

Chapter 2

Strings: Theory, Properties and Applications

Abstract Genomic and proteomic data can be represented as sequences over the nucleotides and amino acids alphabets, and many tasks require algorithms working on strings. This chapter introduces the formalism to deal with sequences and the definition of distance metrics, largely used in string selection methods.

2.1 Introduction

Sequencing projects generate a deluge of data, which is represented by sequences over an alphabet Σ (e.g., the nucleotides alphabet); in this context, string selection is a fundamental task that finds application in different fields, e.g. phylogenetic tree reconstruction, primer design, DNA binding sites identification. Moreover, the ability to synthesize long DNA molecules has increased the need of algorithms to identify sequences that meet specific requirements, e.g. minimize homopolymer segments, design of optimal assembly oligos. Nevertheless, many of these problems are NP-hard, representing challenging tasks both from a computational and biological perspectives. In this chapter, we introduce the basic notations for representing biological sequences; since many tasks involve finding similarities and common patterns, we introduce the most largely used metrics, which are the Levenshtein distance and the Hamming distance. Finally, we give an overview of biological applications that require the solution of string selection problems.

2.2 Basic Definitions

DNA, RNA, and protein sequences can be thought as strings of symbols over a finite alphabet. Specifically, the DNA alphabet consists of four characters, corresponding to the nucleotides, while protein sequences can be viewed as strings over the

20-letter alphabet of amino acids. In this section, we introduce some fundamental concepts, in order to settle the notation for this manuscript.

An alphabet $\Sigma = \{c_1, \dots, c_k\}$ is a finite set of elements, called characters. A string s can be defined as a finite sequence of characters (c_1, \dots, c_m) , $c_i \in \Sigma$, $i = 1, \dots, m$, and we denote the empty string $s^0 = \epsilon$. Given a string s over a finite alphabet Σ , $|s|$ and s_i denote the length of s and the i th character of s , respectively.

2.3 Distance Metrics

Discovering analogies and/or differences in genomic data requires the introduction of metrics to quantify the similarity, alternatively the distance, between two sequences. In general, two metrics are commonly adopted: the Levenshtein or edit distance [5] and the Hamming distance [3].

The Levenshtein edit distance between two strings s and t , denoted as $d_L(s, t)$, is the minimum number of edits, which are insertions, deletions, and substitutions, required to transform one string into the other. Let us define the predicate function $\Phi_L(s, t)$ between s_1, \dots, s_i and t_1, \dots, t_j as follows:

$$\Phi_L(i, j) = \begin{cases} 0 & \text{if } i = j = 1 \\ i & \text{if } j = 1 \wedge i > 1 \\ j & \text{if } i = 1 \wedge j > 1 \\ \min \begin{cases} \Phi_L(i-1, j) + 1 \\ \Phi_L(i, j-1) + 1 \\ \Phi_L(i-1, j-1) + 1 + [s_i \stackrel{?}{=} t_j] \end{cases} & \text{otherwise} \end{cases} \quad (2.1)$$

where $s_i \stackrel{?}{=} t_j$ is equal to 1 if the characters s_i and t_j match, otherwise is 0.

The first element in the minimum corresponds to an insertion from s to t , the second to a deletion and the third to a match or a mismatch, depending on whether the respective symbols are the same. The Levenshtein distance for the string s and t is defined as

$$d_L(s, t) = \Phi_L(|s|, |t|). \quad (2.2)$$

Since the various type of edit operations occur with different probabilities, a metric which takes into account weights can be more reliable: in biology, for instance, some mismatches are less penalizing than others (e.g., mismatch involving amino acids of the same family). In this case, we define the weighted edit distance between two strings s and t as the cost of the cheapest sequence of edit operations needed to transform s into t , where weights are considered.

The Hamming distance represents a special case of the edit distance, where only mismatches between strings are taken into account. Specifically, the Hamming

distance between two strings s and t having equal length, denoted by $d_H(s, t)$, is the number of positions at which s and t differ. Let $\Phi_H : \Sigma \times \Sigma \rightarrow \{0, 1\}$ a predicate function such that $\Phi_H(x, y) = 1$ if and only if $x \neq y$; the Hamming distance between two strings s and t is defined as:

$$d_H(s, t) = \sum_{i=1, \dots, |s|} \Phi_H(s_i, t_i). \quad (2.3)$$

Although the edit distance is a more accurate metric for several genomic operations, many problems are defined in terms of Hamming distance. There are three main reasons: first, the problems we analyze have the objective of finding an exact string or substring, not a modified version of it, therefore gaps are not taken into account. Moreover, since gaps are more destabilizing than substitutions, the use of the Hamming distance is preferable as a distance metric [4]. Last, the Hamming distance can be used to describe the effects of mutations in the protein coding regions of DNA, whereas edit distance is more suitable to measure the evolutionary distance [4].

2.4 Applications

The objective of our analysis are problems addressing the localization of similar features in nucleotide or amino acids sequences, and the identification of patterns enriched in a set of sequences. The first class of problems comprises the closest string problem (CSP), the closest substring problem (CSSP) and its decision version, the common approximate substring problem (CAS), the close to most string problem (CMSP), the center and median string problems; the second class includes the farthest string problem (FSP), the farthest substring problem (FSSP), and the far from most string problem (FFMSP). Another class of problems mentioned in this work aims at finding a pattern that occurs in one set of strings but does not occur in another set, known as distinguishing string selection problem (DSSP), and the d-mismatch problem, which generalizes the concept of closest string to center strings of an aligned set of substrings. These problems arise in many molecular biology tasks and, hence, finding high-quality solutions is challenging both for computer scientists and for biologists; below, we will present some molecular biology problems strictly related to SSP.

2.4.1 *Primer Design for Polymerase Chain Reaction*

Polymerase chain reaction (PCR) is a technique adopted in molecular biology for amplifying a portion of DNA in many copies. PCR has many applications, such as DNA cloning for sequencing, functional analysis of genes, forensic, disease

diagnosis. First, PCR requires the selection of two primers, which are fragments of DNA complementary to the 3' ends of the sense and antisense strands of the regions to amplify, called template DNA; the primers bind the template DNA during the annealing step of PCR, and the polymerase binds to these regions to start DNA synthesis.

Primer selection is a complex task, and it affects the results of PCR experiments. Particularly interesting is the selection of primers that are able to amplify several regions simultaneously; recently, it has been shown that this task can be addressed as a DSSP [1].

2.4.2 Identification of Transcription Factor Binding Sites

A phylogenetic tree or evolutionary tree is a diagram that depicts the evolutionary relationships among various species or other entities, based upon similarities and differences in their characteristics. Phylogenetic trees can be inferred from sequence alignment and used to detect potentially important regions within a DNA sequence: given an alignment, we look for highly conserved regions.

Transcription factors are proteins that bind to specific DNA sequences, controlling the transcription of genetic information from DNA to mRNA, thus affecting the expression of a gene. In particular, transcription factors bind to a specific DNA site, allowing only a small amount of variation; therefore, they can be used to identify conserved regions in biological sequences. Identifying transcription factors among a set of sequences can be reduced to the d -mismatch problem [8]; given a set of sequences, possible binding sites can be identified as a string matching the input sequences with at most d mismatches.

2.4.3 Multiple Tree Alignment Problem

A multiple alignment is a sequence alignment of three or more biological sequences, such as proteins, DNA, or RNA, used to identify regions of similarity that may imply functional, structural, or evolutionary relationships among the sequences. The problem can be viewed as finding a set of patterns which, with some errors, appear in the same order in all the sequences of a given set [2]. A variant of the problem is known as the tree alignment with a given phylogeny; given a set S of n sequences over an alphabet Σ , and an unlabelled tree with n leaves, representing the phylogeny, an evolutionary tree is a labelled tree built on the phylogeny, having labels on the leaves representing the sequences in the input set S , and labels on the other nodes which are the sequences over the alphabet Σ .

The cost of an evolutionary tree is the sum of all the edit distances between the labels of pairs of nodes joined by a vertex, and an optimal evolutionary tree is a tree that minimizes such cost [2]. A star phylogeny is characterized by $n + 1$ nodes,

where n of them are leaves of the tree. This topology is often used to represent a recent population expansion event from a common ancestor. Such problem can be viewed as finding the median string of a set of input sequences.

2.4.4 Design of Diagnostic Probes

The task of designing a string that is able to represent a set of known sequences and, at the same time, is easily distinguishable from another set, is a problem arising in the diagnosis of viruses and bacteria in host organisms. In particular, probes are used to diagnose the presence of viruses and bacteria in biological samples. A probe is a strand of DNA or RNA opportunely treated with a radioactive isotope, dye, or enzyme, and used to easily detect the presence of a specific sequence of nucleotides, called target, on a single-stranded nucleic acid, by performing hybridization experiment. The probe hybridizes to the target if they are complementary to each other. Hence, given a set of sequences representing the virus or the bacteria, and a host, the problem is to discover a sequence that occurs in virus sequences, but does not appear in the host [7]. This problem involves the design of the probe sequence, that has to be as close as possible to the sequences to detect, and, on the other hand, as far as possible from another set of sequences.

2.4.5 Protein Function Prediction

Predicting the function of an unknown protein represents a central problem in bioinformatics, hence an increasing number of methods have been proposed in literature to tackle this problem [6]. One approach consists in analyzing the shared motifs, which are common motifs between two or more sequences and can be associated with a particular function or regulation. String selection problems find application in addressing such problems, by looking for suboptimal matchings among strings.

2.4.6 Drug Design

Another important application of string selection concerns the design of drugs and therapies. Here, given a set of sequences of orthologous genes from a group of closely related pathogens, and a host, the goal is to identify a sequence that is highly conserved in all or most of the pathogens' sequences, but that is not present in the host [4]. In fact, conserved regions might encode relevant biological information, since they seem resilient to mutations. This information can be exploited to identify the chemical components that bind the conserved region, in order to create new effective therapies. In antisense drug design, the same principle is used to create

drugs that inhibit the production of the protein related to the disease, but do not interfere with other proteins [4]. Specifically, antisense therapies focus on impeding the production of proteins that cause the disease: antisense drugs bind mRNA to prevent the genetic code related to the disease be read by the ribosome, which is responsible in assembling proteins based on the instructions carried by mRNA.

References

1. Boucher, C.: Combinatorial and probabilistic approaches to motif recognition. Ph.D. thesis, University of Waterloo (2010)
2. de la Higuera, C., Casacuberta, F.: Topology of strings: Median string is NP-complete. *Theor. Comput. Sci.* **230**(1), 39–48 (2000)
3. Hamming, R.W.: Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**(2), 147–160 (1950)
4. Lancot, J., Li, M., Ma, B., Wang, S., Zhang, L.: Distinguishing string selection problems. In: *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 633–642 (1999)
5. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. In: *Soviet Physics Doklady*, vol. 10, p. 707 (1966)
6. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**(5428), 751–753 (1999)
7. Phillippy, A.M., Mason, J.A., Ayanbule, K., Sommer, D.D., Taviani, E., Huq, A., Colwell, R.R., Knight, I.T., Salzberg, S.L.: Comprehensive dna signature discovery and validation. *PLoS Comput. Biol.* **3**(5), e98 (2007)
8. Stojanovic, N., Berman, P., Gumucio, D., Hardison, R., Miller, W.: A linear-time algorithm for the 1-mismatch problem. In: *Algorithms and Data Structures*, pp. 126–135. Springer, Berlin Heidelberg (1997)

Optimization Approaches for Solving String Selection Problems

Pappalardo, E.; Pardalos, P.; Stracquadanio, G.

2013, VIII, 49 p. 2 illus. in color., Softcover

ISBN: 978-1-4614-9052-4