Refatoração do Arcabouço de Modelos Probabilísticos ToPS

Parte Subjetiva

Renato Cordeiro Ferreira 7990933

31/01/16

A parte subjetiva do Trabalho de Formatura Supervisionado tem como objetivo dar uma visão pessoal - e não técnica - do processo de desenvolvimento do projeto da disciplina. Por esse motivo, neste texto, tomarei a liberdade de usar a primeira pessoa do singular, de modo a expressar meu ponto de vista sobre os desafios e aprendizados do trabalho.

1 Histórico

A ideia deste projeto começou muito antes da disciplina de trabalho de formatura. Ainda em 2012, meu ano de ingresso na USP, comecei a fazer uma iniciação científica com o Prof. Alan Durham, após assistir uma palestra sobre Bioinformática, ministrada por ele, no IV Encontro do BCC. Após realizar algumas leituras sobre biologia molecular, comecei a trabalhar no projeto de validação do preditor de genes MYOP - que durou do início de 2013 até o final de 2014 (com uma bolsa do CNPq entre o meio desses dois anos). Esse período foi, para mim, muito desafiador. No curso do BCC, são raros os casos de alunos que entram numa IC tão cedo, e mais raros ainda os que se mantém num mesmo grupo de pesquisas por tanto tempo. Embora não tenha sido fácil conciliar a carga do currículo com o trabalho científico, o esforço foi recompensador: tive a oportunidade de conhecer mais sobre a pós graduação desde cedo, e pude trabalhar com várias tecnologias muito antes de vê-las no curso.

Nessa época, ouvi falar, pela primeira vez, sobre o ToPS. Como eu possuía poucos conhecimentos sobre probabilidade e programação, era muito difícil, para mim, entender qual era a ideia do sistema. Sabia que ele era feito na linguagem C++ - que eu desconhecia - e que tinha modelos para resolver alguns problemas de genômica e proteômica - que eu entendia superficialmente. Ele era citado nas reuniões semanais do nosso grupo de pesquisas, mas eu não tinha contato direto com o sistema. Embora eu pudesse ter simplesmente ignorado o assunto, minha curiosidade foi mais forte. Aprendi sobre o arcabouço em minhas conversas com meu colega e amigo Ígor Bonadio - com quem dividia a companhia no metrô até a zona leste de São Paulo. Desde o início do seu mestrado, o Ígor trabalhava fazendo extensões e correções no ToPS. Nas nossas discussões, compartilhou comigo as dificuldades de realizar essas tarefas. Justamente numa dessas conversas - agora, no início do meu 6º semestre de curso - surgiu a ideia de fazer uma refatoração. Eu acabara de ler o livro de padrões de projeto da *Gang of*

Four, e estava começando a disciplina de Programação Orientada a Objetos. Oportunamente, eu e Ígor estávamos inscritos, e tínhamos de escolher algum tema para apresentar um seminário. Escolhemos, então, discutir sobre o livro de refatoração do Martin Fowler - uma das primeiras e maiores referências no assunto. Este foi o primeiro passo, tomado conscientemente, para o realizar o que se tornaria o trabalho de conclusão de curso.

Ao discutir a ideia para o trabalho com meu orientador, a proposta feita por ele foi um pouco diferente. Embora ele tenha aceitado, sem problemas, a ideia de fazermos a refatoração, sugeriu que fizéssemos mais. Há alguns anos, ele tinha interesse em implementar no arcabouço o Modelo de Covariância - criado por Richard Durbin e Sean Eddy para fazer a descrição da estrutura secundária de sequências de RNA, e queria que eu o fizesse no trabalho. Após considerar a proposta por algum tempo, aceitei-a, e passei a usar meu tempo livre para estudar o livro *Biological Sequence Analysis* (sobre modelos probabilísticos markovianos) criado pelos dois autores. Convenientemente, nessa época (6º semestre do curso), tive a oportunidade de cursar três disciplinas que me ajudaram, diretamente, a entender os aspectos técnicos do trabalho: Programação Orientada a Objetos (citada acima), Engenharia de Software e Linguagens Formais e Autômatos. No início de 2015, finalmente, comecei o trabalho de formatura.

A primeira proposta de trabalho seguia três ideias: refatorar o ToPS, para deixar o sistema mais amigável para extensões; adicionar o modelo de covariância, com seus algoritmo de treinamento e inferência; e, se possível, fazer alguns testes para validar a implementação, comparando-o com os programas similares da área. No início fazer todas essas tarefas parecia muito factível, e o primeiro título dado ao trabalho foi "Implementação de Modelos de Covariância no arcabouço ToPS".

Por conta da curva de aprendizado para entender o sistema, comecei a fazer programação pareada com meu colega Ígor, ainda em 2014. Assuntos como testes de unidade, integração contínua, desenvolvimento dirigido a testes, qualidade de software, etc., eram termos que eu ouvira em poucas conversas informais no IME. A ajuda do Ígor, nesse período e ao longo do desenvolvimento do projeto, foi muito importantes para que eu aprendesse sobre esses conceitos - posteriormente incorporados na parte teórica da minha monografia.

De todos os meus conhecimentos aprendidos e que foram relevantes para desenvolver o projeto, apenas um não comentado: a linguagem C++. Nos meus primeiros anos do curso, tomei conhecimento sobre a existência da linguagem em palestras e *workshops* da Maratona de Programação e do USPGameDev. Para aprender sobre ela, nas férias de verão de 2013, resolvi usar meu tempo livre para reimplementar, em C++, o jogo feito na disciplina de Laboratório de Programação. Embora eu não tenha continuado o desenvolvimento após essas férias, comecei a programar em C++, e passei a usar a linguagem em EPs de várias disciplinas do 3º e 4º ano de BCC. Para conhecer a sintaxe e os recursos da linguagem, li o livro *The C++ Programming Language*, de Bjarne Stroutsput - criador do C++. Com crescimento do meu interesse, fui me aprofundando, e parti para a leituras mais avançadas, como a série *Effective C++*, de Scott Meyers, e títulos específicos como *C++ Coding Standards*, *C++ Concurrency in Action*, *C++ Hacker's Guide* e, recentemente, *Modern C++ Programming with Test-Driven Development*. Meus estudos e as horas de programação fizeram do C++ a linguagem em que mais me especializei. Todos esses conhecimentos foram essenciais para a refatoração do ToPS, e permitiram que eu criasse a nova arquitetura do sistema.

No repositório do jogo criado nessas férias, comecei a implementar um pequeno Makefile para compilar o código do programa. Graças a alguns desafios que encontrei ao longo do desenvolvimento (criar vários executáveis, gerar documentação, produzir bibliotecas), fui incrementando-o até que ele se tornou um projeto próprio: o All-in-One Makefile. Os recursos que implementei nesse projeto foram importantes para que, no trabalho de conclusão de curso, eu pudesse substituir o CMake como a ferramenta de compilação do ToPS. O trabalho no arcabouço, por outro lado, também ajudou meu projeto de software livre, que cresceu e amadureceu graças às melhorias que implementei pensando em facilitar o seu uso no ToPS.

2 Desafios

Certamente, o maior desafio do projeto foi decidir o que poderia ou não ser feito. Uma refatoração, como técnica para melhorar a qualidade de um programa, pode ser aplicada de forma contínua e ininterrupta. No nosso trabalho, porém, iríamos parar com todas as modificações sendo feitas no arcabouço, para, então, arrumá-lo. Como outros trabalhos do nosso grupo de pesquisa dependem do ToPS, esse processo precisava ser feito com um fim bem definido - quando, então, voltaríamos com as implementações. No próprio trabalho, havíamos nos comprometido a implementar o Modelo de Covariância, que precisaria de um tempo próprio de desenvolvido. Portanto, desde o início, era importante definir até onde iria a refatoração.

Com ajuda do meu colega Ígor, decidimos que portaríamos apenas os modelos relacionados à predição de genes disponíveis na versão publicada do ToPS. Dessa maneira, economizaríamos tempo, e poderíamos colocar o restante dos modelos após o término do trabalho. Nos primeiros meses do projeto, então, dediquei-me a estudar sobre o modelo de covariância, enquanto realizava as primeiras modificações de repositório e código com auxílio do Ígor. Quando ele, finalmente, trouxe o último modelo que portaríamos (o Modelo Oculta de Markov Generalizado), pude começar a mexer nos componentes e na hierarquia do sistema.

Neste ponto do trabalho de conclusão de curso, o ritmo das implementações começou a diminuir: a arquitetura de *front-ends* foi montada em um processo iterativo, e várias propostas foram feitas antes que chegássemos ao mecanismo final. Em termos de implementação, esse foi o desafio mais complexo do projeto, e exigiu que eu melhorasse muito meus conhecimentos sobre metaprogramação em C++. Em particular, os dois últimos *front-ends* feitos foram os mais complexos. Para criar o design dos treinadores, foi necessário mais de um mês de planejamentos e experimentos. Para a criação dos serializadores, a quantidade de código que precisava ser alterada era muito grande. Quando, finalmente, chegou o mês de Agosto - no qual, segundo o planejamento original, a refatoração deveria estar pronta - vi que era necessário decidir se eu implementaria, ou não, o Modelo de Covariância.

Como o trabalho da refatoração estava apresentando bons resultados, e ainda não havia nenhum código para o novo modelo, decidi, junto ao meu orientador, focar na refatoração e em fazer uma análise mais profunda - como software - do arcabouço. Trocamos, então, o título do trabalho para a sua versão final: "Refatoração do arcabouço de modelos probabilísticos ToPS". Essa mudança, feita no início de Setembro, configurou um grande desafio: precisei pesquisar e estudar a literatura específica, e não pude aproveitar os estudos, feitos no início do ano, sobre Modelos de Covariância.

3 Aprendizados

Apesar da mudança de foco do trabalho, pude aprender, neste trabalho de conclusão de curso, muito sobre o Modelo de Covariância. As leituras feitas antes e no início do projeto ajudaram a amadurecer meus conhecimentos sobre probabilidade, e permitiram que eu conhecesse um pouco sobre toda uma subárea da aprendizagem de máquina. Por ser aplicado em sequências de RNA, pude descobrir mais sobre a área de transcriptômica, e compreender sobre alguns mecanismos de funcionamento das células que são relevantes para as modelagens realizadas pela Bioinformática.

Da parte do software, a experiência de mexer com o ToPS me ajudou a, finalmente, entender o funcionamento de testes de unidade, suas vantagens para a manutenção de um programa e o que é necessário para fazê-los. Descobri diversas ferramentas e bibliotecas interessantes, que passei a aplicar no código do ToPS e que levarei para outros projetos em C++. Tive a oportunidade de aplicar várias práticas ágeis de desenvolvimento (programação pareada, revisão de código, organização de tarefas em *kanban*) e sentir a vantagem de utilizá-las. Tornei-me, assim, um programador muito mais experiente, e que é capaz de ligar com um projeto grande, com muito código e que deve ser mantido a longo prazo

Por fim, sobre a perspectiva pessoal, o desenvolvimento do ToPS me ajudou a confirmar meu gosto pelo trabalho científico - que se originou desde a minha iniciação científica começada no início do curso. Por conta desse projeto, decidi fazer um mestrado (com o mesmo orientador) na área de Bioinformática, com a ideia de seguir com os estudos sobre modelos probabilísticos para RNA, implementar o Modelo de Covariância e, se possível, seguir para um doutorado na área.