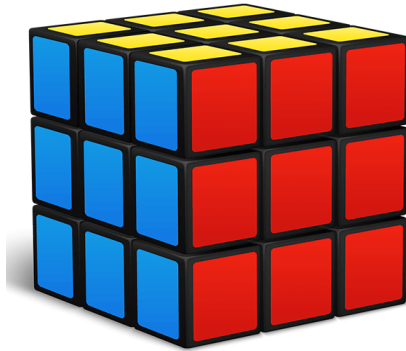


---

# 의생명 문헌 기반 약물 유사도 계산 방법 소개 및 실습

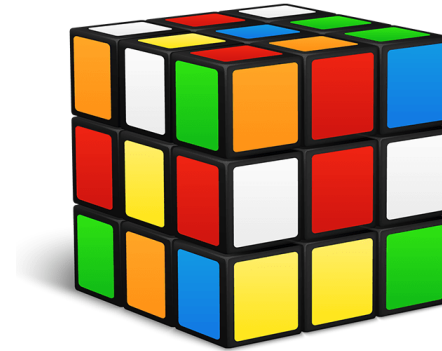
서울대학교 의생명지식공학연구실  
심용선  
yongsun0926@snu.ac.kr

## Structured Data



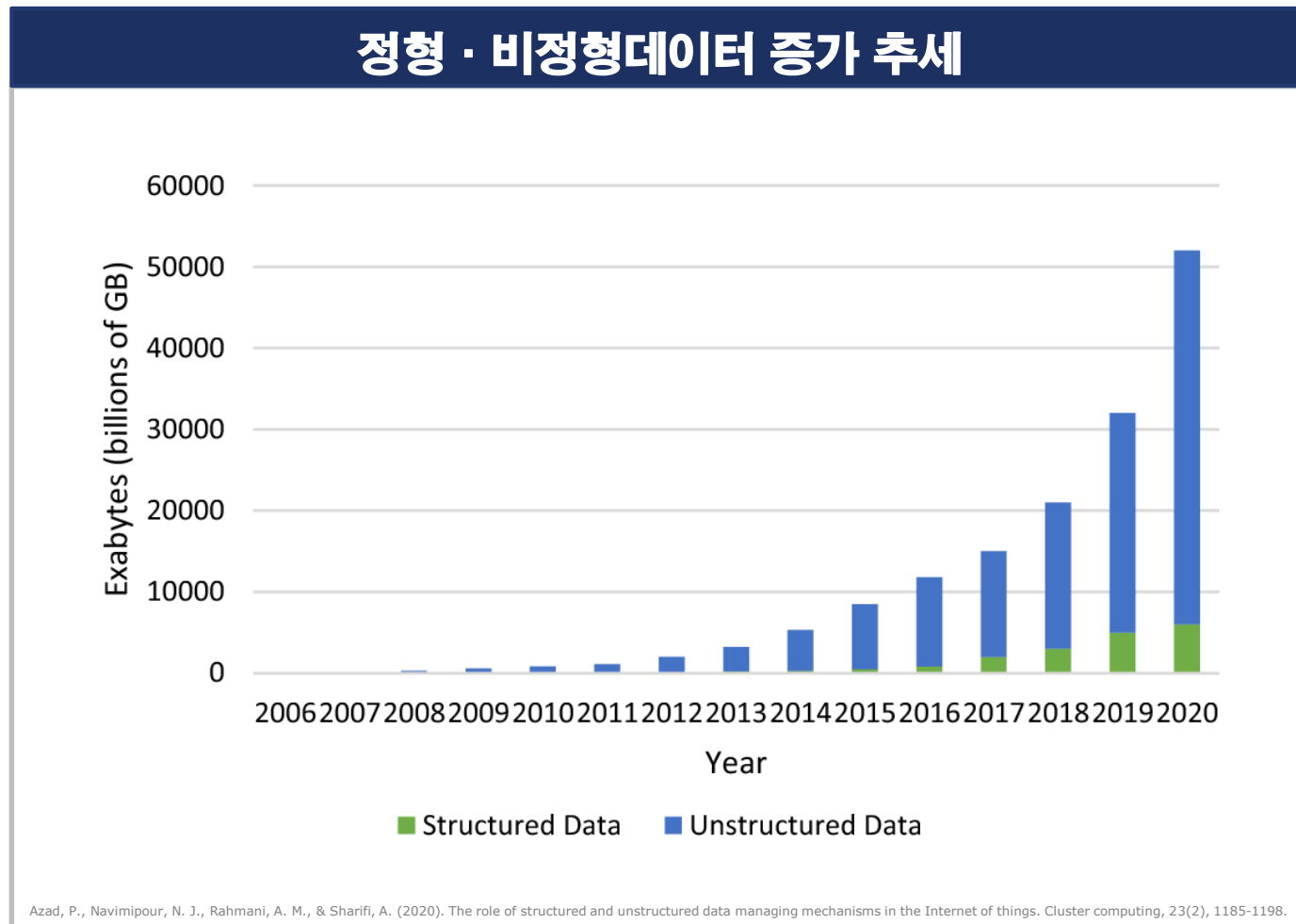
- 고정된 필드에 저장된 데이터
- 관계형 데이터 베이스, 스프레드시트 등
- 구조 변경에 있어 제한됨
- 별도의 분석 처리 기술 없이 간단한 쿼리를 통하여 원하는 데이터 추출 가능

## Unstructured Data



- 고정된 필드에 저장되어 있지 않은 데이터
- 텍스트 문서, 이미지/동영상/음성 데이터 등
- 구조 변경에 있어 자유로움
- 다양하고 방대한 양의 데이터를 처리할 수 있는 별도의 분석 처리 기술 필요

# Structured/Unstructured Data Growth Rate Comparison



# Text Data Collection

---

## Download

### Get PubMed data via FTP

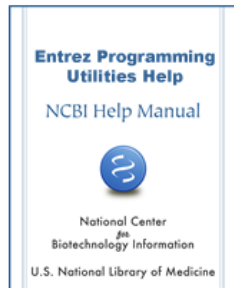
Note: Binary mode must be used when downloading data from our FTP servers.

### Annual baseline

Once a year, NLM produces a baseline set of PubMed citation records in XML format for download; the baseline file is a complete snapshot of PubMed data. When using this data in a local database, the best practice is to overwrite your local data each year with the baseline data.

---

## Use API



### Entrez Programming Utilities Help

< Prev   Next >

Bethesda (MD): [National Center for Biotechnology Information \(US\)](#); 2010-.

[Copyright and Permissions](#)

Search this book

### Introduction to the E-utilities

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&term=Bivalirudin+OR+Argatroban>

# Text Data Collection

← → ↺ 🔒 eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Bivalirudin+OR+Argatroban

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<eSearchResult>
  <script/>
  <Count>3322</Count>
  <RetMax>20</RetMax>
  <RetStart>0</RetStart>
  <IdList>
    <Id>36583436</Id>
    <Id>36571281</Id>
    <Id>36569714</Id>
    <Id>36567588</Id>
    <Id>36556985</Id>
    <Id>36537252</Id>
    <Id>36530956</Id>
    <Id>36520539</Id>
    <Id>36490373</Id>
    <Id>36480797</Id>
    <Id>36478775</Id>
    <Id>36478196</Id>
    <Id>36477689</Id>
    <Id>36477613</Id>
    <Id>36470755</Id>
    <Id>36449392</Id>
    <Id>36447732</Id>
    <Id>36415767</Id>
    <Id>36398035</Id>
    <Id>36386368</Id>
  </IdList>
  <TranslationSet>
    <Translation>
      <From>Bivalirudin</From>
      <To>"bivalirudin"[Supplementary Concept] OR "bivalirudin"[All Fields] OR "bivalirudin's"[All Fields] OR "bivalirudine"[All Fields]</To>
    </Translation>
    <Translation>
      <From>Argatroban</From>
      <To>"argatroban"[Supplementary Concept] OR "argatroban"[All Fields] OR "argatroban's"[All Fields]</To>
    </Translation>
  </TranslationSet>
  <QueryTranslation>"bivalirudin"[Supplementary Concept] OR "bivalirudin"[All Fields] OR "bivalirudin s"[All Fields] OR "bivalirudine"[All Fields]
</eSearchResult>
```

# Text Data Collection

← → ↺ 🔒 eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=36583436&retmode=xml

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<PubmedArticleSet>
  <script/>
  <PubmedArticle>
    <MedlineCitation Status="Publisher" Owner="NLM">
      <PMID Version="1">36583436</PMID>
      <DateRevised>
        <Year>2023</Year>
        <Month>01</Month>
        <Day>03</Day>
      </DateRevised>
      <Article PubModel="Print-Electronic">
        <Journal>
          <ISSN IssnType="Electronic">2047-9980</ISSN>
          <JournalIssue CitedMedium="Internet">
            <Volume>12</Volume>
            <Issue>1</Issue>
            <PubDate>
              <Year>2023</Year>
              <Month>Jan</Month>
              <Day>03</Day>
            </PubDate>
          </JournalIssue>
          <Title>Journal of the American Heart Association</Title>
          <ISOAbbreviation>J Am Heart Assoc</ISOAbbreviation>
        </Journal>
        <ArticleTitle>Femoral Vascular Closure Devices and Bleeding, Hemostasis, and Ambulation Following Percutaneous Coronary Intervention.</ArticleTitle>
        <Pagination>
          <StartPage>e025666</StartPage>
          <MedlinePgn>e025666</MedlinePgn>
        </Pagination>
        <ELocationID EIdType="doi" ValidYN="Y">10.1161/JAHA.122.025666</ELocationID>
        <Abstract>
          <AbstractText>
            Background The effectiveness of vascular closure devices (VCDs) to reduce bleeding after transfemoral percutaneous coronary intervention (TF-PCI) versus bivalirudin on outcomes after percutaneous coronary intervention (PCI) trial who underwent transfemoral percutaneous coronary intervention (TF-PCI) was type 2, 3, or 5 Bleeding Academic Research Consortium access site bleeding at day 3. Univariate and multivariate analyses were compared between groups. Of the 1580 patients who underwent transfemoral percutaneous coronary intervention, 1004 (63.5%) underwent transfemoral PCI and 576 (36.5%) underwent manual compression group (inverse probability weighting-adjusted). There were statistically significant differences between VCD and manual compression groups in bleeding at day 3 (P<0.001), bleeding at day 7 (P<0.001), and bleeding at day 30 (P<0.001). There were no significant differences between VCD and manual compression groups in bleeding at day 90 (P=0.14), bleeding at day 180 (P=0.14), and bleeding at day 360 (P=0.14).
          </AbstractText>
        </Abstract>
      </Article>
    </MedlineCitation>
  </PubmedArticle>
</PubmedArticleSet>
```

# Text Preprocessing

---

## Tokenization

- 문서를 토큰 단위로 분리하는 기법
- 토큰의 단위가 상황에 따라 다르지만, 보통 의미 있는 단위로 토큰을 정의

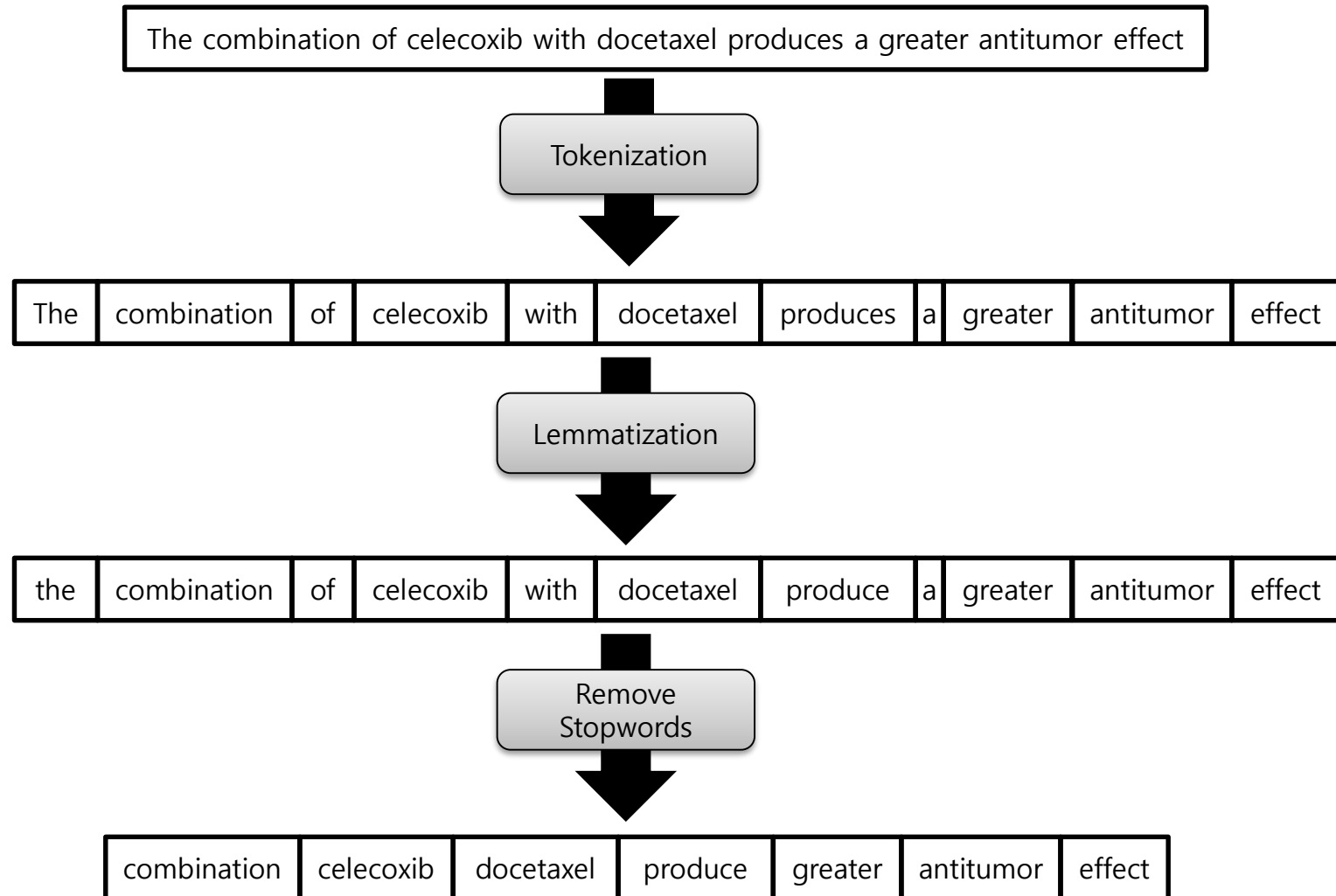
## Lemmatization

- 다양한 형태로 표현되어 있는 단어를 일반형태로 변형하는 기법
- 단어의 의미적인 단위를 고려하고, 형태소 분석을 통해 수행
- Lemmatization을 수행할 경우, 품사 정보가 남아있기 때문에 의미론적 관점에서 더 효과적

## Remove Stopwords

- 자주 등장하지만 분석에 있어 큰 의미가 없는 단어들을 제거하는 과정
- 예를 들면, I, a, the, 조사, 접미사 같은 단어들은 문장에서는 자주 등장하지만 실제 분석에 있어 의미가 거의 없음

# Text Preprocessing

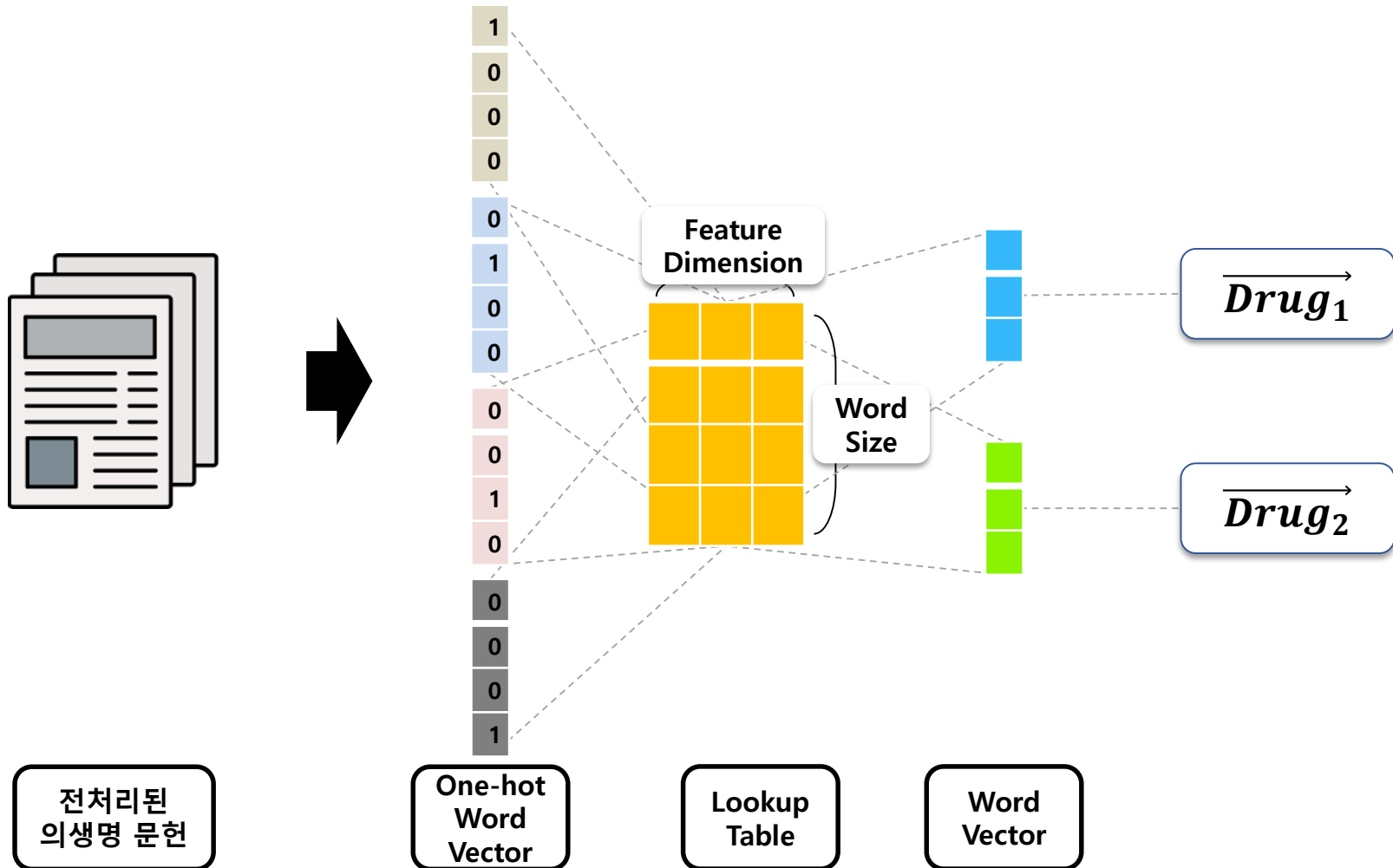




# Word2Vec 모델 기반 약물 단어 임베딩

## 의생명 문헌 기반 약물 벡터 추출

### Word2Vec 모델



## 약물 유사성 연구의 배경

- 약물 유사성은 서로 다른 약물 간의 유사성을 비교하는 연구
- 약물이 유사한 특성을 가질 수 있다는 가정에 기초하여 많은 연구에서 약물-약물 유사성을 활용하여 잠재적인 약물 관련 정보를 발견

## 약물 조합 연구의 방법

- 다양한 약물 관련 데이터를 기반으로 약물 유사성을 측정
- 서로 다른 유형의 약물 관련 정보를 통합하여 다양한 약물-약물 유사성 측정을 설계할 수 있으며 개발된 약물 유사성은 생물 의학 연구를 개선하는 데 추가로 사용

### 약물 관련 데이터 종류



Huang, Lan, et al. "Drug-drug similarity measure and its applications." Briefings in Bioinformatics 22.4 (2021): bbaa265.



Briefings in Bioinformatics, 22(4), 2021, 1–20

---

<https://doi.org/10.1093/bib/bbaa265>

Method Review

---

## Drug–drug similarity measure and its applications

Lan Huang, Huimin Luo, Suning Li, Fang-Xiang Wu<sup>id</sup> and Jianxin Wang<sup>id</sup>

Corresponding author: Jianxin Wang, School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China.  
Tel.: +86-731-88820212; Fax: +86-731-88877936; Email: jxwang@mail.csu.edu.cn

---

### Abstract

Drug similarities play an important role in modern biology and medicine, as they help scientists gain deep insights into drugs' therapeutic mechanisms and conduct wet labs that may significantly improve the efficiency of drug research and development. Nowadays, a number of drug-related databases have been constructed, with which many methods have been developed for computing similarities between drugs for studying associations between drugs, human diseases, proteins (drug targets) and more. In this review, firstly, we briefly introduce the publicly available drug-related databases. Secondly, based on different drug features, interaction relationships and multimodal data, we summarize similarity calculation methods in details. Then, we discuss the applications of drug similarities in various biological and medical areas. Finally, we

<http://vafaeelab.com/drugSimDB.html>

## 스코어 계산 방법

- Structure similarity
  - 화학구조 기반 약물 유사성 계산
- Target Similarity
  - 타겟 단백질 기반 약물 유사성 계산
- Pathway Similarity
  - 패스웨이 기반 약물 유사성 계산
- GO-CC/MF/BP Similarity
  - Gene Ontology Cellular Component
  - Gene Ontology Molecular Function
  - Gene Ontology Biological Process

GenericName	Actions	Target	Groups	Structure Similarity	Target Similarity	Pathway Similarity	GO_CC Similarity	GO_MF Similarity	GO_BP Similarity	Scores
Bivalirudin	View									
Argatroban	approved, investigational			0.04	1	NA	1	1	1	0.81

**A** DrugSimDB Home Search Download Statistics Help Contact

**B** Search a drug name and click View:

GenericName: Bivalirudin Actions: View

Multi-drug query

Upload a list of drugs (CSV):

Browse... No file selected Process

Example #1 Example #2 Example #3

**C** Similarity Network Physicochemical Properties Structure Pharmacology

Network View

Tip: 1. Nodes are colored by their degree of similarity to the selected drug. 2. Nodes are colored by their degree of similarity to the selected drug. 3. Nodes are colored by their degree of similarity to the selected drug. 4. Nodes are colored by their degree of similarity to the selected drug.

Table View

Target	Groups	Structure Similarity	Target Similarity	Pathway Similarity	GO_CC Similarity	GO_MF Similarity	GO_BP Similarity	ADP P-value
Dabigatran etexilate	approved	0.05	1	1	1	1	1	0.02
Argatroban	approved, investigational	0.04	1	NA	1	1	1	0.04
Ximelgepant	approved, investigational, withdrawn	0.04	1	NA	1	1	1	0.04
Prasugrel	approved	0.01	0.17	NA	1	1	1	0.25
Marfentanil	approved, nonrecreational	0	0.03	NA	0.73	0.04	0.04	0.78

**D** PubMed articles for drug pairs:

European Stroke Organization Guidelines on Reversal of Oral Anticoagulation: Hanne Christensen, Charlotte Carstensen, Jessica Kihlström, Lutz Christ Eur Stroke J 2020; doi: 10.1177/2396987319849763 PMID: 31703620

Use of Factor Xa Inhibitors, Common Concomitant Use for the Reversal of Dabigatran Zhai Shan Wang, Meredith Aarnman, Adil Hanna J Intensive Care Med 2019; doi: 10.1177/0885066619882909 PMID: 31766283

Reversal of oral anticoagulation in patients with acute intracerebral haemorrhage

**E** Similarity Network Physicochemical Properties Structure Pharmacology

Drug Name: Bivalirudin

Formula: C15H27N5O10

Molecular weight: NA

LogP: -0.76 (ALOGPS); -14 (ChemSaver)

Hydrogen bond acceptors: 37

**F** Similarity Network Physicochemical Properties Structure Pharmacology

Drug Name: Bivalirudin

By Network Physicochemical Properties Structure Pharmacology

**G** Similarity Network Physicochemical Properties Structure Pharmacology

Drug Name: Bivalirudin

Description: Bivalirudin is a synthetic 20 residue peptide (thrombin inhibitor) which reversibly inhibits thrombin. Once bound to the active site, thrombin cannot activate fibrinogen into fibrin, the crucial step in the formation of thrombus. It is administered intravenously. Because it can cause blood stagnation, it is important to monitor changes in hemostasis, activated partial thromboplastin time, international normalized ratio and blood pressure.

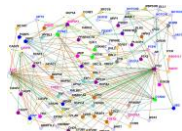
Indication: For treatment of heparin-induced thrombocytopenia and for the prevention of thrombolysis. Bivalirudin is indicated for use in patients undergoing percutaneous coronary intervention (PCI), in patients at moderate to high risk acute coronary syndromes due to unstable angina or non-ST segment elevation in whom a PCI is planned.

Mechanism of action: Inhibits the action of the enzyme by binding both to its catalytic site and to its anion-binding exosite. Thrombin is a serine protease that plays a central role in the thrombotic process, acting to cleave fibrinogen into fibrin monomers and to activate Factor XIII to Factor XIIIa, allowing fibrin to develop a covalently cross-linked framework which stabilizes the thrombus. Thrombin also activates factors V and VIII, promoting further thrombin generation, and activates platelets.

**[https://colab.research.google.com/drive/1xwNZgg78ACBilcYdbykNmIpl6yPlyQdW?usp=share\\_link](https://colab.research.google.com/drive/1xwNZgg78ACBilcYdbykNmIpl6yPlyQdW?usp=share_link)**

# In-silico 기반 의생명 연구 공통 프로세스

## Problem



## Process

### 1 Data Collection

DRUGBANK



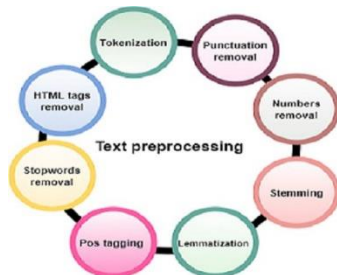
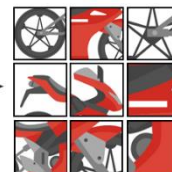
PubMed

### 2 Data Preprocessing

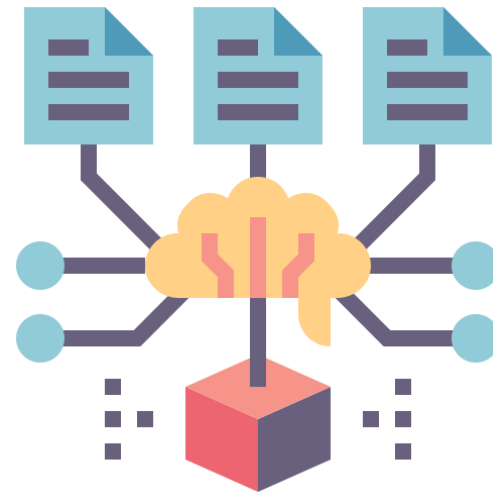


Feature extraction algorithm

Features



### 3 Applying algorithms



Problem resolve

## Output





*Thank*  
**YOU**