

INSTITUTE FOR COMPLEX SCIENTIFIC SOFTWARE

Gene Cooperman

End of the Year Report FY 2004

1 ACTIVITIES

1. **TITLE:** THEMATICIS Group

PI: Mary Jo Ondrechen (Chemistry); co-PIs: Bob Futrelle (CCIS); Ron Williams (CCIS); David Budil (Chemistry); Valentin Ilyin (Biology); Dagmar Ringe (Brandeis)

Supported: Terry Yang (PhD student, A&S), part of 2003-2004; Wenxu Tong (PhD student, CCIS), part of 2003-2004; Dr. Leo Murga (Postdoctoral fellow, A&S), part of 2003-2004; Heathew Brodtkin (PhD student, A&S), Summer I, 2004

Activities: We continue to make progress on several fronts.

- (a) We have now applied THEMATICIS to over 100 proteins.
- (b) With the help of our ICSS collaborators, we are now have software and parallel hardware to analyze protein structures with high throughput.
- (c) THEMATICIS finds not only catalytic residues, but also recognition residues.
- (d) Residues in THEMATICIS clusters are highly conserved in evolution. Furthermore, THEMATICIS clusters are subsets of the residues found to be important by methods based on evolutionary history.
- (e) THEMATICIS can work for theoretical model structures (without the bottleneck of a structure determination in the laboratory).
- (f) Ron Williams and Mary Jo Ondrechen, together with Visiting Assistant Professor Jaeju Ko (who is on sabbatical leave in the THEMATICIS group) and graduate student Terry Yang, have developed statistical metrics for use in the automation of THEMATICIS predictions.
- (g) The machine learning results obtained to date using Neural Nets and SVMs have provided some preliminary results that have been used in our grant proposals. These results are a great help in establishing that our proposed projects are feasible and will increase our chances of a successful funding outcome.

2. **TITLE:** Low-level algorithms to alleviate low-level architectural bottlenecks in high-performance computing.

PI: Gene Cooperman (CCIS)

Supported: Viet Ha Nguyen (PhD Student), Summer I, 2004

Activities: Ha is investigating the use of parallelism and data mining in bioinformatics. As part of some activities not directly supported by ICSS, Ha has worked to improve the parallelization of Geant4, one of the ICSS challenge problems. This has applications to the work of Reucroft and Swain in ICSS. As part of some activities not directly supported by ICSS, Viet Ha had started to work as part of a joint collaboration by Gene Cooperman and Valentin Ilyin. We are currently talking with Massachusetts General Hospital (MGH) about a position for Ha, supported in part by ICSS, to apply data-mining to some bioinformatics data.

3. **TITLE:** Accelerating Sparse Matrix Library Codes

PI: Gene Cooperman (CCIS) and David Kaeli (ECE); co-PI: Misha Kilmer (Tufts University)

Supported: Xiaoqin Ma (PhD student, CCIS), Summer, 2003; Yijian Wang (PhD student, ECE); Chakik Ouarraui (MS student, ECE)

Activities: The ICSS grant obtained by Kaeli and Cooperman was in part a result of these activities. This project contributed to more efficient reuse of clusters, development of a fast algorithm for permuting objects, a highly accurate model for memory-bound computations, and a new algorithm for using multiple CPUs for faster indexing into databases. There are also connections with the previous project. We are currently talking with EMC (one of the three big disk storage companies) about a proposed internship for Xiaoqin Ma there, as part of technology exchange. Chakik Ouarraui now works at EMC, since graduation.

4. **TITLE:** Using Aspect-Oriented Software Development Methods for Re-Design and Performance Optimization of a Legacy Bioinformatics Computation System

PIs: David Kaeli (ECE), David Lorenz (CCIS), Valentin Ilyin (BIO)

Supported: Sergei Kojarski (PhD student, CCIS), part of 2003-2004 year; Chesley Leslin (PhD student BIO), Summer I, 2004; Darren Ng (PhD student, ECE), part of 2003-2004 year; Diego Rivera (PhD student, ECE), Summer I, 2004

Activities: Third party scientific software systems are often hard to extend and maintain. Aspect-Oriented Programming (AOP) is a powerful reflective programming tool. In this project we developed effective AOP constructs for facilitating the process of scientific software program comprehension. We applied these constructs on the Java portion of a sizable legacy system for manipulating and displaying protein sequences entitled Friend. Friend is an integrated analytical front-end application for bioinformatics. Friend was designed to aid scientists visualize protein interactions along multiple alignments, domains, fragments, and binding sites.

In the process of understanding the Friend software, we applied AOP techniques using AspectJ to dynamically identify design flaws. We utilized aspectual reflection (Kojarski and Lorenz, 2003) to support profiling methods and we applied aspects to selectively unplug components. Together these aspect-oriented tools allow profiling beyond traditional Java core reflection and the capability to perform controlled refactoring of legacy code.

In addition to the immediate benefits obtained by understanding a particular body of code, the application of aspects to program comprehension advances the understanding of AOP.

Publications resulting from this project: 2 published papers:

- Sergei Kojarski and David H. Lorenz: “Unplugging Components using Aspects.” In Proceedings of the ECOOP 2003 Eighth International Workshop on Component-Oriented Programming. Darmstadt, Germany, July 21, 2003.

This paper explores the use of AOP for writing aspects that transform exiting hard-coded calls into plugs that can be used to retarget a client to use another subcomponents. By drawing on the code-transformation abilities of AOP (as found in AspectJ), we show how aspects can be used to transform (in “the other way”) from the aspectually woven to the unwoven (unplugged) code. This transformation is sometimes even more useful then AOP in its usual form: Given a monolithic system, one can apply AOP to non-intrusively decouple the system’s components, and may then replace some of the legacy components with alternative third-party components. Hence our approach can help to normalize and refactor existing code.

- Darren Ng, David R. Kaeli, Sergei Kojarski, and David H. Lorenz: “Program Comprehension Using Aspects.” In ICSE 2004 Workshop on Directions in Software Engineering Environments (WoDiSEE’2004).

This paper reports on using aspectual comprehension to understand three bodies of code. The first is the Java portion of a legacy system called Friend. The second is Eclipse, an open source Java IDE. The third is Compress, a SPEC JVM98 Java benchmark.

5. **TITLE:** Constructing Large Scientific Software using Pairwise Composition of Software Components

PIs: David Lorenz (CCIS) and Paul Attie (CCIS)

Supported: Hana Chockler (Postdoctoral fellow, CCIS), part of 2003-2004 year, cost-sharing with NSF

Activities: The project addresses the problem of devising mechanical methods for synthesis and verification of large component-based scientific software. The major impediment to such mechanical analysis is the state-explosion problem. To avoid state-explosion we represent a system comprising of a set of components C_1, C_2, \dots, C_n in a pairwise normal form: the code that synchronizes a pair C_i, C_j of components is cleanly separated from the code that synchronizes other (even overlapping) components. It is then possible to take selected small subsystems consisting of a small number of components and analyze them in isolation to verify a behavioral property f_φ . Our results then enable us to conclude that f_φ is also a behavioral property of the large system. This is achieved without recourse to “interface processes” and other such intermediate constructs, as is typical of “compositional” verification.

Publications resulting from this project: 1 published paper (and several unpublished manuscripts):

- Paul Attie and David H. Lorenz: “Correctness of Model-based Component Composition without State Explosion.” In Proceedings of the ECOOP 2003 Workshop on Correctness of Model-based Software Composition. Darmstadt, Germany, July 22, 2003.

Grant proposals resulting from this the project: Submission of two NSF grant proposals:

- Paul C. Attie and David H. Lorenz. Submitted 12/03; NSF; “A Toolset and Environment for Software Design based on Pairwise Composition of Software Components”; Requested amount: \$320,543.
- David H. Lorenz, Paul C. Attie, and Dana H. Brooks. Submitted 05/04; NSF; “Design Locality: A Concept for Controlling the Design Complexity of Large Software Systems”; Requested amount: \$547,353.

6. **TITLE:** Component-based Design for Large-Scale Computation

PIs: David Lorenz (CCIS), Dana Brooks (ECE)

Supported: Daniel Matysiak (undergraduate student, CCIS), part of 2003-2004; Eric Anderson (MS student, ECE), part of 2003-2004; Sergei Kojarski (PhD student, CCIS), part of Summer I, 2004

Activities: SCIRun is a software package for interactive computation and visualization developed by the Scientific Computing and Imaging (SCI) Institute, University of Utah, who are long-time collaborators of Professor Brooks. The SCIRun system is represented by a network of components, where each component is responsible for a part of the computation, resulting in a visual image. The network of components is defined recursively (that is, each component in the network can itself be a network). This package uses a dataflow paradigm, combined with sophisticated interactive visualization techniques and a modular software interface strategy, to provide a problem-solving environment for bioelectric field problems. Although this interface has a component-like flavor, it has not been designed according to careful component-based principles.

- We illustrated possible modifications of the data structures of current SCIRun modules, in order to increase their component-like behavior and enable new kinds of meta-data functionality. Matysiak has programmed initial “meta-module” wrappers, and has written adjunct documentation that was of interest to the SCI group and which has been used by Professor Lorenz in conjunction with student projects for his course on Component-Based Design. We have reached the point of posing useful component-based questions to the SCI design team, which may well help to influence their own Common-Component Architecture (CCA) standards-based revision of SCIRun.
- We illustrated possible construction of an interface between SCIRun and the CenSSIS toolboxes. The two packages have quite complementary strengths and Prof. Brooks and his collaborators expect to immediately make use of such an interface in their research once it is developed. A prototyped interface is currently being implemented by Danny Gagne, a CCIS undergraduate student, as part of a directed study with Professor Lorenz in Summer 2004.
- We gained a more solid understanding of issues involved with enabling scientists and engineers to connect, verify, and characterize computational solutions that combine existing software coming out of distinct research efforts, whose scientific functionality is complementary but whose software designs may not be immediately compatible.

Grant proposal resulting from this the project: Submission of an NSF grant proposal:

- David H. Lorenz, David R. Kaeli, Dana H. Brooks, and Gene Cooperman. Submitted 05/04; NSF; “A Foundation for an Aspect-Oriented Design Methodology”; Requested amount: \$483,616.

2 GRANT PROPOSALS

1. **David Lorenz** (CCIS, PI); co-PIs: Attie (CCIS), Brooks (ECE); NSF (submitted May, 2004): “Design Locality: A Concept for Controlling the Design Complexity of Large Software Systems”; \$547,353 (3 years)
2. **David Lorenz** (CCIS, PI); co-PIs: Cooperman (CCIS), Brooks (ECE), Kaeli (ECE); NSF (submitted May, 2004): “A Foundation for an Aspect-Oriented Design Methodology” \$483,616 (3 years)
3. **Gene Cooperman** (CCIS, PI); co-PIs: Aslam (CCIS), Kaeli (ECE), Ondrechen (Chemistry), Rappaport (ECE); senior personnel: Eric Miller (ECE), Steve Reucroft (Physics), Ravi Sundaram (CCIS), John Swain (Physics); NSF MRI (submitted Fall, 2003): “A Multi-Disciplinary Testbed for of High Performance Modeling Software”; \$300,000 (+ \$128,000 cost-sharing by N.U.; will obtain vendor discounts)
4. **David Kaeli** (ECE, PI); coPIs: Cooperman (CCIS), Kilmer (Tufts U.); NSF (submitted Jul., 2003; notified of acceptance Dec., 2003; submitted Jan., 2004): “Collaborative Research: Tuning Libraries to Effectively Exploit the Memory Hierarchy”; \$300,000 (“street value” of \$600,000)
5. **David Kaeli and Dana Brooks**; NIH (submitted Fall, 2003; declined, but reviews encouraged a resubmission, which is underway): “An Integrated Content-Indexed Database System for Distributed Access, Mining, Processing and Visualization of Biomedical Datasets”; \$2,563,257
6. **Mary Jo Ondrechen** (Chemistry, PI); co-PIs: Bob Futrelle (CCIS), Ron Williams (CCIS), and Dagmar Ringe (Brandeis University); NSF (submitted Jan., 2004): “ITR: Machine Learning and THEMATICS for Large Scale Protein Function Prediction”; \$3,938,779 (5 years)
7. **Mary Jo Ondrechen** (Chemistry, PI); co-PIs: Bob Futrelle (CCIS), Ron Williams (CCIS), and Dagmar Ringe (Brandeis University); NIH (submitted Spring, 2004): “Prediction of Active Sites from Protein Structure” \$2,285,012 (4 years)

3 PUBLICATIONS ACKNOWLEDGING ICSS SUPPORT

(in alphabetic order)

1. Attie P.C. and Lorenz D.H.: “Correctness of Model-based Component Composition without State Explosion.” In Proceedings of the ECOOP 2003 Workshop on Correctness of Model-based Software Composition. Darmstadt, Germany, July 22, 2003. <http://www.ccs.neu.edu/home/lorenz/papers/cmc03/>
2. Ernst E. and Lorenz D.H.: “Aspects and Polymorphism in AspectJ.” In Proceedings of the 2nd International Conference on Aspect-Oriented Software Development (AOSD), pages 150–157, Boston, Massachusetts, March 17-21, 2003. ACM. <http://www.ccs.neu.edu/home/lorenz/papers/aosd2003polyspect/>
3. Kojarski S. and Lorenz D.H.: “Unplugging Components using Aspects.” In Proceedings of the ECOOP 2003 Eighth International Workshop on Component-Oriented Programming. Darmstadt, Germany, July 21, 2003. <http://www.ccs.neu.edu/home/lorenz/papers/wcop03/>
4. Kojarski S. and Lorenz D.H.: “Domain Driven Web Development With WebJinn.” In Special Track on Domain Driven Development, International Conference on Object-Oriented Programming, Systems and Applications (OOPSLA), pages 53–65, Anaheim, California October 26-30, 2003, ACM Press. <http://www.ccs.neu.edu/home/lorenz/papers/oopsla03b/>
5. Kojarski S., Lieberherr K., Lorenz D.H., and Robert Hirschfeld: “Aspectual Reflection.” In Proceedings of the AOSD 2003 Workshop on Software-engineering Properties of Languages for Aspect Technologies, Boston, Massachusetts, AOSD 2003, March 17-21, 2003. <http://www.ccs.neu.edu/home/lorenz/papers/splat03/>
6. Lorenz D.H. and Vlissides J.: “Pluggable Reflection: Decoupling Meta-Interface and Implementation.” In Proceedings of the 25th International Conference on Software Engineering (ICSE), pages 3–13, Portland, Oregon May 3-10, 2003. IEEE Computer Society. <http://www.ccs.neu.edu/home/lorenz/papers/icse03/>
7. L.F. Murga, Y. Wei, P. Andre, J.G. Clifton, D. Ringe and M.J.Ondrechen, “Physicochemical Methods for Prediction of Functional Information for Proteins,” *Israel Journal of Chemistry*, in press.
8. Ng D., Kaeli D.R., Kojarski S., and Lorenz D.H.: “Program Comprehension Using Aspects.” In Proceedings of the ICSE 2004 Workshop on Directions in Software Engineering Environments (WoDiSEE’2004). <http://www.ccs.neu.edu/home/lorenz/papers/wodisee04/>
9. Ouarraui C. and Kaeli D., “An Object-Oriented Parallel Library,” *International Journal of High Performance Computing and Networking*, accepted for publication, to appear 2004.
10. Ouarraui C. and Kaeli D., “Developing an Object-oriented Parallel Iterative-Methods Library,” *Proceedings of Workshop on Hardware/Software Support for High Performance Scientific and Engineering Computing*, September 2003, pp. 8-15.
11. M.J. Ondrechen, “Identifying Functional Sites Based on Prediction of Charge Group Behavior,” *Current Protocols in Bioinformatics*, in press.
12. D. Ringe, Y. Wei, K.R. Boino and M.J. Ondrechen, “Protein Structure to Function: Insights from Computation,” *Cellular and Molecular Life Sciences*, 61, 387-392 (2004).

13. I.A. Shehadi, A. Uzun, L.F. Murga, V. Ilyin and M.J. Ondrechen, "THEMATICS is Effective for Active Site Prediction in Comparative Model Structures," *Conferences in Research and Practice in Information Technology*, 29, 209-215 (2004). (Only 34 out of 118 submitted papers were accepted.)
14. I.A. Shehadi, A. Abyzov, A. Uzun, Y. Wei, L.F. Murga, V. Ilyin, and M.J. Ondrechen, "Active Site Prediction for Comparative Model Structures with THEMATICS," *Journal of Bioinformatics and Computational Biology*. In Press.
15. Wang Y. and Kaeli D., "Profile-guided File Partitioning on Beowulf Clusters," *Journal of Cluster Computing*, Special Issue on Parallel I/O, accepted for publication, to appear 2004.
16. Wang Y. and Kaeli D.R., "Profile-Guided I/O Partitioning," *Proceedings of the 17th ACM International Symposium on Supercomputing*, June 2003, pp. 252-260. (acceptance rate: 21%)
17. Wang Y. and Kaeli D., "Source Level Transformations to Apply I/O Data Partitioning," *Proceedings of the IEEE Workshop on Storage Network Architecture And Parallel IO*, Oct. 2003, pp.12-21. (acceptance rate: 48%)

4 PATENT FILING

1. THEMATICS: A Simple Computational Method for the Identification and Characterization of Enzyme Active Sites," Mary Jo Ondrechen, James G. Clifton, Dagmar Ringe, Ronald J. Williams, Robert P. Futrelle, Wenxu Tong, David E. Budil, Jaeju Ko, Leonel F. Murga, Ihsan A. Shehadi, Huyuan Yang, U.S. Utility Patent Application, pending.

5 MANUSCRIPTS

(drafts far enough along and close enough to submission.)

1. J. Ko, L.F. Murga, P. Andre, H. Yang, M.J. Ondrechen, R.J. Williams, A. Agunwamba and D.E. Budil, "Statistical Criteria for the Identification of Protein Active Sites Using Theoretical Microscopic Titration Curves." Submitted.
2. W. Tong, R.J. Williams, R.P. Futrelle and M.J. Ondrechen, "Machine Learning and THEMATICS for Prediction of Protein Functional Sites." In preparation.
3. Paul C. Attie and David H. Lorenz: "Establishing Behavioral Compatibility of Software Components without State Explosion." In preparation.
4. Paul C. Attie and Hana Chockler. "Efficiently verifiable sufficient conditions for deadlock-freedom of large concurrent programs." Submitted.
5. Paul C. Attie. "Finite-state concurrent programs can be expressed pairwise." Submitted.
6. Paul C. Attie. "Synthesis of Large Dynamic Concurrent Programs from Dynamic Specifications." Submitted.
7. Paul C. Attie and Nancy A. Lynch. "Dynamic Input/Output Automata: a Formal and Compositional Model for Dynamic Systems." Submitted.
8. Paul C. Attie and Hana Chockler. "Pairwise representation and verification of quorum based distributed algorithms in message-passing systems." Submitted.

9. Jacob Burkhart and David H. Lorenz. “The Road to Aspect Cocoa: Developing an Aspect-Oriented Programming Extension for Objective-C”. In preparation.
10. David H. Lorenz and Therapon Skotiniotis. “Conaj: Generating Contracts as Aspects”. In preparation.
11. David H. Lorenz and Therapon Skotiniotis. “From Contracts to Aspects and Back”. In preparation.

6 MAJOR CONFERENCE PRESENTATIONS

(all include ICSS affiliation and acknowledge ICSS support)

1. Gene Cooperman, Xiaoqin Ma and Viet Ha Nguyen, “Static Performance Evaluation for Memory-Bound Computing: the MBRAM Model” (accepted as full paper in proceedings, to be presented in discussion section, 2004 International Conference on Parallel and Distributed Processing Techniques and Applications; acceptance rate from 2004 conference approximately 1 in 4 for all papers (full and short))
2. Gene Cooperman and Viet Ha Nguyen, “Marshalgen: Marshaling Objects in the Presence of Polymorphism” accepted as full paper, Internet Computing '04, June, 2004, acceptance rate from 2003 conference approximately 1 in 3 for all papers (full and short))
3. Gene Cooperman and Viet Ha Nguyen, “Memory-Based and Disk-Based Algorithms for Very High Degree Permutation Groups,” published as full paper, Proc. of International Symposium on Symbolic and Algebraic Computation (ISSAC '03), ACM Press, 2003, pp. 66–73; premier conference in symbolic algebra; acceptance rate varies between 1 in 2 and 1 in 3.
4. Robert P. Futrelle, Mingyan Shao, Chris Cieslik and Andrea Grimes, “Extraction, layout analysis and classification of diagrams in PDF documents,” ICDAR-2003 (Intl. Conf. on Document Analysis & Recognition). (pp. 1007-1014) Edinburgh, Scotland.
5. Robert P. Futrelle, Andrea Grimes and Mingyan Shao, Extracting structure from HTML documents for language visualization and analysis. In WDA-2003 (Workshop on Web Document Analysis). (pp. 3-6) Edinburgh, Scotland.
6. D. Ng, S. Kojarski, D. Lorenz and D. Kaeli, “Program Comprehension Using Aspect,” Proceedings of WoDiSEE, May 2004, to appear.
7. Mary Jo Ondrechen, “Physicochemical Methods for Protein Function Prediction,” Plenary lecture, 2003 Computational Chemistry Conference, Lexington, KY, October 20, 2003. (This is a national conference with the lectures also heard via the Access Grid at about 20 other sites worldwide.)
8. I.A. Shehadi, A. Uzun, L.F. Murga, V. Ilyin and M.J. Ondrechen, “THEMATICS is Effective for Active Site Prediction in Comparative Model Structures,” Asia Pacific Bioinformatics Conference 2004, Dunedin, New Zealand, January 21, 2004.
9. Wenxu Tong, Ronald J. Williams, Robert P. Futrelle and Mary Jo Ondrechen “Machine Learning and THEMATICS for Protein Function Prediction,” Bioinformatics Gordon Research Conference, Oxford, England, August 26, 2003.
10. Attie P.C. and Lorenz D.H.: “Correctness of Model-based Component Composition without State Explosion.” In Proceedings of the ECOOP 2003 Workshop on Correctness of Model-based Software Composition. Darmstadt, Germany, July 22, 2003. <http://www.ccs.neu.edu/home/lorenz/papers/cmc03/>

11. Ernst E. and Lorenz D.H.: “Aspects and Polymorphism in AspectJ.” In Proceedings of the 2nd International Conference on Aspect-Oriented Software Development (AOSD), pages 150–157, Boston, Massachusetts, March 17-21, 2003. ACM. <http://www.ccs.neu.edu/home/lorenz/papers/aosd2003polyspect/>
12. Kojarski S. and Lorenz D.H.: “Unplugging Components using Aspects.” In Proceedings of the ECOOP 2003 Eighth International Workshop on Component-Oriented Programming. Darmstadt, Germany, July 21, 2003. <http://www.ccs.neu.edu/home/lorenz/papers/wcop03/>
13. Kojarski S. and Lorenz D.H.: “Domain Driven Web Development With WebJinn.” In Special Track on Domain Driven Development, International Conference on Object-Oriented Programming, Systems and Applications (OOPSLA), pages 53–65, Anaheim, California October 26-30, 2003, ACM Press. <http://www.ccs.neu.edu/home/lorenz/papers/oopsla03b/>
14. Kojarski S., Lieberherr K., Lorenz D.H., and Robert Hirschfeld: “Aspectual Reflection.” In Proceedings of the AOSD 2003 Workshop on Software-engineering Properties of Languages for Aspect Technologies, Boston, Massachusetts, AOSD 2003, March 17-21, 2003. <http://www.ccs.neu.edu/home/lorenz/papers/splat03/>
15. Lorenz D.H. and Vlissides J.: “Pluggable Reflection: Decoupling Meta-Interface and Implementation.” In Proceedings of the 25th International Conference on Software Engineering (ICSE), pages 3–13, Portland, Oregon May 3-10, 2003. IEEE Computer Society. <http://www.ccs.neu.edu/home/lorenz/papers/icse03/>