Biyomedikal Veri İçin Metin İşleme Tekniklerinin Kullanımı

Arzucan Özgür
Boğaziçi Üniversitesi
Bilgisayar Mühendisliği Bölümü

arzucan.ozgur@boun.edu.tr

Kuzeybatıda Yapay Öğrenme Yaz Okulu
26.06.2019

Biyomedikal Alanda Metin Tabanlı Veri

- Doğal Dildeki Metinler → Biyomedikal Metin Madenciliği

Chitosan Oligosaccharide Exerts Anti-Allergic Effect against Shrimp Tropomyosin-Induced Food Allergy by Affecting Th1 and Th2 Cytokines.

Jiang T[1,2], Ji H[3], Zhang L[4], Wang Y[5], Zhou H[6].

Author information

Abstract

BACKGROUND: Shrimp-derived allergen has a serious impact on people's health. Chitosan oligosaccharide (COS) has anti-allergic action but its function on shrimp allergen-induced allergy and related molecular mechanisms remain unclear.

METHODS: COS and its degrees of polymerization (DP) were selected to interact with shrimp tropomyosin (TM) and IgE was measured. A mouse model of food allergy was established by receiving shrimp TM intraperitoneally. The models were treated with different concentrations of COS. Fecal and serum histamine, serum IgE, IgG1 and IgG2a, and inflammatory cytokines were measured.

RESULTS: The main products for COS were DP2-6 with the contents of 6, 40, 26, 16, and 4%, respectively, and reacted with shrimp TM increasingly when COS DP was increased. Severe symptoms of food allergy were observed in the TM group (diarrhea, anaphylactic response, and rectal temperature). In contrast, COS treatment improved these symptoms significantly ($p < 0.05$). The sensitized mice were desensitized after they were treated with 1 mg/kg COS. COS treatment significantly reduced serum IgE and IgG1 levels, and increased IgG2a levels ($p < 0.05$). COS consumption decreased fecal and serum histamine. COS treatment reduced Th2 cytokine (IL-4, IL-5, and IL-13) levels and increased the Th1 cytokine (IFN-γ) level ($p < 0.05$).

CONCLUSIONS: COS showed anti-allergy properties by regulating the levels of Th1 and Th2 cytokines.

- DNA Dizisi → Genomik Veri İçin Metin Madenciliği

TTCAGGTGCATAAGACCTTGAC…

- Protein Dizisi ve İlaç Formülleri → İlaç Keşfi İçin Metin Madenciliği

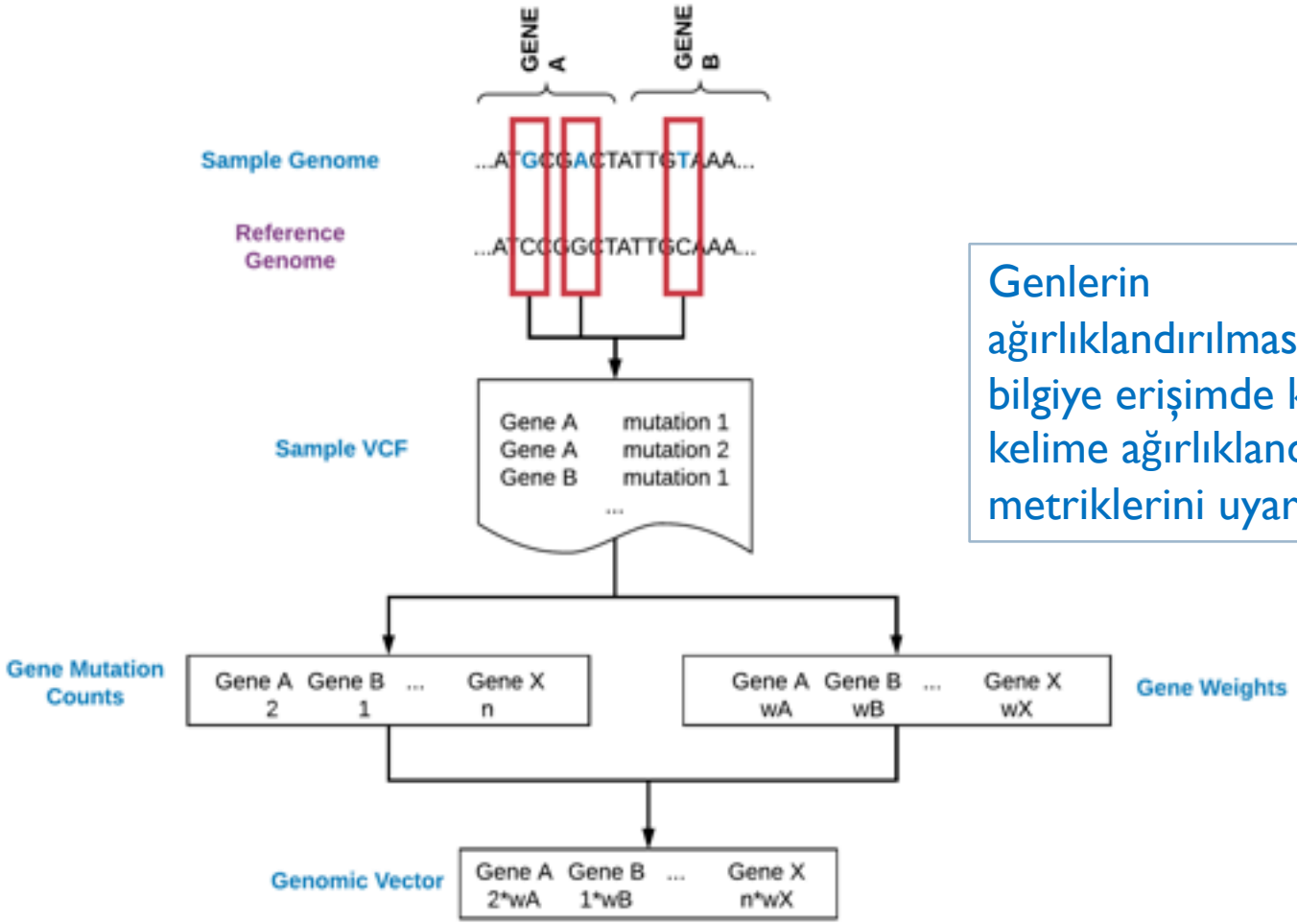
MELPNIMHPVAKLSTALAAALML…

CC1(C(N2C(S1)C(C2=O)NC(=O)…

Genomik Veri İçin Metin Madenciliği

İşbirliği: N. Özlem Özcan Şimşek ve Fikret Gürgen

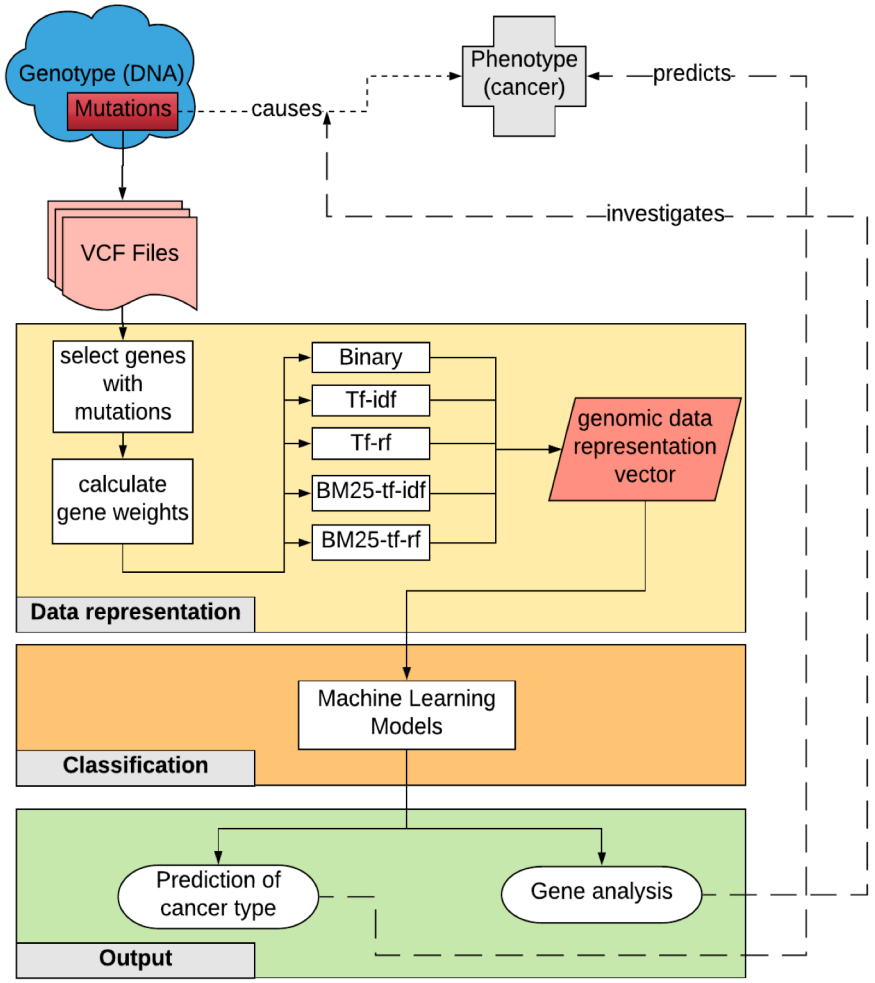
DNA'daki Mutasyonlara Dayanan Hastalık Tahmini



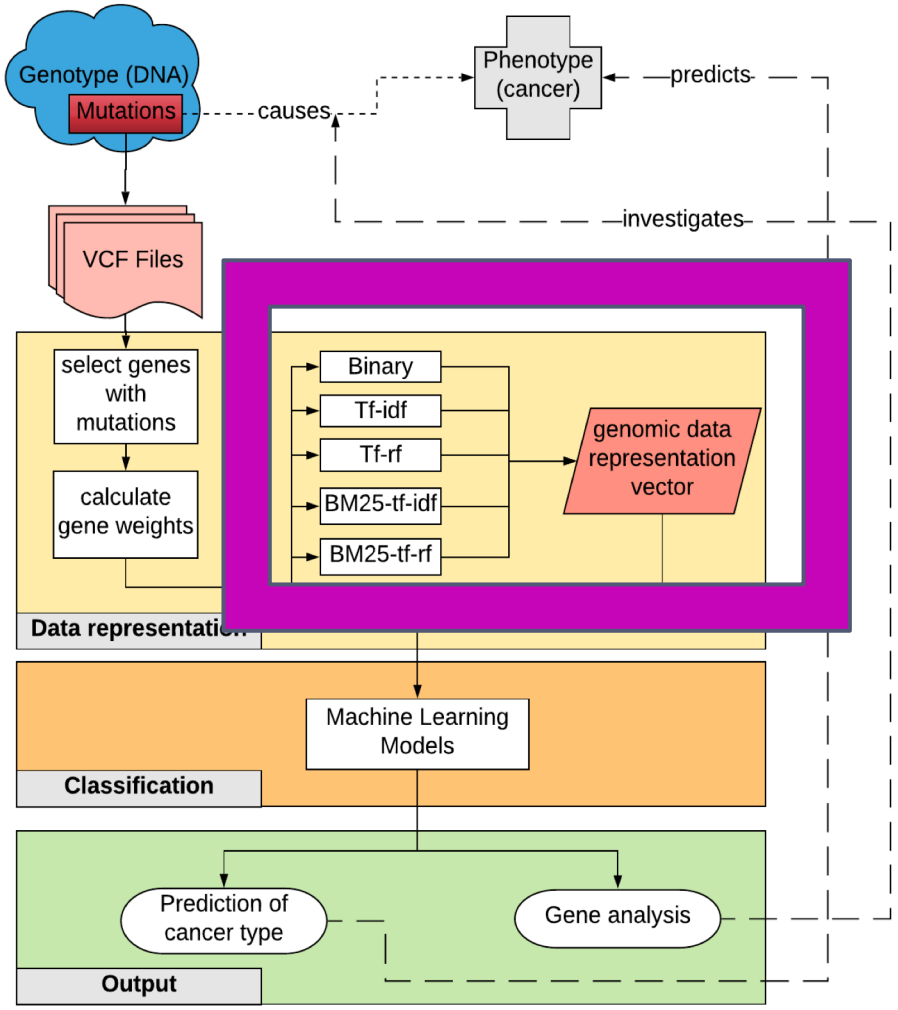
Genlerin ağırlıklandırılması için bilgiye erişimde kullanılan kelime ağırlıklandırma metriklerini uyarladık.

Özcan Şimşek, Ö. N., Özgür, A. & Gürgen, F. (2019). **Statistical representation models for mutation information within genomic data.** *BMC bioinformatics,* 20(1):324.

Sistemin Genel İşleyişi



Veri Gösterim Modelleri



İkili Gösterim

Bir gende mutasyon varsa 1 ile, yoksa 0 ile ifade ediliyor.

GEN-A (mutasyon var)	GEN-B (mutasyon yok)
1	0

Kelime Frekansı – Ters Doküman Frekansı (TF-IDF)

- Bir gende ne kadar çok mutasyon varsa, o gen o kadar önemlidir (tf)

- Nadir mutasyonlar daha ayırdedicidir (idf).

GEN-A (genel mut)	GEN-B (nadir mut)
0.0056	0.035

$$tf\text{-}idf_{g,s} = tf_{g,s} * idf_g$$

mutasyon sayısı

log(N/G) (N: hasta sayısı; G: g geninde mutasyon olan hasta sayısı

Kelime Frekansı – İlgililik Frekansı (TF-RF)

- Sınıf bilgisi kullanan denetimli bir yaklaşım.
- Bir gende belirli bir hastalık türünde diğer hastalık türlerine göre daha fazla mutasyon varsa, o gen daha ayırdedicidir.

	GEN-A	GEN-B
Hastalık-X	0.0056	0.035
Hastalık-Y	1.0056	0.005

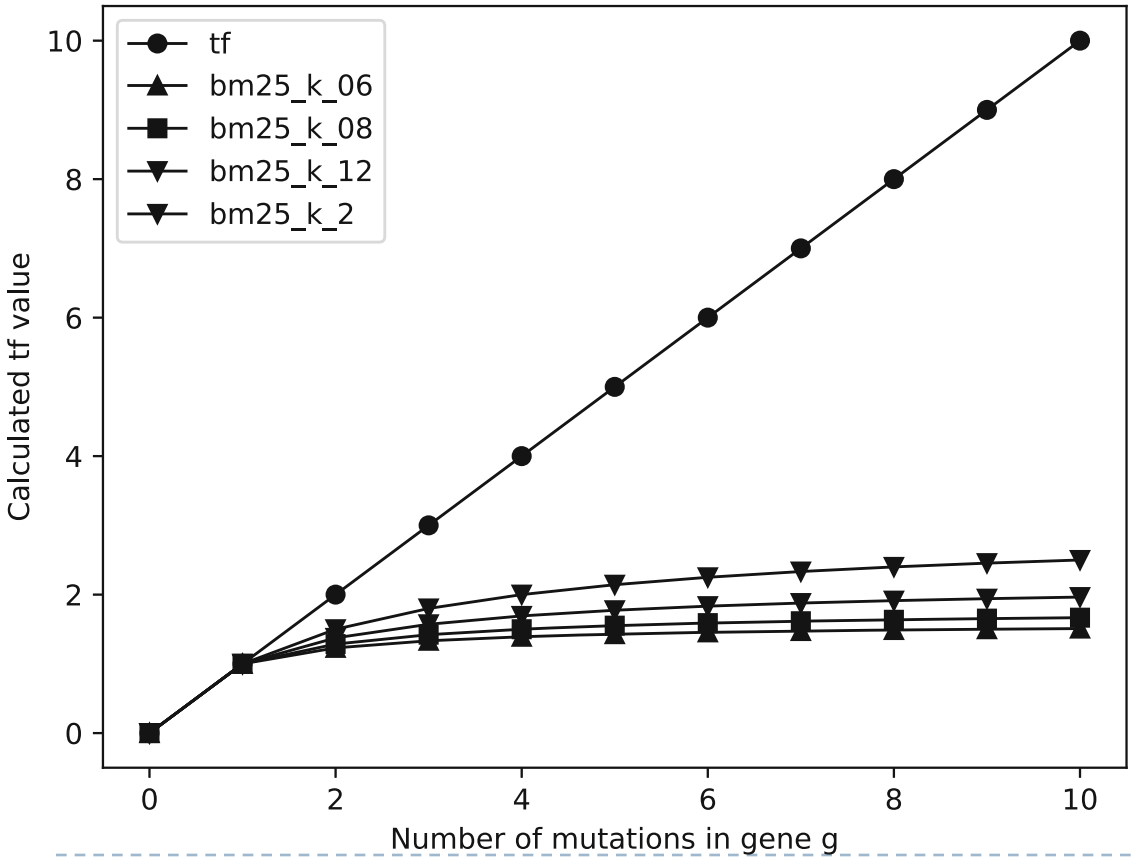
$$tf\text{-}rf_{g,s} = tf_{g,s} * rf_{g,c}$$

mutasyon sayısı

log(2+a/max(1,b))
a: c sınıfında g geninde mutasyon olan hasta sayısı
b: diğer sınıflarda g geninde mutasyon olan hasta sayısı

BM25

tf → **BM25-tf**



Yumuşatma (sınırlama) etkisi

BM25-TF-IDF & BM25-TF-RF

$$BM25\text{-}tf\text{-}idf_{g,s} = BM25\text{-}tf_{g,s} * idf_g$$

yumuşatılmış mutasyon sayısı

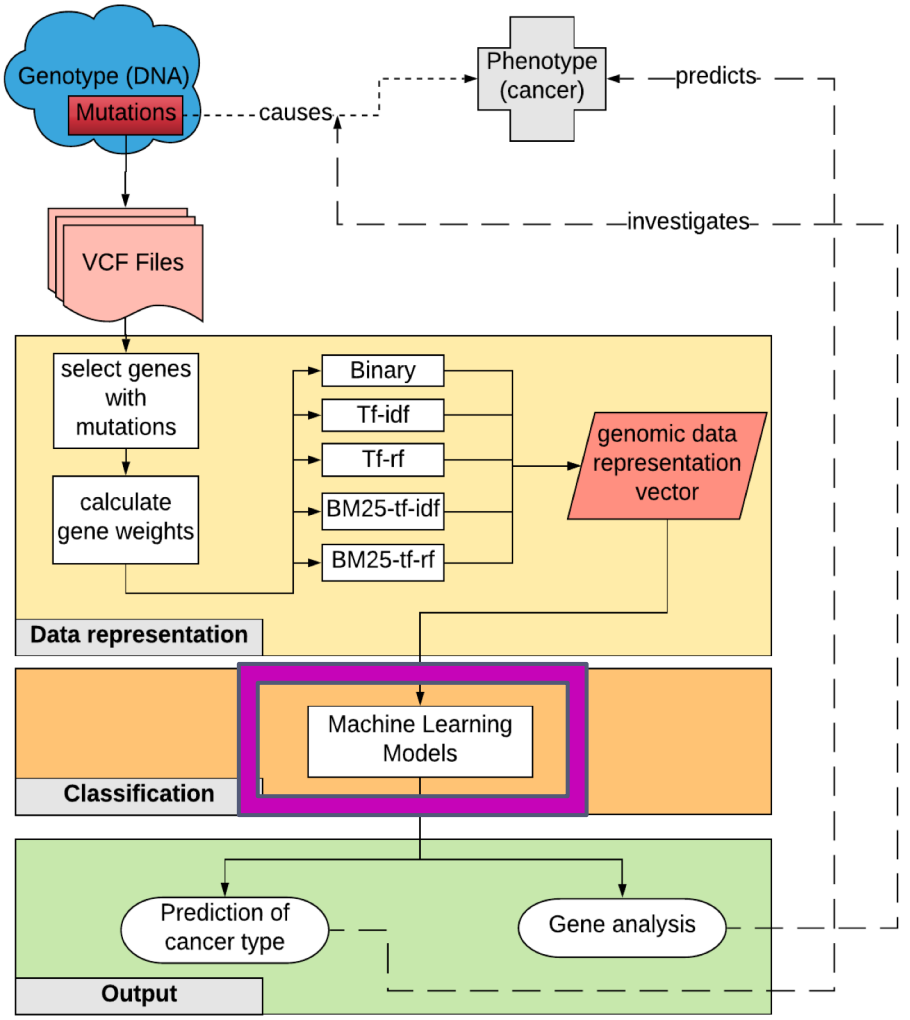
gen ağırlığı

$$BM25\text{-}tf\text{-}rf_{g,s} = BM25\text{-}tf_{g,s} * rf_{g,c}$$

yumuşatılmış mutasyon sayısı

denetimli gen ağırlığı

Sınıflandırma



Sınıflandırma

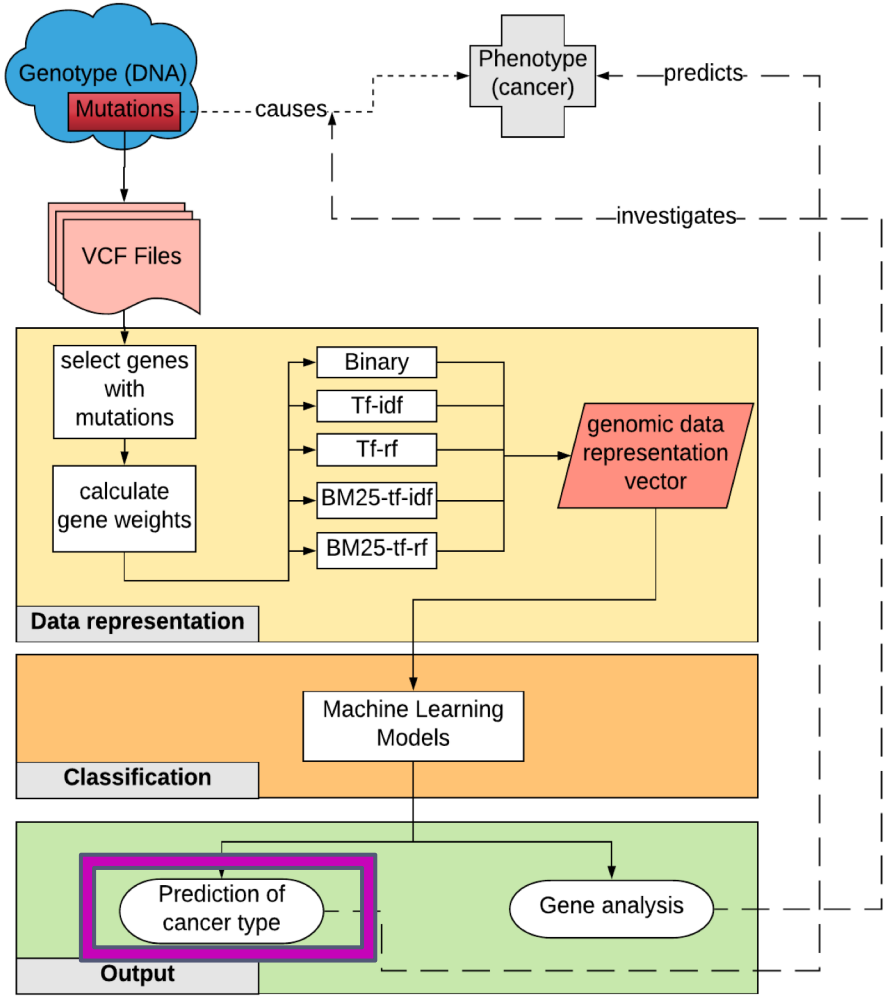
- Naive Bayes (NB)
- K En Yakın Komşu (KNN)
- Destek Vektör Makineleri (SVM)
- Yapısal Bağıntı (Logistic Regression - LR)
- Algılayıcı (Perceptron)
- Çok Katmanlı Yapay Sinir Ağları (NN)

Veri Kümesi

TCGA (The Cancer Gene Atlas)

Cancer type	Sample count
Lung	1232
Breast	1080
Brain	1028
Kidney	734
Colorectal	656
Thyroid	504
Prostate	503
Skin	472
Stomach	441
Liver	378

Kanser Türü Tahmini



Sonuçlar

Bilinen ve tüm genlerin karşılaştırılması

Gen Kümesi	Doğruluk	F-ölçütü	Yanlış Pozitif Oranı
Bilinen	36.74 ± 0.56	36.62 ± 0.83	10.01 ± 0.10
Tüm	68.50 ± 0.48	69.01 ± 0.01	4.07 ± 0.09

* LR sonuçları

Sonuçlar

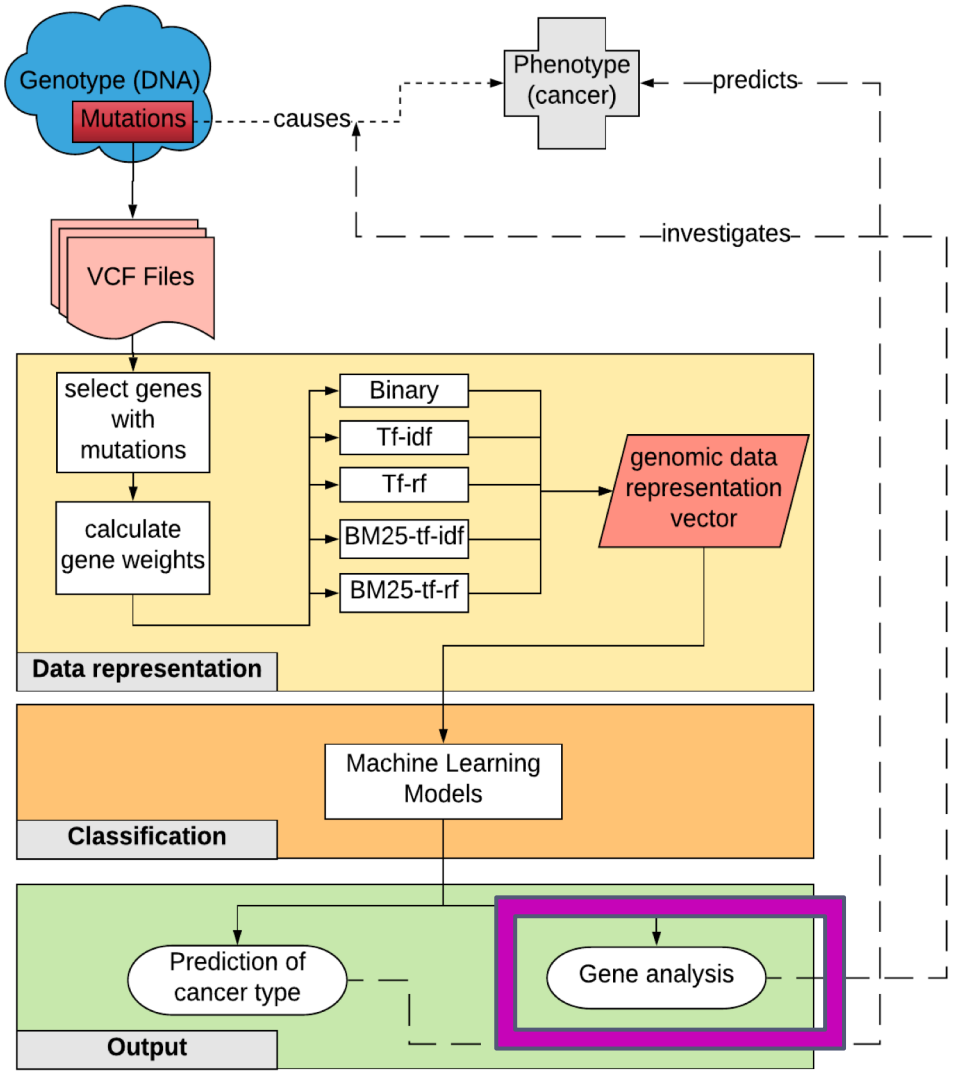
Veri gösterim modellerinin karşılaştırılması

Veri gösterimi	Doğruluk	F-ölçütü	Yanlış Pozitif Oranı
İkili	69.00 ± 0.76	69.52 ± 0.70	3.65 ± 0.17
Tf-idf	62.91 ± 0.79	63.32 ± 0.70	4.00 ± 0.10
Tf-rf	74.13 ± 1.33	74.17 ± 1.47	3.07 ± 0.24
BM25-tf-idf	68.18 ± 1.83	68.79 ± 1.28	4.07 ± 0.54
BM25-tf-rf	76.44 ± 0.66	76.95 ± 0.68	2.75 ± 0.13
C-score	73.74 ± 0.88	74.07 ± 0.73	3.27 ± 0.24

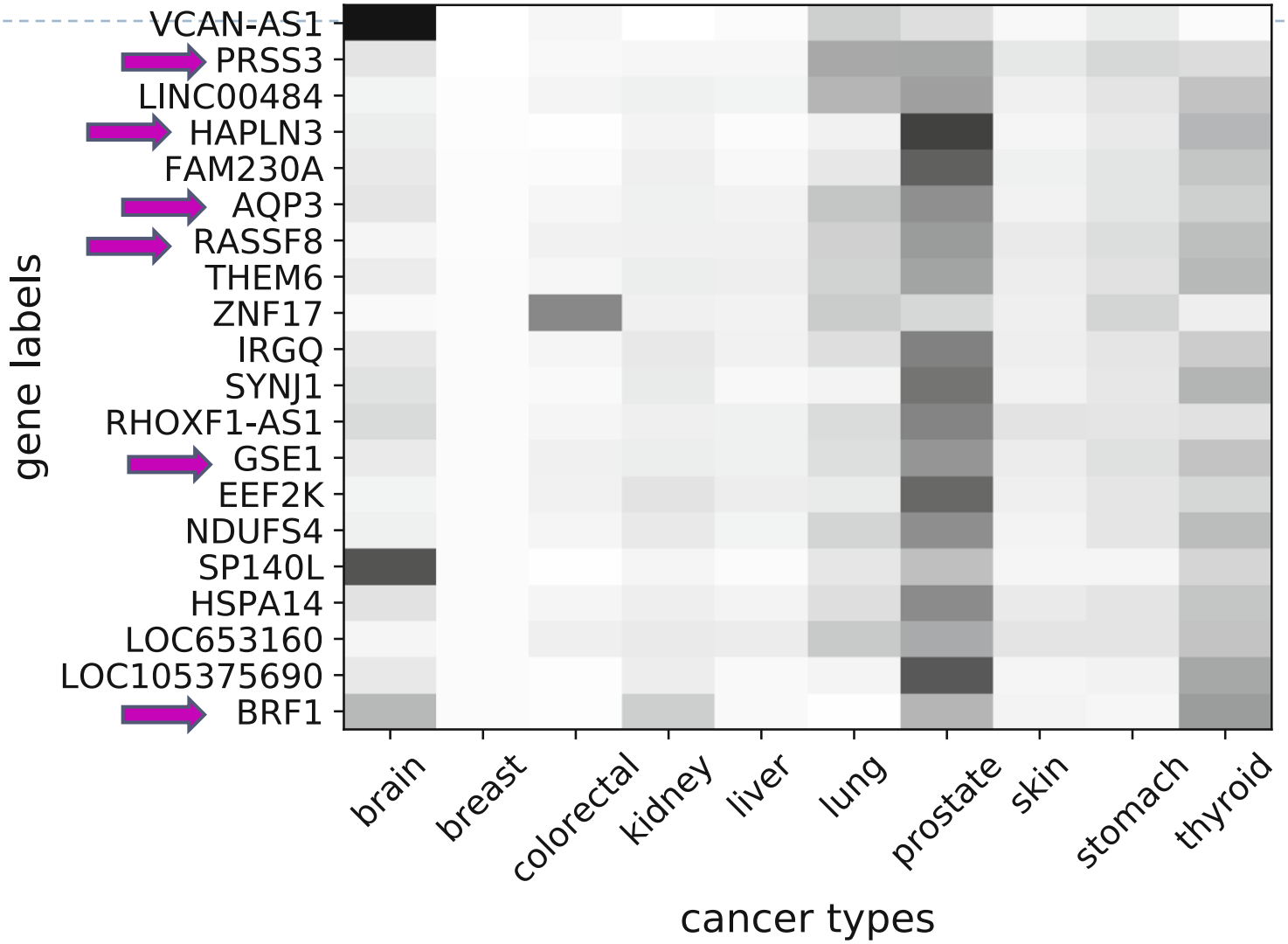
* NN sonuçları

Statistical significance 95% conf. interval p-value: 0.0001

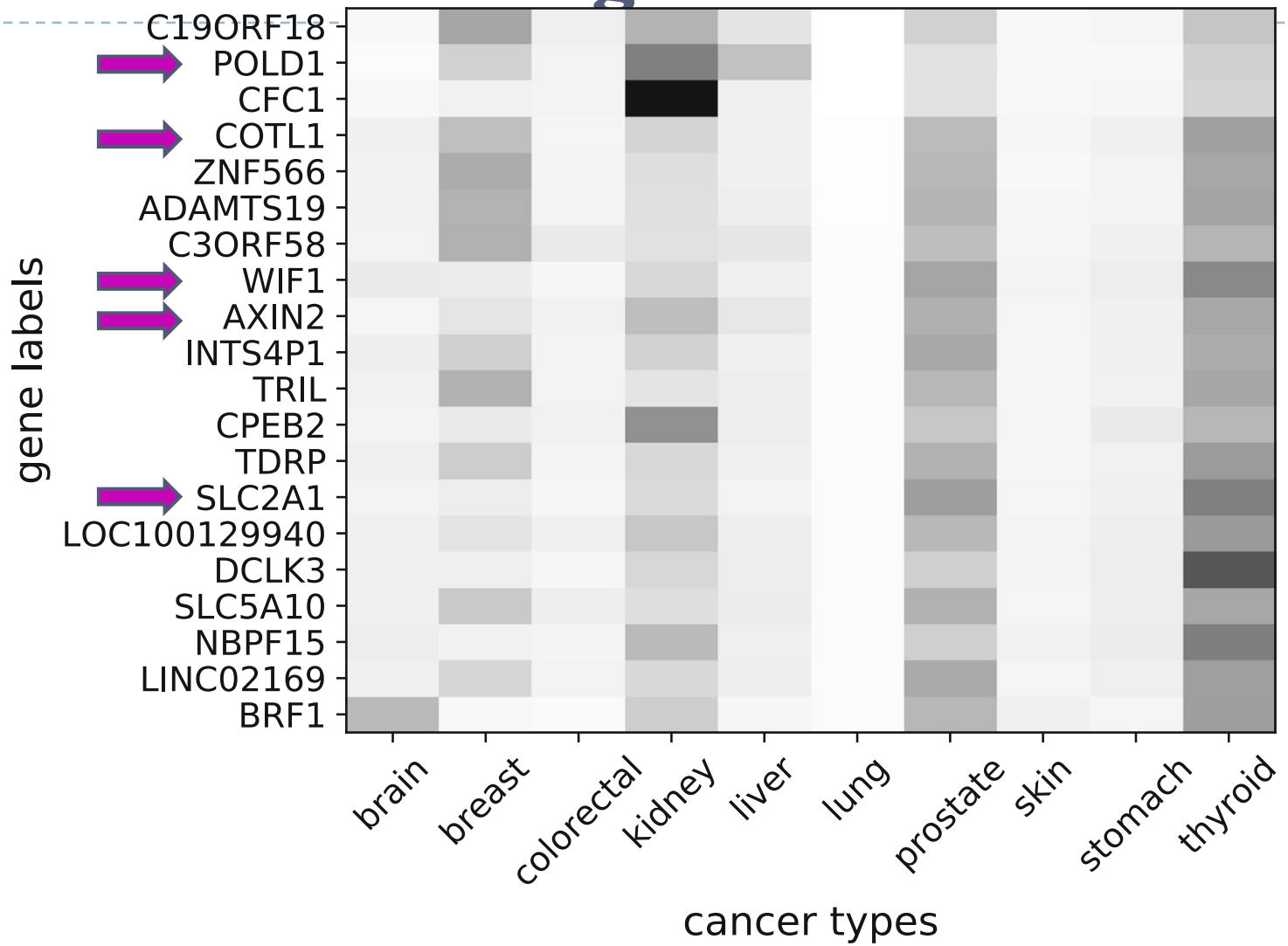
Gen Analizi


Genotype (DNA)
Mutations
causes
Phenotype (cancer)
predicts
investigates
VCF Files
select genes with mutations
calculate gene weights
Binary
Tf-idf
Tf-rf
BM25-tf-idf
BM25-tf-rf
genomic data representation vector
Data representation
Machine Learning Models
Classification
Prediction of cancer type
Gene analysis
Output

Gen Analizi – Meme kanseri



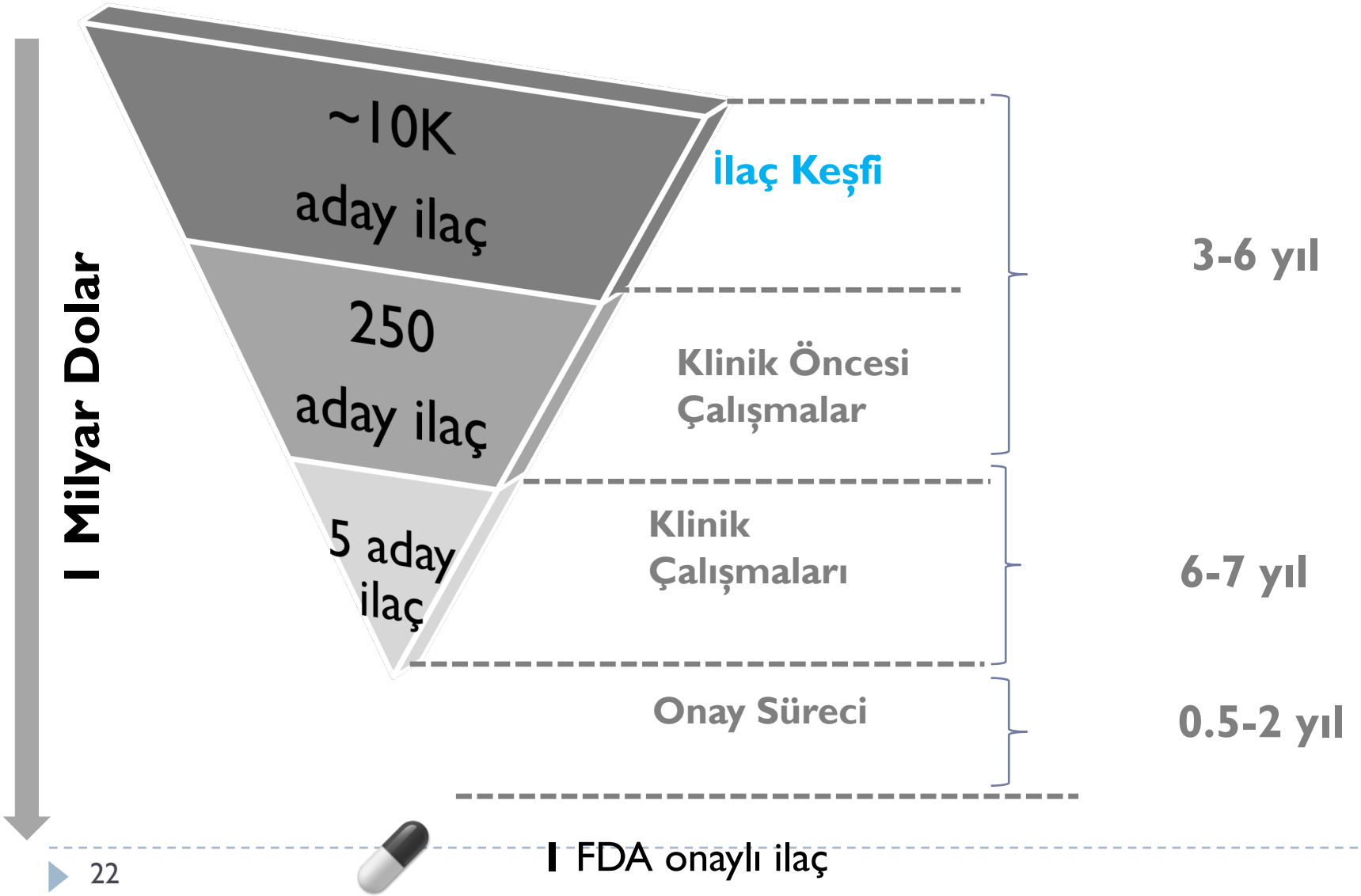
Gen Analizi – Akciğer Kanseri



İlaç Keşfi İçin Metin Madenciliği

İşbirliği: Hakime Öztürk ve Elif Özkırımlı

Motivasyon:
İlaç geliştirme uzun, pahalı ve zor bir süreç



http://www.ogjc.osaka-gu.ac.jp/php/nakagawa/TRIZ/eTRIZ/

Bu süreci nasıl kolaylaştırabiliriz

İlaç adaylarını önceliklendirerek arama uzayını daraltabiliriz.

~93M compounds > 8250 drugs

PubChem DrugBank

~550K proteins > ~10K targets

UniProt DrugBank

2

Mevcut ilaçlar için yeni hedefler belirleyebiliriz.

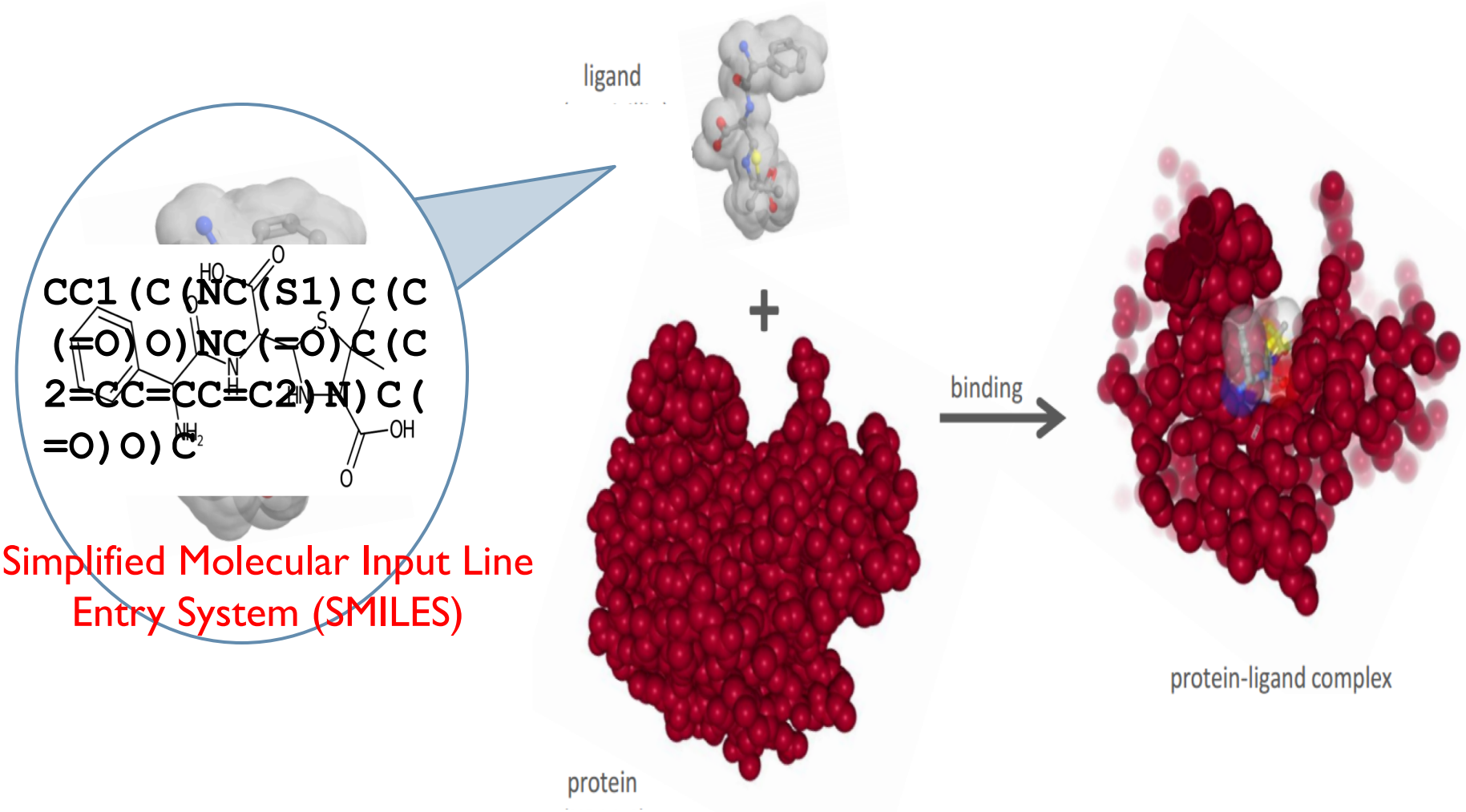
Bilinen ilaçlar

Yeni tedavi yöntemleri

BindingDB
ChEMBL

Yapay öğrenme teknikleri kullanarak

Protein ve ligandların metin tabanlı gösterimi

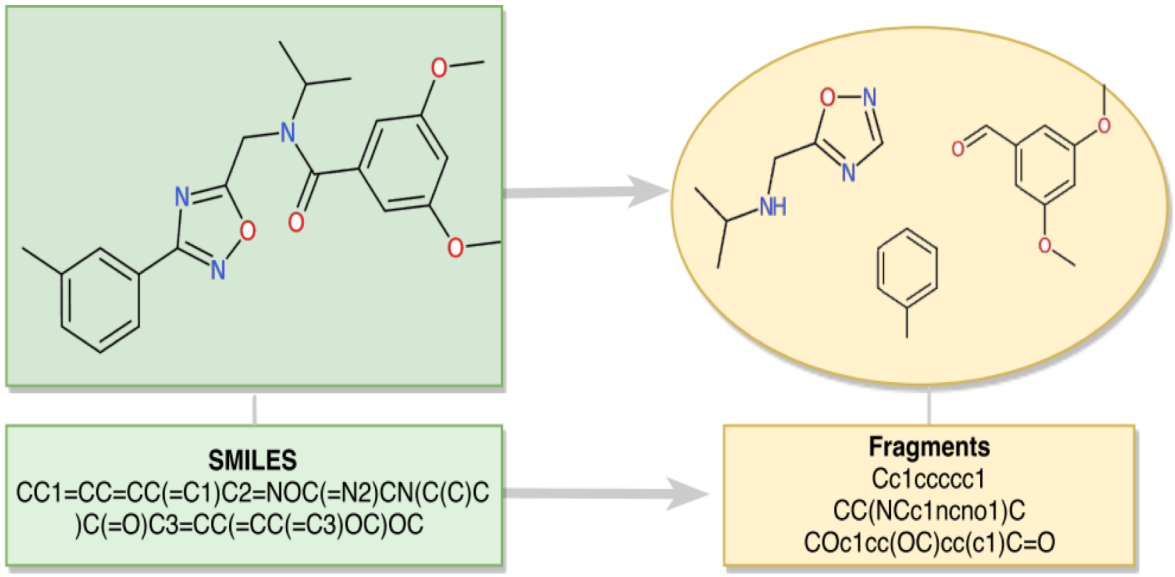


Simplified Molecular Input Line Entry System (SMILES)

SMILES metin tabanlı bir gösterim

- Metin işleme yöntemleri kullanabiliriz
- Ligand gösterimi için iki yaklaşım kullandık
  - TF-IDF tabanlı
  - Dağıtık gösterim tabanlı

Varsayım: Kelimelerden oluşan doküman



- Kelimelerin ne olduğunu bilmiyoruz

Doküman olarak **SMILES**

SMILES:

COC1=C(C=CC(=C1)C=O)O

Doküman olarak **SMILES**

SMILES:

8-karakterlik LINGO

COC1=C(C=CC(=C1)C=O)O

Kimyasal Kelimeler: COC1=C(C

Doküman olarak **SMILES**

SMILES:

8-karakterlik LINGO

COC1=C(C=CC(=C1)C=O)O

Kimyasal Kelimeler: COC1=C(C
OC1=C(C=

Doküman olarak **SMILES**

SMILES:

8-karakterlik LINGO

COC1=C(C=CC(=C1)C=O)O

Kimyasal Kelimeler: COC1=C(C
OC1=C(C=
C1=C(C=C

Doküman olarak **SMILES**

SMILES:

8-karakterlik LINGO

COC1=C(C=CC(=C1)C=O)O

Kimyasal Kelimeler: COC1=C(C
OC1=C(C=
C1=C(C=C
1=C(C=CC

Doküman olarak **SMILES**

SMILES:

8-karakterlik LINGO

COC1=C(C=CC(=C1)C=O)O

Kimyasal Kelimeler: COC1=C(C
OC1=C(C=
C1=C(C=C
1=C(C=CC
…
C1)C=O)O

Kelime Frekansı – Ters Doküman Frekansı **(TF-IDF)**

- Kimyasal kelimeleri ağırlıklandırmak için kullandık
  - TF: Kimyasal kelimenin SMILES'da geçme frekansı
  - IDF: Kimyasal kelimenin derlemde geçtiği SMILES formülü sayısı ile ters orantılı (varsayım: nadir geçen kelimeler daha ayırdedici)

Kelime frekansı (TF)

Kimyasal kelimenin bir SMILES formülünde geçme sıklığı:

Kimyasal kelimeler \ ligand (SMILES)	$SMILES_1$	$SMILES_2$	$SMILES_3$	$SMILES_4$
=CC=CC=C	3	6	4	2
C1)NC(=O	1	2	1	2
COC1=C(C	0	1	0	0
2=C(C(=N	0	1	1	0

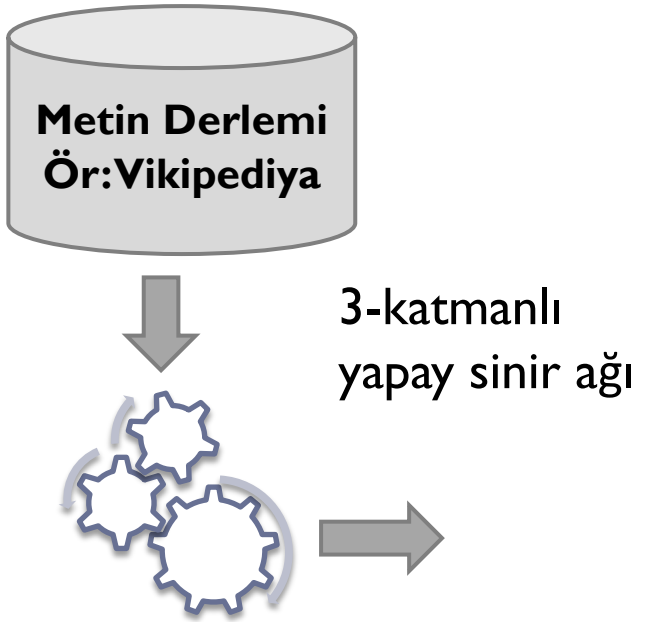
Ters Doküman Frekansı (IDF)

Chemical words \ ligand (SMILES)	$SMILES_1$	$SMILES_2$	$SMILES_3$	$SMILES_4$
=CC=CC=C	3	6	4	2
TÜM ligandlarda geçiyor.				
C1)NC(=O	1	2	1	2
COC1=C(C	0	1	0	0
Tek bir ligandda geçiyor				
2=C(C(=N	0	1	1	0

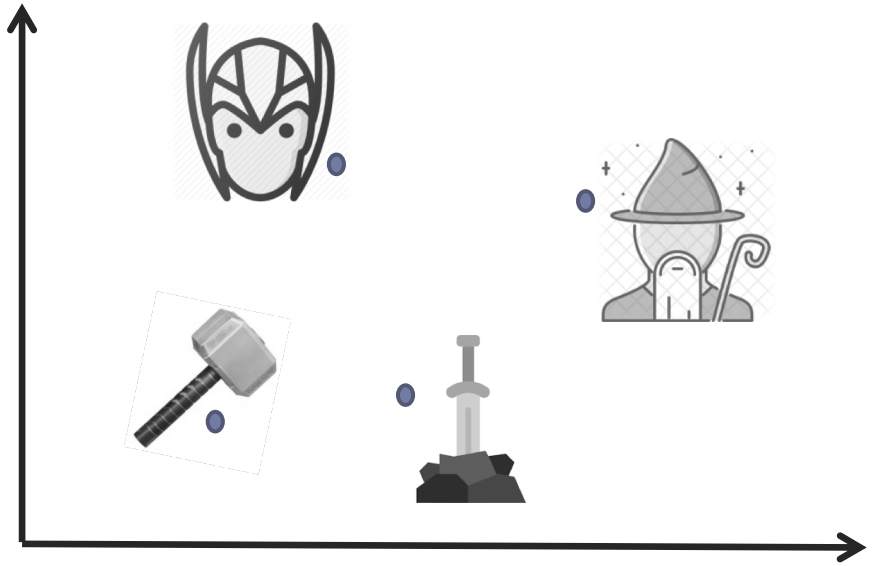
Ligandları TF-IDF tabanlı vektörler olarak ifade ettik.
TF-IDF vektörleri arasındaki kosinüs benzerliğini k en yakın komşu tabanlı algoritma ile kullandık.
İki boyutlu ligand gösterimi ile aynı başarıyı daha az hesaplama gücü ile elde ettik.

Öztürk, H., Özkırımlı, E., Özgür, A. **A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction**. *BMC Bioinformatics*, 17:128, 2016.

Dağıtık Kelime Temsil Modeli



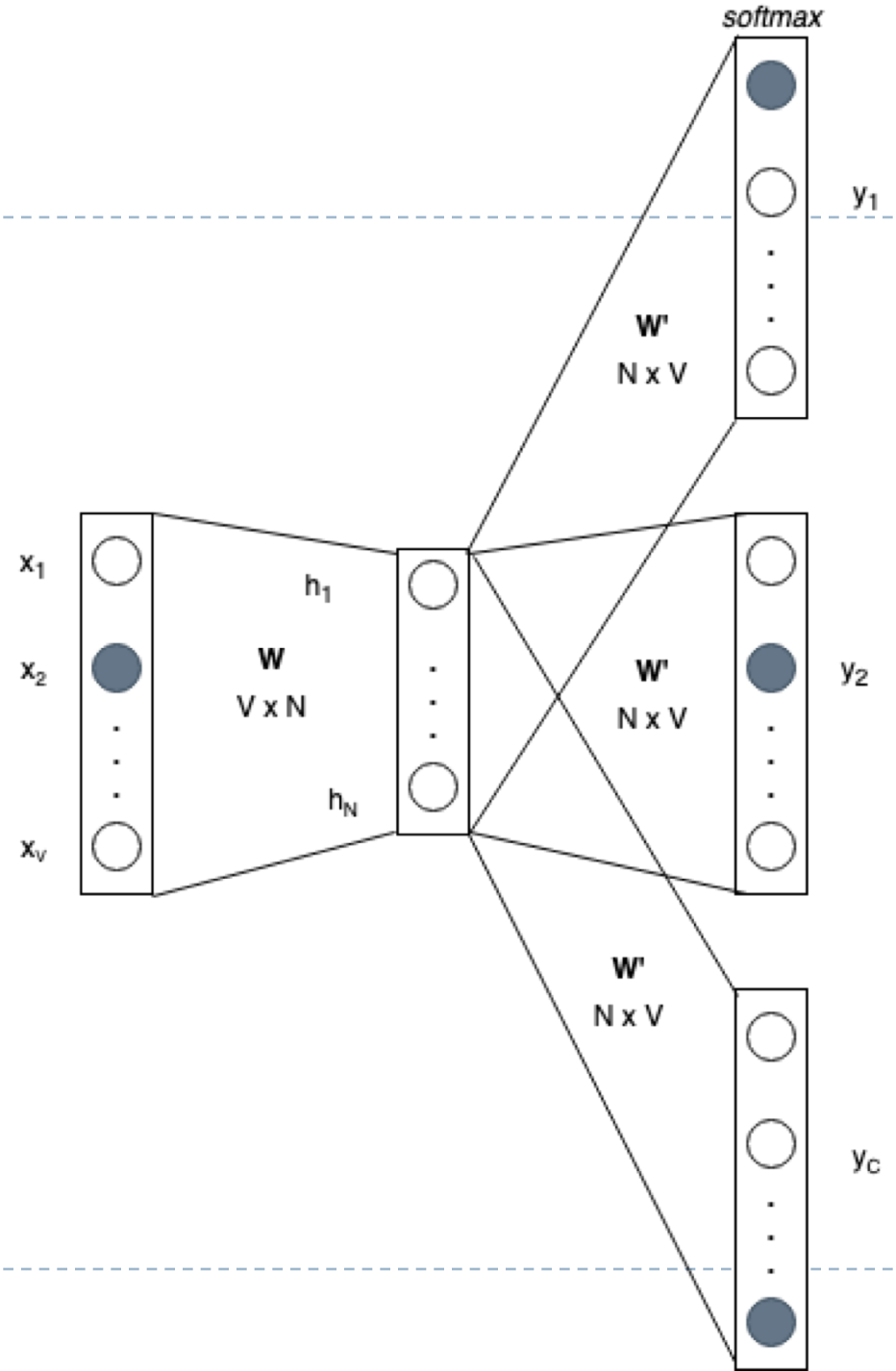
3-katmanlı
yapay sinir ağı



[Thor – Mjolnir] = [Merlin – Excalibur]

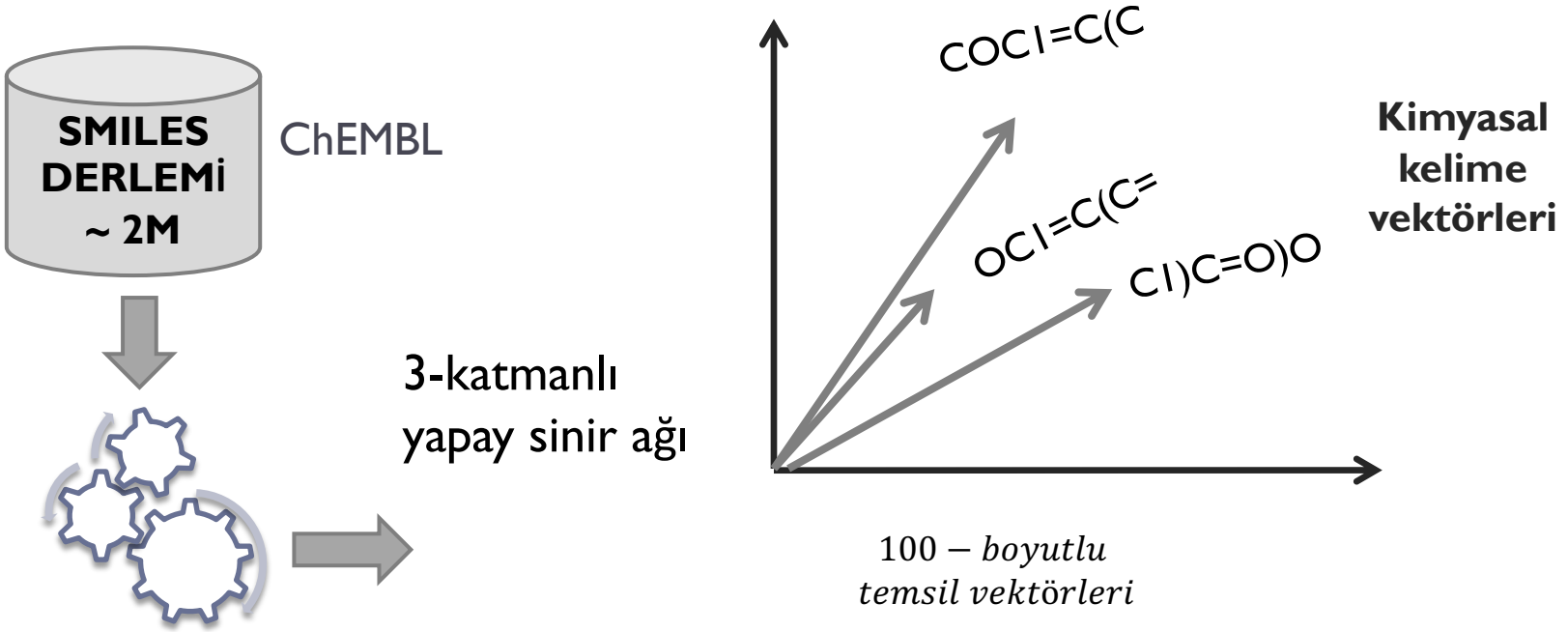
Word2Vec

(Mikolov, Tomas, et al. 2013)

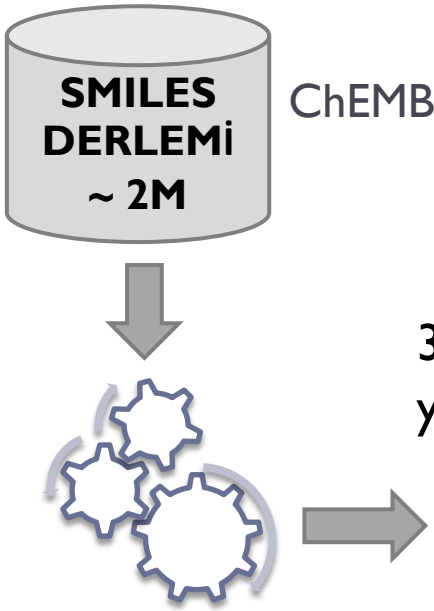


Skip-gram modeli

SMILESVec: Dağıtık ligand temsili

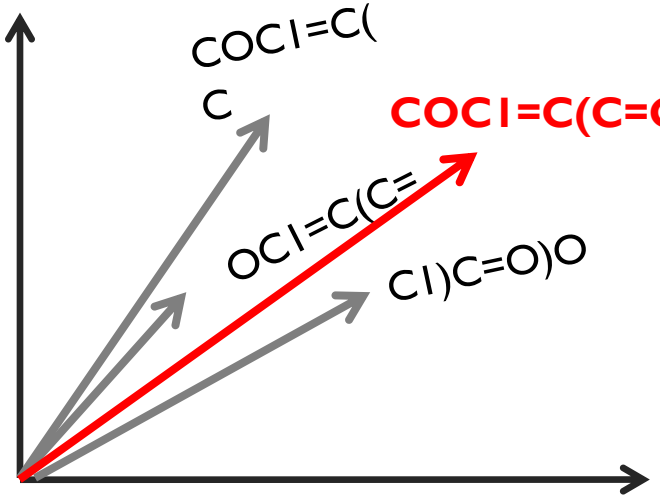


SMILESVec: Dağıtık ligand temsili



ChEMBL

3-katmanlı
yapay sinir ağı



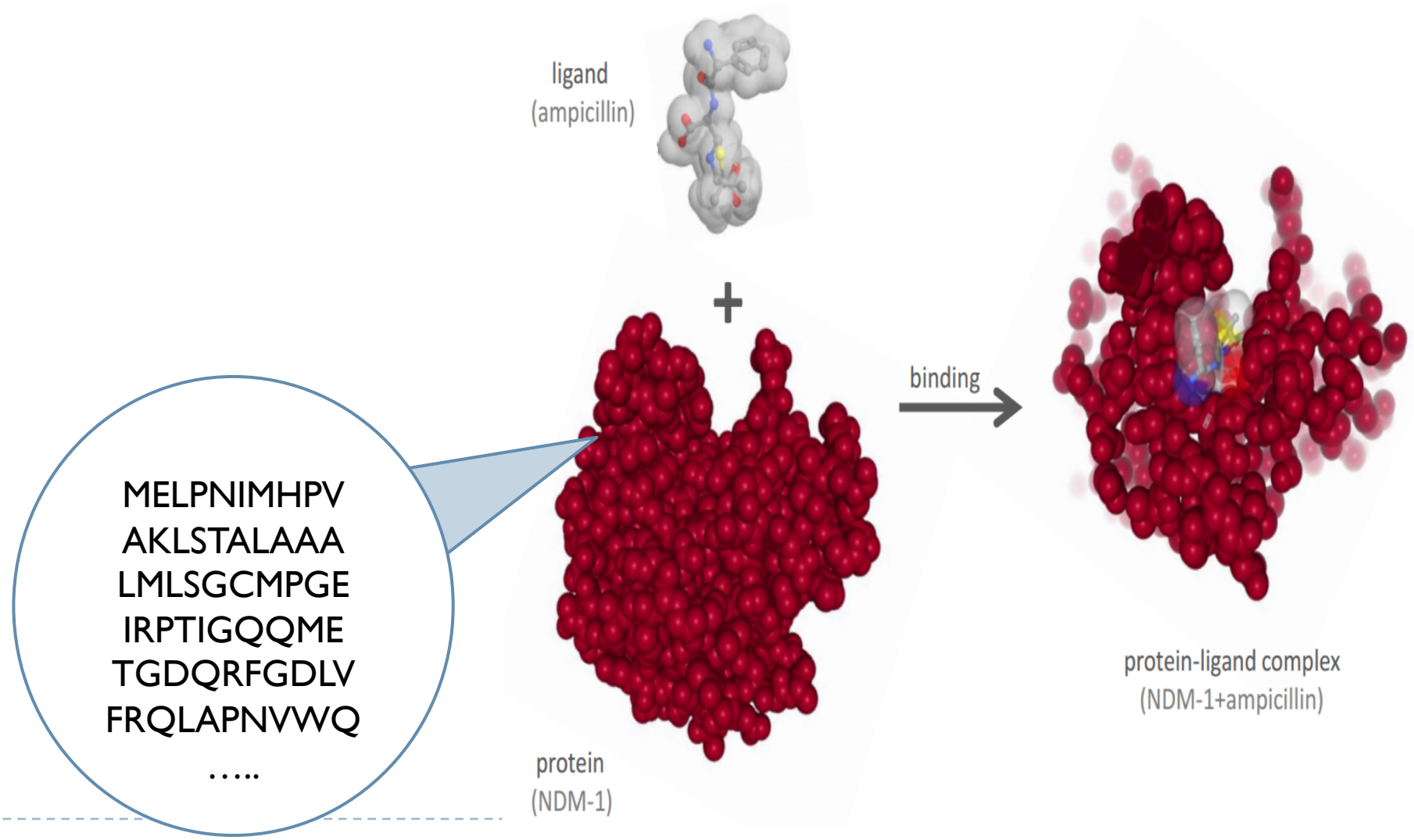
SMILES
Vektörü
(SMILESVec)

$100 - boyutlu\ temsil\ vektörü$

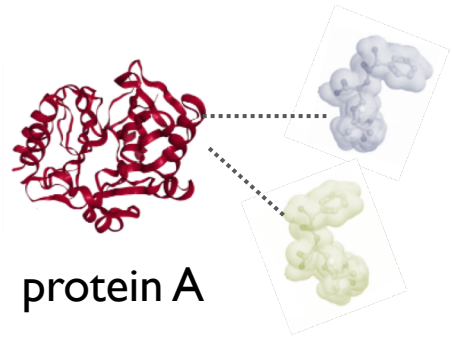
$$\text{SMILESVec} = vector(ligand) = \frac{\sum_{k=1}^{n} vector(word_k)}{n}$$

$n, kimyasal\ kelimelerin\ sayısı$

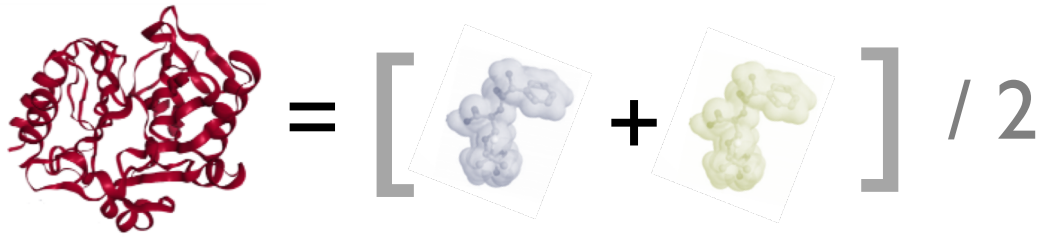
Protein temsili

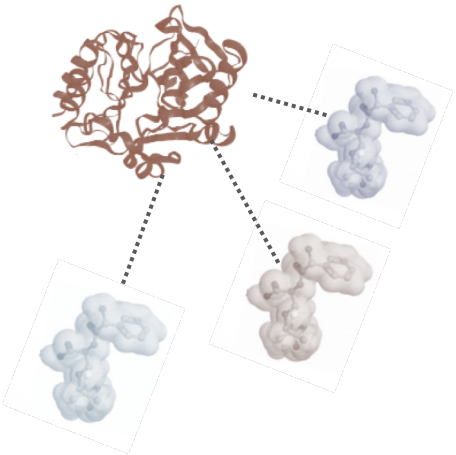


Öneri: Ligand tabanlı protein temsili

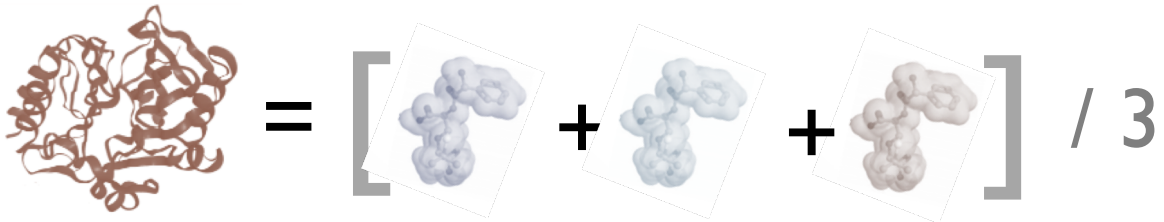


protein A





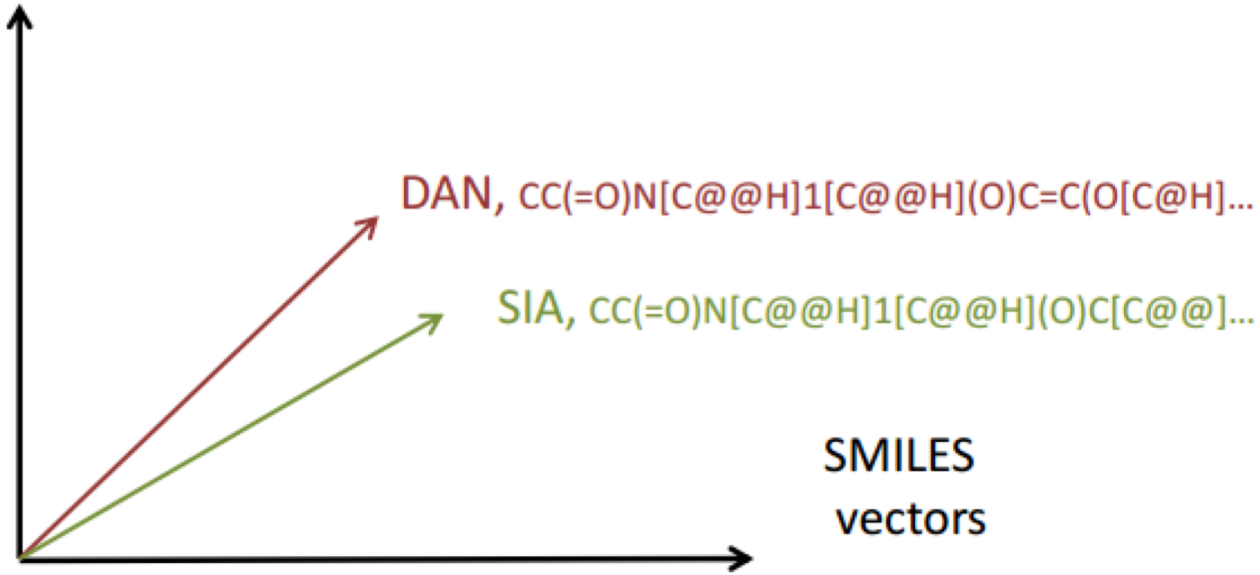
protein B



SMILESVec-tabanlı Protein Temsili

Protein: sialidase

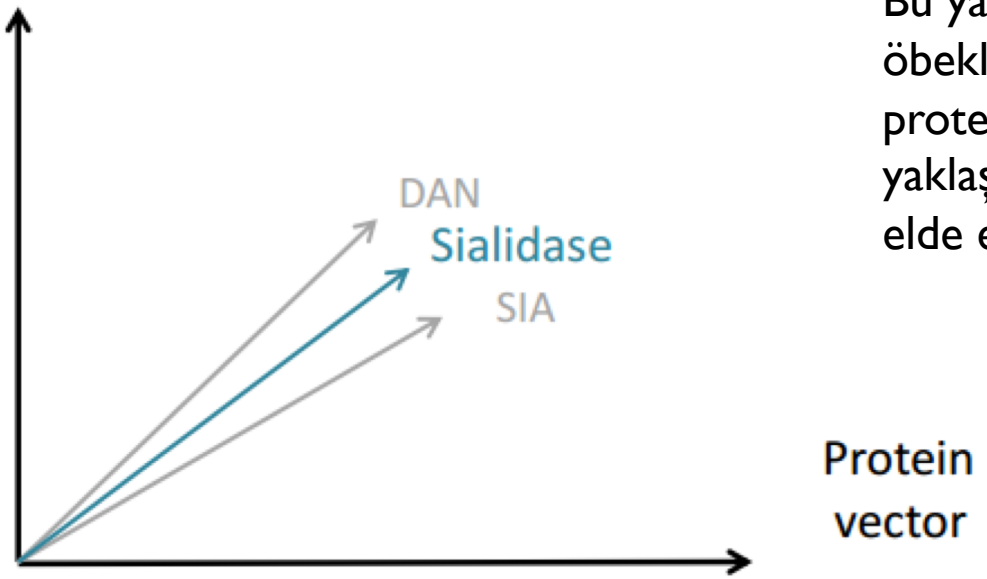
Interacting ligands: DAN, SIA



SMILESVec-based Protein Representation

Protein: sialidase

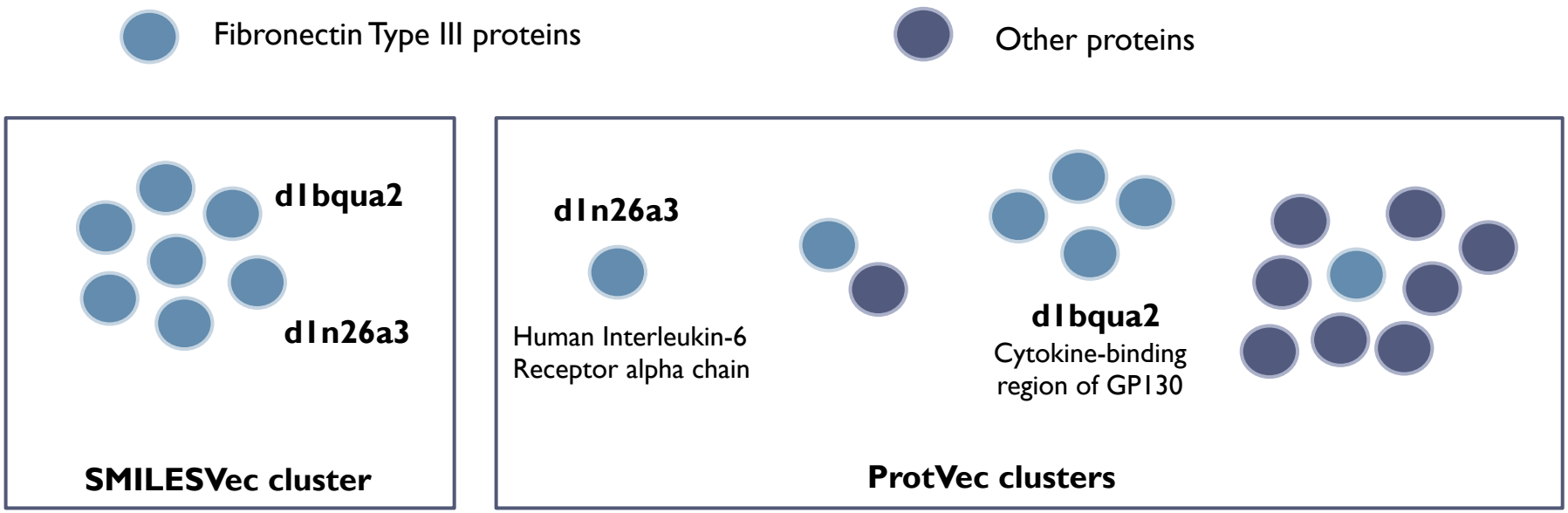
Interacting ligands: DAN, SIA



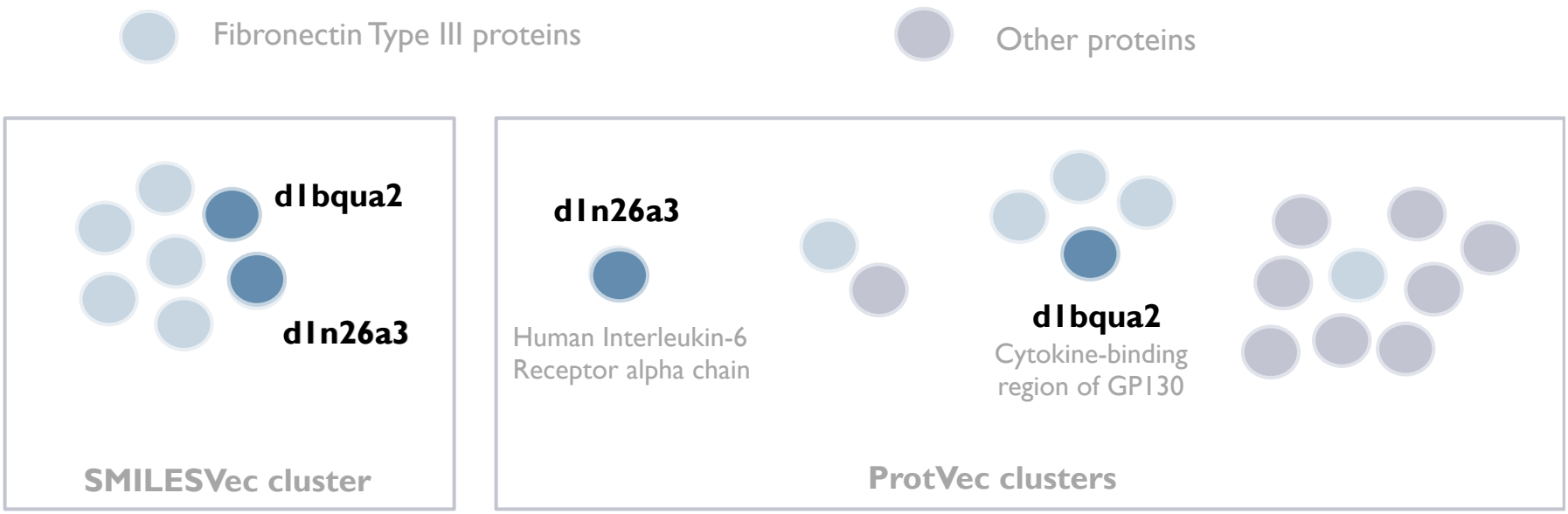
Bu yaklaşımla protein öbekleme probleminde protein dizisini kullanan yaklaşımlarla aynı başarıyı elde ettik.

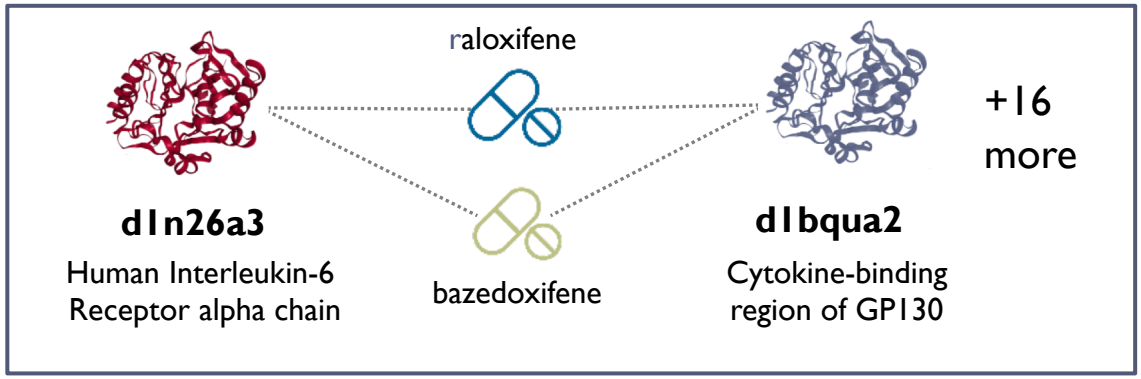
H. Ozturk, E. Ozkirimli, and A. Ozgur. ***A novel methodology on distributed representations of proteins using their interacting ligands.*** *Bioinformatics,* Volume 34, Issue 13, Pages i295-i303, 2018.

SMILESVec ve ProtVec ile öbeklemeye örnek

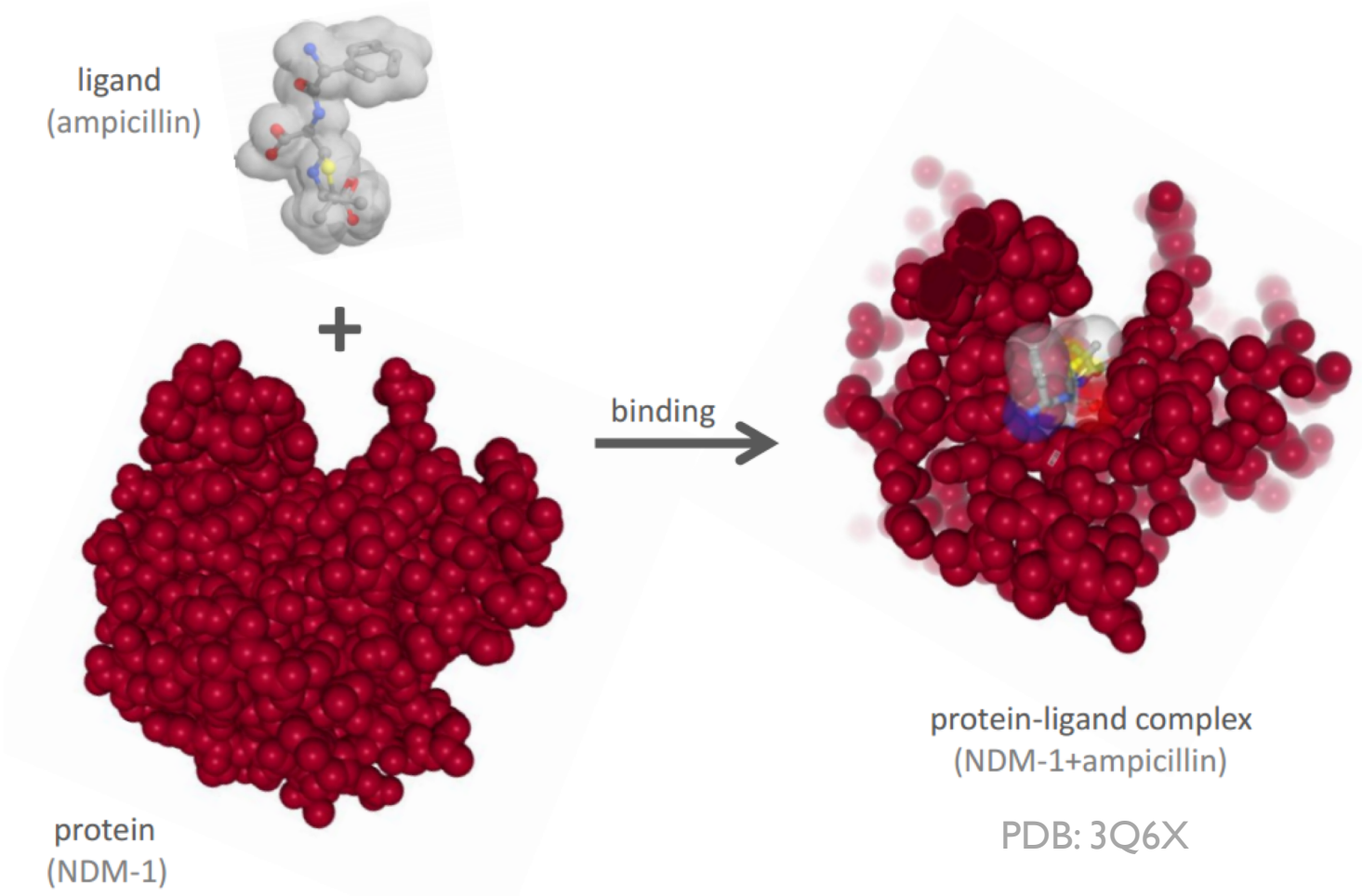


SMILESVec ve ProtVec ile öbeklemeye örnek

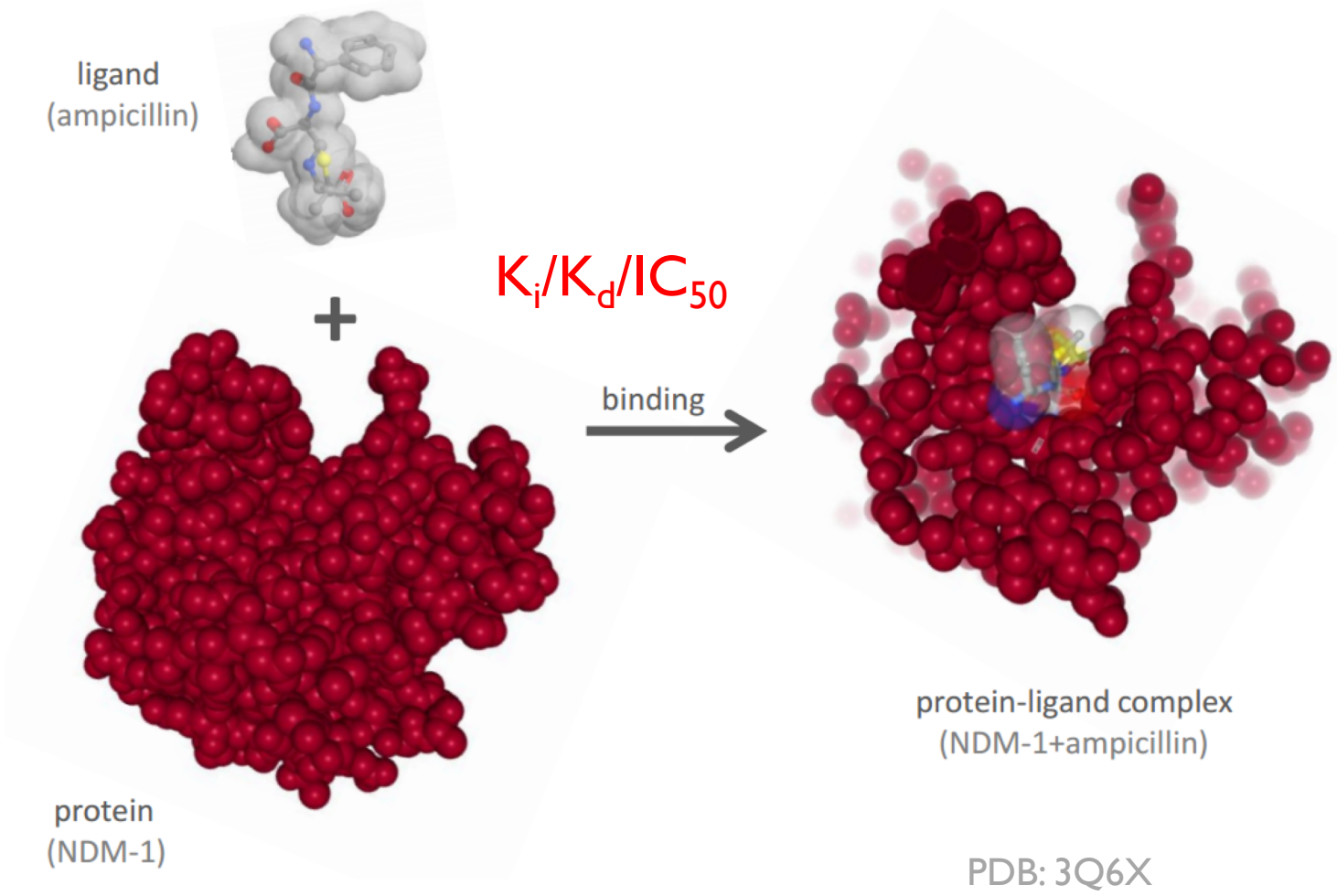




İlaç-protein etkileşim tahmini



İlaç-protein bağlanma ilgisi tahmini

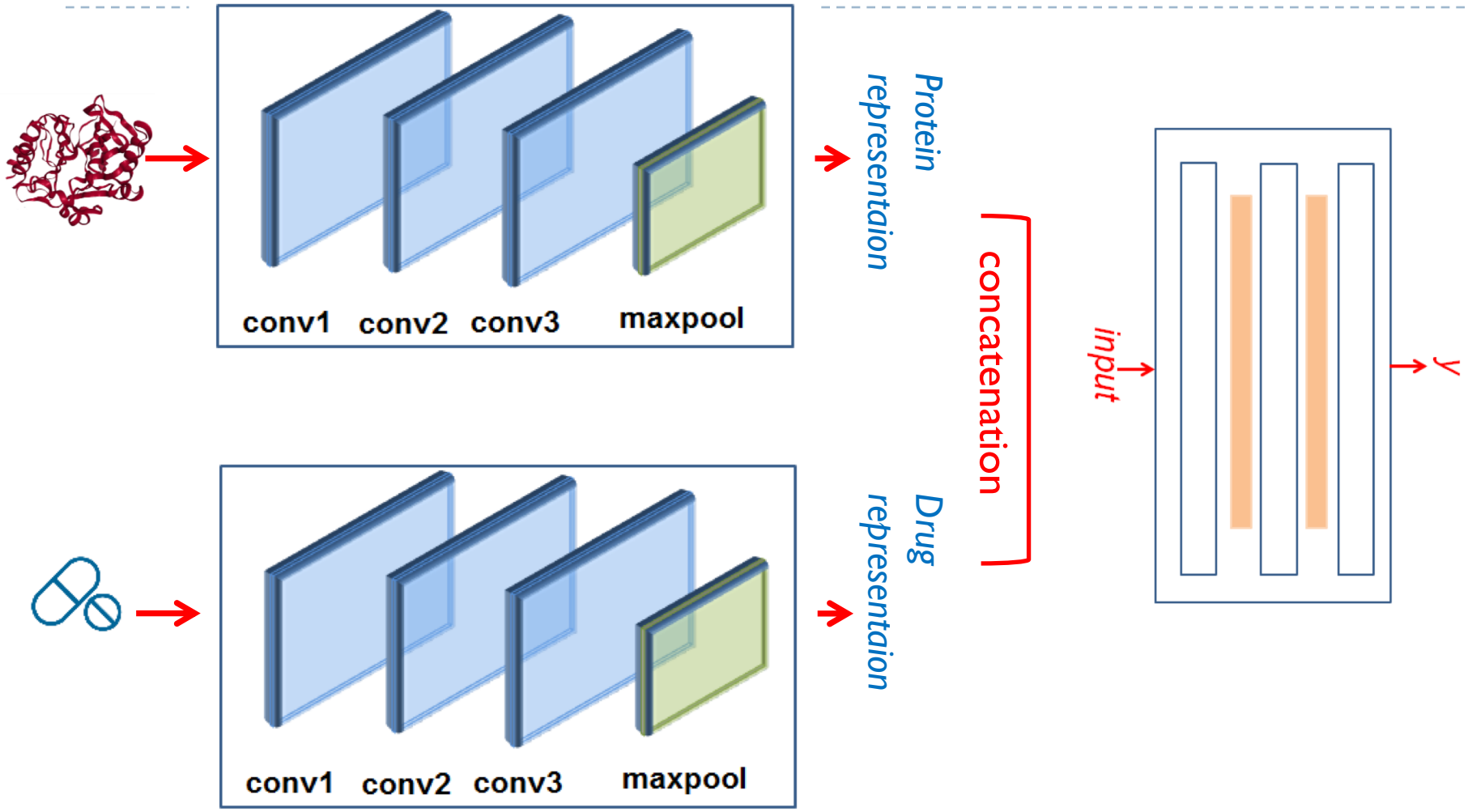


H. Öztürk, A. Özgür, E. Ozkirimli; **DeepDTA: deep drug–target binding affinity prediction,** ***Bioinformatics*****,** Volume 34, Issue 17, Pages i821–i829, 2018.

DeepDTA

Derin öğrenme algoritması ile kimyasal ve proteinin sadece metin tabanlı dizilerini kullanıyor.

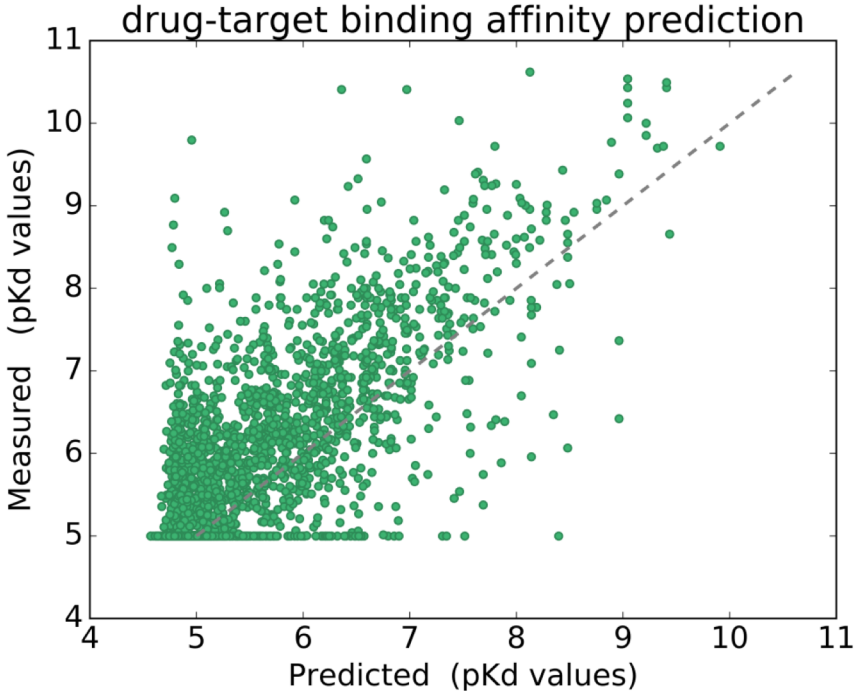
DeepDTA Modeli



Protein-ilaç temsili için Evrişimli Sinir Ağları (CNN)

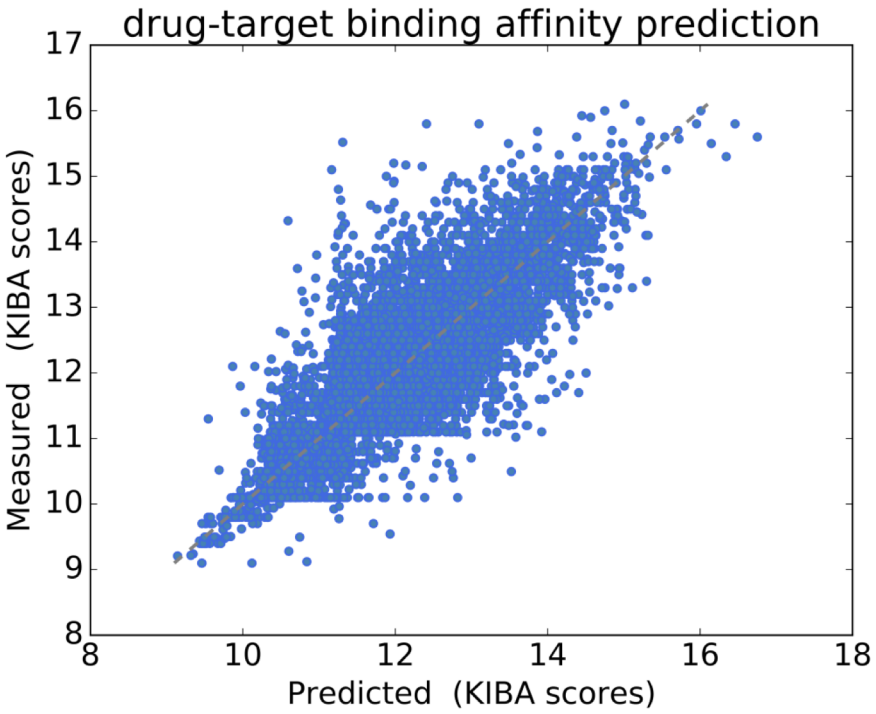
Bağlanma ilgisi tahmini için çok katmanlı yapay sinir ağları

Sonuçlar – **Davis veri kümesi**



Method	**CI**	**AUPR**	**MSE**
KronRLS (Pahikkala et al., 2014)	0.871	0.661	0.379
SimBoost (He et al., 2017)	0.872	0.709	0.282
DeepDTA	**0.878**	**0.714**	**0.261**

Sonuçlar – **KIBA veri kümesi**

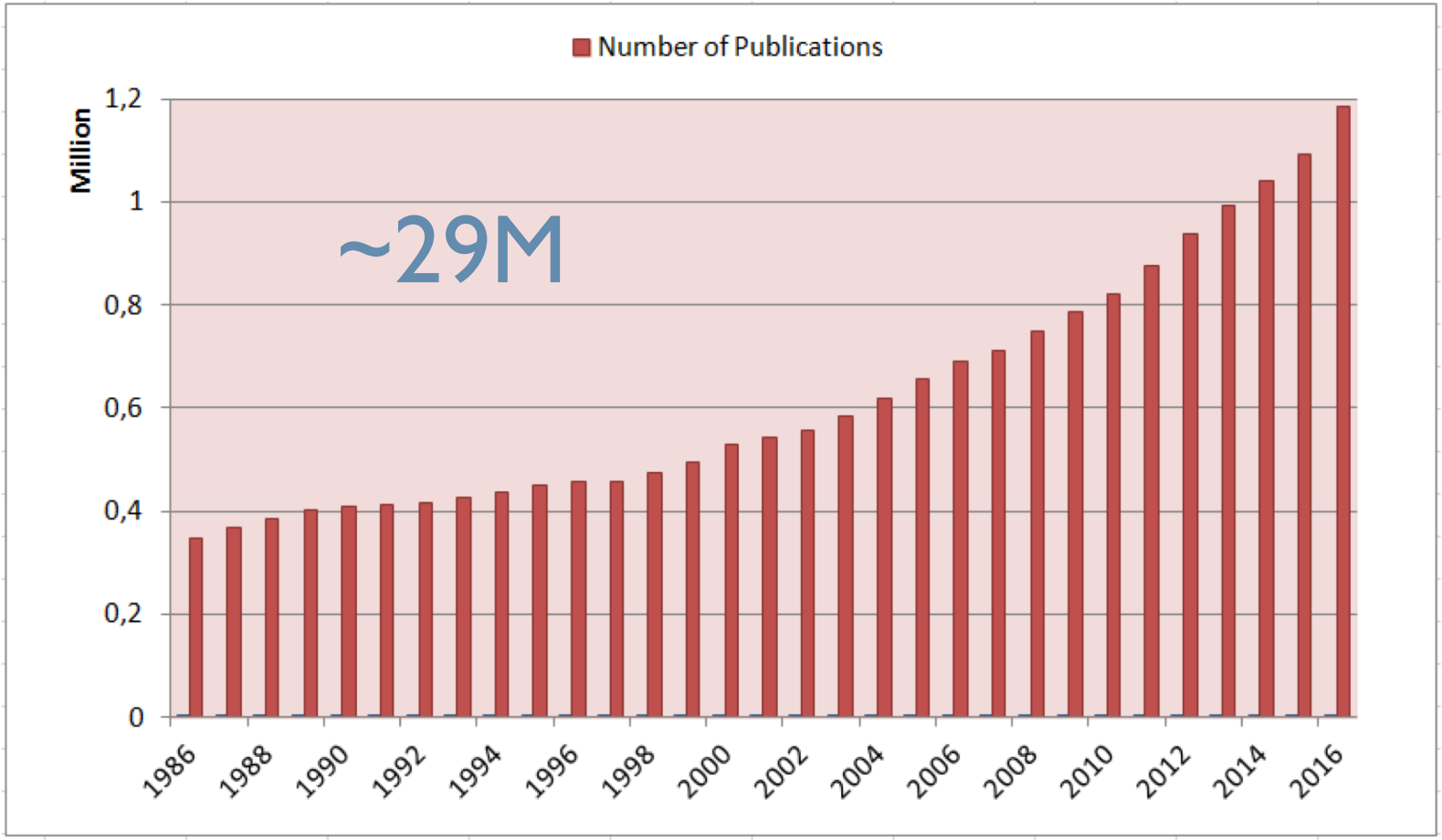


* statistical significance
95% conf. interval
p-value: 0.0001

Method	**CI**	**AUPR**	**MSE**
KronRLS (Pahikkala et al., 2014)	0.782	0.635	0.411
SimBoost (He et al., 2017)	0.836	0.760	0.222
DeepDTA	**0.863***	**0.788***	**0.194**

Biyomedikal Metin Madenciliği

Bilimsel Yayınlarda Artış



Literatürdeki bilgiye erişim zorluğu

Çok zor veya imkansız: bilim insanlarının ilgili yayınları takip edebilmesi.

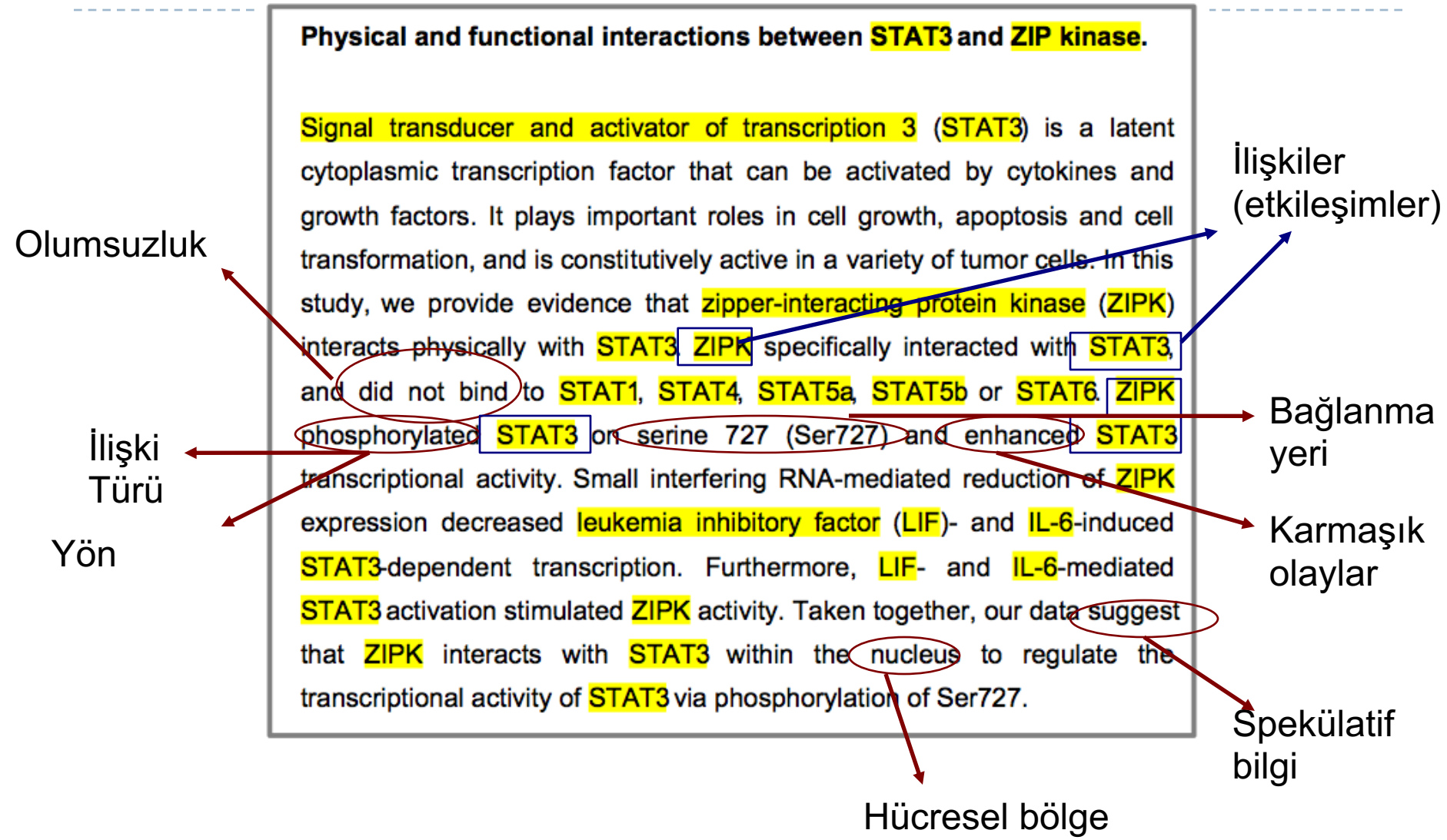
Manuel oluşturulan veritabanları: mevcut bilginin çok küçük bir kısmını kapsamaktadır.

Amaç

- Doğal dil işleme ve yapay öğrenmeye dayanan yöntemlerle önemli bilgilerin otomatik olarak çıkarılması.

- Gizli bağlantıların tespit edilerek yeni bilimsel hipotezlerin oluşturulması.

Ne tür bilgiler çıkarabiliriz?



Ne tür bilgiler çıkarabiliriz?

Bifidobacterium longum NCC2705
Description
Bifidobacterium. Representatives of this genus naturally colonize the human gastrointestinal tract (GIT) and are important for establishing and maintaining homeostasis of the intestinal ecosystem to allow for normal digestion.
Their presence has been associated with beneficial health effects, such as prevention of diarrhea, amelioration of lactose intolerance, or immunomodulation.
The stabilizing effect on GIT microflora is attributed to the capacity of bifidobacteria to produce bacteriocins, which are bacteriostatic agents with a broad spectrum of action, and to their pH-reducing activity.
Most of the ~30 known species of bifidobacteria have been isolated from the mammalian GIT, and some from the vaginal and oral cavity.
All are obligate anaerobes belonging to the Actinomycetales, branch of Gram-positive bacteria with high GC content that also includes Corynebacteria, Mycobacteria, and Streptomycetes.
Description
Bifidobacterium longum.
This organism is found in adult humans and formula fed infants as a normal component of gut flora.
Description
Bifidobacterium longum strain NCC2705.
This strain was isolated from infant feces.
The genome of this strain is being sequenced for comparative genomics.

1) Varlık İsmi Tanıma & Normalizas-yonu

- Bacteria Names
- Habitat Names

2) İlişki Çıkarma

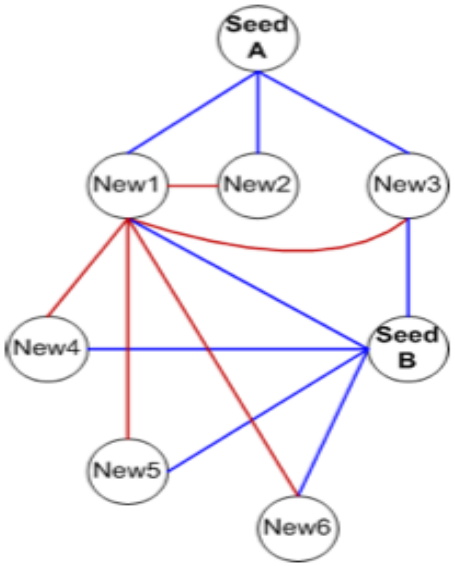
Localization

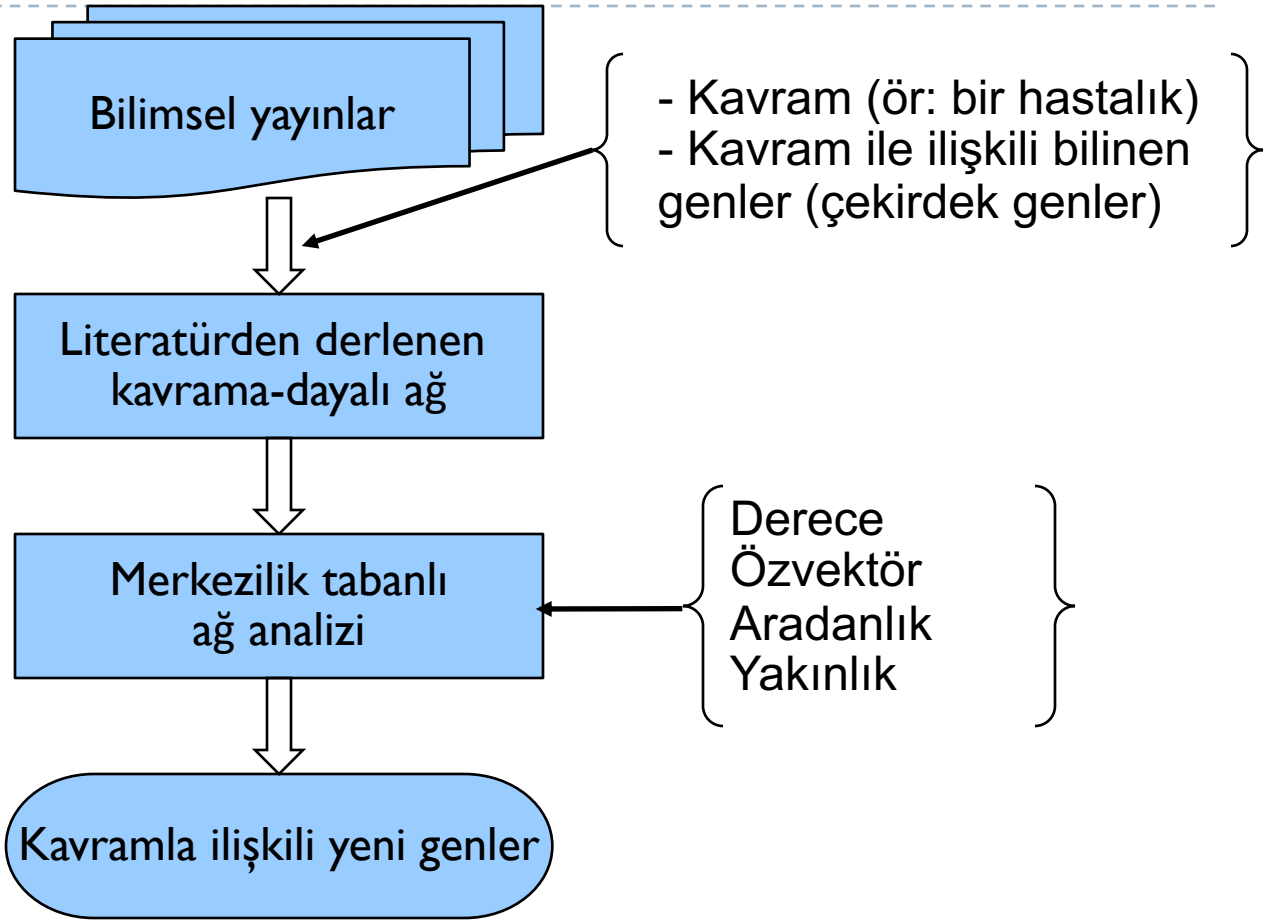
Bacteria ⟷ Habitat

Part-Of

Habitat ⟷ Habitat

Metin ve Ağ Madenciliği ile Yeni Bağlantı Tespiti





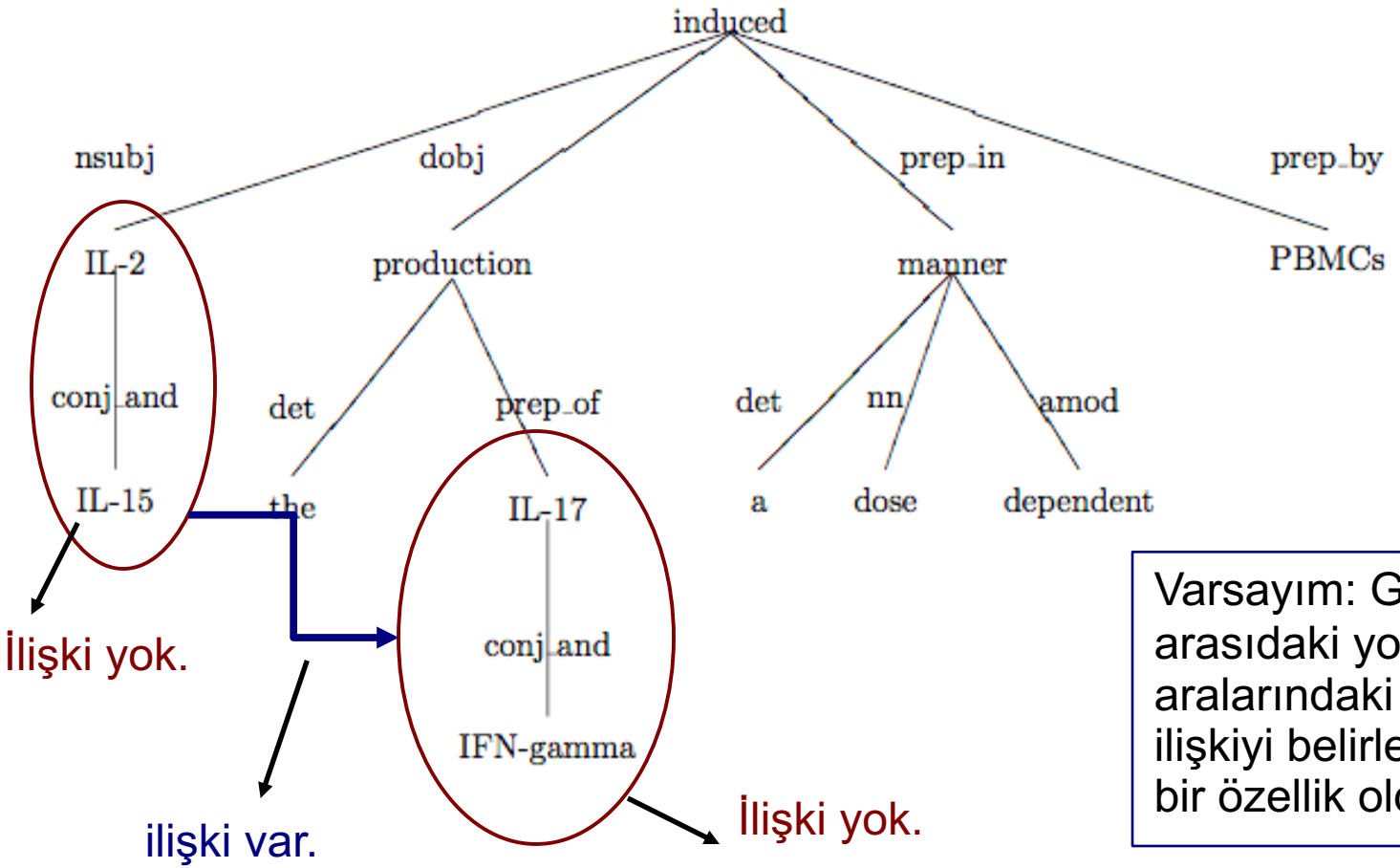
Hipotez:
Kavrama bağlı gen-gen etkileşim ağındaki önemli genler de bu kavramla ilişkili olabilir.

Prostat kanseri, aşıya bağlı bağışıklık ve ateş ile ilişkili genlerin tespiti için kullanılmıştır.

Gen-gen etkileşimlerinin tespiti

IL-2 and IL-15 induced the production of IL-17 and IFN-gamma in a dose dependent manner by PBMCs.

(Genia Tagger: 71% F-measure)



Varsayım: Genler arasıdaki yolun aralarındaki semantik ilişkiyi belirlemek için iyi bir özellik olduğu.

- Bağlam ağaçlarının oluşturulması için Stanford Parser kullanıldı (de Marneffe et al., 2006).

Bağlam Ağacı Yolu Edit Fonksiyonu

- İlk diziyi ikinciye çevirmek için minimum işlem sayısı (işlemler: bir kelimenin eklenmesi, silinmesi veya değiştirilmesi)
  - **IL2** – nsubj – induced – dobj – production – prep_of – **IL-17**
  - **IL2** – nsubj – induced – dobj – production – prep_of – IL-17 – conj_and – **IFN-gamma**
  - **IL-17** – conj_and – **IFN-gamma**
    - Edit mesafesi (Yol1 -> Yol2) = 2 (2 ekleme)
    - Edit mesafesi (Yol1 -> Yol3) = 8 (6 çıkarma + 2 ekleme)

Benzerlik fonksiyonuna dönüştürme:

$$EditSim(p_i, p_j) = e^{[-\gamma(EditDist(p_i, p_j))]}$$

• Çekirdek fonksiyon olarak SVM^{light} paketine (Joachims, 1999) entegre edildi.

Değerlendirme

Veri kümeleri:

Data Set	Sentences	+ Sentences	- Sentences
AIMED[1]	4026	951	3075
CB[2]	4056	2202	1854

Sonuçlar:

	Precision	Recall	F-measure
AIMED	77.52	43.51	55.61
CB	85.15	84.79	84.96

Kesinlik (Precision): Bulunan ilişkiler içinde doğru olanların oranı.
Bulma (Recall): Mevcut ilişkilerden doğru bulunanların oranı.
F-ölçütü: Kesinlik ve bulmanın harmonik ortalaması.

[1] ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/
[2] http://biocreative.sourceforge.net/biocreative_2.html

Prostat Kanseri Genlerinin Tahmini

OMIM Morbid Map'ten 15 prostat kanseri geni çekirdek gen olarak kullanıldı.

Gene	Description
AR	androgen receptor
BRCA2	breast cancer 2, early onset
MSR1	macrophage scavenger receptor 1
EPHB2	EPH receptor B2
KLF6	Kruppel-like factor 6
MAD1L1	MAD1 mitotic arrest deficient-like 1 (yeast)
HIP1	huntingtin interacting protein 1
CD82	CD82 molecule
ELAC2	elaC homolog 2 (E. coli)
MXI1	MAX interactor 1
PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)
RNASEL	ribonuclease L (2',5'-oligoisoadenylate synthetase-dependent)
HPC1	hereditary prostate cancer 1
CHEK2	CHK2 checkpoint homolog (S. pombe)
PCAP	predisposing for prostate cancer

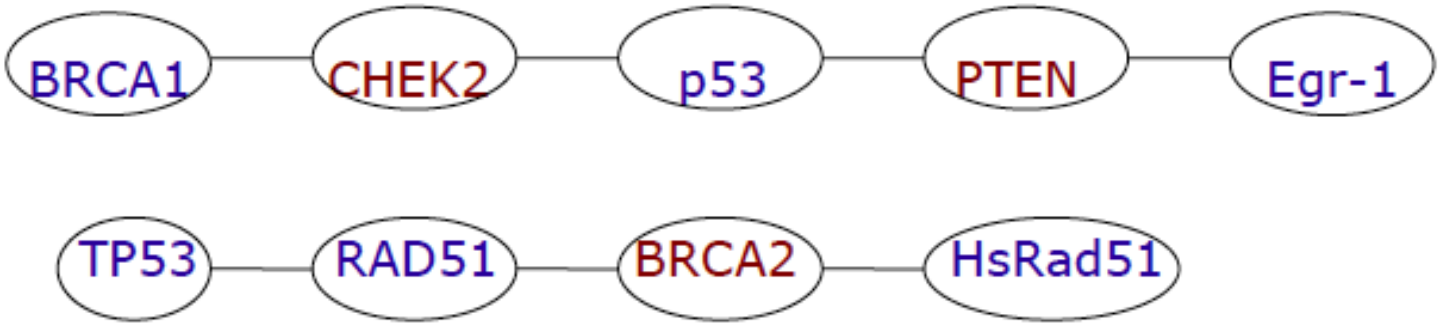
A. Ozgur, T. Vu, G. Erkan, and D. R. Radev. **Identifying gene-disease associations using centrality on a literature mined gene interaction network**. Bioinformatics, Volume 24, Number 13, pp. i277-i285, 2008.

Etkileşim Ağının Oluşturulması

Etkileşim içeren örnek cümleler:

- PTEN is transcriptionally regulated by transcription factors such as p53 and Egr-1.
- In response to DNA damage, the cell-cycle checkpoint kinase CHEK2 can be activated by ATM kinase to phosphorylate p53 and BRCA1, which are involved in cell-cycle control and apoptosis.
- The interactions of RAD51 with TP53, RPA and the BRC repeats of BRCA2 are relatively well understood (see Discussion).
- The interaction of BRCA2 with HsRad51 is significantly more different to both RadA and RecA (Figure 2c).

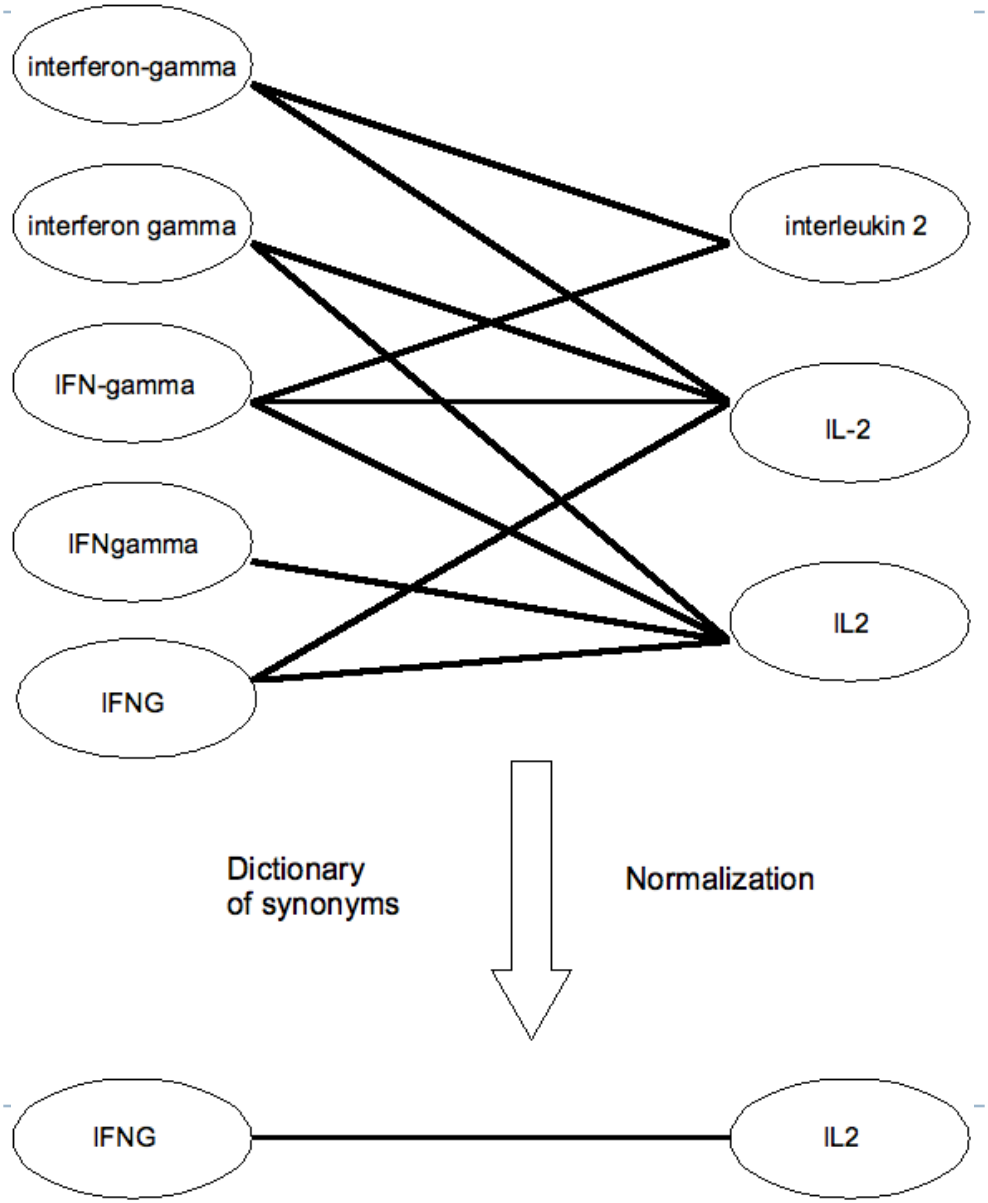
- Oluşturulan ağ:



seed genes:

CHEK2
PTEN
BRCA2

Gene Adı Normalizasyonu



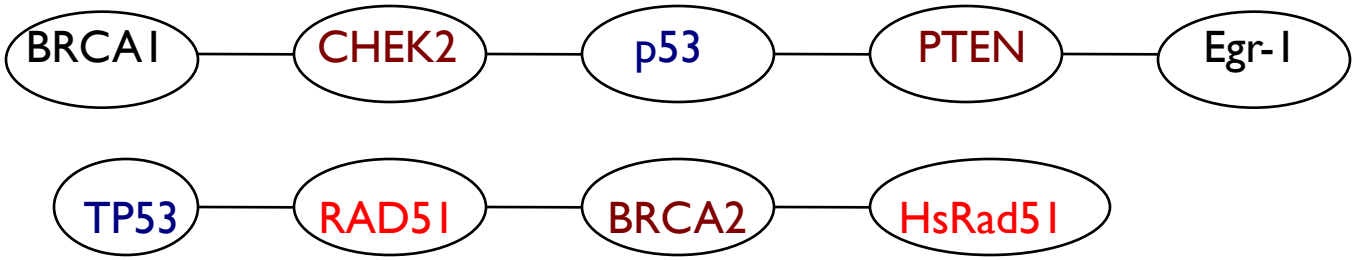
Gen ismi ve eşanlamlıları için HUGO Gene Nomenclature Committee (HGNC) veritabanı sözlük olarak kullanıldı

~28,000 gen girdisi

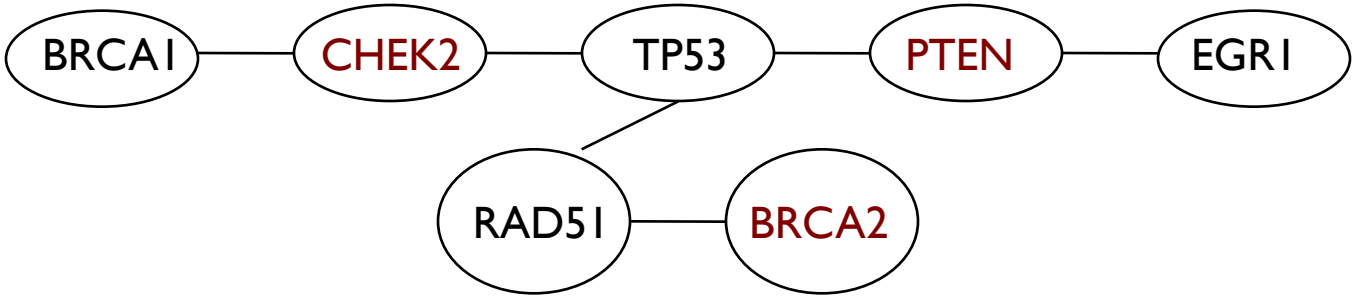
(http://www.genenames.org/)

Etkileşim Ağının Oluşturulması

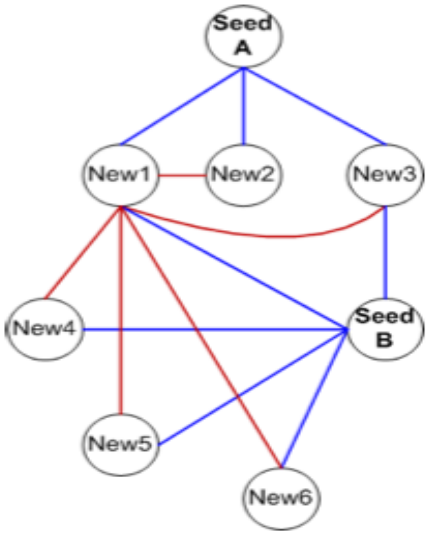
- Gen ismi normalizasyonundan önce:



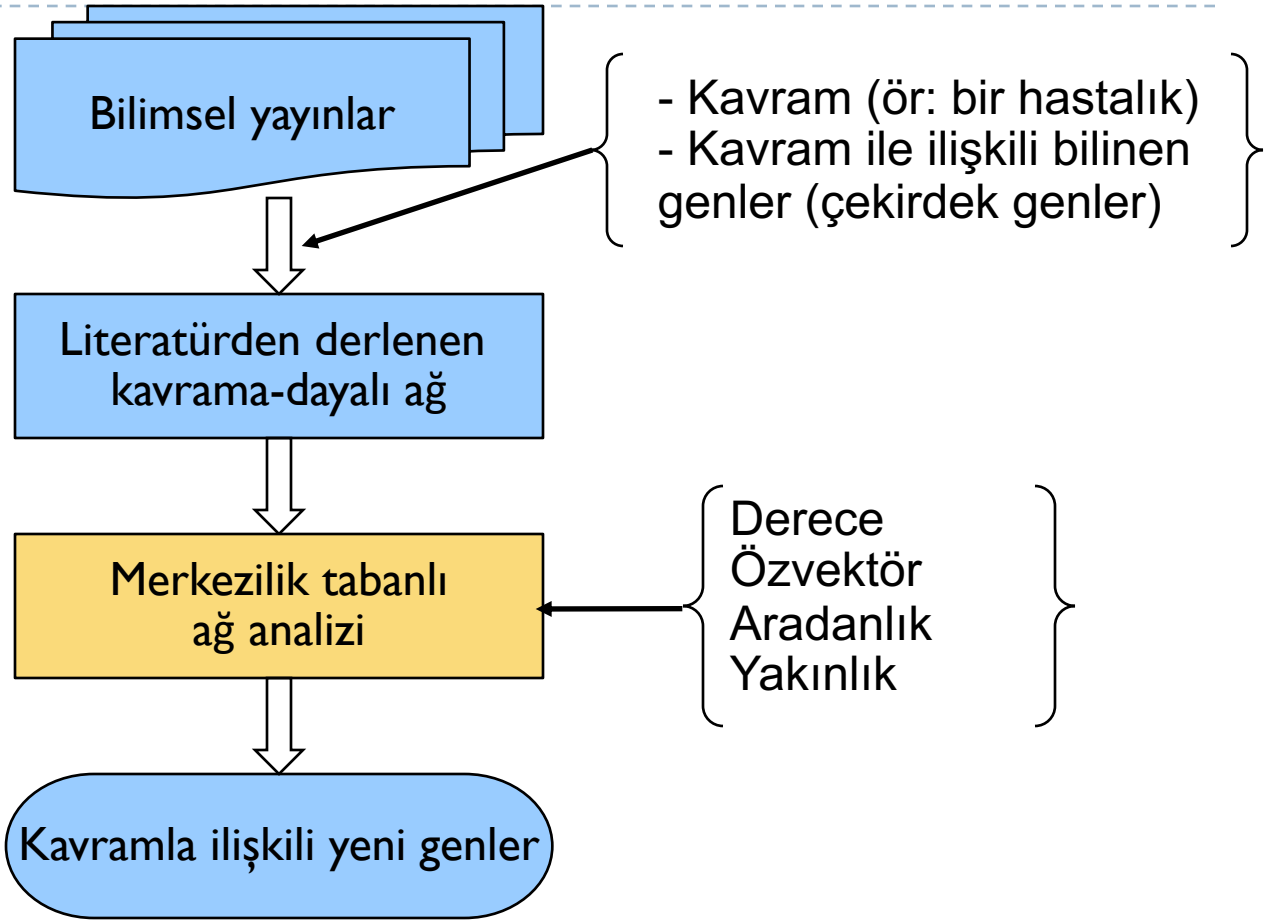
- Gen ismi normalizasyonundan sonra:



Genlerin Ağ Merkezilik Ölçütleri ile Sıralanması



Hipotez:
Kavrama bağlı gen-gen etkileşim ağındaki önemli genler de bu kavramla ilişkili olabilir.



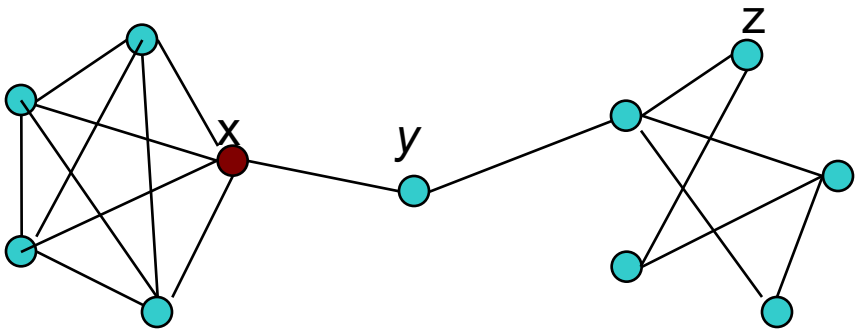
Bir düğümün ağdaki önemi

Derece Merkezilik Ölçütü

- Bir düğümün bağlı olduğu düğüm sayısı.

  A: Komşuluk matrisi

$$k_i = \sum_{j=1}^{n} A_{ij}$$



- Bir düğümün ne kadar çok komşusu varsa, o kadar önemlidir.
- *x* düğümünün derece merkezilik ölçütü 5, y'ninki ise 2.

Özvektör Merkezilik Ölçütü

- Bir düğümün komşularının merkezilik ölçütlerinin toplamıyla orantılıdır.

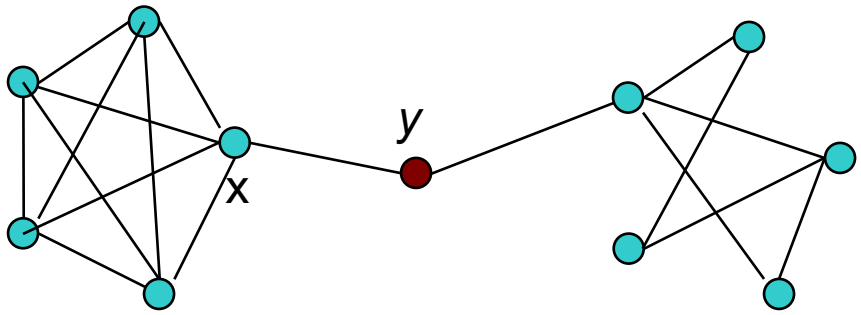
$$x_i = \lambda^{-1} \sum_{j=1}^{n} A_{ij} x_j$$

- Matris gösteriminde: $\lambda \mathbf{x} = \mathbf{A}\mathbf{x}$
  - λ **A** matrisinin en büyük özdeğeri, **x** de ilgili özvektördür.

- Her komşu bir düğümün merkeziliğine eşit oranda katkı sağlamamaktadır.
- Sosyal ağlarda "prestij" olarak ifade edilmektedir.
  - Bir kişinin prestiji sadece kaç arkadaşı olduğuna değil, bu arkadaşların kim (ne kadar prestijli) olduğuna da bağlıdır.

Yakınlık Merkezilik Ölçütü

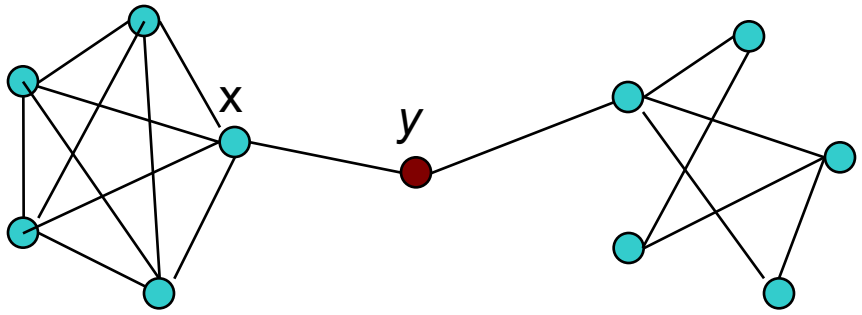
- 1/(düğümün diğer düğümlere olan uzaklıklarının toplamı)



- Bir düğüm diğer düğümlere ne kadar yakınsa, o kadar önemlidir.

Aradanlık Merkezilik Ölçütü

- Bir i düğümü için, i düğümünün üzerinden geçen en kısa yolların sayısının tüm en kısa yollara oranı.



- Bir düğüm ne kadar çok en kısa yol üzerinde yer alırsa, o kadar önemlidir.

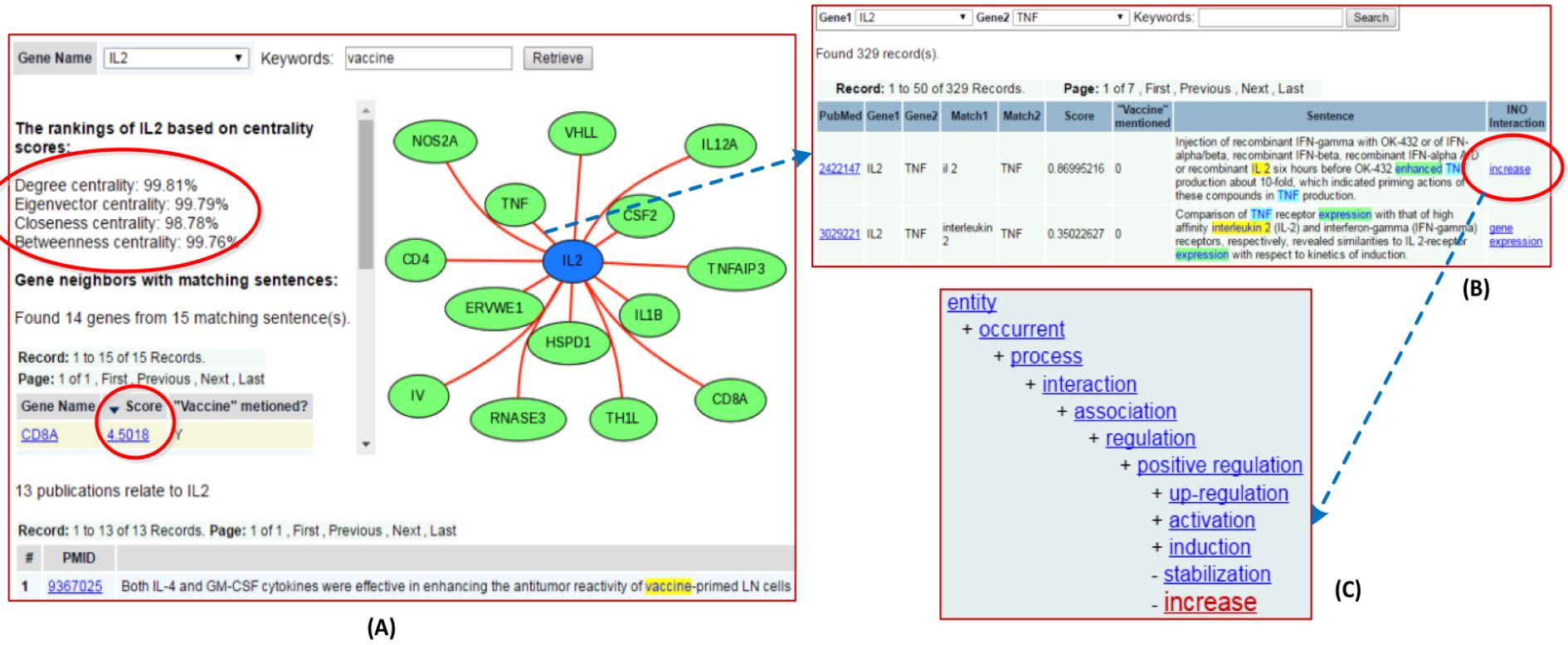
En üst sıradaki 20 Gen

Gene	Degree	Eigenvector	Closeness	Betweenness	Evidence
TP53	+	+	+	+	PGDB
BRCA1	+	+	+	+	PGDB
EREG	+	+	+	+	None
AKT1	+	+	+	+	PGDB
MAPK1	+	+	+	+	Literature (Hao *et al.*, 2007; Sarfaraz *et al.*, 2006)
TNF	+	+	+	+	PGDB
CCND1	+	+	+	+	PGDB
MYC	+	+	+	+	PGDB
APC	+	+	−	−	PGDB
CDKN1B	+	+	+	−	PGDB
MAPK8	+	+	+	+	PGDB
NR3C1	−	+	+	−	Literature (Wei *et al.*, 2007)
VEGFA	+	+	+	−	PGDB
MDM2	+	+	+	−	KEGG and Literature (Wang *et al.*, 2003; Zhang *et al.*, 2003)
POLD1	−	−	+	+	None
SNORA62	−	−	+	+	None
CNTN2	−	−	−	+	None
PPA1	−	−	−	+	None
TMEM37	−	−	+	−	None
FZR1	−	−	+	−	PGDB
SSSCA1	−	−	+	−	None
BCL2	+	−	−	−	PGDB
INS	+	−	−	−	KEGG and Literature (Ho *et al.*, 2003)

12 gen: Prostate Gene DataBase (PGDB)
2 gen: KEGG pathway for prostate cancer ve literatür (MDM2 and INS)
2 gen: literatür (NR3C1 and MAPK1)
7 gen: Olumlu veya olumsuz kanıt bulunamadı.

IGNET: Integrated Gene Network

http://ignet.hegroup.org



Arzucan Özgür, Junguk Hur, Zuoshuang Xiang, Edison Ong, Dragomir R. Radev, Yongqun He: **Ignet: A Centrality and INO-based Web System for Analyzing and Visualizing Literature-mined Networks**. ICBO/BioCreative 2016

Dynamic Ignet Search Program

Keywords Search (throught PubMed search engine)

Keywords: bipolar disorder Search

Keywords: bipolar disorder.

Found 199 gene pairs.

Calculate centrality scores.

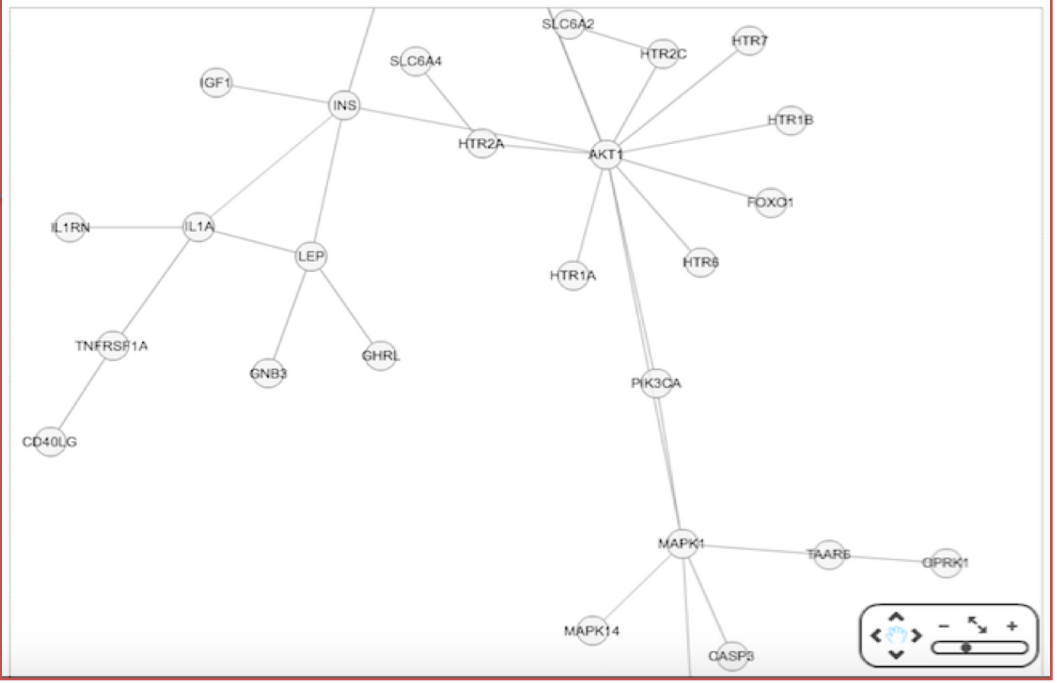
Show network in Cytoscape Web.

#	Gene 1	Gene 2	Number of hits
1	ABCA1	NR1H3	1 hits
2	ACE	AGT	2 hits
3	ACTB	RELN	1 hits
4	ADA	ADARB1	1 hits
5	ADCY7	SUCLG1	1 hits

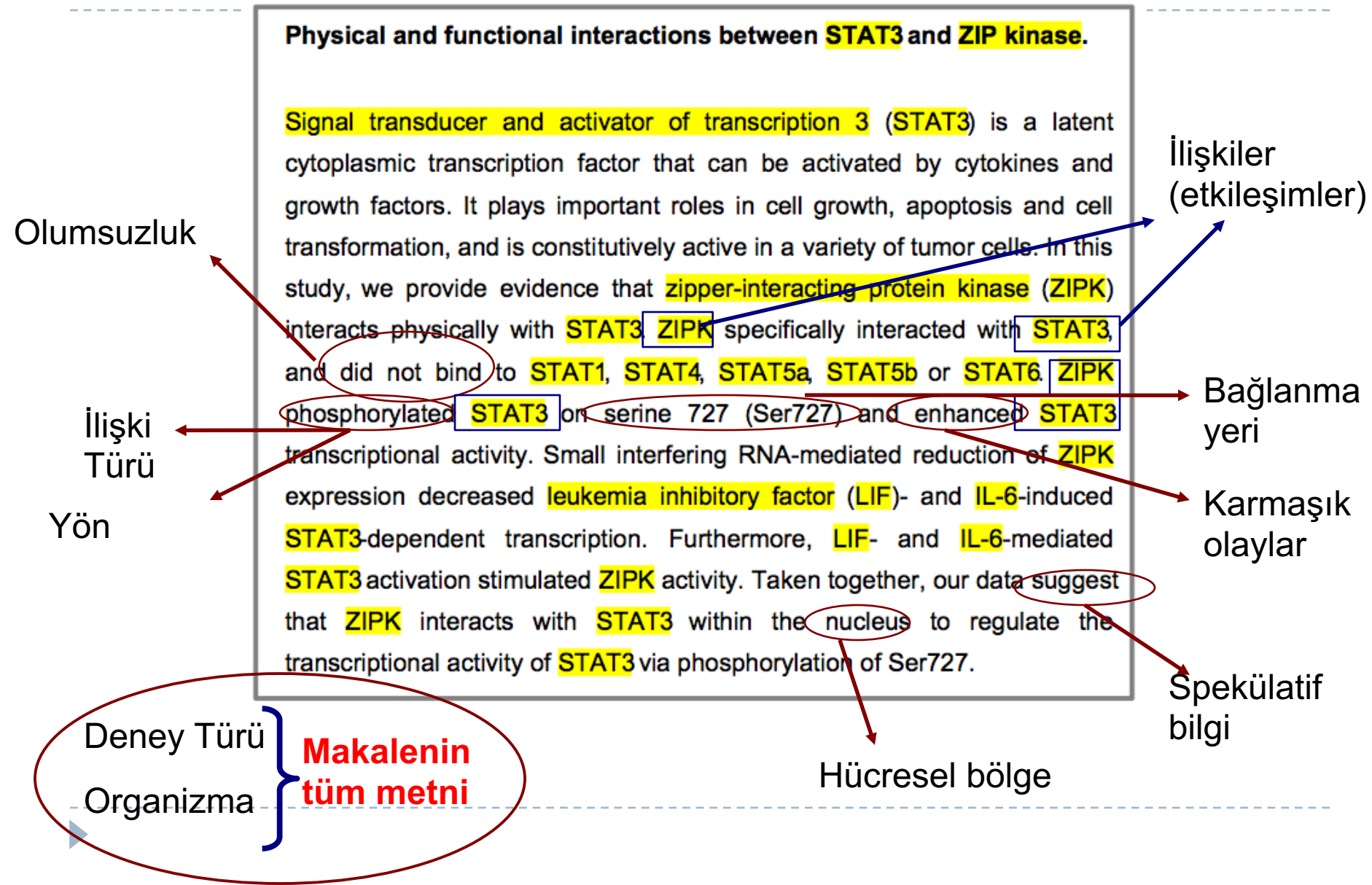
Degree centrality			Eigenvector centrality		
#	Gene	Score	#	Gene	Score
1	GSK3B	0.0645	1	GSK3B	0.0241
2	AKT1	0.0553	2	AKT1	0.0214
3	PER1	0.0415	3	BDNF	0.0188
4	NFIL3	0.0415	4	DISC1	0.0161
5	BDNF	0.0415	5	HSPA5	0.0123

Found 199 gene pairs. Below are the network shown in Cytoscape Web.

Download network in graphml format.



Yerel Olmayan Bağlam Bilgisi Çıkarma



Deneysel Yöntemlerin Açıklandığı Pasajları Bulma

İşbirliği: Ferhat Aydın ve Zehra Melce Hüsünbeyi

Aydın F., Hüsünbeyi Z.M., Özgür A. (2017). **Automatic query generation using word embeddings for retrieving passages describing experimental methods.** *Database*, 2017.

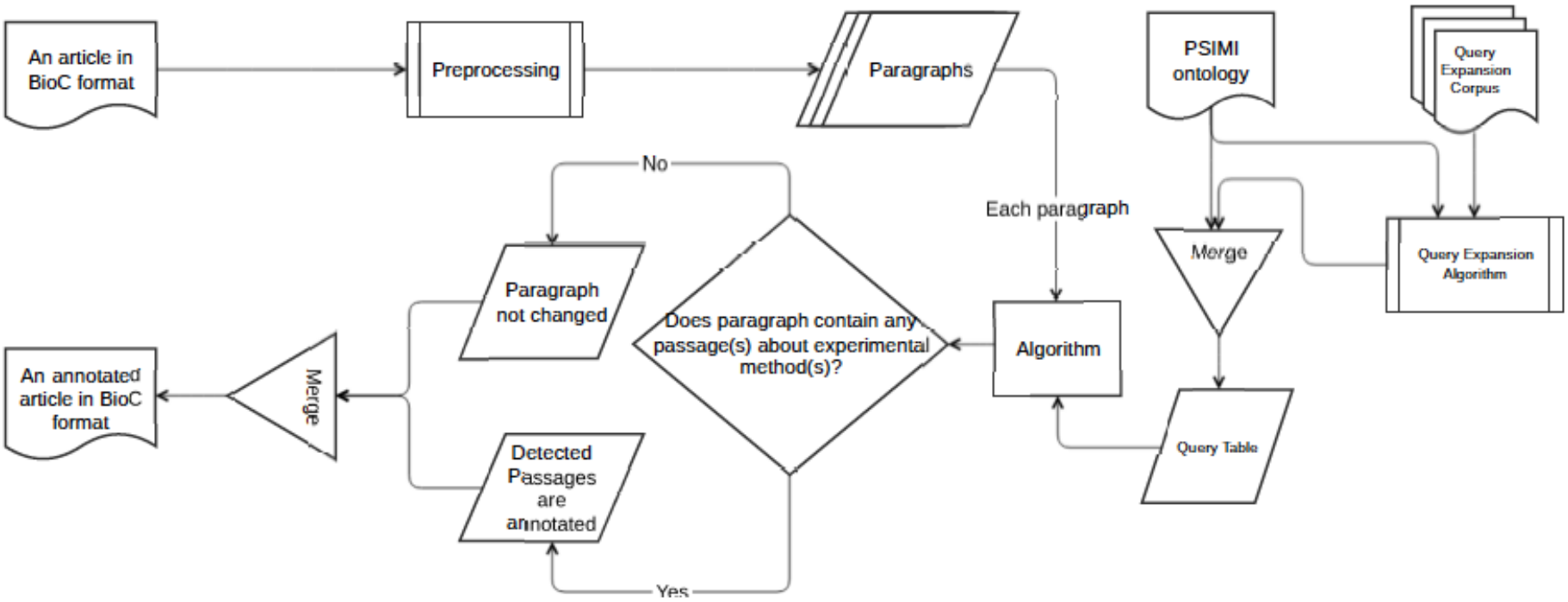
Birden çok deneysel yöntemin anlatıldığı örnek bir paragraf

To identify novel interactors of TBK1, we generated stable RAW264.7 cell lines that express a GS-TAP-tagged version of TBK1 and purified the TBK1 protein complex by tandem affinity purification (Figure 1A and B) (Burckstummer et al, 2006). Mass spectrometry analysis identified TBK1 along with the core complex components TRF family member-associated NF-kappa-B activator (TANK; 17 peptides), TBK-binding protein 1 (TBKBP1, also referred to as SINTBAD; 10 peptides) and TBKBP2 (also referred to as NAP1 or AZI2; 15 peptides), indicating that this native purification was efficient and in agreement with previously published data on the TBK1 core complex (Bouwmeester et al, 2004). In addition, we identified the DEAD-box helicase DDX3X (RefSeq NP_034158) with five peptides. Immunoprecipitation experiments using tagged TBK1 suggested that its interaction with DDX3X and the transcription factor IRF3 are significantly weaker than the interaction between TBK1 and TANK and therefore not detected by coimmunoprecipitation under stringent conditions (Supplementary Figure 1A). Likewise, we could not detect any association of DDX3X with IRF3 (Supplementary Figure 1B).

Manuel Etiketleme

```
<passage>
    <infon key="type">paragraph</infon>
    <offset>17753</offset>
    <text>Infection of eukaryotic cells with large DNA viruses often results in extensive interactions of viral
    gene products with macromolecular pathways of the host cell. By using the yeast two-hybrid system, we
    identified cellular hnRNP-K as an interacting protein with ASFV early protein p30. This interaction was
    further confirmed by an in vitro GST-fusion pull-down assay, using either p30 obtained from baculovirus
    system or ASFV infected cell extracts.</text>
    <annotation id="10">
        <infon key="type">ExperimentalMethod</infon>
        <infon key="PSIMI">0018</infon>
        <location offset="17916" length="123"/>
        <text>By using the yeast two-hybrid system, we identified cellular hnRNP-K as an interacting protein with
        ASFV early protein p30.</text>
    </annotation>
    <annotation id="11">
        <infon key="type">ExperimentalMethod</infon>
        <infon key="PSIMI">0096</infon>
        <location offset="18040" length="163"/>
        <text>This interaction was further confirmed by an in vitro GST-fusion pull-down assay, using either p30
        obtained from baculovirus system or ASFV infected cell extracts.</text>
    </annotation>
</passage>
```

Sistemin genel iş akışı



Sorgu oluşturma

```
{
  name = "0018"
  synonym = [ two hybrid, two-hybrid, yeast two hybrid, 2 hybrid, 2-hybrid, 2h, y2h,
              classical two hybrid, gal4 transcription regeneration, yeast two-hybrid ]
  tier 1 = [ yeast, hybrid, y-2h ]
  tier 2 = [ bait, cdna, gal4, gal, galactosidase ]
}

{
  name = "0019"
  synonym = [ co-immunoprecipitation, coimmunoprecipitation, co-ip, coip, immunoprecipitation,
              anti bait coip, anti bait coimmunoprecipitation, anti tag coip,
              anti tag coimmunoprecipitation ]
  tier 1 = [ immunoprecipitated, immunoprecipitates, precipitated, precipitate, precipitates,
             co-precipitated, ip ]
  tier 2 = [ antibody, antibodies, tag, tagged, bait, immunoblotted]
}

{
  name = "0096"
  synonym  = [ pull down, affinity capture, pulldown, pull-down, pulled-down, pulled down ]
  tier 1 = [ pull, pulled, down, affinity, capture ]
  tier 2 = [ gst, appl1, rab5, gel, glutathione ]
}
```

İsim ve eşanlamlılar PSI-MI ontolojisinden alındı. Tier 1 and Tier 2 terimleri tf.rf ile bulundu.

Sorgu cevaplanması

- To assess whether acetylation of any of these lysines affected Mediator's ability to interact with the Histone H4 or H3 tails we performed the Mediator ***pull down*** assay with Ac-(K5,8,12,16)-H4 and Ac-(K9,14)-H3 peptides. The Ac-(K9,14)-H3 peptide showed no difference in ***affinity*** from the unmodified H3-N peptide (Figure 5A), even at reduced concentrations of peptide and Mediator.

- Cell lysates were subjected to an anti-Flag ***immunoprecipitation*** (***IP***), and ***coprecipitating*** STRAP was detected by immunoblotting (Blot) with anti-HA ***antibodies*** (top section). In the middle section, total lysates were ***immunoprecipitated*** using anti-HA ***antibodies*** and then immunoblotted with anti-Flag ***antibodies***. To confirm expression of Smads, aliquots of total cell lysates were ***immunoblotted*** with anti-Flag ***antibodies*** (bottom section).

Sonuçlar

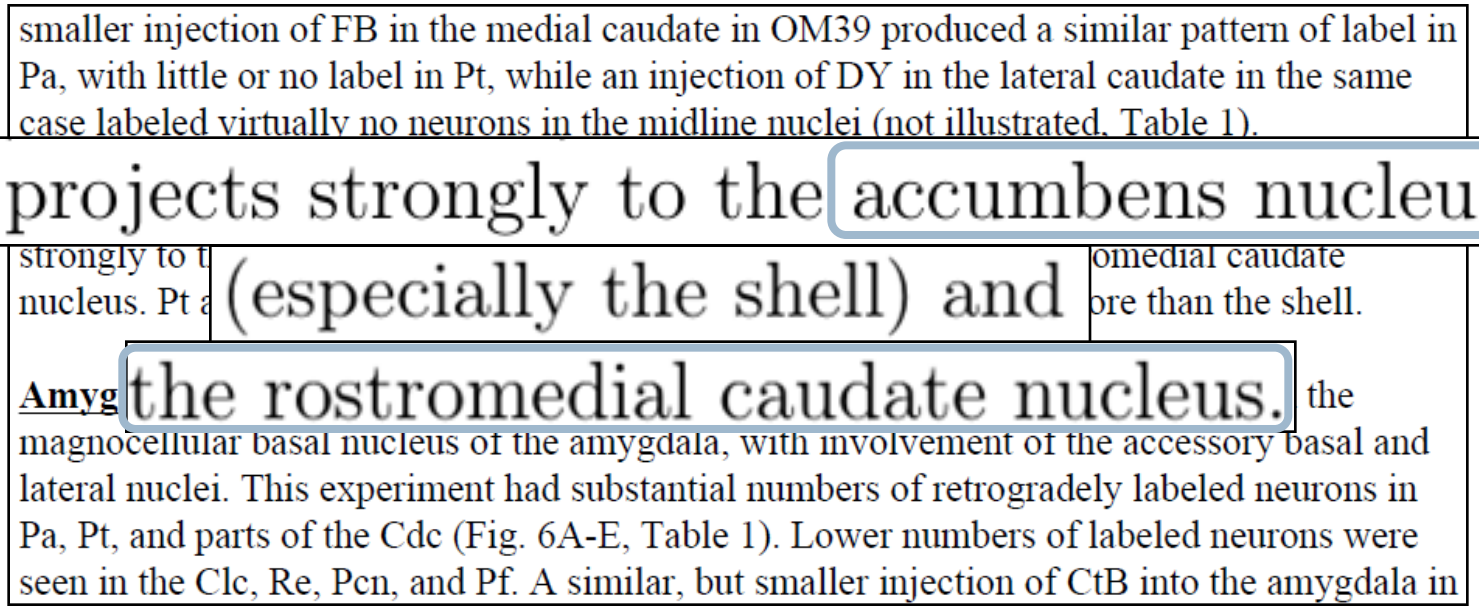
	Precision	Recall	F-measure
baseline	0.424	0.418	0.421
baseline.genia.ino	0.484	0.413	0.446
tf.rf.f7s7	0.120	0.508	0.194
tf.rf.f7s7.genia.ino	0.133	0.502	0.211
tf.rf.f10s10	0.068	0.512	0.119
tf.rf.f10s10.genia.ino	0.074	0.507	0.129
tf.rf.manual	0.315	0.508	0.389
tf.rf.manual.genia.ino	0.357	0.503	0.418
word2vec	0.321	0.618	0.422
word2vec.genia.ino	0.362	0.606	0.453
IAA	0.787	0.937	0.856

Neurobilim Alanında Metin Madenciliği

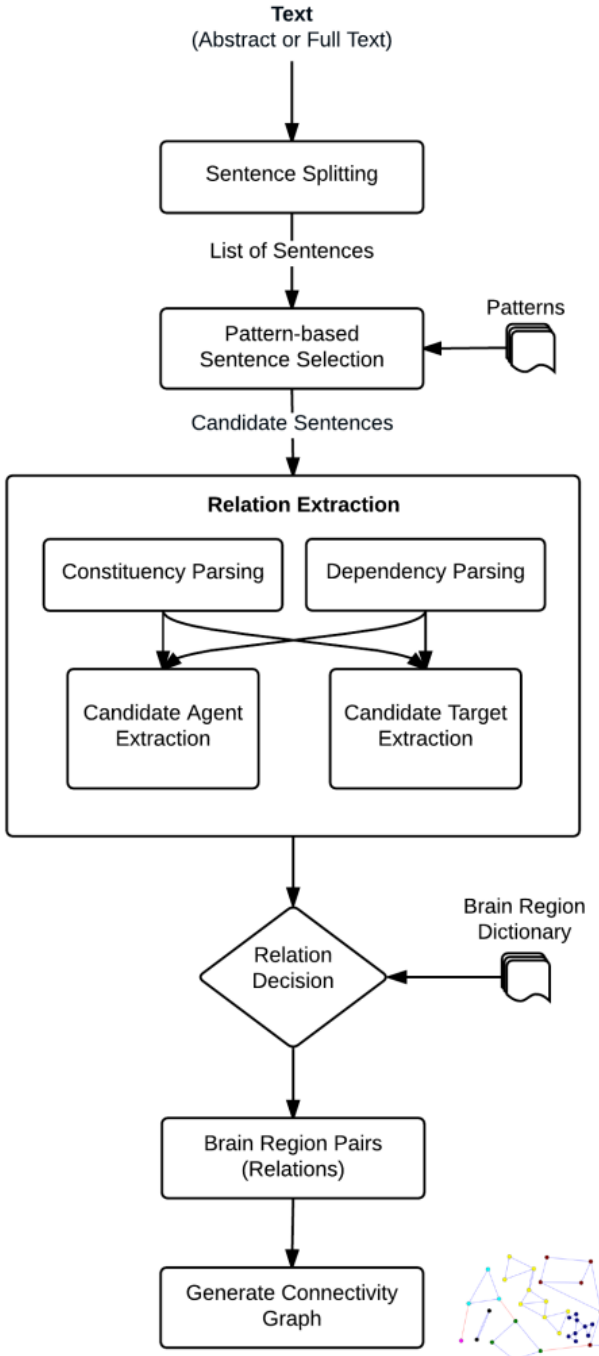
İşbirliği: Erinç Gökdeniz ve Reşit Canbeyli

E. Gokdeniz, A. Ozgur, R. Canbeyli. **Automated Neuroanatomical Relation Extraction: A Linguistically Motivated Approach with a PVT Connectivity Graph Case Study**. *Frontiers in Neuroinformatics*, 10:39, 2016.

Problem Tanımı



Hsu et al.(2009)


Text
(Abstract or Full Text)
Sentence Splitting
List of Sentences
Pattern-based
Sentence Selection
Patterns
Candidate Sentences
Relation Extraction
Constituency Parsing
Dependency Parsing
Candidate Agent
Extraction
Candidate Target
Extraction
Relation
Decision
Brain Region
Dictionary
Brain Region Pairs
(Relations)
Generate Connectivity
Graph

İlişki ifade eden kelimeler

"An anterograde tracer injection into the dorsal midline thalamus revealed strong projections to the accumbens nucleus." [1]

"For example, the pPVT was found to be distinctively innervated by the anterior most aspect of the prelimbic cortex and the agranular portions of the posterior insular cortex ." [2]

Beyin bölgeleri sözlüğü

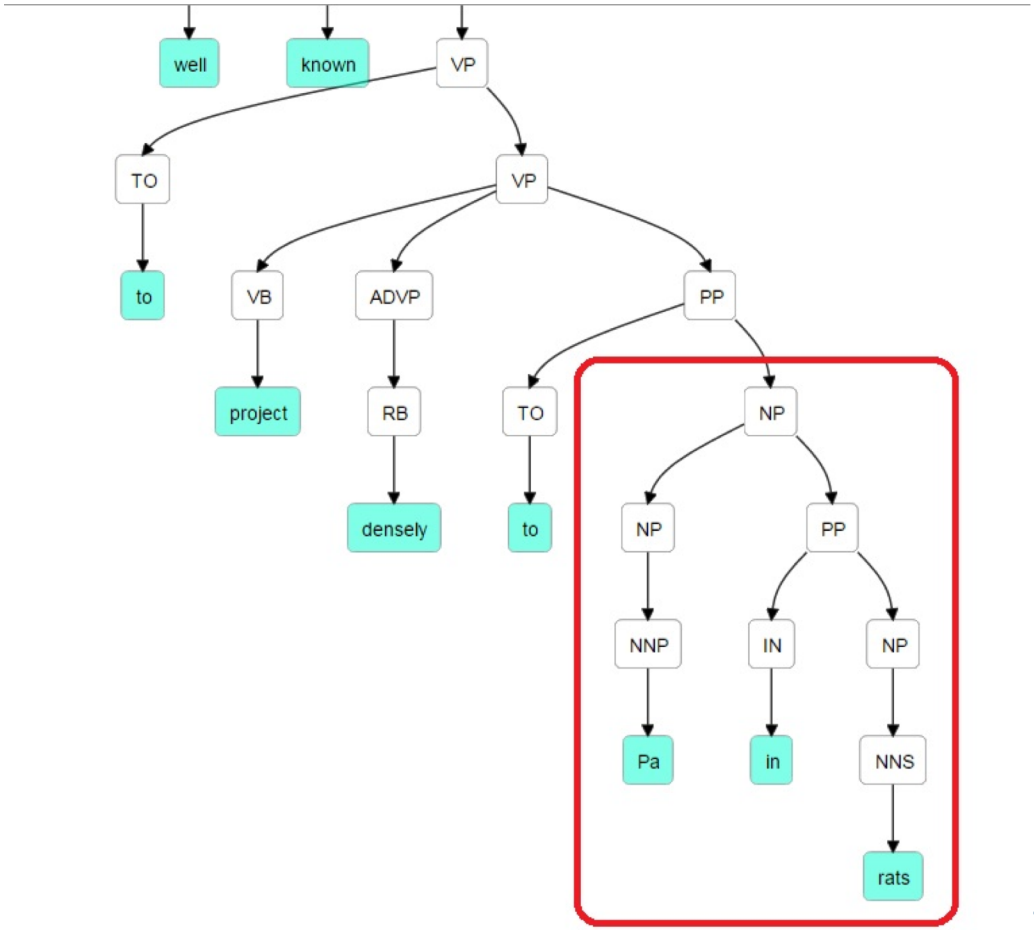
Brain Region	Acronym	Synonym
parietal lobe	PL	parietal cortex, parietal region, lobus parietalis
suprachiasmatic nucleus	SCN	suprachiasmatic nuclei
cingulate gyrus	CgG	cingular gyrus, cingulate area, cingulate region, gyri cinguli, gyrus cinguli
subthalamus	SbTh	subthalamic region, ventral thalamus, thalamus ventralis
superior frontal gyrus	SFG	marginal gyrus, superior frontal convolution, gyrus frontalis superior
parabrachial nucleus	-	parabrachial nuclei,parabrachial
paracentral nucleus	PC	paracentral thalamic nucleus, nucleus paracentralis, paracentral nucleus of the thalamus,paracentral
central medial nucleus	CM	central medial thalamic nucleus, nucleus centralis medialis, centralis medialis,central medial nucleus of the thalamus,central medial

İlişki tespiti

- Sözdizim ve bağlam ağaçları kullanarak cümlelerin gramer analizi yapıldı
  - İlişkinin yönünün tespiti
  - Etkileşen ve etkilenenlerin tespiti

Sözdizim Ağacı

"The suprachiasmatic nucleus is well known to project densely to Pa in rats"



Örnek uygulama: PVT

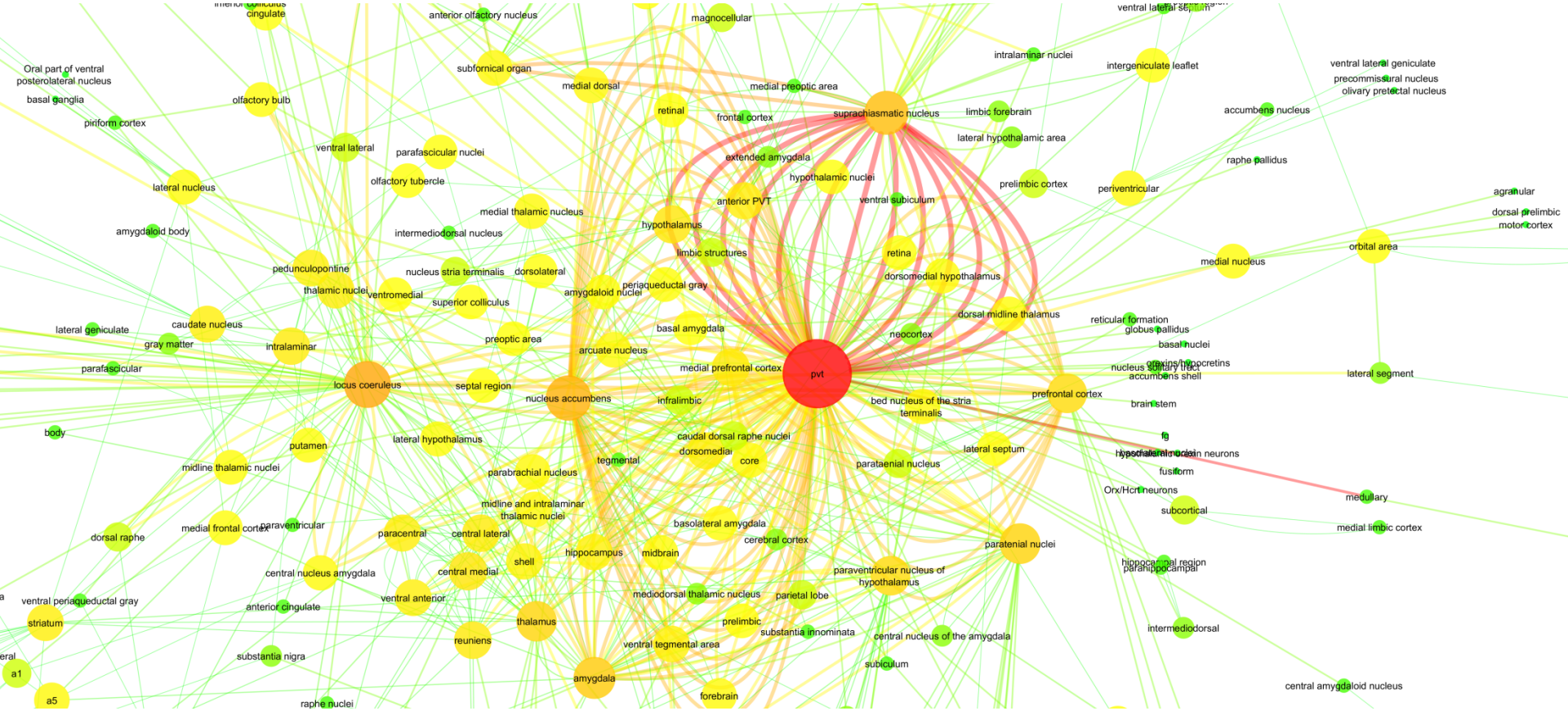
- Elle etiketlenen PVT derlemi
  - 14 tam metinli makale
  - 322 ilişki

- PubMed'deki tüm PVT ile ilgili makaleler
  - 451 özet
  - 107 tam metinli makale

PVT Derlemi Üzerinde Değerlendirme

	Precision	Recall	F-Measure
Strict (Full Match)	66.43%	33.23%	44.30%
Lenient (Full Match + Partial Match)	75.78%	37.89%	50.52%
NLP-based	87.58%	43.79%	58.39%

PVT Bağlantı Ağı



Yeni hipotez

- PVT aşağıdaki beyin bölgeleri ile kuvvetli bağlantılar kuruyor
  - SCN, nucleus accumbens, amygdaloid complex
  - extended amygdala incl.
    - bed nucleus of the stria terminalis
    - ventromedial prefrontal cortex
- Bu beyin kısımları depresyon ve ruh hali ile ilişkilendrilmiştir.
- PVT'nin de depresyon mekanizmasında önemli bir rolü olabilir.
  - Only, Zhu et al. (2011) suggests that PVT neurons might be engaged in acute depressive events

Varlık İsmi Normalizasyonu

İşbirliği: İlknur Karadeniz

Karadeniz, İ., & Özgür, A. (2015). **Detection and categorization of bacteria habitats using shallow linguistic analysis.** *BMC bioinformatics,* 16(10), S5.

Karadeniz, İ., & Özgür, A. (2019). **Linking entities through an ontology using word embeddings and syntactic re-ranking.** *BMC bioinformatics,* 20(1), 156.

Varlık İsmi Normalizasyonu

id:OBT:002307
name: **pediatric patient**
is_a: OBT:002133 ! Patient
is_a: OBT:002146 ! child

id:OBT:000164
name: **respiratory tract**
synonym: "respiratory tree"
synonym: "respiratory"
is_a: OBT:000065 ! animal part

The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice.

To develop a broad understanding of the causes and patterns of occurrence of wheezing associated respiratory infections, we analyzed data from an 11-year study of acute lower respiratory illness in a pediatric practice. Although half of the WARI occurred in children less than 2 years of age, wheezing continued to be observed in 19% of children greater than 9 years of age who had lower respiratory illness. Males experienced LRI 1.25 times more often than did females; the relative risk of males for WARI was 1.35. A nonbacterial pathogen was recovered from 21% of patients with WARI; respiratory syncytial virus, parainfluenza virus types 1 and 3, adenoviruses, and Mycoplasma pneumoniae accounted for 81% of the isolates. Patient age influenced the pattern of recovery of these agents. The most common cause of WARI in children under 5 years of age was RSV whereas Mycoplasma pneumoniae was the most frequent isolate from school age children with wheezing illness. The data expand our understanding of the causes of WARI and are useful to diagnosticians and to researchers interested in the control of lower respiratory disease.

Zorluklar

The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice.

To develop a broad understanding of the causes and patterns of occurrence of wheezing associated respiratory infections, we analyzed data from an 11-year study of acute lower respiratory illness in a pediatric practice. Although half of the WARI occurred in children less than 2 years of age, wheezing continued to be observed in 19% of children greater than 9 years of age who had lower respiratory illness. Males experienced LRI 1.25 times more often than did females; the relative risk of males for WARI was 1.35. A nonbacterial pathogen was recovered from 21% of patients with WARI; respiratory syncytial virus, parainfluenza virus types 1 and 3, adenoviruses, and Mycoplasma pneumoniae accounted for 81% of the isolates. Patient age influenced the pattern of recovery of these agents. The most common cause of WARI in children under 5 years of age was RSV whereas Mycoplasma pneumoniae was the most frequent isolate from school age children with wheezing illness. The data expand our understanding of the causes of WARI and are useful to diagnosticians and to researchers interested in the control of lower respiratory disease.

Sözcüksel benzerlik yok!

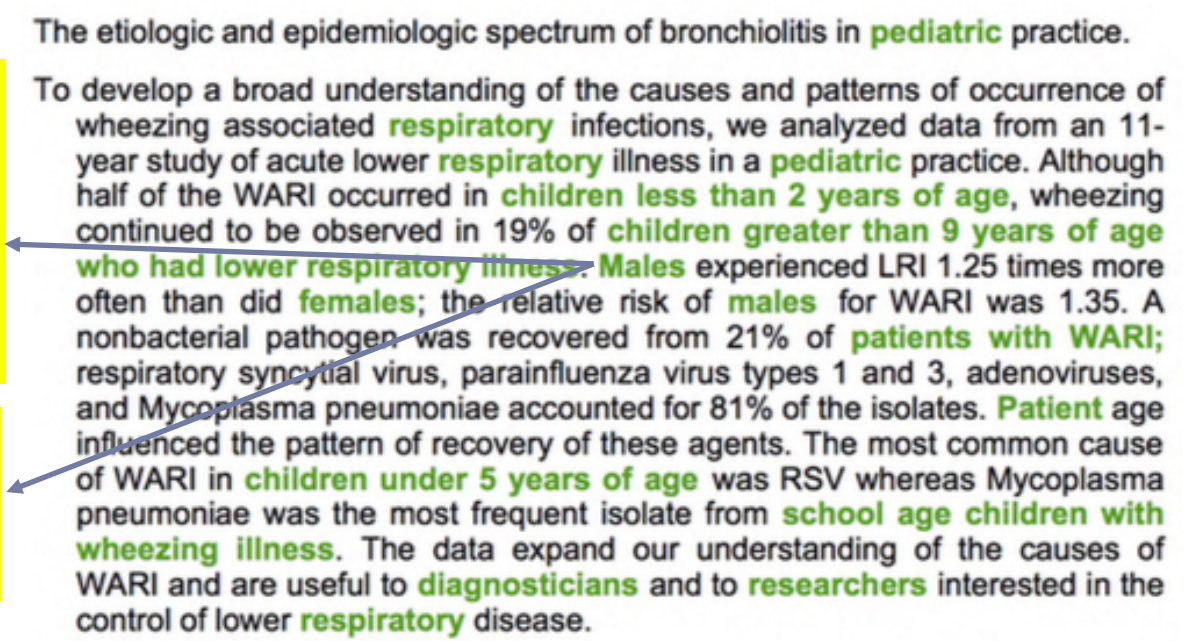
id:OBT:002307
name: **pediatric patient**
is_a: OBT:002133 ! Patient
is_a: OBT:002146 ! child

Zorluklar – Belirsizlik

Doğru kavram

id:OBT:002307
name: **pediatric patient**
is_a: OBT:002133 ! Patient
is_a: OBT:002146 ! child

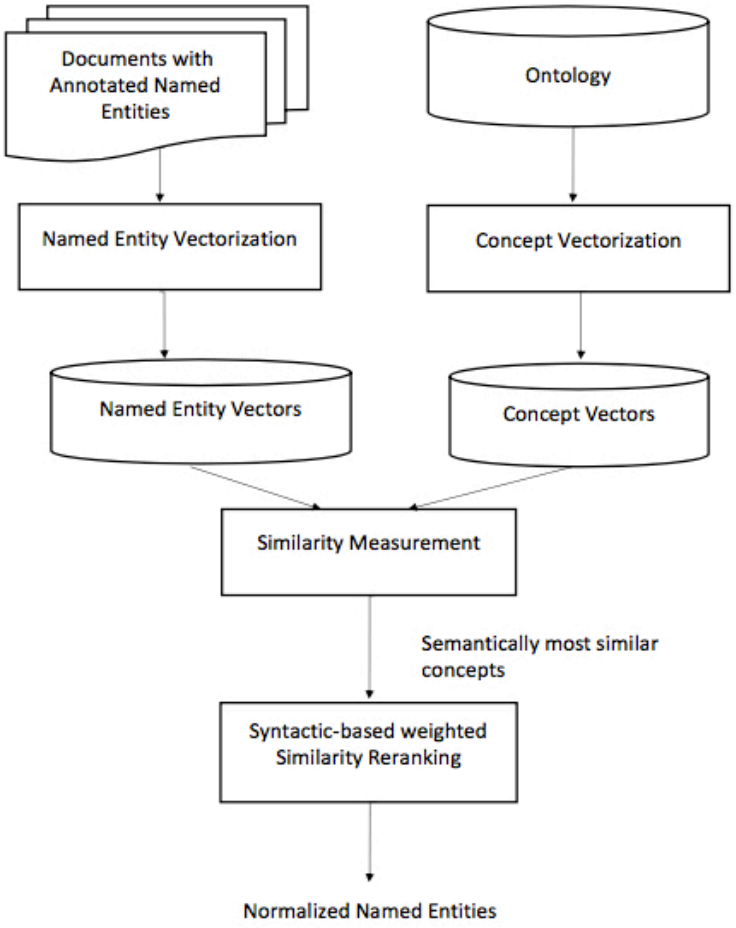
id: OBT:002167
name: **boy**
is_a: OBT:002146 ! child



Sözcüksel olarak en çok benzeyen kavram.

id: OBT:000608
name: **male animal**
synonym: "male"
is_a: OBT:000380 ! animal with age or sex property

Sistemin Genel Yapısı

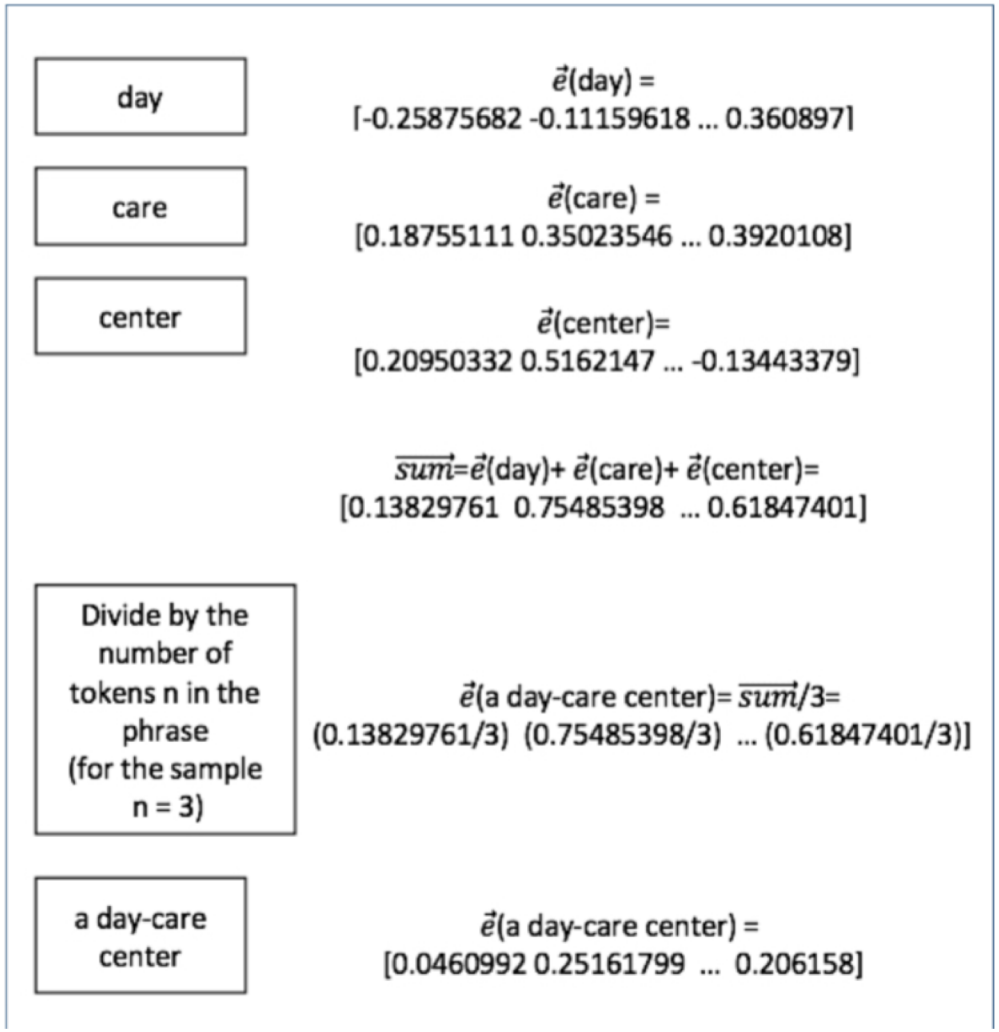


Denetimsiz bir yaklaşım

Elle etiketli veri veya alana özel sözlüklere ihtiyaç duymuyor.

Başka alanlara uyarlanabilir.

Varlık isimlerinin vektörel gösterimi



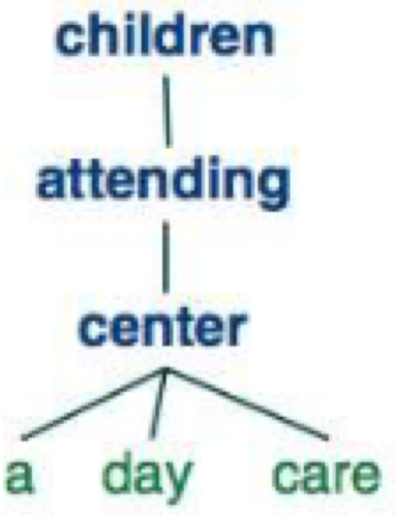
Örnek bir yanlış pozitif

Table 1. Semantically most similar concepts to the entity mention "children attending a day-care center" without reranking

Ranking	Concept	Similarity score
1	OBT:001423 medical center	0.8297
2	OBT:001801 clinic	0.7917
28	OBT:002146 child	0.6979

Yeniden sıralama

- Cümlelerin sözdizimsel analizi kullanıldı
- Aday varlık isminin baş kelimesinin (headword) bulunması
- Aday kavramın baş kelimesinin bulunması
- Baş kelimeler arasındaki benzerliğe ağırlık verilmesi



$$S_{RR}(m,c) = (w * S_S(m_{head}, c_{head})) + ((1-w)*S_S(m,c))$$

Yeniden sıralamanın etkisi

Table 1. Semantically most similar concepts to the entity mention "children attending a day-care center" without reranking

Ranking	Concept	Similarity score
1	OBT:001423 medical center	0.8297
2	OBT:001801 clinic	0.7917
28	OBT:002146 child	0.6979

Table 2. Semantically most similar concepts to the entity mention "children attending a day-care center" after syntactic reranking

Ranking	Concept	Similarity score
1	OBT:002146 child	0.7484
3	OBT:001801 clinic	0.6519
24	OBT:001423 medical center	0.5460

System	Train	Dev
Before Re-ranking	0.601	0.629
After Re-ranking	0.648	0.677

Sonuçlar

System	Precision
BOUNEL	**0.659**
TURKU	0.630
BOUN	0.620
CONTES	0.597
LIMSI	0.438

Biyomedikal alanda cümleler arası semantik benzerlik hesaplama

İşbirliği: Gizem Soğancıoğlu ve Hakime Öztürk

G. Soğancıoğlu, H. Öztürk, A. Özgür. "**BIOSSES: A Semantic Sentence Similarity Estimation System for the Biomedical Domain**", Bioinformatics, 2017.

Biyomedikal alandan örnek

Semantically Highly Similar Sentence Pairs

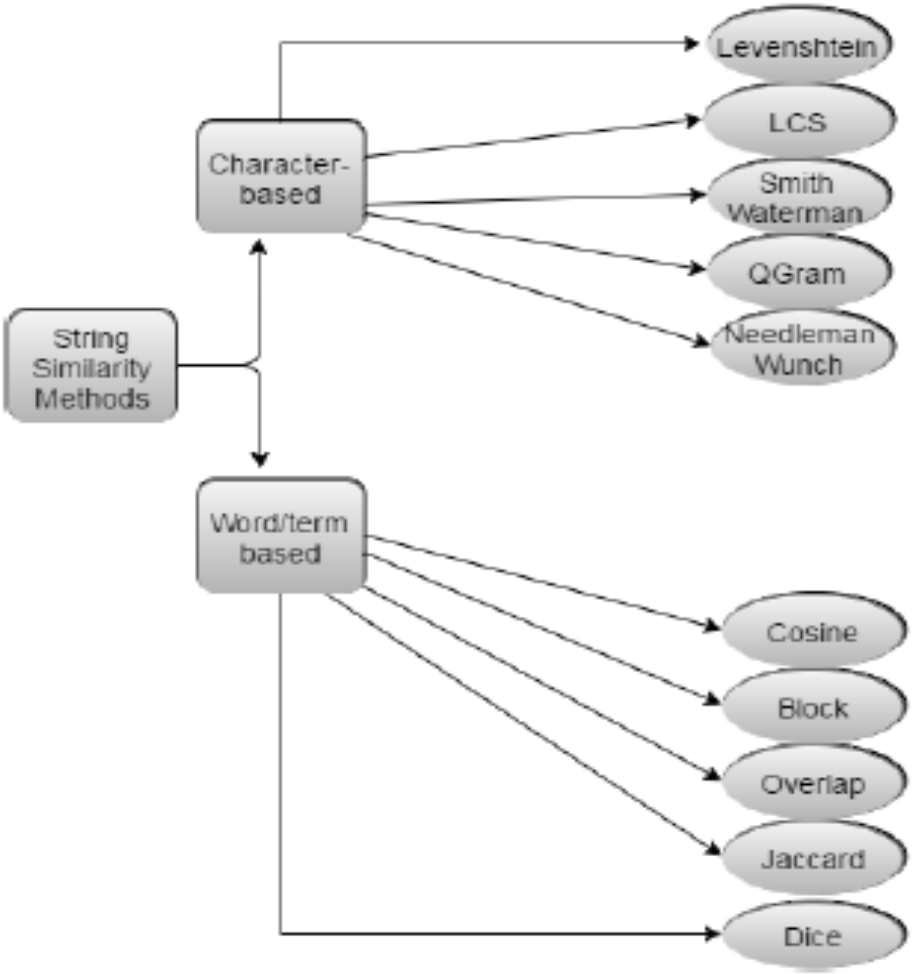
- S1: This form of necrosis, also termed necroptosis, requires the activity of receptor-interacting protein kinase 1 and its related kinase 3 .
- S2: Moreover, other reports have also shown that necroptosis could be induced via modulating RIP1 and RIP3.

Amaç:
Semantik cümle benzerliği hesaplamak için bir sistemin geliştirilmesi.

Benzerlik Yöntemleri

- Karakter dizisi benzerliği
- Dağıtık temsil vektörü benzerliği
- Ontoloji tabanlı benzerlik
- Denetimli makine öğrenmesiyle farklı benzerlik ölçütlerinin birleştirilmesi

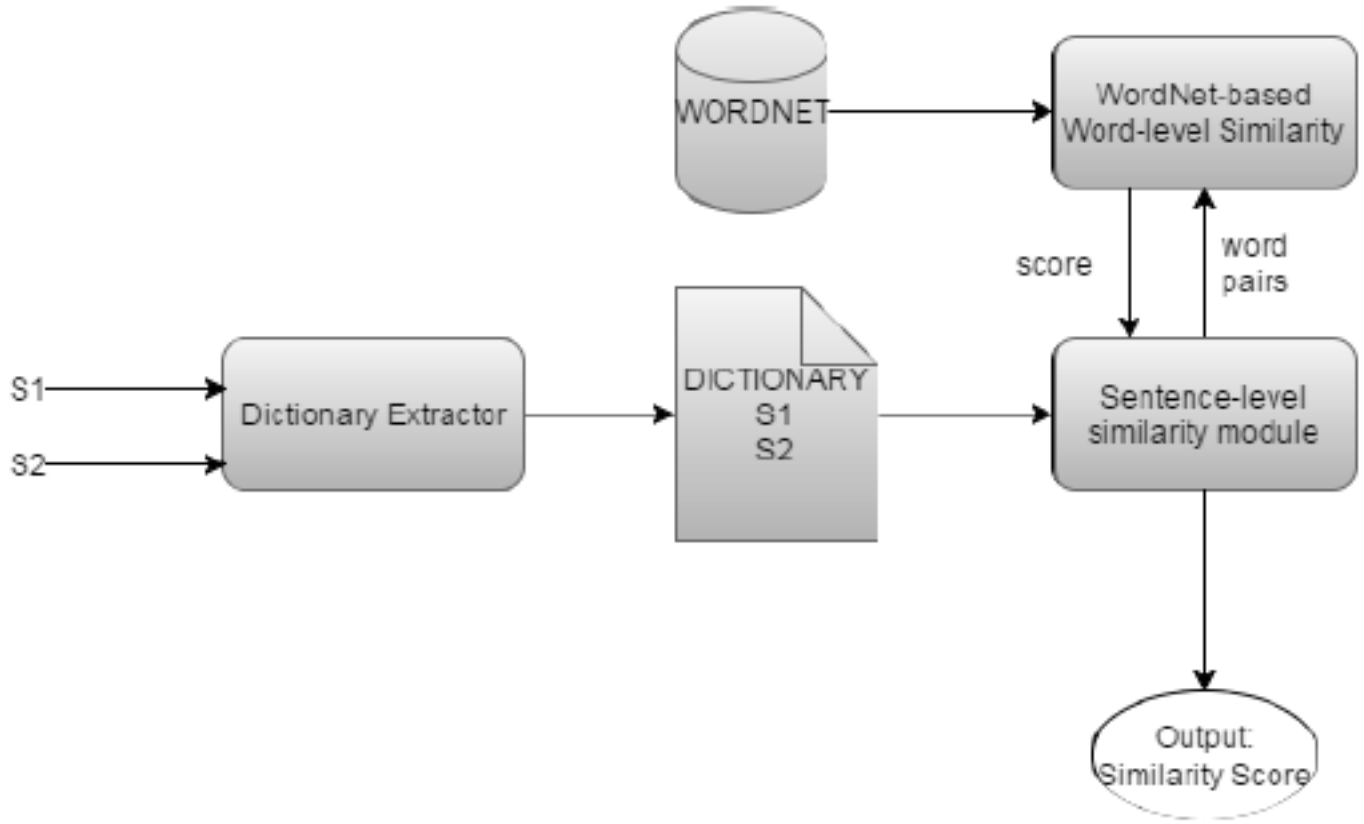
Karakter dizisi benzerliği



Ontoloji tabanlı yöntemler

- WordNet
- UMLS
- Birleştirlmiş Benzerlik Ölçütü

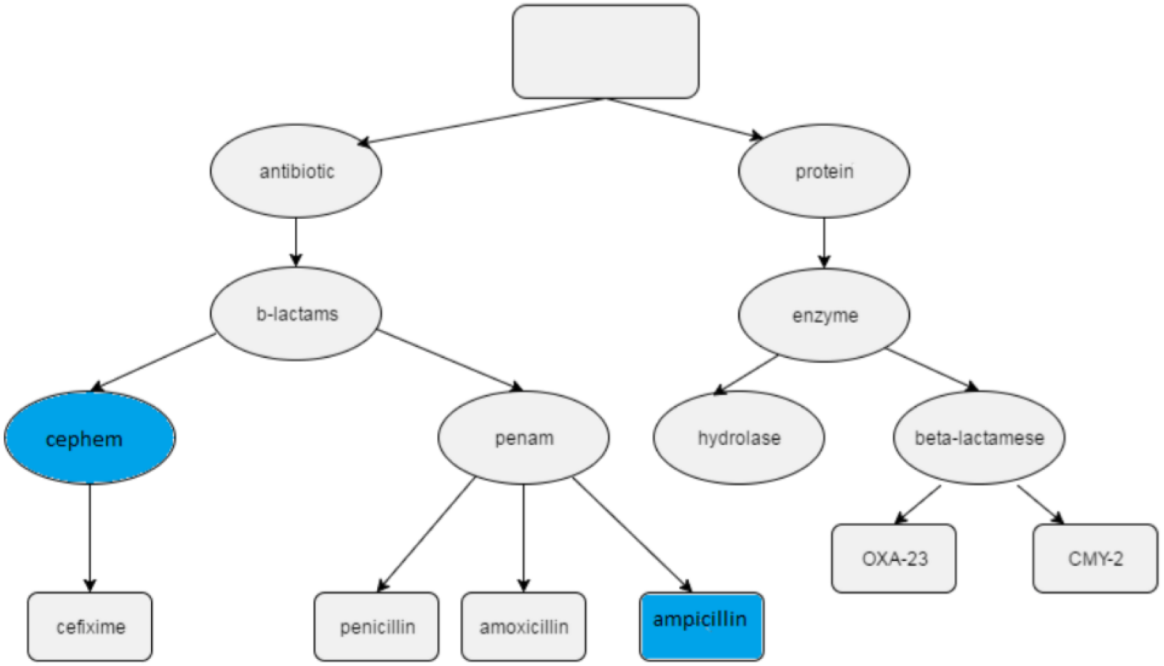
Ontoloji-tabanlı benzerlik



Örnek: Ontoloji-tabanlı benzerlik

$$\texttt{Sim\_Path}(c_1, c_2) = 2 * maxDepth - \texttt{len}(c_1, c_2)$$

$$\texttt{Sim\_Path}(cephem, ampicillin) = 10 - 4 = 6$$



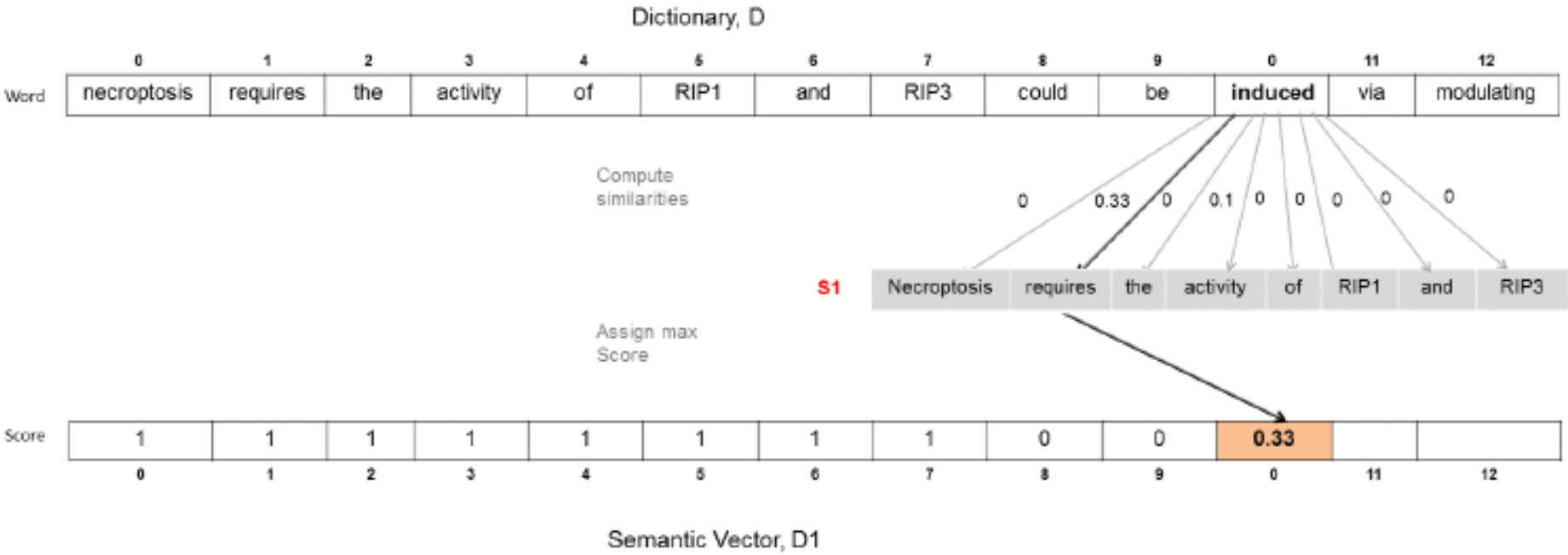
Örnek

S_1 = Necroptosis requires the activity of RIP1 and RIP3.
D_1 = [1,1,1,1,1,1,1,1,0,0,0.33,0,0.1]
S_2 = Necroptosis could be induced via modulating RIP1 and RIP3.
D_2 = [1,0.33,0,0.1,0,1,1,1,1,1,1,1,1]

Standart:
Kosinüs benzerliği(S1, S2)= 0.47

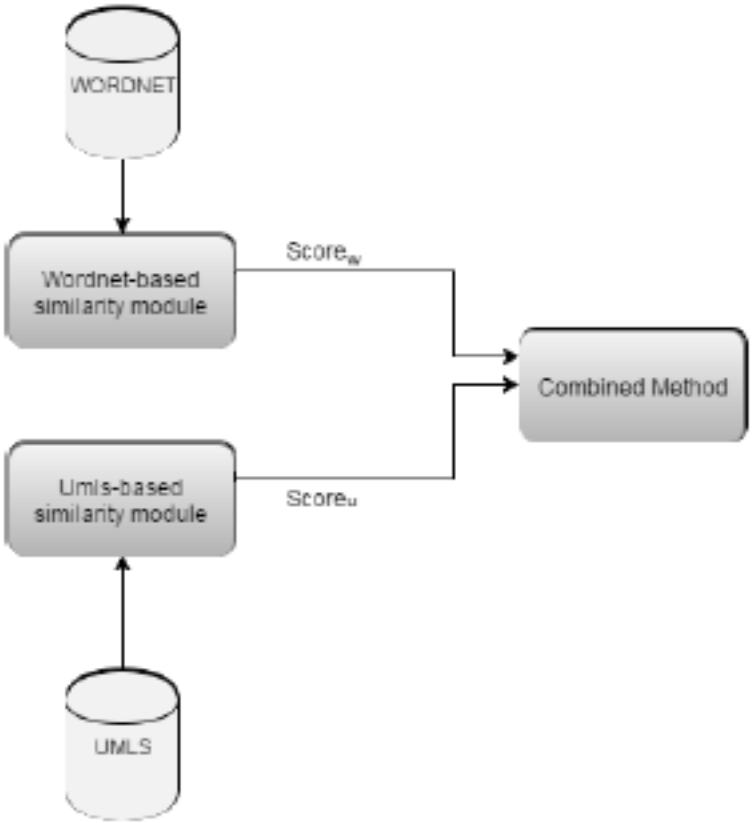
Ontoloji-tabanlı:
Kosinüs benzerliği(D1, D2) = 0.56

Örnek

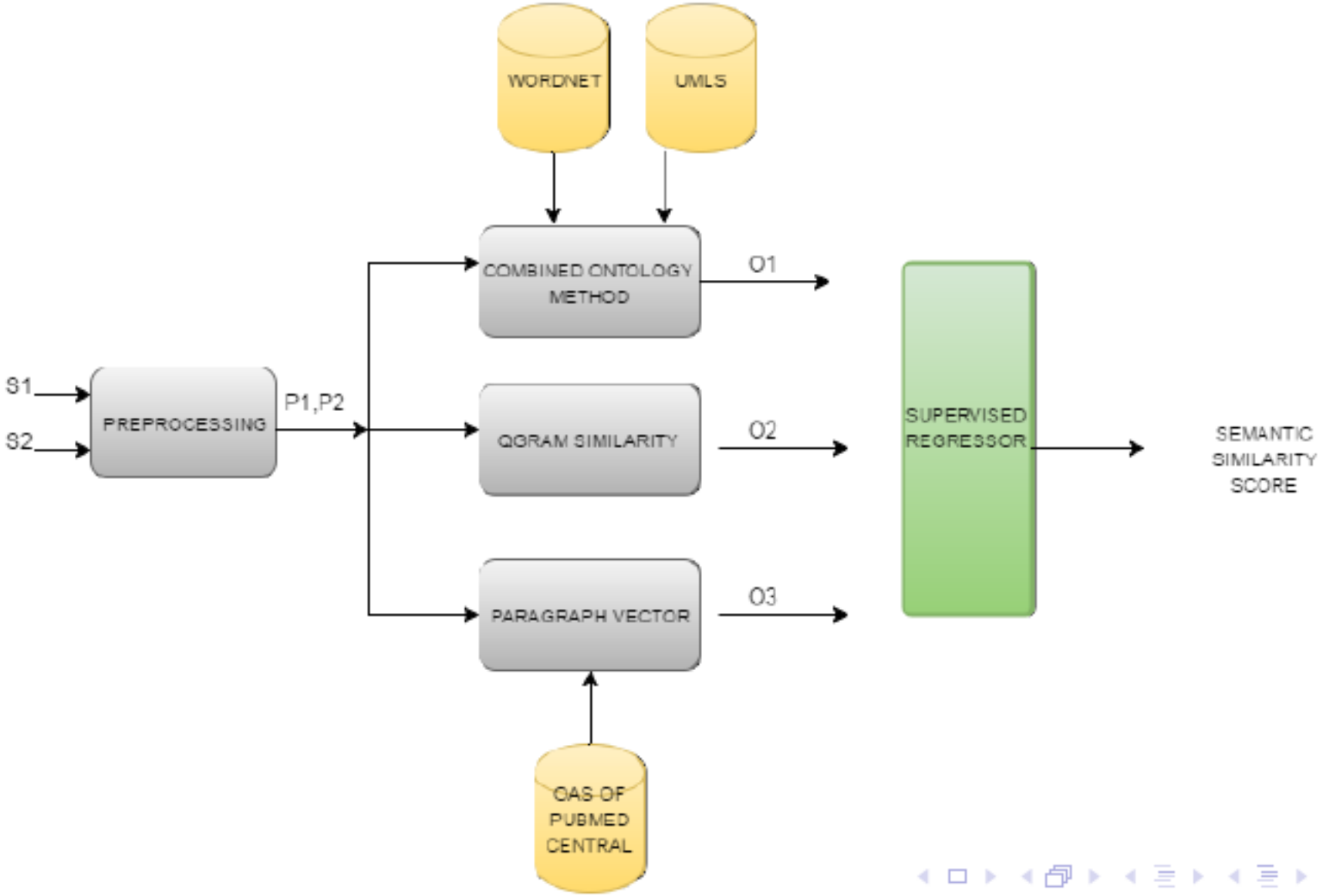

Dictionary, D
0 1 2 3 4 5 6 7 8 9 0 11 12
Word
necroptosis
requires
the
activity
of
RIP1
and
RIP3
could
be
induced
via
modulating
Compute similarities
0
0.33
0
0.1
0
0
0
0
0
S1
Necroptosis
requires
the
activity
of
RIP1
and
RIP3
Assign max Score
Score
1 1 1 1 1 1 1 1 0 0 0.33
0 1 2 3 4 5 6 7 8 9 0 11 12
Semantic Vector, D1

Birleştirilmiş skor

$$combinedscore = Score_{WordNet}.\lambda + Score_{UMLS}.(1 - \lambda)$$



Denetimli Yaklaşım



Veri kümesi etiketlenmesi

Annotation Score	Definition
0	The two sentences are on different topics.
1	The two sentences are not equivalent, but are on the same topic.
2	The two sentences are not equivalent, but share some details.
3	The two sentences are roughly equivalent, but some important information differs/missing.
4	The two sentences are completely or mostly equivalent, as they mean the same thing.

Cümle seçimi

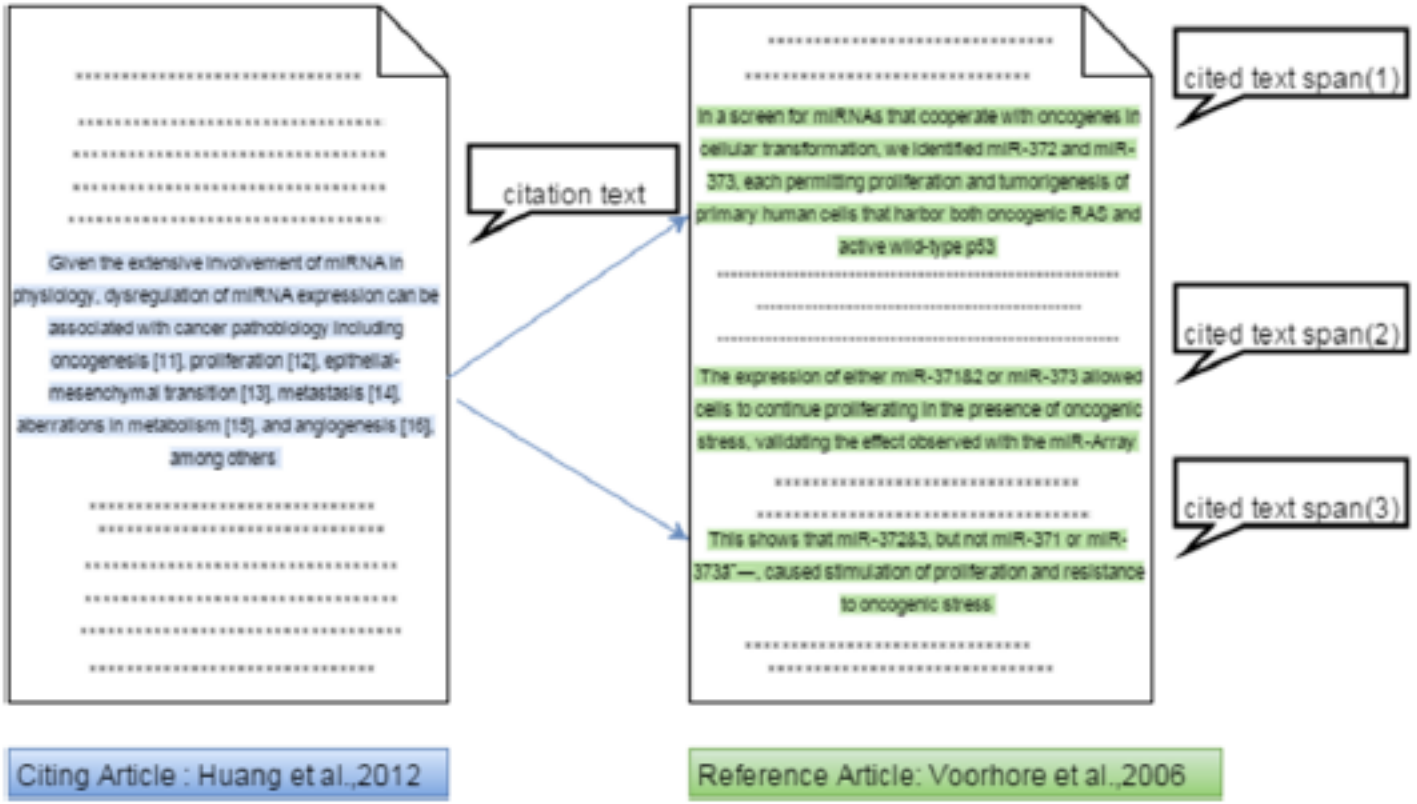


Figure: Example Annotation from TAC^2 training data set

Pair_ID|*Sentence1*|*Sentence2*|*Similarity_Score*|*Annotator_ID*

Sonuçlar

Table: Correlation scores among annotators

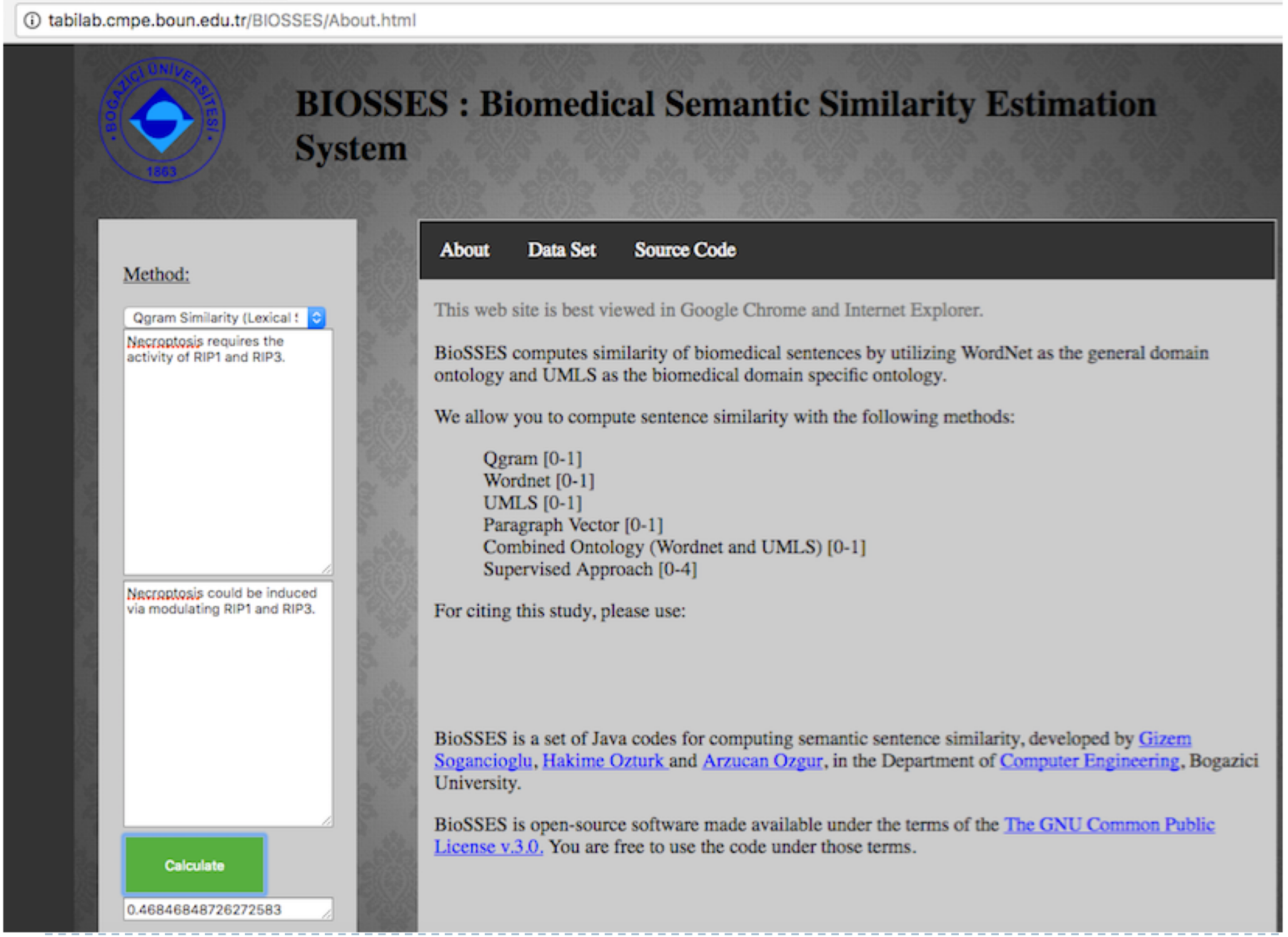
	Correlation r
Annotator A	0.952
Annotator B	0.958
Annotator C	0.917
Annotator D	0.902
Annotator E	0.941

Table: Correlation scores for domain-independent state-of-the-art systems

	Correlation r
ADW	0.586
SEMILAR	0.419

Sonuçlar

Methods	Pearson Correlation
Domain-independent Systems	
ADW	**0.586**
SEMILAR	0.419
String Similarity Measures	
Qgram	**0.754**
Word Embeddings based Similarity	
Paragraph Vector	**0.787**
Ontology-based Similarity	
WordNet based Similarity Module-Path	**0.644**
UMLS based Similarity Module-Path	**0.651**
Combined Ontology Method($[\lambda = 0.5]$)	**0.710**
Supervised Model - Linear Regression	**0.836**


tabilab.cmpe.boun.edu.tr/BIOSSES/About.html
BIOSSES : Biomedical Semantic Similarity Estimation System
Method:
Qgram Similarity (Lexical !
Necroptosis requires the activity of RIP1 and RIP3.
Necroptosis could be induced via modulating RIP1 and RIP3.
Calculate
0.46846848726272583
About
Data Set
Source Code
This web site is best viewed in Google Chrome and Internet Explorer.
BioSSES computes similarity of biomedical sentences by utilizing WordNet as the general domain ontology and UMLS as the biomedical domain specific ontology.
We allow you to compute sentence similarity with the following methods:
Qgram [0-1]
Wordnet [0-1]
UMLS [0-1]
Paragraph Vector [0-1]
Combined Ontology (Wordnet and UMLS) [0-1]
Supervised Approach [0-4]
For citing this study, please use:
BioSSES is a set of Java codes for computing semantic sentence similarity, developed by Gizem Soganciogl, Hakime Ozturk and Arzucan Ozgur, in the Department of Computer Engineering, Bogazici University.
BioSSES is open-source software made available under the terms of the The GNU Common Public License v.3.0. You are free to use the code under those terms.

Teşekkür

Öğrencilerim

İlknur Karadeniz (Işık Ü. Öğretim Üyesi)
Hakime Öztürk
N. Özlem Özcan Şimşek
Erinç Gökdeniz
Ferhat Aydın
Zehra Melce Hüsünbeyi
Gizem Soğancıoğlu

İşbirliği Yaptığımız Araştırmacılar

Elif Özkırımlı (Boğaziçi Ü.)
Reşit Canbeyli (Boğaziçi Ü.)
Fikret Gürgen (Boğaziçi Ü.)
Kutlu Ülgen (Boğaziçi Ü.)
Yongqun Oliver He (University of Michigan)
Junguk Hur (University of North Dakota)
Dragomir Radev (Yale University)

Destekler

EU Marie Curie Career Integration Grant
Bilim Akademisi Genç Bilim İnsanları Ödül Programı (BAGEP)
TUBITAK-BIDEB 2211
Boğaziçi Ü. BAP (No 12304)
Boğaziçi Ü. BAP (No D-13242)
DPT TAM Projesi (No 2007K120610)

Teşekkürler!