Research Portfolio

Contents

2022.03.03.

Yunsoo Kim

de novo Assembly of *Tylosema esculentum* Chloroplast Genome

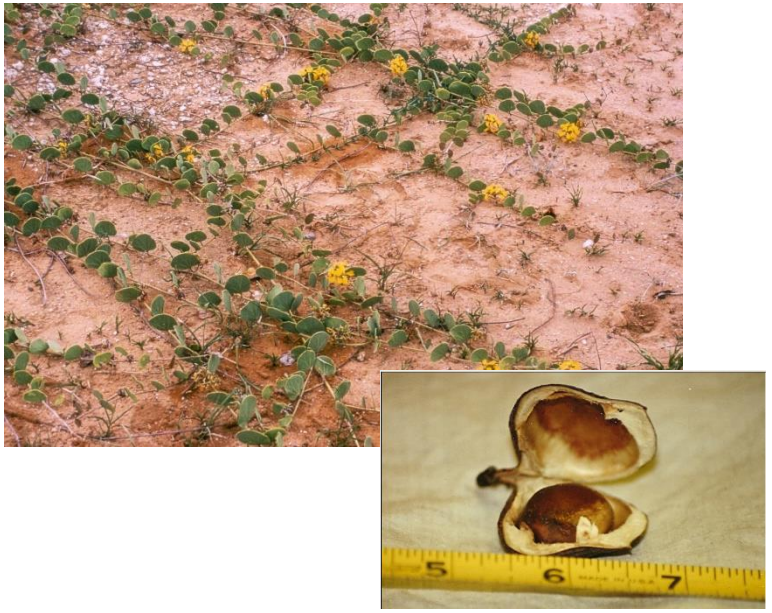
Background

Marama (*Tylosema esculentum*)

- A perennial, wild tuber-producing, and non-nodulating basal legume.
- A possible crop with high drought tolerance.

Hybrid genome assembly

- Hybrid genome assembly to minimize the PCR gap filling step
- Used Illumina short reads and PacBio long reads



Results

Chloroplast genome

- Total Length 161,537 bp
- Large and small single copy (LSC and SSC) region separated by a pair of inverted repeats
- Identified 7479 bp unique inversion in LSC (publication)
  - Not present in any other legumes
  - Possible source of the high drought tolerance

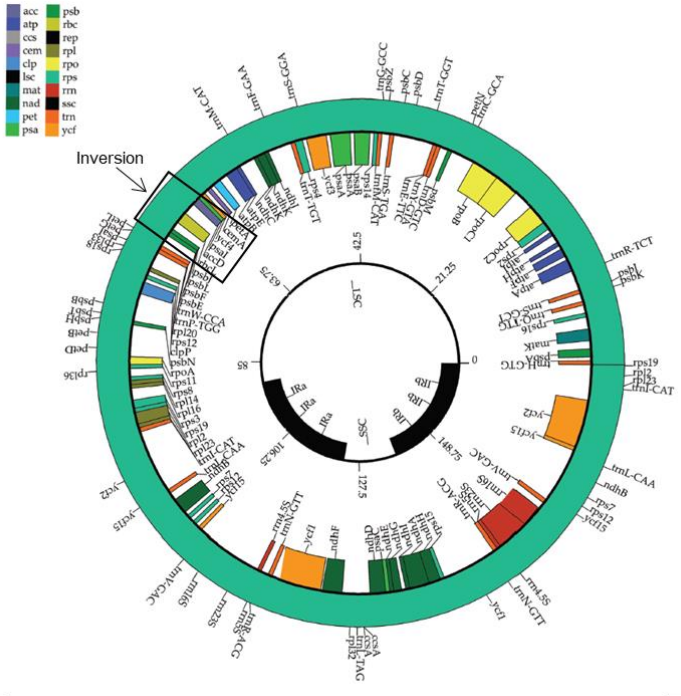


Fig. 1. Circular gene map of the *Tylosema esculentum* (Genistoid; Fabaceae) plastid genome.

A Novel Metagenomics Network Inference Methodology using proportionality measures

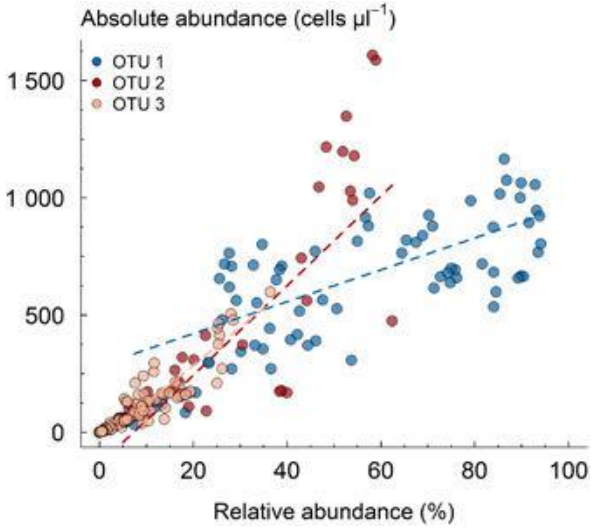
Background

Metagenomics : Relative Abundances

- 16S rRNA Sequences binned by similarity into operational taxonomic units (OTUs)
- OTUs measure of the relative abundance
  - Number of sequences not equal to the true number of cells
  - Caused by the amplification of DNA library

Network between OTUs

- For symbiotic dependencies
- For relative data, proposed a novel network inference method
  - Goodness of fit to proportionality measures



$$\phi_{(A,B)} = \frac{var(log(A/B))}{var(logA)}$$

$$\phi_{(B,A)} = \frac{var(log(B/A))}{var(logB)}$$

Goodness of fit to proportionality measures

Results

Network Inference Methodology

- Differential Network (paper)
  - Tested on multiple EBI datasets including chron's disease
  - Developed web-server
- Co-occurrence network (paper)
  - Benchmarked 8 existing methods: highest precision (0.907)
  - Used Theano to compute on GPU

Methods	Sensitivity	Specificity	Precision
Proportionality	0.259	0.999	0.907
LSA	0.280	1.000	0.586
Spearman	0.299	0.999	0.246
MIC	0.166	1.000	0.725
SparCC	0.337	0.998	0.202
Bray-Curtis	0.106	0.999	0.177
CoNET	0.086	0.998	0.050
RMT	0.039	1.000	0.129
Pearson	0.246	0.973	0.012

Table 1. Performance Summary

Machine learning based omics time series analysis to find significant biomarker in a cohort study

Background and Methods

No established time series method in omics

- Omics is a snapshot of the biological system
  - Need to have samples at different time points
  - Small number of time points and samples (replicates)
  - Other limitations: multivariate, noise, missing values

Machine Learning based analysis

- Clustering of time series and calculate mean trajectory
- For each cluster, apply smooth spline fit to the mean trajectory
  - The smoothing factor is used to fit time series in each cluster
- Distance between mean trajectory and in each time series in a cluster is calculated and used to provide p-value by significance testing
- Extened to include network inference methodology

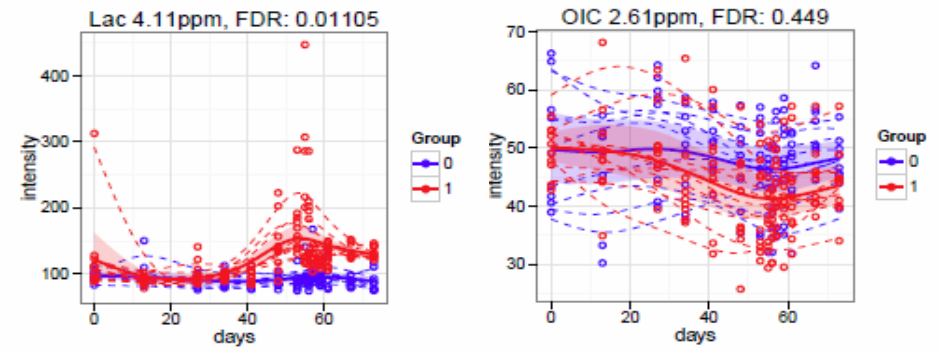


Fig.1. Time series for significant metabolite (left), nonsignificant metabolite (right)

Results

Method tested with simulated scenarios

- Sensitivity tested using simulated scenarios (paper)
  - 81 scenarios with varying parameters
  - Mean sensitivity 0.8
  - 10 replicates are found to be the minimum suggestion for the method to perform well
- Tested on *Schistosoma mansoni* infection dataset
  - Found known metabolites
  - A new metabolite found : Lactate a potential biomarker

metabolites	p-adjusted	Test statistics
Hip_7.55_t	0.011	1097.94
Hip_7.64_t	0.011	474.83
Hip_7.84_d	0.011	1062.85
Lac_4.11_q	0.011	1852.11
N.AG_2.06_s	0.011	745.86
OAP_2.22_t	0.036	510.87
p.CG_2.3_s	0.001	1107.19

Table 1. List of significantly different metabolites for *Schistosoma mansoni* cohort data

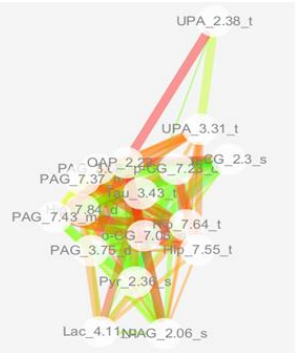


Fig.2. Network Inference Significant Metabolites Red – lower in case, Green – higher in case

Neural network model for protein toxicity prediction from amino acid

Background and Dataset

GM[1] Seed Needs to Pass Safety Approval by Law

- Korea follows the CODEX Alimentarius.
  - Tests on toxicity, allergenicity, nutrient, insert gene property
  - GM seed needs to have substantial equivalence
- Needs pre-screening of safe gene candidate during the discovery phase
  - Global company (such as Monsanto) has in-house safety prediction
  - The prediction results are enclosed for safety approval

Dataset created from SwissProt/TrEMBL

- Downloaded dataset from the previous work for comparison
  - Positive (Toxin) : 8,093 sequences
  - Negative : 62,581 sequences
    : 47,144 (Easy); 8,034 (Moderate); 7,403 (Hard)
  - 75:25 split for train and validation dataset
- Created own database for screening pipeline
  - Positive (Toxin) : 7,227 sequences
  - Negative : 62,581 sequences
    : 50,000 (Easy); 7,229 (Moderate); 6,652 (Hard)

Results

Created a neural network model

- N-Gam based feature extraction from amino acid sequences
- Multilayer perceptron for binary classification

Training Results

- Validation Accuracy 99.38%
- Validation Speed 16,514 sequences / seconds
- For speed comparison, reproduced SVM result
  - Validation Accuracy 96.22%, 22.5 sequences / seconds

Reached SOTA[2]

- Previous SOTA based on SVM (ToxClassifier, 2016)
- 2% Precision Improvement – Patent Application

eval metric	ToxClassifier	Our Method
Specificity	0.99	**0.997**
Sensitivity (Recall)	**0.97**	0.969
Precision	0.96	**0.980**
F1	0.97	**0.975**

1) GM – Gene Modification; 2) SOTA – state of the art

Plant disease severity classification ensemble model to accelerate breeding research

Background and Dataset

Digital Transformation for Breeding

- Disease severity analysis is necessary for disease resistant seed development.
- The severity analysis limitations:
  - Depends on breeder's personal knowledge
  - Lacked systematic management and reproducibility.
- Improvement needed in the process efficiency

Curated Public Dataset with Plant Pathologists

- Collected publicly available tomato bacterial spot datasets.
  - Healthy : 13,110 images; Bacterial Spot : 7,368 images
- Manually labeled subset of the datasets.
  - Randomly sampled and labeled 1,783 images for train
    : Severity from 0 to 5
  - 90 images were sampled for validation

Bacterial Spot of Tomato Severity Scale – From Left to Right 0, 1, 2, 3, 4, 5 Scale

Results

Used Ensemble of CNN models

- ResNext101 trained on ImageNet
- EfficientNet B7 trained on ImageNet

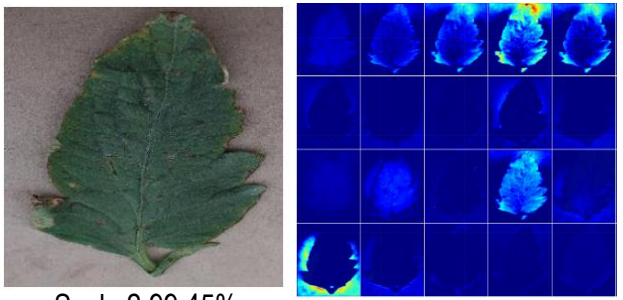
Training Results

- Training Accuracy 98.82%
- Validation Accuracy 94.4%
- Validation Speed 0.078 seconds / image

Researchers Evaluation

- Average Accuracy 89.2% (80.3/90 images)
- Average Speed 432.3 seconds (4.80 seconds / image)

Worked on Grad-CAM for XAI



Scale 2 99.45%

Plant disease classification system for customer (farmers) services

Background and Dataset

Crop Protection Recommendation System

- Needed customer services for inexperienced farmers who suffer from pests and insects
  - Guidance for the control method
  - Pesticide and insecticide recommendation
- Used for crop protection production management
  - Track the number of customers damaged by pests and insects

Made Custom Dataset

- Visited farms with marketing employees
  - Collected 5,800 images
  - Apple, Cucumber, Persimmon were selected
    : Total 7 classes (3 healthy, 4 diseases – 2 for cucumber)
  - 8:2 split for train and validation dataset

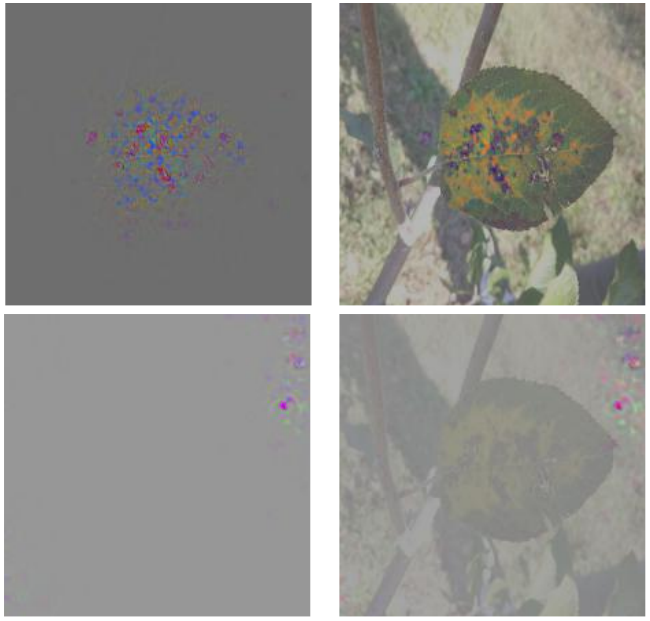
Apple Marssonia Blotch (top) and Healthy Leaves (bottom)

Results

Used Ensemble of CNN models

- ResNext101 trained on ImageNet
- EfficientNet B7 trained on ImageNet
- Validation Accuracy 96.69%
- Test Accuracy 87.5% (105/120 images)

Guided Grad-CAM for XAI



Apple Marssonia Blotch 99.60%

Wrong Class 0.01%

Plant disease classification system for customer (farmers) services

딥러닝을 이용한 작물 병충해 판독 서비스

사진 한 장으로 손쉽게 작물의 병충해 감염 여부와 병충해 이름을 알아보세요.

작물 사진 파일을 끌어다 놓으시거나
여기를 눌러 사진을 골라주세요.

딥러닝을 이용한 작물 병충해 판독 서비스

사진 한 장으로 손쉽게 작물의 병충해 감염 여부와 병충해 이름을 알아보세요.

파일명 : 20181011_135114.jpg

병해 확인하기 >

판독 상세

병해명
오이 흰가루병

확률
99.99%

병해 정보

발병시기 및 발병조건

- 병원균은 병이 발생되었던 피해부위에서 월동하여 다음 해에 공기로 전염되며 줄기, 잎, 과실 등 지상부에 전부 침입한다.
- 하우스 재배 오이에 피해가 많다. 처음에는 잎에 밀가루를 뿌린듯한 흰 병반이 생겼다가 잎 전체로 퍼진다.
- 병세가 진전되면 회색 내지 암회색이 되며 결국에는 황변고사하여 낙엽이 진다.

방제방법

- 병에 걸린 포기는 일찍 제거하여 태운다.
- 광선과 바람이 잘 통하도록 포장관리를 하며 칼리비료를 증시한다.
- 질소질비료의 과용을 피한다.
- 촘촘하게 심지 않도록 하고 통풍에 유의한다.

방제 약제

더마니 수용제 클릭해서 상세보기 | 미리본 액상수화제 클릭해서 상세보기 | 바이코 수화제 클릭해서 상세보기 | 부티나 액상수화제 클릭해서 상세보기

Accelerate research process in LG Chem by chemical and materials science NLP

Background

Artificial Reading System - CLUE

- Recognizes chemical molecules and property from documents
- Use the recognized entities to build knowledge graphs
- Expected to
  - Shorten the incubation time for new projects
  - Prevent sunk costs in the research process
  - Discover research candidates faster

CLUE served to 11 internal organizations

- AI reads and analyzes 1,000 documents per day
- Service for various fields of research
  - Synthetic biology, polymer, battery, *etc*
- Achievements
  - New additive for polymer is found
  - New catalysis candidate is found

Tasks	LGC	Competitor
NER for Chemistry	**96%**	94% (Korea Univ/Naver)
Patent Classification	**98%**	95% (WIPS)

History of Milestones

Pre-trained Model with Custom Corpus

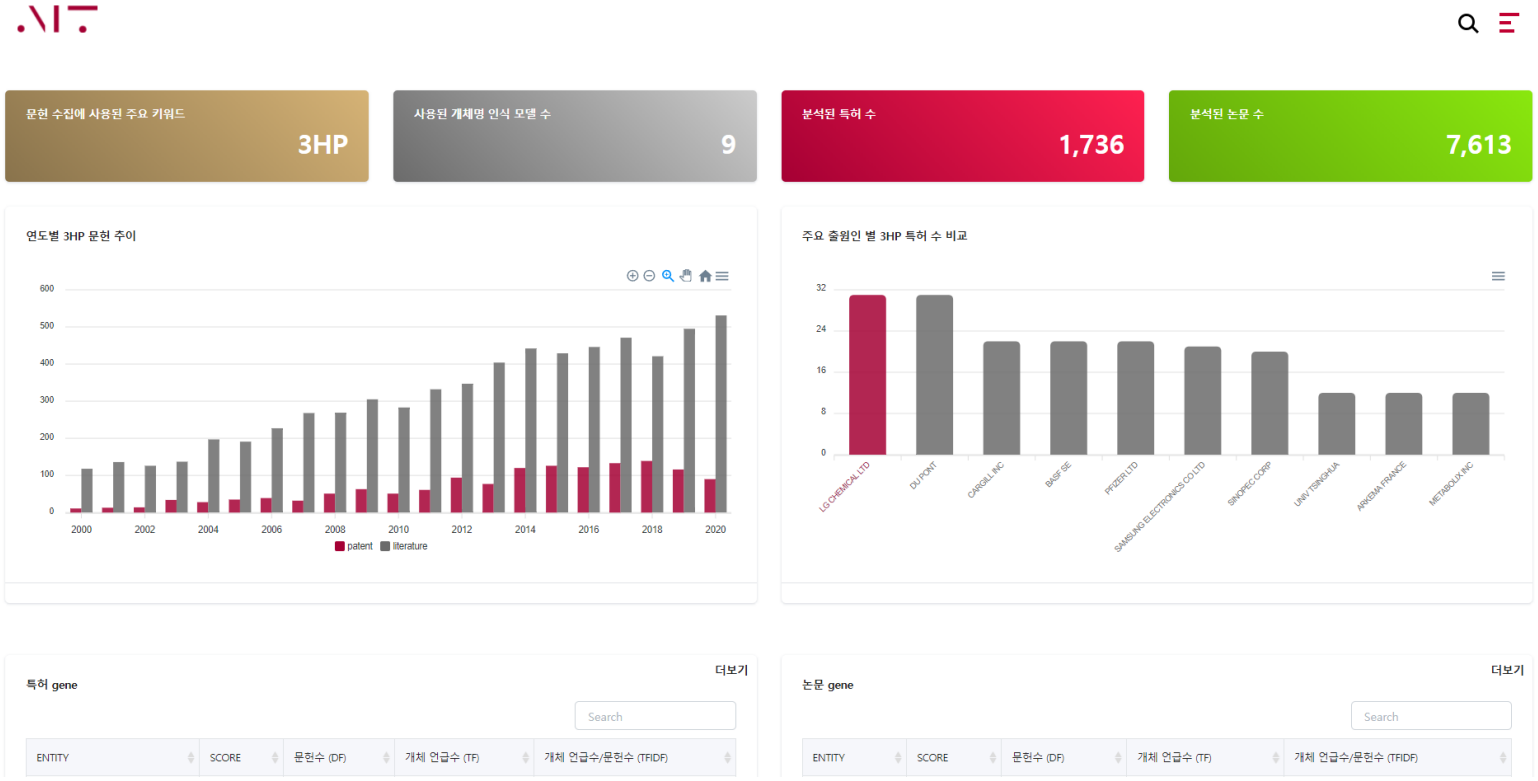
- BERT base (2019)
- BERT large (2020)
- RoBERTa with knowledge distillation (2021)
- Working on DeBERTa-v3 (2022)

Fine-tuning Model with Custom Dataset

- 100 domain specific entity group NER (2019~2021)
- LG Chem business area patent classification (2020)
- Working on Multi-task learning (2022)
- Working on knowledge graphs construction – GNN (2022)

Custom-Entity	F1 (%)
Solvent	90.4
Catalyst	87.1
Monomer	91.2
Temperature	92.0
Time	92.7
pH	79.7
Yield	98.2

Dashboard


문헌 수집에 사용된 주요 키워드
3HP
사용된 개체명 인식 모델 수
9
분석된 특허 수
1,736
분석된 논문 수
7,613
연도별 3HP 문헌 주이
patent
literature
주요 출원인 별 3HP 특허 수 비교
LG CHEMICAL LTD
DU PONT
CARGILL INC
BASF SE
PFIZER LTD
SAMSUNG ELECTRONICS CO LTD
SINOPEC CORP
UNIV TSINGHUA
ARKEMA FRANCE
METABOLIX INC
더보기
특허 gene
Search
ENTITY
SCORE
문헌수 (DF)
개체 언급수 (TF)
개체 언급수/문헌수 (TFIDF)
더보기
논문 gene
Search
ENTITY
SCORE
문헌수 (DF)
개체 언급수 (TF)
개체 언급수/문헌수 (TFIDF)

Research candidate recommendation

더보기

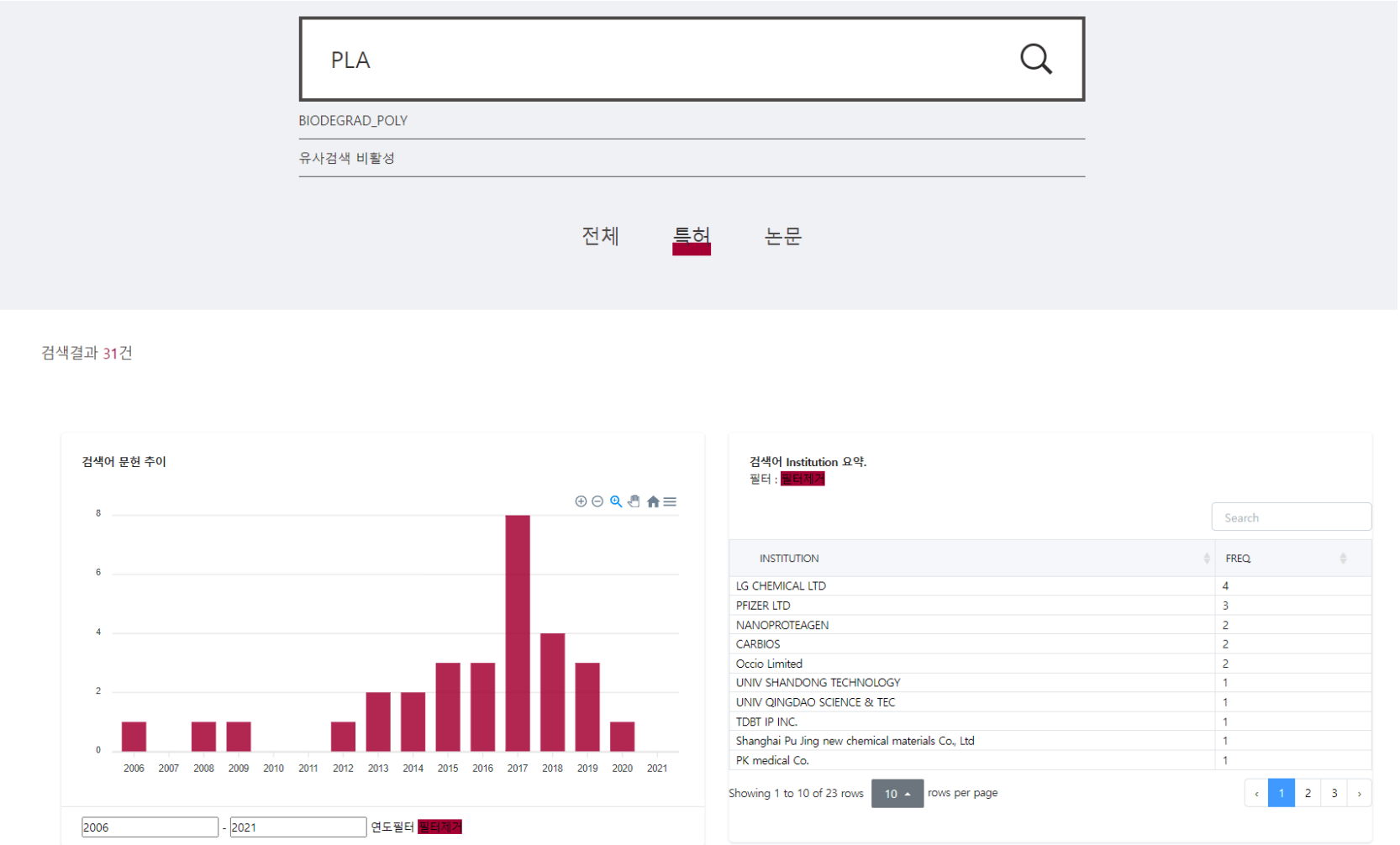
특허 biodegrad_poly

Search

ENTITY	SCORE	문헌수 (DF)	개체 언급수 (TF)	개체 언급수/문헌수 (TFIDF)
[illegible]	1.0	37	219	836.079
[illegible]	0.861	15	129	604.07
[illegible]	0.74	3	73	443.037
[illegible]	0.688	76	169	525.842
[illegible]	0.661	41	123	457.268
[illegible]	0.592	54	117	403.412
[illegible]	0.509	2	42	266.981
PLA	0.436	31	63	251.342
[illegible]	0.389	15	44	206.039
[illegible]	0.366	9	36	185.498

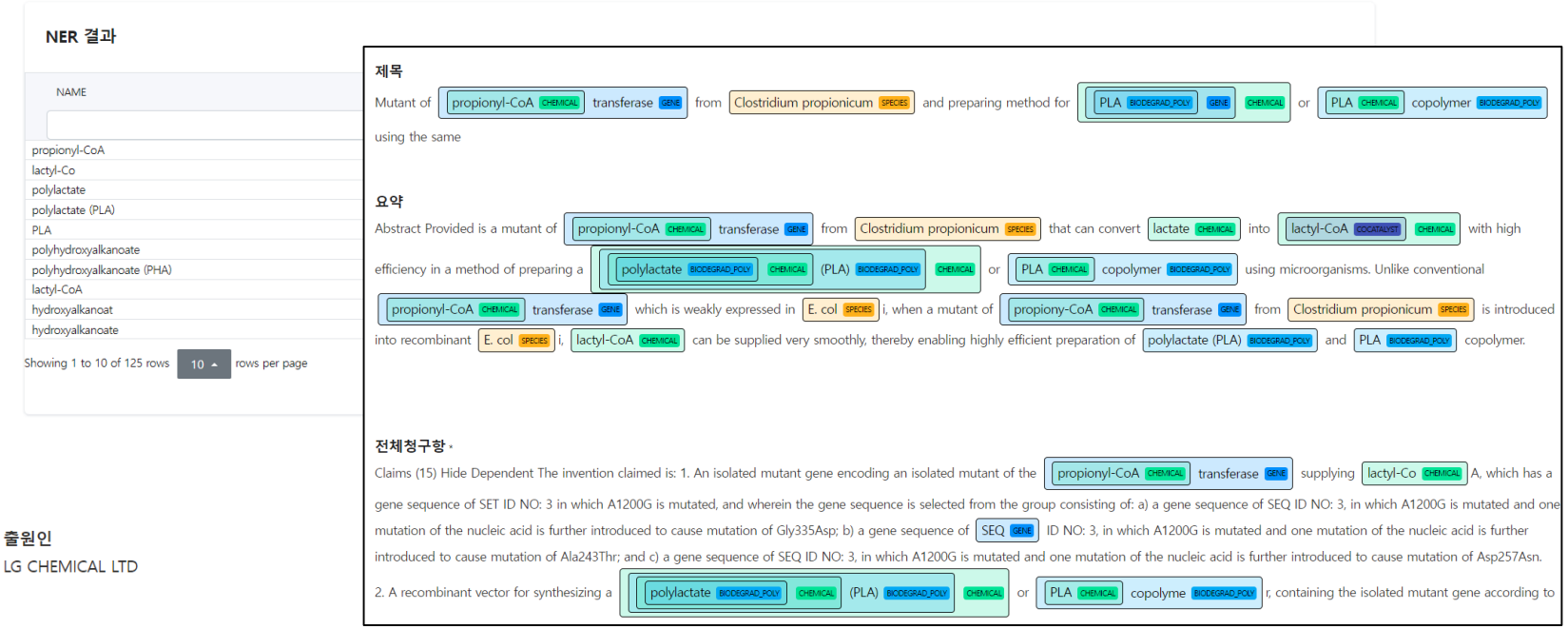
Showing 1 to 10 of 10 rows

Document results for the research candidate



Domain-specific NER tagged for documents

US9957533B2



Further analysis (here co-occurrence analysis is shown as an example)

cocatalyst

methylaluminoxane | tetrakis(pentafluorophenyl)borate

comonomer

pentadiene | 1-butene | styrene | ethylene | 1-hexene | propylene | 1-octene

Co-occurrence Plot