

Mihail R. Halachev

Dept. of Computer Science & Software Engineering
Concordia University
1455 De Maisonneuve Blvd. West,
EV 9.101, Montreal, Quebec
Canada, H3G 1M8

email: m_halach@encs.concordia.ca
http://users.encs.concordia.ca/~m_halach/
Phone: (514) 848-2424 ext.7160 (office)

Education	❖ Ph.D. in Computer Science Concordia University Montreal, Canada	Mar. 2009 (expected)
	❖ MS in Computer Science University of Oklahoma Norman, USA	2003
	❖ MS in Civil Engineering University of Architecture, Civil Engineering, and Geodesy Sofia, Bulgaria	1996

Research Interests	<ul style="list-style-type: none">▪ Bioinformatics Algorithms and Applications▪ Indexing Techniques for Sequence Data▪ Modeling and Management of Biological Sequence Data▪ Data Mining
---------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Teaching Interests	<ul style="list-style-type: none">▪ Databases and Database Management Systems▪ Introduction to Theoretical Computer Science▪ Bioinformatics
---------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Area of Expertise

Searching in Biological Sequence Data: Indexing Techniques and Applications

Problem Description	The amount of publicly available biological sequence data, represented as long strings over the DNA and protein alphabets, is of considerable size and exhibits a stable trend of exponential growth. Providing efficient performance for the various search operations over such data requires development of appropriate indexing techniques tailored to the specifics of the domain.
Existing Techniques	Some of the existing techniques require extensive computational resources and/or lack versatility with respect to the search tasks supported. Further, the proposed solutions provide either efficiency (e.g., Vmatch [1]) or scalability (e.g., Trellis [2]), but not both.
Proposed Solution	<p>Developed and implemented the suffix tree - based HST indexing technique.</p> <ul style="list-style-type: none"> ▪ Designed for typical desktop computers (e.g., single CPU, 2 GB RAM) ▪ Handles sequences of size up to 2^{32} characters (i.e., 4 Gbp) ▪ Reasonable construction times (16 hours for the entire human genome) ▪ Acceptable storage requirements (13 bytes per sequence character) ▪ Currently supports four different bioinformatics search tasks ▪ Provides efficient and scalable search in biological sequence data
Experimental Results	<p>Chromosome Scale Sequences (up to 250 Mbp)</p> <p><u>Exact Match Search</u></p> <ul style="list-style-type: none"> ▪ HST is comparable to Vmatch ▪ HST is an order of magnitude faster than Trellis <p><u>Approximate (k-mismatch) Search</u></p> <ul style="list-style-type: none"> ▪ HST is 3 times faster than Vmatch <p><u>Supermaximal Repeats Search</u></p> <ul style="list-style-type: none"> ▪ HST is 2/9 times faster than Vmatch for repeats of size at least 10/200 bp <p><u>Structured Motif Search</u></p> <ul style="list-style-type: none"> ▪ HST is 5 to 6 times faster than SMOTIF [3], the best known solution <p>Genome Scale Sequences (e.g., entire human genome – 2.8Gbp)</p> <p><u>Exact Match Search</u></p> <ul style="list-style-type: none"> ▪ HST is more than 20 times faster than direct Trellis ▪ HST is 3 times faster than Vmatch (searching at chromosome level) <p><u>Approximate (k-mismatch) Search</u></p> <ul style="list-style-type: none"> ▪ HST is 5 times faster than Vmatch (searching at chromosome level)

[1] Abouelhoda, M., Kurtz, S., and Ohlebusch, E. (2004) “Replacing suffix trees with enhanced suffix arrays”, *Journal of Discrete Algorithms*, 2(1), pp. 53-86.

[2] Phoophakdee, B. and Zaki, M. (2007) “Genome-scale disk-based suffix tree indexing”, In *Proc. of ACM SIGMOD Intl. Conference on Management of Data*, pp. 833-844, Beijing, China, 2007.

[3] Zhang, Y. and Zaki, M. (2006) “SMOTIF: efficient structured pattern and profile motif search”, *Algorithms for Molecular Biology* 2006, 1:22, doi:10.1186/1748-7188-1-22.

Publications	<p>1) Halachev, M. and Shiri, N. (2008). Fast Structured Motif Search in DNA Sequences. In <i>CCIS 13</i>, pp.58-73, 2008, M. Elloumi <i>et al.</i> (Eds.). Proceedings of 2nd Bioinformatics Research and Development Conference (BIRD'08), Vienna, Austria, July 2008.</p> <p>2) Lian, C. L., Halachev, M., Shiri, N. (2008). Searching for Supermaximal Repeats in Large DNA Sequences. In <i>CCIS 13</i>, pp.87-101, 2008, M. Elloumi <i>et al.</i> (Eds.). Proceedings of 2nd Bioinformatics Research and Development Conference (BIRD'08), Vienna, Austria, July 2008.</p> <p>3) Halachev, M., Shiri, N., Thamildurai, A. (2007). Efficient and Scalable Indexing Techniques for Biological Sequence Data. In <i>LNBI 4414</i>, pp.464-479, 2007, S. Hochreiter and R. Wagner (Eds.). Proceedings of 1st Bioinformatics Research and Development Conference (BIRD'07), Berlin, Germany, March 2007.</p> <p>4) Halachev, M., Shiri, N., Thamildurai, A. (2005). Exact Match Search in Sequence Data Using Suffix Trees. In <i>Proc. ACM 14th Int'l Conf. on Information and Knowledge Management (CIKM'05)</i>, pp. 123 – 130, Bremen, Germany, November 2005.</p> <p>5) Halachev, M., Shiri, N., Thamildurai, A. (2005). Exact Match Search in Biological Sequence Data Using Suffix Trees (Poster). At <i>Knowledge-Based Bioinformatics (KBB) Workshop</i>, Montreal, Canada, September 2005.</p> <p>6) Halatchev, M. and Gruenwald, L. (2005). Estimating Missing Values in Related Sensor Data Streams. In <i>Proc. of the 11th Int'l Conf. on Management of Data (COMAD '05)</i>, pp. 83 - 94, Goa, India, January 2005.</p>
---------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Conference Presentations	<p>BIRD 2008 Vienna, Austria, July 2008</p> <ul style="list-style-type: none"> ▪ Fast Structured Motif Search in DNA Sequences ▪ Searching for Supermaximal Repeats in Large DNA Sequences
	<p>BIRD 2007 Berlin, Germany, March 2007</p> <ul style="list-style-type: none"> ▪ Efficient and Scalable Indexing Techniques for Biological Sequence Data
	<p>CIKM 2005 Bremen, Germany, Nov. 2005</p> <ul style="list-style-type: none"> ▪ Exact Match Search in Sequence Data Using Suffix Trees

Work in Progress	<p>Halachev, M., Shiri, N., and Thamildurai, A. Fast and Scalable Search in Biological Sequences. Manuscript in preparation for journal submission.</p>
-------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------

Research Experience	<p>❖ Research Assistant 2004 – Present</p> <p><u>Concordia University, Dept. of Computer Science & Software Engineering</u> Project: “Suitable Indexing Techniques for Biological Sequence Data” Supervisor: Dr. Nematollah Shiri, Concordia University</p> <ul style="list-style-type: none"> ▪ Studied and experimented with various indexing and search techniques for biological sequences ▪ Developed and implemented an efficient, scalable, and versatile suffix tree-based indexing technique suitable for regular desktop computers ▪ Conducted extensive experimental evaluation of the proposed technique and compared it with state-of-the-art solutions using real-life data
	<p>❖ Research Assistant 2001 – 2003</p> <p><u>University of Oklahoma, School of Computer Science</u> Project: “Estimating Missing Values in Related Data Streams” Supervisor: Dr. Le Gruenwald, University of Oklahoma</p> <ul style="list-style-type: none"> ▪ Studied the applicability and suitability of data-mining techniques ▪ Proposed an association rule - based solution for power-aware estimation of missing sensor readings in wireless networks
Teaching Experience	<p>❖ Teaching Assistant</p> <p><u>Concordia University, Dept. of Computer Science & Software Engineering</u> Introduction to Theoretical Computer Science Summer `07 Introduction to Theoretical Computer Science Summer `06 Introduction to Theoretical Computer Science Summer `05 Databases Winter `05</p> <p><u>University of Oklahoma, School of Computer Science</u> Introduction to Database Management Systems Fall `02 Programming for Engineers Spring `01</p> <ul style="list-style-type: none"> ▪ Led weekly tutorial sessions ▪ Helped with homework assignments design, marked solutions, and provided feedback to students ▪ Held office hours to for additional communication with students
	<p><u>Ph.D. Seminar in University Teaching, Concordia University</u> Fall `07</p> <ul style="list-style-type: none"> ▪ Practiced various teaching approaches (e.g., teacher dependence, self-directedness, group discussions) ▪ Formalized goals for student assessment (e.g., exam design/markings) ▪ Designed a database course and developed a syllabus for this course ▪ Planned and lectured a mini-lesson on database indexing ▪ Discussed ethical issues in higher education, such as plagiarism, team dynamics, late submissions ▪ Obtained a certificate upon successful completion of requirements

Programming Languages	C/C++, Java, Fortran, SQL, OWL, HTML, Windows/Linux
------------------------------	-----------------------------------------------------

Software Tools	<p>FASST (Fast and Scalable Search Tool for biological sequence data) http://sepehr.cs.concordia.ca/</p> <p>FASST takes as an input a DNA or protein sequence and constructs its suffix tree - based HST index, which is persistently stored on disk. Subsequently, the index is used to provide efficient and scalable support for various search tasks, including (at this stage): exact match search, approximate (k-mismatch) search, search for structured motifs, and search for exact supermaximal repeats.</p>
-----------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Honors & Awards	<p>Student Travel Grant June 2008 From: School of Graduate Studies, Concordia University For: BIRD'08 Conference</p>
	<p>Teaching Fellowship Winter 2006 From: Dept. of Comp. Sci. & Soft. Eng., Concordia University For: COMP 353 Databases course</p>
	<p>Student Travel Grant Nov. 2005 From: CIKM Conference Organizers For: CIKM'05 Conference</p>
	<p>Teaching Fellowship Winter 2005 From: Dept. of Comp. Sci. & Soft. Eng., Concordia University For: COMP 353 Databases course</p>
	<p>Student Travel Grant Aug. 2004 From: Intel Corporation For: DMSN'04 workshop (in conjunction with VLDB'04)</p>

References

Dr. Nematollaah Shiri

Associate Professor and
Graduate Program Director (Research)
shiri@cs.concordia.ca
(514) 848-2424 ext.3018
[Ph.D. Supervisor]

Department of Computer Science and Software Engineering
Concordia University
1455 De Maisonneuve Blvd. West, EV 3169
Montreal, Quebec, Canada, H3G 1M8

Dr. Gregory Butler

Professor
gregb@cs.concordia.ca
(514) 848-2424 ext.3031
[Dissertation Committee Member]

Department of Computer Science and Software Engineering
Concordia University
1455 De Maisonneuve Blvd. West, EV 3219
Montreal, Quebec, Canada, H3G 1M8

Dr. Le Gruenwald

Director and Presidential and
David W. Franke Professor
ggruenwald@ou.edu
(405) 325-3498
[MS Thesis Supervisor]

School of Computer Science
University of Oklahoma
200 Felgar Street, 116 EL
Norman, OK, 73019, USA