

# Research Portfolio

---

## Contents

- I. Experience from University
- II. Agri-biology Deep Learning
- III. NLP Group Leader

2022.03.03.

Yunsoo Kim



### A Novel Metagenomics Network Inference Methodology using proportionality measures

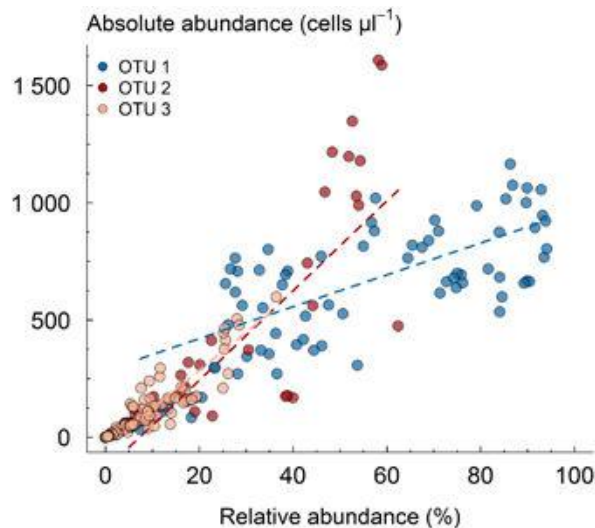
#### Background

##### 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



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

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
- Extended to include network inference methodology

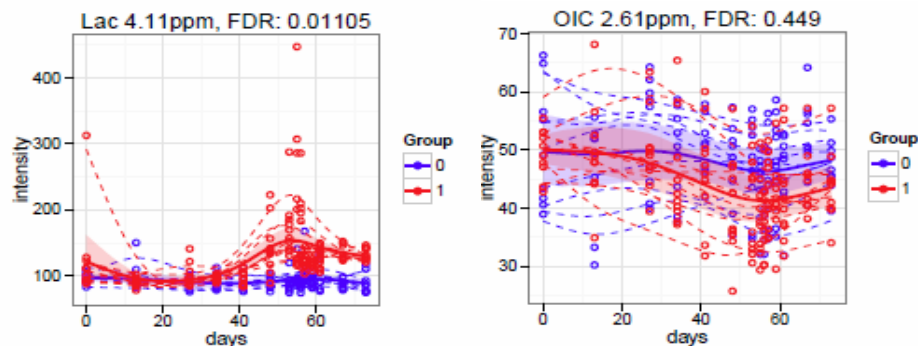


Fig.1. Time series for significant metabolite (left), nonsignificant metabolite (right)

### 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

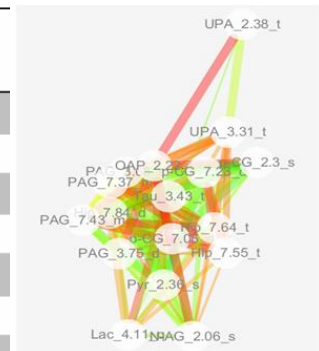


Fig.2. Network Inference Significant Metabolites  
Red – lower in case,  
Green – higher in case

## Neural network model for protein toxicity prediction from amino acid

### Background and Dataset

#### GM<sup>1)</sup> 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<sup>2)</sup>

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

eval metric	ToxClassifier	Our Method
Specificity	0.99	<b>0.997</b>
Sensitivity (Recall)	<b>0.97</b>	0.969
Precision	0.96	<b>0.980</b>
F1	0.97	<b>0.975</b>

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

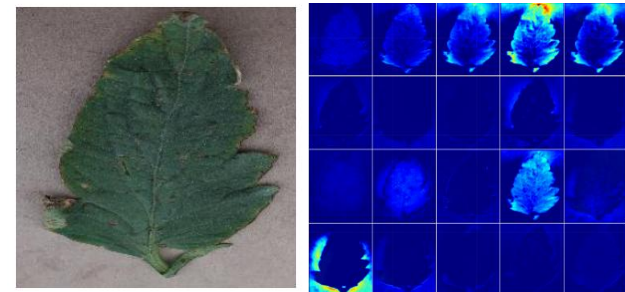
##### Training Results

- 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



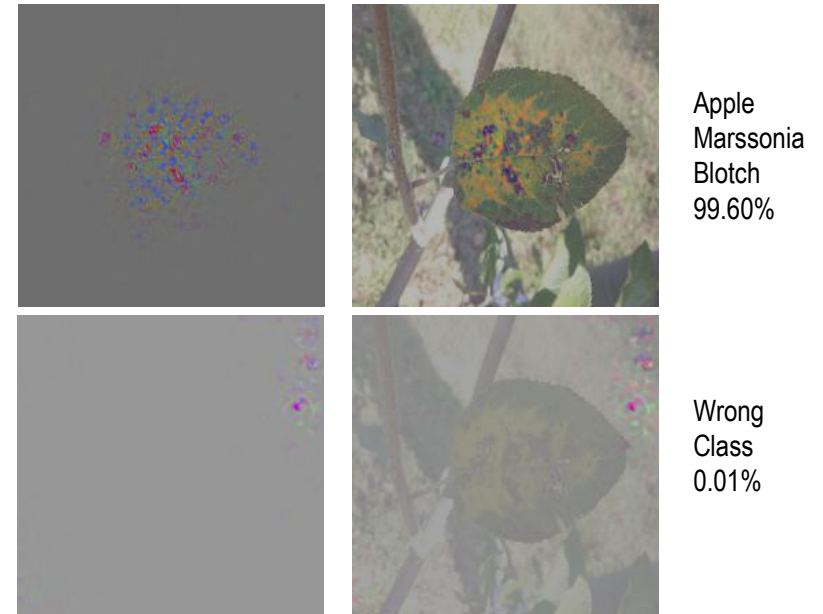
Apple Marssonnia Blotch (top) and Healthy Leaves (bottom)

#### 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



### 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.



#### 딥러닝을 이용한 작물 병충해 판독 서비스

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



#### 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.
- 출출하게 심지 않도록 하고 통풍에 유의한다.

#### 방제 약제





## 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 for further analysis for new insights
- 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

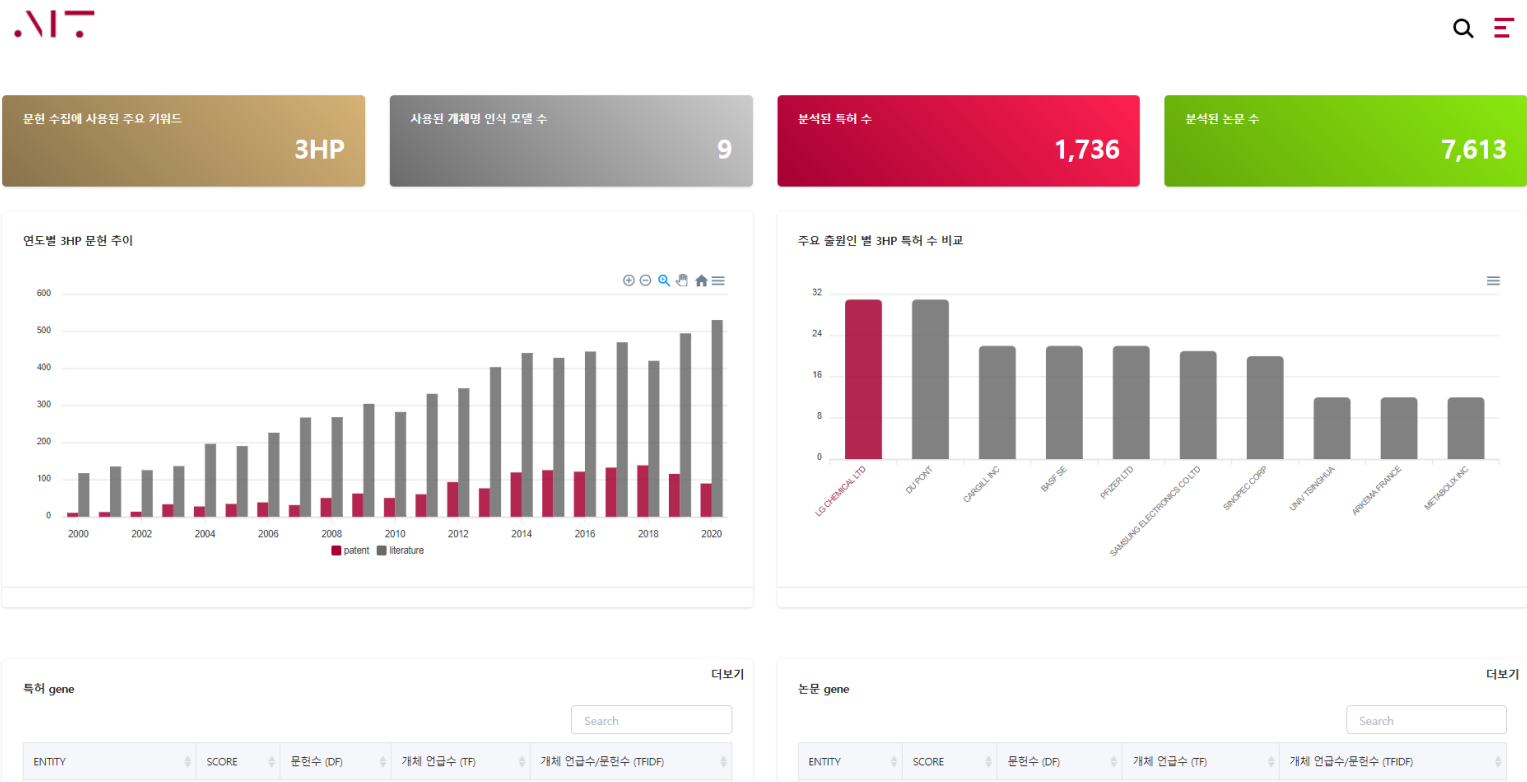
- BERT base (2019)
- BERT large (2020)
- RoBERTa with knowledge distillation (2021)
- Working on DeBERTa-v3 (2022)

#### 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)

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



## Research candidate recommendation

특히 biodegrad\_poly

더보기

Search

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

Showing 1 to 10 of 10 rows

Document results for the research candidate

PLA

BIODEGRAD\_POLY

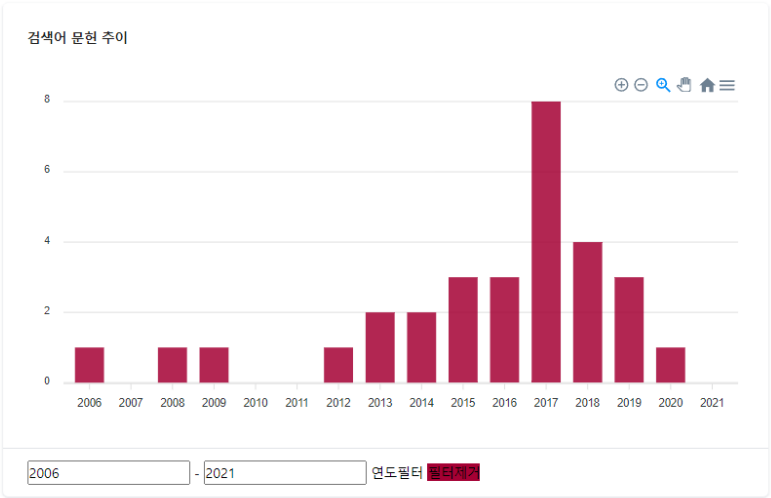
유사검색 비활성

전체

특허

논문

검색결과 31건



검색어 Institution 요약.  
필터:

INSTITUTION	FREQ
LG CHEMICAL LTD	4
PFIZER LTD	3
NANOPROTEAGEN	2
CARBIOS	2
Occio Limited	2
UNIV SHANDONG TECHNOLOGY	1
UNIV QINGDAO SCIENCE & TEC	1
TDBT IP INC.	1
Shanghai Pu Jing new chemical materials Co., Ltd	1
PK medical Co.	1

Showing 1 to 10 of 23 rows 10 rows per page

### Domain-specific NER tagged for documents

US9957533B2

**NER 결과**

NAME

propionyl-CoA

lactyl-Co

polylactate

polylactate (PLA)

PLA

polyhydroxyalkanoate

polyhydroxyalkanoate (PHA)

lactyl-CoA

hydroxyalkanoat

hydroxyalkanoate

Showing 1 to 10 of 125 rows    10 rows per page

**제목**

Mutant of propionyl-CoA transferase from Clostridium propionicum and preparing method for PLA or PLA copolymer using the same

**요약**

Abstract Provided is a mutant of propionyl-CoA transferase from Clostridium propionicum that can convert lactate into lactyl-CoA with high efficiency in a method of preparing a polylactate (PLA) or PLA copolymer using microorganisms. Unlike conventional propionyl-CoA transferase which is weakly expressed in E. coli, when a mutant of propionyl-CoA transferase from Clostridium propionicum is introduced into recombinant E. coli, lactyl-CoA can be supplied very smoothly, thereby enabling highly efficient preparation of polylactate (PLA) and PLA copolymer.

**전체 청구항**

Claims (15) Hide Dependent The invention claimed is: 1. An isolated mutant gene encoding an isolated mutant of the propionyl-CoA transferase supplying lactyl-CoA, which has a gene sequence of SET ID NO: 3 in which A1200G is mutated, and wherein the gene sequence is selected from the group consisting of: a) a gene sequence of SEQ ID NO: 3, in which A1200G is mutated and one mutation of the nucleic acid is further introduced to cause mutation of Gly335Asp; b) a gene sequence of SEQ ID NO: 3, in which A1200G is mutated and one mutation of the nucleic acid is further introduced to cause mutation of Ala243Thr; and c) a gene sequence of SEQ ID NO: 3, in which A1200G is mutated and one mutation of the nucleic acid is further introduced to cause mutation of Asp257Asn.

2. A recombinant vector for synthesizing a polylactate (PLA) or PLA copolymer, containing the isolated mutant gene according to

출원인

LG CHEMICAL LTD



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

—  
cocatalyst

methyaluminoxane    tetrakis(pentafluorophenyl)borate

—  
comonomer

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

—  
Co-occurrence Plot

