CSI5126. Algorithms in Bioinformatics
Review and presentation of a scientific paper — Fall 2018

Instructor: Marcel Turcotte

Version of October 29, 2018

[PDF]

1 Deadline

- **Nov 1, 6, 8**: 20 minutes presentation.

2 Learning outcomes

- **Paraphrase** scientific results
- **Communicate** technical information
- **Develop** good reading habits

3 Directives

Papers in (refereed) journals and conference proceedings are the main vehicles for communicating scientific information. You must select a publication that introduces an **algorithm** or **data structure** to tackle a specific bioinformatics problem. Sometimes the details for the algorithms will be found in the supplement section available on the Web.

The scientic journal *PLOS Computational Biology* has an excellent series of articles entitled "Ten simple rules for…". It touches subjects as writing, starting a company, obtaining funding, etc. In the latest article, the author speaks about developing good reading habits. touches a variei

- Mndez, M. Ten simple rules for developing good reading habits during graduate school and beyond. *PLoS Comput Biol* **14**, e1006467 (2018).

3.1 Deliverable

- One or two-page summary of the publication
- 15-minute presentation + 5-minute for the questions

3.2 Selected publications

You must select a publication in a distinct area than that of your project. Below you will find a list of publications. You are welcomed to propose publications outside of the list. In an appendix, I am including a list of the major journals where bioinformatics research is published. Use the following form to give the reference of the scientific paper that you would like to present:

- https://docs.google.com/document/d/1_ImpmRGSXiFazjC573K4fCADjOYTbnko3jGrnGbWMIc/edit?usp=sharing

- **String algorithms**

   1. Louza, F. A., Gog, S., & Telles, G. P. (2017). Inducing enhanced suffix arrays for string collections. *Theoretical Computer Science*, **678**, 2239. http://doi.org/10.1016/j.tcs.2017.03.039

- **Short read alignment**

   1. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (Oxford, England), 29(1), 1521. http://doi.org/10.1093/bioinformatics/bts635
   2. Mller, A., Hundt, C., Hildebrandt, A., Hankeln, T. & Schmidt, B. MetaCache: Context-aware classification of metagenomic reads using minhashing. Bioinformatics (2017). http://doi.org/10.1093/bioinformatics/btx520
   3. A. Limasset, B. Cazaux, E. Rivals, and P. Peterlongo, Read mapping on de Bruijn graphs, *BMC bioinformatics* (2016), vol. 17, no. 1, p. 237.
   4. Bu, D. & Tang, H. Quasispecies reconstruction based on vertex coloring algorithms. in 6366 (IEEE, 2014). http://doi.org/10.1109/BIBM.2014.6999128
   5. T. Beller, S. Gog, E. Ohlebusch, and T. Schnattinger, Computing the longest common prefix array based on the Burrows-Wheeler transform, *Journal of Discrete Algorithms* (2013), vol. 18, pp. 2231.
   6. L. Huang, V. Popic, and S. Batzoglou, Short read alignment with populations of genomes, *Bioinformatics* (2013), vol. 29, no. 13, pp. i36170.
   7. Y. Liao, G. K. Smyth, and W. Shi, The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote., *Nucleic Acids Res* (2013), vol. 41, no. 10, pp. e108e108.
   8. H. Lee and M. C. Schatz, Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score, Bioinformatics, vol. 28, no. 16, pp. 20972105, 2012.
   9. M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno, SHRiMP2: Sensitive yet Practical Short Read Mapping, *Bioinformatics* (2011), vol. 27, no. 7, pp. 10111012.
   10. S. Misra, A. Agrawal, W. K. Liao, and A. Choudhary, Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing, *Bioinformatics* (2011), vol. 27, no. 2, pp. 189195.
   11. X. Yang, S. Aluru, and K. S. Dorman, Repeat-aware modeling and correction of short read errors, BMC bioinformatics (2011), vol. 12, no. 1, p. S52, 2011.
   12. K. Daily, P. Rigor, S. Christley, X. Xie, and P. Baldi, Data structures and compression algorithms for high-throughput sequencing technologies, *BMC bioinformatics* (2010), vol. 11, p. 514.
   13. Langmead et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* (2009) vol. 10 (3) pp. R25

- **Genome assembly**

   1. Padovani de Souza, K., Setubal, J. C., Ponce de Leon F de Carvalho, A. C., Oliveira, G., Chateau, A., & Alves, R. (2018). Machine learning meets genome assembly. *Briefings in Bioinformatics*, **3**(6), 349. http://doi.org/10.1093/bib/bby072
   2. Di Genova, A., Ruz, G. A., Sagot, M.-F., & Maass, A. (2018). Fast-SG: an alignment-free algorithm for hybrid assembly. *GigaScience*, 7(5). http://doi.org/10.1093/gigascience/giy048

- **Faster sequence alignment**

   1. Charalampopoulos, P., Crochemore, M., Fici, G., Mercas, R. & Pissis, S. R. Alignment-free sequence comparison using absent words. *Information and Computation* **262**, 5768 (2018).

2. B. Liu, D. Guan, M. Teng, and Y. Wang, rHAT: fast alignment of noisy long reads with regional hashing., *Bioinformatics* (2016), vol. 32, no. 11, pp. 16251631.
3. Newberg. Memory-efficient dynamic programming backtrace and pairwise local sequence alignment. *Bioinformatics* (2008) vol. 24 (16) pp. 1772-8
4. M. Cameron, Y. Bernstein, and H. E. Williams. (2007) Clustered sequence representation for fast homology search. *J Comput Biol*, 14(5):594–614.
5. M. Cameron and H. E. Williams. (2007) Comparing compressed sequences for faster nucleotide blast searches. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 4(3):349–64.
6. Cameron M, Williams HE, Cannane A. (2006) A deterministic finite automaton for faster protein hit detection in BLAST. J Comput Biol. 2006 May;13(4):965-78.
7. Rasmussen et al. Efficient q-gram filters for finding all epsilon-matches over a given length. *J Comput Biol* (2006) vol. 13 (2) pp. 296-308
8. X. Cui, T. Vinar, B. Brejová, D. Shasha, and M. Li. (2007) Homology search for genes. *Bioinformatics*, 23(13):i97–103.
9. Ning Z, Cox AJ, Mullikin JC. (2001) SSAHA: a fast search method for large DNA databases. Genome Res. 11(10):1725-9.
10. Myers G, Durbin R. (2003) A table-driven, full-sensitivity similarity search algorithm. J Comput Biol. 2003;10(2):103-17.
11. Itoh M, Goto S, Akutsu T, Kanehisa M. (2005) Fast and accurate database homology search using upper bounds of local alignment scores. Bioinformatics. 2005 Apr 1;21(7):912-21.
12. Zhang H. (2003) Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. Bioinformatics. 2003 Jul 22;19(11):1391-6.
13. W. J. Kent. (2002) Blat–the blast-like alignment tool. *Genome Res*, 12(4):656–64.

- **Pairwise alignment of genomic sequences**

    1. Abouelhoda et al. CoCoNUT: an efficient system for the comparison and analysis of genomes. *BMC Bioinformatics* (2008) vol. 9 (1) pp. 476
    2. A. C.-C. Shih and W.-H. Li. (2003) GS-Aligner: a novel tool for aligning genomic sequences using bit-level operations. *Mol Biol Evol*, 20(8):1299–309.
    3. U. Schulze, B. Hepp, C. S. Ong, and G. Rätsch. (2007) Palma: mRNA to genome alignments using large margin algorithms. *Bioinformatics*, 23(15):1892–900, Aug 2007.
    4. Delcher AL, Phillippy A, Carlton J, Salzberg SL. (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002 Jun 1;30(11):2478-83.
    5. A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg. (1999) Alignment of whole genomes. *Nucleic Acids Res*, 27(11):2369–76, Jun 1999.
    6. Kahveci T, Ljosa V, Singh AK. (2004) Speeding up whole-genome alignment by indexing frequency vectors. Bioinformatics. 2004 Sep 1;20(13):2122-34.
    7. Kalafus KJ, Jackson AR, Milosavljevic A. (2004) Pash: efficient genome-scale sequence anchoring by Positional Hashing. Genome Res. 2004 Apr;14(4):672-8.
    8. Huang W, Umbach DM, Li L. (2006) Accurate anchoring alignment of divergent sequences. Bioinformatics. 2006 Jan 1;22(1):29-34.
    9. C. N. Dewey, P. M. Huggins, K. Woods, B. Sturmfels, and L. Pachter. (2006) Parametric alignment of drosophila genomes. *PLoS Comput Biol*, 2(6):e73.

- **Median string problem**

   1. J. Abreu and J. R. Rico-Juan, A new iterative algorithm for computing a quality approximate median of strings based on edit operations, *Pattern Recognition Letters* (2014), vol. 36, no. 1, pp. 7480.
   2. F. Hufsky, L. Kuchenbecker, K. Jahn, J. Stoye, and S. Bocker, Swiftly computing center strings., *BMC bioinformatics* (2011), vol. 12, p. 106.
   3. N.-P. D. Nguyen, S. Mirarab, K. Kumar, and T. Warnow, Ultra-large alignments using phylogeny-aware profiles., *Genome Biol* (2015), vol. 16, no. 1, p. 124.
   4. Yue and Tang. A Divide-and-Conquer Implementation of Three Sequence Alignment and Ancestor Inference. *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on (2007)* pp. 143 - 150
   5. F. Nicolas and E. Rivals, Hardness results for the center and median string problems under the weighted and unweighted edit distances, *Journal of Discrete Algorithms* (2005), vol. 3, no. 2, pp. 390415.

- **Seeded alignment methods**

   1. Hahn, L., Leimeister, C.-A., Ounit, R., Lonardi, S. & Morgenstern, B. rasbhari : Optimizing Spaced Seeds for Database Searching, Read Mapping and Alignment-Free Sequence Comparison. PLoS Comput Biol 12, e1005107 (2016).
   2. H. Xin, S. Nahar, R. Zhu, J. Emmons, G. Pekhimenko, C. Kingsford, C. Alkan, and O. Mutlu, Optimal seed solver: optimizing seed selection in read mapping., *Bioinformatics* (2016), vol. 32, no. 11, pp. 16321642.
   3. Flannick J, Batzoglou S. (2005) Using multiple alignments to improve seeded local alignment algorithms. Nucleic Acids Res. 2005 Aug 12;33(14):4563-77.
   4. Ma B, Tromp J, Li M. (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics. 2002 Mar;18(3):440-5.
   5. Li M, Ma B, Kisman D, Tromp J. (2003) PatternHunter II: highly sensitive and fast homology search. Genome Inform. 2003;14:164-75.
   6. Gotea V, Veeramachaneni V, Makalowski W. (2003) Mastering seeds for genomic size nucleotide BLAST searches. Nucleic Acids Res. 2003 Dec 1;31(23):6935-41.

- **Applications of suffix trees and other indexing techniques**

   1. Sarkar, H. & Patro, R. Quark enables semi-reference-based compression of RNA-seq data. Bioinformatics (2017). doi:10.1093/bioinformatics/btx428
   2. Cheng, H., Wu, M. & Yun, X. FMtree: A fast locating algorithm of FM-indexes for genomic data. Bioinformatics (2017). doi:10.1093/bioinformatics/btx596
   3. R. Rahn, D. Weese, and K. Reinert, Journaled string tree-a scalable data structure for analyzing thousands of similar genomes on your laptop., *Bioinformatics* (2014), vol. 30, no. 24, pp. 34993505.
   4. Phoophakdee and Zaki. TRELLIS+: an effective approach for indexing genome-scale sequences using suffix trees. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* (2008) pp. 90-101
   5. Khan et al. A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays. *Bioinformatics* (2009) vol. 25 (13) pp. 1609-16
   6. Ohlebusch and Kurtz. Space efficient computation of rare maximal exact matches between multiple sequences. *J Comput Biol* (2008) vol. 15 (4) pp. 357-77
   7. Herold et al. Efficient computation of absent words in genomic sequences. *BMC Bioinformatics* (2008) vol. 9 pp. 167

- **Statistics**

    1. Afreixo, V., Rodrigues, J. M. O. S., Bastos, C. A. C., Tavares, A. H. M. P. & Silva, R. M. Exceptional Symmetry by Genomic Word. Interdiscip Sci Comput Life Sci 9, 1423 (2016).
    2. Liu, S. S., Hockenberry, A. J., Lancichinetti, A., Jewett, M. C. & Amaral, L. A. N. NullSeq: A Tool for Generating Random Coding Sequences with Desired Amino Acid and GC Contents. PLoS Comput Biol 12, e1005184 (2016).
    3. G. Peris and A. Marzal, Statistical significance of normalized global alignment., *J Comput Biol* (2014), vol. 21, no. 3, pp. 257268.
    4. Quinn, T. & Sinkala, Z. A direct method for computing extreme value (Gumbel) parameters for gapped biological sequence alignments. International journal of bioinformatics research and applications 10, 177189 (2014).
    5. P. Ortet and O. Bastien, Where does the alignment score distribution shape come from?, Evol Bioinform Online (2010), vol. 6, no. 6, pp. 159187.
    6. Bastien and Marchal. Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores. *BMC Bioinformatics* (2008) vol. 9 pp. 332
    7. Newberg. Significance of gapped sequence alignments. *J Comput Biol* (2008) vol. 15 (9) pp. 1187-94
    8. S. Wolfsheimer, B. Burghardt, and A. K. Hartmann. (2007) Local sequence alignments statistics: deviations from gumbel statistics in the rare-event tail. *Algorithms for molecular biology : AMB*, 2:9.
    9. G. Landan and D. Graur. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, 24(6):1380–3, Jun 2007.
    10. D. Metzler. (2006) Robust e-values for gapped local alignments. *J Comput Biol*, 13(4):882–96, May 2006.
    11. A. Y. Mitrophanov and M. Borodovsky. (2006) Statistical significance in biological sequence analysis. *Brief Bioinform*, 7(1):2–24, Mar 2006.

- **Profile alignment methods**

    1. E. Giaquinta, S. Grabowski, and E. Ukkonen, Fast Matching of Transcription Factor Motifs Using Generalized Position Weight Matrix Models, *J Comput Biol* (2013), vol. 20, no. 9, pp. 621630.
    2. Pizzi and Ukkonen. Fast profile matching algorithms - A survey. *Theoretical Computer Science* (2008) vol. 395 (2-3) pp. 137-157
    3. M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz. (2006) Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7:389.
    4. Yona G, Levitt M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol. 2002 Feb 1;315(5):1257-75.

- **Rearrangements**

    1. M. Brudno, S. Malde, A. Poliakov, C. B. Do, O. Couronne, I. Dubchak, and S. Batzoglou. (2003) Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:i54–62.
    2. T. M. Phuong, C. B. Do, R. C. Edgar, and S. Batzoglou. Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic Acids Res*, 34(20):5932–42, Jan 2006.

- **Palindromes**

    - Kim, H. & Han, Y.-S. OMPPM: online multiple palindrome pattern matching. NCBI. Bioinformatics 32, 11511157 (2016).

- **Repetitive elements or words**

    1. Tavares, A. H. M. P. et al. DNA word analysis based on the distribution of the distances between symmetric words. **Sci Rep 7**, 127 (2017).
    2. Pickett, B. D., Miller, J. B. & Ridge, P. G. Kmer-SSR: a fast and exhaustive SSR search algorithm. Bioinformatics (2017). doi:10.1093/bioinformatics/btx538
    3. M. Federico, P. Peterlongo, N. Pisanti, and M.-F. Sagot, Rime: Repeat identification, *Discrete Applied Mathematics* (2014), vol. 163, no. 3, pp. 275286.
    4. Reneker J, Shyu CR. (2005) Refined repetitive sequence searches utilizing a fast hash function and cross species information retrievals. BMC Bioinformatics. 2005 May 3;6:111.
    5. Edgar RC, Myers EW. (2005) PILER: identification and classification of genomic repeats. Bioinformatics. 2005 Jun;21 Suppl 1:i152-8.
    6. Price AL, Jones NC, Pevzner PA. (2005) De novo identification of repeat families in large genomes. Bioinformatics. 2005 Jun;21 Suppl 1:i351-8.
    7. Sagot MF, Myers EW. (1998) Identifying satellites and periodic repetitions in biological sequences. J Comput Biol. 1998 Fall;5(3):539-53.
    8. Martin DE. (2006) The Exact Joint Distribution of the Sum of Heads and Apparent Size Statistics of a “Tandem Repeats Finder” Algorithm. Bull Math Biol. 2006 Aug
    9. Delgrange O, Rivals E. (2004) STAR: an algorithm to Search for Tandem Approximate Repeats. Bioinformatics. 2004 Nov 1;20(16):2812-20. Epub 2004 Jun 4.
    10. Krishnan A, Tang F. (2004) Exhaustive whole-genome tandem repeats search. Bioinformatics. 2004 Nov 1;20(16):2702-10. Epub 2004 May 14.
    11. Parisi V, De Fonzo V, Aluffi-Pentini F. (2003) STRING: finding tandem repeats in DNA sequences. Bioinformatics. 2003 Sep 22;19(14):1733-8.
    12. Hauth AM, Joseph DA. (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinformatics. 2002;18 Suppl 1:S31-7.
    13. Adebiyi EF, Jiang T, Kaufmann M. (2001) An efficient algorithm for finding short approximate non-tandem repeats. Bioinformatics. 2001;17 Suppl 1:S5-S12.
    14. Landau GM, Schmidt JP, Sokol D. (2001) An algorithm for approximate tandem repeats. J Comput Biol. 2001;8(1):1-18.
    15. Apostolico A, Bock ME, Lonardi S, Xu X. (2001) Efficient detection of unusual words. J Comput Biol. 2000 Feb-Apr;7(1-2):71-94.
    16. Allison L, Stern L, Edgoose T, Dix TI. (2000) Sequence complexity for biological sequence analysis. Comput Chem. 2000 Jan;24(1):43-55.

- **Sequence motifs**

    1. Tong, H., Schliekelman, P. & Mrzek, J. Unsupervised statistical discovery of spaced motifs in prokaryotic genomes. *BMC Genomics* **18**, (2017).
    2. Wang, X., Lin, P. & Ho, J. W. K. Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest. **BMC Genomics 19**, (2018).
    3. liopoulos, C. S., Mohamed, M., Pissis, S. P. & Vayani, F. in *Combinatorial Pattern Matching* **11147**, 191205 (Springer International Publishing, 2018).
    4. Pan, X. & Shen, H.-B. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* **34**, 34273436 (2018).
    5. Salekin, S., Zhang, J. M. & Huang, Y. Base-pair resolution detection of transcription factor binding site by deep deconvolutional network. *Bioinformatics* **34**, 34463453 (2018).

- **Multiple sequence alignment**

    1. N.-P. D. Nguyen, S. Mirarab, K. Kumar, and T. Warnow, Ultra-large alignments using phylogeny-aware profiles., *Genome Biol*, vol. 16, no. 1, p. 124, 2015.

2. Q. Zou, Q. Hu, M. Guo, and G. Wang, HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy., *Bioinformatics* (2015), vol. 31, no. 15, pp. 24752481.
3. S. Mirarab, N. Nguyen, S. Guo, L.-S. Wang, J. Kim, and T. Warnow, PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences, *J Comput Biol* (2015), vol. 22, no. 5, pp. 377386.
4. Bradley et al. Fast statistical alignment. *PLoS Comput Biol* (2009) vol. 5 (5) pp. e1000392
5. C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–40.
6. A. S. Konagurthu and P. J. Stuckey. (2006) Optimal sum-of-pairs multiple sequence alignment using incremental carrillo and lipman bounds. *J Comput Biol*, 13(3):668–85.
7. T. J. Wheeler and J. D. Kececioglu. (2007) Multiple alignment by aligning alignments. *Bioinformatics*, 23(13):i559–68.
8. M. Kruspe and P. F. Stadler. (2007) Progressive multiple sequence alignments from triplets. *BMC Bioinformatics*, 8:254, Jul 2007.
9. I. Elias. (2006) Settling the intractability of multiple alignment. *J Comput Biol*, 13(7):1323–39.
10. Higgins DG, Blackshields G, Wallace IM. (2005) Mind the gaps: progress in progressive alignment. Proc Natl Acad Sci U S A. 2005 Jul 26;102(30):10411-2. 18.
11. Loytynoja A, Goldman N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 2005 Jul 26;102(30):10557-62.
12. Pei J, Sadreyev R, Grishin NV. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics. 2003 Feb 12;19(3):427-8.
13. Hudek AK, Brown DG. (2005) Ancestral sequence alignment under optimal conditions. BMC Bioinformatics. 2005 Nov 17;6:273.
14. Notredame C, Higgins DG, Heringa J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000 Sep 8;302(1):205-17.
15. Lassmann T, Sonnhammer EL. (2005) Kalign–an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005 Dec 12;6:298.
16. Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792-7.
17. Grasso C, Lee C. (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 2004 Jul 10;20(10):1546-56.

- **Multiple sequence alignment of genomic sequence**

1. Dubchak et al. Multiple whole-genome alignments without a reference organism. *Genome Res* (2009) vol. 19 (4) pp. 682-9
2. Paten et al. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics* (2009) vol. 25 (3) pp. 295-301
3. C. Wang and E. J. Lefkowitz. (2005) Genomic multiple sequence alignments: refinement using a genetic algorithm. *BMC Bioinformatics*, 6:200.
4. M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–15.
5. Bray N, Pachter L. (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 2004 Apr;14(4):693-9.
6. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E; NISC Comparative Sequencing Program; Green ED, Sidow A, Batzoglou S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 2003 Apr;13(4):721-31.

7. M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. C. S. Program, E. D. Green, A. Sidow, and S. Batzoglou. (2003) Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Res*, 13(4):721–31.
8. Brudno M, Chapman M, Gottgens B, Batzoglou S, Morgenstern B. (2003) Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics. 2003 Dec 23;4:66.
9. Wang C, Lefkowitz EJ. (2005) Genomic multiple sequence alignments: refinement using a genetic algorithm. BMC Bioinformatics. 2005 Aug 8;6:200.
10. Brudno M, Chapman M, Gottgens B, Batzoglou S, Morgenstern B. (2003) Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics. 2003 Dec 23;4:66.
11. Hohl M, Kurtz S, Ohlebusch E. (2002) Efficient multiple genome alignment. Bioinformatics. 2002;18 Suppl 1:S312-20.
12. Zhang Y, Waterman MS. (2003) An Eulerian path approach to global multiple alignment for DNA sequences. J Comput Biol. 2003;10(6):803-19.

- **Graph algorithms**

1. Jain, S., Bayrak, C. S., Petingi, L. & Schlick, T. Dual graph partitioning highlights a small group of pseudoknot-containing RNA submotifs. *Genes* **9**, (2018).
2. Turner, I., Garimella, K. V., Iqbal, Z. & McVean, G. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics* **34**, 25562565 (2018).

- **Machine learning applications in bioinformatics or that could be applicable to bioinformatics**

1. Bacciu, D., Micheli, A. & Sperduti, A. Generative Kernels for Tree-Structured Data. *IEEE Trans. Neural Netw. Learning Syst.* **29**, 49324946.
2. Dal Pal, A., Dovier, A., Formisano, A. & Pontelli, E. Exploring life: answer set programming in bioinformatics. *Declarative Logic Programming: Theory, Systems, and Applications* 359412 (Association for Computing Machinery and Morgan & Claypool, 2018). http://doi.org/10.1145/3191315.3191323

- **RNA**

1. Arslan, A. N. et al. Efficient RNA structure comparison algorithms. *JBCB* 15, (2017).

- **Storage**

1. Roguski, L., Ochoa, I., Hernaez, M., & Deorowicz, S. (2018). FaStore - a space-saving solution for raw sequencing data. *Bioinformatics*, **34**(16), 27482756. http://doi.org/10.1093/bioinformatics/bty205

- **Parallel, quantum or hardware implementations**

1. Rahn, R. et al. Generic accelerated sequence alignment in SeqAn using vectorization and multi-threading. *Bioinformatics* **34**, 34373445 (2018).
2. Cinti, A., Bianchi, F. M., 1712.03560, A. R. A. P. A., 2017. (n.d.). A novel algorithm for online inexact string matching and its FPGA implementation. *Arxiv.org*. .
3. Y. Chen, B. Schmidt, and D. L. Maskell, A hybrid short read mapping accelerator., *BMC bioinformatics* (2013), vol. 14, no. 1, p. 67.
4. J. Blom, T. Jakobi, D. Doppmeier, S. Jaenicke, J. Kalinowski, J. Stoye, and A. Goesmann, Exact and complete short-read alignment to microbial genomes using Graphics Processing Unit programming., Bioinformatics (2011), vol. 27, no. 10, pp. 13511358.

5. E. Fernandez, W. Najjar, and S. Lonardi, String matching in hardware using the FM-Index, *IEEE International Symposium on Field-Programmable Custom Computing Machines,* FCCM 2011 (2011), pp. 218225.

6. E. Fernandez, W. Najjar, and S. Lonardi, String Matching in Hardware Using the FM-Index, presented at the *2011 IEEE 19th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)* , pp. 218225.

7. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 12971303 (2010).

8. Trapnell and Schatz. Optimizing data intensive GPGPU computations for DNA sequence alignment. *Parallel Computing* (2009) vol. 35 (8-9) pp. 429-440

9. Saebo PE, Andersen SM, Myrseth J, Laerdahl JK, Rognes T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W535-9.

10. Rognes T, Seeberg E. (2000) Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. Bioinformatics. 2000 Aug;16(8):699-706.

11. Rognes T. (2001) ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches. Nucleic Acids Res. 2001 Apr 1;29(7):1647-52.

12. Qi Y, Lin F. (2005) Parallelisation of the blast algorithm. Cell Mol Biol Lett. 2005;10(2):281-5.

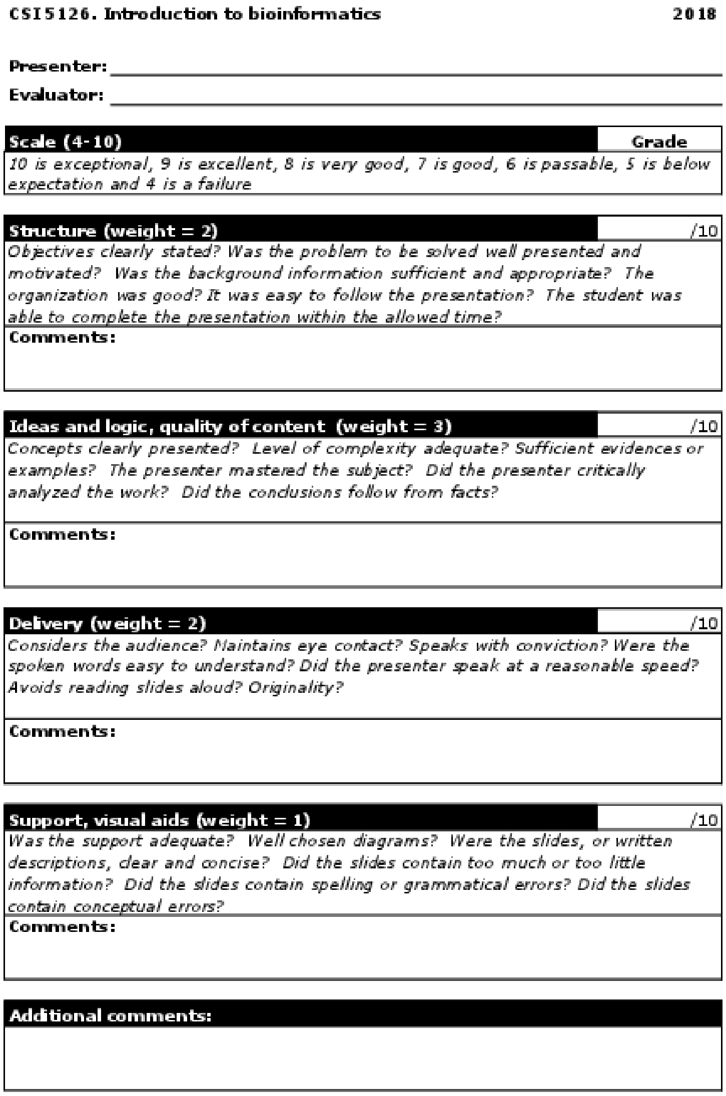
13. P. E. Saebø, S. M. Andersen, J. Myrseth, J. K. Laerdahl, and T. Rognes. (2005) Paralign: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res*, 33(Web Server issue):W535–9.

14. Hollenberg LC. (2000) Fast quantum search algorithms in protein sequence comparisons: quantum bioinformatics. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 2000 Nov;62(5 Pt B):7532-5.

15. R. P. Jacobi, M. Ayala-Rincón, L. G. A. Carvalho, C. H. Llanos, and R. W. Hartenstein. (2005) Reconfigurable systems for sequence alignment and for general dynamic programming. *Genet Mol Res*, 4(3):543–52.

A Evaluation criteria

The evaluation is as follows: my evaluation of your presentation (80%), evaluation the presentation by your peers (10%), summary of the publication (10 %).



CSI5126. Introduction to bioinformatics **2018**

Presenter: ______________________

Evaluator: ______________________

Scale (4-10)	**Grade**
10 is exceptional, 9 is excellent, 8 is very good, 7 is good, 6 is passable, 5 is below expectation and 4 is a failure	
Structure (weight = 2)	/10
Objectives clearly stated? Was the problem to be solved well presented and motivated? Was the background information sufficient and appropriate? The organization was good? It was easy to follow the presentation? The student was able to complete the presentation within the allowed time?	
Comments:	
Ideas and logic, quality of content (weight = 3)	/10
Concepts clearly presented? Level of complexity adequate? Sufficient evidences or examples? The presenter mastered the subject? Did the presenter critically analyzed the work? Did the conclusions follow from facts?	
Comments:	
Delivery (weight = 2)	/10
Considers the audience? Maintains eye contact? Speaks with conviction? Were the spoken words easy to understand? Did the presenter speak at a reasonable speed? Avoids reading slides aloud? Originality?	
Comments:	
Support, visual aids (weight = 1)	/10
Was the support adequate? Well chosen diagrams? Were the slides, or written descriptions, clear and concise? Did the slides contain too much or too little information? Did the slides contain spelling or grammatical errors? Did the slides contain conceptual errors?	
Comments:	
Additional comments:	

B 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 inter-disciplinary 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

C Resources

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

D Frequently Asked Questions (FAQ)

1. None, yet.