DIMACS

Center for Discrete Mathematics &

Theoretical Computer Science

DIMACS EDUCATIONAL MODULE SERIES

MODULE 09-2

F*ind*ing Repeats With*in* Str*in*gs

Date Prepared: November, 2009

Dina Sokol[1]
Department of Computer and Information Science
Brooklyn College of the City University of NY
Brooklyn, NY 11210
sokol@sci.brooklyn.cuny.edu

Frederick Adkins
Mathematics Department
Indiana University of Pennsylvania
Indiana, Pennsylvania 15705
fadkins@iup.edu

Zhongyuan Che
Department of Mathematics
Penn State University, Beaver Campus
Monaca, PA 15061
zxc10@psu.edu

Kristin Pfabe
Department of Mathematics and Computer Science
Nebraska Wesleyan University
Lincoln, NE 68504
kpfabe@nebrwesleyan.edu

DIMACS Center, CoRE Bldg., Rutgers University, 96 Frelinghuysen Road, Piscataway, NJ 08854-8018
TEL: 732-445-5928 • FAX: 732-445-5932 • EMAIL: center@dimacs.rutgers.edu Web:
http://dimacs.rutgers.edu/

Founded as a National Science Foundation Science and Technology Center and a Joint Project of Rutgers University, Princeton University, AT&T Labs - Research, Bell Labs, NEC Laboratories America and Telcordia Technologies with affiliated members Avaya Labs, Georgia Institute of Technology, HP Labs, IBM Research, Microsoft Research, Rensselaer Polytechnic Institute, and Stevens Institute of Technology.

---

[1] This work has been supported in part by the *National Science Foundation* Grant DB&I 0542751.

Module Description Information

- **Title:**
  F*indin*g Repeats With*in* Str*in*gs

- **Authors:**

  1. Dina Sokol, Brooklyn College of the City University of NY, Brooklyn, NY 11210, sokol@sci.brooklyn.cuny.edu
  2. Frederick Adkins, Indiana University of Pennsylvania, Indiana, PA 15705, fadkins@iup.edu
  3. Zhongyuan Che, Pennsylvania State University Beaver Campus, Monaca, PA 15061, zxc10@psu.edu
  4. Kristin Pfabe, Nebraska Wesleyan University, Lincoln, NE 68504, kpfabe@nebrwesleyan.edu

- **Abstract:**

  Genomic sequences often contain copies of patterns called *repeats.* Repeats occurring in the genome are important genetic markers for disease diagnosis and mapping studies, as well as for human identity testing. This module presents several algorithms for finding repeats within biological sequences. Both *tandem* repeats, i.e. repeats in which copies are contiguous, and non-tandem repeats are discussed. Dynamic programming is described and a modification of the Smith-Waterman algorithm is shown for finding non-tandem repeats with errors. Algorithms are presented in pseudocode and illustrated with examples, including carefully diagrammed matrices. Each algorithm is analyzed for its asymptotic time complexity, motivating the selection of more efficient techniques. Several exercises and suggestions for additional explorations are given. Finally, programs in C++ or Java are included for the algorithms presented, and are available for running on the web at: http://tandem.sci.brooklyn.cuny.edu/SWrepeats.

- **Informal Description:**
  This module provides an introduction to several computational genomic techniques within the framework of finding repeats within a sequence. It is designed for students who are not necessarily experts in either biology or computer science. In Section 1 we describe some basic background in biology and the importance of repeats found in biological sequences. In Section 2 we present two algorithms for finding tandem

repeats within a sequence. Asymptotic time complexity is explained for the benefit of those who are unfamiliar with "big Oh" notation. In Section 3 we discuss more general repeats, allowing non-tandem repeats that include insertions, deletions, and mismatches. We describe how dynamic programming is used to generate a 2-sequence alignment using the Smith-Waterman algorithm. We then show how this can be modified to find all repeats occurring in a sequence. Section 5 contains additional exercises and Section 6 contains supplementary material including the solutions to all of the exercises. C++ or java code is included for each algorithm in the appendix.

- **Target Audience:**
  Undergraduate students at the sophomore level or above in a computer science, mathematics, or biology department.

- **Prerequisites:**
  An introductory computer science class and familiarity with counting techniques is required. In a computational biology course, this material would follow nicely after coverage of the Smith-Waterman method for local alignment of two strings. However, this presentation is self-contained and requires no specific background in computational biology.

- **Mathematical Field:**
  Computational Biology, Discrete Mathematics, Computer Science

- **Application Areas:**
  Sequence analysis of biological sequences such as DNA, RNA and protein sequences

- **Mathematics Subject Classification:**
  MSC (2010): 68W32, 68Q25, 92D20, 62P10

- **Contact Information:**
  Dina Sokol, Brooklyn College of the City University of NY, Brooklyn, NY 11210,
  sokol@sci.brooklyn.cuny.edu

- **Other DIMACS modules related to this module:**
  None