 편미분하면 } \frac{\partial S}{\partial b} = -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = 0$$

위의 편미분한 두 식을 -2로 나누고 a, b 에 대하여 정리하면

$$\sum_{i=1}^n y_i - Na - \sum_{i=1}^n x_i b = 0 \quad (1)$$

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i a - \sum_{i=1}^n x_i^2 b = 0 \quad (2)$$

$$(1) \Rightarrow Na + \sum_{i=1}^n x_i b = \sum_{i=1}^n y_i$$

$$(2) \Rightarrow \sum_{i=1}^n x_i a + \sum_{i=1}^n x_i^2 b = \sum_{i=1}^n x_i y_i$$

위의 (1)과 (2)의 연립방정식을 행렬로 나타내면

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

a, b 를 구하기 위하여 Cramer's rule에 적용하면

* Determinant $D = N \sum x_i^2 - (\sum x_i)^2$ 를 구한 후 D_a, D_b 로부터

$$a = \frac{D_a}{D}, b = \frac{D_b}{D} \text{를 얻게 됨}$$

Fig. 1. Schematic design for cover-all procedure of germplasm information by NIRS analysis.

$$Da = \begin{bmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{bmatrix} = \sum y_i \sum x_i^2 - \sum x_i y_i \sum x_i$$

$$Db = \begin{bmatrix} N & \sum y_i \\ \sum x_i & \sum x_i y_i \end{bmatrix} = N \sum x_i y_i - \sum x_i \sum y_i$$

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N \sum x_i^2 - (\sum x_i)^2}, \quad b = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

∴ $\hat{Y} = a + bX$ 식을 최종적으로 결정함

* Cramer's rule

$$a_1 X + b_1 Y = c_1$$

$$a_2 X + b_2 Y = c_2$$

$$D: \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} = a_1 b_2 - a_2 b_1$$

$$D_x: \begin{bmatrix} c_1 & b_1 \\ c_2 & b_2 \end{bmatrix} = c_1 b_2 - c_2 b_1, \quad D_y: \begin{bmatrix} a_1 & c_1 \\ a_2 & c_2 \end{bmatrix} = a_1 c_2 - a_2 c_1$$

$$X = \frac{D_x}{D}, \quad Y = \frac{D_y}{D}$$

$$X = \frac{c_1 b_2 - c_2 b_1}{a_1 b_2 - a_2 b_1}, \quad Y = \frac{a_1 c_2 - a_2 c_1}{a_1 b_2 - a_2 b_1}$$

상관계수는 회귀계수에 XY 두 변량 사이의 표준편차의 비율을 곱한 값과 같다. 상관계수 또한 결정계수와 관계에서 상관계수 r 이 0에 가까우면 XY 관계가 의미가 없어지므로 r 값이 $0 < r < 1$ 일때 의미를 가지게 되고, 이 r 을 제곱함으로써 결정계수(R^2)로 사용하게 되며, 이는 회귀분석에서 설명변수 비율 값과 동일한 의미를 가지게 된다.

상관계수 제곱값(r^2) = 결정계수(R^2) = 회귀분석에서 설명변수 비율 값(SS_R/SS_T)

$$\begin{aligned} COV_{xy} &= \{ \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i Y_i - X_i \bar{Y} - \bar{Y} X_i + \bar{X} \bar{Y}) \} \\ &= (\sum X_i Y_i - \bar{Y} \sum X_i - \bar{X} \sum Y_i + \sum \bar{X} \bar{Y}) \\ &= \left(\sum X_i Y_i - \frac{N \bar{Y} \sum X_i}{N} - \frac{N \bar{X} \sum Y_i}{N} + \sum \frac{\sum X_i \sum Y_i}{N^2} \right) \\ &= \left(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N} - \frac{\sum X_i \sum Y_i}{N} + \sum \frac{\sum X_i \sum Y_i}{N^2} \right) \\ &= \left(\sum X_i Y_i - 2 \frac{\sum X_i \sum Y_i}{N} + \sum \frac{\sum X_i \sum Y_i}{N^2} \right) \\ &= \left(\sum X_i Y_i - 2 \frac{\sum X_i \sum Y_i}{N} + N \frac{\sum X_i \sum Y_i}{N^2} \right) \\ &= \left(\sum X_i Y_i - 2 \frac{\sum X_i \sum Y_i}{N} + \frac{\sum X_i \sum Y_i}{N} \right) \end{aligned}$$

$$= \left(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N} \right)$$

$$\begin{aligned} (\sqrt{Sx^2 * Sy^2})^2 &= Sx^2 * Sy^2 = \sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2 \\ \Rightarrow \sum (X_i - \bar{X})^2 &= \sum (X_i^2 - 2\bar{X} X_i + \bar{X}^2) = \sum X_i^2 - 2\bar{X} \sum X_i + \sum \bar{X}^2 \\ &= \sum X_i^2 - 2N\bar{X}^2 + N\bar{X}^2 = \sum X_i^2 - N\bar{X}^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{N} \\ \sum (Y_i - \bar{Y})^2 &= \sum (Y_i^2 - 2\bar{Y} Y_i + \bar{Y}^2) = \sum Y_i^2 - 2\bar{Y} \sum Y_i + \sum \bar{Y}^2 \\ &= \sum Y_i^2 - 2N\bar{Y}^2 + N\bar{Y}^2 = \sum Y_i^2 - N\bar{Y}^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{N} \end{aligned}$$

$$\begin{aligned} \frac{SS_R}{SS_T} = R^2 &= \left\{ \frac{COV_{xy}}{\sqrt{Sx^2 * Sy^2}} \right\}^2 = \frac{\left\{ \sum (x_i - \bar{X})(y_i - \bar{Y}) \right\}^2}{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2} \\ &= \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{N} \right)^2}{\left\{ \sum x_i^2 - \frac{(\sum x_i)^2}{N} \right\} \left\{ \sum y_i^2 - \frac{(\sum y_i)^2}{N} \right\}} \end{aligned}$$

회귀식에서 SS_T 와 SS_R , SS_E 관계는 다음과 같다.

$$SS_T = SS_R + SS_E$$

$$\sum (y_i - \bar{Y})^2 = \sum (\hat{y}_i - \bar{Y})^2 + \sum (y_i - \hat{y}_i)^2$$

R^2 는 NIRS 예측모델의 결정계수이고 작성된 검량식이 얼마만큼 정확하게 미지시료의 성분 함량을 예측할 수 있는지를 알려준다. 결정계수 값은 0과 1사이의 값을 가지며 1에 가까울수록 예측모델의 신뢰성은 높아진다. x_i 는 NIRS에 의해 측정된 성분 함량, $\bar{X}$ 는 x_i 의 평균값이고, y_i 는 습식분석법에 의해 측정된 성분 함량, $\bar{Y}$ 는 y_i 의 평균값, N 은 전체 관측값 개수이다. 평균값과 관측치와의 차이를 SS_T , 평균값과 예측치와의 차이를 SS_R 이라고 하면 Fig. 2와 같다.

작성된 검량식 중 습식분석값을 가장 잘 예측할 수 있는 최적 NIRS 예측모델을 RSQ (R^2), Standard error of calibration (SEC), Standard error of prediction (SEP), Slope, Bias, Standard error of cross-validation (SECV) 및 One minus the ratio of unexplained variance to total variance (1-VR) 등의 통계치를 고려하여 선별하였다. 최적 NIRS 예측모델의 미지시료 적용가능성을 검증하기 위해 검량식 작성에 이용되지 않은 자원들을 대상으로 근적외선 스펙트럼 측정과 아밀로스 및 단백질 함량 분석을 수행하여 validation file을 작성하였다. 작성된 validation file은 WINISI III project manager (FOSS, ver. 1.50e)의 monitor result 기능을 사용하여 Slope, RSQ, SEP, Bias 등의 통계치를

Fig. 2. Analysis of regression principle on two dimensions for NIRS prediction formula

계산하였고 이를 바탕으로 NIRS 예측모델의 미지시료 적용가능성과 정확성에 대하여 검증 실험을 진행하였다.

NIRS를 이용한 아밀로스 및 단백질 함량 측정

현미가루시료 약 600 mg을 micro insert ring을 끼운 mini sample cup에 채운 후 sample cup backs를 이용하여 시료 내 공극을 없애준 후 실온조건 가시광선 및 근적외선 대역(400-2500 nm)에서 한 자원 당 2반복하여 스펙트럼을 측정하였다. 스펙트럼 측정은 NIRS (FOSS, XRD near-infrared)를 사용하였고, 아밀로스 및 단백질 함량 계산은 개발된 NIRS 예측모델이 입력된 NIRS 구동 전용프로그램인 ISI scan (FOSS, ver. 4.2.0)을 사용하였다.

자원들의 다양성 지수 계산

다양성 지수는 자원들 간의 차이를 나타내기 위한 수단이지만, 수치화된 지표가 없으므로 분자생물학분야에서 사용되는 계산식을 적용하여 이를 다양성지수로 표시하고자 하였다. 아밀로스 및 단백질 함량구간을 2% 함량단위로 나누고 각 함량구간에 속하는 자원수(n_i)를 전체 자원수(9,771)로 나누어 다양성 확률값($P_i = \frac{n_i}{9,771}$)을 계산했다. 다양성 확률값 각각을 제곱하고 모두 합산 후 1에서 뺀 값으로 다양성지수 VIV (Variability Index Value)를 구하였다. 집단 내 개체들의 특성을 표현하는 유전자는 2n 형태의 대립유전자로 존재한다. 유전적 다양성 계산에서는 대립유전자를 homogeneous type으로 가정하기 때문에 P_i^2 값은

homogeneous type의 유전자가 발현될 확률을 의미하며, 이를 1에서 뺀 값은 heterogeneous type의 유전자가 발현될 확률이다. 따라서 다양성 지수가 1에 가까울수록 heterogeneous type의 유전자가 발현될 확률이 크다는 의미이므로 집단의 다양성은 커진다.

$$\text{Variability Index Value} = 1 - \sum_{i=1}^k P_i^2$$

정규분포의 작성

히스토그램과 정규분포 작성 과정에 관한 이론적 배경을 Fig. 3과 같이 나타냈다. 연속 계급값을 x축에 이들의 확률값을 y축에 나타내어 확률분포 막대그래프로 나타냈다(Fig. 3A). 막대그래프의 y축에 표시된 확률값을 막대그래프의 면적으로 전환하기 위해서 계급폭으로 나누어 {P(X)/계급폭} 표시하고 확률값을 면적으로 환산된 확률밀도 히스토그램을 작성했다(Fig. 3B). 히스토그램의 확률밀도분포가 정규분포를 따르면 임의의 확률변수 구간에 해당하는 확률값은 정규분포곡선이 이루는 면적값과 동일하다. 정규분포 함수 $f(x)$ 는 성분 함량의 평균값과 분산값에 의해 결정되며 엑셀프로그램의 정규분포함수(NORM.DIST)를 사용하면 x축 계급값은 연속된 실수값으로 환산 처리되고 이에 대응되는 $f(x)$ 값을 계산하여 정규분포곡선을 작성할 수 있다. NIRS 측정결과 얻어진 함량 자료를 바탕으로 작성된 계단모양의 확률밀도 히스토그램과 곡선 형태의 확률밀도함수로 표현된 정규분포를 비교하여 함량별 자원분포 특성을 비교하였다(Fig. 3C). 히스토그램과 정규분포에 관한 이론을 바탕으로

Fig. 3. Relationship between probability density histogram and normal distribution.

육성계통 메벼 9,771 (N) 자원에 대하여 아밀로스 및 단백질 함량 분포를 나타내는 도수분포표와 확률밀도를 나타내는 히스토그램을 구하였다. 계급폭(n, class width)은 아밀로스는 1, 단백질은 0.25, 계급은 아밀로스는 1 함량구간, 단백질은 0.25 함량구간으로 하였다. 아밀로스 및 단백질 함량의 확률밀도함수를 나타내기 위해 히스토그램의 y축 값은 확률밀도(probability density)를 나타내는 $\frac{p_i}{n}$ (확률값/계급폭)값으로, x축 값은 계급값들을 연속 확률변수형태(contents range)로 표시하였다.

분산분석(ANOVA)

육성계통 메벼 9,771자원 중 높은 자원비율을 나타낸 한국 원산지 2,386자원, 중국 원산지 2,136자원, 일본 원산지 1,219자원, 필리핀 원산지 1,213자원을 대상으로 분산분석을 이용하여 평균함량들 간의 유의성을 평가하였다. 독립변수는 국가별 집단(i)으로, 종속변수는 아밀로스와 단백질 함량(X)으로 설정하였다. 세 집단(그룹) 이상의 집단 간 평균을 비교할 경우 분산분석(analysis of variance, ANOVA, F-검정)을 실시하게 되는데 F-분포를 이용하여 집단 간의 분산과 집단 내의 분산을 비교하여 평균들 간 차이유무를 검정하는 방법이다. ANOVA 검정 절차를 아래 Fig. 4에 나타냈다. 편차를 기반으로 하는 분산은 평균을 중심으로 관찰값들이 얼마나 넓게 분포하는지를 나타내는 통계량으로 계산 과정에서 모든 관찰값이 사용되기 때문에 평균과 함께 가장 많이

Fig. 4. Schematic diagram for procedure of method in analysis of variance.

쓰인다. x_i 는 관찰값, N 은 관찰값 수, $\bar{x}$ 는 관찰값 평균이다. 편차 ($x_i - \bar{x}$)는 표본평균($\bar{x}$)과 관찰값(x_i)과의 차이를 말하며 분산은 편차제곱의 평균 $\left\{ \frac{1}{N} \sum (x_i - \bar{x})^2 \right\}$ 으로 정의되며, 관찰값 제곱의 평균에

서 평균의 제곱값을 뺀 값이다. 즉, $\left\{ \frac{1}{N} \sum (x_i - \bar{X})^2 = \frac{1}{N} \sum x_i^2 - \left(\frac{1}{N} \sum x_i \right)^2 \right\}$ 에서 편차제곱의 평균은 관찰값제곱의 평균에서 평균의 제곱값을 뺀 것으로 표현된다. 편차제곱합은 관찰값들 간에 차이가 클수록 큰 값이 나온다. 따라서 편차제곱합을 이용하는 분산은 관찰값의 분포정도를 나타내는 통계량으로서 그 표현력이 크다고 말할 수 있다.

분산분석은 귀무가설 설정, 제곱합과 평균제곱합 계산, 분산분석표 작성, 유의수준과 기각값 설정, 귀무가설 채택 또는 기각 여부 결정, 분산분석 결과 해석 과정을 거친다. 귀무가설(H_0)은 ‘모든 평균은 다 같다’이고, 대립가설(H_1)은 ‘평균들이 모두 같지는 않다’이다. 가설검정과정에서 H_0 이 맞지만 기각하게 되는 제 I 종 오류가 발생할 확률의 최대 허용한계를 유의수준(α)이라고 한다. $\alpha=0.01$ 은 제 I 종 오류의 최대 크기가 0.01 즉, 잘못된 판단을 할 확률이 0.01라는 것이며 ‘1%유의수준’이라고 말한다. ‘각각의 관찰값(x_i)은 전체평균($\bar{X} = \bar{X}_t$), 군간 편차($\bar{X} - \bar{X}_t$), 군내 편차($x_i - \bar{X}_t$)의 합과 같다’라는 정의는 분산분석의 이론적 배경이다. X 는 확률변수(아밀로스 또는 단백질 함량)를, $\sum x_i$ 는 한 집단 내 관찰값들의 총합을, $\sum \sum x_i$ 는 전체 집단 내 관찰값들의 총합을 의미하며 $\sum x (= \sum x_t)$ 과 동일한 표현이다. $\bar{X}_t$ 는 전체 관찰값의 평균을 나타내며 $\bar{x}_t$ 는 한 집단 내 관찰값들의 평균을 나타낸다. n 은 전체 집단(그룹)의 수, n_t 는 한 집단 내 관찰값의 총 개수(집단의 크기)를 N 은 전체 집단 내 관찰값의 총 개수를 의미하며 n 과 n_t 의 곱의 값과 동일하다. 각 제곱합들의 정의식(좌변)과 실제 계산에 사용되는 계산식(우변)을 식(1), 식(2), 식(3)에 나타냈다. 실제 분산분석을 위한 각 제곱합들의 계산은 $\frac{1}{N} (\sum x_t)^2$ 항이 SST 와 SSB 에 동일하게 나타나서 계산 과정이 단순해지므로 SST 와 SSB 를 먼저 구한 후 $SST=SSB+SSW$ 관계를 활용하여 SSW 를 구한다.

$$\begin{aligned} SST &= \sum (x_t - \bar{X}_t)^2 = \sum x_t^2 - \frac{(\sum x_t)^2}{N} \\ &= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) - \frac{(\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4)^2}{(n_1 + n_2 + n_3 + n_4)} \end{aligned} \quad \text{식(1)}$$

Table 2. ANOVA formula definition.

SV	df	SS	MSS	F_s
SSB	$K - 1$	$SSB = \sum \left\{ \frac{(\sum x_t)^2}{n_t} \right\} - \frac{(\sum x_t)^2}{N}$	$MSSB = \frac{SSB}{K - 1}$	$F_s = \frac{MSSB}{MSSW}$
SSW	$N - K$ or $K(n - 1)$	$SSW = \sum x_t^2 - \sum \left\{ \frac{(\sum x_t)^2}{n_t} \right\}$	$MSSW = \frac{SSW}{N - K}$	
SST	$N - 1$	$SST = \sum x_t^2 - \frac{(\sum x_t)^2}{N}$		

$$\begin{aligned} SSB &= \sum \sum (\bar{X}_t - \bar{X}_t)^2 = \sum \left\{ \frac{(\sum x_t)^2}{n_t} \right\} - \frac{(\sum x_t)^2}{N} \\ &= \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} \\ &\quad - \frac{(\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4)^2}{(n_1 + n_2 + n_3 + n_4)} \end{aligned} \quad \text{식(2)}$$

$$\begin{aligned} SSW &= \sum \sum (\bar{X}_t - \bar{X}_t)^2 = \sum x_t^2 - \sum \left\{ \frac{(\sum x_t)^2}{n_t} \right\} \\ &= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) \\ &\quad - \left\{ \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} \right\} \end{aligned} \quad \text{식(3)}$$

제곱합(SST , SSB , SSW)은 각각의 자유도를 가진다. 전체 집단의 수를 K (Σ Group), 전체 집단의 크기(전체 집단의 확률변수 총 개수)를 N 이라고 하면 SST 자유도는 $N-1$, SSB 는 $K-1$, SSW 는 $N-K$ 이다. SST 자유도는 SSB 자유도와 SSW 자유도의 합이다. 편차제곱합을 자유도로 나눈 것을 평균편차제곱합(MSS)이라 하며 $MSSB$ 와 $MSSW$ 로 분류한다. 군간평균제곱합은 $\frac{SSB}{(K-1)}$, 군내평균제곱합은 $\frac{SSW}{(N-K)}$ 이다. 검정통계량(F_s)은 군간평균제곱합을 군내평균제곱합으로 나눈 값이며 아래 식과 같다.

$$F_s = \frac{MSSB}{MSSW} = \frac{\frac{SSB}{K-1}}{\frac{SSW}{N-K}}$$

위 과정에서 계산된 통계량들을 정리한 분산분석표는 아래 Table 2와 같다.

Duncan's multiple range test (DMRT)

분산분석과정에서 귀무가설을 기각되어 각 집단 간 ‘평균들이 모두 같지는 않다’는 가설이 성립되면 이는 두 개 이상의 평균이 같지 않다는 뜻이며 어느 집단 간의 평균들이 차이가 있는지는 알 수 없다. 이 같은 경우 평균의 다중비교 방법을 사용하여 각 집단들 간 평균 비교를 할 수 있다. 이론적 배경은 여러 개의 평균들을 상호 짝을 지어 반복 비교하더라도 제I종 오류(평균간 차이가 없음에도 평균간 유의차가 있는 것으로 판단)가 일정

유의수준(0.01 또는 0.05)을 넘지 않는 구간(유의범위)을 정하여 평균 간 차이와 비교하는 방법이다. 평균 간 차이가 유의하다고 판단할 수 있는 ‘최소유의차(LSD)’ 또는 ‘최소유의범위(LSR)’를 구하여 두 평균 간 차이를 비교하여 두 평균 간 차이가 최소유의차 또는 최소유의범위보다 크면 두 평균은 유의하게 다른 것이고 작으면 유의차가 없는 것이다. Least Significant Difference (LSD)는 t 검정이며 이는 두 집단 간 비교에 이용된다. 따라서 다중검정에 적용할 경우 제1종 오류를 범할 위험성이 있다. 이를 보완한 방법이 DMRT이며 비교하는 평균들 간의 거리에 따라서 다른 유의범위를 적용한다. 농학 연구 분야에서는 주로 LSD와 DMRT가 많이 적용되며 본 연구에서는 DMRT를 사용하였다. DMRT는 평균을 크기순으로 정렬하고 가장 큰 평균부터 시작하여 가장 작은 평균까지 순차적으로 비교하게 되는데 이때 기준이 되는 값이 필요하다. 이 값을 Least Significant Range (LSR)이라고 하며 전체 평균수(전체 집단수) 보다 1개 적은 LSR 값이 필요하다. LSR은 아래 식(4)와 같이 정의된다. S_x 는 표준오차이며 군내평균제곱합(MSSW)을 집단의 크기(n)로 나눈 수치의 양의 제곱근 값이다(식(5)). 각 집단의 크기(n)가 서로 다를 경우는 식(5)의 n 대신 n_h 값을 적용한다. Significant Studentized Range (SSR)값은 오차자유도, 전체 평균수, 유의수준(α)을 정한 후 ‘다중검정을 위한 스튜던트화한 범위’표를 참조하여 구할 수 있다. 예를 들면 오차자유도 60, 유의수준(α) 0.05, 전체 집단수 4인 경우 SSR값은 2.83, 2.98, 3.08이다.

$$LSR = SSR \times S_x \quad \text{식(4)}$$

$$S_x = \sqrt{\frac{MSSW}{n(\text{or } n_h)}}, \quad n_h = \frac{K}{\sum \frac{1}{n_i}} \quad \text{식(5)}$$

※ 전체 집단수: 3,

$$\text{집단의 크기: } 438, 308, 131 \text{ 이면 } n_h = \frac{3}{\frac{1}{438} + \frac{1}{308} + \frac{1}{131}}$$

식(4)에서 구한 LSR값을 크기순으로 나열하고 각 집단의 평균값

을 크기순으로 나열하는 것으로 유의성검정과정을 시작한다. 가장 큰 평균에서 가장 큰 LSR값을 빼 값은 가장 큰 평균과 나머지 평균들을 비교하는 기준값이고 이 보다 작은 평균값들이 존재하면 가장 큰 평균값과 5% 유의수준에서 유의차가 있다고 판단하고 가장 큰 평균값 아래 밑선을 긋거나 위 첨자 영문 소문자로 표시한다. 두 번째 큰 평균에서 두 번째 큰 LSR값을 빼 값은 두 번째 큰 평균과 나머지 평균들을 비교하는 기준값이고 이 보다 작은 평균값들이 존재하면 두 번째 큰 평균값과 5% 유의수준에서 유의차가 있다고 판단하고 두 번째 큰 평균값 아래 밑선을 긋거나 위 첨자 영문 소문자로 표시한다. 이와 같은 과정을 모든 LSR값을 사용하여 검정과정을 끝낼 때까지 반복한다. 만일 비교 기준값 보다 큰 평균값들이 존재하여 평균 간 차이가 유의하지 않는 경우에는 비교하는 평균값들 아래 밑선 또는 동일한 문자표시를 하여 유의하지 않는 평균값을 가지는 동일 집단임을 표시한다.

Fig. 5. Schematic diagram for procedure of Duncan's multiple range test.

Table 3. Descending arrangement of average values ($\bar{X}$) in each group and the determination of Studentized range p value.

Alphabeticals	$\bar{X}$		Order in size	$\bar{X}$		B	C	E	A	D
A	13.29		B	30.86		D	5	4	3	2
B	30.86		C	29.14	p value	A	4	3	2	
C	29.14	$\bar{X}$ Array $\Rightarrow$	E	21.00	$\Rightarrow$	E	3	2		
D	11.00		A	13.29		C	2			
E	21.00		D	11.00		B

Table 4. Creating DMRT table by comparing with $\bar{X}$ and $|\bar{X} - LSR_{0.01}|$ value in each group.

p	Group	$\bar{X}$	$LSR_{0.01}$	Reference ($\bar{X} - LSR_{0.01}$)	Comparison Reference and $\bar{X}$	DMRT Result
5	B	30.86	4.64	26.22	$29.14^C > 26.22^r > 21.00^E, 13.29^A, 11.00^D$	<u>B B C</u>
4	C	29.14	4.58	24.56	$24.56^r > 21.00^E, 13.29^A, 11.00^D$	<u>C</u>
3	E	21.00	4.47	16.53	$16.53^r > 13.29^A, 11.00^D$	<u>E</u>
2	A	13.29	4.28	9.01	$9.01^r < 11.00^D$	<u>A D</u>
	D	11.00	-	-	-	

Table 5. Interpretation of DMRT result.

DMRT Result	Interpretation
<u>B C E A D</u>	Equal to B and C at level of 1%
	Equal to A and D at level of 1%
A ^a B ^b C ^b D ^a E ^c	Grouping as A&D, B&C, and E at level of 1%

DMRT에 의한 검정절차를 Fig. 5에 요약했다. 그룹별 관찰값 정렬과 평균값계산, 분산분석표 작성을 통해 집단 간 유의성이 확인되면(1단계) DMRT를 통해 각 집단 간 평균을 비교할 수 있다. 집단들의 평균값을 내림차순으로 정렬한 후 집단들이 이루는 평균값 구간에 존재하는 평균값 개수(집단개수, p 값)를 확인하여 DMRT에 필요한 p 값을 정한다(2단계, Table 3). 평균 간 비교를 위한 기준이 되는 LSR 값은 상기 식(4), 식(5)에 의해 구할 수 있다. SSR 값은 p 값, 검정유의수준(α), 오차자유도(SSW 의 df)를 결정하면 ‘다중검정을 위한 스튜던트화한 범위’표를 참조하여 구할 수 있다. $S_{\bar{x}}$ 는 표준오차이며 분산분석표의 $MSSW$ 값(8.44)과 집단의 크기($n=7$)를 알면 식(5)에 의해 구할 수 있다. 각 집단들의 평균값($\bar{X}$)과 LSR 값들을 크기순으로 나열하여 DMRT 표를 작성한다(3단계, Table 4). 평균($\bar{X}$)과 LSR 값의 차($|\bar{X} - LSR_{0.01}|$)를 기준값으로 사용하여 다른 집단의 평균과 비교하거나 집단 간 평균 차($|\bar{X}_i - \bar{X}_j|$)와 $LSR_{0.01}$ 값을 비교하는 두 가지 방식을 사용할 수 있다. 평균과 LSR 값의 차($|\bar{X} - LSR_{0.01}|$)를 기준값으로 하는 경우 기준값이 다른 집단의 평균보다 크면 유의성이 있다고 판단하고 작으면 유의성이 없다고 판단한다. 집단 간 평균 차($|\bar{X}_i - \bar{X}_j|$)와 $LSR_{0.01}$ 값을 비교하는 경우 집단 간 평균간 차보다 LSR 값이 크면 유의성이 없다고 판단하고 작으면 유의성이 있다고 판단한다. DMRT 결과를 실험 목적 맞게 해석한다(4단계, Table 5).

결과 및 고찰

NIRS 예측모델설정 및 검증

검량식 관련 주요 통계지표에는 SEC, SECV, Slope, RSQ

(R^2), 1-VR가 있다. SEC는 검량식 표준오차이고 SECV는 교차 검증 표준오차이다. 표준오차값이 0에 가까울수록 예측모델의 정확도는 높아진다. Slope은 작성된 검량식의 기울기이고 NIRS 측정값과 습식분석법에 의한 분석값과의 차이의 비율을 반영하며 1에 가까울수록 두 값들의 차는 작아진다. R^2 는 NIRS 예측모델의 결정계수이고 1-VR은 교차검증 결정계수이다. 결정계수는 작성된 검량식이 얼마만큼 정확하게 미지시료의 성분 함량을 예측할 수 있는지를 알려준다. 결정계수 값은 0과 1사이의 값을 가지며 1에 가까울수록 예측모델의 신뢰성은 높아진다. 100자원, 300자원, 500자원과 같이 자원수를 달리하여 각각의 검량식을 수동 분석 방법으로 작성하였다. 각각의 수동 작성된 검량식의 R^2 값과 slope 값을 비교하여 검량식의 예측성능을 비교하였다. 아밀로스 검량식의 R^2 값은 100자원 검량식은 0.814, 300자원은 0.966, 500자원은 0.972였다. 단백질 검량식의 R^2 값은 100자원 검량식은 0.969, 300자원은 0.972, 500자원은 0.978이었다. 아밀로스 검량식의 slope 값은 100자원 검량식은 0.814, 300자원은 0.955, 500자원은 0.981이었다. 단백질 검량식의 slope 값은 100자원 검량식 0.980, 300자원은 0.981, 500자원은 0.991이었다. 아밀로스 100자원 검량식, 300자원 검량식, 500자원 검량식에서 R^2 값과 slope 값은 검량식 작성에 사용된 자원수가 증가함에 따라 1에 가까워졌다. 단백질 100자원 검량식, 300자원 검량식, 500자원 검량식에서 R^2 값과 slope 값은 검량식 작성에 사용된 자원수가 증가할수록 1에 가까워졌다. 검량식 자원수와 R^2 , slope 값의 변화 정도를 고려하여 예측성능이 우수하다고 판단된 500자원을 사용하여 작성된 검량식을 기초로 하여 이후 보완과정을 진행하였다. NIRS 분석방법은 습식분석에 비해 정확성은 낮기 때문에 검량식 자원들을 적정 농도 구간에서 균등한 자원

밀도를 나타내도록 구성하는 것이 중요하다(Kim et al. 2008). 자원밀도가 낮은 구간은 검량식 자원을 추가분석하여 보완하는 방법으로 NIRS 예측모델을 최적화하였다. 검량식 작성에 사용된

벼 자원들의 성분별 함량 분포 구간은 아밀로스는 6.15-32.25%, 단백질은 4.72-14.48%였다(Fig. 6). 검량식 작성과 보완에 사용된 511자원 중 최적화를 위해 이상치 자원은 제외시켰다. WINISI

Fig. 6. Amylose (A) and protein (B) content in the milled brown rice for NIRS prediction model (n=511).

Fig. 7. Comparison of two methods of plotting NIRS equation graph between WINISI III program (A) and manual Excel program (B) based on the milled brown rice germplasm.

III project manager 프로그램을 사용하여 작성된 NIRS 검량식 그래프와 수동 방법으로 작성된 그래프를 상호 비교한 결과 관련 통계수치들이 동일한 값을 나타내어 작성된 검량식이 동일함을 확인하였다(Fig. 7). Table 6에 검량식 수동 계산 과정을 나타냈다. 1-VR은 검량식 작성에 이미 이용된 자원을 재차 이용하여 정확도를 평가하는 역검정 방법이다. 간편하게 평가가 이뤄지는 장점이 있으나 1-VR 만으로는 개발된 NIRS 예측모델의 미지시료 분석 시 정확도를 평가하기에 부족하다(Bagchi et al. 2016). 따라서 검량식 작성에 사용된 자원 외의 별도자원들을 이용하여 NIR 기기내의 external validation 프로그램을 이용한

외부검증과정을 거쳤다. 검량식 작성에 사용된 511자원 이외의 재래종 2,000자원을 최적 예측모델이 적용된 NIRS를 이용하여 아밀로스 및 단백질 함량을 분석하였다. NIRS로 분석된 자원들을 저, 중, 고 함량 세 구간으로 나누고 각 함량의 구간별 대표자원을 선택하는 방법으로 아밀로스는 132자원, 조단백질은 124자원을 선정하여 NIRS 예측모델 검증자원으로 사용하였다. 선정된 검증자원들의 습식분석값을 NIRS 구동 프로그램의 lab data 항목에 입력하여 external validation set을 구성하였고, 최적 NIRS 예측모델에 external validation set이 적용된 검정 결과를 확인하였다(Table 7). 아밀로스와 단백질의 RSQ (r^2) 값은 각각

Table 6. Statistic data table for manual calculating NIRS equation.

Number of Observed values (N)	NIRS equation for protein content					NIRS equation for amylose content				
	Number of accessions (n=473)					Number of accessions (n=464)				
	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	6.703	6.650	6.703 ²	6.650 ²	6.703*6.650	5.228	6.380	5.228 ²	6.380 ²	5.228*6.3
2	8.844	8.840	8.844 ²	8.840 ²	8.844*8.840	6.457	6.150	6.457 ²	6.150 ²	6.457*6.1
3	8.268	8.220	8.268 ²	8.220 ²	8.268*8.220	7.714	6.480	7.714 ²	6.480 ²	7.714*6.4
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
1,392	•	•	•	•	•	11.234	11.130	11.234 ²	11.130 ²	11.234*11.1
1,419	8.845	8.935	8.845 ²	8.935 ²	8.845*8.935					
Σx_i	10,905.6					25,989.3				
Σy_i		10,910.9					25,989.3			
Σx_i^2			85,860.8					530,699.9		
Σy_i^2				85,972.6					532,102.7	
$\Sigma x_i y_i$					85,897.0					530,700.1
$(\Sigma x_i)^2$	10,905.615 ²					25,989.378 ²				
N	1,419					1,392				
$\frac{(\Sigma x_i)^2}{N}$	$\frac{10,905.615^2}{1,419}$					$\frac{25,989.378^2}{1,392}$				
$R^2 = \frac{\left(\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{N} \right)^2}{\left\{ \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N} \right\} \left\{ \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{N} \right\}}$ $= \frac{\left(85,897.043 - \frac{10,905.615 \times 10,910.970}{1,419} \right)^2}{\left\{ 85,860.890 - \frac{(10,905.615)^2}{1,419} \right\} \left\{ 85,972.685 - \frac{(10,910.970)^2}{1,419} \right\}}$ $= 0.981$						$= \frac{\left(\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{N} \right)^2}{\left\{ \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N} \right\} \left\{ \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{N} \right\}}$ $= \frac{\left(530,700.122 - \frac{25,989.378 \times 25,989.386}{1,392} \right)^2}{\left\{ 530,699.924 - \frac{(25,989.378)^2}{1,392} \right\} \left\{ 532,102.792 - \frac{(25,989.386)^2}{1,392} \right\}}$ $= 0.970$				
$\text{Slope} = \frac{N \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N \Sigma x_i^2 - (\Sigma x_i)^2}$ $= \frac{1,419 \times 85,897.043 - 10,905.615 \times 10,910.970}{1,419 \times 85,860.890 - (10,905.615)^2}$ $= 0.998$						$= \frac{N \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N \Sigma x_i^2 - (\Sigma x_i)^2}$ $= \frac{1,392 \times 530,700.122 - 25,989.378 \times 25,989.386}{1,392 \times 530,699.924 - (25,989.378)^2}$ $= 1.000$				

0.962, 0.986이었고 SEP 값은 각각 2.349, 0.415였다. 단백질 SEP 값이 아밀로스 SEP 값보다 작으므로 단백질 측정 정확도가 높았으며 이는 아밀로스 SEP 0.882, 단백질 SEP 0.280 값으로 보고된 Hwang et al. (1994)의 결과와도 유사한 경향으로 나타났다. SEP 값이 아밀로스가 단백질에 비해 높게 나타나는 것은 현미에는 아밀로스와 관련이 적은 성분이 많이 존재하는 등 미강의 혼입 및 그 분산 상태가 백미에 비해 다르기 때문인 것으로 생각된다(Moon et al. 1994). 이상의 결과들을 종합해 볼 때 개발된 최적 NIRS 예측모델은 기존의 습식분석에 의한 성분 분석 결과와 높은 상관도를 가지며, 분석정확도 또한 큰 차이가 없는 것으로 나타났다(Oh et al. 2017a).

벼 유전자원의 집단분석

육성계통 벼 9,998자원에 대한 아밀로스 함량분포의 정규분포 함수에서는 확률밀도함수와 일치하지 않는 부분이 나타났으며, 정규분포곡선 속에서 뚜렷한 두 개의 집단으로 구분됨을 확인할 수 있었다. 이는 개략적으로 4.0-15.0% 구간과 16.0-35.0% 구간으로 대별되었다. 두 개의 집단으로 구분되게 하는 아밀로스 함량구간은 15.0-16.0% 구간이었으며, 이 구간에는 수집자원이 거의 없었다. 이로 인하여 전체 아밀로스 함량분포가 정규분포를 이루지 않았다. Oh et al. (2017b)은 육종분야에서 벼 분류 기준으로 사용되는 아밀로스 함량범위(Kim et al. 1991, Song et al.

2008, Jeong et al. 2013)를 바탕으로 NIRS 측정 아밀로스 함량에 대하여 11.0%이하를 찰벼, 11.0%이상을 메벼로 분류하였다. 이 분류기준과 국립식량과학원의 분류기준을 참고하여 본 연구에서는 아밀로스 함량 10.0% 이상 자원을 메벼로 분류하였고, 전체 정규분포에서 찰벼집단과 메벼집단으로 두 개의 정규분포 집단을 형성하였으나 자원분포가 많은 메벼 집단만을 대상으로 하여 함량 분포분석과 통계처리 하였다.

육성계통 벼 유전자원 9,998자원 중 메벼 9,771자원의 아밀로스 및 단백질 함량 별 자원분포와 정규분포를 비교한 결과는 Fig. 8과 같았다. 메벼 자원은 아밀로스 15-19% 함량구간에서 자원밀도가 정규분포함수에 비해 낮았으나, 전반적으로 자원분포는 평균값 23.6, 표준편차 4.0인 정규분포와 유사하였고, 자원의 다양성 지수는 0.83이었다(Fig. 8A). 이와 같이 일부 구간에서 자원밀도가 낮게 나타나는 것은 유전자원 수집이 부족하였거나 이에 해당하는 육성계통 벼 자원이 없다는 의미로 판단된다. 57개국에서 수집된 육성계통 메벼 9,771자원 중 한국 원산지 자원은 2,386자원, 중국 원산지 자원은 2,136자원, 일본 원산지 자원은 1,219자원, 필리핀 원산지 자원은 1,213자원으로 이들 자원들이 전체 육성계통 자원의 71.2%를 차지하였다(Table 1). 자원 비중이 큰 한국, 중국, 일본, 필리핀의 4개국 자원들을 대상으로 아밀로스 함량 분포특성을 상호 비교하였다(Fig. 9). 필리핀 원산지 자원은 아밀로스 22-23% 함량구간에서 정규분포

Fig. 8. Normal distribution and probability density of non-glutinous rice amylose (A) and protein (B) contents in total breeding line germplasm (n=9,771).

Table 7. External validation results of NIRS equation model for the amylose and protein content in the milled brown rice.

Constituent	External validation				NIRS equation model			
	No.	r^2	SEP	Slope	No.	R^2	SEC	Slope
Amylose	132	0.962	2.349	0.811	464	0.970	1.010	1.000
Protein	124	0.986	0.415	1.038	473	0.981	0.158	0.998

Fig. 9. Normal distribution and probability density of non-glutinous rice amylose and protein contents in breeding line germplasm of Philippines (a), China (b), Korea (c), Japan (d).

함수에 비해 자원밀도가 낮았고 26-27% 함량구간에서 정규분포 함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 25.2, 표준편차 2.8인 정규분포와 유사하였고, 자원의 다양성 지수는 0.75였다(Fig. 9a: L). 중국 원산지 자원은 아밀로스 15-20% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 28-29% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 24.5, 표준편차 4.0인 정규분포와 유사하였고, 자원의 다양성 지수는 0.83이었다(Fig. 9b: L). 한국 원산지 자원은 아밀로스 15-20% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 21-25% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 22.1, 표준편차 3.8인 정규분포와 유사하였고, 자원의 다양성 지수는 0.78이었다(Fig. 9c: L). 일본 원산지 자원은 아밀로스 15-19% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 21-22% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 21.5, 표준편차 2.5인 정규분포와 유사하였고, 자원의 다양성 지수는 0.70이었다(Fig. 9d: L). 한국, 중국, 일본, 필리핀 원산지 자원들의 아밀로스 함량분포와 다양성 지수는 상호 유사성이 없었다.

육성계통 메벼 9,771자원의 단백질 함량분포를 나타냈다(Fig. 8B). 단백질 8-9% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 6.5-7.5% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 7.9, 표준편차 1.3인 정규분포와 유사하였고, 자원의 다양성 지수는 0.54였다. 자원 비중이 큰 한국, 중국, 일본, 필리핀의 4개국 원산지 자원들을 대상으로 단백질 함량 분포특성을 상호 비교하였다(Fig. 9). 필리핀 원산지 자원은 단백질 9.5% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 7.0-8.5% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 8.2, 표준편차 1.2인 정규분포와 유사하였고, 자원의 다양성 지수는 0.54였다(Fig. 9a: R). 중국 원산지 자원은 8-10% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 6.5-7.5% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 7.8, 표준편차 1.3인 정규분포와 유사하였고, 자원의 다양성 지수는 0.52였다(Fig. 9b: R). 한국 원산지 자원은 단백질 8.3-10.3% 함량구간에서 정규분포함수에 비해 자원밀도가 낮았고 6.3-7.3% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 7.6, 표준편차 1.1인 정규분포와 유사하였고, 자원의 다양성 지수는 0.48이었다(Fig. 9c: R). 일본 원산지 자원은 아밀로스 7.8-8.3% 함량구간에서 정규분포함

수에 비해 자원밀도가 낮았고 6.3-7.0% 함량구간에서 정규분포함수에 비해 자원밀도는 높았다. 전반적으로 자원분포는 평균값 7.4, 표준편차 0.9인 정규분포와 유사하였고, 자원의 다양성 지수는 0.41이었다(Fig. 9d: R). 한국, 중국, 일본, 필리핀 원산지 자원들의 단백질 함량분포와 다양성 지수는 상호 유사성이 없었다.

벼 유전자원의 국가별 품종 기준(ANOVA)

독립변수를 몇 개의 수준(또는 범주)으로 나누고 각 수준에 따라 나누어진 집단 간의 평균차를 검정하는 방법(차의 검정)으로써 서로 다른 세 개 이상의 집단에서 평균간 차이가 있는지를 확인하고 자 할 때 분산분석이 이용된다. 독립변수는 아밀로스 또는 단백질 함량으로, 종속변수는 품종 재배의 기준점으로 정하여 국가별 품종 재배의 기준점이 상이한가를 검정해보고자 수집보존자원을 대상으로 통계분석을 실시하였다. 검정과정에서 필요한 유의성은 계산된 검정통계량(F_s)을 기각값(critical value, F_c)과의 크기 비교과정을 통해 결정된다. 이러한 과정은 정의된 수확공식에 관찰값과 관찰값 개수를 엑셀프로그램에 입력하면 비교적 용이하게 얻을 수 있다. 하지만 검정통계량의 개념과 그 개념의 기반이 되는 총제곱합, 군내제곱합, 군간제곱합을 이해하기에는 다소 어려움이 있으므로 분산분석 절차와 통계량을 순서도(Fig. 4)와 표(Table 2)로 정리하여 상호간의 관계를 보다 명확하게 제시하고자 하였다. 육성계통 메벼 9,771 자원 중 높은 자원비율을 나타낸 한국 원산지 2,386자원, 중국 원산지 2,136자원, 일본 원산지 1,219자원, 필리핀 원산지 1,213자원을 대상으로 아밀로스 성분에 대한 분산분석을 실시한 결과 SSB 15,259, SSW 85,767을 얻었고, 또한 단백질 성분에 대한 분산분석 결과 SSB 429, SSW 9,162를 얻었다(Tables 8, 10). 귀무가설(H_0)은 ‘국가별 자원들의 함량 간 차이는 없다’로, 대립가설(H_1)은 ‘국가별 자원들의 함량 간 차이는 있다’로 설정하였다. 아밀로스함량의 경우 검정통계량 F_s 는 412.2였으며, 유의수준 0.01, SSB 자유도 3, SSW 자유도 6,950일 때 F -분포표의 기각값($F_{0.01(3,6950)}$)은 3.78이므로 검정통계량과 기각값의 비교에서 검정통계량의 값이 월등히 높으므로 귀무가설이 기각되고 대립가설이 채택되었다. 따라서 벼 유전자원의 아밀로스 함량은 국가별 품종 기준이 다르다는 것을 수치화 할 수 있으며, ‘1% 유의수준에서 차이가 인정된다’는 결론을 내릴 수 있었다(Table 9). 단백질 함량의 경우 검정통계량 F_s 는 108.4, 유의수준 0.01, SSB 자유도 3, SSW 자유도 6,950일 때 F -분포표의 기각값($F_{0.01(3,6950)}$)은 3.78이므로 검정통계량과 기각값의 비교에서 검정통계량의 값이 월등히 높으므로 이 또한 귀무가설이 기각되고 대립가설이 채택

되었다. 따라서 벼 유전자원의 단백질 함량은 국가별 품종 기준이 결론을 내릴 수 있었다(Table 11). 다르다고 할 수 있으며, ‘1% 유의수준에서 차이가 인정된다’는

Table 8. Statistic data table for ANOVA test on non-glutinous rice amylose content by countries.

	Korea		China		Japan		Philippines	
	$n_1=2,386$		$n_2=2,136$		$n_3=1,219$		$n_4=1,213$	
No.	A	A^2	B	B^2	C	C^2	D	D^2
1	10.01	$(10.01)^2$	10.01	$(10.01)^2$	10.02	$(10.02)^2$	10.03	$(10.03)^2$
2	10.04	$(10.04)^2$	10.01	$(10.01)^2$	10.12	$(10.12)^2$	10.13	$(10.13)^2$
3	10.04	$(10.04)^2$	10.06	$(10.06)^2$	10.19	$(10.19)^2$	10.29	$(10.29)^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1,213	⋮	⋮	⋮	⋮	⋮	⋮	30.55	$(30.55)^2$
1,219	⋮	⋮	⋮	⋮	29.62	$(29.62)^2$		
2,136			32.52	$(32.52)^2$				
2,386	31.32	$(31.32)^2$						
$\sum x_i$	52,671.52		52,424.63		26,167.50		30,543.05	
$\sum x_i^2$		1,197,053.19		1,320,967.92		569,107.85		778,839.05
$(\sum x_i)^2$	$(52,671.52)^2$		$(52,424.63)^2$		$(26,167.50)^2$		$(30,543.05)^2$	
n_i	2,386		2,136		1,219		1,213	
$\frac{\sum x_i}{n_i}$	$\frac{52,671.52}{2,386}$		$\frac{52,424.63}{2,136}$		$\frac{26,167.50}{1,219}$		$\frac{30,543.05}{1,213}$	
$\frac{(\sum x_i)^2}{n_i}$	$\frac{(52,671.52)^2}{2,386}$		$\frac{(52,424.63)^2}{2,136}$		$\frac{(26,167.50)^2}{1,219}$		$\frac{(30,543.05)^2}{1,213}$	
$= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) - \frac{(\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4)^2}{(n_1 + n_2 + n_3 + n_4)}$								
SST	$= (1,197,053.19+1,320,967.92+569,107.85+778,839.05)$ $= \{(52,671.52+52,424.63+26,167.50+30,543.05)^2/(2,386+2,136+1,219+1,213)\}$ $= 101,025.64$							
	$= \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} - \frac{(\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4)^2}{(n_1 + n_2 + n_3 + n_4)}$							
SSB	$= \{(52,671.52)^2/2,386+(52,424.63)^2/2,136+(26,167.50)^2/1,219+(30,543.05)^2/1,213\}$ $= \{(52,671.52+52,424.63+26,167.50+30,543.05)^2/(2,386+2,136+1,219+1,213)\}$ $= 15,258.86$							
	$= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) - \left\{ \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} \right\}$							
SSW	$= (1,197,053.19+1,320,967.92+569,107.85+778,839.05)$ $= \{(52,671.52)^2/2,386+(52,424.63)^2/2,136+(26,167.50)^2/1,219+(30,543.05)^2/1,213\}$ $= 85,766.78$							

Table 9. ANOVA table of non-glutinous rice amylose content in breeding line germplasm by countries.

SV	df	SS	MSS	F_s	Critical value ($F_{0.01(3,6950)}$)
SSB	3	15,259	5,086	412.2**	3.78
SSW	6,950	85,767	12		
SST	6,953				

벼 유전자원의 국가별 특성(DMRT)

분산분석 결과 집단 간 평균 간 유의미한 차이가 확인되면 어느 집단에서 유의한 차가 있는 지를 검정하여야 하며 이때

사용되는 통계방법 중 하나가 DMRT이다. 이 방법은 각 집단들의 평균($\bar{x}$)과 LSR 값을 비교하여 집단 간 평균차를 검정하는 방법이다. $\bar{x}$ 와 LSR 값의 차($|\bar{x} - LSR|$)를 기준값으로 사용하

Table 10. Statistic data table for ANOVA test on non-glutinous rice protein content by countries.

	Korea		China		Japan		Philippines	
	$n_1=2,386$		$n_2=2,136$		$n_3=1,219$		$n_4=1,213$	
No.	A	A^2	B	B^2	C	C^2	D	D^2
1	5.31	$(5.31)^2$	5.26	$(5.26)^2$	5.37	$(5.37)^2$	5.45	$(5.45)^2$
2	5.36	$(5.36)^2$	5.26	$(5.26)^2$	5.44	$(5.44)^2$	5.48	$(5.48)^2$
3	5.39	$(5.39)^2$	5.40	$(5.40)^2$	5.50	$(5.50)^2$	5.50	$(5.50)^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1,213	⋮	⋮	⋮	⋮	⋮	⋮	15.69	$(15.69)^2$
1,219	⋮	⋮	⋮	⋮	12.55	$(12.55)^2$		
2,136	⋮	⋮	14.46	$(14.46)^2$				
2,386	14.04	$(14.04)^2$						
$\sum x_i$	18,095.91		16,648.91		9,048.07		9,931.36	
$\sum x_i^2$		139,946.2		133,362.0		68,229.1		83,108.2
$(\sum x_i)^2$	$(18,095.91)^2$		$(16,648.91)^2$		$(9,048.07)^2$		$(9,931.36)^2$	
n_i	2,386		2,136		1,219		1,213	
$\frac{\sum x_i}{n_i}$	$\frac{18,095.91}{2,386}$		$\frac{16,648.91}{2,136}$		$\frac{9,048.07}{1,219}$		$\frac{9,931.36}{1,213}$	
$\frac{(\sum x_i)^2}{n_i}$	$\frac{(18,095.91)^2}{2,386}$		$\frac{(16,648.91)^2}{2,136}$		$\frac{(9,048.07)^2}{1,219}$		$\frac{(9,931.36)^2}{1,213}$	
$= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) - \frac{(\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4)^2}{(n_1 + n_2 + n_3 + n_4)}$								
SST	$= (139,946.23+133,362.00+68,229.10+83,108.24)$ $= \{(18,095.91+16,648.91+9,048.07+9,931.36)^2/(2,386+2,136+1,219+1,213)\}$ $= 9,590.08$							
$= \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} - \frac{(\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4)^2}{(n_1 + n_2 + n_3 + n_4)}$								
SSB	$= \{(18,095.91)^2/2,386+(16,648.91)^2/2,136+(9,048.07)^2/1,219+(9,931.36)^2/1,213\}$ $= \{(18,095.91+16,648.91+9,048.07+9,931.36)^2/(2,386+2,136+1,219+1,213)\}$ $= 428.51$							
$= (\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2) - \left\{ \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} \right\}$								
SSW	$= (139,946.23+133,362.00+68,229.10+83,108.24)$ $= \{(18,095.91)^2/2,386+(16,648.91)^2/2,136+(9,048.07)^2/1,219+(9,931.36)^2/1,213\}$ $= 9,161.57$							

Table 11. ANOVA table of non-glutinous rice protein content in breeding line germplasm by countries.

SV	df	SS	MSS	F_s	Critical value ($F_{0.01(3,6950)}$)
SSB	3	429	142.84	108.4**	3.78
SSW	6,950	9,162	1.32		
SST	6,953				

Table 12. DMRT table of non-glutinous rice amylose content in breeding line germplasm by countries.

Origin	Amylose (%)	SSR _{0.01}	LSR _{0.01}	Amylose-LSR _{0.01}	DMRT
Philippines	25.18	3.90	0.35	24.83 (25.18-0.35)	<u>25.18</u>
China	24.54	3.80	0.34	24.20 (24.54-0.34)	<u>24.54</u>
Korea	22.08	3.64	0.32	21.76 (22.08-0.32)	<u>22.08</u>
Japan	21.47	-	-	-	<u>21.47</u>

$$※ S_x = \sqrt{\frac{MSSW}{n_h}} = \sqrt{\frac{12.34}{1579.76}} = 0.09$$

Table 13. DMRT table of non-glutinous rice protein content in breeding line germplasm by countries.

Origin	Protein (%)	SSR _{0.01}	LSR _{0.01}	Protein-LSR _{0.01}	DMRT
Philippines	8.19	3.90	0.12	8.07 (8.19-0.12)	<u>8.19</u>
China	7.79	3.80	0.11	7.68 (7.79-0.11)	<u>7.79</u>
Korea	7.58	3.64	0.11	7.47 (7.58-0.11)	<u>7.58</u>
Japan	7.42	-	-	-	<u>7.42</u>

$$※ S_x = \sqrt{\frac{MSSW}{n_h}} = \sqrt{\frac{1.32}{1579.76}} = 0.03$$

여 다른 집단의 평균과 비교하거나 집단 간 평균 차($|\bar{x}_i - \bar{x}_j|$)와 LSR값을 직접 비교하는 두 가지 방식을 사용할 수 있다. 이들 방법 간의 차이와 분석절차를 단계별 표(Tables 3, 4, 5)로 정리하여 나타냈고, DMRT 절차 순서도(Fig. 5)를 나타내는 등 시각적 표현을 통해 명확하게 제시하고자 하였다. 분산분석결과 한국, 중국, 일본, 필리핀 자원의 아밀로스 및 단백질 평균들 간 유의성이 인정되어 국가별 자원들 간의 유의차가 있는지를 DMRT를 사용하여 확인하였다. 아밀로스 함량의 경우 계산된 S_x 는 0.09였고, 군내 자유도 6,950, 유의수준 0.01일 때 SSR값은 통계교과서에 수록된 ‘다중검정을 위한 스튜던트화한 범위’표를 참조하여 얻었다. 그 결과 계산된 LSR값과 아밀로스 평균함량 차의 값을 기준값으로 정하여 다중범위검정을 실시하였다. 아밀로스 분석의 경우 필리핀, 중국, 한국, 일본 자원의 네 집단으로 나눌 수 있었으며, 국가별 평균 함량은 각각 필리핀 25.18%, 중국 24.54%, 한국 22.08%, 일본 21.47%였다. 필리핀 자원의 DMRT 기준값(Amylose-LSR_{0.01})은 24.83이었고 나머지 세 자원들의 평균 아밀로스 함량 보다 크기 때문에 유의성이 있으며 결과적으로 필리핀 자원은 독립집단을 이룬다고 할 수 있었다. 이에 대한 결과를 구체적으로 해석해 보면 필리핀 자원은 한국, 일본의 자포니카 품종과는 구분되는 아밀로스 함량 특성을 나타낸 것으로 사료된다. 중국 자원의 기준값이 24.20이었고 한국 자원 평균값이 22.08, 일본 자원 평균값이 21.47이었으므로(Japan 21.47 < Korea 22.08 < China 24.20) 중국 자원과 한국, 일본 자원은 유의성이

있으므로 중국 자원은 독립집단을 이뤘다. 한국 자원의 기준값이 21.76이었고 일본 자원의 평균값이 21.47이었으므로(Japan 21.47 < Korea 21.76) 한국 자원과 일본 자원은 서로 다른 독립집단을 이뤘다(Table 12). 단백질 함량의 경우 계산된 S_x 는 0.03이었고, 군내 자유도 6,950, 유의수준 0.01일 때 SSR값은 ‘다중검정을 위한 스튜던트화한 범위표’를 참조하여 얻었다. 그 결과 계산된 LSR값과 단백질 평균함량 차의 값을 기준값으로 정하여 다중범위검정을 실시하였다. 단백질 분석의 경우 필리핀, 중국, 한국, 일본 자원의 네 집단으로 나눌 수 있었으며, 국가별 평균 함량은 각각 필리핀 8.19%, 중국 7.79%, 한국 7.58%, 일본 7.42%였다. 필리핀 자원의 DMRT 기준값(Amylose-LSR_{0.01})은 8.07이었고 나머지 세 자원들의 평균 단백질 함량 보다 크기 때문에 유의성이 있으며 결과적으로 필리핀 자원은 독립집단을 이룬다고 할 수 있었다. 이에 대한 결과를 구체적으로 해석해 보면 필리핀 자원은 한국, 일본의 자포니카 품종과는 구분되는 단백질 함량 특성을 나타낸 것으로 사료된다. 중국 자원의 기준값이 7.68이었고 한국 자원 평균값이 7.58, 일본 자원 평균값이 7.42이었으므로(Japan 7.42 < Korea 7.58 < China 7.68) 중국 자원과 한국, 일본 자원은 유의성이 있으므로 중국 자원은 독립집단을 이뤘다. 한국 자원의 기준값이 7.47이었고 일본 자원의 평균값이 7.42이었으므로(Japan 7.42 < Korea 7.47) 한국 자원과 일본 자원은 서로 다른 독립집단을 이뤘다(Table 13). 필리핀 자원의 아밀로스 평균함량은 25.18±2.84%였고, 단백질 평균함량은 8.19±1.22%였다. 중

Table 14. Non-glutinous rice amylose and protein content in breeding line germplasm by countries.

Origin	Amylose (%)			Protein (%)		
	Mean ²	Min.	Max.	Mean	Min.	Max.
Philippines	25.18±2.84 ^a	10.03	30.55	8.19±1.22 ^a	5.45	15.69
China	24.54±4.01 ^b	10.01	32.52	7.79±1.30 ^b	5.26	14.46
Korea	22.08±3.79 ^c	10.01	31.32	7.58±1.06 ^c	5.31	14.04
Japan	21.47±2.46 ^d	10.02	29.62	7.42±0.94 ^d	5.37	12.55

²Mean mean±standard deviation.³Mean separation within columns by Duncan's multiple range test at $p=0.01$.

국 자원의 아밀로스 평균함량은 24.54±4.01%였고, 단백질 평균 함량은 7.79±1.30%였다. 한국 자원의 아밀로스 평균함량은 22.08±3.79%였고, 단백질 평균함량은 7.58±1.06%였다. 일본 자원의 아밀로스 평균함량은 21.47±2.46%였고, 단백질 평균함량은 7.42±0.94%였다. 한국 자원과 일본 자원은 필리핀과 중국 자원에 비해 낮은 아밀로스 및 단백질 함량 분포를 나타냈다 (Table 14). 필리핀 벼 자원은 대부분 인디카 종이며, 한국과 일본의 벼 품종은 대부분 자포니카 종이므로 이와 같은 결과는 각각의 품종 특성이 반영된 것으로 사료된다. 일본의 경우 벼 품종의 육종 기준은 밥맛, 품질, 복합 내병성, 직과 적응성, 수량성의 5가지 항목에 중점을 두고 있으며 특히 밥맛검정 과정에서 NIRS를 이용하여 아밀로스, 단백질, 지방, 수분함량을 평가하고 있다(Park 2005). 일본 소비자들은 아밀로스 함량이 낮은 벼 품종을 선호하기 때문에 일본자원의 벼 품종의 육종 기준은 다른 국가별 자원들에 비해 낮은 아밀로스 함량을 나타내는 것으로 보인다.

적 요

본 연구는 선행연구에서 개발된 근적외선 분광분석(NIRS) 예측 모델을 활용하여 측정된 국내외 육성계통 벼 유전자원의 아밀로스 및 단백질 함량 자료를 통계처리 하여 자원의 지리적 특성과 성분 함량에 대한 정확한 정보를 제공하기 위해 실시하였다. 정규분포분석 결과 벼 유전자원의 아밀로스 평균은 23.6%였고, 단백질 평균은 7.9%였으며 전체 자원의 95%를 차지하는 자원들의 함량범위는 아밀로스가 15.7-31.5%, 단백질이 5.3-10.5%였다. 자원의 다양성지수는 아밀로스가 0.83, 단백질은 0.54였다. ANOVA, DMRT에 사용된 자원 수는 한국 자원이 2,386, 중국은 2,136, 일본은 1,219, 필리핀은 1,213자원이었다. 국가별 아밀로스 평균 함량은 한국 자원이 22.1%, 중국 자원은 24.5%, 일본 자원은 21.5%, 필리핀 자원은 25.2%였다. 단백질 평균함량은 한국 자원

이 7.6%, 중국 자원은 7.8%, 일본 자원은 7.4%, 필리핀 자원은 8.2%였다. ANOVA 결과 벼 유전자원의 아밀로스 및 단백질 함량은 국가별 차이가 있었고 1% 유의수준에서 차이가 인정되었다. DMRT 결과 국가별 아밀로스 함량은 한국, 중국, 일본, 필리핀의 네 집단으로 나눌 수 있었으며 각 집단 간 아밀로스 함량차이는 1% 유의수준에서 차이가 인정되었다. 단백질 함량은 한국, 중국, 일본, 필리핀의 네 집단으로 나눌 수 있었으며 각 집단 간 단백질 함량차이는 1% 유의수준에서 차이가 인정되었다. 일본 자원은 가장 낮은 아밀로스와 단백질 평균함량을 나타냈고, 필리핀 자원은 가장 높은 아밀로스와 단백질 평균함량을 나타냈다. 이러한 지리적 분포에 따른 벼 자원 간 함량차이는 각 지역별 자원 선호도와 계통 특성이 반영된 결과라고 할 수 있다.

사 사

본 연구는 농촌진흥청 농업과학기술연구 개발사업(과제번호: PJ01353904)의 지원에 의해 이루어졌습니다.

REFERENCES

1. Ahn HG, Kim YH. 2012. Discrimination of Korean domestic and foreign soybeans using near-infrared reflectance spectroscopy. Korean J Crop Sci 57: 296-300.
2. Bagchi TB, Sharma S, Chattopadhyay K. 2016. Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. Food Chem 191: 21-27.
3. Champagne ET, Aett KL, Vinyard BT, Webb BD, McClung AM, Barton FE. 1997. Effects of drying conditions, final moisture content and degree of milling on rice flavor. J Cereal Chem 74: 566-570.
4. Choi KS, Jung GH. 2017. Study on vinyl coating cultivation

- of potatoes under low temperature conditions. Kor J Plant Res 30: 556-564.
5. Hwang HG, Cho RK, Sohn JK, Lee SK. 1994. Rapid evaluation of chemical components of rice grain using near infrared spectroscopy. Korean J Crop Sci 39: 7-14.
6. Jang MR, Moon HK, Kim TR, Yuk DH, Kim EH, Hong CK, Choi CM, Hwang IS, Kim JH, Kim MS. 2011. The survey on pesticide residues in vegetables collected in Seoul. Kor J Pesticide Sci 15: 114-124.
7. Jeong JM, Jeung JU, Lee SB, Kim MK, Kim BK, Sohn JK. 2013. Physicochemical properties of rice endosperm with different amylose contents. Korean J Crop Sci 58: 274-282.
8. Juliano BO. 2003. Rice: In Caballero, B., L. Trugo, P. Finglas (eds.), Encyclopedia of Food Sciences and Nutrition (2nd ed.), Academic Press, London, UK. pp. 4995-5001.
9. Kim DJ, Oh SK, Lee JH, Yoon MR, Choi IS, Lee DH, Kim YG. 2012. Changes in quality properties of brown rice after germination. Kor J Food Sci Technol 44: 300-305.
10. Kim HI. 2004. Comparison of Korean and Japanese rice by NIR and chemical analysis. J East Asian Soc Dietary Life 14: 135-144.
11. Kim JS, Song MH, Choi JE, Lee HB, Ahn SN. 2008. Quantification of protein and amylose contents by near-infrared reflectance spectroscopy in aroma rice. Kor J Food Sci Technol 40: 603-610.
12. Kim KH, Heu MH, Park SZ, Koh HJ. 1991. New mutants for endosperm and embryo characters in rice. Korean J Crop Sci 36: 197-203.
13. Kim MK, Kim MY, Choi CH, Ko BG, Kwon SI, Weon HY. 2010. Pathogenic *E. coli* inactivation in upland soils to a change of soil moisture content and temperature. Kor J Soil Sci Fert 43: 90-95.
14. Kim SJ. 2016. Analysis of freshness of rice depending on packing material using MANOVA. Korean J Applied Statistics 29: 1421-1428.
15. Kim SS, Cha HC. 2017. Comparison of the total phenolic and flavonoid contents and antioxidant activities of four kinds of sand dune plants living in Taejeon. Korea. Kor J Plant Res 30: 8-16.
16. Lee CK, Kim JH, Choi MK, Kwak KS, Shin JC. 2010. Nitrogen application method for high quality and labor saving in rice production under amended standard N application level. Korean J Crop Sci 55: 70-75.
17. Lee I, Joo JC, Lee CS, Kim GY, Woo DY, Kim JH. 2017. Evaluation of the water quality changes in agricultural reservoir covered with floating photovoltaic solar-tracking system. J Kor Soc Environ Eng 39: 255-264.
18. Lim HI, Kim GN, Jang KH, Park WG. 2015. Effect of wet cold and gibberellin treatments on germination of dwarf stone pine seeds. Kor J Plant Res 28: 253-258.
19. Moon SS, Lee KH, Cho RK. 1994. Application of near infrared reflectance spectroscopy in quality evaluation of domestic rice. Kor J Food Sci Technol 26: 718-725.
20. Nam JI, Choi GE, Kim YM, Park JI. 2015. Analysis of morphological characteristics and variation among six populations of *Pourthiaea villosa* (Thunb.) Decne. var. *villosa* in Korea. Kor J Plant Res 28: 26-34.
21. Oh HS, Park YH, Kim JH. 2002. Isoflavone contents, antioxidative and fibrinolytic activities of some commercial cooking-with-rice soybeans. Korean J Food Sci Technol 34: 498-504.
22. Oh SJ, Lee MC, Choi YM, Lee SK, Oh MW, Ali A, Chae BS, Hyun DY. 2017a. Development of near-infrared reflectance spectroscopy (NIRS) model for amylose and crude protein contents analysis in rice germplasm. Kor J Plant Res 30: 38-49.
23. Oh SJ, Lee MC, Choi YM, Lee SK, Oh MW, Ali A, Chae BS, Hyun DY. 2017b. Fast systemic evaluation of amylose and protein contents in collected rice landrace germplasm using near-infrared reflectance (NIRS). Kor J Plant Res 30: 450-465.
24. Pak SI, Oh TH. 2010. The Application of analysis of variance (ANOVA). J Veterinary Clinics 27: 71-78.
25. Park KH. 2005. A recent rice breeding strategy in Japan. Kor J Intl Agri 17: 1-5.
26. Park SZ, Hahn WS, Chung NJ. 2016. Comparison of several groups In Biostatistics, KNOU Press, Seoul, Korea. pp. 230-250.
27. Sim MO, Lee HJ, Jang JH, Lee HE, Jung HK, Kim TM, No JH, Jung J, Jung DE, Cho HW. 2017. Anti-inflammatory and antioxidant effects of *Spiraea prunifolia* Sieb. et Zucc. var. *simpliciflora* Nakai in raw 264.7 cells. Kor J Plant Res 30: 335-342.
28. Son JR, Kim JH, Lee JI, Kim JK, Hwang HG, Moon HP. 2002. Trend and further of rice quality evaluation. Korean J Crop Sci 47: 33-54.
29. Son KA, Im GJ, Hong SM, Kim CS, Gil GH, Jin YD, Kim J, Ihm YB, Ko HS, Kim JE. 2013. The extrapolations to reduce the need for pesticide residues trials on continuous

- harvesting leafy vegetables. Kor J Pesticide Sci 17: 293-301.
30. Song JH. 2015. Effect of pretreatment methods on seed germination of honey plant, *Phellodendron amurense*. J Apic 30: 337-341.
31. Song J, Kim JH, Kim DS, Lee CK, Youn JT, Kim SL, Suh SJ. 2008. Physicochemical properties of starches in japonica rices of different amylose content. Korean J Crop Sci 53: 285-291.
32. Sung MK, Han SJ, Seo HJ, Choi SW, Nam SH, Chung JI. 2014. Genotype and environment influence on raffinose and stachyose content of soybean seed. Korean J Crop Sci 59: 319-324.
33. Williams P, Norris K. 1987. Near-Infrared Technology in Agricultural and Food Industries. American Association of Cereal Chemists, Inc., MN (USA). p. 330.
34. Yang CI, Yang SJ, Jeoung YP, Choi HC, Shin YB. 2001. Genotype $\times$ environment interaction of rice yield in multi-location trials. Korean J Crop Sci 46: 453-458.