CO 598A: Computational Biology
Bibliography

R. Ravi [*] Andrew Yao [†]

Spring 1995

Abstract

This bibliography contains references to all the articles that were used as material for a graduate course at Princeton. The progression of the material tries to roughly follow the various steps in assembly and analysis of genetic information from small to big pieces. After a brief introduction to the various basic biological concepts, we described an alternative sequencing strategy (SBH). Then we turned to how how sequenced pieces are assembled into fragments, and subsequently these fragments into clones. Then we studied problems arising in the the process of putting these clones into a physical map using probes. At this point, we took a detour to study algorithms to search and analyze this information efficiently, by looking at methods for efficient comparison of two or more molecular sequences and their statistical significance. Then, we illustrated two applications of sequence data in Biology: building evolutionary trees and also investigating rearrangements of whole genomes. During the course of the semester, we also sampled the work of several guest speakers on physical mapping techniques, prediction of the tertiary structure of proteins from their sequence, prediction of genes from raw sequence data, perfect phylogeny, reconstructing evolutionary trees from distance data, and viral shell assembly models.

1 Surveys

The following is a nonexhaustive list of survey articles in the literature relating to several topics we covered in the course.

P. A. Pevzner and M. S. Waterman, "Open Combinatorial problems in computational molecular biology," in *Proc. Third Israel Symp. on Theory of Computing and Systems*, (1995). This contains a good survey of open problems in genome rearrangements, DNA physical

[*]Address: Department of Computer Science, Princeton DIMACS postdoctoral fellowship for the special year on Mathematical Support for Molecular Biology. Email: ravi@cs.princeton.edu.

[†]Address: Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08544-2087. Email: yao@cs.princeton.edu.

mapping, DNA sequencing and sequence comparison.

R. M. Karp, "Mapping the genome: some combinatorial problems arising in molecular biology," in *Proc. STOC 93*, pp. 278-285. Survey of physical mapping in an invited talk.

M. S. Waterman, "General methods of sequence comparison," *Bulletin of Mathematical Biology* V. 46, No. 4, 473-500 (1984).

Algorithmica (1995), Vol. 13, Nos. 1/2. This is a special issue on Computational Biology edited by Eugene W. Myers. Contains several articles covered in the course.

S. C. Chan, A. K. C. Wong, and D. K. Y. Chiu, "A survey of multiple sequence comparison methods," *Bull. Math. Biol.* V. 54, No. 4, 563-598, (1992). Survey of multiple alignment methods until 1992.

A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," *J. Mol. Biol.* V. 235, 1501-1531 (1994). Survey of HMM applications, fairly technical.

A neat web page containing archives and links for biological software and databases around the world is in the home page "http://www.gdb.org/Dan/softsearch/biol-links.html"

2 Introduction to Molecular Biology protocols

Here are four references on basic molecular biology protocols.

The Human Genome Program Primer on Molecular Genetics, June 1992. Maintained at the Oakridge National Lab. This document can be accessed in the World Wide Web at the site "http://www.ornl.gov/TechResources/Human_Genome/home.html" via Mosaic. Serves as a basic introduction and reference to all that we covered in the first two classes.

The Cartoon Guide to Genetics, Larry Gonick and Mark Wheelis, Barnes and Nobles, New York, (1983). A very enjoyable and informative introduction to Molecular Biology. Highly recommended.

Introduction to Genetic Engineering, William Sofer, Butterworth-Heinemann, Stoneham, MA (1991). A crisp and logical introduction to the subject. Fairly easy to read and follow.

The Nov. 20, 1992 issue of *Los Alamos Science.* This is a special issue on the Human Genome Project. It contains informative chapters on several aspects of the project including many of the topics we covered in class, as well as transcription of panel discussions between experts in the field. A valuable reference. Worth browsing a couple of times for topics that specially interest you.

The relevant articles on SBH are the following.
P. A. Pevzner, "l-tuple DNA sequencing: Computer Analysis," *Journal of Biomolecular Structure and Dynamics* V. 7, No. 1, 63-73 (1989). This paper contains the first reformulation of the reconstruction problem for SBH as an Eulerian rather than Hamiltonian problem and was responsible for renewed interest in the technique.

P. A. Pevzner and R. J. Lipshutz, "Towards DNA Sequencing by Hybridization," *Proc. MFCS 1994*, Sprincer-Verlag LNCS 841, 143-158. Contains all definitions relevant to SBH chips and reviews attempts to fabricate such chips.

Material from the above survey is summarized and further open problems are mentioned in the Pevzner-Waterman survey paper below. These problems relate to reconstruction algorithms for complex SBH chips, handling errors and also designing chips with good resolving power.
P. A. Pevzner and M. S. Waterman, "Open Combinatorial problems in computational molecular biology," in *Proc. Third Israel Symp. on Theory of Computing and Systems*, (1995). This contains a good survey of open problems in genome rearrangements, DNA physical mapping, DNA sequencing and sequence comparison.

Pavel Pevzner, "DNA physical mapping and alternating Eulerian cycles in colored graphs," *Algorithmica* (1995), Vol. 13, Nos. 1/2. This article addresses a simple transformation that can convert any Euler tour in an Eulerian graph to any other. Plus interpretations of this transformation in terms of traditional conjectures.

A more algorithmic work motivated by reconstruction problems from SBH is the following.
S. S. Skiena and G. Sundaram, "Reconstructing strings from substrings," *Proc. 1993 WADS*, appears as Springer-Verlag LNCS 709, 565-576.

For the truly motivated, here are two leads on two issues that are open in SBH. The first is the design of better SBH chips, and the second is on new techniques to resolve ambiguities in the reconstruction.
P. A. Pevzner et al., "Improved chips for SBH," *J. Biomolecular Structure and Dynamics*, Vol. 9, No. 2 (1991), 399-410.
A. B. Chetverin and F. R. Kramer, "Sequencing of pools of nucleic acids on oligonucleotide arrays," *Biosystems*, 30 (1993), 215-231.

3 Fragment assembly

Fragment assembly problem is addressed in clear and complete detail in the following article.
J. D. Kececioglu and E. W. Myers, "Combinatorial algorithms for DNA sequence assembly," TR 92-37, U. Arizona, Tucson. Appears in *Algorithmica* Vol. 13, Nos. 1/2 (special issue on computational molecular biology edited by Gene Myers), 7-51.

A theoretical problem that has been motivated by the Fragment Assembly problem is that of finding a shortest common superstring of a set of strings. The first constant-factor approximation algorithm appeared in the following paper.
A. Blum, T. Jiang, M. Li, J. Tromp, and M. Yannakakis, "Linear approximation of shortest superstrings," *Proc. 23rd STOC*, 328-336 (1991).
There have been several papers improving the performance ratio in this paper and the current best performance ratio for the SCS problem stands at 2.75. For the list of papers contributing to this line of work, see the references in the Pevzner-Waterman survey.

Tackling the fragment assembly problem that arises in practice is addressed in clear and complete detail in the following article.
J. D. Kececioglu and E. W. Myers, "Combinatorial algorithms for DNA sequence assembly," in *Algorithmica* Vol. 13, Nos. 1/2 (special issue on computational molecular biology edited by Gene Myers), 7-51.

4 Physical mapping

Physical mapping is at the next level of coarseness in biological maps. A probabilistic model for the problem is proposed and analyzed in the following paper. This work has become very popular and is often used as a thumb-rule by practitioners for estimating parameters in mapping experiments.
E. S. Lander and M. S. Waterman, "Genomic mapping of fingerprinting random clones: a mathematical analysis," *Genomics* 2, 231-239 (1988) .

The error-free formulation of the probe-ordering problem is testing for consecutive ones property of a matrix a problem that has well known solutions (using PQ-trees). Work by Karp's group in Berkeley on the physical mapping problem is in the following two papers.

F. Alizadeh, R. M. Karp, L. A. Newberg and D. K. Weisser, "Physical mapping of chromosomes: A combinatorial problem in molecular biology," *Proc. 4th SODA*, 371-381 (1993).

F. Alizadeh, R. M. Karp, D. K. Weisser and G. Zweig, "Physical mapping of chromosomes using unique probes," *Proc. 5th SODA*, 489-500 (1994).

The following article gives a precise characterization of the minimum number of probes required to construct an unambiguous physical map under the Lander-Waterman probabilistic model, and thus explains a probabilistic threshold phenomenon reported in the empirical results in the Alizadeh et al. paper on nonunique probes.
M. Dyer, A. Frieze, and S. Suen, "Ordering clone libraries in computational biology," preprint (November 1993).

Using Interval Graphs and Parameterized Complexity in Physical Mapping.

Characterizations and Algorithms for Chordal Graphs and Interval Graphs are covered in:
M.C. Golumbic. *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, New York, 1980.

Interval graph sandwich problems are discussed in:
M.C. Golumbic, Haim Kaplan and Ron Shamir, *On the complexity of DNA physical mapping*, Advances in Applied Mathematics 15,251-261 (1994).
You could find there a biological discussion motivating the problems. NPC proofs and open problems.

Interval/proper interval graph deletion problems were proven to be NPC in:

Paul Goldberg, M.C. Golumbic, Haim Kaplan, Ron Shamir, *"Four strikes against Physical Mapping of DNA"*. to appear in Journal of Computational Biology.

This reference also contains NPC proof for the following problem: Given a zero one matrix can you permute rows and columns an obtain the consecutive ones property with at most two consecutive blocks per column.

A related paper in which the chordal completion problem was proved to be NPC:

M. Yannakakis. *Computing the minimum fill-in is NP-complete.* SIAM J. Alg. Disc. Meth, 2, 1981.

The general framework for proving parametrized intractability was developed recently by Rod Downey and Mike Fellows. They wrote a series of papers the first of which (which covers everything necessary to use the theory) is the following:
R.G. Downey and M.R. Fellows, *Fixed-Parameter Tractability and Completeness I: Basic Results*, 1991.

Another interesting paper written by the same authors goes in the other direction and try to develop some techniques for proving parameterized tractability, i.e, designing algorithms for the fixed parameter version of a problem.

R.G. Downey and M.R. Fellows, *Parameterized Computational Feasibility*, in Feasible Mathematics II, (ed P. Clote and J. Remmel) Birkhauser, Boston, to appear.

Parameterized complexity analysis of problems in biology has been done in the following papers:

F.R. McMorris, T. Warnow, and T. Wimer. *Triangulating vertex-colored graphs*, SIAM J. Disc.. Math. 5:249-258, 1992.

H.L. Bodlaender, M.R. Fellows, and M. Hallet. *Beyond NP-completeness for problems of bounded width: Hardness for the W hierarchy*, in STOC 1994 (26) p. 449-458.

H. Kaplan, R. Shamir and R. E. Tarjan, *Tractability of parameterized completion problems on chordal and interval graphs: Minimum Fill-in and Physical Mapping*, Proceedings of the 35th annual symposium on Foundations of Computer Science (FOCS), November 1994, IEEE Computer Science Press, 780–791.

H.L. Bodlaender and B.L.E. de Fluiter, *Intervalizing k-Colored Graphs*, to appear in ICALP 95.

H.L. Bodlaender, R.G. Downey, M.R. Fellows, M.T. Hallett and H.T. Wareham, *Parameterized Complexity Analysis in Computational Biology*, IEEE Computer Society Workshop on Shape and Pattern Matching in Computational Biology, June 20, 1994.

5 Sequence Comparison

Alignments are of two types: global where two whose sequences are compared in their entirety and local where substrings of the two that have the maximum similarity are sought. Both problems can be solved using dynamic programming (DP). The first algorithm for global alignment is in the following article.

S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.* 48, 443-453 (1970).

The algorithm implied by this article is cubic rather than the direct quadratic method

that one obtains by DP. The first clear exposition of the DP method is attributed to the following article.
T. F. Smith and M. S. Waterman, "The identification of common molecular subsequences," *J. Mol. Biol.* 147, 195-197 (1981).

Equivalence of similarity versus distance between two sequences is demonstrated in
T. F. Smith and M. S. Waterman, "Comparison of biosequences," *Adv. in App. Math.* 2, 482-489 (1981).

A trick used often to cut down the space of the dynamic programming to linear from quadratic is in the following article.
Hirschberg *CACM 18*, 341-343 (1975).

The following is a nice survey article on the history and development of sequence comparison methods until 1984.
M. S. Waterman, "General methods of sequence comparison," *Bulletin of Mathematical Biology* V. 46, No. 4, 473-500 (1984).

Parametric alignment refers to computing the alignment between two sequences parametrized by the costs for substitutions, indels, and gaps, and doing this in an efficient way. The following article has some nice results in this direction.
D. Gusfield, K. Balasubramanian, and D. Naor, "Parametric optimization for sequence alignment," *Proceedings of the third ACM-SIAM Symp. on Disc. Algorithms* (1992), 432-439.

We dealt with two important types of alignments in class: global and local. The Gusfield et al. paper on parametric alignments gave the set of all possible optimal alignments under all the possible choices for mismatch and indel penalties. This gives some understanding of the set of all possible relevant global alignments.

Most biological database searches however must use local alignments to pick out small regions of homogeneity, so the choice of the scoring matrix and gap penalties used in such searches and the relevance of the resulting scores are important issues in performing meaningful searches. Dayhoff and co-workers (1978) came up with the popular PAM matrices that are widely used in protein sequence comparisons these days. These matrices were essentially log-likelihood matrices as we described in class.

A newly-discovered rationale behind the relevance of these log-likelihood scores is presented in the following paper.
S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proc. Natl. Acad. Sci. USA* 87, 2264-2268, (1990).

The following paper has a good interpretation of scoring matrices used in searches, and an interpretation of the scores obtained in any search in terms of information content measured in bits.
S. F. Altschul, "Amino acid substitution Matrices from an information theoretic perspective," *J. Mol. Biol.* 219, 555-565 (1991).

On the other hand, for global alignments, very little is known in terms of statistics, the most notable result in this area being the analysis of the expected length of the longest common subsequence (LCS) of two random sequences.
C. Chvatal and D. Sankoff, "Longest common subsequences of two random sequences," *J. Appl. Prob.* 12, 306-315, (1975).

Statistics of local versus global alignments show very different behaviors. Roughly, local alignments may be thought of as finding alignments allowing no gaps or penalizing very heavily for gaps and mismatches. In this case, the size of the longest local alignment between two random sequences grows roughly logarithmically with the lengths of the two sequences compared. On the other hand, global alignments are those you get with very small values for the gap penalties, since you are looking to find similarity between the two whole sequences. In this case, the expected match length (or similarity score) grows roughly linearly (like longest common subsequences of Chvatal-Sankoff) with the product of the lengths of the compared sequences. A phase transition can thus be noticed when the statistics are observed with the gap costs varying (formally, the gap initiation adn gap extension costs). This has been studied recently in the following paper.
R. A. Arratia and M. S. Waterman, "A phase transition for the score in matching random sequences allowing deletions," *Annals of Applied Probability* 4, 200-225 (1994).

An intuitive explanation of this phase transition and its relevance to choosing scores for biological homology searches is in the following paper.
M. Vingron and M. S. Waterman, "Sequence alignment and penalty choices," *J. Mol. Biol*, (1994) 235, 1-12.

Along the practical side, there are two widely used database searching packages: BLAST and FASTA. BLAST is described in the article below.
S. A. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *J. Mol. Biol.* 215, 403-410 (1990).

To learn more about BLAST, try sending email to the BLAST email server "blast@ncbi.nlm.nih.gov" with the single line HELP in the body of the message. An extended description of the statistical significance parameters reported by BLAST is in the following article.
S. F. Altschul, M. S. Boguski, W. Gish and J. C. Wootton, "Issues in searching molecular sequence databases," *Nature Genetics* 6, 119-129, (1994).

FASTA is first described in the following article.
D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity searches," *Science* 227, 1435-1440 (1985).
A more recent follow-up is appears in the paper below.
W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proc. Natl. Acad. Sci. USA* 85, 2444-2448, (1988).

6 Multiple Alignments

Multiple sequence alignments are used in many ways: most notably, in phylogenetic comparison, to identify motifs and common profiles over many related sequences and sometime in error correction in fragment assembly. An exact algorithm for the multiple alignment problem using dynamic programming is in the following article.
H. Carrillo and D. Lipman, "The multiple sequence alignment problem in biology," *SIAM J. App. Math.* 48, 1073-1082, (1988).

A folklore result on building a multiple alignment that is compatible with a tree spanning the given sequences is attributed to the article below.
[FD] D. Feng and R. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *J. Molec. Evol.* 25, 351-360 (1987).

The first approximation algorithm for multiple sequence alignment under the SP or sum-of-pairs objective is due to Gusfield.
D. Gusfield, "Efficient methods for multiple sequence alignment with guaranteed error bounds," *Bull. Math. Biol.* 55, 141-154, (1993).

An improved heuristic is presented in the paper below.
V. Bafna, E. L. Lawler, and P. A. Pevzner, "Approximation algorithms for multiple sequence alignment," in *Proc. Comb. Patt. Matching Conf.*, LNCS 807, 43-53, (1994).

7 Phylogenetic Reconstruction

The intent in this area is to recreate the branching pattern of species evolution from data based on distinguishing characteristics of the species. There are two broad subareas, depending on whether the data is given as characters of different species, or as distances between species. The two methods are dubbed "character-based" and "distance-based" respectively.

Richa Agarwala lectured about a very significant positive result she was instrumental in deriving. Her work appears in the following article.

R. Agarwala and D. Fernández-Baca, "A polynomial-time algorithms for the perfect phylogeny problem when the number of character states is fixed," *Proc. 34th FOCS*, 140-147 (1993). Journal version in SIAM J. on Computing, Vol. 23, No. 6, 1216-1224 (1994).
She also sketched an algorithm that is drawn from another reference in this paper for the case when the recorded characters are binary.

Turning to distance-based methods, there are two notions of weighted trees used in representing evolutionary history. An unrooted tree where the weights represent divergence measures in evolution is called an *additive tree* while a rooted tree with the sum of weights from the root to all the leaves being identical is called an *ultrametric tree*. Reconstructing a matrix of distances into an additive tree was solved by many different authors at different times. An old method appears in the following article.
M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer, "Additive evolutionary trees," *J. Theor. Biol.* 64, 199-213, (1977).

The following paper solves several reconstruction and best-fit problems for ultrametric trees and derives some hardness results as well.
M. Farach, S. Kannan and T. Warnow, "A robust model for finding optimal evolutionary trees," *Proc. 25th STOC*, 137-145 (1993).

Using phylogenetic trees to guide multiple alignments is a popular theme, and this gives rise to several problems that are addressed in the following book.
D. Sankoff and J. Kruskal (eds.) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Addison-Wesley, reading, Mass., 1983.

The following papers address the problem of inferring ancestral sequences in a phylogeny relating aligned sequences.
[F] W. M. Fitch, "Towards defining the course of evolution: minimum change for a specific tree topology," *Syst. Zool.* 20, 406-416 (1971).
[H] J. A. Hartigan, "Minimum mutation fits to a given tree," *Biometrics*, 29, 53-65 (1973).

An approximation algorithm for reconstructing ancestral sequences in a given phylogenetic tree so as to minimize the weight of the tree is the subject of the following paper.
[JLW] T. Jiang, E. L. Lawler, and L. Wang, "Aligning sequences via an evolutionary tree: complexity and approximation," *Proc. 26th STOC*, 760-769.

The taxonomy below is due to Gene Lawler for classifying Phylogenetic inference problems depending on what is available (+) and what is not (-).

Mult. Algmt.	Tree	Anc. seq.	Problem name
-	-	-	Steiner Tree problem in edit space [JLW]
-	-	+	Spanning Tree problem in edit space
-	+	-	Tree Alignment [JLW]
-	+	+	Phylogeny Induced Alignment [FD]
+	-	-	Steiner Tree problem in Hamming space /Perfect Phylogeny
+	-	+	Spanning Tree problem in Hamming space
+	+	-	Ancestral sequences from aligned phylogeny [F,H]
+	+	+	null problemo!

8 Genome Rearrangements

Genome rearrangements refer to large-scale chromosomal edit operations that alter the orientation and position of genes in chromosomes. The most common and important ones are inversions, transpositions and translocations. Distances between genomes under inversions is the subject of the following paper.
Kececioglu, John and David Sankoff."Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement," *Algorithmica*, Vol. 13, Nos. 1/2, 180-210 (1994). (An earlier version appeared as "Exact and approximation algorithms for the inversion distance between two chromosomes," Proceedings of the 4th *Symposium on Combinatorial Pattern Matching*, Springer-Verlag Lecture Notes in Computer Science 684, 87–105, June 1993.)

The performance ratio of the approximation algorithm in the above article is improved in the following paper. This paper also introduced for the first time, an analysis of the distance using the alternating cycle structure in a breakpoint graph constructed from the genomes.
Bafna, Vineet and Pavel A. Pevzner, "Genome rearrangements and sorting by reversals," Proceedings of the 34th *Symposium on Foundations of Computer Science*, 148–157, November 1993.

Transpositions and translocations were first studied in the following papers.
Bafna, Vineet and Pavel A. Pevzner, "Sorting by transpositions," Proceedings of the 6th ACM-SIAM *Symposium on Discrete Algorithms*, 614- 623, (1995).
J. D. Kececioglu, and R. Ravi, "Of mice and men: Algorithms for evolutionary distances between genomes with translocations," Proceedings of the 6th ACM-SIAM *Symposium on Discrete Algorithms*, 604-613, (1995).

Divergence between man and mouse is estimated in the following biological treatise.
Nadeau, Joseph H. and Benjamin A. Taylor."Lengths of chromosomal segments conserved

since divergence of man and mouse," *Proceedings of the National Academy of Sciences USA* 81, 814–818, (1984).

There have been some recent exciting developments in the area with recent work on determining the exact inversion distance between two permutations when the data is signed, i.e., unique orientations can be assigned to all the markers. This work is reported first in the following paper, and the application of the techniques to signed translocations is discussed in the second article below.
S. Hannenhalli and P. Pevzner, "Transforming cabbage into turnip (Polynomial algorithm for sorting signed permutations by reversals)," To appear in proceedings of the *Symposium on the Theory of Computing*, May 1995. Available as a technical report CSE-95-004, from the Department of Computer Science and Engineering, Pennsylvania State University.
S. Hannenhalli, "Polynomial algorithm for computing translocation distance between genomes," To appear in proceedings of the *Combinatorial Pattern Matching Conference*, July 1995. Available as a technical report CSE-95-005, from the Department of Computer Science and Engineering, Pennsylvania State University.

9 Miscellaneous topics

The gene prediction problem is that of identifying possible gene boundaries from raw sequence data. We heard a talk based on the following paper in class.
Steven Salzberg, "Locating protein coding regions in human DNA using a decision tree algorithm," J. Computational Biology, to appear, Fall 1995.

The talk by Mona Singh in class developed on earlier work on protein structure prediction methods that appeared in the following article
Bonnie Berger and David B. Wilson, "Improved Algorithms for Protein Motif Recognition," *Proc. Symp. on Disc. Algorithms*, 58-67, (1995).

Acknowledgements

Special thanks to the guest speakers in the course: Faridh Alizadeh, RUTCOR, Mona Singh, MIT, Steven Salzberg, Johns Hopkins University, Richa Agarwala, DIMACS, Martin Farach, Rutgers University, Vineet Bafna, DIMACS, and Bonnie Berger, MIT.