Kun-Mao Chao
Louxin Zhang

Sequence Comparison: Theory and Methods

– Monograph –

Springer

Foreword

My first thought when I saw a preliminary version of this book was: Too bad there was nothing like this book when I really needed it.

Around 20 years ago, I decided it was time to change my research directions. After exploring a number of possibilities, I decided that the area of overlap between molecular biology and computer science (which later came to be called "bioinformatics") was my best bet for an exciting career. The next decision was to select a specific class of problems to work on, and the main criterion for me was that algorithmic methods would be the main key to success. I decided to work on sequence analysis. A book like this could have, so to speak, straightened my learning curve.

It is amazing to me that those two conclusions still apply: bioinformatics is a tremendously vibrant and rewarding field to be in, and sequence comparison is (arguably, at least) the subfield of bioinformatics where algorithmic techniques play the largest role in achieving success. The importance of sequence-analysis methods in bioinformatics can be measured objectively, simply by looking at the numbers of citations in the scientific literature for papers that describe successful developments; a high percentage of the most heavily cited scientific publications in the past 30 years are from this new field. Continued growth and importance of sequence analysis is guaranteed by the explosive development of new technologies for generating sequence data, where the cost has dropped 1000-fold in the past few years, and this fantastic decrease in cost means that sequencing and sequence analysis are taking over jobs that were previously handled another way.

Careful study of this book will be valuable for a wide range of readers, from students wanting to enter the field of bioinformatics, to experienced users of bioinformatic tools wanting to use tool options more intelligently, to bioinformatic specialists looking for the killer algorithm that will yield the next tool to sweep the field. I predict that you will need more that just mastery of this material to reach stardom in bioinformatics – there is also a huge amount of biology to be learned, together with a regular investment of time to keep up with the latest in data-generation technology and its applications. However, the material herein will remain useful for years, as new sequencing technologies and biological applications come and go.

I invite you to study this book carefully and apply ideas from it to one of the most exciting areas of science. And be grateful that two professionals with a combined 30 years of experience have taken the time to open the door for you.

State College, Pennsylvania *Webb Miller*

June 2008

Preface

Biomolecular sequence comparison is the origin of bioinformatics. It has been extensively studied by biologists, computer scientists, and mathematicians for almost 40 years due to its numerous applications in biological sciences. Today, homology search is already a part of modern molecular biology. This book is a monograph on the state-of-the-art study of sequence alignment and homology search.

Sequence alignment, as a major topic of bioinformatics, is covered in most bioinformatics books. However, these books often tell one part of the story. The field is evolving. The BLAST program, a pearl of pearls, computes local alignments quickly and evaluates the statistical significance of any alignments that it finds. Although BLAST homology search is done more than 100,000 times per day, the statistical calculations used in this program are not widely understood by its users. In fact, these calculations keep on changing with advancement of alignment score statistics. Simply using BLAST without a reasonable understanding of its key ideas is not very different from using a PCR without knowing how PCR works. This is one of the motivations for us to write this book. It is intended for covering in depth a full spectrum of the field from alignment methods to the theory of scoring matrices and to alignment score statistics.

Sequence alignment deals with basic problems arising from processing DNA and protein sequence information. In the study of these problems, many powerful techniques have been invented. For instance, the filtration technique, powered with spaced seeds, is shown to be extremely efficient for comparing large genomes and for searching huge sequence databases. Local alignment score statistics have made homology search become a reliable method for annotating newly sequenced genomes. Without doubt, the ideas behind these outstanding techniques will enable new approaches in mining and processing structural information in biology. This is another motivation for us to write this book.

This book is composed of eight chapters and three appendixes. Chapter 1 works as a tutorial to help all levels of readers understand the connection among the other chapters. It discusses informally why biomolecular sequences are compared through alignment and how sequence alignment is done efficiently.

Chapters 2 to 5 form the method part. This part covers the basic algorithms and methods for sequence alignment. Chapter 2 introduces basic algorithmic techniques that are often used for solving various problems in sequence comparison.

In Chapter 3, we present the Needleman-Wunsch and Smith-Waterman algorithms, which, respectively, align a pair of sequences globally and locally, and their variants for coping with various gap penalty costs. For analysis of long genomic sequences, the space restriction is more critical than the time constraint. We therefore introduce an efficient space-saving strategy for sequence alignment. Finally, we discuss a few advanced topics of sequence alignment.

Chapter 4 introduces four popular homology search programs: FASTA, BLAST family, BLAT, and PatternHunter. We also discuss how to implement the filtration idea used in these programs with efficient data structures such as hash tables, suffix trees, and suffix arrays.

Chapter 5 covers briefly multiple sequence alignment. We discuss how a multiple sequence alignment is scored, and then show why the exact method based on a dynamic-programming approach is not feasible. Finally, we introduce the progressive alignment approach, which is adopted by ClustalW, MUSCLE, YAMA, and other popular programs for multiple sequence alignment.

Chapters 6 to 8 form the theory part. Chapter 6 covers the theoretic aspects of the seeding technique. PatternHunter demonstrates that an optimized spaced seed improves sensitivity substantially. Accordingly, elucidating the mechanism that confers power to spaced seeds and identifying good spaced seeds become new issues in homology search. This chapter presents a framework of studying these two issues by relating them to the probability of a spaced seed hitting a random alignment. We address why spaced seeds improve homology search sensitivity and discuss how to design good spaced seeds.

The Karlin-Altschul statistics of optimal local alignment scores are covered in Chapter 7. Optimal segment scores are shown to follow an extreme value distribution in asymptotic limit. The Karlin-Altschul sum statistic is also introduced. In the case of gapped local alignment, we describe how the statistical parameters of the distribution of the optimal alignment scores are estimated through empirical approach and discuss the edge-effect and multiple testing issues. We also relate theory to the calculations of the Expect and P-values in BLAST program.

Chapter 8 is about the substitution matrices. We start with the reconstruction of popular PAM and BLOSUM matrices. We then present Altschul's theoretic-information approach to scoring matrix selection and recent work on compositional adjustment of scoring matrices for aligning sequences with biased letter frequencies. Finally, we discuss gap penalty costs.

This text is targeted to a reader with a general scientific background. Little or no prior knowledge of biology, algorithms, and probability is expected or assumed. The basic notions from molecular biology that are useful for understanding the topics covered in this text are outlined in Appendix A. Appendix B provides a brief introduction to probability theory. Appendix C lists popular software packages for pairwise alignment, homology search, and multiple alignment.

This book is a general and rigorous text on the algorithmic techniques and mathematical foundations of sequence alignment and homology search. But, it is by no means comprehensive. It is impossible to give a complete introduction to this field because it is evolving too quickly. Accordingly, each chapter concludes with the bibliographic notes that report related work and recent progress. The reader may ultimately turn to the research articles published in scientific journals for more information and new progress.

Most of the text is written at a level that is suitable for undergraduates. It is based on lectures given to the students in the courses in bioinformatics and mathematical genomics at the National University of Singapore and the National Taiwan University each year during 2002 – 2008. These courses were offered to students from biology, computer science, electrical engineering, statistics, and mathematics majors. Here, we thank our students in the courses we have taught for their comments on the material, which are often incorporated into this text.

Despite our best efforts, this book may contain errors. It is our responsibility to correct any errors and omissions. A list of errata will be compiled and made available at http://www.math.nus.edu.sg/~matzlx/sequencebook.

Taiwan & Singapore — *Kun-Mao Chao*
June 2008 — *Louxin Zhang*

Contents

References

1. Altschul, S.F. (1989) Generalized affine gap costs for protein sequence alignment. *Proteins* **32**, 88-96.
2. Altschul, S.F. (1989) Gap costs for multiple sequence alignment. *J. Theor. Biol.* **138**, 297-309.
3. Altschul, S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J Mol. Biol.* **219**, 555-65.
4. Altschul, S.F., Boguski, M.S., Gish, W., and Wootton, J.C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119-129.
5. Altschul, S.F., Bundschuh, R., Olsen, R., and Hwa, T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.* **29**, 351-361.
6. Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Methods Enzymol.* **266**, 460-480.
7. Altschul, S.F., Gish, W., Miller, W., Myers, E., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
8. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped Blast and Psi-Blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
9. Altschul, S.F., Wootton, J.C., Gertz, E.M., Agarwala, R., Morgulis, A., Schaffer, A.A., and Yu, Y.K. (2005) Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **272**, 5101-5109.
10. Arratia, R., Gordon, L., and Waterman, M.S. (1986) An extreme value theory for sequence matching. *Ann. Stat.* **14**, 971-983.
11. Arratia, R., Gordon, L., and Waterman, M.S. (1990) The Erdös-Renýi law in distribution, for coin tossing and sequence matching. *Ann. Stat.* **18**, 539-570.
12. Arratia, R. and Waterman, M.S. (1985) An Erdös-Renýi law with shifts. *Adv. Math.* **55**, 13-23.
13. Arratia, R. and Waterman, M.S. (1986) Critical phenomena in sequence matching. *Ann. Probab.* **13**, 1236-1249.
14. Arratia, R. and Waterman, M.S. (1989) The Erdös-Renýi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17**, 1152-1169.
15. Arratia, R. and Waterman, M.S. (1994) A phase transition for the scores in matching random sequences allowing deletions. *Ann. Appl. Probab.* **4**, 200-225.
16. Arvestad, L. (2006), Efficient method for estimating amino acid replacement rates. *J. Mol. Evol.* **62**, 663-673.
17. Baase, S. and Gelder, A.V. (2000) *Computer Algorithms - Introduction to Design and Analysis*. Addison-Wesley Publishing Company, Reading, Massachusetts.
18. Bafna, V., Lawler, E.L., and Pevzner, P.A. (1997) Approximation algorithms for multiple sequence alignment. *Theor. Comput. Sci.* **182**, 233-244.

19. Bafna, V. and Pevzner, P.A. (1996) Genome rearrangements and sorting by reversals. *SIAM J. Comput.* **25**, 272-289.
20. Bahr, A., Thompson, J.D., Thierry, J.C., and Poch, O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats. transmembrane sequences and circular permutations. *Nucleic Acids Res.* **29**, 323-326.
21. Bailey, T.L. and Gribskov, M. (2002) Estimating and evaluating the statistics of gapped local-alignment scores. *J. Comput. Biol.* **9**, 575-593.
22. Balakrishnan, N. and Koutras, M.V. (2002) *Runs and Scans with Applications*. John Wiley & Sons, New York.
23. Batzoglou, S. (2005) The many faces of sequence alignment. *Brief. Bioinform.* **6**, 6-22.
24. Bauer, M., Klau, G.W., and Reinert, K. (2007) Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics* **8**, 271.
25. Bellman, R. (1957) *Dynamic Programming*. Princeton University Press, Princeton, New Jersey.
26. Benner, S.A., Cohen, M.A., and Gonnet, G.H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteinsm. *J. Mol. Biol.* **229**, 1065-1082.
27. Bentley, J. (1986) *Programming Pearls*. Addison-Wesley Publishing Company, Reading, Massachusetts.
28. Bray, N. and Pachter, L. (2004) MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14**, 693-699.
29. Brejovà, B., Brown D., and Vinař, T. (2004) Optimal spaced seeds for homologous coding regions. *J. Bioinform. Comput. Biol.* **1**, 595-610.
30. Brejovà, B., Brown, D., and Vinař, T. (2005) Vector seeds: an extension to spaced seeds. *J. Comput. Sys. Sci.* **70**, 364–380.
31. Brown, D.G. (2005) Optimizing multiple seed for protein homology search. *IEEE/ACM Trans. Comput. Biol. and Bioinform.* **2**, 29-38.
32. Brudno, M., Do, C., Cooper, G., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721-731.
33. Buhler, J. (2001) Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics* **17,** 419-428.
34. Buhler, J., Keich, U., and Sun, Y. (2005) Designing seeds for similarity search in genomic DNA. *J. Comput. Sys. Sci.* **70**, 342-363.
35. Bundschuh, R. (2002) Rapid significance estimation in local sequence alignment with gaps. *J. Comput. Biol.* **9**, 243-260.
36. Burkhardt, S. and Kärkkäinen, J. (2003) Better filtering with gapped q-grams. *Fund. Inform.* **56**, 51-70.
37. Califano, A. and Rigoutsos, I. (1993) FLASH: A fast look-up algorithm for string homology. In *Proc. 1st Int. Conf. Intell. Sys. Mol. Biol.*, AAAI Press, pp. 56-64.
38. Carrilo, H. and Lipman, D. (1988) The multiple sequence alignment problem in biology. *SIAM J. Applied Math.* **48**, 1073-1082.
39. Chan, H.P. (2003) Upper bounds and importance sampling of p-values for DNA and protein sequence alignments. *Bernoulli* **9**, 183-199
40. Chao, K.-M., Hardison, R.C., and Miller, W. (1993) Locating well-conserved regions within a pairwise alignment. *Comput. Appl. Biosci.* **9**, 387-396.
41. Chao, K.-M., Hardison, R.C., and Miller, W. (1994) Recent developments in linear-space alignment methods: a survey. *J. Comput. Biol.* **1**, 271-291.
42. Chao, K.-M. and Miller, W. (1995) Linear-space algorithms that build local alignments from fragments. *Algorithmica* **13**, 106-134.
43. Chao, K.-M., Pearson, W.R., and Miller, W. (1992) Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.* **8**, 481-487.
44. Chen, L. (1975) Poisson approximation for dependent trials. *Ann. Probab.* **3**, 534-545.
45. Chiaromonte, F., Yap, V.B., and Miller, W. (2002) Scoring pairwise genomic sequence alignments. In *Proc. Pac. Symp. Biocomput.*, 115-126.

46. Chiaromonte, F., Yang, S., Elnitski, L., Yap, V.B., Miller, W., and Hardison, R.C. (2001) Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc. Nat'l. Acad. Sci. USA* **98**, 14503-8.
47. Choi, K.P. and Zhang, L.X. (2004) Sensitivity analysis and efficient method for identifying optimal spaced seeds. *J. Comput. Sys. Sci.* **68**, 22-40.
48. Choi, K.P., Zeng, F., and Zhang L.X. (2004) Good spaced seeds for homology search. *Bioinformatics* **20**, 1053-1059.
49. Chv́tal V and Sankoff D. (1975) Longest common subsequence of two random sequences. *J. Appl. Probab.* **12**, 306-315.
50. Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, UK.
51. Collins, J.F., Coulson, A.F.W., and Lyall, A. (1998) The significance of protein sequence similarities. *Comput. Appl. Biosci.* **4**, 67-71.
52. Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C. (2001) *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts.
53. Csürös, M., and Ma, B. (2007) Rapid homology search with neighbor seeds. *Algorithmica* **48**, 187-202.
54. Darling, A., Treangen, T., Zhang, L.X., Kuiken, C., Messeguer, X., and Perna, N. (2006) Procrastination leads to efficient filtration for local multiple alignment. In *Proc. 6th Int. Workshop Algorithms Bioinform. Lecture Notes in Bioinform.*, vol. 4175, pp.126-137.
55. Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). A model of evolutionary changes in proteins. In *Atlas of Protein Sequence and Structure* vol. 5, suppl 3 (ed. M.O. Dayhoff), 345-352, Nat'l Biomed. Res. Found, Washington.
56. Dembo, A., Karlin, S., and Zeitouni, O. (1994) Critical phenomena for sequence matching with scoring. *Ann. Probab.* **22**, 1993-2021.
57. Dembo, A., Karlin, S., and Zeitouni, O. (1994) Limit distribution of maximal non-aligned two sequence segmental score. *Ann. Probab.* **22**, 2022-2039.
58. Deonier, R.C., Tavaré, S., and Waterman, M.S. (2005), *Computational Genome Analysis*. Springer, New York.
59. Do, C.B., Mahabhashyam, M.S.P., Brudno, M., and Batzoglou, S. (2005) PROBCONS: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330-340.
60. Dobzhansky, T. and Sturtevant, A.H. (1938) Inversions in the chromosomes of Drosophila pseudoobscura. *Genetics* **23**, 28-64.
61. Durbin, R., Eddy, S., Krogh, A., and Mitichison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
62. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797.
63. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, no. 113.
64. Ewens, W.J. and Grant, G.R. (2001) *Statistical Methods in Bioinformatics: An Introduction*. Springer-Verlag, New York.
65. Farach-Colton, M., Landau, G., Sahinalp, S.C., and Tsur, D. (2007) Optimal spaced seeds for faster approximate string matching. *J. Comput. Sys. Sci.* **73**, 1035-1044.
66. Fayyaz, A.M., Mercier, S., Ferré, Hassenforder, C. (2008) New approximate *P*-value of gapped local sequence alignments. *C. R. Acad. Sci. Paris Ser. I* **346**, 87-92.
67. Feller, W. (1966) *An introduction to Probability Theory and its Applications*. Vol. 2 (1st edition), John Wiley & Sons, New York.
68. Feller, W. (1968) *An introduction to Probability Theory and its Applications*. Vol. I (3rd edition), John Wiley & Sons, New York.
69. Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360.
70. Feng, D.F. Johnson, M.S. and Doolittle, R.F. (1985) Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.* **21**, 112-125.

71. Fitch, W.M. and Smith, T.F. (1983) Optimal sequence alignments. *Proc. Nat'l. Acad. Sci. USA* **80**, 1382-1386.
72. Flannick, J., and Batzoglou, S. (2005) Using multiple alignments to improve seeded local alignment algorithms. *Nucleic Acids Res.* **33**, 4563-4577.
73. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967-974.
74. Gertz, E.M. (2005) BLAST scoring parameters. *Manuscript*(ftp://ftp.ncbi.nlm.nih.gov/blast/documents/developer/scoring.pdf).
75. Giegerich, R. and Kurtz, S. (1997) From Ukkonen to McCreight and Weiner: A unifying view of linear-time suffix tree construction. *Algorithmica* **19**, 331-353.
76. Gilbert, W. (1991) Towards a paradigm shift in biology. *Nature* **349**, 99.
77. Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443-5.
78. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708.
79. Gotoh, O. (1989) Optimal sequence alignment allowing for long gaps. *Bull. Math. Biol.* **52**, 359-373.
80. Gotoh, O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264**, 823-838.
81. Gribskov, M., Luthy, R., and Eisenberg, D. (1990) Profile analysis. In R.F. Doolittle (ed.) *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, Methods in Enzymol., vol. 183, Academic Press, New York, pp. 146-159.
82. Grossman, S. and Yakir, B. (2004) Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignment. *Bernoulli* **10**, 829-845.
83. Guibas, L.J. Odlyzko, A.M. (1981) String overlaps, pattern matching, and nontransitive games. *J. Combin. Theory* (series A) **30**, 183-208.
84. Gupta, S.K., Kececioglu, J., and Schaffer, A.A. (1995) Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J. Comput. Biol.* **2**, 459-472.
85. Gusfield, D. (1997) *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, Cambridge, UK.
86. Hannenhalli, S. and Pevzner, P.A. (1999) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *J. Assoc. Comput. Mach.* **46**, 1-27.
87. Hardy, G., Littlewood, J.E., and Pólya, G. (1952) *Inequalities*, Cambridge University Press, Cambridge, UK.
88. Henikoff, S. and Henikoff, JG (1992) Amino acid substitution matrices from protein blocks. *Proc. Nat'l Acad. Sci. USA* **89**, 10915-10919.
89. Henikoff, S. and Henikoff, JG (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49-61.
90. Hirosawa, M., Totoki, Y., Hoshida, M., and Ishikawa, M. (1995) Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput. Appl. Biosci.* **11**, 13-18.
91. Hirschberg, D.S. (1975) A linear space algorithm for computing maximal common subsequences. *Comm. Assoc. Comput. Mach.* **18**, 341-343.
92. Huang, X. and Miller, W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* **12**, 337-357.
93. Huang, X. and Chao, K.-M. (2003) A generalized global alignment algorithm. *Bioinformatics* **19**, 228-233.
94. Huffman, D.A. (1952) A method for the construction of minimum-redundancy codes. *Proc. IRE* **40**, 1098-1101.
95. Ilie, L., and Ilie, S. (2007) Multiple spaced seeds for homology search. *Bioinformatics* **23**, 2969-2977
96. Indyk, P. and Motwani, R. (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proc. 30th Ann. ACM Symp. Theory Comput.*, 604-613.

97. Jones D.T., Taylor, W.R., and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275-82.
98. Jones, N.C. and Pevzner, P.A. (2004) *Introduction to Bioinformatics Algorithms*. The MIT Press, Cambridge, Massachusetts.
99. Karlin, S. (2005) Statistical signals in bioinformatics. *Proc Nat'l Acad Sci USA.* **102**, 13355-13362.
100. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat'l. Acad. Sci. USA* **87**, 2264-2268.
101. Karlin, S. and Altschul, S.F. (1993) Applications and statistics fro multiple high-scoring segments in molecular sequences. *Proc. Nat'l Acad. Sci. USA* **90**, 5873-5877.
102. Karlin, S. and Dembo, A. (1992) Limit distribution of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Probab.* **24**, 113-140.
103. Karlin, S. and Ost, F. (1988) Maximal length of common words among random letter sequences. *Ann. Probab.* **16**, 535-563.
104. Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., Miller, W., Pedersen, J.S., Pohl, A., Raney, B.J., Rhead, B., Rosenbloom, K.R., Smith, K.E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A.S., Haussler, D., Kent, W.J. (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.* **36**, D773-779.
105. Karp, R.M., and Rabin, M.O. (1987) Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.* **31**, 249-260.
106. Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic Acids Res.* **20**, 511-518.
107. Kececioglu, J. and Starrett, D. (2004) Aligning alignments exactly. In *Proc. RECOMB*, 85-96.
108. Keich, U., Li, M., Ma, B., and Tromp, J. (2004) On spaced seeds for similarity search. *Discrete Appl. Math.* **3**, 253-263.
109. Kent, W.J. (2002) BLAT: The BLAST-like alignment tool. *Genome Res.* **12**, 656-664.
110. Kent, W.J. and Zahler, A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans Genomic Alignment. *Genome Res.* **10**, 1115-1125.
111. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK.
112. Kisman, D., Li, M., Ma, B., and Wang, L. (2005) tPatternHunter: gapped, fast and sensitive translated homology search. *Bioinformatics* **21**, 542-544.
113. Knuth, D.E. (1973) *The art of computer programming*. Vol. 1, Addison-Wesley Publishing Company, Reading, Massachusetts.
114. Knuth, D.E. (1973) *The art of computer programming*. Vol. 3, Addison-Wesley Publishing Company, Reading, Massachusetts.
115. Kong, Y. (2007), Generalized correlation functions and their applications in selection of optimal multiple spaced seeds for homology search. *J. Comput. Biol.* **14**, 238-254.
116. Korf, I., Yandell, M., and Bedell, J. (2003) *BLAST*. O'reilly, USA.
117. Kschischo, M., Lässig, M., and Yu, Y.-K. (2005) Towards an accurate statistics of gapped alignment. *Bull. Math. Biol.* **67**, 169-191.
118. Kucherov G., Noè, L. and M. Roytberg M. (2005) Multiseed lossless filtration. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2**, 51-61.
119. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biology* **5**, R12.
120. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2007) ClustalW and ClustalX version 2. *Bioinformatics* **23**, 2947-2948.
121. Lassmann, T. and Sonnhammer, L.L. (2005) Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**, 298.

122. Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257-260.
123. Li, M., Ma, B., Kisman, D., and Tromp, J. (2004) PatternHunter II: Highly sensitive and fast homology search. *J. Bioinform. Comput. Biol.* **2**, 417-439.
124. Li, M., Ma, B. Kisman, D. and Tromp, J. (2004) PatternHunter II: Highly Sensitive and Fast Homology Search. *J. Bioinform Comput. Biol.* **2** (3),417-439.
125. Li, M., Ma, B. and Zhang, L.X. (2006) Superiority and complexity of spaced seeds. In *Proc. 17th SIAM-ACM Symp. Discrete Algorithms*. 444-453.
126. Li, W.H., Wu, C.I., and Luo, C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150-174.
127. Lipman, D.J., Altschul, S.F., and Kececioglu, J.D. (1989) A tool for multiple sequence alignment. *Proc. Nat'l. Acad. Sci. USA* **86**, 4412-4415.
128. Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441.
129. Lothaire, M. (2005) *Applied Combinatorics on Words*. Cambridge University Press, Cambridge, UK.
130. Ma, B. and Li, M. (2007) On the complexity of spaced seeds. *J. Comput. Sys. Sci.* **73**, 1024-1034.
131. Ma, B., Tromp, J., and Li, M. (2002) PatternHunter-faster and more sensitive homology search. *Bioinformatics* **18**, 440-445.
132. Ma, B., Wang, Z., and Zhang, K. (2003) Alignment between two multiple alignments. *Proc. Combin. Pattern Matching*, 254-265.
133. Manber, U. (1989) *Introduction to Algorithms*. Addison-Wesley Publishing Company, Reading, Massachusetts.
134. Manber, U. and Myers, E. (1991) Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.* **22**, 935-948.
135. McCreight, E.M. (1976) A space-economical suffix tree construction algorithm. *J. Assoc. Comput. Mach.* **23**, 262-272.
136. McLachlan, A.D. (1971) Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. *J. Mol. Biol.* **61**, 409-424.
137. Metzler, D. (2006) Robust E-values for gapped local alignment. *J. Comput. Biol.* **13**, 882-896.
138. Miller, W. (2001) Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17**, 391-397.
139. Miller, W. and Myers, E. (1988) Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50**, 97-120.
140. Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S.L., Nekrutenko, A., Giardine, B., Harris, R.S., Tyekucheva, S., Diekhans, M., Pringle, T.H., Murphy, W.J., Lesk, A., Weinstock, G.M., Lindblad-Toh, K., Gibbs, R.A., Lander, E.S., Siepel, A., Haussler, D., and Kent, W.J. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17**,1797-1808.
141. Mitrophanov, A.Y. and Borodovsky, M. (2006) Statistical significance in biological sequence analysis. *Brief. Bioinform.* **7**, 2-24.
142. Mohana Rao, J.K. (1987) New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. Peptide Protein Res.* **29**, 276-281.
143. Morgenstern, B., French, K., Dress, A., and Werner, T. (1998) DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics* **14**, 290-294.
144. Mott, R. (1992) Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59-75.
145. Mott, R. (2000) Accurate formula for P-values for gapped local sequence and profile alignment. *J. Mol. Biol.* **276**, 71-84.
146. Mott, R and Tribe, R. (1999) Approximate statistics of gapped alignments. *J. Comput. Biol.* **6**, 91-112.

147. Müller, T. and Vingron, M. (2000) Modeling amino acid replacement. *J. Comput. Biol.* **7**, 761-776.
148. Müller, T., Spang, R., and Vingron, M. (2002) Estimating amino acid substitution models: A comparison of Dayoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.* **19**, 8-13.
149. Myers, G. (1999) Whole-Genome DNA sequencing. *Comput. Sci. Eng.* **1**, 33-43.
150. Myers, E. and Miller, W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11-17.
151. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443-453.
152. Nicodème, P., Salvy, B., and Flajolet, P. (1999) Motif Statistics. In *Lecture Notes in Comput. Sci.*, vol. 1643, 194-211, New York.
153. Noè, L., and Kucherov, G. (2004) Improved hit criteria for DNA local alignment. *BMC Bioinformatics* **5**, no.159.
154. Notredame, C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.* **3**, e123.
155. Notredame, C., Higgins, D., and Heringa, J. (2000) T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.* **302**, 205-217.
156. Overington, J., Donnelly, D., Johnson, M.S., Sali, A., and Blundell, T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216-26.
157. Park, Y. and Spouge, J.L. (2002) The correction error and finite-size correction in an ungapped sequence alignment. *Bioinformatics* **18**, 1236-1242.
158. Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures, *J. Mol. Biol.* **224**, 461-471.
159. Pearson, W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Science* **4**, 1145-1160.
160. Pearson, W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.
161. Pearson, W.R. and Lipman, D. (1988) Improved tools for biological sequence comparison. *Proc. Nat'l. Acad. Sci USA* **85**, 2444-2448.
162. Pearson, W.R. and Wood, T.C. (2003) Statistical Significance in Biological Sequence Comparison. In *Handbook of Statistical Genetics* (edited by D.J. Balding, M. Bishop and C. Cannings), 2nd Edition. John Wiley & Sons, West Sussex, UK.
163. Pei, J. and Grishin, N.V. (2006) MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucleic Acids Res.* **34**, 4364-4374.
164. Pei, J. and Grishin, N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23**, 802-808.
165. Pevzner, P.A. (2000) *Computational Molecular Biology: An Algorithmic Approach*. The MIT Press, Cambridge, Massachusetts.
166. Pevzner, P.A. and Tesler, G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* **13**, 37-45.
167. Pevzner, P.A. and Waterman, M.S. (1995) Multiple filtration and approximate pattern matching. *Algorithmica* **13**, 135-154.
168. Preparata, F.P., Zhang, L.X., and Choi, K.P. (2005) Quick, practical selection of effective seeds for homology search. *J. Comput. Biol.* **12**, 1137-1152.
169. Reich, J.G., Drabsch, H., and Däumler, A. (1984) On the statistical assessment of similarities in DNA sequences. *Nucleic Acids Res.* **12**, 5529-5543.
170. Rényi, A. (1970) *Foundations of Probability*. Holden-Day, San Francisco.
171. Reinert, G., Schbath, S., and Waterman, M.S. (2000), Probabilistic and statistical properties of words: An overview. *J. Comput. Biol.* **7**, 1 - 46.
172. Risler, J.L., Delorme, M.O., Delacroix, H., and Henaut, A. (1988) Amino Acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol.* **204**, 1019-1029.

173. Robinson, A.B. and Robinson, L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Nat'l. Acad. Sci USA* **88**, 8880-8884.
174. Sankoff, D. (2000) The early introduction of dynamic programming into computational biology. *Bioinformatics* **16**, 41-47.
175. Sankoff, D. and Kruskal, J.B. (eds.) (1983). *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparisons*. Addison-Wesley, Reading, Massachusetts.
176. Schwager, S.J. (1983) Run probabilities in sequences of Markov-dependent trials, *J. Amer. Stat. Assoc.* **78**, 168-175.
177. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003) Human-mouse alignment with BLASTZ. *Genome Res.* **13**, 103-107.
178. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R.C., and Miller, W. (2000) PipMaker - a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577-86.
179. Siegmund, D. and Yakir, B. (2000) Approximate p-value for local sequence alignments. *Ann. Stat.* **28**, 657-680.
180. Smith, T.F. and Waterman, M.S., (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
181. Smith, T.F., Waterman, M.S., and Burks, C. (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.* **13**, 645-656.
182. Spang, R. and Vingron, M. (1998) Statistics of large-scale sequences searching. *Bioinformatics* **14**, 279-284.
183. Spitzer, F. (1960) A tauberian theorem and its probability interpretation. *Trans. Amer Math Soc.* **94**, 150-169.
184. States, D.J., Gish, W., and Altschul, S.F. (1991), Imporved sensitivity of nucleic acid databases searches using application-specific scoring matrices. *Methods* **3**, 61-70.
185. Sun, Y. and Buhler, J. Designing multiple simultaneous seeds for DNA similarity search. *J. Comput. Biol.* **12**, 847-861.
186. Sun, Y., and Buhler, J. (2006) Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinformatics* **7**, no. 133.
187. Sze, S.-H., Lu, Y., and Yang, Q. (2006) A polynomial time solvable formulation of multiple sequence alignment. *J. Comput. Biol.* **13**, 309-319.
188. Taylor, W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.* **119**, 205-218.
189. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
190. Thompson, J.D., Plewniak, F., and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27**, 2682-2690.
191. Ukkonen, E. (1995) On-line construction of suffix trees. *Algorithmica* **14**, 249-260.
192. Vingron, M. and Argos, P. (1990) Determination of reliable regions in protein sequence alignment. *Protein Eng.* **3**, 565-569.
193. Vingron, M. and Waterman, M.S. (1994) Sequence alignment and penalty choice: Review of concepts, case studies and implications. *J. Mol. Biol.* **235**, 1-12.
194. Wagner, R.A. and Fischer, M.J. (1974) The string-to-string correction problem. *J. Assoc. Comput. Mach.* **21**, 168-173.
195. Wang, L. and Jiang, T. (1994) On the complexity of multiple sequence alignment. *J. Comput. Biol.* **1**, 337-348.
196. Waterman, M.S. (1984) Efficient sequence alignment algorithms. *J. Theor. Biol.* **108**, 333-337.
197. Waterman, M.S. (1995) *Introduction to Computational Biology*. Chapman and Hall, New York.
198. Waterman, M. S. and Eggert, M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Mol. Biol.* **197**, 723-728.

199. Waterman, W.S., Gordon, L., and Arratia, R. (1987) Phase transition in sequence matches and nucleic acid structure. *Proc. Nat'l. Acad. Sci. USA* **84**, 1239-1243.
200. Waterman, M.S. and Vingron, M. (1994) Rapid and accurate estimate of statistical significance for sequence database searches. *Proc. Nat'l. Acad. Sci. USA* **91**, 4625-4628.
201. Webber, C. and Barton, G.J. (2001) Estimation of *P*-value for global alignments of protein sequences. *Bioinformatics* **17**, 1158-1167.
202. Weiner, P. (1973) Linear pattern matching algorithms. In *Proc. the 14th IEEE Annu. Symp. on Switching and Automata Theory*, pp. 1-11.
203. Wilbur, W.J. (1985), On the PAM matrix model of protein evolution. *Mol. Biol. Evol.* **2**, 434-447.
204. Wilbur, W. and Lipman, D. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Nat. Acad. Sci. USA* **80**, 726-730.
205. Wilbur, W. and Lipman, D. (1984) The context dependent comparison of biological sequences. *SIAM J. Appl. Math.* **44**, 557-567.
206. Xu, J., Brown, D. Li, M., and Ma, M. (2006) Optimizing multiple spaced seeds for homology search. *J. Comput. Biol.* **13**, 1355-1368.
207. Yang, I.-H., Wang, S.-H., Chen, Y.-H., Huang, P.-H., Ye, L., Huang, X. and Chao, K.-M. (2004) Efficient methods for generating optimal single and multiple spaced seeds. In *Proc. IEEE 4th Symp. on Bioinform. and Bioeng.*, pp. 411-418.
208. Yang J.L., and Zhang, L.X. (2008) Run probabilities of seed-like patterns and identifying good transition seeds. *J. Comput. Biol.* (in press).
209. Yap, V.B. and Speed, T. (2005), Estimating substitution matrices. In *Statistical Methods in Molecular Evolution* (ed. R. Nielsen), Springer.
210. Ye, L. and Huang, X. (2005) MAP2: Multiple alignment of syntenic genomic sequences. *Nucleic Acids Res.* **33**, 162-170.
211. Yu, Y.K., Wootton, J.C., and Altschul, S.F. (2003) The compositional adjustment of amino acid substitution matrices. *Proc. Nat'l. Acad. Sci. USA.* **100**, 15688-93.
212. Yu, Y.K. and Altschul, S.F. (2005) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics* **21**, 902-11.
213. Zachariah, M.A., Crooks, G.E., Holbrook, S.R., Brenner, S.E. (2005) A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins* **58**, 329-38.
214. Zhang, L.X. (2007) Superiority of spaced seeds for homology search. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4**, 496-505.
215. Zhang, Z, Schwartz, S, Wagner, L, and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203-214.
216. Zhou L. and Florea, L. (2007) Designing sensitive and specific spaced seeds for cross-species mRNA-to-genome alignment. *J. Comput. Biol.* **14**, 113-130,
217. Zuker, M. and Somorjal, R.L. (1989) The alignment of protein structures in three dimensions. *Bull. Math. Biol.* **50**, 97-120.