References

Abouelhoda, M. I. (2007), A chaining algorithm for mapping cDNA sequences to multiple genomic sequences, in *14th International Symposium on String Processing and Information Retrieval*, Vol. 4726 of Lecture Notes in Computer Science. Berlin: Springer, pp. 1–13.

Adelson-Velskii, G. & Landis, E. M. (1962), 'An algorithm for the organization of information', *Proceedings of the USSR Academy of Sciences* **146**, 263—266 [in Russian]. English translation by Myron J. Ricci in *Soviet Mathematics Doklady*, **3**, 1259–1263 (1962).

Aho, A. V. & Corasick, M. J. (1975), 'Efficient string matching: An aid to bibliographic search', *Communications of the ACM* **18**(6), 333–340.

Ahuja, R., Goldberg, A., Orlin, J. & Tarjan, R. (1992), 'Finding minimum-cost flows by double scaling', *Mathematical Programming* **53**, 243–266.

Ahuja, R. K., Magnanti, T. L. & Orlin, J. B. (1993), *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Alanko, J., Belazzougui, D., Cunial, F. & Mäkinen, V. (2015), Scalable clustering of metagenomic reads. To be published.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Loman, N. J., Andersson, A. F. & Quince, C. (2013), 'CONCOCT: Clustering cONtigs on COverage and ComposiTion', arXiv:1312.4038.

Alstrup, S., Husfeldt, T. & Rauhe, T. (1998), Marked ancestor problems, in *Proceedings of the 39th Annual Symposium on Foundations of Computer Science*, IEEE, pp. 534–543.

Apostolico, A. (2010), Maximal words in sequence comparisons based on subword composition, in *Algorithms and Applications*. Berlin: Springer, pp. 34–44.

Apostolico, A. & Bejerano, G. (2000), 'Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space', *Journal of Computational Biology* **7**(3–4), 381–393.

Apostolico, A. & Denas, O. (2008), 'Fast algorithms for computing sequence distances by exhaustive substring composition', *Algorithms for Molecular Biology* **3**, 13.

Apostolico, A. & Lonardi, S. (2002), 'A speed-up for the commute between subword trees and DAWGs', *Information Processing Letters* **83**(3), 159–161.

Arge, L., Fischer, J., Sanders, P. & Sitchinava, N. (2013), On (dynamic) range minimum queries in external memory, in *Algorithms and Data Structures*, Vol. 8037 of Lecture Notes in Computer Science. Berlin: Springer, pp. 37–48.

Arlazarov, V., Dinic, E., Kronrod, M. & Faradzev, I. (1970), 'On economic construction of the transitive closure of a directed graph', *Doklady Akademii Nauk SSSR* **194**(11), 487–488 [in Russian]. English translation in *Soviet Mathematics Doklady* **11**, 1209–1210 (1975).

Babenko, M., Gawrychowski, P., Kociumaka, T. & Starikovskaya, T. (2015), Wavelet trees meet suffix trees, in *Symposium on Discrete Algorithms, SODA 2015*, pp. 572–591.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Baeza-Yates, R. A. & Ribeiro-Neto, B. A. (2011), *Modern Information Retrieval – The Concepts and Technology behind Search*, 2nd edn. New York, NY: Addison-Wesley.

Baker, B. S. (1993), On finding duplication in strings and software, Technical report, AT&T Bell Laboratories, New Jersey.

Bang-Jensen, J. & Gutin, G. (2008), *Digraphs: Theory, Algorithms and Applications*, 2nd edn. Springer Monographs in Mathematics. Berlin: Springer.

Bao, E., Jiang, T. & Girke, T. (2013), 'BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences', *Bioinformatics* **29**(10), 1250–1259.

Baran, Y. & Halperin, E. (2012), 'Joint analysis of multiple metagenomic samples', *PLoS Computational Biology* **8**(2), e1002373.

Baum, L. E. (1972), An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, in *Inequalities III: Proceedings of the Third Symposium on Inequalities*. New York: Academic Press, pp. 1–8.

Bayer, R. (1972), 'Symmetric binary B-trees: Data structure and maintenance algorithms', *Acta Informatica* **1**, 290–306.

Beerenwinkel, N., Beretta, S., Bonizzoni, P., Dondi, R. & Pirola, Y. (2014), Covering pairs in directed acyclic graphs, in *Language and Automata Theory and Applications*, Vol. 8370 of Lecture Notes in Computer Science. Berlin: Springer, pp. 126–137.

Behnam, E., Waterman, M. S. & Smith, A. D. (2013), 'A geometric interpretation for local alignment-free sequence comparison', *Journal of Computational Biology* **20**(7), 471–485.

Belazzougui, D. (2014), Linear time construction of compressed text indices in compact space, in *Symposium on Theory of Computing, STOC 2014*, pp. 148–193.

Belazzougui, D., Cunial, F., Kärkkäinen, J. & Mäkinen, V. (2013), Versatile succinct representations of the bidirectional Burrows-Wheeler transform, in *21st Annual European Symposium on Algorithms (ESA 2013)*, Vol. 8125 of Lecture Notes in Computer Science. Berlin: Springer, pp. 133–144.

Belazzougui, D. & Puglisi, S. (2015), Range predecessor and Lempel–Ziv parsing. To be published.

Beller, T., Berger, K. & Ohlebusch, E. (2012), Space-efficient computation of maximal and supermaximal repeats in genome sequences, in *19th International Symposium on String Processing and Information Retrieval (SPIRE 2012)*, Vol. 7608 of Lecture Notes in Computer Science. Berlin: Springer, pp. 99–110.

Beller, T., Gog, S., Ohlebusch, E. & Schnattinger, T. (2013), 'Computing the longest common prefix array based on the Burrows–Wheeler transform', *Journal of Discrete Algorithms* **18**, 22–31.

Bellman, R. (1958), 'On a routing problem', *Quarterly of Applied Mathematics* **16**, 87–90.

Bender, M. A. & Farach-Colton, M. (2000), The LCA problem revisited, in *4th Latin American Symposium on Theoretical Informatics (LATIN 2000)*, Vol. 1776 of Lecture Notes in Computer Science, Berlin: Springer, pp. 88–94.

Bernard, E., Jacob, L., Mairal, J. & Vert, J.-P. (2014), 'Efficient RNA isoform identification and quantification from RNA-seq data with network flows', *Bioinformatics* **30**(17), 2447–2455.

Blazewicz, J. & Kasprzak, M. (2003), 'Complexity of DNA sequencing by hybridization', *Theoretical Computer Science* **290**(3), 1459–1473.

Blumer, A., Blumer, J., Haussler, D., Ehrenfeucht, A., Chen, M.-T. & Seiferas, J. (1985), 'The smallest automation recognizing the subwords of a text', *Theoretical Computer Science* **40**, 31–55.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Bowe, A., Onodera, T., Sadakane, K. & Shibuya, T. (2012), Succinct de Bruijn graphs, in *Algorithms in Bioinformatics*, Vol. 7534 of Lecture Notes in Computer Science. Berlin: Springer, pp. 225–235.

Brodal, G. S., Davoodi, P. & Rao, S. S. (2011), Path minima queries in dynamic weighted trees, in *Algorithms and Data Structures*, Vol. 6844 of Lecture Notes in Computer Science. Berlin: Springer, pp. 290–301.

Burrows, M. & Wheeler, D. (1994), A block sorting lossless data compression algorithm, Technical report 124, Digital Equipment Corporation.

Cancedda, N., Gaussier, E., Goutte, C. & Renders, J. M. (2003), 'Word sequence kernels', *The Journal of Machine Learning Research* **3**, 1059–1082.

Chairungsee, S. & Crochemore, M. (2012), 'Using minimal absent words to build phylogeny', *Theoretical Computer Science* **450**, 109–116.

Chaisson, M., Pevzner, P. A. & Tang, H. (2004), 'Fragment assembly with short reads', *Bioinformatics* **20**(13), 2067–2074.

Chan, H.-L., Lam, T.-W., Sung, W.-K., Tam, S.-L. & Wong, S.-S. (2006), A linear size index for approximate pattern matching, in *Proceedings of the Annual Symposium on Combinatorial Pattern Matching*, Vol. 4009 of Lecture Notes in Computer Science. Berlin: Springer, pp. 49–59.

Chan, R. H., Chan, T. H., Yeung, H. M. & Wang, R. W. (2012), 'Composition vector method based on maximum entropy principle for sequence comparison', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**(1), 79–87.

Charikar, M., Lehman, E., Liu, D., Panigrahy, R., Prabhakaran, M., Rasala, A., Sahai, A. & Shelat, A. (2002), Approximating the smallest grammar: Kolmogorov complexity in natural models, in *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, ACM, pp. 792–801.

Chazelle, B. (1988), 'A functional approach to data structures and its use in multidimensional searching', *SIAM Journal on Computing* **17**(3), 427–462.

Chikhi, R., Limasset, A., Jackman, S., Simpson, J. & Medvedev, P. (2014), 'On the representation of de Bruijn graphs', in *Proceeedings of the Annual International Conference on Research in Computational Molecular Biology*, Vol. 8394 of Lecture Notes in Computer Science. Berlin: Springer, pp. 35–55.

Chikhi, R. & Rizk, G. (2012), Space-efficient and exact de Bruijn graph representation based on a Bloom filter, in B. J. Raphael & J. Tang, eds., *Algorithms in Bioinformatics*, Vol. 7534 of Lecture Notes in Computer Science. Berlin: Springer, pp. 236–248.

Cilibrasi, R., Iersel, L., Kelk, S. & Tromp, J. (2005), On the complexity of several haplotyping problems, in R. Casadio & G. Myers, eds., *Algorithms in Bioinformatics*, Vol. 3692 of Lecture Notes in Computer Science. Berlin: Springer, pp. 128–139.

Cilibrasi, R. & Vitányi, P. M. (2005), 'Clustering by compression', *IEEE Transactions on Information Theory* **51**(4), 1523–1545.

Clark, D. (1996), Compact Pat Trees, PhD thesis, University of Waterloo, Canada.

Cole, R., Gottlieb, L. A. & Lewenstein, M. (2004), Dictionary matching and indexing with errors, in *Proceedings of the Symposium on Theory of Computing*, pp. 91–100.

Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E. & Mueller, S. (2008), 'Virus attenuation by genome-scale changes in codon pair bias', *Science* **320**(5884), 1784–1787.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2009), *Introduction to Algorithms*, 3rd edn. Cambridge, MA: MIT Press.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Cristianini, N. & Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*. New York, NY: Cambridge University Press.

Crochemore, M., Landau, G. & Ziv-Ukelson, M. (2002), A sub-quadratic sequence alignment algorithm for unrestricted cost matrices, in *Proceedings of the 13th ACM–SIAM Symposium on Discrete Algorithms (SODA 2002)*, pp. 679–688.

Crochemore, M., Mignosi, F. & Restivo, A. (1998), 'Automata and forbidden words', *Information Processing Letters* **67**(3), 111–117.

Crochemore, M., Mignosi, F., Restivo, A. & Salemi, S. (2000), 'Data compression using antidictionaries', *Proceedings of the IEEE* **88**(11), 1756–1768.

Crochemore, M. & Rytter, W. (2002), *Jewels of Stringology*. Singapore: World Scientific.

Crochemore, M. & Vérin, R. (1997a), Direct construction of compact directed acyclic word graphs, in *Proceeding of the 8th Annual Symposium on Combinatorial Pattern Matching (CPM)*, Vol. 1264 of Lecture Notes in Computer Science. Berlin: Springer, pp. 116–129.

Crochemore, M. & Vérin, R. (1997b), On compact directed acyclic word graphs, in *Structures in Logic and Computer Science*. Berlin: Springer, pp. 192–211.

Davoodi, P. (2011), Data Structures: Range Queries and Space Efficiency, PhD thesis, Department of Computer Science, Aarhus University.

Dayarian, A., Michael, T. P. & Sengupta, A. M. (2010), 'SOPRA: Scaffolding algorithm for paired reads via statistical optimization', *BMC Bioinformatics* **11**, 345.

de Berg, M., van Kreveld, M., Overmars, M. & Schwarzkopf, O. (2000), *Computational Geometry – Algorithms and Applications*, Vol. 382 of Lecture Notes in Computer Science. Berlin: Springer.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O. & Salzberg, S. L. (1999), 'Alignment of whole genomes', *Nucleic Acids Research* **27**(11), 2369–2376.

Deorowicz, S. & Grabowski, S. (2011), 'Robust relative compression of genomes with random access', *Bioinformatics* **27**(21), 2979–2986.

Dietz, P. F. (1989), Optimal algorithms for list indexing and subset rank, in *Algorithms and Data Structures (WADS' 89)*, Vol. 382 of Lecture Notes in Computer Science. Berlin: Springer, pp. 39–46.

Dilworth, R. P. (1950), 'A decomposition theorem for partially ordered sets', *The Annals of Mathematics* **51**(1), 161–166.

Do, H. H., Jansson, J., Sadakane, K. & Sung, W.-K. (2012), Fast relative Lempel–Ziv self-index for similar sequences, in *Joint International Conference on Frontiers in Algorithmics and Algorithmic Aspects in Information and Management (FAW-AAIM)*, Vol. 7285 of Lecture Notes in Computer Science. Berlin: Springer, pp. 291–302.

Donmez, N. & Brudno, M. (2013), 'SCARPA: Scaffolding reads with practical algorithms', *Bioinformatics* **29**(4), 428–434.

Drineas, P., Mahoney, M. W., Muthukrishnan, S. & Sarlós, T. (2011), 'Faster least squares approximation', *Numerische Mathematik* **117**(2), 219–249.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Eddy, S. R. (2011), 'Accelerated profile HMM searches', *PLoS Computational Biology* **7**(10), e1002195.

Edmonds, J. & Karp, R. (1972), 'Theoretical improvements in algorithmic efficiency for network flow problems', *Journal of the Association for Computing Machinery* **19**, 248–264.

Edwards, R. A., Olson, R., Disz, T., Pusch, G. D., Vonstein, V., Stevens, R. & Overbeek, R. (2012), 'Real time metagenomics: Using *k*-mers to annotate metagenomes', *Bioinformatics* **28**(24), 3316–3317.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Elias, P. (1975), 'Universal codeword sets and representations of the integers', *IEEE Transactions on Information Theory* **21**(2), 194–203.

Eppstein, D., Galil, Z., Giancarlo, R. & Italiano, G. F. (1992a), 'Sparse dynamic programming I: Linear cost functions', *Journal of the ACM* **39**(3), 519–545.

Eppstein, D., Galil, Z., Giancarlo, R. & Italiano, G. F. (1992b), 'Sparse dynamic programming II: Convex and concave cost functions', *Journal of the ACM* **39**(3), 546–567.

Farach, M. (1997), Optimal suffix tree construction with large alphabets, in *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 137–143.

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A. & Ziv, J. (1995), On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence, in *Proceedings of the Sixth Annual ACM–SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, pp. 48–57.

Feng, J., Li, W. & Jiang, T. (2011), 'Inference of isoforms from short sequence reads', *Journal of Computational Biology* **18**(3), 305–321.

Ferrada, H., Gagie, T., Hirvola, T. & Puglisi, S. J. (2014), 'Hybrid indexes for repetitive datasets', *Philosophical Transactions of the Royal Society A* **372**(2016), 20130137.

Ferragina, P., Luccio, F., Manzini, G. & Muthukrishnan, S. (2009), 'Compressing and indexing labeled trees, with applications', *Journal of the ACM* **57**(1), Article 4.

Ferragina, P. & Manzini, G. (2000), Opportunistic data structures with applications, in *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, IEEE, pp. 390–398.

Ferragina, P. & Manzini, G. (2005), 'Indexing compressed texts', *Journal of the ACM* **52**(4), 552–581.

Ferragina, P., Nitto, I. & Venturini, R. (2009), On the bit-complexity of Lempel-Ziv compression, in *Proceedings of the Twentieth Annual ACM–SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, pp. 768–777.

Ferragina, P., Nitto, I. & Venturini, R. (2013), 'On the bit-complexity of Lempel–Ziv compression', *SIAM Journal on Computing* **42**, 1521–1541.

Fischer, J. & Heun, V. (2011), 'Space-efficient preprocessing schemes for range minimum queries on static arrays', *SIAM Journal on Computing* **40**(2), 465–492.

Fischer, J., Mäkinen, V. & Välimäki, N. (2008), Space-efficient string mining under frequency constraints, in *Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM 2008)*, IEEE Computer Society, pp. 193–202.

Ford, L. R. (1956), Network flow theory, Technical Report Paper P-923, The RAND Corporation, Santa Monica, CA.

Ford, L. R. & Fulkerson, D. R. (1956), 'Maximal flow through a network', *Canadian Journal of Mathematics* **8**, 399–404.

Fredman, M. L. (1975), 'On computing the length of longest increasing subsequences', *Discrete Mathematics* **11**(1), 29–35.

Fredman, M. L. & Willard, D. E. (1994), 'Trans-dichotomous algorithms for minimum spanning trees and shortest paths', *Journal of Computer and System Sciences* **48**(3), 533–551.

Fulkerson, D. R. (1956), 'Note on Dilworth's decomposition theorem for partially ordered sets', *Proceedings of the American Mathematical Society* **7**(4), 701–702.

Gabow, H. N. (1990), Data structures for weighted matching and nearest common ancestors with linking, in *Proceedings of the First Annual ACM–SIAM Symposium on Discrete Algorithms*, *SODA '90*, Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 434–443.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Gabow, H. N., Bentley, J. L. & Tarjan, R. E. (1984), Scaling and related techniques for geometry problems, in *Proceedings of the 16th Annual ACM Symposium on Theory of Computing (STOC 1984)*, ACM, pp. 135–143.

Gabow, H. N. & Tarjan, R. E. (1989), 'Faster scaling algorithms for network problems', *SIAM Journal on Computing* **18**(5), 1013–1036.

Gagie, T., Gawrychowski, P., Kärkkäinen, J., Nekrich, Y. & Puglisi, S. J. (2012), A faster grammar-based self-index, in *6th International Conference on Language and Automata Theory and Applications (LATA 2012)*, Vol. 7183 of Lecture Notes in Computer Science. Berlin: Springer, pp. 240–251.

Gallé, M. (2011), Searching for Compact Hierarchical Structures in DNA by Means of the Smallest Grammar Problem, PhD thesis, Université Rennes 1.

Gao, S., Sung, W.-K. & Nagarajan, N. (2011), 'Opera: Reconstructing optimal genomic scaffolds with high-throughput paired-end sequences', *Journal of Computational Biology* **18**(11), 1681–1691.

Garcia, S. P. & Pinho, A. J. (2011), 'Minimal absent words in four human genome assemblies', *PLoS One* **6**(12), e29344.

Garcia, S. P., Pinho, A. J., Rodrigues, J. M., Bastos, C. A. & Ferreira, P. J. (2011), 'Minimal absent words in prokaryotic and eukaryotic genomes', *PLoS One* **6**(1), e16065.

Garey, M. R. & Johnson, D. S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY: W. H. Freeman & Co.

Giancarlo, R., Rombo, S. E. & Utro, F. (2014), 'Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies', *Briefings in Bioinformatics* **15**(3), 390–406.

Göke, J., Schulz, M. H., Lasserre, J. & Vingron, M. (2012), 'Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts', *Bioinformatics* **28**(5), 656–663.

Goldberg, A. & Tarjan, R. (1990), 'Finding minimum-cost circulations by successive approximation', *Mathematics of Operations Research* **15**, 430–466.

Goldberg, A. V. & Tarjan, R. E. (1987), Solving minimum-cost flow problems by successive approximation, in A. V. Aho, ed., *STOC '87 Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, ACM, pp. 7–18.

Goldberg, A. V. & Tarjan, R. E. (1988), Finding minimum-cost circulations by canceling negative cycles, in J. Simon, ed., *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, ACM, pp. 388–397.

Goldberg, A. V. & Tarjan, R. E. (1989), 'Finding minimum-cost circulations by canceling negative cycles', *Journal of the ACM* **36**(4), 873–886.

Gonnet, G., Baeza-Yates, R. & Snider, T. (1992), *Information Retrieval: Data Structures and Algorithms*. Upper Saddle River, NJ: Prentice-Hall, Chapter 3: New indices for text: Pat trees and Pat arrays, pp. 66–82.

Gori, F., Folino, G., Jetten, M. S. & Marchiori, E. (2011), 'MTR: Taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks', *Bioinformatics* **27**(2), 196–203.

Gotoh, O. (1982), 'An improved algorithm for matching biological sequences', *Journal of Molecular Biology* **162**(3), 705–708.

Grossi, R., Gupta, A. & Vitter, J. (2003), High-order entropy-compressed text indexes, in *Proceedings of the 14th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA)*, pp. 841–850.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Grossi, R. & Vitter, J. (2000), Compressed suffix arrays and suffix trees with applications to text indexing and string matching, in *Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC)*, pp. 397–406.

Grossi, R. & Vitter, J. (2006), 'Compressed suffix arrays and suffix trees with applications to text indexing and string matching', *SIAM Journal on Computing* **35**(2), 378–407.

Guisewite, G. & Pardalos, P. (1990), 'Minimum concave-cost network flow problems: Applications, complexity, and algorithms', *Annals of Operations Research* **25**(1), 75–99.

Gusfield, D. (1997), *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

Gusfield, D., Landau, G. M. & Schieber, B. (1992), 'An efficient algorithm for the all pairs suffix-prefix problem', *Information Processing Letters* **41**(4), 181–185.

Hampikian, G. & Andersen, T. (2007), Absent sequences: nullomers and primes, in *Pacific Symposium on Biocomputing*, Vol. 12, pp. 355–366.

Haque, M. M., Ghosh, T. S., Komanduri, D. & Mande, S. S. (2009), 'SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences', *Bioinformatics* **25**(14), 1722–1730.

Harel, D. & Tarjan, R. E. (1984), 'Fast algorithms for finding nearest common ancestors', *SIAM Journal on Computing* **13**(2), 338–355.

Hartman, T., Hassidim, A., Kaplan, H., Raz, D. & Segalov, M. (2012), How to split a flow?, in A. G. Greenberg & K. Sohraby, eds., *Proceedings of INFOCOM 2012*, IEEE, pp. 828–836.

Haubold, B., Pierstorff, N., Möller, F. & Wiehe, T. (2005), 'Genome comparison without alignment using shortest unique substrings', *BMC Bioinformatics* **6**(1), 123.

Haussler, D. (1999), Convolution kernels on discrete structures, Technical report UCSC-CRL-99-10, UC Santa Cruz.

Herold, J., Kurtz, S. & Giegerich, R. (2008), 'Efficient computation of absent words in genomic sequences', *BMC Bioinformatics* **9**(1), 167.

Hirschberg, D. S. (1975), 'A linear space algorithm for computing maximal common subsequences', *Communications of the ACM* **18**(6), 341–343.

Hoare, C. A. R. (1962), 'Quicksort', *Computer Journal* **5**(1), 10–15.

Hon, W.-K. & Sadakane, K. (2002), Space-economical algorithms for finding maximal unique matches, in *Proceedings of the 13th Annual Symposium on Combinatorial Pattern Matching (CPM '02)*, Vol. 2373 of Lecture Notes in Computer Science. Berlin: Springer, pp. 144–152.

Hon, W.-K., Sadakane, K. & Sung, W.-K. (2003), Breaking a time-and-space barrier in constructing full-text indices, in *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science (FOCS)*, IEEE, pp. 251–260.

Hoobin, C., Puglisi, S. J. & Zobel, J. (2011), 'Relative Lempel–Ziv factorization for efficient storage and retrieval of web collections', *Proceedings of the VLDB Endowment* **5**(3), 265–273.

Hopcroft, J. E. & Karp, R. M. (1973), 'An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs.', *SIAM Journal on Computing* **2**(4), 225–231.

Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. (2009), 'Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes', *Genome Research* **19**(7), 1270–1278.

Hu, T. C. (1966), 'Minimum-cost flows in convex-cost networks', *Naval Research Logistics Quarterly* **13**(1), 1–9.

Huang, K., Brady, A., Mahurkar, A., White, O., Gevers, D., Huttenhower, C. & Segata, N. (2013), 'MetaRef: A pan-genomic database for comparative and community microbial genomics', *Nucleic Acids Research* 42(D1), D617–D624.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Huang, L., Popic, V. & Batzoglou, S. (2013), 'Short read alignment with populations of genomes', *Bioinformatics* **29**(13), 361–370.

Hui, L. C. K. (1992), Color set size problem with application to string matching, in *Proceedings of the Annual Symposium on Combinatorial Pattern Matching (CPM)*, Vol. 644 of Lecture Notes in Computer Science. Berlin: Springer, pp. 230–243.

Hunt, J. W. & Szymanski, T. G. (1977), 'A fast algorithm for computing longest common subsequences', *Communications of the ACM* **20**(5), 350–353.

Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. (2011), 'Integrative analysis of environmental sequences using MEGAN4', *Genome Research* **21**(9), 1552–1560.

Huson, D. H., Reinert, K. & Myers, E. W. (2001), The greedy path-merging algorithm for sequence assembly, in *Proceedings of RECOMB 2001*, pp. 157–163.

Hyyrö, H. (2001), Explaining and extending the bit-parallel approximate string matching algorithm of Myers, Technical report A2001-10, Department of Computer and Information Sciences, University of Tampere, Finland.

Jacobson, G. (1989), Space-efficient static trees and graphs, in *Proceedings of the 30th IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 549–554.

Jenkins, P., Lyngsø, R. B. & Hein, J. (2006), How many transcripts does it take to reconstruct the splice graph?, in *Proceedings of Algorithms in Bioinformatics, 6th International Workshop, WABI 2006*, Vol. 4175 of Lecture Notes in Computer Science. pp. 103–114.

Jiang, B., Song, K., Ren, J., Deng, M., Sun, F. & Zhang, X. (2012), 'Comparison of metagenomic samples using sequence signatures', *BMC Genomics* **13**, 730.

Kärkkäinen, J. & Na, J. C. (2007), Faster filters for approximate string matching, in *Proceedings of the 9th Workshop on Algorithm Engineering and Experiments (ALENEX07)*, SIAM, pp. 84–90.

Kärkkäinen, J., Sanders, P. & Burkhardt, S. (2006), 'Linear work suffix array construction', *Journal of the ACM* **53**(6), 918–936.

Kasai, T., Lee, G., Arimura, H., Arikawa, S. & Park, K. (2001), Linear-time longest-common-prefix computation in suffix arrays and its applications, in *Proceedings of the Annual Symposium on Combinatorial Pattern Matching (CPM)*, Vol. 2089 of Lecture Notes in Computer Science. Berlin: Springer, pp. 181–192.

Kececioglu, J. (1991), Exact and Approximation Algorithms for DNA Sequence Reconstruction, PhD thesis, The University of Arizona.

Kececioglu, J. D. & Myers, E. W. (1993), 'Combinatorial algorithms for DNA sequence assembly', *Algorithmica* **13**(1–2), 7–51.

Kim, D., Sim, J., Park, H. & Park, K. (2005), 'Constructing suffix arrays in linear time', *Journal of Discrete Algorithms* **3**(2–4), 126–142.

Knuth, D. E., Morris, J. & Pratt, V. R. (1977), 'Fast pattern matching in strings', *SIAM Journal of Computing* **6**(2), 323–350.

Ko, P. & Aluru, S. (2005), 'Space efficient linear time construction of suffix arrays', *Journal of Discrete Algorithms* **3**(2–4), 143–156.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994), 'Hidden Markov models in computational biology: Applications to protein modeling', *Journal of Molecular Biology* **235**(5), 1501–1531.

Kuksa, P. P., Huang, P.-h. & Pavlovic, V. (2008), V.: Scalable algorithms for string kernels with inexact matching, in *Proceedings of the Neural Information Processing Systems 21 (NIPS 2008)*, pp. 881–888.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Kuksa, P. P. & Pavlovic, V. (2012), Efficient evaluation of large sequence kernels, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 759–767.

Kulekci, M. O., Vitter, J. S. & Xu, B. (2012), 'Efficient maximal repeat finding using the Burrows–Wheeler transform and wavelet tree', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**(2), 421–429.

Kuruppu, S., Puglisi, S. J. & Zobel, J. (2011a), Optimized relative Lempel–Ziv compression of genomes, in *Proceedings of the Thirty-Fourth Australasian Computer Science Conference*, Australian Computer Society, Inc., pp. 91–98.

Kuruppu, S., Puglisi, S. J. & Zobel, J. (2011b), Reference sequence construction for relative compression of genomes, in *String Processing and Information Retrieval*, Vol. 7024 of Lecture Notes in Computer Science. Berlin: Springer, pp. 420–425.

Lam, T. W., Sung, W. K., Tam, S. L., Wong, C. K. & Yiu, S. M. (2008), 'Compressed indexing and local alignment of DNA', *Bioinformatics* **24**(6), 791–797.

Landau, G. & Vishkin, U. (1988), 'Fast string matching with *k* differences', *Journal of Computer and System Sciences* **37**, 63–78.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009), 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biology* **10**(3), R25.

Lee, C., Grasso, C. & Sharlow, M. F. (2002), 'Multiple sequence alignment using partial order graphs', *Bioinformatics* **18**(3), 452–464.

Lehman, E. & Shelat, A. (2002), Approximation algorithms for grammar-based compression, in *Proceedings of the Thirteenth Annual ACM–SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, pp. 205–212.

Leslie, C. & Kuang, R. (2003), Fast kernels for inexact string matching, in *Learning Theory and Kernel Machines*, Vol. 2777 of Lecture Notes in Computer Science. Berlin: Springer, pp. 114–128.

Levenshtein, V. (1966), 'Binary codes capable of correcting deletions, insertions and reversals', *Soviet Physics Doklady* **10**, 707.

Li, H. & Durbin, R. (2009), 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics* **25**(14), 1754–1760.

Li, J. J., Jiang, C. R., Brown, J. B., Huang H. & Bickel, P. J. (2011a), 'Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation', *Proceedings of the National Academy of Sciences* **108**(50), 19867–19872.

Li, M., Chen, X., Li, X., Ma, B. & Vitányi, P. M. (2004), 'The similarity metric', *IEEE Transactions on Information Theory* **50**(12), 3250–3264.

Li, M. & Vitányi, P. M. (2008), *An Introduction to Kolmogorov Complexity and its Applications*. Berlin: Springer.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. & Wang, J. (2009), 'SOAP2: An improved ultrafast tool for short read alignment', *Bioinformatics* **25**(15), 1966–1967.

Li, W., Feng, J. & Jiang, T. (2011b), 'IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly', *Journal of Computational Biology* **18**(11), 1693–1707.

Lifshits, Y., Mozes, S., Weimann, O. & Ziv-Ukelson, M. (2009), 'Speeding up HMM decoding and training by exploiting sequence repetitions', *Algorithmica* **54**(3), 379–399.

Lin, Y.-Y., Dao, P., Hach, F., Bakhshi, M., Mo, F., Lapuk, A., Collins, C. & Sahinalp, S. C. (2012), CLIIQ: Accurate comparative detection and quantification of expressed isoforms in a population, in B. J. Raphael & J. Tang, eds., *Algorithms in Bioinformatics*, Vol. 7534 of Lecture Notes in Computer Science. Berlin: Springer, pp. 178–189.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Lindsay, J., Salooti, H., Zelikovsky, A. & Măndoiu, I. (2012), Scalable genome scaffolding using integer linear programming, in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, BCB '12*, ACM, New York, NY, pp. 377–383.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. (2011), 'Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences', *BMC Genomics* **12**(Suppl. 2), S4.

Lo, C., Kim, S., Zakov, S. & Bafna, V. (2013), 'Evaluating genome architecture of a complex region via generalized bipartite matching', *BMC Bioinformatics* **14**(Suppl. 5), S13.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. & Watkins, C. (2002), 'Text classification using string kernels', *The Journal of Machine Learning Research* **2**, 419–444.

Löytynoja, A., Vilella, A. J. & Goldman, N. (2012), 'Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm', *Bioinformatics* **28**(13), 1684–1691.

Maier, D. (1978), 'The complexity of some problems on subsequences and supersequences', *Journal of the ACM* **25**(2), 322–336.

Maillet, N., Lemaitre, C., Chikhi, R., Lavenier, D. & Peterlongo, P. (2012), 'Compareads: Comparing huge metagenomic experiments', *BMC Bioinformatics* **13**(Suppl. 19), S10.

Mäkinen, V. & Navarro, G. (2007), 'Rank and select revisited and extended', *Theoretical Computer Science* **387**(3), 332–347.

Mäkinen, V., Navarro, G., Síren, J. & Välimäki, N. (2009), Storage and retrieval of individual genomes, in *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2009)*, Vol. 5541 of Lecture Notes in Bioinformatics. Berlin: Springer, pp. 121–137.

Mäkinen, V., Navarro, G., Sirén, J. & Välimäki, N. (2010), 'Storage and retrieval of highly repetitive sequence collections', *Journal of Computational Biology* **17**(3), 281–308.

Mäkinen, V., Navarro, G. & Ukkonen, E. (2003), Algorithms for transposition invariant string matching, in *Proceedings of the 20th International Symposium on Theoretical Aspects of Computer Science (STACS '03)*, Vol. 2607 of Lecture Notes in Computer Science. Berlin: Springer, pp. 191–202.

Mäkinen, V., Salmela, L. & Ylinen, J. (2012), 'Normalized N50 assembly metric using gap-restricted co-linear chaining', *BMC Bioinformatics* **13**, 255.

Manber, U. & Myers, G. (1993), 'Suffix arrays: A new method for on-line string searches', *SIAM Journal on Computing* **22**(5), 935–948.

Marschall, T., Costa, I. G., Canzar, S., Bauer, M., Klau, G. W., Schliep, A. & Schönhuth, A. (2012), 'CLEVER: Clique-enumerating variant finder', *Bioinformatics* **28**(22), 2875–2882.

Masek, W. & Paterson, M. (1980), 'A faster algorithm for computing string edit distances', *Journal of Computer and System Sciences* **20**(1), 18–31.

Mathé, C., Sagot, M.-F., Schiex, T. & Rouzé, P. (2002), 'Current methods of gene prediction, their strengths and weaknesses', *Nucleic Acids Research* **30**(19), 4103–4117.

McCreight, E. (1976), 'A space-economical suffix tree construction algorithm', *Journal of the ACM* **23**(2), 262–272.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. (2010), 'Detecting copy number variation with mated short reads', *Genome Research* **20**(11), 1613–1622.

Medvedev, P., Georgiou, K., Myers, G. & Brudno, M. (2007), Computability of models for sequence assembly, in *Workshop on Algorithms in Bioinformatics (WABI 2007)*, Vol. 4645 of Lecture Notes in Computer Science. Berlin: Springer, pp. 289–301.

Minoux, M. (1986), Solving integer minimum cost flows with separable convex cost objective polynomially, in G. Gallo & C. Sandi, eds, *Netflow at Pisa*, Vol. 26 of Mathematical Programming Studies. Berlin: Springer, pp. 237–239.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Moore, E. (1959), The shortest path through a maze, in H. Aiken, ed., *Proceedings of an International Symposium on the Theory of Switching, 2–5 April 1957, Part II*. Cambridge, MA: Harvard University Press, pp. 285–292.

Morris, J. & Pratt, V. R. (1970), A linear pattern-matching algorithm, Technical Report 40, University of California, Berkeley.

Munro, I. (1996), Tables, in *Proceedings of the 16th Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, Vol. 1180 of Lecture Notes in Computer Science. Berlin: Springer, pp. 37–42.

Myers, E. W. (2005), 'The fragment assembly string graph', *Bioinformatics* **21**(Suppl. 2), ii79–ii85.

Myers, G. (1999), 'A fast bit-vector algorithm for approximate string matching based on dynamic programming', *Journal of the ACM* **46**(3), 395–415.

Nagarajan, N. & Pop, M. (2009), 'Parametric complexity of sequence assembly: Theory and applications to next generation sequencing', *Journal of Computational Biology* **16**(7), 897–908.

Navarro, G. (2001), 'A guided tour to approximate string matching', *ACM Computing Surveys* **33**(1), 31–88.

Navarro, G. (2012), Wavelet trees for all, in J. Kärkkäinen & J. Stoye, eds., *Combinatorial Pattern Matching*, Vol. 7354 of Lecture Notes in Computer Science. Berlin: Springer, pp. 2–26.

Navarro, G. & Mäkinen, V. (2007), 'Compressed full-text indexes', *ACM Computing Surveys* **39**(1), Article 2.

Navarro, G. & Raffinot, M. (2002), *Flexible Pattern Matching in Strings – Practical On-Line Search Algorithms for Texts and Biological Sequences*. Cambridge: Cambridge University Press.

Needleman, S. B. & Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of Molecular Biology* **48**(3), 443–453.

Nevill-Manning, C. G. (1996), Inferring Sequential Structure, PhD thesis.

Ntafos, S. & Hakimi, S. (1979), 'On path cover problems in digraphs and applications to program testing', *IEEE Transactions on Software Engineering* **5**(5), 520–529.

Nykänen, M. & Ukkonen, E. (2002), 'The exact path length problem', *Journal of Algorithms* **42**(1), 41–52.

Ohlebusch, E., Gog, S. & Kügel, A. (2010), Computing matching statistics and maximal exact matches on compressed full-text indexes, in *String Processing and Information Retrieval*, Vol. 6393 of Lecture Notes in Computer Science. Berlin: Springer, pp. 347–358.

Onodera, T., Sadakane, K. & Shibuya, T. (2013), Detecting superbubbles in assembly graphs, in *Algorithms in Bioinformatics*, Vol. 8126 of Lecture Notes in Computer Science. Berlin: Springer, pp. 338–348.

Orlin, J. (1993), 'A faster strongly polynomial minimum cost flow algorithm', *Operations Research* **41**, 338–350.

Orlin, J. B. (1988), A faster strongly polynominal minimum cost flow algorithm, in J. Simon, ed., *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, ACM, pp. 377–387.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J. & Trajanoski, Z. (2013), 'A survey of tools for variant analysis of next-generation genome sequencing data', *Briefings in Bioinformatics* **15**(2), 256–278.

Parida, L. (2007), *Pattern Discovery in Bioinformatics: Theory and Algorithms*. New York, NY: Chapman & Hall/CRC.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G. W. & Schönhuth, A. (2014), WhatsHap: Haplotype assembly for future-generation sequencing reads, in *18th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2014)*, Vol. 8394 of Lecture Notes in Computer Science. Berlin: Springer, pp. 237–249.

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M. & Brown, C. T. (2012), 'Scaling metagenome sequence assembly with probabilistic de Bruijn graphs', *Proceedings of the National Academy of Sciences* **109**(33), 13272–13277.

Pellicer, J., Fay, M. F. & Leitch, I. J. (2010), 'The largest eukaryotic genome of them all?', *Botanical Journal of the Linnean Society* **164**(1), 10–15.

Pevzner, P. (1989), 'ℓ-tuple DNA sequencing: Computer analysis', *Journal of Biomolecular Structure and Dynamics* **7**(1), 63–73.

Pevzner, P. A., Tang, H. & Waterman, M. S. (2001), A new approach to fragment assembly in DNA sequencing, in *Proceedings of RECOMB 2001*, pp. 256–267.

Pijls, W. & Potharst, R. (2013), 'Another note on Dilworth's decomposition theorem', *Journal of Discrete Mathematics* **2013**, 692645.

Policriti, A. & Prezza, N. (2014a), 'Fast randomized approximate string matching with succinct hash data structures', *BMC Bioinformatics*, in press.

Policriti, A. & Prezza, N. (2014b), Hashing and indexing: Succinct data structures and smoothed analysis, in *ISAAC 2014 – 25th International Symposium on Algorithms and Computation*, Vol. 8889 of Lecture Notes in Computer Science. Berlin, Springer, pp. 157–168.

Policriti, A., Tomescu, A. I. & Vezzi, F. (2012), 'A randomized numerical aligner (rNA)', *Journal of Computer and System Sciences* **78**(6), 1868–1882.

Puglisi, S. J., Smyth, W. F. & Turpin, A. (2007), 'A taxonomy of suffix array construction algorithms', *ACM Computing Surveys* **39**(2).

Qi, J., Wang, B. & Hao, B.-I. (2004), 'Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach', *Journal of Molecular Evolution* **58**(1), 1–11.

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., Yu, J. & Chinnaiyan, A. M. (2010), 'HPeak: An HMM-based algorithm for defining read-enriched regions in ChIP-Seq data', *BMC Bioinformatics* **11**, 369.

Rabin, M. O. & Scott, D. (1959), 'Finite automata and their decision problems', *IBM Journal of Research and Development* **3**(2), 114–125.

Raffinot, M. (2001), 'On maximal repeats in strings', *Information Processing Letters* **80**(3), 165–169.

Räihä, K.-J. & Ukkonen, E. (1981), 'The shortest common supersequence problem over binary alphabet is NP-complete', *Theoretical Computer Science* **16**, 187–198.

Reinert, G., Chew, D., Sun, F. & Waterman, M. S. (2009), 'Alignment-free sequence comparison (I): Statistics and power', *Journal of Computational Biology* **16**(12), 1615–1634.

Rizzi, R., Tomescu, A. I. & Mäkinen, V. (2014), 'On the complexity of minimum path cover with subpath constraints for multi-assembly', *BMC Bioinformatics* **15**(Suppl. 9), S5.

Rødland, E. A. (2013), 'Compact representation of k-mer de Bruijn graphs for genome read assembly', *BMC Bioinformatics* **14**(1), 313.

Rousu, J., Shawe-Taylor, J. & Jaakkola, T. (2005), 'Efficient computation of gapped substring kernels on large alphabets', *Journal of Machine Learning Research* **6**(9), 1323–1344.

Rytter, W. (2003), 'Application of Lempel–Ziv factorization to the approximation of grammar-based compression', *Theoretical Computer Science* **302**(1), 211–222.

Sacomoto, G. (2014), Efficient Algorithms for de novo Assembly of Alternative Splicing Events from RNA-seq Data, PhD thesis, Université Claude Bernard Lyon 1, Lyon.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Sadakane, K. (2000), Compressed text databases with efficient query algorithms based on the compressed suffix array, in *Proceedings of the 11th International Symposium on Algorithms and Computation (ISAAC)*, Vol. 1969 of Lecture Notes in Computer Science. Berlin: Springer, pp. 410–421.

Sahlin, K., Street, N., Lundeberg, J. & Arvestad, L. (2012), 'Improved gap size estimation for scaffolding algorithms', *Bioinformatics* **28**(17), 2215–2222.

Sakharkar, M. K., Chow, V. T. & Kangueane, P. (2004), 'Distributions of exons and introns in the human genome', *In Silico Biology* **4**(4), 387–393.

Salikhov, K., Sacomoto, G. & Kucherov, G. (2013), Using cascading Bloom filters to improve the memory usage for de Brujin graphs, in *Algorithms in Bioinformatics*, Vol. 8126 of Lecture Notes in Computer Science. Berlin: Springer, pp. 364–376.

Salmela, L., Mäkinen, V., Välimäki, N., Ylinen, J. & Ukkonen, E. (2011), 'Fast scaffolding with small independent mixed integer programs', *Bioinformatics* **27**(23), 3259–3265.

Salmela, L., Sahlin, K., Mäkinen, V. & Tomescu, A. I. (2015), Gap filling as exact path length problem, in *Proceedings of RECOMB 2015, 19th International Conference on Research in Computational Molecular Biology*. To be published.

Salmela, L. & Schröder, J. (2011), 'Correcting errors in short reads by multiple alignments', *Bioinformatics* **27**(11), 1455–1461.

Sankoff, D. (1975), 'Minimal mutation trees of sequences', *SIAM Journal on Applied Mathematics* **28**, 35–42.

Schnattinger, T., Ohlebusch, E. & Gog, S. (2010), Bidirectional search in a string with wavelet trees, in *21st Annual Symposium on Combinatorial Pattern Matching (CPM 2010)*, Vol. 6129 of Lecture Notes in Computer Science. Berlin: Springer, pp. 40–50.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. & Weigel, D. (2009), 'Simultaneous alignment of short reads against multiple genomes', *Genome Biology* **10**, R98.

Schrijver, A. (2003), *Combinatorial Optimization. Polyhedra and Efficiency. Vol. A*, Vol. 24 of Algorithms and Combinatorics. Berlin: Springer.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012), 'Metagenomic microbial community profiling using unique clade-specific marker genes', *Nature Methods* **9**(8), 811–814.

Seth, S., Välimäki, N., Kaski, S. & Honkela, A. (2014), 'Exploration and retrieval of whole-metagenome sequencing samples', *Bioinformatics* **30**(17), 2471–2479.

Sharma, V. K., Kumar, N., Prakash, T. & Taylor, T. D. (2012), 'Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin', *PLoS One* **7**(4), e34030.

Shawe-Taylor, J. & Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Shimbel, A. (1955), Structure in communication nets, in *Proceedings of the Symposium on Information Networks (New York, 1954)*. Brooklyn, NY: Polytechnic Press of the Polytechnic Institute of Brooklyn, pp. 199–203.

Simpson, J. T. & Durbin, R. (2010), 'Efficient construction of an assembly string graph using the FM-index', *Bioinformatics [ISMB]* **26**(12), 367–373.

Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. (2009), 'Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions', *Proceedings of the National Academy of Sciences* **106**(8), 2677–2682.

Sirén, J., Välimäki, N. & Mäkinen, V. (2011), Indexing finite language representation of population genotypes, in *11th International Workshop on Algorithms in Bioinformatics (WABI 2011)*, Vol. 6833 of Lecture Notes in Computer Science. Berlin: Springer, pp. 270–281.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Sirén, J., Välimäki, N. & Mäkinen, V. (2014), 'Indexing graphs for path queries with applications in genome research', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11**(2), 375–388.

Smith, T. & Waterman, M. (1981), 'Identification of common molecular subsequences', *Journal of Molecular Biology* **147**(1), 195–197.

Smola, A. J. & Vishwanathan, S. (2003), Fast kernels for string and tree matching, in S. Becker, S. Thrun & K. Obermayer, eds., *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press, pp. 585–592.

Sobih, A., Belazzougui, D. & Cunial, F. (2015), Space-efficient substring kernels. To be published.

Sobih, A., Tomescu, A. I. & Mäkinen, V. (2015), High-throughput whole-genome metagenomic microbial profiling. To be published.

Song, K., Ren, J., Zhai, Z., Liu, X., Deng, M. & Sun, F. (2013), 'Alignment-free sequence comparison based on next-generation sequencing reads', *Journal of Computational Biology* **20**(2), 64–79.

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S. & Sun, F. (2014), 'New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing', *Briefings in Bioinformatics* **15**(3), 343–353.

Song, L. & Florea, L. (2013), 'CLASS: Constrained transcript assembly of RNA-seq reads', *BMC Bioinformatics* **14**(Suppl. 5), S14.

Spang, R., Rehmsmeier, M. & Stoye, J. (2002), 'A novel approach to remote homology detection: Jumping alignments', *Journal of Computational Biology* **9**(5), 747–760.

Su, C.-H., Wang, T.-Y., Hsu, M.-T., Weng, F. C.-H., Kao, C.-Y., Wang, D. & Tsai, H.-K. (2012), 'The impact of normalization and phylogenetic information on estimating the distance for metagenomes', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**(2), 619–628.

Su, X., Wang, X., Xu, J. & Ning, K. (2013), GPU-meta-storms: Computing the similarities among massive microbial communities using GPU, in *Proceedings of the 7th International Conference on Systems Biology (ISB)*, IEEE, pp. 69–74.

Tanaseichuk, O., Borneman, J. & Jiang, T. (2011), Separating metagenomic short reads into genomes via clustering, in *Algorithms in Bioinformatics*, Vol. 6833 of Lecture Notes in Computer Science. Berlin: Springer, pp. 298–313.

Tardos, E. (1985), 'A strongly polynomial algorithm to solve combinatorial linear programs', *Combinatorica* **5**, 247–255.

Teo, C. H. & Vishwanathan, S. (2006), Fast and space efficient string kernels using suffix arrays, in *Proceedings of the 23rd International Conference on Machine Learning*, ACM, pp. 929–936.

Tomescu, A. I., Kuosmanen, A., Rizzi, R. & Mäkinen, V. (2013a), 'A novel min-cost flow method for estimating transcript expression with RNA-Seq', *BMC Bioinformatics* **14**(Suppl. 5), S15.

Tomescu, A. I., Kuosmanen, A., Rizzi, R. & Mäkinen, V. (2013b), A novel combinatorial method for estimating transcript expression with RNA-Seq: Bounding the number of paths, in *Algorithms in Bioinformatics*, Vol. 8126 of Lecture Notes in Computer Science. Berlin: Springer, pp. 85–98.

Trapnell, C., Pachter, L. & Salzberg, S. L. (2009), 'TopHat: Discovering splice junctions with RNA-Seq', *Bioinformatics* **25**(9), 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. (2010), 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology* **28**(5), 511–515.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Tutte, W. (1954), 'A short proof of the factor theorem for finite graphs', *Canadian Journal of Mathematics* **6**(1954), 347–352.

Ukkonen, E. (1985), 'Algorithms for approximate string matching', *Information and Control* **64**(1–3), 100–118.

Ukkonen, E. (1995), 'On-line construction of suffix trees', *Algorithmica* **14**(3), 249–260.

Ulitsky, I., Burstein, D., Tuller, T. & Chor, B. (2006), 'The average common substring approach to phylogenomic reconstruction', *Journal of Computational Biology* **13**(2), 336–350.

Valenzuela, D., Välimäki, N. & Mäkinen, V. (2015), Variation calling over pan-genomes. To be published.

Välimäki, N. (2012), Least random suffix/prefix matches in output-sensitive time, in *23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012)*, Vol. 7354 of Lecture Notes in Computer Science. Berlin: Springer, pp. 269–279.

Välimäki, N., Ladra, S. & Mäkinen, V. (2012), 'Approximate all-pairs suffix/prefix overlaps', *Information and Computation* **213**, 49–58.

Välimäki, N. & Puglisi, S. J. (2012), Distributed string mining for high-throughput sequencing data, in *12th International Workshop on Algorithms in Bioinformatics (WABI 2012)*, Vol. 7534 of Lecture Notes in Computer Science. Berlin: Springer, pp. 441–452.

Välimäki, N. & Rivals, E. (2013), Scalable and versatile *k*-mer indexing for high-throughput sequencing data, in *Bioinformatics Research and Applications*, Vol. 7875 of Lecture Notes in Computer Science. Berlin: Springer, pp. 237–248.

van Emde Boas, P. (1977), 'Preserving order in a forest in less than logarithmic time and linear space', *Information Processing Letters* **6**(3), 80–82.

Vatinlen, B., Chauvet, F., Chrétienne, P. & Mahey, P. (2008), 'Simple bounds and greedy algorithms for decomposing a flow into a minimal set of paths', *European Journal of Operational Research* **185**(3), 1390–1401.

Vazirani, V. V. (2001), *Approximation Algorithms*. New York, NY: Springer.

Vinga, S. & Almeida, J. (2003), 'Alignment-free sequence comparison – a review', *Bioinformatics* **19**(4), 513–523.

Viterbi, A. (1967), 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Transactions on Information Theory* **13**(2), 260–269.

Wagner, R. A. & Fischer, M. J. (1974), 'The string-to-string correction problem', *Journal of the Association for Computing Machinery* **21**, 168–173.

Wan, L., Reinert, G., Sun, F. & Waterman, M. S. (2010), 'Alignment-free sequence comparison (II): Theoretical power of comparison statistics', *Journal of Computational Biology* **17**(11), 1467–1490.

Wandelt, S., Starlinger, J., Bux, M. & Leser, U. (2013), 'RCSI: Scalable similarity search in thousand(s) of genomes', *PVLDB* **6**(13), 1534–1545.

Wang, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. (2012a), 'MetaCluster 4.0: A novel binning algorithm for NGS reads and huge number of species', *Journal of Computational Biology* **19**(2), 241–249.

Wang, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. (2012b), 'MetaCluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample', *Bioinformatics* **28**(18), i356–i362.

Wang, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. (2014a), 'MetaCluster-TA: Taxonomic annotation for metagenomic data based on assembly-assisted binning', *BMC Genomics* **15**(Suppl. 1), S12.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press

Wang, Y., Liu, L., Chen, L., Chen, T. & Sun, F. (2014b), 'Comparison of metatranscriptomic samples based on *k*-tuple frequencies', *PLoS One* **9**(1), e84348.

Weiner, P. (1973), Linear pattern matching algorithm, in *Proceedings of the 14th Annual IEEE Symposium on Switching and Automata Theory*, pp. 1–11.

Weintraub, A. (1974), 'A primal algorithm to solve network flow problems with convex costs', *Management Science* **21**(1), 87–97.

West, D. B. (2000), *Introduction to Graph Theory*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall, Inc.

Willner, D., Thurber, R. V. & Rohwer, F. (2009), 'Metagenomic signatures of 86 microbial and viral metagenomes', *Environmental Microbiology* **11**(7), 1752–1766.

Yang, X., Chockalingam, S. P. & Aluru, S. (2012), 'A survey of error-correction methods for next-generation sequencing', *Briefings in Bioinformatics* **14**(1), 56–66.

Zerbino, D. & Birney, E. (2008), 'Velvet: Algorithms for de novo short read assembly using de Bruijn graphs', *Genome Research* **18**, 821–829.

Ziv, J. & Lempel, A. (1977), 'A universal algorithm for sequential data compression', *IEEE Transactions on Information Theory* **23**(3), 337–343.

Ziv, J. (2008), 'On finite memory universal data compression and classification of individual sequences', *IEEE Transactions on Information Theory* **54**(4), 1626–1636.

https://doi.org/10.1017/CBO9781139940023.019 Published online by Cambridge University Press