IDENTIFYING FEATURES IN BIOLOGICAL SEQUENCES:

FIFTH WORKSHOP REPORT

(ASPEN CENTER FOR PHYSICS, MAY 30 - JUNE 19, 1994)

Eugene Myers[1,2] , Christian Burks[3] , and Gary Stormo[4]

TR 94-36
Dept. of Computer Science
University of Arizona
Tucson, AZ 85721
December 19, 1994

Abstract: This report is for the fifth of an annual series of workshops held at the Aspen Center for Physics concentrating on the identification of features in DNA sequence, and more broadly on related topics in computational molecular biology. Over the last five years the workshop has been cited as supporting and/or inspiring over 40 papers, has provided training and inspiration to Ph.D students and post-doctoral fellows early in their careers, and provided senior scientists the unique opportunity to seriously engage in interdisciplinary collaborations over the 3 weeks of the workshop.

---

[1]To whom correspondence should be addressed:
telephone, 602-621-6612
fax, 602-621-4246
email: gene@cs.arziona.edu

[2]Dep't of Computer Science
University of Arizona
Tucson, AZ 85721

[3]Theoretical Biology and Biophysics Group
Los Alamos National Laboratory
Los Alamos, NM 87545

[4]Dep't of Molecular, Cellular, and Developmental Biology
University of Colorado
Boulder, CO 80309

Summary

The Aspen Center for Physics (ACP), in Aspen, Colorado, sponsored a three-week workshop from May 30 to June 19, 1994, with 23 scientists participating, 13 for their first time. The workshop, entitled *Identifying Features in Biological Sequences* was the fifth (IF-V) in a yearly series hosted by ACP on this topic. The previous workshops (RG-I[1], RG-II[2], RG-III[3], and IF-IV) occurred in years 1990 through 1993.

The workshop focused on discussion of current needs and future strategies for developing the ability to identify and predict the presence of complex functional units on sequenced, but otherwise uncharacterized, genomic DNA. We addressed the need for computationally-based, automatic tools for synthesizing available data about individual consensus sequences and local compositional patterns into the composite objects (e.g., genes) that are -- as composite entities -- the true object of interest when scanning DNA sequences. The general background and justification for a workshop on this topic was discussed earlier in the report on RG–I[1]. Of particular interest over the past year has been the maturation of previously described[5,6,9,14] as well as the emergence of several new[4,7,24,29,45–49] approaches to predicting the presence and location of genes (or at least protein-coding exons).

The workshop was structured to promote sustained informal contact and exchange of expertise between molecular biologists, computer scientists, and mathematicians. No participant stayed for less than one week, and most attended for two or three weeks. Computers, software, and databases were available for use as "electronic blackboards" and as the basis for collaborative exploration of ideas being discussed and developed at the workshop.

There have been no recent (or even not-so-recent) meetings devoted to precisely the topic that provided the theme of our workshop. Though there have been a number of workshops over the years devoted to DNA sequence analysis, none have focused on the recognition and characterization of composite objects (such a genes) exclusively; for this reason, RG-I and RG-II have provided a unique approach to addressing a very important challenge in making use of the data coming out of large-scale sequencing projects.

There are very few meetings that:

- last for several weeks
- with scientists from disparate disciplines
- and a meeting structure promoting informal interaction and primary emphasis on pairwise (or small group) intensive exchange of ideas and insights.

As with previous years, this workshop provided an unusual and very facilitating, environment and time for scientists to address the necessary and considerable learning curves associated with unfamiliar disciplines. Several long-term formal and informal collaborative interdisciplinary projects resulted from this workshop that would not otherwise have developed, partly because of the mixture of scientists from different disciplines and partly because of the sustained interaction that the workshop afforded.

The educational value of the workshop is worth noting. Exchange of information, and particularly description of problems to those most able to provide the appropriate tools were a central theme of this workshop. Furthermore, since several of the participants were post-doctoral fellows or graduate students, the workshop provided interdisciplinary training that will presumably be very useful to them, even -- in some cases -- influencing them in deciding to pursue one of the interdisciplinary research paths.

Administrative Details

Dates of workshop. The workshop extended from May 30 to June 19, 1994.

Location of the workshop. The ACP hosted the workshop, and provided facilities and administrative support staff, including an Administrative Vice-President and four full-time secretaries on site for responding to housing, word-processing, and secretarial requests. ACP provided offices for use by participants, with two scientists per office. Several lecture/conference rooms with projection capabilities and seating capacity for a group of our size were available (and used), including an outside patio conference room for informal seminars and discussions. Condominium style residences were provided (through the ACP) that allow for routine meal preparation, informal gatherings, and family/guests. A more detailed description of ACP and its support for workshops such as ours was given previously[1].

Organizing Committee. E. Myers, C. Burks, and G. Stormo constituted the Organizing Committee, with E. Myers serving as Chairperson, for the workshop's technical agenda. T. Appelquist, as President of the Aspen Center for Physics, was the formal contact for administrative aspects of the workshop.

Funding. Funding was provided through the standing NSF grant (NSF-86-06266) to ACP for its summer program, and from NSF grant BIR-9406201 specifically awarded to this years workshop. These funds were used to partially offset the participants' costs in attending the workshop and provide general administrative support for the workshop, consistent with the general approach used in the ACP summer program.

Call for Participation and Selection. ACP did their traditional wide-spread mailing and made publicly-advertised announcements regarding their summer program (and this workshop in particular, which was on the publicized agenda for the summer program). To augment these announcements, the Organizing Committee made a direct mailing to over two hundred scientists active in molecular biology, computer science, or mathematics (or interdisciplinary research among these three disciplines), encouraging them either to apply or to pass information about the workshop along to other potential participants. Selection of participants was made as described in the RG-I report[1].

Workshop organization. Two formal talks were scheduled each day; the remainder of the time was devoted to small discussion groups, initiation of collaborations, research, and informal presentations.

Computational facilities. We set up a small, temporary workstation network for use by our workshop, including three Sun workstations (1 LX, 1 Classis, and 1 Sparc 1) and three additional Sun X-terminals. All but one Sun workstation were loaned to us for the duration of the workshop by Sun Microsystems Denver office to whom we are very grateful. The Sparc 1 and a post-script capable laser writer were loaned to us by Los Alamos National Laboratory along with an additional gigabyte of disk to accompany a gigabyte loaned by Sun. System administration of this network was provided by M. Engle, a workshop participant and a systems programmer in the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory. The workstations and X-terms were configured in a network that included a direct "slip-link" to the Internet (with the required hardware, installation, and systems consulting very purchased from the Colorado SuperNet), and modems for dialing remote computers and allowing remote computers to dial in.

Software and databases that were available included:

- standard UNIX utilities (file editors and management, etc.);
- the TROFF electronic typesetting suite;
- copies of the GenBank, PIR, SwissProt, and Entrez databases;
- a copy of D. Higgins' CLUSTAL for multiple sequence alignments;
- a copy of W. Pearson's FASTA for scanning databases;
- a copy of S. Henikoff's BLOCKS software;
- a copy of X. Huang's CAP dna sequencing software;

- a copy of C. Burks' GENFRAG dna sequencing software;
- a copy of S. Smith's GDE software suite;
- several windowing systems, including OpenWindows 2.0, X-11 R5, and SunView;
- the xfig drawing program;
- and compilers, including FORTRAN, Pascal, and C.

The computational resources served as an advanced "electronic blackboard" where new ideas and data sets could be laid out and modified more rapidly than would otherwise be possible. As such, it was a relatively unique aspect of this workshop compared to others (though several "genome" workshops now routinely provide workstations for demonstrating software already developed, they are rarely used as a focus of initiating and carrying through on new ideas during the meeting). In addition, the facility allowed participants to communicate via email with their home institution and other colleagues, an important consideration when one is leaving their workplace for up to three weeks.

Participants

The majority of applicants made a substantial commitment of time, with the minimum stay being one week and many attending for the full three weeks. The following is a list of those who participated in the workshop (unfortunately, there were more applicants than there were slots available). This group represented an excellent cross-section of the disciplines and expertises the workshop drew on, and included four graduate students and four post-doctoral fellows. One-half of the participants in IF-V had not attended a previous Aspen workshop.

Name	Institution	Formal Background	Research Focus
Tim Bailey	UCSD	Computer Science	pattern-learning,
George Bell	LANL	Physics, Biophysics	biocomputing overview, intron structure, experimental biology
Christian Burks	LANL	Databases, Software, Biology	pattern matching, data structures, high speed searches
Sean Eddy	MRC-LMB	Biology, Software	sequence alignment, HMM models
Michael Engle	LANL	Computer Science	software engineering, database search algorithms
Rob Farber	LANL	Computer Science	neural nets, protein folding
D'vorah Graeser	Wash U.	Biology	hybridization technology, physical mapping
George Hartzell	Berkeley	Computer Science, Biology, Software	dna sequencing, physical mapping
Jim Holloway	Oregon St.	Biology, Computer Science	sequence comparison, genetic mapping

Paul Horton	Berkeley	Computer Science	multi-alignment
Gordon Hutchinson	U. British Columbia	Medicine, Biology	gene recognition, repetitive elements
Toni Kazic	Washington U.	Biology, Computer Science	genome informatics, metabolic pathways
Jim Knight	U. Arizona	Computer Science	pattern matching
Leonid Kruglyak	Whitehead (MIT)	Physics, Software	physical mapping
Alan Lapedes	LANL	Physics, Mathematics	neural nets, protein folding
Richard Lathrop	MIT AI	Computer Science	pattern learning & matching, protein folding
Suzanna Lewis	LBL	Biology, Software, Computer Science	physical mapping databases
Catherine Macken	LANL	Biology, Software	gene recognition
Gene Myers	U. Arizona	Computer Science, Software	high speed searches, pattern languages
Bill Pearson	U. Virginia	Biochemistry, Software	scoring schemes, high speed searches
Victor Solovyev	Baylor	Biology, Software	sequence analysis software
Gary Stormo	U. Colorado	Biology, Software	predicting functional regions, pattern-learning, sequence signals
Michael Storrie-Lombardi	Cambridge	Astronomy, Physics	protein folding

Presentations made

The table below is a list of the formally scheduled talks given during the workshop (there were a number of more informal presentations/discussions). The focus was both on work already accomplished and on current problems and future strategies to address them. Several of the talks were tutorial in nature.

Speaker	Topic
George Bell	A taxonomy of repetitive DNA elements
Gordon Hutchinson	A taxonomy of Alus

Gordon Hutchinson	Identifying genes by content *and* signal
Victor Solovyev	Using discriminant analysis to recognize acceptor|donor sites
Tim Bailey	EM algorithms for finding a signal common to a set of sequences
Gary Stormo	2nd order neural nets for pattern detection
Sean Eddy	Hidden markov models for alignment and pattern detection
James Knight	Language for super pattern matching: combining signals
Rick Lathrop	Learning pattern parameters and a CM2 implementation
Gene Myers	Approximate matching of context free patterns
Sean Eddy	Hidden markov models for RNA folding
Rick Lathrop	Motif threading for the inverse protein folding problem
Toni Kazic	Computational physiology
Paul Horton	A* algorithm for multi-alignment
Jim Holloway	Alignment allowing inversions and transpositions
Bill Pearson	How to carefully discriminate distant homologies in DB searches
Jim Holloway	SAR method for genetic mapping
Suzanna Lewis	STS content mapping
Leonid Kruglyak	Radiation hybrids for mapping markers
D'vorah Graeser	Technology for sensitive hybridization probes
Christian Burks	Stochastic methods for shotgun assembly & constraints
Gene Myers	Constraints in shotgun assembly
Gene Myers	Fast methods for overlap detection in shotgun assembly

Workshop citations.

We know of over 40 papers[10–51] that explicitly acknowledge the Aspen Center for Physics on work initiated or advanced during attendance at RG-I, RG-II, RG-III, IF-IV, or IF-V. In addition, several other pieces of work[52–57] were influenced by the workshops (but do not cite it directly). We anticipate that the number of such examples will continue to increase as ongoing work is published.

In addition, the Organizing Committee wrote this summary report[8].

Acknowledgements.

We are extremely grateful to ACP for sponsoring this workshop, and to the Center for Human Genome Studies and the Theoretical Division at Los Alamos National Laboratory, for providing funds for this workshop. In addition, we are indebted to Colorado SuperNet, Inc. for donating the hardware and staff support for establishing and maintaining the slip-link connection to the Internet during the workshop; and to Sun Microsystems for donating a Sun workstation server during the workshop. The Theoretical Biology and Biophysics Group at Los Alamos National Laboratory loaned several Sun workstations and peripheral hardware to us; in addition, M. Engle and F. Martinez in that group provided systems planning and support for the workshop's workstation network. We also acknowledge the able administrative support from the ACP staff.

References.

1. Burks, C., Myers, E. and Stormo, G.D. (1990) *Recognizing Genes and Other Components of Genomic Structure: Workshop Report (Aspen Center for Physics, 28 May - 15 June, 1990)*, Los Alamos National Laboratory, technical report LA-UR-91-1713.
2. Burks, C., Fields, C. and Myers, E. (1991) *Recognizing Genes and Other Components of Genomic Structure: Second Workshop Report (Aspen Center for Physics, 20 May - 07 June, 1991)*, Los Alamos National Laboratory, technical report LA-UR-92-645.
3. Burks, C., Fields, C., Henikoff, S., Joseph, D., and Stormo, G. (1992) *Recognizing Genes and Other Components of Genomic Structure: Third Workshop Report (Aspen Center for Physics, 18 May - 5 June, 1992).* Los Alamos National Laboratory, technical report.
4. Durbin, R., Dear, S., Gleeson, T., Green, P., Hillier, L., Lee, C., Staden, R. and Thierry-Mieg, J. (1991) Software for the C. elegans genome project. *Abstracts of Paper Presented at the 1991 Meeting on Genome Mapping and Sequencing*, Olson, M., Cantor, C. and Roberts, R., Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 126-126.
5. Fields, C. and Soderlund, C. (1990) gm: A practical tool for automating DNA sequence analysis. *CABIOS*, **6**, 263-270.
6. Fields, C.A., Soderlund, C.A. and Shanmugam, P. (1991) Performance of **gm**, Version 2.0. *Abstracts of Paper Presented at the 1991 Meeting on Genome Mapping and Sequencing*, Olson, M., Cantor, C. and Roberts, R., Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 127-127.
7. Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1991) A rule-based approach to the prediction of gene structure. *Abstracts of Paper Presented at the 1991 Meeting on Genome Mapping and Sequencing*, Olson, M., Cantor, C. and Roberts, R., Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 128-128.
8. Myers, E., Burks, C., and Stormo, G. (1994) *Identifying Features in Biological Sequences Fifth Workshop Report (Aspen Center for Physics, 30 May - 19 June, 1994).* Technical Report TR 94-36, Dept. of Computer Science, U. of Arizona, Tucson, AZ 85721 (this report).
9. Soderlund, C.A., Shanmugam, P. and Fields, C.A. (1991) New functions in **gm**, Version 2.0. *Abstracts of Paper Presented at the 1991 Meeting on Genome Mapping and Sequencing*, Olson, M., Cantor, C. and Roberts, R., Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 193-193.

Papers That Cite The ACP Workshops:

10. Burks, C., Parsons, R.J., and Engle, M.L. (1994) Integration of competing ancillary assertions in genome assembly. *Proceedings: Second International Conference on Intelligent Systems for Molecular Biology*, Altman, R., Brutlag, D., Karp, P., Lathrop, R., and Searls, D., Eds., AAAI Press, Menlo Park, CA, pp. 62-69.
11. Churchill, G., Burks, C., Eggert, M., Engle, M.L., and Waterman, M.S. (1993) Assembling DNA sequence fragments by shuffling and simulated annealing. Los Alamos National Laboratory, Technical Report LA-UR-93-2287
12. Engle, M.L., and Burks, C. (1993) Artificially generated data sets for testing DNA sequence assembly algorithms. *Genomics* **16**, pp. 286-288.
13. Engle, M.L. and Burks, C. (1994) GenFrag 2.1: new features for more robust fragment assembly benchmarks. *Comp. Applic. Biosci.* **10**, 567-568.

14. Fichant, G.A. and Burks, C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, **220**, 659-671.
15. Fickett, J.W. and Tung, C.S. (1992) Assessment of protein coding measures. *Nucl. Acids Res.* **20**, pp. 6441-6450.
16. Fickett, J.W. (1995) A developer's introduction to the gene identification problem. Manuscript submitted.
17. Henikoff, S. (1991) Playing with blocks: some pitfalls of forcing multiple alignments. *New Biol.*, **3**, pp. 1148-1154.
18. Henikoff, S. and Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. *Nucl. Acids Res.*, **19**, pp. 6565-6572.
19. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**, pp. 10915-10919.
20. Henikoff, S. and Henikoff, J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics* **19**, pp. 97-107.
21. Henikoff, S. and Henikoff, J.G. (1994) A protein family classification method for analysis of large DNA sequences. *Proc. 27th Ann. Hawaii Conf. on Systems Sciences*, pp. 265-274.
22. Holloway, J.L. and Cull, P. (1994) Aligning Genomes with Inversions and Swaps. *Proceedings, Second International Conference on Intelligent Systems for Molecular Biology*, pp. 195--202.
23. Honda, S., Parrott, N.W., Smith, R., and Lawrence, C. (1993) An object model for genome information at all levels of resolution. *Proc. 26th Hawaii International Conference on System Sciences* Vol. 1 (eds. T.Mudge, V. Milutinovic, and L.Hunter) IEEE Computer Society Press, Los Alamitos, CA.
24. Hutchinson, G.B. (1994) Towards the Automation of Feature Recognition in DNA Sequence. Ph.D. Thesis. The University of British Columbia. Vancouver, British Columbia, Canada.
25. Kazic, T. (1993) Representation, reasoning and the intermediary metabolism of *E. Coli*. *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, T.N. Mudge, V. Milutinovic and L. Hunter (eds.), vol. I, pp. 853-862.
26. Kazic, T. (1993) Reasoning about biochemical compounds and processes. *Second International Conference on Bioinformatics, Supercomputing and the Human Genome Project*, H. A. Lim, J. W. Fickett, C. R. Cantor and R. J. Robbins, eds. World Scientific, Singapore, pp. 35-49.
27. Kazic, T. (1994) Biochemical databases: challenges and opportunities. *New Data Challenges in Our Information Age. Proceedings of the Thirteenth International CODATA Conference*. P. S. Glaeser and M. T. L. Millward, eds., pp. C133-C140.
28. Kazic, T. (1995) Representing biochemistry for modeling organisms. *Molecular Modeling: From Virtual Tools to Real Problems*, T. Kumosinksi and M. N. Liebman, eds. American Chemical Society, in press.
29. Knight, J. and Myers, E. (1995) Super Pattern Matching. *Algorithmica* **13**, pp. 211-243.
30. Lathrop, R.H. and Temple F. Smith, T.F. (1995) A Branch and Bound Search Algorithm for Finding the Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Potentials. Submitted for publication.
31. Li, W. (1991) Expansion-modification systems: a model for spatial 1/f spectra. *Phys. Rev. A*, **43**, 5240-5260.
32. Li, W. (1992) Generating non-trivial long-range correlations and 1/f spectra by replication and mutation. *Internat. J. Bifurc. Chaos.* **2**, pp. 137-154.
33. Li, W. and Kaneko, K. (1992) Long-range correlation and partial $1/f^{\alpha}$ spectrum in a non-coding DNA sequence. *Europhysics Letters* **17**, pp. 655-660.
34. Mehldau, G. and Myers, E. (1993) A System for Pattern Matching Applications on Biosequences. *CABIOS* **9**, pp. 299-314.
35. Mount, S.M., Burks, C., Hertz, G., Stormo, G., White, O., and Fields, C. (1992) Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucl. Acids Res.*, **20**, pp. 4255-4262.
36. Mount, S. M. (1993) Messenger RNA splicing signals in Drosophila genes. *An Atlas of Drosophila Genes*, G. Maroni, ed.. Oxford University Press.
37. Myers, E. (1994) A Sublinear Algorithm for Approximate Keyword Matching. *Algorithmica* **12**, pp. 345-374.
38. Myers, E. (1994) Approximately Matching Context Free Languages. *Information Processing Letters*, submitted for publication.
39. Myers, E. (1994) Approximate Matching of Network Expressions with Spacers. *Journal of Computational Biology*, submitted for publication.
40. Myers, E. (1994) Algorithmic Advances for Searching Biosequence Databases. in *Computational Methods in Genome Research* (S. Suhai, ed.), Plenum Press (New York), pp. 121-135.

41. Myers, E. and Miller, W. (1995) Chaining Multiple-Alignment Fragments in Sub-Quadratic Time. *Proceedings of the Sixth ACM-SIAM Symposium on Discrete Algorithms*, accepted for publication.
42. Prestridge, D.S. and Burks, C. (1993) The density of transcriptional elements in promoter and non-promoter sequences. *Human Molec. Genet.* **2**, 1449-1453.
43. Salamov A.A., and Solovyev V.V. (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiply sequence alignments. *J. Mol. Biol.*, submitted.
44. Schöniger, M. and von Haeseler, A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3**, pp. 240-247.
45. Snyder, E.E. and Stormo, G.D. (1994) Identification of Protein Coding Regions in Genomic DNA. *J. Mol. Biol.*, accepted.
46. Soderlund, C., Shanmugam, P., White, O., and Fields, C. (1991) **gm**: A tool for exploratory analysis of DNA sequence data. *Proceedings of the Twenty-Fifth Hawaii International Conference On System Sciences: Vol I*. Milutinovic, V. and Shriver, B.D., Eds., IEEEE Computer Society Press, Washington, pp. 653-662.
47. Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of open reading frames. *Nucl. Acids Res.*, in press.
48. Solovyev V.V., Salamov A.A., and Lawrence C.B. (1995) Prediction of human gene structure based on discriminant analysis and graph algorithms. *J. Mol. Biol.*, submitted.
49. Solovyev V.V., and Lawrence C.B. (1995) Analysis and prediction of messenger RNA splicing signals in Mammalia genes. *CABIOS*, to be submitted.
50. Storrie-Lombardi, M.C. and Lahav, O. (1995) Neural network applications in astronomy. *Handbook of Brain Theory and Neural Networks*, M.A. Arbib, Ed., Boston: M.I.T. Press, in press.
51. Storrie-Lombardi, M.C., Irwin, M.J., von Hippel, T., and Storrie-Lombardi, L.J. (1994) Spectral classification with principal component analysis and artificial neural networks. *Vistas in Astronomy* **38**, in press.

Papers That Were Influenced By Participation At The ACP Workshops:

52. Burks, C., Engle, M.L., Forrest, S., Parsons, R.J., Soderlund, C.A., and Stolorz, P.E. (1994) Stochastic optimization tools for genomic sequence assembly. *Automated DNA Sequencing and Analysis*, Adams, M.D., Fields, C., and Venter, J.C., Eds., Academic Press, New York, pp. 249-259.
53. Guo, M., P. C. H. Lo and S.M. Mount (1993) Species-specific signals for the splicing of a short Drosophila intron in vitro. *Mol. Cell. Biol.* **13**, pp. 1104-1118.
54. Guo, M. and S. M. Mount (1995) Localization of sequences required for size-specific splicing of a small Drosophila intron in vitro. Submitted.
55. Kazic, T. and Tsur, S. (1993) Modeling and simulating biological processes as logical enterprises. *Proceedings of the NSF Scientific Database Projects, 1991-1993*, W. W. Chu, A. F. Cardenas and R. K. Taira, eds., pp. 16-22.
56. Mount, S. M. and S. Henikoff (1993) Nested genes take flight. *Current Biology* **3**, 372-374.
57. Schöniger, M. and Waterman, M.S. (1992) A local algorithm for DNA sequence alignment with inversions. *Bulletin of Mathematical Biology* **54**, pp. 521-536.