Anthony S. Kosky, Ph.D.

CONTACT INFORMATION

Home Address:	1010 Shattuck Avenue Berkeley, CA 94707
Phone:	(510) 499-1673
E-mail:	anthony_kosky@yahoo.com
Web:	http://www.anthonykosky.com

SUMMARY

My background is in computer science research and I have over twenty years' experience working with scientific data, particularly within the bioinformatics and life sciences industries. I have extensive experience both as a technical contributor to, and manager of large-scale commercial and research-oriented software development projects, including work in FDA-regulated environments. I have made significant contributions, both theoretical and practical, to the fields of computer science and bioinformatics, specializing in big data analytics, database systems and data integration problems, as well as laboratory support and infrastructure, and have had my work published in leading computer science and bioinformatics journals and conferences. I have experience in designing and implementing systems for analysis, exploration and management of molecular biology, clinical and genomics data from wide varieties of heterogeneous data sources, particularly gene expression, sequence and related data. I have managed large software engineering teams working on multiple projects under challenging circumstances, established software development processes and best practices, and have worked as an individual developer and in smaller development groups. I have excellent written communication and presentation skills, especially for technical documentation, and have extensive experience planning and running projects.

TECHNICAL SKILLS

- Programming: Extensive experience in C++ (including C++11, STL, Boost, multi-threading, etc), and Shell scripting (bash, awk, sed, etc.); moderate experience in Ruby (including Ruby on Rails), and R (including Bioconductor, creating R packages and S4 objects); some Python, Scala, etc.
- Tools: Build tools (CMake, make); debuggers (gdb, Visual Studio); CProf, Valgrind, static code analysis (Klocwork, Coverity), test frameworks (GoogleTest, CPPUnit, TestThat for R); continuous integration systems (Bamboo, Jenkins); source control (SVN, GIT); configuration management (Ansible).
- Frameworks and platforms: Ruby on Rails; Docker; Amazon Web Services; some exposure to SalesForce; Hadoop, Heroku.
- Databases: PostgreSQL, SQLServer, MySQL, some NoSQL, Lucene.
- Issue tracking and planning tools: JIRA, FogBugz, Microsoft Project.
- Experience implementing software development lifecycles and processes (Scrum, Kanban, etc.)

EMPLOYMENT

- February 2016 - January 2018: **GenomeDx Biosciences**
  Senior Director, Software Engineering

  GenomeDx is a company running a CAP/CLIA certified laboratory performing cancer diagnosis, using DNA microarrays, in order to help predict outcomes and plan treatment for cancer. Responsible for software engineering department at GenomeDx, including groups doing software development for internal systems, development and support of enterprise systems (primarily SalesForce), and software QA and regulatory compliance.

Projects include:

- *ART* (*Accessioning, Requisitioning and Tracking tool*). ART is a proprietary laboratory information system developed using the SalesForce Force.com platform. ART replaced a number of disconnected systems and adhoc processes. ART provides support for the full lifespan of an order within GenomeDx, including order submission, sample acquisition, laboratory processing, analysis and report generation, and billing. The system includes integration with GenomeDx's analytics pipeline (CIP), third party billing systems (XIFIN) and customer EMRs (through Elkay Bridge).
- C*IP* (*Cancer Informatics Pipeline*): CIP provides a pipeline for normalization and analysis of microarray data generated by GenomeDx's assay. Analyses include GenomeDx's *Decipher* scores and QC metrics. CIP is implemented as workflow of Docker images running on AWS, with data stored in a PostgreSQL database.
- *GRID* (*Genomic Resource Information Database*): GRID represents an infrastructure for developing new classifiers and research access to GenomeDx's reference set of 1.6M markers for each of over 50,000 cancer tissue samples. GRID includes a platform for running new classifiers, implemented in R or Python, across all samples, and a library of 190 established classifiers on all new samples. In addition, it provides bioinformaticians with access to 1.6M data points for all existing samples for ad hoc analysis in R, and provides automated report generation for studies and individual samples.

- July 2010 – January 2016: **Affymetrix Inc.**
  Manager, Scientific Programming

  Led group responsible for development and maintenance of Affymetrix Power Tools (APT) – a set of command-line tools developed in C++ for analysis of Affymetrix microarray data. Also responsible for management of ChAS (Chromosome Analysis Suite) applications – a JAVA-based visualization tool for data generated by Affymetrix cytogenetic arrays.

  Projects included:

  - *APT2*: (Project lead, architect and contributor). The Affymetrix Power Tools 2.0 framework was a replacement for the existing 1.* framework which had incurred a great deal of technical debt. Key features included:
    - Workflow architecture allowing reuse of processing nodes between applications.
    - Unified approach to parameter declaration and parsing.
    - Self-documenting applications
    - File-format agnostic file I/O - supporting HDF5, text and proprietary binary formats
    - Built in serial/resume features for debugging and test automation.
  - OncoScan analysis pipeline: (Project lead and contributor). OncaScan is a microarray platform designed for for whole genome copynumber and LOH analysis for oncology samples. The analysis pipelines for these arrays were the first pipeline developed from scratch with the APT2 framework.
  - CytoScan Copynumber analysis pipeline: (Project lead and contributor). CytoScan is a copynumber and LOH analysis microarray design. In order to add various improvements to the algorithms, the analysis pipeline was ported to the APT2 framework, which proved considerably more efficient than dealing with legacy code.
  - Axiom Genotyping analysis pipelines: (Project lead) The Axiom family of arrays targeted SNPs (single nucleotide polymorphisms) within the human genome, and were used for genotyping very large sample sets (thousands of samples). Since the amount of data involved was too large to fit in machine memory, considerable effort put into the organization of data on disk to ensure efficient processing.
  - ChAS 2.* and 3.0: (Project lead) ChAS (Chromosome Analysis Suite) provided a visualization tool for exploring copynumber variations and cytogenetic abnormalities in the human genome using data generated by Affymetrix cytogenetic arrays. Version 3.0 added a database backend to ChAS, allowing users to identify regions of interest and search the database for previously loaded samples with similar or overlapping copynumber variations.

- CytoScan IVD and ChAS DX: (Project lead and contributor). CytoScan IVD is a version of the Affymetrix cytogenetics array for diagnostic purposes, and was the first microarray assay approved by the FDA for genetic disease diagnostics. As such versions of the analysis pipelines and ChAS had to be developed in accordance with FDA regulatory requirements.
- Genotyping for UK Biobank: (Project lead). The UK Biobank ran 500,000 Affymetrix SNP arrays, and required Affymetrix to run genotyping algorithms on the data. Existing genotyping code needed to be adapted and ported to a Hadoop cluster in order to make the analysis feasible.

- November 2007 – June 2010: **DOE Joint Genome Institute**
  Group Lead, Scientific and Institutional Applications
  Lead group responsible for developing web-based systems using Ruby on Rails and Java (SEAM) to support the Project Management Organization of JGI, and for support of business and enterprise applications. Projects and tasks included:
  - Promoting improved software architecture using approaches such as Service Oriented Architecture (SOA), and improved software development practices such as Test Driven Development (TDD) and Agile software development.
  - *Work Initiation Process (WIP)* application, supporting the submission of sequencing project proposals, the scientific and technical review of proposals, notification of proposal acceptance or rejection, and populating internal project tracking databases with accepted proposals.
  - *Collaborator Sample Information (CSI)* application, supporting the submission of meta-data for samples, including taxonomy, bio-safety certification and QC images, prior to submission to JGI, and allowing for review of sample data, generation of shipping instructions and populating internal sample tracking databases.
  - *Single Sign-on and User Administration* system, providing support for user account creation and updates, and for authentication for multiple web-based applications within JGI.
  - *IMG-ACT,* a system supporting educational programs in microbial genome annotation, including support for assigning genome annotation tasks, managing classes, and collecting student annotations using a combination of internally developed web applications (Ruby on Rails) and wiki pages (Atlassian Confluence).
  - Supporting institutional and enterprise software applications including internally developed applications for employee management, Atlassian Confluence (wiki) and JIRA (issue tracking) and legacy purchase management and HR applications.
  - Support and development of JGI institutional and intranet web pages.

- October 2007: **Affymetrix Inc.**
  Consultant
  Porting analysis methods for copy number estimation to Linux environment as part of Affymetrix APT tools.

- June 2007 – November 2007: **Corimbia Inc.**
  Consultant
  Consulting and implementation on Laboratory Information Management System (*LIMS*) for Childrens Hospital Los Angeles. System implemented using Microsoft SQL Server and ASP.NET technologies.

- August 2005 – April 2007: **Axiope Inc.**
  Chief Technical Officer (August 2005 – April 2007)
  Consultant (March 2005 – July 2005)
  Strategic and technical planning and leadership for Catalyzer, an XML-based data management system for biomedical and scientific research applications with both web-based and thick-client interfaces. Responsible for working with management team to set technical and business directions, as well as leadership of software developers, development plans and many other tasks

as needed. Consulted on SBIR grant proposal for Data Transformation and Integration Tools to extend Axiope's *Catalyzer* product line.
Catalyzer is a general purpose data management system which has been used for Electronic Lab Notebook (ELN), Electronic Data Capture (EDC), Image Management, Inventory Management and Specimen Tracking applications. It provides object-oriented data modeling capabilities, a graphical user interface allowing non-technical users to rapidly create and evolve databases, and interfaces to various image and document file formats, scientific equipment such as confocal microscopes, bar-code scanners and printers.

- September 1997 – December 2004: **Gene Logic Inc.**
  ***Senior Director**, Enterprise Application Systems* (November 2003 – December 2004),
  ***Director**, Data Management Tools* (August 2001 – November 2003),
  Senior Group Leader (August 1998 – August 2001),
  Senior Computer Scientist (September 1997 – August 1998).
  I took on a wide variety of roles while working in the *Software Development* division of Gene Logic for seven years. These included:
  - Overall responsibility for overall project planning and lead on Gene Logic's *Genesis/GeneExpress* software systems (versions 1.2, 1.3, 2.0, 2.5, 2.6 and various maintenance releases). This included coordinating the various teams producing the different software components, working with marketing and higher management to plan monitor development, anticipating, assessing and dealing with technical issues, and so on. The project required synchronization or several subgroups including internal LIMS and database development, integration/data migration tools (*GX Connect*) and data analysis/query/exploration tools (*GeneExpress*). Also responsible for analysis engine/server development and defining analysis APIs for GeneExpress.
  - *ASCENTA 1.0 Software:* Overall responsibility for software development for initial release of ASCENTA product. ASCENTA is Gene Logic's "entry-level" gene-expression data product, providing access to aggregate gene expression and related data through web based interfaces
  - *Probe Intensity Analysis*: Project lead for Probe Intensity Analysis R&D project. This involves developing tools for rapid access of Affymetrix GeneChip probe-level data (i.e. CEL files) from various statistical and programming languages (R, S-plus, C++, Perl).
  - *Various Data Integration and Customization Projects*: Responsible for a variety of customization and professional services projects, including one data integration project that involved loading gene expression, gene annotation and sample annotation data from a custom LIMS system, with support for custom chips and incremental updates. Also responsible for certain customer interactions and providing technical support for custom projects.
  - *OPM Multidatabase Query System*: Responsible for development of Multidatabase Query System based on the Object Protocol Model (licensed from LBNL). Included design, development and maintenance of the multidatabase query engine, development of query servers for accessing Oracle and Sybase databases, XML/SQL hybrid databases, BLAST and other bioinformatics tools. The system was deployed at Smith-Kline Beecham and a number of academic and research institutions. The project also involved managing customer interactions and requests.
  - *Site manager* for Berkeley branch of Gene Logic Inc. (Nov 2003 – Oct 2004): 25 employees including software developers, Software QA testers, project managers, technical writers, managers and administrative assistants.

- September 1995 – September 1997: **Lawrence Berkeley National Laboratory**
  ***Research Scientist**.*
  Working in the *Information and Computer Science Division, Data Management Research and Development Group*, on the *Object-Protocol Model* (OPM) toolset.

- November 1988 – August 1989: **STC Technology Ltd., Harlow, England.**
  Software Engineer.
  Working in *Formal Methods Dept.* on *RAISE* (Rigorous Approach to Industrial Software Engineering) project – and ESPRIT funded project. Involved with the implementation of a language based editor and tool set for the *RAISE Specification Language.*

EDUCATION

- September 1989 – August 1995:
  Dept. of Computer and Information Science, University of Pennsylvania.
  Ph.D. in *Computer Science*. Thesis "Transforming Databases with Recursive Data Structures", supervised by Prof. Peter Buneman and Prof. Susan Davidson.

- October 1987 – September 1988:
  Dept. of Computing, Imperial College of Science and Technology.
  MSc./DIC (Distinction) in *Foundations of Advanced Information Technology.*
  Subjects studied include semantics of programming languages, domain theory, theory of functions, functional programming technology, logic programming, models of concurrent computation.
  Dissertation on "Semantics of Object Oriented Programming Languages", supervised by Prof. Samson Abramsky.

- October 1984 – June 1987:
  University of Kent at Canterbury.
  BSc. (First Class Honors) in *Mathematics.*
  Winner of 1987 Rotary and I.M.A. prizes.
  Subjects studied include real analysis, complex analysis, numerical and computational mathematics, topology, discrete mathematics, relativity theory, fractal geometry and complex analytic dynamics.

REFERENCES

Available upon request.

SELECTED PUBLICATIONS

A full list and links to all my publications are available at http://www.anthonykosky.com/bib.html.

- "Declarative Languages for Advanced Information Technology", *Journal of Information Technology,* Vol. 3, No. 2, June 1988.

- "A Formal Model for Databases with Applications to Schema Merging", in *Specification of Database Systems*, Glasgow 1991, Harper and Norrie (eds.).

- "A Basis for Interactive Schema Merging", (with P. Buneman, S. Davidson and M. VanInwegen), in *Proc. Hawaii International Conference on Systems Sciences*, 1992.

- "Theoretical Aspects of Schema Merging", (with P. Buneman and S. Davidson), in *Proc. Extending Database Technology (EDBT)*, Vienna, 1992.

- "Facilitating Transformation in a Human Genome Project Database", (with S. Davidson and B. Eckman), in *Proc. Third International Conference on Information and Knowledge Management (CIKM)*, Gaithersburg, MD, 1994.

- "Observational Distinguishability of Databases with Object Identity", in *Proc. 5th International Workshop of Database Programming Languages (DBPL5)*, Gubbio, Italy, 1995.

- "Exploring Heterogeneous Molecular Biology Databases in the Context of the Object-Protocol Model", (with V. M. Markowitz and I. A. Chen), in *Theoretical and Computational Methods in Genome Research*, Suhai, S. (Ed), Plenum Press, 1997.

- "Facilities for Exploring Molecular Biology Databases on the Web: A Comparative Study", (with V. M. Markowitz, I. A. Chen and E. Szeto), in *Proc. of the Pacific Symposium on Biocomputing*, Hawaii, January 1997.

- "WOL: A Language for Database Transformations and Constraints", (with S. Davidson), in *Proc. 13th International Conference on Data Engineering (ICDE)*, Birmingham, United Kingdom, 1997.

- "Semantics of Database Transformations", (with S. Davidson and P. Buneman), in *Semantics in Databases*, Springer Lecture Notes in Computer Science, 1998, Thalheim and Libkin (eds.).

- "Constructing and Maintaining Scientific Database Views" (with I. A. Chen, V. M. Markowitz and E. Szeto), in *Proc. of the 9th International Conference on Scientific and Statistical Database Management*, Olympia, WA, 1997, Hansen and Ioannidis (eds.).

- "Exploring Heterogeneous Biological Databases: Tools and Applications" (with I. A. Chen, V. M. Markowitz, and E. Szeto), in *Proc. of the 6th International Conference on Extending Database Technology (EDBT'98)*, Valencia, Spain, 1998.

- "Advanced Query Mechanisms for Biological Databases" (with I. A. Chen, V. M Markowitz, E. Szeto and T. Topaloglou), in *Proc. of the 6th International Conference on Intelligent Systems for Molecular Biology (ISMB'98)*, June 1998.

- "Object-Protocol Model Data Management Tools '97". (with Markowitz, V.M., Chen, I.A., and Szeto), in *Bioinformatics Databases and Systems*, Stan Letovsky (ed), Kluwer Academic Publishers, 1999, pp. 187-199.

- "Seamless Integration of Biological Applications within a Database Framework". (with Topaloglou,T. and Markowitz,V.M.), in *Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB'99)*. Heidelberg, Germany, June 1999.

- "Extending traditional query-based integration approaches for functional characterization of post-genomic data", (with B. Eckman and L. Laroco), in *Journal of Bioinformatics*, Vol. 17, No. 7, 2001, pp. 587-601.

- "Gene Expression Data Management: A Case Study", (with V.M. Markowitz and I.A. Chen), in *Proc. Eighth International Conference on Extending Database Technology"*, Prague, 2002.

- "Integration Challenges in Gene Expression Data Management". (with Markowitz V.M., Campbell, J., Chen, I.A., Kosky, A., Palaniappan, K., and Topaloglou, T.), to appear as a chapter in *Bioinformatics: Managing Scientific Data*, Morgan Kauffman / Elsevier Science, May 2003.