서울대학교 SEOUL NATIONAL UNIVERSITY **보 도 자 료** http://www.snu.ac.kr

2021. 5. 11.(화) / 즉시

문의 : 담당자 연락처(02-880-4864)
연구책임자 최상호 교수(02-880-4857) / 교신저자
연구진 임한혁 박사과정(02-880-4864) / 제1저자

국내 연구진 햄버거병 원인균 장출혈성대장균의 새로운 치료제 개발 길 열어, 미국국립과학원회보 (PNAS) 논문 게재

- **장출혈성대장균은 인간의 신장 기능을 떨어뜨려 흔히 햄버거병이라고도 불리는 용혈성요독증후군을 일으키는 치명적인 식중독 감염균**이다. 서울대 농업생명과학대학 최상호 교수 연구팀과 이화여대 김병식 교수는 **감염균의 유전정보로부터 독성 여부를 판별할 수 있는 인공지능(AI, artificial intelligence) 모델을 개발하고, 이를 통해 장출혈성대장균의 새로운 독소인자들을 발굴**한 결과를 미국국립과학원회보(PNAS) 논문을 통해 5월 10일(미국시각 오후 03시) 발표하였다. 이 결과는 기후위기 속 **새롭게 출현하는 장출혈성대장균의 독성 예측과 신종 치료제 개발에 유용하게 활용될 수 있다는 점**에서 의미가 크다.

- 장출혈성대장균은 주로 오염된 소고기, 채소 등의 식품 섭취를 통해 영유아나 고령의 환자를 감염시키며, 신장 기능을 손상시켜 심할 경우 사망에까지 이르게 하는 용혈성요독증후군을 유발한다. 국내 장출혈성대장균 감염에 의한 환자는 해마다 100명 넘게 발생하고 있으며 매년 증가하고 있다. 장출혈성대장균은 독소인자인 부착단백질 인티민 (intimin)을 사용하여 인간에게 병을 일으킨다고 알려져 있어, 지금까지는 유전체 상 인티민 유전자의 존재 여부를 기준으로 장출혈성대장균의 독성을 판별해 왔다. 하지만, 2011년 유럽에서 약 3,000명을 감염시켜 50명의 사망자를 낳은 장출혈성대장균은 인티민 유전자를 보유하지 않은 변이주들로 밝혀진 바 있다. 이러한 변이주들에 의한 감염사태는 장출혈성대장균들의 잠재적 독성을 판별하는 기준으로 새로운 독소인자들을 발굴할 필요성을 제기하였다.

- 본 연구는 유전정보로부터 장출혈성대장균의 독성을 판별할 수 있는 인공지능 모델을 개발하였다. 개발한 인공지능 모델은 지금까지 유전정보가 공개된 2,646개의 장출혈성대장균들의 독성 여부를 약 99%의 높은 정확도로 판별하였다. 또한 인공지능 모델은 기존의 방법으로는 판별할 수 없었던 장출혈성대장균 변이주들의 독성을 정확하게 판별하였다. 나아가 **본 연구는 인공지능 모델을 분석해 장출혈성대장균의 독성 판별에 기준이 되는 새로운 독소인자들도 발굴**하였다. 발굴한 새로운 기준 독소인자들 중 상당수는 아직 그 기능이 밝혀져 있지 않다. 하지만, 이 신규 발굴 기준 독소인자들의 기능을 규명함으로써 장출혈성대장균들의 새로운 발병기전을 제시할 수 있고, 이들을 타겟으로 하는 새로운 치료기술 개발 역시 가능할 것으로 예상된다.

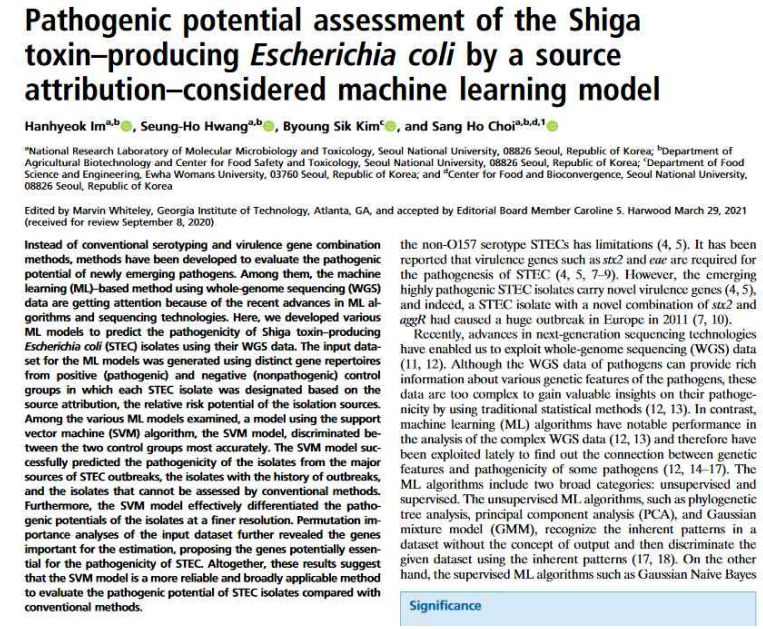
- 결론적으로, 본 연구는 유전정보만으로 장출혈성대장균의 독성을 판별할 수 있는 인공지능 모델을 개발하였다. 개발한 인공지능 모델은 지금까지 독성 판별이 불가능했던 장출혈성대장균 변이주들의 독성을 판별할 수 있었다. 또한, 장출혈성대장균에 독성을 부여하는 기준 독소인자들을 발굴함으로써 새로운 치료법을 개발할 수 있는 길을 열었다. 이러한 연구결과는 장출혈성대장균 뿐만 아니라, 다른 감염균들의 독성을 판별하고 기준 독소인자들을 발굴하기 위한 인공지능 모델을 개발하는 새로운 연구들의 기초자료로도 활용될 수 있을 것이다.

- 본 연구는 한국연구재단 이공분야기초연구사업 (세부과제번호: 2017R1E1A1A01074639)의 지원으로 이루어졌다.

- 참여국가 및 기관: 대한민국 (서울대학교, 이화여자대학교)

[붙임] 1. 연구결과

연 구 결 과



Pathogenic potential assessment of the Shiga toxin–producing *Escherichia coli* by a source attribution–considered machine learning model

Hanhyeok Im[a,b], Seung-Ho Hwang[a,b], Byoung Sik Kim[c], and Sang Ho Choi[a,b,d,1]

[a]National Research Laboratory of Molecular Microbiology and Toxicology, Seoul National University, 08826 Seoul, Republic of Korea; [b]Department of Agricultural Biotechnology and Center for Food Safety and Toxicology, Seoul National University, 08826 Seoul, Republic of Korea; [c]Department of Food Science and Engineering, Ewha Womans University, 03760 Seoul, Republic of Korea; and [d]Center for Food and Bioconvergence, Seoul National University, 08826 Seoul, Republic of Korea

Edited by Marvin Whiteley, Georgia Institute of Technology, Atlanta, GA, and accepted by Editorial Board Member Caroline S. Harwood March 29, 2021 (received for review September 8, 2020)

Instead of conventional serotyping and virulence gene combination methods, methods have been developed to evaluate the pathogenic potential of newly emerging pathogens. Among them, the machine learning (ML)–based method using whole-genome sequencing (WGS) data are getting attention because of the recent advances in ML algorithms and sequencing technologies. Here, we developed various ML models to predict the pathogenicity of Shiga toxin–producing *Escherichia coli* (STEC) isolates using their WGS data. The input dataset for the ML models was generated using distinct gene repertoires from positive (pathogenic) and negative (nonpathogenic) control groups in which each STEC isolate was designated based on the source attribution, the relative risk potential of the isolation sources. Among the various ML models examined, a model using the support vector machine (SVM) algorithm, the SVM model, discriminated between the two control groups most accurately. The SVM model successfully predicted the pathogenicity of the isolates from the major sources of STEC outbreaks, the isolates with the history of outbreaks, and the isolates that cannot be assessed by conventional methods. Furthermore, the SVM model effectively differentiated the pathogenic potentials of the isolates at a finer resolution. Permutation importance analyses of the input dataset further revealed the genes important for the estimation, proposing the genes potentially essential for the pathogenicity of STEC. Altogether, these results suggest that the SVM model is a more reliable and broadly applicable method to evaluate the pathogenic potential of STEC isolates compared with conventional methods.

the non-O157 serotype STECs has limitations (4, 5). It has been reported that virulence genes such as *stx2* and *eae* are required for the pathogenesis of STEC (4, 5, 7–9). However, the emerging highly pathogenic STEC isolates carry novel virulence genes (4, 5), and indeed, a STEC isolate with a novel combination of *stx2* and *aggR* had caused a huge outbreak in Europe in 2011 (7, 10).

Recently, advances in next-generation sequencing technologies have enabled us to exploit whole-genome sequencing (WGS) data (11, 12). Although the WGS data of pathogens can provide rich information about various genetic features of the pathogens, these data are too complex to gain valuable insights on their pathogenicity by using traditional statistical methods (12, 13). In contrast, machine learning (ML) algorithms have notable performance in the analysis of the complex WGS data (12, 13) and therefore have been exploited lately to find out the connection between genetic features and pathogenicity of some pathogens (12, 14–17). The ML algorithms include two broad categories: unsupervised and supervised. The unsupervised ML algorithms, such as phylogenetic tree analysis, principal component analysis (PCA), and Gaussian mixture model (GMM), recognize the inherent patterns in a dataset without the concept of output and then discriminate the given dataset using the inherent patterns (17, 18). On the other hand, the supervised ML algorithms such as Gaussian Naive Bayes

Significance

MICROBIOLOGY

(2021년 03월 29일, PNAS Accept 됨)

Pathogenic potential assessment of the Shiga toxin-producing *Escherichia coli* by a source attribution-considered machine learning model

본 연구는 **유전정보를 활용하여 장출혈성대장균들의 잠재적 독성을 판별하는 새로운 방법을 제시**하였다. 이제까지는 인티민 등의 특정 독소인자의 존재 여부를 분석하여 장출혈성대장균들의 독성을 판별하였으나, 이러한 방법들은 새로운 독소인자를 가지는 장출혈성대장균 변이주들의 독성을 판별할 수 없었다. 본 연구는 인공지능 알고리즘 기술과 유전정보 분석기술을 활용하여 **장출혈성대장균 분리주들의 유전정보로부터 독성을 판별할 수 있는 인공지능 모델을 개발**하였다.

개발한 인공지능 모델은 약 99%의 높은 정확도로 지금까지 유전정보가 공개된 2,646개의 장출혈성대장균 분리주들에 대하여 실제 병을 일으킨 전력이 있는 임상분리주들은 독성이 있다고, 병을 일으킨 전력이 없는 식품 혹은 환경 유래의 분리주들은 독성이 없다고 판별하였다. 또한, 개발한 인공지능 모델은 기존의 독소인자 분석 방법으로 판별할 수 없었던 장출혈성대장균 변이주들의 독성을 정확하게 판별하였다.

추가로 본 연구에서는 개발한 인공지능 모델을 심층 분석해 장출혈성대장균 분리주들의 독성을 판별하는데 기준이 되는 새로운 독소인자들을 밝혀냈다. 결론적으로, 본 연구는 환자나 식품, 환경 등에서 분리된 장출혈성대장균 분리주들의 독성뿐 아니라, 기존의 독소인자 분석 방법으로는 독성 판별이 불가능했던 변이주들의 독성 또한 판별할 수 있는 정확도 높은 인공지능 모델을 개발하였다.

나아가 본 연구는 장출혈성대장균의 독성 판별에 기준이 되는 새로운 독소인자들을 발굴하여 이들을 타겟으로 하는 새로운 치료기술 개발의 가능성을 제시하였다. 이러한 연구결과는 다른 감염균들의 독성을 판별하는 인공지능 모델을 개발하고, 치료 타겟으로 새로운 독소인자를 발굴하는 미래의 연구들에도 활용될 수 있다는 점에서 그 의의가 있다.