Information and Computation 213 (2012) 59–69

Contents lists available at SciVerse ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Optimizing restriction site placement for synthetic genomes

Pablo Montes [a,1], Heraldo Memelli [a], Charles B. Ward [a,*,2], Joondong Kim [b], Joseph S.B. Mitchell [b], Steven Skiena [a,2]

[a] Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, United States
[b] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, United States

ARTICLE INFO

Article history:
Available online 3 February 2012

Keywords:
Synthetic biology
Restriction enzyme placement
Genome refactoring

ABSTRACT

Restriction enzymes are the workhorses of molecular biology. We introduce a new problem which arises in the course of our project to design virus variants to serve as potential vaccines: we wish to modify virus-length genomes to introduce large numbers of unique restriction enzyme recognition sites while preserving wild-type function by substitution of synonymous codons. We show that the resulting problem is NP-Complete, give an exponential-time algorithm, as well as well-performing heuristics, and give excellent results for five sample viral genomes. Our resulting modified genomes have several times more unique restriction sites and reduce the maximum gap between adjacent sites by three to nine-fold.

1. Introduction

Synthetic biology is an exciting emerging field with the goal of designing novel living organisms at the genetic level. DNA sequencing technology can be thought of as reading DNA molecules, so as to describe them as strings on {*ACGT*} for computational analysis, while DNA *synthesis* is the inverse operation, constructing DNA molecules with exactly a specified sequence. Commercial vendors such as GeneArt (http://www.geneart.com) and Blue Heron (http://www.blueheronbio.com) today charge as little as 60 cents per base to synthesize virus-length sequences, and prices are rapidly dropping [4,8]. The advent of cheap synthesis will have many exciting new applications throughout the life sciences, and the need to design new sequences to specification leads to a variety of new algorithmic problems on sequences.

In this paper, we introduce a new problem which arises in the course of our project to design virus variants to serve as potential vaccines [7,16,23]. In particular, *restriction enzymes* are laboratory reagents which cut DNA at specific patterns. For example, the enzyme EcoRI cuts at the pattern *GAATTC*. Each enzyme cuts at a particular pattern, and over 3000 restriction enzymes have been studied in detail, with more than 600 of these being available commercially [20].

Each occurrence of a pattern within a given DNA target sequence is called a *restriction enzyme recognition site* or *restriction site*. Unique restriction sites within a given target are particularly prized, as they cut the sequence unambiguously in exactly one place. Many techniques for manipulating DNA make use of unique restriction sites [12,19]. In particular, *subcloning* is an important method of inserting a new sequence between restriction sites unique for two different enzymes.

* Corresponding author.
E-mail addresses: pmontes@cs.sunysb.edu (P. Montes), hmemelli@cs.sunysb.edu (H. Memelli), charles@cs.sunysb.edu (C.B. Ward), jdkim@ams.sunysb.edu (J. Kim), jsbm@ams.sunysb.edu (J.S.B. Mitchell), skiena@cs.sunysb.edu (S. Skiena).

[1] On leave from and supported in part by Politécnico Grancolombiano, Bogotá, Colombia.
[2] Supported in part by NIH Grant 5R01AI07521903, NSF Grant DBI-0444815, and IC Postdoctoral Fellowship HM1582-07-BAA-0005.

0890-5401/$ – see front matter
doi:10.1016/j.ic.2012.02.003

Thus, a genomic sequence which contains unique restriction sites at regular intervals will be easy to manipulate in the laboratory. Traditionally, DNA sequences manipulated in laboratories were from living organisms, so the experimenter had no choice but to work with what they were given, but low-cost, large-scale DNA synthesis changes this equation. Refactoring [15] is a software engineering term for redesigning a program to improve its internal structure for better ease of maintenance, while leaving its external behavior unchanged. Genome synthesis technology provides us the means to refactor biological organisms, restructuring the genome of an organism to be easier to manipulate while preserving its natural biological functions.

The redundancy of the genetic code (64 three-base codons coding for 20 distinct amino acids) gives us the freedom to insert and remove restriction sites without changing the protein coded for by a given gene. Identifying the locations of both current and potential sites can be done using conventional pattern matching algorithms. Much more challenging is the problem of finding well-spaced unique placements for many different enzymes to facilitate laboratory manipulation of synthesized sequences. Our contributions in this paper are:

- *Problem Definition* – We abstract a new optimization problem on sequences to model this sequence design task: the *Unique Restriction Site Placement Problem* (URSPP). We show this problem is NP-Complete and give approximability results.
- *Algorithm Design* – We present a series of algorithms and heuristics for the Unique Restriction Site Placement Problem. In particular give an $O(n^2 2^m)$-time dynamic programming algorithm for URSPP, which is practical for designs with small number of enzymes. We also give an efficient greedy heuristic as well as a heuristic based on weighted bipartite matching, both of which are polynomial in the sequence length n and the number of restriction enzymes m, and both of which construct good designs in practice.
- *Sequence Design Tool* – Our design algorithms have been integrated with the Aho–Corasick pattern matching algorithm to yield a sequence design tool we anticipate will be popular within the synthetic biology community. In particular, we have developed this tool as part of a project underway to design a candidate vaccine for a particular agricultural pathogen.
- *Experimental Results for Synthetic Viruses* – The URSPP problem abstraction to some extent obscures the practical aspects of sequence design. The critical issue is how regularly unique restriction sites can be inserted into the genomes of representative viruses. We perform a series of experiments to demonstrate that impressive numbers of regularly-spaced, unique restriction sites can be engineering into viral genomes.
  Indeed, our system produces genomes with three to four-fold more unique restriction enzymes than a baseline algorithm (details given in the results section) and reduces the maximum gap size between restriction sites three to nine-fold. Fig. 1 shows example results for Polio virus and Equine Arteritis virus.

This paper is organized as follows. In Section 2 we briefly review related work on genome refactoring and sequence design. In Section 3 we discuss our algorithmic approach to the problem. Finally, in Section 4 we give the results for our refactored viral genomes.

2. Related work

Synthetic biology is an exciting new field of growing importance. The synthesis of virus-length DNA sequences (8 to 20 thousand bases), a difficult task just a decade ago [5], is now a relatively inexpensive commercialized service. This enables a tremendous number of applications, notably the manipulation of viral genomes to produce attenuated viruses for vaccine production [7,16,23]. This work is in support of genome refactoring efforts related to this project.

Broadly, the field of genome refactoring seeks to expand our understanding of genetics through the construction of an engineering toolkit to easily modify genomes. Chan et al. [6], for example, refactored the bacteriophage T7 so "that is easier to study, understand, and extend." A number of different tools for genome refactoring exist: GeneJAX [3] is a JavaScript web application CAD tool for genome refactoring. SiteFind is a tool which seeks to introduce a restriction enzyme as part of a point mutation using site-directed mutagenesis [11]. However, SiteFind considers the much more restricted problem of introducing a single restriction site into a short (< 400b) sequence specifically to serve as a marker for successful mutagenesis, in contrast with our efforts to place hundreds of sites in several kilobase genomes.

GeneDesign is a tool which aids in the design of synthetic genes [18]. One of its functionalities is the silent insertion of restriction sites, allowing the user to manually choose enzymes and sites from possible places where they can be inserted, or doing this automatically. The latter is similar to our tool in trying to automate the creation of restriction sites in the sequence, but the process is done quite differently. GeneDesign only attempts to insert restriction sites of enzymes that do not appear anywhere in the sequence, and further follows only a simple heuristic to try to space the introduced consecutive sites at an interval specified by the user, without any attempt or guarantee to optimize the process.

Other relevant work includes work by Skiena [21], which gives an algorithm for optimally removing restriction sites from a given coding sequence. The problem here differs substantially, in that (1) we seek to remove all but one restriction sites per cutter, and (2) we seek to introduce cut sites of unrepresented enzymes in the most advantageous manner.