Solution Quality and Efficiency in Discrete Optimization

2 Scientific Work

2.1 Summary

The most interesting discrete optimization problems are computationally very hard. That is why we cannot hope for efficient (practicable) algorithms that would guarantee to compute appropriate exact optimal solutions. The only way to attack these hard problems is to relax our demand on acceptable solutions and to hope for the efficient solvability of the resulting problem variants. To shift from intractability to efficiency with minimal deviation from our initial requirements, we have the following three main approaches:

1. To specify large subclasses of problem instances for which solutions can be computed efficiently. This allows for more sophisticated problem analysis where the hardness of problem instances can be examined. Problem instances can thus be classified with respect to computational hardness.

2. To deviate from the requirement to compute an optimal solution by permitting solutions of sub-optimal yet satisfactory quality.

3. To replace the algorithms' guarantee of optimal behavior by a guarantee to behave desirably with a high probability.

We shall consider all of these approaches and their combinations and investigate their potential.

Our main interest is in the development of concrete techniques for the design of algorithms for hard problems. In particular, we plan to focus on the concept of stability of approximation algorithms introduced in our previous work. This concept belongs to the first category of the approaches above and enables us to find good approximate solutions for problems considered to be hard even for approximation algorithms.

Regarding the problem classes, we focus on graph-theoretical problems with relevance to logistics and operations research and to problems in computational biology.

2.2 Research Plan

2.2.1 State of the Art

In the following sub-sections, the state of the art for the different problem classes will be presented. We start with graph-theoretical problems in the context of logistics and operations research. This is followed by a description of algorithms for problems in the field of compuataional biology. Our own contributions are marked with a star ($\star$).

2.2.1.1 Stability of Approximation and Graph-theoretical Problems

The goal behind the concept of stability of approximation is to attack hard problems, for which no good algorithms exist that would yield a constant approximation ratio. The approach is to define an additional constraint to the problem, such that the constrained class of the problem allows for good approximation algorithms. Employing a distance measure on the problem, we can prove that good approximation ratios can also be achieved for problems outside this constrained class that lie close to it with respect to the distance measure. The resulting increase in approximation ratio is a function of the distance measure, but not of the input size. This concept was introduced in [BHK+02$\star$], where it was successfully applied to the TSP.

In the case of the TSP, the distance measure is defined by means of the deviation from the triangle inequality, measured in the parameter β. The metric TSP is 1.5-approximable by the well-known Christofides algorithm [Chr76]. Although this algorithm is not stable for input instances satisfying a relaxed triangle inequality only, it can be modified as proposed in [BHK+02$\star$], leading to a stable algorithm for the general TSP with an approximation ratio of $\frac{3}{2}\beta^2$. This improves on a result by Andreae and Bandelt [AB95], who gave a $\frac{3}{2}\beta^2 + \frac{1}{2}\beta$ approximation algorithm. The same problem was also considered in [BC00], where a 4β-approximation was presented, and by [And01], that gave a $\beta^2 + \beta$-approximation algorithm.

On the other hand, also the TSP with sharpened triangle inequality has been considered. [BS00$\star$] presents algorithms that provide approximation ratios that converge to 1 in accordance with the sharpening of the triangle inequality. The asymmetric TSP with sharpened triangle inequality was investigated in [CR02, Blä03]. [EK01] gives some hardness results for the TSP with sharpened triangle inequality.

Approximation algorithms for connectivity problems in graphs satisfying a sharpened triangle inequality have been proposed in [BBH+02$\star$, BBH+03$\star$, BBH+04$\star$, FP03].

2.2.1.2 Computational Biology

Molecular biology has emerged as one of the most important application areas for computer science over the recent years. The interpretation and analysis of data generated in biological experiments often requires non-trivial algorithmic techniques. Although some biological problems admit efficient algorithmical solutions, e.g., the alignment algorithms for comparing the primary structure of biological molecules, the mathematical modeling of many other biological experiments often leads to computationally hard problems. An overview of the

successful applications of algorithmics to biological problems can be found in the textbooks by Böckenhauer and Bongartz [BB03⋆], Gusfield [Gus97], Pevzner [Pev00], or Setubal and Meidanis [SM97].

Mainly, we are interested in these three biological applications:

1. *Haplotyping:* Humans and many other living organisms (including all vertebrates) have a diploid genome, i.e., they have two copies for each chromosome. For an individual organism, these two copies usually differ from each other as a result of their inheritance from two separate parent entities. The most common difference is the so-called single nucleotide polymorphisms (SNPs), where a single nucleotide is replaced by another. For a given chromosome, the complete sequence information of one copy is called a haplotype.

   Knowing both haplotypes of a chromosome can be very helpful, e.g., to detect genetic diseases. However, most of the existing sequencing data does not distinguish between the haplotypes. Instead, it only yields information about a mixture of the haplotypes, the so-called genotype. The task is to reconstruct the two haplotypes of an individual from the genotype. The genotype is usually given as a set of sequenced DNA fragments for which especially the positions of the SNPs are known. To determine the haplotypes, the set of fragments has to be partitioned into two sets corresponding to the two haplotypes.

   This task is readily accomplishable if the given sequencing data is error-free. In this case we just have to check, for any two fragments, if they coincide at their common SNP positions. However, in real life situations, the data is error-prone, turning this task into a challenging problem. The data can be represented by a matrix, where the columns correspond to the SNPs and the rows correspond to the fragments. Each entry of this matrix is either 0 or 1 (corresponding to the two possible values of the SNP), if the SNP occurs in the fragment, or a gap symbol otherwise. There are (at least) three possible types of errors: erroneously identified SNP sites (corresponding columns in the matrix should be removed), false fragments due to contamination (corresponding rows of the matrix should be ignored) and reading errors during the sequencing of the fragments (corresponding matrix entries should be altered). The problems of removing either a minimum set of fragments or a minimum set of SNPs to make the data consistent were first investigated by Lancia et al. [LBILS01]. They showed the NP-hardness of the problems in general and presented polynomial-time algorithms for some special cases. Parameterized algorithms for both problems were presented by Rizzi et al. [RBIL02]. Greenberg et al. suggested a model for dealing with reading errors, namely to remove a minimum number of zeros and ones from the matrix. This problem was further investigated by Panconesi and Sozio [PS04], who showed its hardness and proposed some heuristics for it.

   As for determining the haplotypes within a given (small) population of organisms, an overview can be found in the conference proceedings of a DIMACS/RECOMB satellite workshop on haplotype inference [IWC02] or in the survey paper by Bonizzoni et al. [BDDL03].

2. *Protein Structure Prediction:* The protein folding problem, i.e., the computational prediction of the three-dimensional structure of a protein from its amino acid sequence, is one of the most important and challenging problems in computational biology. This is because the function of the protein relies heavily on its three-dimensional structure. Thus, knowledge of three-dimensional structures of proteins is of great importance, e.g., in pharmaceutics for the design of new drugs.

   Since a complete simulation of the folding process of a protein is far too complex to handle, approximate solutions by using simplified, abstract models are studied. One of the most popular models is the HP model introduced by Dill [Dil85, DBY+95], where the hydrophobic interactions between the amino acids are considered to be the main force driving the folding process, and the folding space is discretized to a two- or three-dimensional grid lattice.

   The problem of protein folding in the HP model was shown to be NP-hard for the two-dimensional [CGPPY98] as well as for the three-dimensional grid lattice [BL98]. Many attempts to attack the problem include approximation algorithms [HI96, New02, NR04], constraint programming algorithms, local search, genetic algorithms, and other heuristics. An overview of the existing literature is given in the survey articles [CDK03, GHL03].

   The HP model has some severe disadvantages. One major drawback of the original HP model is the bipartiteness of the underlying grid, which severely restricts the set of possible foldings. Hence, several refined discrete models, which are more adequate for the protein folding problem, have been investigated. In a first step towards a more realistic model of protein folding that overcomes at least the restrictions of the underlying grid, HP-like models on various extended lattices have been studied [ABD+97, HI97, Heu99, BB04⋆].

3. *Genome Rearrangements:* For closely related species, the DNA sequences of homologous genes are often very similar, only the order of the genes on the chromosomes may vary a lot. The differences in gene order are caused by high-level mutations—genome rearrangements—that cut out pieces of a chromosome and insert them elsewhere or in reversed orientation. The number of such mutations needed to transform the genome of one species into the genome of another species can be used as to measure the relationship between these species, e.g., for constructing a phylogenetic tree.

   Formally, this problem can be modeled as the task of sorting a given permutation (or a given string, in case of multiple occurrences of genes) by using certain operations like reversals (cutting out a block and inserting it in reversed orientation) or transpositions (exchanging the positions of two consecutive blocks).

   Sorting a permutation by reversals has extensively been studied in the literature, both for the case of ordinary unoriented permutations and for oriented permutations, where each element is assigned an orientation indicating the reading direction of the corresponding gene. This modeling dates back to Watterson et al. [WEHM82]. Caprara [Cap97] has shown the NP-hardness of the problem in the case of unoriented permutations, and a series of approximation algorithms for the problem were designed,

starting with a 2-approximation found by Kececioglu and Sankoff [KS95]. The best algorithm currently known, designed by Berman, Hannenhalli, and Karpinski [BHK02], achieves an approximation ratio of 1.375. A polynomial-time algorithm for sorting an oriented permutation by reversals was invented by Hannenhalli and Pevzner [HP95], a simplified proof can be found in the paper by Bergeron [Ber01].

In contrast, the complexity of sorting a permutation by transpositions is still an open problem. This problem was first investigated by Bafna and Pevzner [BP96], who presented a 1.5-approximation algorithm for this problem. A somewhat simpler algorithm achieving the same approximation ratio was described by Hartman [Har03]. A detailed overview of the different models of genome rearrangements mentioning also several other approaches can be found in the paper by Sankoff and El-Mabrouk [SE02].

One drawback of all of the models listed above is that the length of the executed reversals or transpositions is not taken into account, whereas in nature, shorter reversals or transpositions are more likely to occur. To overcome this drawback, length-weighted reversals were considered by Pinter and Skiena [PS02], Bender et al. [BGH+04], and Swidan et al. [SBG+04].

2.2.2 Current State of Our Research

We have actively been pursuing research on the concept of stability of approximation, and publications include [BHK+00⋆] and [BHK+02⋆]. Extensions of this concept for polynomial-time approximation schemes can be found in the book [Hro01⋆]. Our recent work on the TSP includes [BHK+00a⋆, BHK+00b⋆, BS00⋆]. Graph connectivity problems are dealt with in [BBH+02⋆, BBH+03⋆, BBH+04⋆].

Computational biology is an up and coming field. Although our involvement has spanned a few years only, our strong algorithmic background will provide valuable insight and advancement to the field. [BB03⋆] provides a general overview to the approach in computational biology. For the shortest common superstring problem, we refer to [Bon00⋆] and [Bon01⋆]. [Bon04⋆] addresses the protein similarity search under mRNA structure constraints. An investigation into the protein folding problem on a modified grid lattice with diagonals is presented in [BB04⋆].

2.2.3 Detailed Research Plan

We plan to employ the three main techniques described in the summary to graph-theoretical problems with relevance to logistics and operations research and to problems in computational biology.

2.2.3.1 Logistics and Operations Research

Our conceptual idea for a general method development is based on merging our concept of stability of approximation with the concept of parameterized complexity. In parameterized approaches, it becomes permissible to have the time complexity grow fast with respect to a parameter (but still slowly with respect to input size). Our concept of stability of approximation leads to algorithms

of a time complexity independent from input size, but the quality of their output deteriorates at a growing parameter. For problems considered to be very hard even for approximation algorithms, it could be helpful to merge these approaches, i.e., to allow for growth of time complexity and deterioration of output quality (approximation ratio) in terms of the parameter. In this way, we hope to find interesting subclasses of input instances of hard problems. For these, we could then guarantee to compute reasonably good solutions in reasonable time.

We consider to apply this approach mainly to some generalizations of the famous TSP. For problems like the TSP with windows, where some vertices must be visited within preassigned time intervals, there are no known approximation algorithms, not even for the metric version. In operations research, it is common to use some heuristics based on linear programming, but so far, nobody has been able to provide guarantees for the quality of output by means of analysis. This very general problem formulation looks too complex for any current analysis technique in the area of approximation algorithms to be applicable. Our idea is to consider two versions of the problem.

In the first generalization, windows are replaced by some requirements on the order in which some special vertices are to be visited. For this problem, we have begun to apply the concept of approximation stability. This approach looks very promising and may help us succeed for the above-stated generalization of the TSP.

The second generalization is the TSP with deadlines. This is a very realistic problem and actually pretty close to the TSP with windows. Here, for some special nodes, deadlines may be imposed implying that these nodes must have been visited before a given time. Compared to the TSP with windows, it means that we remove the constraint not to visit nodes before the opening time of their windows. It looks promising to take the number of nodes with deadlines as a parameter and to develop parameterized approximation algorithms for the metric TSP with deadlines, where the approximation ratio grows with the parameter, but not with the input size. Another idea is to consider the distance to the metric property as an additional parameter and to investigate the growth of the complexity and the deterioration of output quality for both parameters simultaneously. We aim not only to provide a theoretical analysis that ensures efficiency and solution quality in most cases, but also to present experimental evaluation of our designed algorithms for some real data, e.g., the scheduling of airport services.

We intend to look at further general optimization tasks, for which only heuristical solutions currently exist. We aim to design algorithms with some quality guarantee for such optimization tasks. The reason is not only to develop new algorithms, but also to capture the nature of the hardness of the considered problems.

2.2.3.2 Computational Biology

Molecular biology is an expanding research field and with it, the need for mathematical models for biological processes also grows rapidly. Our main focus is on the three main biological problems mentioned here. However, we will not exclusively restrict our interest to these. We hope to exploit our algorithmic expertise to help solve other biologically relevant problems.

1. *Haplotyping:* In dealing with haplotypes for data from a single individual, our first goal is to further investigate the model for haplotyping introduced by Greenberg et al. and to find some algorithms with performance guarantees for it. The next goal is to find a satisfactory model which simultaneously incorporates all of the error types occurring in haplotyping. Another approach tries to determine the haplotypes within a given (small) population of organisms. This task leads to a variety of interesting algorithmic problems which we also plan to grapple with.

2. *Protein Structure Prediction:* Here, our research emphasis includes the investigation of extensions of the restricted HP model, e.g., by considering the different levels of hydrophobicity and additional forces in the folding process, or the different sizes of the amino acids and other possible spatial constraints in the model of the folding space.

3. *Genome Rearrangements:* We plan to work on the length-weighted model for genome rearrangements. In particular, we aim to extend this model to incorporate transpositions. We also plan to look into refined models that can deal with different (length-weighted) operations simultaneously.

2.2.4 Timetable

Starting on October 1st, 2005, we are going to attack all of the listed problems simultaneously.

In the first year, we shall primarily focus on the development of proof techniques and then, in the two remaining years, on concrete research on the stated problems.

2.2.5 Importance of this Work

Many problems in the field of logistics and operations research can be seen as graph-related problems. For these hard problems, common approaches are based on heuristics, and little is known about the quality of these approaches. Our work will aid to establish a more theoretical analysis of the problems, and to develop algorithms that are less reliant on heuristics.

Computational biology is a field of tremendous importance in understanding and working with nature. Our understanding of the biological processes can be improved using more sophisticated models. Finding algorithms of practical usability that run in feasible time can lead to better interpretation and analysis of the data generated in biological experiments. The resulting models can provide valuable anchor points to guide future research by providing better initial values, boundaries of solutions, approximate and/or exact solutions for several purposes, e.g., in pharmaceutics for the design of new drugs.

References

[AB95] T. Andreae, H.-J. Bandelt: Performance Guarantees for Approximation Algorithms Depending on Parametrized Triangle Inequalities. *SIAM Journal on Discrete Mathematics* 8(1), pp. 1–16.

[ABD+97] R. Agarwala, S. Batzoglou, V. Dančík, S. E. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, and S. Skiena: Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model. *Journal of Computational Biology* 4(2), 1997, pp. 275–296.

[And01] T. Andreae: On the traveling salesman problem restricted to inputs satisfying a relaxed triangle inequality. *Networks* 38(2), 2001, pp. 59–67.

[BP96] V. Bafna and P. Pevzner: Sorting by transpositions. *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'96)*, 1996, pp. 614–623. Full version in *SIAM Journal on Discrete Mathematics* 11, 1998, pp. 224–240.

[BGH+04] M. A. Bender, D. Ge, S. He, H. Hu, R. Y. Pinter, S. Skiena, and F. Swidan: Improved bounds on sorting with length-weighted reversals (extended abstract). *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA'04)*, 2004, pp. 919–928.

[BC00] M. A. Bender and C. Chekuri: Performance guarantees for the TSP with a parameterized triangle inequality. *Information Processing Letters* 73(1-2), 2000, pp. 17–21

[BL98] B. Berger and F. T. Leighton: Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Proc. of the 2nd Annual International Conference on Research in Computational Molecular Biology (RECOMB'98)*, 1998, pp. 30–39.

[Ber01] A. Bergeron: A very elementary presentation of the Hannenhalli-Pevzner theory. *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching (CPM'01)*, Springer LNCS 2089, 2001, pp. 106–117.

[BHK02] P. Berman, S. Hannenhalli, and M. Karpinski: 1.375-approximation algorithm for sorting by reversals. *Proceedings of the 10th Annual European Symposium on Algorithms (ESA'02)*, Springer LNCS 2461, 2002, pp. 200–210.

[BB03⋆] H.-J. Böckenhauer and D. Bongartz: *Algorithmische Grundlagen der Bioinformatik*. Teubner, 2003.

[BB04⋆] H.-J. Böckenhauer and D. Bongartz: Protein folding in the HP model on grid lattices with diagonals. *Proceedings of the 29th International Symposium on Mathematical Foundations of Computer Science (MFCS 2004)*, Springer LNCS 3153, 2004, pp. 227–238.

[BBH+02⋆] H.-J. Böckenhauer, D. Bongartz, J. Hromkovič, R. Klasing, G. Proietti, S. Seibert, W. Unger: On the hardness of constructing minimal 2-connected spanning subgraphs in complete graphs with sharpened triangle inequality. *Proceedings of the 22nd Conference on Foundations of Software Technologz and Theoretical Computer Science (FSTTCS 2002)*, Springer LNCS 2002, pp. 59–70

[BBH+03⋆] H.-J. Böckenhauer, D. Bongartz, J. Hromkovič, R. Klasing, G. Proietti, S. Seibert, W. Unger: On k-edge-connectivity problems with sharpened triangle inequality. *Proceedings of the 5th Italian Conference on Algorithms and Complexity (CIAC 2003)*, Springer LNCS 2653, 2003, pp. 189–200.

[BBH+04⋆] H.-J. Böckenhauer, D. Bongartz, J. Hromkovič, R. Klasing, G. Proietti, S. Seibert, W. Unger: On the hardness of constructing minimal 2-connected spanning subgraphs in complete graphs with sharpened triangle inequality. *Theoretical Computer Science* 326(1–3), 2004, pp. 137–153.

[BDDL03] P. Bonizzoni, G. Della Vedova, R. Dondi, and J. Li: The haplotyping problem: an overview of computational models and solutions. *Journal of Computer Science and Technology* 18(6), 2003, pp. 675–688.

[BHK+00⋆] H.-J. Böckenhauer, J. Hromkovič, R. Klasing, S. Seibert, W. Unger: Towards the notion of stability of approximation for hard optimization tasks and the traveling salesman problem (extended abstract). *Proceedings of the 4th Italian Conference on Algorithms and Complexity (CIAC 2000)*,Springer LNCS 1767, 2000, pp. 72–86

[BHK+00a⋆] H.-J. Böckenhauer, J. Hromkovič, R. Klasing, S. Seibert, W. Unger: Approximation algorithms for the TSP with sharpened triangle inequality. *Information Processing Letters* 75, 2000, pp. 133–138.

[BHK+00b⋆] H.-J. Böckenhauer, J. Hromkovič, R. Klasing, S. Seibert, W. Unger: An improved lower bound on the approximability of metric TSP and approximation algorithms for the TSP with sharpened triangle inequality (extended abstract). *Proceedings of the 17th Symposium on Theoretical Aspects of Computer Science (STACS 2000)*, Springer LNCS 1770, 2000, pp.382–394.

[BHK+02⋆] H.-J. Böckenhauer, J. Hromkovič, R. Klasing, S. Seibert, W. Unger: Towards the notion of stability of approximation for hard optimization tasks and the traveling salesman problem. *Theoretical Computer Science* 285(1), 2002, pp. 3–24.

[Blä03] M. Bläser: An Improved Approximation Algorithm for the Asymmetric TSP with Strengthened Triangle Inequality. *Proceedings of the 30th International Colloquium on Automata, Languages and Programming (ICALP 2003)*, LNCS 2719, pp. 157–163.

[Bon00⋆] D. Bongartz: On the approximation ratio of the group-merge algorithm for the shortest common superstring problem. *Proceedings of the 27th Conference on Current Trends in Theory and Practice of Informatics (SOFSEM 2000)*, Springer LNCS 1963, 2000, pp. 298–307.

[Bon01⋆] D. Bongartz: On the approximation ratio of the group-merge algorithm for the shortest common superstring problem. *Computing and Informatics*, Vol. 20/4, 2001 pp. 325–357.

[Bon04⋆] D. Bongartz: Some notes on the complexity of protein similarity search under mRNA structure constraints. *Proceedings of the 30th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2004)*, Springer LNCS 2932, 2004, pp. 174–183.

[BS00⋆] H.-J. Böckenhauer and S. Seibert: Improved lower bounds on the approximability of the travelling salesman problem. *RAIRO Theoretical Informatics and Applications* 34, 2000, pp. 213–255.

[Cap97] A. Caprara: Sorting by reversals is difficult. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*, 1997, pp. 75–83.

[CDK03] V. Chandra, A. DattaSharma, and V. S. A. Kumar: The algorithmics of folding proteins on lattices. *Discrete Applied Mathematics* 127, 2003, pp. 145–161.

[CGPPY98] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis: On the complexity of protein folding. *Journal of Computational Biology* 5(3), 1998, pp. 423-466. Extended Abstract in *Proc. of the 30th Annual ACM Symposium on the Theory of Computing (STOC'98)*, 1998, pp. 597–603.

[Chr76] N. Christofides: Worst-case analysis of a new heuristic for the traveling salesman problem. Technical Report 388, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, 1976.

[CR02] L. S. Chandran, L. S. Ram: Approximations for ATSP with Parametrized Triangle Inequality. *Proceedings of the 19th Annual Symposium on Theoretical Aspects of Computer Science (STACS 2002)*, LNCS 2285, 2002, pp. 227–237.

[DBY+95] K. A. Dill, S. Bromberg, K. Yue, K. Fiebig, D. Yee, P. Thomas, and H. Chan: Principles of protein folding – a perspective from simple exact models. *Protein Science* 4, 1995, pp. 561–602.

[Dil85] K. A. Dill: Theory for the folding and stability of globular proteins. *Biochemistry* 24, 1985, p. 1501.

[EK01] L. Engebretsen and M. Karpinski: Approximation Hardness of TSP with Bounded Metrics *Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP 2001)*, LNCS 2076, pp. 201–212.

[FP03] A. Ferrante, M. Parente: An Approximation Algorithm for the Minimum Weight Vertex-Connectivity Problem in Complete Graphs with Sharpened Triangle Inequality. *Proceedings of the 8th Italian Conference on Theoretical Computer Science (ICTCS 2003)*, LCNS 2841, 2003, pp. 137–149.

[GHL03] H.J. Greenberg, W.E. Hart, and G. Lancia: Opportunities for combinatorial optimization in computational biology. *INFORMS Journal of Computing* 16(3), 2003, pp. 211–231.

[Gus97] D. Gusfield: *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.

[Har03] T. Hartman: A simpler 1.5-approximation algorithm for sorting by transpositions. *Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching (CPM'03)*, 2003, Springer LNCS 2676, pp. 156–169.

[Heu99] V. Heun: Approximate protein folding in the HP side chain model on extended cubic lattices. *Discrete Applied Mathematics* 127(1), 2003, pp. 163-177. Extended abstract in *Proc. of the 7th Annual European Symposium on Algorithms (ESA'99)*, Springer LNCS 1643, 1999, pp. 212–223.

[HI96] W. E. Hart and S. Istrail: Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal. *Journal of Computational Biology* 3(1), 1996, pp. 53–96.

[HI97] W. E. Hart and S. Istrail: Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86% of optimal. *Journal of Computational Biology* 4(3), 1997, pp. 241–259. Extended Abstract in *Proc. of the 1st Annual International Conference on Research in Computational Molecular Biology (RECOMB'97)*, 1997, pp. 137–146.

[HP95] S. Hannenhalli and P. Pevzner: Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the 27th ACM Symposium on the Theory of Computing (STOC'95)*, 1995, pp. 178–189. Full version in *Journal of the ACM* 46(1), 1999, pp. 1–27.

[Hro01⋆] J. Hromkovič: *Algorithmics for Hard Problems*. Springer 2001 (2nd edition, 2003).

[IWC02] S. Istrail, M. Waterman, and A. Clark (Eds.): *Computational Methods for SNPs and Haplotype Inference*, Springer LNBI 2983, 2002.

[KS95] J. Kececioglu and D. Sankoff: Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica* 13, 1995, pp. 180–210.

[LBILS01] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz: SNPs problems, complexity, and algorithms. *Proceedings of the 9th European Symposium on Algorithms (ESA 2001)*, Springer LNCS 2161, 2001, pp. 182–193.

[New02] A. Newman: A new algorithm for protein folding in the HP model. *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'02)*, 2002, pp. 876–884.

[NR04] A. Newman and M. Ruhl: Combinatorial Problems on Strings with Applications to Protein Folding. *Proc. of the 6th Latin American Symposium on Theoretical Informatics (LATIN'04)*, Springer LNCS 2976, 2004, pp. 369–378.

[Pev00] P. A. Pevzner: *Computational Molecular Biology – An Algorithmic Approach*. MIT Press, 2000.

[PS02] R. Y. Pinter and S. Skiena: Genomic sorting with length-weighted reversals. *Proceedings of the 13th International Conference on Genome Informatics (GIW'02)*, pp. 173–182. Universal Academic Press, 2002.

[PS04] A. Panconesi and M. Sozio: Fast Hare: a fast heuristic for single individual SNP haplotype reconstruction. *Proceedings of the 4th Workshop on Algorithms in Bioinformatics (WABI 2004)*, Springer LNCS 3240, 2004, pp. 266–277.

[RBIL02] R. Rizzi, V. Bafna, S. Istrail, and G. Lancia: Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI 2002)*, Springer LNCS 2452, 2002, pp. 29–43.

[SBG+04] F. Swidan, M. A. Bender, D. Ge, S. He, H. Hu, and R. Y. Pinter: *Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM'04)*, 2004, Springer LNCS 3109, pp. 32–46.

[SE02] D. Sankoff and N. El-Mabrouk: Genome Rearrangement. In: T. Jiang, Y. Xu und M. Q. Zhang (Eds.): *Current Topics in Computational Molecular Biology*. MIT Press, 2002.

[SM97] J. Setubal and J. Meidanis: *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[WEHM82] G. A. Watterson, W. J. Ewens, T. E. Hall und A. Morgan: The chromosome inversion problem. *Journal of Theoretical Biology* 99, 1982, pp. 1–7.