JAVA FOR BIOINFORMATICS AND BIOMEDICAL APPLICATIONS

JAVA FOR BIOINFORMATICS AND BIOMEDICAL APPLICATIONS

by

Harshawardhan Bal
Booz Allen Hamilton, Inc., Rockville, MD

and

Johnny Hujol
Vertex Pharmaceuticals, Inc., Cambridge, MA

Springer

Library of Congress Control Number: 2006930294

ISBN-10: 0-387-37235-0 e-ISBN-10: 0-387-37237-7
ISBN-13: 978-0-387-37237-8

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springer.com

Contents

Foreword

April 2006

Introduction

Bioinformatics is at a crossroads. We work in a field that is changing every day, increasingly moving from specific solutions created by single researchers working alone or in small groups to larger, often geographically dispersed programs enabled by collaborative computing and open software. This book represents an important development, giving the reader an opportunity to discover how the use of open and reusable Java code can solve large bioinformatics problems in a software engineered and robust way. I work with one of the authors of this book every day, on the National Cancer Institute's cancer Biomedical Informatics Grid (caBIG™) project, and I can attest that they are well suited to share with their readers both their experience in the development and use of bioinformatics software, as well as their interest in solid software engineering and interoperability.

Background and history

In its short history, bioinformatics has become an increasingly important part of how scientists involved in biological research go about their work. This has lead to an explosion of interest in the subject, and a similar explosion in tools and data resources for researchers to learn and use in their work. Historically, tools for bioinformatics have been idiosyncratic and are custom-developed by the end-users (or those close to them) in an iterative fashion until the specific immediate problem is solved. This has led to a balkanization of informatics systems, sometimes yielding multiple, incompatible systems at a single institution for a single application. This trend is beginning to change, with groups throughout the research community developing standards and shared data models, in areas ranging from gene expression arrays to pathways and proteomics. With a range of emerging software capabilities and a growing interest in interoperable tools and standards, bioinformatics practitioners have an ever-expanding toolbox from which to draw on to develop the basic software infrastructure behind their work. Similarly, with the increasing interest within the biomedical informatics community in the use of well-defined software engineering methodologies, and disciplines like design patterns and model-driven architecture, the software developed there will increasingly last longer, be easier to maintain, foster interoperability and reuse, and ultimately be more robust and cost effective.

Interfaces and standards

Interfaces and standards, as well as the use of well established development platforms, especially object-oriented programming, allow the bioinformatics practitioner to solve problems faster, with fewer lines of reusable, well-documented code than before. Through access to and study of well-established principles of software engineering and computer science, the solutions to problems in biomedical informatics will also be solid and optimally designed. With the increasing size of the datasets used in biomolecular informatics, derived from all manner of new high-throughput technologies and online databases, it is increasingly important to use thoughtful, efficient and well-established algorithms in the analysis of that data. Informatics students who can decompose complex, biologically significant informatics problems into simpler models, for which there are corresponding, validated and pre-existing software objects, will be amply rewarded for their efforts. It is by building on well-supported software platforms, using established and tested methodologies, that the most favorable balance can be achieved between effort and benefit.

Java as a platform

This book will teach you ways to make use of the Java programming language as a platform for your work in biomedical informatics, and in doing so, will open you up to the possibility of using a wide range of software objects in use throughout the large software engineering and computer science communities. Java is, of course, not the only object-oriented platform that is appropriate for bioinformatics. Perl is very well established, and are python, C++ and many others. The lessons that you can learn in Java are transferable to any object-oriented system, and Java is proving to be a solid platform for work throughout the informatics community. In the caBIG™ project that both Harshawardhan and I are a part of, Java is one of the main (but far from the only) programming languages used in that project. As a result, there is a lot of infrastructure available in the form of open-source code and open-content resources that are available for the busy researcher, serious student, or interested hobbyist. The latter chapters in this book detail how to connect with and make use of those resources to solve your own informatics programs.

The future

Through the efforts of a global community of biomedical informatics researchers, and through the prevalence of the Internet, it has become possible for any interested person to learn enough about biology, software engineering, and computer science, to contribute meaningfully to the emerging science of informatics. With the amount of openly available raw biological data growing by leaps and bounds every day, there is every reason to believe that you can contribute too, and the book that you hold in your hand is a great way to join in. Bon voyage!

Mark Adams
Program Manager
NCI Cancer Biomedical Informatics Grid (caBIG™)
Booz Allen Hamilton
Rockville, MD

Preface

On April 15, 2003, the International Human Genome Sequencing Consortium (IHGSC) – an association of laboratories from around the world which had jointly undertaken the Human Genome Project formally announced the completion of the colossal task they had set out to accomplish: the sequencing and assembly of the 3 billion bases that comprise the human genome. This was a truly landmark achievement for science and medicine. Today, the word "genome" has become a household term and together with bioinformatics has revolutionized how we approach biomedical research. The human genome project has led to identification of thousands of disease genes and paved the way for the development of newer drugs and treatments. Undoubtedly, the sequencing of the human and other genomes is just the beginning of the revolution that is unfolding right in front of our eyes. We are moving towards a paradigm shift in medicine, from just-in-time treatment that is given after the onset of symptoms to predictive and personalized treatment where the determination of the genetic factors predisposing an individual to disease is made right at birth and treatment started much before the onset of disease.

There is also a fundamental shift in how biomedical research is going to be conducted and funded in the years to come, especially, in areas such as cancer research and heart disease where there is a critical need to bring newer and better treatments for patients. Cancer has passed heart disease as the number one killer in UK and US and has been recognized by the World Health Organization as a major health problem across the globe. To meet this challenge, the US National Cancer Institute (NCI) has launched the biggest collaborative research program in 2003 called the cancer Biomedical Informatics Grid (caBIG™). In the words of NCI Director, Dr. Andrew von Eschenbach, "...caBIG will become the 'World Wide Web' of cancer research informatics and will accelerate the development of exciting discoveries in all areas of cancer research". Thus started the journey towards the NCI Challenge Goal, "To eliminate the suffering and death due to cancer by 2015" and together with it the efforts of more than 50 NCI-designated cancer centers, scores of research laboratories, Universities and public and private institutions across the country.

Where does J2EE come in the picture? The healthcare and medical research enterprise that we see today with its complex distributed Internet-enabled architecture is dependent on technologies that provide the critical infrastructure components necessary to fulfill its patient data safety, security and regulatory compliance requirements. Java has emerged as a powerful programming language for developing secure, scalable and robust web-enabled applications and is particularly well suited for building the many interrelated components of the geographically dispersed biomedical research and business engine. Together with support from a number of open source standards, J2EE offers a number of advantages for such applications and is the major platform for development efforts under caBIG™.

Why now?

We were confronted with this question early on in the writing of the book. The answer lies in the way the biomedical research enterprise has been transforming itself over the past decade or so and in doing so, promising to revolutionize the way we provide patient care. caBIG™ is based on the principles of open source, open access, open development and federation and uses J2EE and open source technologies for all software development efforts under the program. CaBIG™ is perhaps the next major landmark in the making in the history of biomedical research. Consequently, the time for a closer look at J2EE and open source technologies in a way that combines industry standard software engineering and design principles, genomics, bioinformatics and cancer research, is ripe.

This book is an attempt to fill that critical need. The main differentiating feature of the book is its focus on creating and integrating practical, useful tools for the scientific community in the context of real-life, real-value biomedical problems that researchers encounter on a routine basis. The book leverages technologies for molecular biology, genomics, bioinformatics, clinical research and cancer research developed by the National Cancer Institute Center for Bioinformatics (NCICB), the National Center for Biotechnology Information (NCBI, a division of the

National Library of Medicine (NLM) at the NIH), and scores of research organizations across the nation.

The book begins with an overview of the state of biomedical research today and the challenges it faces due to the silo model that has perpetuated over decades across universities and research centers across the world. It establishes a case for and the rationale behind the current move towards integrative, collaborative and standards based research platform through an introduction to the NCI caBIG™ program. It next provides an overview of emerging architectural trends such as Web Services and Service-Oriented Architecture. The book is not as much about the J2EE platform as it is about its *application* to building useful software and does not dwell on the theoretical aspects of the language or the platform; the authors (as well as the readers) recognize that several excellent works on that topic already exist. Instead the uniqueness of this book is that after just a short introduction, it takes a deep dive into demonstrating how to build highly functional graphical user interfaces for common and widely used bioinformatics tools that most researchers are familiar with and find indispensable for any kind of research activity. The reader is led through a step-wise and incremental software development approach with two goals in mind – to demonstrate a systematic standard software engineering approach to application development and, to activate a thoughtful design process in the mind of the developer that is aimed at exploring ways to enhance the functionality and usefulness for end-users. The applications that are considered the backbone of modern genomic and bioinformatics-driven research – Basic Local Alignment Search Tool (BLAST), Genscan gene prediction tool and others are used to illustrate this process. The reader will notice a significant amount of code in this book and realize that this is so by design. Although there are many ways of architecting a solution for a particular problem, we have illustrated one such approach while encouraging users to build their own. In doing so, we have also attempted to promote the reuse of tried and tested code from existing software libraries based on open source projects such as Apache, BioJava, caBIG™, and others.

Another differentiating feature of the book, best described by a reviewer, is we "...take a gradual and applied approach to combining Java and Bioinformatics". This statement, in fact, represents the very fabric of our strategy. By the same design, we have devoted little time on describing features and individual programming elements for which excellent and easily accessible documentation already exists. Our approach has also been to create pipelines where two applications are combined together along logical workflows that researchers normally use in their research environments to produce an enhanced application that has more utility than the individual applications.

The book does not profess to be the comprehensive tome on J2EE; instead, it is designed to cover a few of the important topics that lend themselves to use in the situations that are commonly encountered in this domain. It is hoped that a more focused approach would lead to a better and clearer understanding of the core capabilities of the platform than would be achieved by a lengthier treatment of the subject that cover all its different aspects. Indeed, the vastness and the complexity of the biomedical space and the pace and profundity with which science, technology, policy and legislation affect it is at times daunting. The authors acknowledge the challenge of writing on a topic this difficult and hope to address the concerns of the readers of this volume to identify gaps and produce a more inclusive title while providing time for the emerging technologies described in this book and others beyond the scope of this book to mature and gain wider acceptance by the user community.

With this background in mind, the book is especially tailored towards graduate students majoring in computer science, or information technology and who intend to take up careers in architecting software solutions for biomedicine and healthcare. It is also meant for practicing professionals who are actively involved in developing, maintaining or enhancing biomedical software and need to remain on the cutting edge of trends and standards in medicine and information. Finally, it will also be useful to molecular biologists, life scientists and clinicians who have a strong commitment towards understanding how software technologies can be put to use in solving the unique demands presented by the modern post-genomic translational research landscape.

This work would not be possible but for the many people who helped us get our thoughts together and organized to this point. We thank the many initial reviewers of this book who represent both private as well as public companies and research organizations including thought leaders in the field, many of whom are closely associated with the latest movements in information and biomedical technologies, and in their application to initiatives such as caBIG™. We thank Dr. Mark Adams, the caBIG™ Program Manager, for his wholehearted support for the book from concept to conclusion and for lending his expert insight into the future of biomedicine as captured in the Foreword for this book. We thank the good people at Springer – especially, Joseph Burns and Marcia Kidston and their team – for sticking with us throughout the process and coming to our assistance whenever we had the slightest of troubles. We also thank our individual families – the grown-ups (our wives) Nathalie Hujol and Snehal Bal, and not so grown-up (Arnav Bal, just 3 at the time of this writing), who knowingly or unknowingly – but by no means reluctantly – allowed us both to pursue this adventure and leave the life outside our small world for the better part of the 2005-2006 to flourish without our intercession for the most part.

To all our readers – whether you are an end-user or a developer, a biologist, a clinician or a bioinformatician or, indeed, one of the many documented cross-disciplinary "hybrid professionals" - we hope this book serves the small but meaningful purpose we began with in our minds and that it provides a vignette into the fast and exciting world of biomedical research. We value your feedback and will continue to incorporate your suggestions and work hard to meet your expectations in partnership with you throughout the lifetime of this book. We hope to hear from you!

Bon chance and bonne journee.

Harshawardhan Bal
Johnny Hujol

April 2006