

Sangseon Lee (이상선)

BK21 FOUR Intelligence Computing
Seoul National University
Seoul, Korea
Phone. +82-10-2928-3740
E-mail. sangseon486@snu.ac.kr

EDUCATION

Seoul National University, Korea

- Ph.D., Computer Science and Engineering (Advisor: Professor Sun Kim) Mar. 2014 – Feb. 2020
- B. S., Computer Science and Engineering (*Summa Cum Laude*) Mar. 2010 – Aug. 2013

EXPERIENCE

Seoul National University, Korea

- Postdoctoral research fellow, SNU BK21 FOUR Intelligence Computing (Advisor: Professor Sun Kim) Dec. 2020 – Present
- Postdoctoral research fellow, SNU Bioinformatics Institute (Advisor: Professor Sun Kim) Mar. 2020 – Nov. 2020

RESEARCH INTERESTS

My general research area is applied machine learning & deep learning for interpretation of biological systems. Focuses on modeling interactions between biological elements of living organisms from DNA sequences to gene expression or epigenomes. Applications include subpath mining, sequence similarity measurement, interpretable deep learning model, and computational biology with multi-omics analysis.

RESEARCH EXPERIENCE

Graduate research in Bio & Health Informatics Lab, Seoul National University, Korea

< Deep Learning models for interpreting biological data with prior knowledge >

- Cancer subtype classification and modeling by pathway attention and propagation
 - Designing a pathway-based explainable deep learning model by graph convolutional network and attention mechanism
 - Predicting cancer subtypes using gene expression data and pathway information
 - Identifying subtype related transcription factors using network propagation algorithm
 - Analyzed RNA-seq dataset of five cancers with multiple subtypes
 - Paper accepted by *Bioinformatics* journal (SCI, IF=4.531)

- Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning with Regulatory Latent Space
 - Designing multi-modal deep learning model for learning gene regulation mechanism
 - Exploring properties of each omics by different neural network structures (ex. CNN, RNN)
 - Elucidating epigenetic gene regulation mechanism by various attention layers
 - Paper accepted by *Frontiers in Genetics* journal (SCIE, IF=3.258)
- Deep Hierarchical Embedding for Simultaneous Modeling of GPCR Proteins in a Unified Metric Space
 - Learning hierarchy in GPCR proteins using a multi-task convolutional neural network
 - Utilizing two types of loss: softmax-loss (for inter-class) & center-loss (for intra-class)
 - Encapsulating biological properties (ex. Phylogeny, Motif) in embedded hidden features
 - Paper submitted in *Scientific Reports* journal (Under review: Major revision)

< Pathway analysis for high-dimension & small sample-size data >

- Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data
 - Suggesting five criteria for measuring usefulness of pathway activity inference tools
 - Analyzed 13 pathway activity inference tools on PAN-cancer data
 - Paper accepted by *Briefings in Bioinformatics* journal (SCI, IF=9.101)
- MIDAS: Mining differentially activated subpaths of KEGG pathways from multi-class RNA-seq data
 - Kinds of subgraph mining for identifying differentially activated subpaths among multi-class data
 - Handling multi-class data by statistical test & greedy-based expansion technique
 - Paper accepted by *Methods* journal (SCI, IF=3.998)

< Information theoretic approach in sequence-type data >

- Ranked k-spectrum kernel for comparative and evolutionary comparison of exons, introns, and CpG islands
 - Suggesting a sequence similarity kernel based on rank and evolutionary information
 - Reconstructing phylogenetic trees of 10 mammalian species using exon, intron, and CpG islands sequences
 - Finding relationship across three regions in terms of evolutionary information contents: exon > CpG island > intron
 - Paper accepted by *IEEE/ACM Transactions on Computational Biology and Bioinformatics* journal (SCIE, IF=2.896)
- SpliceHetero: An information theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq
 - Measuring intratumor heterogeneity in transcriptome level by handling spliceome
 - Utilizing Shannon entropy and Jensen-Shannon Divergence to capture heterogeneity
 - Paper accepted by *PloS ONE* journal (SCI, IF=2.776)

< Application of computational algorithms on epigenomic data >

- DNMT1 maintains metabolic fitness of adipocytes through acting as an epigenetic safeguard of mitochondrial dynamics
 - Showing that DNA methylation directs distal enhancer-mediated transcriptomic features of adipocytes

- Using a sliding-window based DMR (Differentially Methylated Region) detecting algorithm and a combinatorial analysis of ChIP-seq and PCHi-C data
- Paper accepted by *Proceedings of the National Academy of Sciences (PNAS)* journal (SCI, IF=9.412)
- BioVLAB-mCpG-SNP-EXPRESS: A system for multi-level and multi-perspective analysis and exploration of DNA methylation, sequence variation (SNPs), and gene expression from multi-omics data
 - Platform for analyzing complex relationships between omics data
 - Guiding interpretation of results by multi-level and multi-perspective panels
 - Paper accepted by *Methods* journal (SCIE, IF=3.802)
- PRISM: Methylation Pattern-based, Reference-free Inference of Subclonal Makeup
 - Inferring the composition of epigenetically distinct subclones of a tumor from methylation patterns
 - Proofreading of methylation pattern based on DNMT1-like HMM
 - Inferencing epigenetic subclonal information by k-mixture decomposition with EM algorithm
 - Paper accepted by *ISMB 2019* conference, *Bioinformatics* journal (SCI, IF=4.531)

TEACHING EXPERIENCE

Guest lecturer: Machine Learning in Bioinformatics, 2021 Spring, Seoul National University, Korea

Guest lecturer: 생물정보학을 위한 IT기초, 2020 Spring, Seoul National University, Korea

Guest lecturer: 휴먼바이오공학기초, 2020 Spring, Ewha Womans University, Korea

Guest lecturer: 기계학습의 이해, 2019 Oct, Big Data Academy, Seoul National University, Korea

Guest lecturer: 휴먼바이오공학기초, 2019 Spring, Ewha Womans University, Korea

Guest lecturer: 기계학습의 이해, 2018 Jul, Big Data Academy, Seoul National University, Korea

Guest lecturer: Deep Learning for Bioinformatics, 2018 Feb, KSBi-BIML 2018 (생물정보학&머신러닝/인공지능 워크샵), Korea

PUBLICATIONS

Research Highlights

4 papers published as Deep Learning in Bioinformatics (국제 2편, 국내 2편)

2 papers published in *Bioinformatics* (The flagship journal in the bioinformatics field)

1 paper presented at *Intelligent Systems for Molecular Biology (ISMB)* (한국정보과학회 선정 최우수 학술대회)

International Journal (*: equal contributors)

[1] Sung I*, Lee S*, Pak M, Shin Y, Kim S. AutoCoV: Learning the Spatial and Temporal Dynamics of COVID-19 Spread Patterns from Embedding Space by k-mer Based Deep Learning. *International Conference on Intelligent Biology and Medicine (ICIBM 2021)*, Under review.

[2] Lee T, Lee S, Kang M, Sun Kim. Deep Hierarchical Embedding for Simultaneous Modeling of GPCR Proteins in a Unified Metric Space. *Scientific Reports*, Under review.

[3] Kim MS, Lee S, Lim S, Lee DY, Kim S. AttentiveCancer: an attention framework using subnetwork

level representation of transcriptome for predicting lymph node metastasis in early oral cancer. *Bioinformatics*, Under review.

[4] Kim I, **Lee S**, Kim Y, Namkoong H, Kim S. Probabilistic model for pathway-guided gene set selection. *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2021)*, Under review.

[5] Moon JH, **Lee S**, Hur B, Kim S. MLDEG: A network-based ensemble model to identify differentially expressed genes consistently. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (SCIE, IF=3.015), 2021 March; in press.

[6] Jeong D, Lim S, **Lee S**, Oh M, Cho C, Seong H, Jung W, Kim S. Construction of Condition-Specific Gene Regulatory Network using Kernel Canonical Correlation Analysis. *Frontiers in Genetics* (SCIE, IF=3.258), 2021 March; in press.

[7] Park YJ, **Lee S***, Lim S*, Hahn N, Ji Y, Huh JY, Alfadda AA, Kim S, Kim JB. DNMT1 maintains metabolic fitness of adipocytes through acting as an epigenetic safeguard of mitochondrial dynamics. *Proceedings of the National Academy of Sciences (PNAS)* (SCI, IF=9.412). 2021 Mar 16; 118(11):e2021073118.

[8] Oh M, Park S, **Lee S**, Lee D, Lim S, Jeong D, Jo K, Jung I, Kim S. DRIM: A web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration. *Frontiers in Genetics* (SCIE, IF=3.258), 2020 Nov; 11:564792.

[9] Kang M*, **Lee S***, Lee D, Kim S. Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning with Regulatory Latent Space. *Frontiers in Genetics* (SCIE, IF=3.258), 2020 Sep; 11:869.

[10] **Lee S**, Lim S, Lee T, Sung I, Kim S. Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics* (SCI, IF=5.610), 2020 Jun; 36(12):3818-3824.

[11] Lim S, **Lee S**, Jung I, Rhee S, Kim S. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in Bioinformatics* (SCI, IF=8.990), 2020 Jan; 21(1):36-46.

[12] Hur B, Kang D, **Lee S**, Moon JH, Lee G, Kim S. Venn-diaNet: Venn diagram based network propagation analysis framework for computing multiple biological experiments. *BMC Bioinformatics* (SCIE, IF=3.242), 2019 Dec 27; 20(23):1-12.

[13] Kang D, Ahn H, **Lee S**, Lee CJ, Hur J, Jung W, Kim S. StressGenePred: a twin prediction model architecture for classifying the stress types of samples and discovering stress-related genes in Arabidopsis. *BMC Genomics* (SCIE, IF=3.594), 2019 Dec 20; 20(11):949.

[14] Kim MS, **Lee S**, Lim S, Kim S. SpliceHetero: An information theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq. *PloS ONE* (SCI, IF=2.740), 2019 Oct 23; 14(10).

[15] **Lee S**, Lee T, Noh YK, Kim S. Ranked k-spectrum kernel for comparative and evolutionary comparison of exons, introns, and CpG islands. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (SCIE, IF=3.015), 2019 Sep 3, in press.

[16] Lee DH, **Lee S**, Kim S. PRISM: Methylation Pattern-based, Reference-free Inference of Subclonal Makeup. *Bioinformatics* (SCI, IF=5.610), 2019 Jul 5; 35(14):i520-i529.

[17] Lee CJ, Kang D, **Lee S**, Lee S, Kang J, Kim S. In silico experiment system for testing hypothesis on gene functions using three condition specific biological networks. *Methods* (SCI, IF=3.812), 2018 May 25; 145:10-15.

[18] **Lee S**, Park Y, Kim S. MIDAS: Mining differentially activated subpaths of KEGG pathways from multi-

class RNA-seq data. *Methods* (SCI, IF=3.812), 2017 Jul 15; 124:13-24.

[19] Moon JH, Lim S, Jo K, **Lee S**, Seo S, Kim S, PINTnet: construction of condition-specific pathway interaction network by computing shortest paths on weighted PPI. *BMC Systems Biology* (SCI, IF=2.048), 2017 Mar 14; 11(2):15.

[20] Chae H, **Lee S**, Nephew KP, Kim S, Subtype-specific CpG island shore methylation and mutation patterns in 30 breast cancer cell lines. *BMC Systems Biology* (SCI, IF=2.048), 2016 Dec 23; 10(4):116.

[21] Chae H, **Lee S**, Seo S, Jung D, Chang H, Nephew KP, Kim S. BioVLAB-mCpG-SNP-EXPRESS: A system for multi-level and multi-perspective analysis and exploration of DNA methylation, sequence variation (SNPs), and gene expression from multi-omics data. *Methods* (SCI, IF=3.802), 2016 Jul 28; 111:64-71.

[22] Jeong HM, **Lee S**, Chae H, Kim R, Kwon MJ, Oh E, Choi YL, Kim S, Shin YK. Efficiency of methylated DNA immunoprecipitation bisulphite sequencing for whole-genome DNA methylation analysis. *Epigenomics* (SCIE, IF=4.112), 2016 Jun 8; 8(8):1061-1077.

International Conference

[1] Hur B, Kang D, **Lee S**, Moon JH, Lee G, Kim S. Venn-diaNet: Venn diagram based network propagation analysis framework for computing multiple biological experiments. *30th International Conference on Genome Informatics (GIW)* 2019, Dec 9-11, Sydney, Australia.

[2] Lee DH, **Lee S**, Kim S. PRISM: Methylation Pattern-based, Reference-free Inference of Subclonal Makeup. *27th conference on Intelligent Systems for Molecular Biology (ISMB)* 2019, July 21-25, Basel, Switzerland.

[3] Kang D, Ahn H, **Lee S**, Lee CJ, Hur J, Jung W, Kim S. Identifying stress-related genes and predicting stress types in Arabidopsis using logical correlation layer and CMCL loss through time-series data. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2018, Dec 3-6, Madrid, Spain.

[4] **Lee S**, Moon JH, Park Y, Kim S, Flow maximization analysis of cell cycle pathway activation status in breast cancer subtypes, *IEEE International Conference on Big Data and Smart Computing (BigComp)* 2017, Feb 13-16, Jeju, Korea.

[5] Moon JH, Lim S, Jo K, **Lee S**, Seo S, Kim S, PINTnet: construction of condition-specific pathway interaction network by computing shortest paths on weighted PPI. *15th Asia Pacific Bioinformatics Conference (APBC)* 2017, Jan 16-18, Shenzhen, China.

[6] Chae H, **Lee S**, Nephew KP, Kim S, Subtype-specific CpG island shore methylation and mutation patterns in 30 breast cancer cell lines. *27th International Conference on Genome Informatics (GIW)* 2016, Oct 3-5, Shanghai, China.

Book Chapters

[1] Pak M, Jeong D, Moon JH, Ann H, Hur B, **Lee S**, Kim S. Network Propagation for the Analysis of Multi-Omics Data. *Recent Advances in Biological Network Analysis* (pp. 185-217). 2021

Domestic Journal

[1] 박은화, 이상선, 김선. 헬스케어 예측을 위한 전자 건강 기록 기반 그래프 컨볼루션 모델. 한국정보과학회 2020 한국컴퓨터종합학술대회 논문집, Vol.2020, pp.569-571. 2020 July.

[2] 강동원, 안홍렬, 이상선, 정우석, 김선. 시계열 유전자 발현 데이터의 스트레스 특이 반응

유전자 검출 및 스트레스 예측 알고리즘 개발. 한국정보과학회 2018 한국컴퓨터종합학술대회
논문집, Vol.2018, pp.1063-1065. 2018 Jun.

GRANTS / AWARDS

Google Education/Course Grant (Co-instructor), Google, Oct 2020.

스탠다임 우수논문상, 한국생명정보학회, Aug 20, 2020.

Best Paper Award (Bronze), 한국정보과학회 인공지능소사이어티 The AI KOREA 2019, Jul 26, 2019.

Best Paper Award on Bioinformatics, IEEE International Conference on Big Data and Smart Computing (BigComp) 2017, Feb 16, 2017.

Domestic Scholarship, 관정 이종환 교육재단, 2014-2015.