Alinhamentos e comparação de sequências

Francisco Eloi Soares de Araujo

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Ciência da Computação

Orientador: Prof. Dr. José Augusto Ramos Soares

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, 23 de julho de 2012

Alinhamentos
e comparação de sequências

Esta versão definitiva da tese
contém as correções e alterações sugeridas pela
Comissão Julgadora durante a defesa realizada
por Francisco Eloi Soares de Araujo em 24 de maio de 2012.

Comissão Julgadora:

- Prof. Dr. José Augusto Ramos Soares (Presidente) – IME - USP
- Prof. Dr. Carlos Eduardo Ferreira – IME - USP
- Prof. Dr. Luiz Carlos da Silva Rozante – CMCC - UFABC
- Prof. Dr. Zanoni Dias – IC - UNICAMP
- Prof. Dr. Fábio Henrique Viduani Martinez – FACOM - UFMS

Agradecimentos

Muitos parentes e amigos foram importantes para o nosso doutorado. Certamente cada um desses sabe da sua importância nesse processo. Se eu esquecer de referenciar alguém direta ou indiretamente, perdoe-me pois a memória às vezes falha.

Agradeço inicialmente e principalmente à minha família: à Ione, minha esposa e aos meus filhos Gabriel, Sarah e Pedro que me deram amor e apoio logístico durante todos os anos de meu doutorado; e ao José Augusto, meu orientador e amigo, pelos ensinamentos e pela paciência sobrenatural. Certamente, sem o apoio da família e a paciência do José Augusto, esse trabalho não teria terminado.

Agradeço aos irmão e cunhados, sobrinhos e sobrinhas, sogro e sogra pelo apoio moral, em particular à minha cunhada Cláudia pelo incentivo e também pelas caronas do e para o IME. Agradeço também à minha família que estou redescobrindo em Campo Grande pela acolhida e pelo carinho recebido quando cheguei naquela cidade: meu tio Jair e meus primos Aily, Aise e Fernando, bem como os respectivos companheiros e descendentes.

Pela torcida e pela amizade, agradeço aos amigos do colégio São Luís; aos colegas de Jaboticabal incluindo em particular os irmãos de coração da república Amoribunda; aos amigos com quem trabalhei, Paulo Sérgio na Unisa, Hirata, Joyce, Nelson e Shirley no Senac, e Ana Lúcia, Cláudia Melo, Gabiru, Hamilton e Helena na Metodista; aos atuais amigos e colegas de trabalho da UFMS, em particular aos companheiros da FACOM-2 Carlos Higa, Fábio Iaione e Vagner Pedrotti que acompanharam os últimos momentos desse trabalho; aos amigos de doutorado Alexandre, Álvaro, Cao, Charlie Brown, Fábio Viduani, Jair Donadelli, Marco Aurélio, Mário Leston, Rogério e Said, e aos professores do IME Coelho, Cristina, José Augusto, Nami, Paulo Feofiloff, Yoshiharu e Yoshiko que me ensinaram muito. Também agradeço ao Professor Carlinhos do IME que esteve presente em todas as etapas de avaliação deste trabalho com sugestões motivadoras e enriquecedoras.

Agradeço ainda à CAPES pelo auxílio financeiro durante o doutorado.

Elói Araújo

Aos meus pais (in memoriam),
à minha esposa Ione e aos meus queridos filhos Gabriel, Sarah e Pedro,
que sempre me apoiaram
e ao meu orientador,
a quem fico grato pela oportunidade de estudar computação.

Resumo

Araujo, F. E. S. **Alinhamentos e comparação de sequências**. 2012. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

A comparação de sequências finitas é uma ferramenta que é utilizada para a solução de problemas em várias áreas. Comparamos sequências inferindo quais são as *operações de edição* de *substituição*, *inserção* e *remoção* de símbolos que transformam uma sequência em uma outra. As *matrizes de pontuação* são estruturas largamente utilizadas e que definem um custo para cada tipo de operação de edição. Uma matriz de pontuação γ é indexada pelos símbolos do alfabeto. A entrada de γ na linha `a`, coluna `b` mede o custo da operação de edição para substituir o símbolo `a` pelo símbolo `b`. As matrizes de pontuação induzem funções que atribuem uma pontuação para um conjunto de operações de edição. Algumas dessas funções para a comparação de duas e de várias sequências são estudadas nesta tese.

Quando cada símbolo de cada sequência é editado exatamente uma vez para transformar uma sequência em outra, o conjunto de operações de edição pode ser representado por uma estrutura conhecida por *alinhamento*. Descrevemos uma estrutura para representar o conjunto de operações de edição que não pode ser representado por um alinhamento convencional e descrevemos um algoritmo para encontrar a pontuação de uma sequência ótima de operações de edição usando um algoritmo conhecido para encontrar a pontuação de um alinhamento convencional ótimo.

Considerando três diferentes funções induzidas de pontuação, caracterizamos, para cada uma delas, a classe das matrizes para as quais as funções induzidas de pontuação são métricas nas sequências.

Dadas duas matrizes de pontuação γ e δ, dizemos que elas são *equivalentes* para uma dada função que é induzida por uma matriz de pontuação e que avalia a qualidade de um alinhamento se, para quaisquer dois alinhamentos A e B, vale o seguinte: o alinhamento A é "melhor" do que o alinhamento B considerando a matriz γ se e somente se A é "melhor" do que o alinhamento B considerando a matriz δ. Neste trabalho, determinamos condições necessárias e suficientes para que duas matrizes de pontuação sejam equivalentes.

Finalmente, definimos três novos critérios para pontuar alinhamentos de várias sequências. Todos os critérios consideram o comprimento do alinhamento além das operações de edição por ele representadas. Para cada um dos critérios definidos, propomos um algoritmo e o problema de decisão correspondente mostramos ser NP-completo.

Palavras-chave: métrica, matrizes equivalentes, custo normalizado de alinhamentos, distância de edição, alinhamento estendido, alinhamento de sequências, alinhamento de várias sequências.

Abstract

Araujo, F. E. S. **Alignment and comparison of sequences**. 2012. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

Comparison of finite sequences is a tool used to solve problems in several areas. In order to compare sequences, we infer which are the *edit operations* of *substitution*, *insertion* and *deletion* of symbols that transform one sequence into another. Scoring matrices are a widely used structure to define a cost for each type of edit operation. A scoring matrix γ is indexed by symbols of an alphabet. The entry in γ in row `a` and column `b` measures the cost of the edit operation for replacing symbol `a` by symbol `b`. Scoring matrices induce functions that assign a score for a set of edit operations. Some of these functions for comparing two and multiple sequences are studied in this thesis.

If each symbol is edited exactly once for transforming a sequence into another, the set of edit operations can be represented by a structure called *alignment*. We describe a structure to represent the set of edit operations that cannot be represented by a conventional alignment and we design an algorithm to find the cost of an optimal sequence of edit operations by using a known algorithm to find the cost of an optimal alignment.

Considering three different kinds of induced scoring functions, we characterize, for each one of them, the class of matrices for which the induced scoring functions are metrics on sequences.

Given two scoring matrices γ and δ, we say they are *equivalent* for a given function that is induced by a scoring matrix and that evaluates the quality of an alignment if, for any two alignments A and B of two sequences, we have the following: alignment A is "better" than B considering scoring matrix γ if and only if A is "better" than B considering scoring matrix δ. In this work, we determine necessary and sufficient conditions for scoring matrices to be equivalent.

Finally, we define three new criteria for scoring alignments of several sequence. Every criterion considers the length of the alignment and the edit operations represented by it. An algorithm for each criterion is studied and the corresponding decision problem is shown to be NP-complete.

Keywords: metric, equivalent matrices, normalized alignment cost, edit distance, extended alignment, sequence alignment, multiple sequence alignments.