BIOL 75302 (phytoinformatics)

Dr. Damon P. Little

The Graduate Center, City University of New York & New York Botanical Garden

718-817-8521; dlittle@nybg.org

https://github.com/dpl10/phytoinformatics2023

[Teams office hours by appointment]

Tuesdays & Wednesdays 2:00–5:00 PM

2023 January 25 through 2023 May 11

all course meetings via Teams

Objectives

This course will provide students of plant organismal biology the computational tools needed to process and extract data from text files; basic POSIX command–line tools; introductory BASH and Python scripting; basic processing and interpretation of DNA/RNA/AA sequences; relational database structure; introductory Simple Query Language (SQL); and basic classification and regression machine learning tasks. By the end of the course you should be:

(1) comfortable using the BASH command–line interface

(2) able to extract and manipulate data in text files/streams at scale using text processing tools and pipes

(3) able to run programs in batch mode in a single user environment

(4) able write basic efficient BASH and Python scripts

(5) able to query and retrieve sequences from GenBank using the API

(6) able to assemble sequencing reads into useful contigs

(7) able to extract useful data from assembled contigs

(8) able to conduct sequence analyses including similarity and feature searches

(9) able to write basic SQL queries for MariaDB

(10) able to design a relational MariaDB database

(11) able to train and use classification, regression, and segmentation machine learning models in TensorFlow

Texts

Abascal, F., R. Zardoya & M. J. Telford. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Research 38: W7–W13.

Allam, A., P. Kalnis & V. Solovyev. 2015. Karect: accurate correction of substitution, insertion and deletion errors for next–generation sequencing data. Bioinformatics 31: 3421–3428.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers & D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller & D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.

Arbuthnott, J. 1710. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. Philosophical Transactions 27: 186–190.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25–29.

Beyer, L., X. Zhai, A. Royer, L. Markeeva, R. Anil & A. Kolesnikov. 2021. Knowledge distillation:a good teacher is patient and consistent. arXiv 2106.05237.

Buchfink, B., K. Reuter & H.-G. Drost. 2021. Sensitive protein alignments at tree–of–life scale using DIAMOND. Nature Methods 18: 366–368.

Chen, D., F. Hu, G. Nian & T. Yang. 2020. Deep residual learning for nonlinear regression. Entropy 22: 193.

Codd, E. F. 1970. A relational model of data for large shared data banks. Communications of the ACM 13: 377–387.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson & T. R. Gingeras. 2013. STAR: ultrafast universal RNA–seq aligner. Bioinformatics 29: 15–21.

Dwibedi, D., Y. Aytar, J. Tompson, P. Sermanet & A. Zisserman. 2021. With a little help from my friends: nearest–neighbor contrastive learning of visual representations. arXiv 2104.14548.

Eddy, S. 2011. Accelerated profile hmm searches. PLOS Computational Biology 7: e1002195.

Eddy, S. R. 2004. What is a hidden markov model? Nature Biotechnology 22: 1315–1316.

Edgar, R. C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

———. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.

Eitner, K., U. Koch, T. Gawęda & J. Marciniak. 2010. Statistical distribution of amino acid sequences: a proof of Darwinian evolution. Bioinformatics 26: 2933–2935.

Hassani, A., S. Walton, N. Shah, A. Abuduweili, J. Li & H. Shi. 2021. Escaping the big data paradigm with compact transformers. arXiv 2104.05704.

Hinton, G., O. Vinyals & J. Dean. 2015. Distilling the knowledge in a neural network. arXiv 1503.02531.

Iandola, F., S. Han, M. Moskewicz, K. Ashraf, W. Dally & K. Keutzer. 2016. SqueezeNet: AlexNet–level accuracy with 50x fewer parameters and <0.5mb model size. arXiv 1602.07360.

Jackman, S. D., B. P. Vandervalk, H. Mohamadi, J. Chu, S. Yeo, S. A. Hammond, G. Jahesh, H. Khan, L. Coombe, R. L. Warren & I. Birol. 2017. ABySS 2.0: resource–efficient assembly of large genomes using a Bloom filter. Genome Research 27: 768–777.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez & S. Hunter. 2014. Interproscan 5: genome–scale protein function classification. Bioinformatics 30: 1236–1240.

Katoh, K., K. Misawa, K. Kuma & T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059–3066.

Katoh, K. & H. Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9: 286–298.

Khandelwal, G. & J. Bhyravabhotla. 2010. A phenomenological model for predicting melting temperatures of DNA sequences. PLOS ONE 5: e12433.

Kulmanov, M. & R. Hoehndorf. 2020. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36: 422–429.

Lassmann, T. & E. L. Sonnhammer. 2005. Kalign—an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6: 298.

LeCun, Y., Y. Bengio & G. Hinton. 2015. Deep learning. Nature 521: 436–444.

Li, Y., C. Huang, L. Ding, Z. Li, Y. Pan & X. Gao. 2019. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods 166: 4–21. Deep Learning in Bioinformatics.

Lin, Y., J. Li, H. Shen, L. Zhang, C. J. Papasian & H.-W. Deng. 2011. Comparative studies of de novo assembly tools for next–generation sequencing technologies. Bioinformatics 27: 2031–2037.

Matthes, E. 2019. Python Crash Course. 2nd ed. No Starch Press, San Francisco.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller & E. Teller. 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21: 1087–1092.

Min, S., B. Lee & S. Yoon. 2016. Deep learning in bioinformatics. Briefings in Bioinformatics 18: 851–869.

Needleman, S. B. & C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48: 443–453.

Pertsemlidis, A. & J. W. Fondon. 2001. Having a BLAST with bioinformatics (and avoiding BLAST-phemy). Genome Biology 2: 1–10.

Phillips, A., D. Janies & W. Wheeler. 2000. Multiple sequence alignment in phylogenetic analysis. Molecular Phylogenetics and Evolution 16: 317–330.

Qiu, Y., Y. Liu, S. Li & J. Xu. 2021. MiniSeg: an extremely minimum network for efficient COVID-19 segmentation. AAAI Conference on Artificial Intelligence 35: 6.

Rychlik, W., W. J. Spencer & R. E. Rhoads. 1990. Optimization of the annealing temperature for DNA amplification *in vitro*. Nucleic Acids Research 18: 6409–6412.

Schuler, G. D. 1997. Sequence mapping by electronic PCR. Genome Research 7: 541–550.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones & I. Birol. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19: 1117–1123.

Smith, T. F. & M. S. Waterman. 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147: 195–197.

Sobell, M. G. & M. Helmke. 2018. A practical guide to LINUX commands, editors, and shell programming. 4th ed. Addison–Wesly, Boston.

Szegedy, C., S. Ioffe, V. Vanhoucke & A. Alemi. 2016. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv 1602.07261.

Ter-Hovhannisyan, V., A. Lomsadze, Y. O. Chernoff & M. Borodovsky. 2008. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Research 18: 1979–1990.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser & I. Polosukhin. 2017. Attention is all you need. arXiv 1706.03762.

Warren, R. L., G. G. Sutton, S. J. M. Jones & R. A. Holt. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23: 500–501.

Yu, W., M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng & S. Yan. 2021. MetaFormer is actually what you need for vision. arXiv 2111.11418.

Zhang, H., M. Cisse, Y. Dauphin & D. Lopez-Paz. 2018. Mixup: beyond empirical risk minimization. arXiv 1710.09412.

Grading

weekly laboratory exercises (5% each, 60% total, the two lowest exercise scores will be dropped)

take–home final exam (40%)

Exam questions are mostly based on the laboratory exercises. Therefore it is very important that the laboratory exercises be completed.

Assignments are due at the beginning of class on the date specified. No late assignments will be accepted.

Course schedule

Week 1 (January 25). Overview of grading, exams, and other logistics; phytoinformatics defined; overview of LINUX systems and distributions; the BASH shell. **Readings:** Arbuthnott (1710); Eitner et al. (2010); Sobell & Helmke (2018: chapters 1 & 2). **Laboratory:** installing Ubuntu LINUX & BASH basics.

Week 2 (January 31 & February 1). Software installation; moving data: files, streams, and pipes. **Readings:** Sobell & Helmke (2018: chapters 3, 4, 5, & 8; appendix C). **Laboratory:** installing software.

Week 3 (February 7 & 8). BASH: scripts, parallelism, and job control; the power of command–line text tools. **Readings:** Sobell & Helmke (2018: chapters 10 & 14); What is code? (https://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/). **Laboratory:** BASH scripts, parallelism, and job control.

Week 4 (February 14 & 15). Basic scripting in AWK, Perl, and Python3. **Readings:** Sobell & Helmke (2018: chapters 11 & 12; appendix A). **Laboratory:** LINUX command–line text processing tools.

Week 5 (February 21 & 22). Basic Python3, data structures, operators, and conditionals; DNA–DNA binding. **Readings:** Khandelwal & Bhyravabhotla (2010); Matthes (2019: chapters 1, 2, 3, 4, 5, & 6); Rychlik et al. (1990). **Laboratory:** beginning Python3 (PCR primer annealing temperature calculations).

Week 6 (February 28 & March 1). Python3 loops and functions; DNA/RNA/AA sequence search. **Laboratory:** more beginning Python3 (BLAST & FASTA processing). Matthes (2019: chapters 7, 8, & 10)

Week 7 (March 7 & 8). DNA/RNA/AA sequence search and alignment. **Readings:**; Altschul et al. (1990, 1997); Buchfink et al. (2021); Eddy (2004, 2011); Metropolis et al. (1953); Needleman & Wunsch (1970); Pertsemlidis & Fondon (2001); Schuler (1997); Smith & Waterman (1981). **Laboratory:** BLAST and e-PCR.

Week 8 (March 14 & 15). More DNA/RNA/AA sequence alignment. **Readings:** Abascal et al. (2010); Edgar (2004a,b); Katoh et al. (2002); Katoh & Toh (2008); Lassmann & Sonnhammer (2005); Phillips et al. (2000). **Laboratory:** multiple sequence alignment.

Week 9 (March 21 & 22). Raw DNA sequence quality, processing, assembly, and mapping. **Readings:** Allam et al. (2015); Dobin et al. (2013); Jackman et al. (2017); Lin et al. (2011); Simpson et al. (2009); Warren et al. (2007). **Laboratory:** sequence assembly and remapping.

Week 10 (March 28 & 29). Open reading frame identification, GO, and InterPro. **Readings:** Ashburner et al. (2000); Jones et al. (2014); Kulmanov & Hoehndorf (2020); Ter-Hovhannisyan et al. (2008). **Laboratory:** getting GO.

Week 11 (April 4). An overview of database types, the structure of relational databases, and introduction to SQL. **Readings:** Codd (1970); Sobell & Helmke (2018: chapter 13). **Laboratory:** basic MariaDB.

Week 12 (April 18 & 19). SQL queries of relational databases. **Readings:** the MariaDB manual (https://mariadb.com/kb/en/documentation/). **Laboratory:** intermediate MariaDB.

Week 13 (April 25 & 26). Machine learning input, output, layers, and basic training. **Readings:** the TensorFlow manual (https://www.tensorflow.org/guide); LeCun et al. (2015); Min et al. (2016); Li et al. (2019). **Laboratory:** TensorFlow image classification.

Week 14 (May 2 & 3). Intermediate training, basic model structures, and TensorFlow. **Readings:** Beyer et al. (2021); Chen et al. (2020); Dwibedi et al. (2021); Hassani et al. (2021); Hinton et al. (2015); Zhang et al. (2018). **Laboratory:** TensorFlow sequence regression.

Week 15 (May 9 & 10). Intermediate model structures; Semester review. **Readings:** Iandola et al. (2016); Qiu et al. (2021); Szegedy et al. (2016); Vaswani et al. (2017); Yu et al. (2021). **Laboratory:** TensorFlow image segmentation.

Take home final exam distributed May 10, due May 23.