CSI 5180. Topics in Artificial Intelligence
Machine Learning for Bioinformatics Applications
Fall 2019

Projects

Version of October 29, 2019

1 Learning outcomes

- **Critically assess** a scientific publication
- **Further develop** lifelong learning skills
- **Communicate** technical information effectively in writing

2 Deadlines

- 2019-10-08 — Project proposal
- 2019-12-03 — Report

2.1 Deliverable

You must select a recent scientific publication where a machine learning algorithm was applied to a bioinformatics problem. You can select a paper from the bibliography below or one of your own choice. **This paper must be different than the one you will be presenting in class**.

You must recreate in part or in full the dataset used in that publication [1]. As seen in our lecture on essential bioinformatics skills, write scripts to automate this work. Make sure to describe the steps needed to prepare your data (cleaning, normalization, encoding...).

Apply two machine learning algorithms to your dataset and analyze your results. One of these two algorithms should be the same, or similar algorithm, to what was proposed in the publication. Were you able to reproduce the results from the paper? If not, why? Compare the two machine-learning algorithms. Is one approach better than the other? Why?

2.2 Teamwork

I am expecting that teams will be made of 1 or 2 members. Larger teams are possible and will have to produce proportionally more work! Complementary work between teams is also welcomed, i.e. two or more teams working on a related but complementary topic, leading to a more realistic application.

2.3 Report

The project is worth 20% of your final mark. Its marking will be based on the outline, a written report, as well as the source code submitted along with your project. Reports should be sufficiently detailed that it should be possible to rerun the analysis on the basis of the text alone. Having said that, you should also make every conceivable effort to keep the report concise. Assuming a team of size 2, a 10–15 page report should be appropriate. Suggested structure for the reports:

[1] If recreating the data happens to be too challenging for one reason or another, you can simulate the data, but only as a last resort.

- Introduction
  - Background
  - Problem definition
  - Describing the data
    - Where did you get the data?
    - What were the file formats?
    - Did you clean the data? If so, how?
    - Did you encode the data? What choices did you make?
- Methods
- Results
- Conclusions
- Full list of references

2.4 Selected publications

You must select a publication in a distinct area than that of your presentation in class. Below you will find a list of publications. You are welcomed proposing publications outside of the list. In an appendix, I am including a list of the major journals where bioinformatics research is published.

Essential Cell Biology

[1] Lawrence Hunter. Life and its molecules: A brief introduction. *AI Magazine*, 25(1):9–22, 2004.

Introduction to Applications of Machine Learning in Bioinformatics

[1] Michael Molla, Michael Waddell, David Page, and Jude W. Shavlik. Using machine learning to design and interpret gene-expression microarrays. *AI Magazine*, 25(1):23–44, 2004.

Reviews and Comparative Analyses

[1] Anne-Laure Boulesteix. Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Comput Biol*, 11(4):e1004191, Apr 2015.

[2] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData Min*, 10:35, 2017.

[3] Pedro M. Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87, 2012.

[4] Jianwen Fang. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief Bioinform*, Jul 2019.

[5] Jalil Nourmohammadi Khiarak, Rana VALİZADEH-KAMRAN, Ahmad Heydariyan, and Najmeh Damghani. Big data analysis in plant science and machine learning tool applications in genomics and proteomics. *International Journal of Computational and Experimental Science and Engineering*, 4(2):23–31.

[6] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 03 2006.

[7] Yumeng Liu, Xiaolong Wang, and Bin Liu. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform*, 20(1):330–346, 01 2019.

[8] Chuang Ma, Hao Helen Zhang, and Xiangfeng Wang. Machine learning for big data analytics in plants. *Trends Plant Sci*, 19(12):798–808, Dec 2014.

[9] Mufti Mahmud, Mohammed Shamim Kaiser, Amir Hussain, and Stefano Vassanelli. Applications of deep learning and reinforcement learning to biological data. *IEEE Transactions on Neural Networks and Learning Systems*, 29:2063–2079, 2018.

[10] Balachandran Manavalan, Shaherin Basith, Tae Hwan Shin, Leyi Wei, and Gwang Lee. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics*, 35(16):2757–2765, Aug 2019.

[11] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Brief Bioinform*, 18(5):851–869, 09 2017.

[12] Dai Hai Nguyen, Canh Hao Nguyen, and Hiroshi Mamitsuka. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief Bioinform*, Aug 2018.

[13] Randal S Olson, William La Cava, Zairah Mustahsan, Akshay Varik, and Jason H Moore. Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput*, 23:192–203, 2018.

[14] Kleber Padovani de Souza, João Carlos Setubal, André Carlos Ponce de Leon F de Carvalho, Guilherme Oliveira, Annie Chateau, and Ronnie Alves. Machine learning meets genome assembly. *Brief Bioinform*, Aug 2018.

[15] Arwa B Raies and Vladimir B Bajic. In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip Rev Comput Mol Sci*, 6(2):147–172, Mar 2016.

[16] Georgina Stegmayer, Leandro E Di Persia, Mariano Rubiolo, Matias Gerard, Milton Pividori, Cristian Yones, Leandro A Bugnon, Tadeo Rodriguez, Jonathan Raad, and Diego H Milone. Predicting novel microRNA: a comprehensive comparison of machine learning approaches. *Brief Bioinform*, May 2018.

[17] Adi L Tarca, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, Jun 2007.

[18] Chunming Xu and Scott A Jackson. Machine learning and complex biological data. *Genome Biol*, 20(1):76, 04 2019.

[19] Guido Zampieri, Supreeta Vijayakumar, Elisabeth Yaneske, and Claudio Angione. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput Biol*, 15(7):e1007084, Jul 2019.

[20] Yi-Hui Zhou and Paul Gallins. A review and tutorial of machine learning methods for microbiome host trait prediction. *Front Genet*, 10:579, 2019.

Possible projects (new)

[1] Xinzhong Li, Haiyan Wang, Jintao Long, Genhua Pan, Taigang He, Oleg Anichtchik, Robert Belshaw, Diego Albani, Paul Edison, Elaine K Green, and James Scott. Systematic analysis and biomarker study for Alzheimer's disease. *Sci Rep*, 8(1):17394, 11 2018.

[2] Yuan Lin, Yinyin Cai, Juan Liu, Chen Lin, and Xiangrong Liu. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC Bioinformatics*, 20(Suppl 8):291, Jun 2019.

Evaluating Learning Algorithms

[1] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: a classification perspective*. Cambridge University Press, Cambridge, 2011.

[2] Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome Biol*, 20(1):118, 06 2019.

[3] Samaneh Kouchaki, Yang Yang, Timothy M Walker, A Sarah Walker, Daniel J Wilson, Timothy E A Peto, Derrick W Crook, CRyPTIC Consortium, and David A Clifton. Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, 35(13):2276–2282, 11 2018.

[4] Lukas M Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D Robinson. Essential guidelines for computational method benchmarking. *Genome Biol*, 20(1):125, Jun 2019.

[5] Ying Zeng, Hongjie Yuan, Zheming Yuan, and Yuan Chen. A high-performance approach for predicting donor splice sites based on short window size and imbalanced large samples. *Biol Direct*, 14(1):6, 04 2019.

Dimensionality Reduction, Feature Selection, and Feature Engineering

[1] Syed Faraz Ahmed, Ahmed A Quadeer, David Morales-Jimenez, and Matthew R McKay. Sub-dominant principal components inform new vaccine targets for HIV Gag. *Bioinformatics*, 35(20):3884–3889, Oct 2019.

[2] Óscar Álvarez, Juan Luis Fernández-Martínez, Celia Fernández-Brillet, Ana Cernea, Zulima Fernández-Muñiz, and Andrzej Kloczkowski. Principal component analysis in protein tertiary structure prediction. *J Bioinform Comput Biol*, 16(2):1850005, 04 2018.

[3] Tallulah S Andrews and Martin Hemberg. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics*, 35(16):2865–2867, Aug 2019.

[4] Zafer Aydin, Oğuz Kaynar, and Yasin Görmez. Dimensionality reduction for protein secondary structure and solvent accesibility prediction. *J Bioinform Comput Biol*, 16(5):1850020, 10 2018.

[5] Gonzalo Cerruela García and Nicolás García-Pedrajas. Boosted feature selectors: a case study on prediction P-gp inhibitors and substrates. *J Comput Aided Mol Des*, 32(11):1273–1294, 11 2018.

[6] Weng Howe Chan, Mohd Saberi Mohamad, Safaai Deris, Nazar Zaki, Shahreen Kasim, Sigeru Omatu, Juan Manuel Corchado, and Hany Al Ashwal. Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme. *Comput Biol Med*, 77:102–15, 10 2016.

[7] Wei Chen, Hao Lv, Fulei Nie, and Hao Lin. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*, 35(16):2796–2800, Aug 2019.

[8] Yuehui Chen and Yaou Zhao. A novel ensemble of classifiers for microarray data classification. *Appl. Soft Comput.*, 8(4):1664–1669, 2008.

[9] Shahana Yasmin Chowdhury, Swakkhar Shatabda, and Abdollah Dehzangi. iDNAProt-ES: Identification of DNA-binding proteins using evolutionary and structural features. *Sci Rep*, 7(1):14938, 11 2017.

[10] Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14):i427–i435, 07 2019.

[11] Lakshmipadmaja D and B. Vishnuvardhan. Classification performance improvement using random subset feature selection algorithm for data mining. *Big Data Research*, 12:1–12, 2018.

[12] Patrick Deelen, Sipko van Dam, Johanna C Herkert, Juha M Karjalainen, Harm Brugge, Kristin M Abbott, Cleo C van Diemen, Paul A van der Zwaag, Erica H Gerkes, Evelien Zonneveld-Huijssoon, Jelkje J Boer-Bergsma, Pytrik Folkertsma, Tessa Gillett, K Joeri van der Velde, Roan Kanninga, Peter C van den Akker, Sabrina Z Jan, Edgar T Hoorntje, Wouter P Te Rijdt, Yvonne J Vos, Jan D H Jongbloed, Conny M A van Ravenswaaij-Arts, Richard Sinke, Birgit Sikkema-Raddatz, Wilhelmina S Kerstjens-Frederikse, Morris A Swertz, and Lude Franke. Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. *Nat Commun*, 10(1):2837, Jun 2019.

[13] John M Giorgi and Gary D Bader. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094, Dec 2018.

[14] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[15] Lan Huong Nguyen and Susan Holmes. Ten quick tips for effective dimensionality reduction. *PLoS Comput Biol*, 15(6):e1006907, Jun 2019.

[16] Ahmad Abu Shanab and Taghi Khoshgoftaar. Is gene selection enough for imbalanced bioinformatics data? In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 346–355. IEEE, 2018.

[17] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, 35(12):2133–2140, Jun 2019.

[18] Jing Tang, Yunxia Wang, Jianbo Fu, Ying Zhou, Yongchao Luo, Ying Zhang, Bo Li, Qingxia Yang, Weiwei Xue, Yan Lou, Yunqing Qiu, and Feng Zhu. A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies. *Brief Bioinform*, Jun 2019.

[19] Leyi Wei, Chen Zhou, Huangrong Chen, Jiangning Song, and Ran Su. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, 34(23):4007–4016, Dec 2018.

[20] Jing Xu, Peng Wu, Yuehui Chen, Qingfang Meng, Hussain Dawood, and Muhammad Murtaza Khan. A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. *IEEE Access*, 7:22086–22095, 2019.

[21] Silu Zhang, Junqing Wang, Torumoy Ghoshal, Dawn Wilkins, Yin-Yuan Mo, Yixin Chen, and Yunyun Zhou. lncRNA gene signatures for prediction of breast cancer intrinsic subtypes and prognosis. *Genes (Basel)*, 9(2), Jan 2018.

Data Imputation

[1] Xuesi Dong, Lijuan Lin, Ruyang Zhang, Yang Zhao, David C Christiani, Yongyue Wei, and Feng Chen. TOBMI: trans-omics block missing data imputation using a k-nearest neighbor weighted approach. *Bioinformatics*, 35(8):1278–1283, Apr 2019.

[2] Timothy J Durham, Maxwell W Libbrecht, J Jeffry Howbert, Jeff Bilmes, and William Stafford Noble. PREDICTD PaRallel epigenomics data imputation with cloud-based tensor decomposition. *Nat Commun*, 9(1):1402, 04 2018.

[3] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol*, 33(4):364–76, Apr 2015.

[4] Kohbalan Moorthy, Aws Naser Jaber, Mohd Arfian Ismail, Ferda Ernawan, Mohd Saberi Mohamad, and Safaai Deris. Missing-values imputation algorithms for microarray gene expression data. *Methods Mol Biol*, 1986:255–266, 2019.

[5] Kohbalan Moorthy, Mohd Saberi Mohamad, and Safaai Deris. A review on missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics,* 9(1):18–22, 2014.

[6] Qian Qin and Jianxing Feng. Imputation for transcription factor binding predictions based on deep learning. *PLoS Comput Biol,* 13(2):e1005403, 02 2017.

[7] Jacob Schreiber, Timothy Durham, Jeffrey Bilmes, and William Stafford Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *bioRxiv,* 2018.

[8] Aiguo Wang, Ye Chen, Ning An, Jing Yang, Lian Li, and Lili Jiang. Microarray missing value imputation: A regularized local learning method. *IEEE/ACM Trans Comput Biol Bioinform,* Feb 2018.

Unsupervised Learning

[1] Davide Chicco and Marco Masseroli. Ontology-based prediction and prioritization of gene functional annotations. *IEEE/ACM Trans Comput Biol Bioinform,* 13(2):248–60, 2016.

[2] Pietro Coretto, Angela Serra, and Roberto Tagliaferri. Robust clustering of noisy high-dimensional gene expression data for patients subtyping. *Bioinformatics,* 34(23):4064–4072, 12 2018.

[3] Pallavi Gaur and Anoop Chaturvedi. Clustering and candidate motif detection in exosomal miRNAs by application of machine learning algorithms. *Interdiscip Sci,* 11(2):206–214, Jun 2019.

[4] Troy P Hubbard, Jonathan D D'Gama, Gabriel Billings, Brigid M Davis, and Matthew K Waldor. Unsupervised learning approach for comparing multiple transposon insertion sequencing studies. *mSphere,* 4(1), 02 2019.

[5] Benjamin T James, Brian B Luczak, and Hani Z Girgis. MeShClust: an intelligent tool for clustering DNA sequences. *Nucleic Acids Res,* 46(14):e83, Aug 2018.

[6] Rachel Jeitziner, Mathieu Carrière, Jacques Rougemont, Steve Oudot, Kathryn Hess, and Cathrin Brisken. Two-Tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis. *Bioinformatics,* 35(18):3339–3347, Sep 2019.

[7] G Kerr, H J Ruskin, M Crane, and P Doolan. Techniques for clustering gene expression data. *Comput Biol Med,* 38(3):283–93, Mar 2008.

[8] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet,* 20(5):273–282, 05 2019.

[9] Lokesh Kumar and Matthias E Futschik. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation,* 2(1):5–7, May 2007.

[10] Xiangtao Li, Shixiong Zhang, and Ka-Chun Wong. Single-cell RNA-seq interpretations using evolutionary multiobjective ensemble pruning. *Bioinformatics,* 35(16):2809–2817, Aug 2019.

[11] Hang Liu, Lei Peng, Joan So, Ka Hing Tsang, Chi Ho Chong, Priscilla Hoi Shan Mak, Kui Ming Chan, and Siu Yuen Chan. TSPYL2 regulates the expression of EZH2 target genes in neurons. *Mol Neurobiol,* 56(4):2640–2652, Apr 2019.

[12] Matthew J Michalska-Smith and Stefano Allesina. Telling ecological networks apart by their structure: A computational challenge. *PLoS Comput Biol,* 15(6):e1007076, Jun 2019.

[13] Hung Nguyen, Sangam Shrestha, Sorin Draghici, and Tin Nguyen. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics,* 35(16):2843–2846, Aug 2019.

[14] Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, and Ezekiel Adebiyi. Clustering algorithms: Their application to gene expression data. *Bioinform Biol Insights,* 10:237–253, 2016.

[15] Ren Qi, Anjun Ma, Qin Ma, and Quan Zou. Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinform*, Jul 2019.

[16] Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, page 622803, 2019.

[17] Grzegorz Rorbach, Olgierd Unold, and Bogumil M Konopka. Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods. *Sci Rep*, 8(1):7560, 05 2018.

[18] Debajyoti Sinha, Akhilesh Kumar, Himanshu Kumar, Sanghamitra Bandyopadhyay, and Debarka Sengupta. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res*, 46(6):e36, 04 2018.

[19] Xiaoping Su, Gabriel G Malouf, Yunxin Chen, Jianping Zhang, Hui Yao, Vicente Valero, John N Weinstein, Jean-Philippe Spano, Funda Meric-Bernstam, David Khayat, and Francisco J Esteva. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget*, 5(20):9864–76, Oct 2014.

[20] Tian Tian, Ji Wan, Qi Song, and Zhi Wei. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191, 2019.

[21] Qiu Xiao, Jiawei Luo, Cheng Liang, Jie Cai, Guanghui Li, and Buwen Cao. CeModule: an integrative framework for discovering regulatory patterns from genomic data in cancer. *BMC Bioinformatics*, 20(1):67, Feb 2019.

Supervised Learning

[1] Toby Dylan Hocking, Patricia Goerner-Potvin, Andreanne Morin, Xiaojian Shao, Tomi Pastinen, and Guillaume Bourque. Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning. *Bioinformatics*, 33(4):491–499, 02 2017.

[2] Kevin Li, Rachel Chen, William Lindsey, Aaron Best, Matthew DeJongh, Christopher Henry, and Nathan Tintle. Implementing and evaluating a Gaussian mixture framework for identifying gene function from TnSeq data. *Pac Symp Biocomput*, 24:172–183, 2019.

[3] Yuan Lin, Yinyin Cai, Juan Liu, Chen Lin, and Xiangrong Liu. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC Bioinformatics*, 20(Suppl 8):291, Jun 2019.

Linear, Logistic Regression, naïve Bayes

[1] Kseniia Cheloshkina and Maria Poptsova. Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. *BMC Cancer*, 19(1):434, May 2019.

[2] Simon Dirmeier, Christiane Fuchs, Nikola S Mueller, and Fabian J Theis. netReg: network-regularized linear models for biological association studies. *Bioinformatics*, 34(5):896–898, 03 2018.

[3] Daniel W Kennedy, Nicole M White, Miles C Benton, Andrew Fox, Rodney J Scott, Lyn R Griffiths, Kerrie Mengersen, and Rodney A Lea. Critical evaluation of linear regression models for cell-subtype specific methylation signal from mixed blood cell DNA. *PLoS One*, 13(12):e0208915, 2018.

[4] Justin R Klesmith and Benjamin J Hackel. Improved mutant function prediction via PACT: Protein Analysis and Classifier Toolkit. *Bioinformatics (Oxford, England)*, 35(16):2707–2712, 2019.

[5] Wenyuan Li, Chun-Chi Liu, Shuli Kang, Jian-Rong Li, Yu-Ting Tseng, and Xianghong Jasmine Zhou. Pushing the annotation of cellular activities to a higher resolution: Predicting functions at the isoform level. *Methods*, 93:110–8, Jan 2016.

[6] Thomas Luechtefeld, Dan Marsh, Craig Rowlands, and Thomas Hartung. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol Sci*, 165(1):198–212, 09 2018.

[7] Trisevgeni Rapakoulia, Xin Gao, Yi Huang, Michiel de Hoon, Mariko Okada-Hatakeyama, Harukazu Suzuki, and Erik Arner. Genome-scale regression analysis reveals a linear relationship for promoters and enhancers after combinatorial drug treatment. *Bioinformatics*, 33(23):3696–3700, Dec 2017.

[8] Liang Sun, Haitao Luo, Dechao Bu, Guoguang Zhao, Kuntao Yu, Changhai Zhang, Yuanning Liu, Runsheng Chen, and Yi Zhao. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res*, 41(17):e166, Sep 2013.

[9] Heng Xiong, Dongbing Liu, Qiye Li, Mengyue Lei, Liqin Xu, Liang Wu, Zongji Wang, Shancheng Ren, Wangsheng Li, Min Xia, Lihua Lu, Haorong Lu, Yong Hou, Shida Zhu, Xin Liu, Yinghao Sun, Jian Wang, Huanming Yang, Kui Wu, Xun Xu, and Leo J Lee. RED-ML: a novel, effective RNA editing detection method based on machine learning. *Gigascience*, 6(5):1–8, 05 2017.

Decision Trees, Random forests and eXtreme Gradient Boosting

[1] Xing Chen, Li Huang, Di Xie, and Qi Zhao. EGBMMDA: Extreme gradient boosting machine for miRNA-disease association prediction. *Cell Death Dis*, 9(1):3, 01 2018.

[2] Sayamon Hongjaisee, Chanin Nantasenamat, Tanawan Samleerat Carraway, and Watshara Shoombuatong. HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. *Comput Biol Chem*, 80:419–432, Jun 2019.

[3] Dharm Skandh Jain, Sanket Rajan Gupte, and Raviprasad Aduri. A data driven model for predicting RNA-Protein interactions based on gradient boosting machine. *Sci Rep*, 8(1):9552, 06 2018.

[4] Michael Lee, Erdahl T Teber, Oliver Holmes, Katia Nones, Ann-Marie Patch, Rebecca A Dagg, Loretta M S Lau, Joyce H Lee, Christine E Napier, Jonathan W Arthur, Sean M Grimmond, Nicholas K Hayward, Peter A Johansson, Graham J Mann, Richard A Scolyer, James S Wilmott, Roger R Reddel, John V Pearson, Nicola Waddell, and Hilda A Pickett. Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic Acids Res*, 46(10):4903–4918, 06 2018.

[5] Maxwell W Libbrecht, Oscar L Rodriguez, Zhiping Weng, Jeffrey A Bilmes, Michael M Hoffman, and William Stafford Noble. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol*, 20(1):180, 08 2019.

[6] Joseph Luttrell, 4th, Tong Liu, Chaoyang Zhang, and Zheng Wang. Predicting protein residue-residue contacts using random forests and deep networks. *BMC Bioinformatics*, 20(Suppl 2):100, Mar 2019.

[7] Arturo Magana-Mora and Vladimir B Bajic. OmniGA: Optimized omnivariate decision trees for generalizable classification models. *Sci Rep*, 7(1):3898, 06 2017.

[8] Raghvendra Mall, Luigi Cerulo, Luciano Garofano, Veronique Frattini, Khalid Kunji, Halima Bensmail, Thais S Sabedot, Houtan Noushmehr, Anna Lasorella, Antonio Iavarone, and Michele Ceccarelli. RGBM: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. *Nucleic Acids Res*, 46(7):e39, 04 2018.

[9] Bethany Signal, Brian S Gloss, Marcel E Dinger, and Tim R Mercer. Machine learning annotation of human branchpoints. *Bioinformatics*, 34(6):920–927, 03 2018.

[10] Junhui Wang and Michael Gribskov. IRESpy: an XGBoost model for prediction of internal ribosome entry sites. *BMC Bioinformatics*, 20(1):409, Jul 2019.

[11] Lei Wang, Zhu-Hong You, Xing Chen, Yang-Ming Li, Ya-Nan Dong, Li-Ping Li, and Kai Zheng. LMTRDA: Using logistic model tree to predict miRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput Biol*, 15(3):e1006865, 2019.

[12] Xin Wang, Peijie Lin, and Joshua W K Ho. Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random forest. *BMC Genomics*, 19(Suppl 1):929, 01 2018.

[13] Jing Xu, Peng Wu, Yuehui Chen, Qingfang Meng, Hussain Dawood, and Muhammad Murtaza Khan. A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. *IEEE Access*, 7:22086–22095, 2019.

Hidden Markov Models (HMM)

[1] Cristian V Crisan, Aroon T Chande, Kenneth Williams, Vishnu Raghuram, Lavanya Rishishwar, Gabi Steinbach, Samit S Watve, Peter Yunker, I King Jordan, and Brian K Hammer. Analysis of Vibrio cholerae genomes identifies new type VI secretion system gene clusters. *Genome Biol*, 20(1):163, 08 2019.

[2] Thomas Harrison, Jaime Ruiz, Daniel B Sloan, Asa Ben-Hur, and Christina Boucher. aPPRove: An HMm-based method for accurate prediction of RNA-pentatricopeptide repeat protein binding events. *PLoS One*, 11(8):e0160645, 2016.

[3] Ka-Chun Wong. DNA motif recognition modeling from protein sequences. *iScience*, 7:198–211, Sep 2018.

[4] Tobias Zehnder, Philipp Benner, and Martin Vingron. Predicting enhancers in mammalian genomes using supervised hidden markov models. *BMC Bioinformatics*, 20(1):157, Mar 2019.

Extreme Learning Machines

[1] Yanjuan Li, Mengting Niu, and Quan Zou. ELM-MHC: An improved MHC identification method with extreme learning machine algorithm. *J Proteome Res*, 18(3):1392–1401, Mar 2019.

[2] M Rubiolo, D H Milone, and G Stegmayer. Extreme learning machines for reverse engineering of gene regulatory networks from expression time series. *Bioinformatics*, 34(7):1253–1260, 04 2018.

[3] S Zhang, T Zhang, and C Liu. Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine. *SAR QSAR Environ Res*, 30(3):209–228, Mar 2019.

Support Vector Machine (SVM)

[1] Eran Barash, Neta Sal-Man, Sivan Sabato, and Michal Ziv-Ukelson. BacPaCS-bacterial pathogenicity classification via sparse-SVM. *Bioinformatics*, 35(12):2001–2008, Jun 2019.

[2] Jialu Hu, Jingru Wang, Jianan Lin, Tianwei Liu, Yuanke Zhong, Jie Liu, Yan Zheng, Yiqun Gao, Junhao He, and Xuequn Shang. MD-SVM: a novel SVM-based algorithm for the motif discovery of transcription factor binding sites. *BMC Bioinformatics*, 20(Suppl 7):200, May 2019.

[3] Yuchen Li, Jingjing Zhao, Shulin Yu, Zhen Wang, Xigan He, Yonghui Su, Tianan Guo, Haoyue Sheng, Jie Chen, Qiupeng Zheng, Yan Li, Weijie Guo, Xiaohong Cai, Guohai Shi, Jiong Wu, Lu Wang, Peng Wang, Xianghuo He, and Shenglin Huang. Extracellular vesicles long RNA sequencing reveals abundant mRNA, circRNA, and lncRNA in human blood as potential biomarkers for cancer diagnosis. *Clin Chem*, 65(6):798–808, Jun 2019.

[4] Maria Littmann, Tatyana Goldberg, Sebastian Seitz, Mikael Bodén, and Burkhard Rost. Detailed prediction of protein sub-nuclear localization. *BMC Bioinformatics*, 20(1):205, Apr 2019.

[5] Xin Ma, Jing Guo, and Xiao Sun. Prediction of microRNA-binding residues in protein using a Laplacian support vector machine based on sequence information. *J Bioinform Comput Biol*, 16(3):1840009, 06 2018.

[6] Saskia Metzler and Olga V Kalinina. Detection of atypical genes in virus families using a one-class SVM. *BMC Genomics*, 15:913, Oct 2014.

[7] Qiao Ning, Miao Yu, Jinchao Ji, Zhiqiang Ma, and Xiaowei Zhao. Analysis and prediction of human acetylation using a cascade classifier based on support vector machine. *BMC Bioinformatics*, 20(1):346, Jun 2019.

[8] Julia Rahman, Md Nazrul Islam Mondal, Md Khaled Ben Islam, and Md Al Mehedi Hasan. Feature fusion based svm classifier for protein subcellular localization prediction. *J Integr Bioinform*, 13(1):288, Dec 2016.

[9] Avanti Shrikumar, Eva Prakash, and Anshul Kundaje. GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics*, 35(14):i173–i182, 07 2019.

[10] Hao Yang, Hao Chi, Wen-Feng Zeng, Wen-Jing Zhou, and Si-Min He. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, 35(14):i183–i190, 07 2019.

Kernel Methods

[1] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:i38–46, Jun 2005.

[2] Dexiong Chen, Laurent Jacob, and Julien Mairal. Biological sequence modeling with convolutional kernel networks. *Bioinformatics*, 35(18):3294–3302, Sep 2019.

[3] Xing Chen, Lei Wang, Jia Qu, Na-Na Guan, and Jian-Qiang Li. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics*, 34(24):4256–4265, 12 2018.

[4] Mehmet Gönen and Adam A. Margolin. Localized data fusion for kernel k-Means clustering with application to cancer biology. In *NIPS*, pages 1305–1313, 2014.

[5] George Karypis. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins*, 64(3):575–86, Aug 2006.

[6] Seonho Kim, Juntae Yoon, Jihoon Yang, and Seog Park. Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11:107, Feb 2010.

[7] Rui Kuang, Eugene Ie, Ke Wang, Kai Wang, Mahira Siddiqi, Yoav Freund, and Christina Leslie. Profile-based string kernels for remote homology detection and motif extraction. *J Bioinform Comput Biol*, 3(3):527–50, Jun 2005.

[8] Bin Liu, Deyuan Zhang, Ruifeng Xu, Jinghao Xu, Xiaolong Wang, Qingcai Chen, Qiwen Dong, and Kuo-Chen Chou. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics*, 30(4):472–9, Feb 2014.

[9] Dan Liu, Xiaohua Hu, Tingting He, and Xingpeng Jiang. Virus-host association prediction by using kernelized logistic matrix factorization on heterogeneous networks. In *BIBM*, pages 108–113. IEEE Computer Society, 2018.

[10] Dawei Liu, Debashis Ghosh, and Xihong Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9:292, Jun 2008.

[11] Yingiun Ma, Limin Yu, Tingting He, Xiaohua Hu, and Xingpeng Jiang. Prediction of long non-coding RNA-protein interaction through kernel soft-neighborhood similarity. In *BIBM*, pages 193–196. IEEE Computer Society, 2018.

[12] Peter Meinicke, Maike Tech, Burkhard Morgenstern, and Rainer Merkl. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, 5:169, Oct 2004.

[13] Jesper Salomon and Darren R Flower. Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores. *BMC Bioinformatics*, 7:501, Nov 2006.

[14] Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–80, Jul 2006.

[15] Jean-Philippe Vert, Jian Qiu, and William S Noble. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S8, 2007.

[16] Dennis Wylie, Hans A Hofmann, and Boris V Zemelman. SArKS: de novo discovery of gene expression regulatory motif sites and domains by suffix array kernel smoothing. *Bioinformatics*, Mar 2019.

[17] Han Zhang, Xueting Huo, Xia Guo, Xin Su, Xiongwen Quan, and Chen Jin. A disease-related gene mining method based on weakly supervised learning model. In *BIBM*, pages 169–174. IEEE Computer Society, 2018.

Artificial Neural Networks (ANN)

[1] Maria Schelling, Thomas A Hopf, and Burkhard Rost. Evolutionary couplings and sequence variation effect predict protein binding sites. *Proteins*, 86(10):1064–1074, 10 2018.

Deep Learning

[1] Tanvir Alam, Mohammad Tariqul Islam, Mowafa Househ, Samir Brahim Belhaouari, and Ferdaus Ahmed Kawsar. DeepCNPP: Deep learning architecture to distinguish the promoter of human long non-coding RNA genes and protein-coding genes. *Stud Health Technol Inform*, 262:232–235, Jul 2019.

[2] Tanvir Alam, Mohammad Tariqul Islam, Mowafa Househ, Abdesselam Bouzerdoum, and Ferdaus Ahmed Kawsar. DeepDSSR: Deep learning structure for human donor splice sites recognition. *Stud Health Technol Inform*, 262:236–239, Jul 2019.

[3] Jose Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(24):4049, 12 2017.

[4] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, Nov 2017.

[5] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell Syst*, 8(4):292–301.e3, Apr 2019.

[6] Noorul Amin, Annette McGrath, and Yi-Ping Phoebe Chen. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence*, 1(5):246, 2019.

[7] Matthew Amodio, David van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, Anita Desai, V Ravi, Priti Kumar, Ruth Montgomery, Guy Wolf, and Smita Krishnaswamy. Exploring single-cell data with deep multitasking neural networks. *Nat Methods*, Oct 2019.

[8] Genta Aoki and Yasubumi Sakakibara. Convolutional neural networks for classification of alignments of non-coding RNA sequences. *Bioinformatics,* 34(13):i237–i244, 07 2018.

[9] Žiga Avsec, Mohammadamin Barekatain, Jun Cheng, and Julien Gagneur. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *Bioinformatics,* 34(8):1261–1269, 04 2018.

[10] Žiga Avsec, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, Daniel S Kim, Thorsten Beier, Lara Urban, Anshul Kundaje, Oliver Stegle, and Julien Gagneur. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat Biotechnol,* 37(6):592–600, Jun 2019.

[11] Junghwan Baek, Byunghan Lee, Sunyoung Kwon, and Sungroh Yoon. LncRNAnet: long non-coding RNA identification using deep learning. *Bioinformatics,* 34(22):3889–3897, 11 2018.

[12] Ilan Ben-Bassat, Benny Chor, and Yaron Orenstein. A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics,* 34(17):i638–i646, 09 2018.

[13] Christopher F Blum and Markus Kollmann. Neural networks with circular filters enable data efficient inference of sequence motifs. *Bioinformatics,* 35(20):3937–3943, Oct 2019.

[14] Philipp Bongartz and Siegfried Schloissnig. Deep repeat resolution-the assembly of the Drosophila Histone Complex. *Nucleic Acids Res,* 47(3):e18, 02 2019.

[15] Hannes Bretschneider, Shreshth Gandhi, Amit G Deshwar, Khalid Zuberi, and Brendan J Frey. COSSMO: predicting competitive alternative splice site selection using deep learning. *Bioinformatics,* 34(13):i429–i437, 07 2018.

[16] Stefan Budach and Annalisa Marsico. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics,* 34(17):3035–3037, 09 2018.

[17] LA Bugnon, C Yones, DH Milone, and G Stegmayer. Deep neural architectures for highly imbalanced data in bioinformatics. *IEEE transactions on neural networks and learning systems,* 2019.

[18] Zhen Cao and Shihua Zhang. Simple tricks of convolutional neural network architectures improve DNA-protein binding prediction. *Bioinformatics,* 35(11):1837–1843, Jun 2019.

[19] P Chang, J Grinband, B D Weinberg, M Bardis, M Khy, G Cadena, M-Y Su, S Cha, C G Filippi, D Bota, P Baldi, L M Poisson, R Jain, and D Chow. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am J Neuroradiol,* 39(7):1201–1207, 07 2018.

[20] Hao Chen, Dipan Shaw, Jianyang Zeng, Dongbo Bu, and Tao Jiang. DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. *Bioinformatics,* 35(14):i284–i294, 07 2019.

[21] Muhao Chen, Chelsea J T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics,* 35(14):i305–i314, 07 2019.

[22] Kevin M Cherry and Lulu Qian. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature,* 559(7714):370–376, 07 2018.

[23] Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics,* 34(23):4046–4053, 12 2018.

[24] Mattia Di Gangi, Giosuè Lo Bosco, and Riccardo Rizzo. Deep learning architectures for prediction of nucleosome positioning from sequences data. *BMC Bioinformatics,* 19(Suppl 14):418, Nov 2018.

[25] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet*, 20(7):389–403, Jul 2019.

[26] Rui Fa, Domenico Cozzetto, Cen Wan, and David T Jones. Predicting human protein function with multi-task deep neural networks. *PLoS One*, 13(6):e0198216, 2018.

[27] Chao Fang, Yi Shang, and Dong Xu. MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins*, 86(5):592–598, 05 2018.

[28] Antonino Fiannaca, Massimo La Rosa, Laura La Paglia, Riccardo Rizzo, and Alfonso Urso. nRC: non-coding RNA classifier based on structural features. *BioData Min*, 10:27, 2017.

[29] Lex Flagel, Yaniv Brandvain, and Daniel R Schrider. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol Biol Evol*, 36(2):220–238, 02 2019.

[30] Xin Gao, Jie Zhang, Zhi Wei, and Hakon Hakonarson. Deeppolya: A convolutional neural network approach for polyadenylation site prediction. *IEEE Access*, 6:24340–24349, 2018.

[31] Brian L Gudenas and Liangjiang Wang. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Sci Rep*, 8(1):16385, Nov 2018.

[32] Yanbu Guo, Weihua Li, Bingyi Wang, Huiqing Liu, and Dongming Zhou. DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC Bioinformatics*, 20(1):341, Jun 2019.

[33] Yanbu Guo, Bingyi Wang, Weihua Li, and Bei Yang. Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. *J Bioinform Comput Biol*, 16(5):1850021, 10 2018.

[34] Yang Guo, Xuequn Shang, and Zhanhuai Li. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. *Neurocomputing*, 324:20–30, 2019.

[35] Geoffrey D Hannigan, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, Michael Wurst, Jakub Kotowski, Dan Chang, Rurun Wang, Grazia Piizzi, Gergely Temesi, Daria J Hazuda, Christopher H Woelk, and Danny A Bitton. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res*, Aug 2019.

[36] Jack Hanson, Kuldip Paliwal, Thomas Litfin, Yuedong Yang, and Yaoqi Zhou. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, 34(23):4039–4045, Dec 2018.

[37] Somaye Hashemifar, Behnam Neyshabur, Aly A Khan, and Jinbo Xu. Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810, 09 2018.

[38] Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, Yuedong Yang, and Yaoqi Zhou. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep*, 5:11476, Jun 2015.

[39] Rhys Heffernan, Kuldip Paliwal, James Lyons, Jaswinder Singh, Yuedong Yang, and Yaoqi Zhou. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J Comput Chem*, 39(26):2210–2216, 10 2018.

[40] Rhys Heffernan, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 33(18):2842–2849, Sep 2017.

[41] Steven T Hill, Rachael Kuintzle, Amy Teegarden, Erich Merrill, 3rd, Padideh Danaee, and David A Hendrix. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res*, 46(16):8105–8113, 09 2018.

[42] Jiajun Hong, Yongchao Luo, Yang Zhang, Junbiao Ying, Weiwei Xue, Tian Xie, Lin Tao, and Feng Zhu. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Briefings in bioinformatics*, 2019.

[43] Hailin Hu, An Xiao, Sai Zhang, Yangyang Li, Xuanling Shi, Tao Jiang, Linqi Zhang, Lei Zhang, and Jianyang Zeng. Deephint: understanding hiv-1 integration via deep learning with attention. *Bioinformatics*, 35(10):1660–1667, May 2019.

[44] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, Eric D Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, Jan 2019.

[45] Anupama Jha, Matthew R Gazzara, and Yoseph Barash. Integrative deep models for alternative splicing. *Bioinformatics*, 33(14):i274–i282, Jul 2017.

[46] Shuangxi Ji, Tuğçe Oruç, Liam Mead, Muhammad Fayyaz Rehman, Christopher Morton Thomas, Sam Butterworth, and Peter James Winn. DeepCDpred: Inter-residue distance and contact prediction for improved prediction of protein structure. *PLoS One*, 14(1):e0205214, 2019.

[47] David T Jones and Shaun M Kandathil. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, 34(19):3308–3315, 10 2018.

[48] Vanessa Isabell Jurtz, Alexander Rosenberg Johansen, Morten Nielsen, Jose Juan Almagro Armenteros, Henrik Nielsen, Casper Kaae Sønderby, Ole Winther, and Søren Kaae Sønderby. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics*, 33(22):3685–3690, Nov 2017.

[49] Manal Kalkatawi, Arturo Magana-Mora, Boris Jankovic, and Vladimir B Bajic. DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. *Bioinformatics*, 35(7):1125–1132, Apr 2019.

[50] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, Sep 2019.

[51] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res*, 26(7):990–9, 07 2016.

[52] Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, and Raghvendra Mall. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, 08 2018.

[53] Savvas Kinalis, Finn Cilius Nielsen, Ole Winther, and Frederik Otzen Bagger. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC Bioinformatics*, 20(1):379, Jul 2019.

[54] Nguyen Quoc Khanh Le, Edward Kien Yee Yapp, and Hui-Yuan Yeh. ET-GRU: using multi-layer gated recurrent units to identify electron transport proteins. *BMC Bioinformatics*, 20(1):377, Jul 2019.

[55] Nung Kion Lee, Farah Liyana Azizan, Yu Shiong Wong, and Norshafarina Omar. DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery. *Biotechnology & Biotechnological Equipment*, 32(3):759–768, 2018.

[56] Yifeng Li, Wenqiang Shi, and Wyeth W Wasserman. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinformatics*, 19(1):202, 05 2018.

[57] Ge Liu, Haoyang Zeng, and David K Gifford. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC Bioinformatics*, 20(1):401, Jul 2019.

[58] Qiao Liu, Hairong Lv, and Rui Jiang. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics*, 35(14):i99–i107, 07 2019.

[59] Yang Liu, Perry Palmedo, Qing Ye, Bonnie Berger, and Jian Peng. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst*, 6(1):65–74.e3, Jan 2018.

[60] Yang Liu, Qing Ye, Liwei Wang, and Jian Peng. Learning structural motif representations for efficient protein structure search. *Bioinformatics*, 34(17):i773–i780, 09 2018.

[61] Jan Ludwiczak, Aleksander Winski, Krzysztof Szczepaniak, Vikram Alva, and Stanislaw Dunin-Horkawicz. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, 35(16):2790–2795, Aug 2019.

[62] Fenglin Luo, Minghui Wang, Yu Liu, Xing-Ming Zhao, and Ao Li. DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, 35(16):2766–2773, Aug 2019.

[63] Xu Min, Wanwen Zeng, Shengquan Chen, Ning Chen, Ting Chen, and Rui Jiang. Predicting enhancers with deep convolutional neural networks. *BMC Bioinformatics*, 18(Suppl 13):478, Dec 2017.

[64] Tatsuhiko Naito. Human splice-site prediction with deep neural networks. *J Comput Biol*, 25(8):954–961, 08 2018.

[65] James O'Connell, Zhixiu Li, Jack Hanson, Rhys Heffernan, James Lyons, Kuldip Paliwal, Abdollah Dehzangi, Yuedong Yang, and Yaoqi Zhou. SPIN2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins*, 86(6):629–633, 06 2018.

[66] Zhangyi Ouyang, Feng Liu, Chenghui Zhao, Chao Ren, Gaole An, Chuan Mei, Xiaochen Bo, and Wenjie Shu. Accurate identification of RNA editing sites from primitive sequence with deep neural networks. *Sci Rep*, 8(1):6005, 04 2018.

[67] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 09 2018.

[68] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*, 35(18):3313–3319, Sep 2019.

[69] Joseph M Paggi and Gill Bejerano. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA*, 24(12):1647–1658, 12 2018.

[70] Xiaoyong Pan and Hong-Bin Shen. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, 34(20):3427–3436, 10 2018.

[71] Albert Pla, Xiangfu Zhong, and Simon Rayner. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLoS Comput Biol*, 14(7):e1006185, 07 2018.

[72] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean, and Mark A DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*, 36(10):983–987, 11 2018.

[73] Blake Pyman, Alireza Sedghi, Shekoofeh Azizi, Kathrin Tyryshkin, Neil Renwick, and Parvin Mousavi. Exploring microRNA regulation of cancer with context-aware deep cancer classifier. *Pac Symp Biocomput*, 24:160–171, 2019.

[74] Sayed Mohammad Ebrahim Sahraeian, Ruolin Liu, Bayo Lau, Karl Podesta, Marghoob Mohiyuddin, and Hugo Y K Lam. Deep convolutional neural networks for accurate somatic mutation detection. *Nat Commun*, 10(1):1041, 03 2019.

[75] Sirajul Salekin, Jianqiu Michelle Zhang, and Yufei Huang. Base-pair resolution detection of transcription factor binding site by deep deconvolutional network. *Bioinformatics*, 34(20):3446–3453, 10 2018.

[76] Arshdeep Sekhon, Ritambhara Singh, and Yanjun Qi. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics*, 34(17):i891–i900, 09 2018.

[77] Seokjun Seo, Minsik Oh, Youngjune Park, and Sun Kim. DeepFam: deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics*, 34(13):i254–i262, 07 2018.

[78] Dipan Shaw, Hao Chen, and Tao Jiang. DeepIsoFun: A deep domain adaptation approach to predict isoform functions. *Bioinformatics*, Dec 2018.

[79] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Sci Rep*, 8(1):15270, Oct 2018.

[80] Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, 18(1):277, May 2017.

[81] Binhua Tang, Zixiang Pan, Kang Yin, and Asif Khateeb. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in Genetics*, 10:214, 2019.

[82] Ameni Trabelsi, Mohamed Chaabane, and Asa Ben-Hur. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics*, 35(14):i269–i277, 07 2019.

[83] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A*, Jul 2017.

[84] Ramzan Umarov, Hiroyuki Kuwahara, Yu Li, Xin Gao, and Victor Solovyev. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics*, 35(16):2730–2737, Aug 2019.

[85] Ramzan Kh Umarov and Victor V Solovyev. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One*, 12(2):e0171410, 2017.

[86] Fan Wang, Pranik Chainani, Tommy White, Jin Yang, Yu Liu, and Benjamin Soibam. Deep learning identifies genome-wide DNA binding sites of long noncoding RNAs. *RNA Biol*, 15(12):1468–1476, 2018.

[87] Linyu Wang, Yuanning Liu, Xiaodan Zhong, Haiming Liu, Chao Lu, Cong Li, and Hao Zhang. DMfold: A novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Front Genet*, 10:143, 2019.

[88] Sheng Wang, Zhen Li, Yizhou Yu, and Jinbo Xu. Folding membrane proteins by deep transfer learning. *Cell Syst*, 5(3):202–211.e3, 09 2017.

[89] Shunfang Wang, Mingyuan Li, Lei Guo, Zicheng Cao, and Yu Fei. Efficient utilization on PSSM combining with recurrent neural network for membrane protein types prediction. *Comput Biol Chem*, 81:9–15, Aug 2019.

[90] Tong Wang, Yanhua Qiao, Wenze Ding, Wenzhi Mao, Yaoqi Zhou, and Haipeng Gong. Improved fragment sampling for ab initio protein structure prediction using deep neural networks. *Nature Machine Intelligence*, August 2019.

[91] Tongxin Wang, Travis S Johnson, Wei Shao, Zixiao Lu, Bryan R Helm, Jie Zhang, and Kun Huang. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol*, 20(1):165, 08 2019.

[92] Wafaa Wardah, M G M Khan, Alok Sharma, and Mahmood A Rashid. Protein secondary structure prediction using neural networks and deep learning: A review. *Comput Biol Chem*, 81:1–8, Aug 2019.

[93] Sarah Webb. Deep learning for biology. *Nature*, 554(7693):555–557, 02 2018.

[94] Guo-Wei Wei. Protein structure prediction beyond AlphaFold. *Nature Machine Intelligence*, August 2019.

[95] Ming Wen, Peisheng Cong, Zhimin Zhang, Hongmei Lu, and Tonghua Li. DeepMirTar: a deep-learning approach for predicting human miRNA targets. *Bioinformatics*, 34(22):3781–3787, 11 2018.

[96] Ryan R Wick, Louise M Judd, and Kathryn E Holt. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*, 20(1):129, Jun 2019.

[97] Hongjie Wu, Chengyuan Cao, Xiaoyan Xia, and Qiang Lu. Unified deep learning architecture for modeling biology sequence. *IEEE/ACM Trans Comput Biol Bioinform*, 15(5):1445–1452, 2018.

[98] Zhihao Xia, Yu Li, Bin Zhang, Zhongxiao Li, Yuhui Hu, Wei Chen, and Xin Gao. DeeReCT-PolyA: a robust and generic deep learning method for PAS identification. *Bioinformatics*, 35(14):2371–2379, Nov 2018.

[99] Yungang Xu, Yongcui Wang, Jiesi Luo, Weiling Zhao, and Xiaobo Zhou. Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to esc fate decision. *Nucleic Acids Res*, 45(21):12100–12112, Dec 2017.

[100] Kazunori D. Yamada and Kengo Kinoshita. De novo profile generation based on sequence context specificity with the long short-term memory network. *BMC Bioinformatics*, 19(1):272:1–272:11, 2018.

[101] Zichao Yan, Eric Lécuyer, and Mathieu Blanchette. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*, 35(14):i333–i342, 07 2019.

[102] Cheng Yang, Longshu Yang, Man Zhou, Haoling Xie, Chengjiu Zhang, May D Wang, and Huaiqiu Zhu. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, 34(22):3825–3834, 11 2018.

[103] Yang Yang, Mingyu Zhou, Qingwei Fang, and Hong-Bin Shen. AnnoFly: annotating Drosophila embryonic images based on an attention-enhanced RNN model. *Bioinformatics*, 35(16):2834–2842, Aug 2019.

[104] Jiaying You, Robert D McLeod, and Pingzhao Hu. Predicting drug-target interaction network using deep learning model. *Comput Biol Chem*, 80:90–101, Jun 2019.

[105] Ning Yu, Zeng Yu, and Yi Pan. A deep learning method for lincRNA detection using auto-encoder algorithm. *BMC Bioinformatics*, 18(Suppl 15):511, Dec 2017.

[106] Haoyang Zeng, Matthew D Edwards, Ge Liu, and David K Gifford. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, 32(12):i121–i127, 06 2016.

[107] Zhao-Hui Zhan, Li-Na Jia, Yong Zhou, Li-Ping Li, and Hai-Cheng Yi. BGFE: A deep learning model for ncRNA-protein interaction predictions based on improved sequence information. *Int J Mol Sci*, 20(4), Feb 2019.

[108] Buzhong Zhang, Jinyan Li, and Qiang Lü. Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics*, 19(1):293, 08 2018.

[109] Jingpu Zhang, Zuping Zhang, Zixiang Wang, Yuting Liu, and Lei Deng. Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Bioinformatics*, 34(10):1750–1757, 05 2018.

[110] Juhua Zhang, Wenbo Peng, and Lei Wang. LeNup: learning nucleosome positioning from DNA sequences with improved convolutional neural networks. *Bioinformatics*, 34(10):1705–1712, 05 2018.

[111] Sai Zhang, Hailin Hu, Jingtian Zhou, Xuan He, Tao Jiang, and Jianyang Zeng. Analysis of ribosome stalling and translation elongation dynamics by deep learning. *Cell Syst*, 5(3):212–220.e6, 09 2017.

[112] Shao-Wu Zhang, Ya Wang, Xi-Xi Zhang, and Jia-Qi Wang. Prediction of the RBP binding sites on lncRNAs using the high-order nucleotide encoding convolutional neural network. *Anal Biochem*, 583:113364, Jul 2019.

[113] Zijun Zhang, Zhicheng Pan, Yi Ying, Zhijie Xie, Samir Adhikari, John Phillips, Russ P Carstens, Douglas L Black, Yingnian Wu, and Yi Xing. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat Methods*, 16(4):307–310, 04 2019.

[114] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R Shayakhmetov, Alexander Zhebrak, Lidiya I Minaeva, Bogdan A Zagribelnyy, Lennart H Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*, Sep 2019.

[115] Jian Zhou, Christopher Y Park, Chandra L Theesfeld, Aaron K Wong, Yuan Yuan, Claudia Scheckel, John J Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B Darnell, and Olga G Troyanskaya. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet*, 51(6):973–980, Jun 2019.

[116] Jian Zhou and Olga G. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 745–753. JMLR.org, 2014.

[117] Jiyun Zhou, Hongpeng Wang, Zhishan Zhao, Ruifeng Xu, and Qin Lu. CNNH_PSS: protein 8-class secondary structure prediction by convolutional neural network with highway. *BMC Bioinformatics*, 19(Suppl 4):60, 05 2018.

[118] Kai Yin Zhou, Yu Xing Wang, Sheng Zhang, Mina Gachloo, Jin Dong Kim, Qi Luo, Kevin Bretonnel Cohen, and Jing Bo Xia. GOF/LOF knowledge inference with tensor decomposition in support of high order link discovery for gene, mutation and disease. *Math Biosci Eng*, 16(3):1376–1391, 02 2019.

[119] Jianwei Zhu, Sheng Wang, Dongbo Bu, and Jinbo Xu. Protein threading using residue co-variation and deep learning. *Bioinformatics*, 34(13):i263–i273, 07 2018.

[120] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nat Genet*, 51(1):12–18, 01 2019.

[121] Jasper Zuallaert, Fréderic Godin, Mijung Kim, Arne Soete, Yvan Saeys, and Wesley De Neve. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics*, 34(24):4180–4188, 12 2018.

Deep Learning: Word Embeddings

[1] Ehsaneddin Asgari and Mohammad R K Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, 10(11):e0141287, 2015.

[2] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

[3] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(Suppl 1):82, Feb 2019.

[4] Aparajita Dutta, Tushar Dubey, Kusum Kumari Singh, and Ashish Anand. SpliceVec: Distributed feature representations for splice junction prediction. *Comput Biol Chem*, 74:434–441, Jun 2018.

[5] Md-Nafiz Hamid and Iddo Friedberg. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics*, 35(12):2009–2016, Jun 2019.

[6] Romain Menegaux and Jean-Philippe Vert. Continuous embeddings of dna sequencing reads and application to metagenomics. *J Comput Biol*, 26(6):509–518, Jun 2019.

[7] Stephen Woloszynek, Zhengqiao Zhao, Jian Chen, and Gail L Rosen. 16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses. *PLoS Comput Biol*, 15(2):e1006721, 02 2019.

[8] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 08 2018.

[9] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(23):4138, 12 2018.

[10] Haoyang Zeng and David K Gifford. DeepLigand: accurate prediction of MHC class I ligands using peptide embedding. *Bioinformatics*, 35(14):i278–i283, 07 2019.

Concept and Rule-based Learning

[1] Tirtharaj Dash, Ashwin Srinivasan, Lovekesh Vig, Oghenejokpeme I. Orhobor, and Ross D. King. Large-scale assessment of deep relational machines. In *ILP*, volume 11105 of *Lecture Notes in Computer Science*, pages 22–37. Springer, 2018.

[2] Alexandre Drouin, Gaël Letarte, Frédéric Raymond, Mario Marchand, Jacques Corbeil, and François Laviolette. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci Rep*, 9(1):4071, Mar 2019.

[3] Jerome Feret and Kim Quyen Ly. Local traces: An over-approximation of the behavior of the proteins in rule-based models. *IEEE/ACM Trans Comput Biol Bioinform*, 15(4):1124–1137, 2018.

[4] Torsten Gross, Matthew J Wongchenko, Yibing Yan, and Nils Blüthgen. Robust network inference using response logic. *Bioinformatics*, 35(14):i634–i642, 07 2019.

[5] Saurav Mallik and Zhongming Zhao. Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data. *Brief Bioinform*, Jan 2019.

Learning Graphs

[1] Seyed Ziaeddin Alborzi, David W Ritchie, and Marie-Dominique Devignes. Computational discovery of direct associations between GO terms and protein domains. *BMC Bioinformatics*, 19(Suppl 14):413, Nov 2018.

[2] Qingfeng Chen, Chaowang Lan, Baoshan Chen, Lusheng Wang, Jinyan Li, and Chengqi Zhang. Exploring consensus RNA substructural patterns using subgraph mining. *IEEE/ACM Trans Comput Biol Bioinform*, 14(5):1134–1146, 2017.

[3] Jimmy Ka Ho Chiu, Tharam S Dillon, and Yi-Ping Phoebe Chen. Large-scale frequent stem pattern mining in rna families. *J Theor Biol*, 455:131–139, Oct 2018.

[4] Fabrizio Costa, Dominic Grün, and Rolf Backofen. GraphDDP: a graph-embedding approach to detect differentiation pathways in single-cell-data using prior class knowledge. *Nat Commun*, 9(1):3685, 09 2018.

[5] Yunchuan Kong and Tianwei Yu. A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics*, 34(21):3727–3737, 11 2018.

[6] Meng Li, Jianmei Zhao, Xuecang Li, Yang Chen, Chenchen Feng, Fengcui Qian, Yuejuan Liu, Jian Zhang, Jianzhong He, Bo Ai, Ziyu Ning, Wei Liu, Xuefeng Bai, Xiaole Han, Zhiyong Wu, Xiue Xu, Zhidong Tang, Qi Pan, Liyan Xu, Chunquan Li, Qiuyu Wang, and Enmin Li. HiFreSP: A novel high-frequency sub-pathway mining approach to identify robust prognostic gene signatures. *Brief Bioinform*, Jul 2019.

[7] Stefan Mautner, Soheila Montaseri, Milad Miladi, Martin Raden, Fabrizio Costa, and Rolf Backofen. ShaKer: RNA SHAPE prediction using graph kernel. *Bioinformatics*, 35(14):i354–i359, 07 2019.

[8] Mohammad Mohebbi, Liang Ding, Russell L Malmberg, Cory Momany, Khaled Rasheed, and Liming Cai. Accurate prediction of human miRNA targets via graph modeling of the miRNA-target duplex. *J Bioinform Comput Biol*, 16(4):1850013, 08 2018.

[9] Aida Mrzic, Pieter Meysman, Wout Bittremieux, Pieter Moris, Boris Cule, Bart Goethals, and Kris Laukens. Grasping frequent subgraph mining for bioinformatics applications. *BioData Min*, 11:20, 2018.

[10] Nicolò Navarin and Fabrizio Costa. An efficient graph kernel method for non-coding RNA functional prediction. *Bioinformatics*, 33(17):2642–2650, Sep 2017.

[11] Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, and Fei Wang. Graph convolutional networks for computational drug development and discovery. *Brief Bioinform*, Jun 2019.

Ensemble Learning

[1] Zhen Cao, Xiaoyong Pan, Yang Yang, Yan Huang, and Hong-Bin Shen. The lncLocator: a subcellular localization predictor for long non-coding rnas based on a stacked ensemble classifier. *Bioinformatics*, 34(13):2185–2194, 07 2018.

[2] Xing Chen, Chi-Chi Zhu, and Jun Yin. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput Biol*, 15(7):e1007209, Jul 2019.

[3] Maria Colomé-Tatché and Fabian J Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, 2018.

[4] Yuming Ma, Yihui Liu, and Jinyong Cheng. Protein secondary structure prediction based on data partition and semi-random subspace method. *Sci Rep*, 8(1):9856, Jun 2018.

[5] Prabina Kumar Meher, Tanmaya Kumar Sahu, Shachi Gahoi, Subhrajit Satpathy, and Atmakuri Ramakrishna Rao. Evaluating the performance of sequence encoding schemes and machine learning methods for splice sites recognition. *Gene*, 705:113–126, Jul 2019.

[6] Hui Peng, Yi Zheng, Zhixun Zhao, Tao Liu, and Jinyan Li. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions. *Bioinformatics*, 34(17):i757–i765, 09 2018.

[7] Anand Pratap Singh, Sarthak Mishra, and Suraiya Jabin. Sequence based prediction of enhancer regions from DNA random walk. *Sci Rep*, 8(1):15912, 10 2018.

[8] Jaswinder Singh, Jack Hanson, Rhys Heffernan, Kuldip Paliwal, Yuedong Yang, and Yaoqi Zhou. Detecting proline and non-proline cis isomers in protein structures from sequences using deep residual ensemble learning. *J Chem Inf Model*, 58(9):2033–2042, 09 2018.

[9] Weijia Su, Xun Gu, and Thomas Peterson. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol Plant*, 12(3):447–460, 03 2019.

[10] Xiaoying Wang, Bin Yu, Anjun Ma, Cheng Chen, Bingqiang Liu, and Qin Ma. Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics*, 35(14):2395–2402, 12 2018.

[11] Jialin Yu, Shaoping Shi, Fang Zhang, Guodong Chen, and Man Cao. PredGly: predicting lysine glycation sites for homo sapiens based on XGboost feature optimization. *Bioinformatics*, 35(16):2749–2756, Aug 2019.

[12] Xiangxiang Zeng, Yue Zhong, Wei Lin, and Quan Zou. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief Bioinform*, Oct 2019.

[13] Long Zhang, Guoxian Yu, Dawen Xia, and Jun Wang. Protein-protein interactions prediction based on ensemble deep neural networks. *Neurocomputing*, 324:10–19, 2019.

[14] Xuan Zhang, Jun Wang, Jing Li, Wen Chen, and Changning Liu. Crlncrc: a machine learning-based method for cancer-related long noncoding rna identification using integrated features. *BMC Med Genomics*, 11(Suppl 6):120, Dec 2018.

[15] Ruiqing Zheng, Min Li, Xiang Chen, Fang-Xiang Wu, Yi Pan, and Jianxin Wang. BiXGBoost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics*, 35(11):1893–1900, Jun 2019.

Learning to rank

[1] Hao Yang, Hao Chi, Wen-Feng Zeng, Wen-Jing Zhou, and Si-Min He. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, 35(14):i183–i190, 07 2019.

[2] Ronghui You, Zihan Zhang, Yi Xiong, Fengzhu Sun, Hiroshi Mamitsuka, and Shanfeng Zhu. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473, 07 2018.

Semi-Supervised Learning

[1] Tahmid F Mehdi, Gurdeep Singh, Jennifer A Mitchell, and Alan M Moses. Variational infinite heterogeneous mixture model for semi-supervised clustering of heart enhancers. *Bioinformatics*, 35(18):3232–3239, Sep 2019.

[2] Heinrich Roder, Carlos Oliveira, Lelia Net, Benjamin Linstid, Maxim Tsypin, and Joanna Roder. Robust identification of molecular phenotypes using semi-supervised learning. *BMC Bioinformatics*, 20(1):273, May 2019.

[3] Ioannis A Tamposis, Konstantinos D Tsirigos, Margarita C Theodoropoulou, Panagiota I Kontou, and Pantelis G Bagos. Semi-supervised learning of Hidden Markov Models for biological sequence analysis. *Bioinformatics*, 35(13):2208–2215, 11 2018.

[4] Joris Tavernier, Jaak Simm, Karl Meerbergen, Jörg Kurt Wegner, Hugo Ceulemans, and Yves Moreau. Fast semi-supervised discriminant analysis for binary classification of large data sets. *Pattern Recognition*, 91:86–99, 2019.

[5] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Comput Methods Programs Biomed*, 166:99–105, Nov 2018.

[6] Haruka Yonemoto, Kiyoshi Asai, and Michiaki Hamada. A semi-supervised learning approach for RNA secondary structure prediction. *Comput Biol Chem*, 57:72–9, Aug 2015.

Data Integration

[1] Hadi Fanaee-T and Magne Thoresen. Multi-insight visualization of multi-omics data via ensemble dimension reduction and tensor factorization. *Bioinformatics*, 35(10):1625–1633, May 2019.

[2] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform*, 19(2):325–340, 03 2018.

[3] Bilal Mirza, Wei Wang, Jie Wang, Howard Choi, Neo Christopher Chung, and Peipei Ping. Machine learning and integrative analysis of biomedical big data. *Genes (Basel)*, 10(2), Jan 2019.

[4] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res*, 46(20):10546–10562, 11 2018.

[5] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res*, 47(2):1044, 01 2019.

[6] Nimrod Rappoport and Ron Shamir. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356, Sep 2019.

[7] Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, 07 2019.

[8] Cen Wu, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. A selective review of multi-level omics data integration using variable selection. *High Throughput*, 8(1), Jan 2019.

Reinforcement Learning

[1] Peter Eastman, Jade Shi, Bharath Ramsundar, and Vijay S Pande. Solving the RNA design problem with reinforcement learning. *PLoS Comput Biol*, 14(6):e1006176, 06 2018.

[2] Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design RNA. *CoRR*, abs/1812.11951, 2018.

Statistics

[1] Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Cambridge University Press, 2018.

Tools and Data Sets

[1] Mohammed AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311, Jun 2019.

[2] Jan P Buchmann and Edward C Holmes. Entrezpy: A Python library to dynamically interact with the NCBI Entrez databases. *Bioinformatics*, May 2019.

[3] Kathleen M Chen, Evan M Cofer, Jian Zhou, and Olga G Troyanskaya. Selene: a PyTorch-based deep learning library for sequence data. *Nat Methods*, 16(4):315–318, 04 2019.

[4] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, Johannes Köster, and Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*, 15(7):475–476, 07 2018.

[5] Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Min*, 10:36, 2017.

[6] Ladislav Rampasek and Anna Goldenberg. TensorFlow: Biology's gateway to deep learning? *Cell Syst*, 2(1):12–4, 01 2016.

[7] Guohua Wang, Ximei Luo, Jianan Wang, Jun Wan, Shuli Xia, Heng Zhu, Jiang Qian, and Yadong Wang. MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res*, 46(D1):D146–D151, 01 2018.

[8] Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R S Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. The Ensembl REST API: Ensembl data for any language. *Bioinformatics*, 31(1):143–5, Jan 2015.

Puzzles or puzzling

[1] Rohan V Koodli, Benjamin Keep, Katherine R Coppess, Fernando Portela, Eterna participants, and Rhiju Das. EternaBrain: Automated RNA design through move sets and strategies from an Internet-scale RNA videogame. *PLoS Comput Biol*, 15(6):e1007059, Jun 2019.

[2] Sara Reardon. How machine learning could keep dangerous DNA out of terrorists' hands. *Nature*, 566(7742):19, 02 2019.

Automated Scientific Discovery

[1] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–9, May 2015.

[2] Ross D. King, Vlad Schuler Costa, Chris Mellingwood, and Larisa N. Soldatova. Automating sciences: Philosophical and social dimensions. *IEEE Technol. Soc. Mag.*, 37(1):40–46, 2018.

[3] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, Andrew Sparkes, Kenneth E Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–9, Apr 2009.

[4] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, Andrew Sparkes, Kenneth E Whelan, and Amanda Clare. Make way for robot scientists. *Science*, 325(5943):945, Aug 2009.

[5] Ross D King, Kenneth E Whelan, Ffion M Jones, Philip G K Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–52, Jan 2004.

[6] Andrew Sparkes, Wayne Aubrey, Emma Byrne, Amanda Clare, Muhammed N Khan, Maria Liakata, Magdalena Markham, Jem Rowland, Larisa N Soldatova, Kenneth E Whelan, Michael Young, and Ross D King. Towards robot scientists for autonomous scientific discovery. *Autom Exp*, 2:1, Jan 2010.

A Scientific journals

This section lists the scientific journals where bioinformatics research is most often published. The numbers in parentheses are the impact factors of these journals. Major contributions and/or interdisciplinary research are published in journals such as the following:

- Nature (40.137)
- Science (37.205)

- Nature Communications (12.353)
- Proceedings of the National Academy of Sciences of the United States of America (PNAS) (9.661)
- PLOS One (2.766)

The following life science journals are known to publish bioinformatics research on a regular basis.

- Nature Reviews Genetics (40.282)
- Cell (31.398)
- Genome Biology (11.908)
- Nucleic Acids Research (11.561)
- Molecular Biology and Evolution (10.217)
- Molecular Systems Biology (8.447)
- GigaScience (7.463)

The following journals are dedicated to bioinformatics research.

- Bioinformatics (5.481)
- Briefings in Bioinformatics (5.134)
- Computational and Structural Biotechnology Journal (4.148)
- PLOS Computational Biology (3.995)
- Database (3.978)
- BMC Bioinformatics (2.213)
- IEEE/ACM Transactions on Computational Biology and Bioinformatics (1.955)
- Bulletin of Mathematical Biology (1.484)
- Computers in Biology and Medicine (2.115)
- Journal of Theoretical Biology (2.049)
- Evolutionary Bioinformatics (1.877)
- Journal of Mathematical Biology (1.846)
- Statistical Applications in Genetics and Molecular Biology (1.77)
- Journal of Proteomics & Bioinformatics (1.57)
- Algorithms for Molecular Biology (1.536)
- Computational Biology and Chemistry (1.331)
- Journal of Data Mining in Genomics & Proteomics (1.16)
- Journal of Computational Biology (1.032)
- Journal of Bioinformatics and Computational Biology (0.931)
- Current Bioinformatics (0.770)

Lists of bioinformatics journals can be found here:

- https://en.wikipedia.org/wiki/List_of_bioinformatics_journals
- https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_bioinformatics

B Resources

- http://www.bioinformatics.org/wiki/journals
- https://en.wikipedia.org/wiki/List_of_bioinformatics_journals

Last Modified: October 29, 2019