Dennis E. Shasha – Publication Highlights

Biology and Bioinformatics

Most of the work has had to have with the design of experiments using combinatorial design, analysis in order to find important genes or experimental techniques. Papers are listed in reverse chronological order with commentary.

1. "DNA Hash Pooling and its Application" Dennis Shasha and Martyn Amos International Journal of Nanotechnology and Molecular Computation 1(1), 18-32, January-March 2009
   Commentary: a technique for identifying whether two multi-species samples have common species even if the species have never been sequenced.

2. "An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis" Karen E Thum, Michael J Shin, Rodrigo Gutierrez, Indrani Mukherjee, Manpreet S Katari, Damion Nero, Dennis Shasha and Gloria M Coruzzi BMC Systems Biology 2008, 2:31 (04 Apr 2008)
   Commentary: analysis of nitrogen and light in plants.

3. "Insights into the genomic nitrate response using genetics and the Sungear software system" Rodrigo A. Gutirrez, Miriam L. Gifford, Chris Poultney, Rongchen Wang, Dennis E. Shasha, Gloria M. Coruzzi and Nigel M. Crawford JXB Advance Access published online on April 29, 2007 Journal of Experimental Botany, doi:10.1093/jxb/erm079
   Commentary: Use of visualization software Sungear to analyze the similarity among different experiments in the literature.

4. Rodrigo A. Gutierrez, Laurence V. Lejay, Alexis Dean, Francesca Chiaromonte, Dennis E. Shasha, and Gloria M. Coruzzi, "Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis" Genome Biology, 2007, 8:R7 (doi: 10.1186/gb-2007-8-1-r7) 11 January 2007. Cited as a "must read" in the Faculty of 1000.
   Commentary: Interaction of carbon and nitrogen in plants.

5. Christopher S. Poultney, Rodrigo A. Gutirrez, Manpreet S. Katari, Miriam L. Gifford, W. Bradford Paley, Gloria M. Coruzzi and Dennis E. Shasha "Sungear: Interactive visualization and functional analysis of genomic datasets" Bioinformatics, 2007; Jan 15;23(2):259-61 doi: 10.1093/bioinformatics/btl496
   Commentary: Visualization system called Sungear to compare the results of multiple experiments.

6. Charles J. Colbourn, Sosina S. Martirosyan, Gary L. Mullen, Dennis Shasha, George B. Sherwood, Joseph L. Yucas "Products of Mixed Covering Arrays of Strength Two" Journal of Combinatorial Designs Volume 14, Issue 2, Date: March 2006, Pages: 124-138
Commentary: theoretical paper generalizing the use of combinatorial design already used in math.

7. Rodrigo Gutierrez, Dennis Shasha, and Gloria Coruzzi, "Systems Biology for the Virtual Plant" Plant Physiology, June 2005, vol. 38, pp. 550-554.
Commentary: description of a database system for experiments in plants.

8. "A gene expression map of the Arabidopsis root" Kenneth Birnbaum, Dennis E. Shasha, Jean Y. Wang, Jee W. Jung, Georgina M. Lambert, David W. Galbraith, and Philip N. Benfey Science, Dec 12 2003: 1956-1960 (A review article in the Research Focus section of Trends in Biotechnology called the article "At the end of 2003, the root biology community was blessed with what has become today already a historical paper that described for the first time a genome wide expression analysis of Arabidopsis root development [2].")

9. Mitchell Levesque, Dennis Shasha, Wook Kim, Michael G. Surette, and Philip N. Benfey "Trait-To-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes" *Current Biology*, vol. 13, 129-133, January 21, 2003. Discussed in: http://www.the-scientist.com/yr2003/jun/hot_030603.html
Commentary: A technique for inferring the function of an unknown gene based on species traits. Verified in flagella.

Tree and graph matching/nearest neighbor searching

This work extends approximate string matching to searching and inexact comparison in trees and graphs.

1. A. Ferro, R. Giugno, G. Pigola, A. Pulvirenti, D. Skripin, G. D. Bader, D. Shasha "NetMatch: a Cytoscape Plugin for Searching Biological Networks" Bioinformatics, 2007 23(7):910-912; doi:10.1093/bioinformatics/btm032
Commentary: A system for finding subgraphs in bigger graphs. This is heuristic because the problem is NP-complete, but is very effective.

2. Jason T. L. Wang, Huiyuan Shan, Dennis Shasha and William H. Piel, "Fast Structural Search in Phylogenetic Databases," Evolutionary Bioinformatics Online, Vol. 1, October 2005, pp. 37-46.
Commentary: phylogenetic databases consist of unordered trees. This system permits search in those databases.

3. "Homology search for genes" Xuefeng Cui; Tomas Vinar; Brona Brejova; Dennis Shasha; Ming Li Bioinformatics. 2007 Jul 1;23 (13):i97-i103 17646351 (P,S,E,B,D)
   Commentary: genes in higher organisms contain parts that are excised out. Doing a matching in that case requires special approximation techniques.

4. "Antipole Tree Indexing to Support Range Search and K-Nearest Neighbor Search in Metric Spaces" Domenico Cantone, Alfredo Ferro, Alfredo Pulvirenti, Diego Reforgiata, Dennis Shasha IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 5 (April 2005), pp. 535-550.
   Commentary: A technique for rapid search in metric spaces.

5. "Algorithmics and Applications of Tree and Graph Searching" Dennis Shasha, Jason Wang, Rosalba Giugno ACM Pods 2002, May 2002. (Invited) pp. 39-52.
   Commentary: A review of previous work that has since led to a substantial body of work in tree and graph comparison within the database community.

6. "Approximate Tree Matching in the Presence of Variable Length Don't Cares" K. Zhang, T-L. Wang, and Dennis Shasha *Journal of Algorithms*, vol. 16, pp. 33-66 (1994).
   Commentary: approximate matching in ordered trees (sibling order matters) but with wildcards.

7. "Fast Algorithms for the Unit Cost Editing Distance Between Trees" Dennis Shasha and K. Zhang *Journal of Algorithms*, vol. 11, pp. 581-621 (1990).

8. "Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems" K. Zhang and Dennis Shasha *Siam Journal of Computing*, vol. 18, no. 6, pp. 1245-1262, December 1989.
   Commentary: Algorithm whose complexity is theoretically and practically close to approximate string matching. Used in various IBM's XMLdiff software.

Concurrency Control, Data Structures, Database Tuning, and Scheduling

Most of my initial systems work in databases has had to do with concurrency control and data structures. Later, this evolved into work about database tuning with a short interlude in real-time scheduling for overloaded systems.

1. "Making Snapshot Isolation Serializable" Alan Fekete, Dimitrios Liarokapis, Elizabeth O'Neil, Patrick O'Neil, Dennis Shasha ACM TODS, June 2005 vol. 30, number 2. pp. 492-528
   Commentary: Snapshot isolation as currently practiced does not guarantee serializability. This paper showed the theory underlying a fix.

2. *Database Tuning : Principles Experiments and Troubleshooting Techniques* Dennis Shasha and Philippe Bonnet, Morgan Kaufmann Publishers, June 2002, ISBN 1-55860-753-6, Paper, 464 Pages. (Translations to Russian (Kudits obraz), simplified Chinese (Publishing House of Electronics Industry, phei), and Korean (Brain Korea) are complete.)
   Commentary: Second and vastly expanded version of my single author 1992 book on the subject. Since the publication of these books and our tutorials at Sigmod and VLDB, there has been substantial work in database tuning.

3. "Transaction Chopping: Algorithms and Performance Studies" Dennis Shasha, F. Llirbat, E. Simon, P. Valduriez *ACM Transactions on Database Systems*, October 1995, pp. 325-363.
   Commentary: A technique for improving the concurrency by chopping transactions into smaller transactions but still maintaining the serializability of the original transactions.

4. "D-Over: An Optimal On-Line Scheduling Algorithm for Overloaded Uniprocessor Real-Time Systems" G. Koren and Dennis Shasha *Siam Journal on Computing*, April 1995, pp. 318-339, vol. 24, no. 2.
   Commentary: An algorithm for firm real-time systems (in which tasks receive value for finishing by the deadline and no value for finishing after the deadline) that is provably optimal under overload.

5. "Inserts and Deletes on B-trees: why free-at-empty is better than merge-at-half" T. Johnson and Dennis Shasha *Journal of Computer Sciences and Systems*, invited, vol. 47, no. 1, pp. 45-76, Aug. 1993.
   Commentary: showing that utilization is nearly as good for B-trees that never do merges as long as there are more inserts than deletes. Since used by many implementors.

6. "2Q: a low overhead high performance buffer replacement algorithm" T. Johnson and Dennis Shasha, *Very Large Database Systems Conference 1994*, September, 1994.
   Commentary: A constant time approximation to the LRU/k buffer management algorithm of O'Neil and Weikum. Has been implemented widely among search engine companies.

7. "Concurrent Search Structure Algorithms" Dennis Shasha and N. Goodman, *ACM Transactions on Database Systems*, vol. 13, no. 1, pp. 53-90,

March 1988.
Commentary: A framework for verifying non-conflict preserving serializable concurrent algorithms on data structures.

Cryptography and Data

The goal of preserving privacy in a technological world has led me to work on different platforms and for different purposes.

1. Nicolas Anciaux, Mehdi Benzine, Luc Bouganim, Philippe Pucheral, Dennis Shasha: "GhostDB: querying visible and hidden data without leaks" SIGMOD Conference 2007: 677-688
   Commentary: Use of flash drives to hold and process data secretly when using untrusted computers.

2. Michael Rabin and Dennis Shasha "Preventing Piracy while Preserving Privacy" Dr. Dobb's Journal, October 2005.
   Commentary: an approach to digital rights management that preserves privacy. Also the subject of several patents.

3. Jinyuan Li, Maxwell Krohn, David Mazieres, and Dennis Shasha "Secure Untrusted Data Repository (SUNDR)" Proceedings of the 6th Symposium On Operating Systems Design and Implementation (OSDI '04) San Francisco, CA. December, 2004.
   Commentary: An efficient implementation of SUNDR – a file system sitting on top of an untrusted data repository.

4. "Building secure file systems out of Byzantine storage", David Mazieres and Dennis Shasha, Principles of Distributed Computing, 2002. pp. 108-117.
   Commentary: Theory underlying SUNDR including the notion of fork consistency.

Time Series

This works consists of attempts to compute incremental statistics very fast on streaming time series data.

1. "Better Burst Detection" IEEE International Conference on Data Engineering, April 2006 p. 146ff Xin Zhang and Dennis Shasha
   Commentary: A one-sweep approach to detecting bursts at multiple time scales from milliseconds to hours.

2. "Fast Window Correlations Over Uncooperative Time Series" Richard Cole, Dennis Shasha, and Xiaojian Zhao, ACM Knowledge and Data Discovery 2005, pp. 743-749.
Commentary: an efficient method for incrementally detecting correlated time series.

3. *High Performance Discovery in Time Series: techniques and case studies* Dennis Shasha and Yunyue Zhu, Springer Verlag Publishers, Monographs in Computer Science, June 2004, ISBN 0387008578, 270 Pages.
Commentary: Book describing techniques for incremental correlation, burst detection, and approximate matching in time series. This is the basis for the algorithms above.

4. "AQuery: Query Language for Ordered Data, Optimization Techniques, and Experiments", Alberto Lerner and Dennis Shasha, VLDB 2003
Commentary: Language that is a conservative extension to SQL to handle ordered data (e.g. time series data). Paper also includes algorithmic techniques.

Mathematical Puzzles for the General Public

Through columns and books, I've tried to evangelize computational thinking among the mathematically adept.

1. Dr. Dobb's Journal: *Omniheurist Puzzle Corner* April 1998 to September, 2002
April 2004 to December 2005.
Puzzle column for programmers

2. Scientific American: *Puzzling Adventures* April 2001 to May 2004 in magazine and on web.
June 2004 and forward at www.sciam.com.
Puzzle column for general scientific literates.

3. *Puzzles for Programmers and Pros* John Wiley/Wrox, May 2007. Puzzles and methods to solve puzzles. Translations to Korean and Japanese.
Commentary: Directed towards programmers as it tries to explain fundamental puzzle-solving techniques.

4. *The Puzzler's Elusion: A Tale of Fraud, Pursuit, and the Art of Logic* Thunder's Mouth Press, March 2006.
Commentary: A further collection of puzzles from Scientific American and Dr. Dobb's Journal. Translation to traditional Chinese 2007.

5. *Puzzling Adventures* W. W. Norton, January 2005. A collection of puzzles from Scientific American and Dr. Dobb's Journal.

Commentary: Puzzle contest was won by Jeremiah Farrell and written about on May 19, 2006 in the NY Sun under the title "A Washington Square Park Puzzle is Solved" by Gary Shapiro in the Arts and Letters section.

6. *Dr. Ecco's Cyberpuzzles : 36 Puzzles for Hackers and Other Mathematical Detectives* Dennis E. Shasha W. W. Norton, 2002. ISBN 0-393-05120-X, 231 pages. (Translations to Korean, Simplified Chinese, Korean, Polish, and Turkish currently in progress. Translation to Czech, Hungarian, French, German, Portugese, and traditional Chinese complete.)

7. *Codes, Puzzles, and Conspiracy* by Dennis Shasha, W. H. Freeman, New York 1992. Republished by Dover in 2004.
(Adventures of a mathematical detective whose problems are often algorithmic or combinatoric in nature. Second in series. Translated to French, Portuguese, Slovenian, and Turkish. Also translated to Spanish in two separate editions.)

8. *The Puzzling Adventures of Dr. Ecco* by Dennis Shasha, W. H. Freeman, New York 1988. Dover, 1997.
(Adventures of a mathematical detective whose problems are often algorithmic or combinatoric in nature. First in series. Translated to Chinese, French, German, Spanish, Japanese, Portuguese, Turkish, Slovenian, and Hungarian. Cited by Professor Andy Liu, selected the Canadian Professor of the Year in 1999, as his favorite book for teaching mathematics at the University of Alberta. The book is also used at Grant MacEwan College.)

Great Discoveries in Computer Science

This book attempts to explain great inventions in computer science through the lives of the inventors.

1. *Out of Their Minds: the lives and discoveries of 15 great computer scientsts* by Dennis Shasha and Cathy Lazere, Springer-Verlag, New York, August, 1995.
(Book of short biographies and research philosophies. Translated to Japanese, Korean, traditional Chinese (Taiwan), and simplified Chinese (China).)

Software Available

Much of the research work has resulted in software which we make available to the public.

1. Web sites:

   http://cs.nyu.edu/cs/faculty/shasha/papers/tree.html [ordered tree matching]

   http://cs.nyu.edu/cs/faculty/shasha/papers/treesearch.html [unordered tree matching]

   http://cs.nyu.edu/cs/faculty/shasha/papers/graphgrep/index.html [graph matching]

   http://cs.nyu.edu/cs/faculty/shasha/papers/statstream.html [time series matching]

   http://cs.nyu.edu/cs/faculty/shasha/fintime.html [benchmark for financial databases]

   http://cs.nyu.edu/cs/faculty/shasha/spytime/spytime.html [benchmark for bitemporal databases]