지원 포지션 : Data Scientist

이기욱 1987 년 (35 세) | 남성 | 기혼

leek0502@gmail.com 010-4220-9622

linkedin.com/in/기욱-이-7793807b https://bimlkw.medium.com/

경기 광명시 도덕공원로 35

전문 분야	기계학습, 딥러닝, 유전체, 유전자가위
총 경력	4 년 11 개월, 컴퓨터 기술 기반 바이오 데이터 분석
현재 연봉	7,500 만원
희망 연봉	협의 후 결정

경력

기간	회사명 / 부서명 / 직위(직책)	담당 업무
2019.07 ~ 현재	지플러스 생명과학 / BI,AI 팀 / 선임 연구원 (팀장)	- 딥러닝을 이용한 유전자 가위 관련 모델 개발 - 기계학습을 이용한 스마트팜 데이터 분석
2016.07 ~ 2019.07 (3 년 1 개월)	삼성 유전체 연구소 / QC,Sequencing 팀 / 연구원	- 기계학습을 이용한 유전체 데이터 분석 - 유전체 분석 파이프라인 개발

학력

재학기간	구분	학교명(소재지)	전공	학점
2017.03 ~ 2019.08	수료	성균관대학교 일반대학원 (박사) (서울)	융합의과학원	4 / 4.5
	졸업	숭실대학교 일반대학원 (석사) (서울)	인공지능	4 / 4.5

2014.03 ~ 2016.08		논문&졸업작품 기계학습을 이용한 신규변이 필터링 기법		
2005.09 ~ 2009.06	졸업	숭실대학교 (학사) (서울)	컴퓨터공학	3.1 / 4.5
		논문&졸업작품 - 영상 변환을 이용한 보안 감시 시스템 및 그 방법 (출원번호: 10-2013-0117905) - 영상 변환을 이용한 보안 감시 시스템의 설계 및 구현(한국정보과학회 2013.11, 781-183)		

핵심 역량

- 기계학습/딥러닝 기반 유전자 가위 효율 예측 모델 개발
- 기계학습을 이용한 유전체 분석 모델 개발
- DNA/RNA/Epigenome 등의 각종 Omics 분석 파이프라인 구축 및 분석
- C/R/Python/Perl 을 이용한 Back-end, Front-end 개발자

논문 리스트

년도	저자/제목/저널
2021 (manuscript correction)	**KW Lee***, DG Won, Crispr/Cas efficiency prediction using transfer learning.
2019	HT Shin, NY Kim, JW Yun, BR Lee, SK Kyung, **KW Lee**, DE Ryu, JH Kim, JS Bae, DH Park, YL Choi, SH Lee, MJ Ahn, Keunchill Park, WY Park*,"Junction Location Identifier (JuLI) Accurate Detection of DNA Fusions in Clinical Sequencing for Precision Oncology", ***Journal of Molecular Diagnostics***, vol. 22, Issue 3, pp. 304-318, 2019.
	DH Seong, JS Chung, **KW Lee**, SY Kim, BS Kim, JK Song, SW Jung, TS Lee, DH Park BK Yi*, WY Park* DS Son*, "Benchmark Database for Process Optimization and Quality Control of Clinical Cancer Panel Sequencing", ***Biotechnology and Bioprocess Engineering***, vol 24, pp. 793-798, 2019
	JS Chung, **KW Lee**, C Lee, SH Shin, SK Kyung, HJ Jeon, SY Kim, EJ Cho, CE Yoo, DS Son, WY Park*, DH Park*, "Performance evaluation of commercial library construction kits for PCR-

	based targeted sequencing using a unique molecular identifier", ***BMC Genomics***, vol. 20, Article number : 216, 2019
	JH Cho, SM Ahn, DS Son, NY Kim, **KW Lee**, ST Kim, JY Lee, SH Park, JO Park, JY An, MG Choi, JH Lee, TS Sohn, JM Bae, S Kim, KM Kim*, "Bridging genomics and phenomics of gastric carcinoma", ***Cancer Genetics and Epigenetics***, Vol. 145, Issue 9, p. 2407-2417, 2019
2018	HJ Lee, **KW Lee (co-first author)**, TS Lee, DH Park, JS Chung, C Lee, WY Park, DS Son*, "Performance evaluation method for read mapping tool in clinical panel sequencing", ***Genes & Genomics***, Vol. 40, 189-197, 2018

상세 경력기술서

(1) 지플러스 생명과학 / BI,AI팀 / 선임 연구원/ 근무기간: 2019.07 ~ 현재까지

1. 딥러닝 기반 유전자 가위 예측 도구 개발

- **목 적**
  - **Issue 1. Crispr/Cas12 데이터의 부재**
    - Essential gene 을 확인하기 위해, Crispr/Cas9 를 이용한 screening 프로젝트가 Broad, Sanger 를 주축으로, 많은 연구가 있었고, 그로 인하여 Crispr/Cas9 에 대한 많은 연구와 함께, 예측 모델을 만들기 위한 충분한 양의 데이터가 공개되어 있음.
    - 그렇지만, 상대적으로 절단 효율이 떨어지는 Crispr/Cas12 의 경우, 공개 데이터가 미비한 상태이며, 이로 인하여 정확한 예측 모델을 만들기 쉽지 않은 실정임.
  - **Issue 2. Cell line 에 따른 절단 효율의 편차 존재.**
    - 유전자 가위의 학습 모델을 어떤 Cell line 을 기준으로 만들지에 따라, 절단 효율에 대한 예측값에 큰 편차가 존재.
    - 따라서, 특정 Cell line 에 대한 절단 효율을 예측하고자 하면, Cell line 별 데이터와 이에 대한 모델 개발이 필요한 실정임.
- **방 법**
  - **특징 1 : Transfer learning** 을 이용, 1) Cas9 모델 → Cas12 전이. 2) Cell line 전이
  - **특징 2 : CNN-RNN 통합 프레임 워크**를 이용한 유전자 가위 예측 모델 개발. CNN 은 weight-sharing 전략을 사용하여, local pattern 을 capturing 하는데 우수한 반면, sequential correlation 을 학습하는데는 좋지 않음. 반대로, RNN 은 sequential correlation 을 학습하는데는 우수한 성능을 보이는 반면, parallel 하게 feature 를

derive 하는데는 불리함. CNN 과 RNN 의 장점을 살려, Convolution module stage 에서 1D convolution filter 를 사용, 시퀀스 패턴을 capture, 스캔하고, RNN 단계는 motif 간의 방향과 공간적 관계를 고려하여 복잡한 high-level 관계를 학습하는데 사용.

- **개발언어**

  - Pythorn, PyTorch

(2) 삼성유전체 연구소 / QC, Sequencing 팀 / 연구원/ 근무기간: 2016.08 ~ 2019.05

1. 유전자 패널 검사에서 복제수 변이 탐지를 위한 성능 비교 평가 및 검출기 개발

- **목적**

- 유전자 패널 검사에서 적합한 복제수 변이 탐지 도구 개발

- **작업내용**

1. 복제수 변이 검출기의 성능 평가를 위한 참조 데이터 확보

Breast cancer cell line 데이터를 serial dillution 을 진행한 후, Purity 별로, Weighted univariate k-means clustering 을 진행하여 Neutral 과 Mosaicism, Amplification, Deletion 을 구분함. 이 데이터를 참조 데이터를 기반으로 복제수 변이 검출기 성능 평가를 진행함.

2. 암 패널 시퀀싱에 적합한 CNV 검출기 성능 비교 평가

계산된 참조 데이터를 기반으로 8 개의 복제수 변이 검출기의 성능을 비교하여 암 패널 시퀀싱에 적합한 CNV 검출기 선정함.

3. 패널 시퀀싱용 CNV 검출기 개발

선정된 복제수 변이 검출기의 성능의 Read depth normalization 과 Tumor purity estimation 을 부정확하게 추정하는 문제를 발견하였고, 이러한 문제를 개선하고자, Normalization 은 loess regression 대신 haar wavelet 방법을 통해, 성능을 개선하였고, Tumor purity estimation 은 Gaussian mixture 를 도입하여, extreme low/high purity 에서 정확한 예측을 가능하도록 알고리즘을 개발함.

- **개발언어 :** C/R Language

2. Alignment 알고리즘 성능 평가 및 NGS 파이프라인 개발

- **목적**
  - 유전자 패널 검사에서 적합한 Alignment 알고리즘에 대한 성능 평가

- **작업내용**

  1. Alignment 성능 측정을 위한 시뮬레이션 데이터 생성

  참조 유전체 데이터를 기반으로, SNV, InDel 을 삽입 가능한 시뮬레이터를 개발하여, 이를 이용, 다양한 조건의 변이 시뮬레이션 데이터를 생성함.

  2. 시뮬레이션 데이터를 기반으로 Alignment 성능 비교 평가

  생성된 데이터를 기반으로, 변이 개수, 변이 길이 변화에 따라 Alignment 의 성능을 비교하였고, 그 결과 BWA-mem 이 가장 좋은 성능을 보임.

  3. 패널 시퀀싱에 적합한 파이프라인 작성

  성능 비교 결과를 참고하여, 패널 시퀀싱에 적합하도록, NGS 파이프라인을 개발함.

- **개발언어 :** Shell scripts / C Language
- **참고**

  https://link.springer.com/article/10.1007/s13258-017-0621-9

3. UMI 를 사용한 library construction kit 성능 평가

- **목적**
  - PCR artifacts 를 효과적으로 제거하기 위해 UMI 를 사용한 commercial library construction kit 의 성능 평가

- **작업내용**

  1. Commercial library construction kit 성능 평가

  Archer® Reveal ctDNA™ 28 Kit, NEBNext Direct® Cancer HotSpot Panel, Nugen Ovation® Custom Target Enrichment System, Qiagen Human Comprehensive Cancer Panel(HCCP), Qiagen Human Actionable Solid Tumor Panel(HASTP) 성능 비교

  2. 다양한 측면에서의 성능 비교 측정

  UMI 길이, DNA 양, Mutation frequency 변화에 따라, PCR artifact 가 가장 효과적으로 제거가되는 Commercial library construction kit 별 성능을 비교함.

- **개발언어 :** Shell scripts

- **참고**

  https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5583-7

4. QC benchmarking database 개발

- **목적**
  - NGS 기반 임상진단기술 개발 및 핵심기술에 대한 성능 분석 결과와 데이터를 제공하는 벤치마킹 데이터베이스 개발 및 운영

- **작업내용**

1. NIST 표준물질을 이용한 생식세포 변이 검출 평가 자동화 시스템 개발

사용자의 검출 범위 파일(Browser Extensible Data, BED)과 변이 검출 결과(Variant calling format, VCF)를 입력받음으로써 Panel sequencing 및 whole exome sequencing 비교 결과 제공함.

2. 정도관리 지표 리포트 및 시퀀싱 rawdata DB 구축

NGS 임상검사의 단계별 QC measure 에 대한 평가 자료 제공 및 핵심 지표에 대한 기존에 생산된 NGS 데이터를 이용한 실험적 검증 자료 제공함.

3. 임상적 방법으로 확인된 변이를 가진 임상시료의 시퀀싱 rawdata DB 구축

FISH/IHC, Cobas 와 같은 임상적 방법으로 확인된 변이 정보 및 실험과정에 포함된 rawdata 와 결과 파일 제공함.

- **개발언어 :** Perl
- **참고**

  https://link.springer.com/article/10.1007/s12257-019-0202-7

5. SV 시뮬레이터 개발 및 caller 성능 평가

- **목적**
  - Target sequencing 에 적합한 SV caller 성능 비교 및 성능 분석에 사용하기 위한 SV 시뮬레이터 개발

- **작업내용**

1.SV 시뮬레이터 개발

SV caller 의 성능 평가를 위해, Variant, Tumor purity, Mean coverage, Variant length 를 조건으로 하는 SV 시뮬레이션 파이프라인 개발

2. SV caller 성능 평가

연구소에서 개발한 SV caller 인 JuLI 와 SvABA, Delly, Manta, LUMPY, novoBreak 에 대하여 생성한 시뮬레이션 데이터를 이용하여 SV 탐지 성능을 평가함.

- **개발언어 :** R/Shell script
- **참고**

  https://linkinghub.elsevier.com/retrieve/pii/S1525157819304568

6. 위암 데이터 분석

- **목적**
  - 위암의 임상 결과와 병리학적 결과를 비교하기 위해, 330 개 샘플의 381 개 암 관련 유전자를 분석하여, 어느 유전자가 유의미한지 확인함.

- **작업내용**

  1. 330개의 위암 샘플 시퀀싱 작업 진행.
  2. 통계분석을 통한 위암 관련 유전자 및 유의성 확인함.

- **개발언어 :** R Language
- **참고**

  https://www.ncbi.nlm.nih.gov/pubmed/30801717

(3) 숭실대학교 / 기계학습 연구실 / 석사 과정 / 근무기간: 2014.03 ~ 2016.08

1. Annotation tool 성능 향상

- **목적**

  연구실에서 개발한 annotation 툴인 gSearch 및 gNOME 의 속도 향상.

- **작업 내용:**

숭실대학교 기계학습 연구실에서 개발한 annotation 소프트웨어 도구인 gSearch 와 질병 관련 pathway 를 추적할 수 있는 도구인 gNOME 은 그 성능면에서 속도가 느리다는 단점이 존재함.

따라서 속도를 향상시키는 작업을 진행함.

1. **문제의 파악**

   해당 도구를 분석해본 결과 느린 속도는 사용하는 vcf 형태의 데이터베이스에 불필요한 저장공간이 많아 데이터베이스를 읽어 들이는 작업이 많아 생기는 문제로 파악함.

2. **해쉬 테이블 생성 및 binary 화**

   따라서 중복되는 항목을 최소화하기 위해 동일하게 반복되는 패턴을 해쉬테이블에 저장 및 binary 화 하여 데이터베이스의 용량을 1/5 수준으로 줄임.

3. **결과**

   데이터 binary 화로 데이터베이스의 읽기 시간을 단축하였으며, 이에 병렬처리 작업을 개선하여 기존 4 시간의 수행시간을 20 분으로 단축시킴.

- **개발언어 :** C Language
- **참고**

gSearch: http://ml.ssu.ac.kr/gSearch/index.html

gNOME: http://gnome.tchlab.org (http://220.70.0.234:8080)

2. 생존율 분석 도구 iSURV 개발

- **목적**

기존에는 사용자의 mRNA 데이터를 통한 암 환자의 생존율 분석 도구가 없으므로 이를 개발함.

- **작업과정**

1. 사용자의 mRNA 데이터와 TCGA 의 데이터를 병합 후, normalization 을 진행함.
2. mRNA 의 발현 분포를 바탕으로 PCA 를 진행하여 샘플과 TCGA 환자, 정상인과의 연관 정도 파악.
3. TCGA 데이터와 비교하여 생존율을 분석 후 각종 차트로 표현.

- **결과**

사용자의 mRNA 발현 데이터를 입력하면, TCGA 데이터를 사용하여 PCA(principal components analysis)를 진행하여 사용자의 샘플이 알려져 있는 cancer 결과와 얼마나 유사한지를 알려줌.

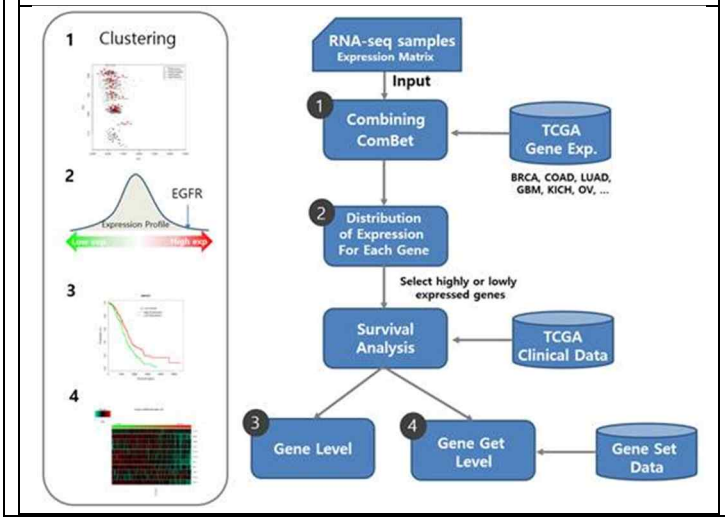
사용자의 샘플정보에서 cancer 와 관련된 gene 의 expression 정보 및 KM-plot 등을 제공함으로써 생존분석을 가능하게끔 도와줌.

- **개발언어 :** Perl / R Language

- **참고**

http://220.70.0.232:8080/

iSURV workflow



3. *de novo* mutation 데이터 분석

- **목적**

서울대학교 병원의 신경질환 소아청소년 유전체 데이터 샘플 분석.

- **작업내용**

1. 병의 원인이 되는 *de novo* mutation 후보들을 찾기 위해 annotation 정보와 *de novo* mutation 탐지 도구인 PolyMutt 을 이용함.
2. PolyMutt 의 결과에는 false positive 가 상당수 포함되어있어, 이를 줄이기 위해 기계학습 방법을 연구함.

- **결과**

소아 청소년의 신경질환을 유발하는 *de novo* mutation 후보 목록을 작성.

4. *de novo* mutation 필터링 – 석사논문 주제

- **목적**

서울대병원과 진행한 소아 청소년의 신경질환에 연관 있는 *de novo* mutation 을 찾는 연구에서 false positive 를 줄이는 방법이 필요하다고 느껴 석사 논문 주제로 정하여 연구를 시작.

- **작업과정**

1. **기존 도구들의 false positive 비율 확인**
   de novo mutation 을 탐지하는 기존 도구인 PolyMutt, DeNovoGear, TrioDenovo 의 false positive 비율이 90%이상임을 확인.
2. **기계학습을 진행하기 위한 자질(feature) 확인**
   NGS pipeline 을 진행하며 얻는 여러 정보들 중 *de novo* mutation 과 연관이 큰 정보들을 이후 기계학습에 사용하기 위하여 선택.
3. **적절한 기계학습 방법 선택**
   de novo mutation 을 분류하기 위해서 기계학습 방법들을 비교하여 Logistic regression 을 선택함.
4. **여러 데이터에 검증**
   European 데이터와 African 데이터를 사용하여 새로 만든 분류기법을 검증함.

- **개발언어 :** R Language

- **결과**

False positive 의 비율을 1/3 이상 감소 시키는 성과를 보임.
인종별 (CEU vs YRI) 신규변이 탐지 비율의 차이와 Filtering 결과가 차이 나는 원인에 관심이 생겨 새로운 연구를 시작함.

기타 사항

외국어능력	영어 (보통 수준)
컴퓨터활용능력	워드/엑셀/파워포인트
프로그램 언어	C, R, Python, Perl 등