

# Parameterized Problems in Bioinformatics

Liliana Félix Ávila    Alina García    María José Serna    Dimitrios M. Thilikos

June 27, 2006

## Abstract

## 1 Introduction

## 2 Genome Rearrangement

### 2.1 $k$ -BREAKPOINT MEDIAN

Biological Motivation:

With breakpoint distance, the genome rearrangement field delivered one of the currently most popular measures in phylogenetic studies for related species. Here, breakpoint median, whose genomes are represented as signed orderings, is the core basic problem [GN02].

Problem Definition:

Given a set  $S = \{1, \dots, n\}$ , an *ordering*  $\pi$  on  $S$  is a  $1 : 1$  function  $\pi : S \rightarrow S$ . It is required that every ordering is extended by two special elements namely  $s$ , marking the start, and  $t$ , marking the end, and write the ordering  $\pi$  as  $\langle s \ \pi(1) \ \pi(2) \ \dots \ \pi(n) \ t \rangle$ . Then  $S_s$  is  $S \cup \{s\}$  ( $S_t$  and  $S_{s,t}$ , analogously) [GN02].

An ordering  $\pi$  is *signed* iff every  $\pi(x)$ ,  $x \in S$ , is equipped with a sign  $\{+, -\}$ , denoting the “orientation” of the element, such that  $\pi(x)$  can be, for  $y \in S$ , a “positive” element  $+y$  (or, only  $y$ ), having left-to-right orientation, or a “negative” element  $-y$ , having right-to-left orientation. Note that a signed ordering contains either  $y$  or  $-y$ , but not both at the same time. The special elements  $s$  and  $t$  are always unsigned. We write  $S^\pm$  for the set  $\{-1, 1, -2, 2, \dots, -n, n\}$  and  $S_s^\pm$  for  $S^\pm \cup \{s\}$  ( $S_t^\pm$  and  $S_{s,t}^\pm$  analogously) [GN02].

We use  $\text{succ}_\pi(x)$ , for signed ordering  $\pi$  and  $x \in S_s$ , to denote the *successor*  $y \in S_{s,t}^\pm$  of element  $x$  in  $\pi$ , which is defined w.r.t.  $x$ ’s direction: For an element  $x \in G$  occurring positively in  $\pi$ , the successor is the element following  $x$ . An  $x \in G$  occurring negatively, however, has “reverse” orientation; hence, from  $x$ ’s point of view, its successor is the “reverse version” of the element preceding  $x$  [GN02].

Given two signed orderings  $\pi_1$  and  $\pi_2$ , both over  $S$ , we call a pair  $(x, y)$ ,  $x \in S_s^\pm$  and  $y \in S_t^\pm$ , a *breakpoint* of  $\pi_1$  w.r.t.  $\pi_2$ , if

1.  $x = s$  or  $\pi_1(l) = x$  for some  $l \in S$ , and
2.  $\text{succ}_{\pi_1}(x) = y$  and  $\text{succ}_{\pi_2}(x) \neq y$

Using the notion of breakpoints, we define the *breakpoint distance*  $d_{bp}$  between two signed orderings as follows:  $d_{bp}(\pi_1, \pi_2) = |\{(x, y) | x, y \in S_{s,t}^\pm, x, y \text{ is breakpoint of } \pi_1 \text{ w.r.t. } \pi_2\}|$  [GN02].

Due to symmetry,  $d_{bp}(\pi_1, \pi_2) = d_{bp}(\pi_2, \pi_1)$  [GN02].

Herein,  $d_{bp}(\pi_i, \pi)$  denotes the *breakpoint distance* between orderings  $\pi_i$  and  $\pi$ .

*Instance:* Signed orderings  $\pi_1, \pi_2, \dots, \pi_m$  on  $n$  elements, and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a signed ordering  $\pi$  such that  $\sum_{i=1}^m d_{bp}(\pi_i, \pi) \leq k$ ?

Complexity:

NP-complete [PS].

Parameterized Complexity:

FPT,  $O(2.15^k \cdot mn)$  time [GN02].

## 2.2 SORTING BY REVERSALS

Biological Motivation:

Studies of genomes evolving by rearrangements lead to combinatorial problem of *sorting permutation by reversals*. Physical maps usually do not provide information about directions of genes and, therefore lead to representation of a genome as an *unsigned* permutation  $\pi$ . Biologists implicitly try to derive a signed permutation from this representation by assigning a positive (negative) sign to increasing (decreasing) strips of  $\pi$ . Biologists have to choose the desired degree of resolution of the constructed physical maps. Low-resolution physical maps usually contain many *singletons*

(strips of size one) and, as a result, rearrangement scenarios for such maps are hard to analyze [HP96].

$O(\log n)$  singletons is the desired trade-off of resolution for cross-hybridization physical mapping in molecular evolution studies. If the number of singletons is large, a biologist might choose additional experiments (i.e. sequencing of some areas) to resolve the ambiguities in gene directions [HP96].

**Problem Definition:**

A reversal  $\rho = \rho(i, j)$  on a permutation  $\pi = \pi_1 \dots \pi_{i-1} \pi_i \dots \pi_j \pi_{j+1} \dots \pi_n$  reverses the order of elements  $\pi_i \dots \pi_j$  and transforms  $\pi$  into permutation  $\pi \cdot \rho = \pi_1 \dots \pi_{i-1} \pi_j \dots \pi_i \pi_{j+1} \dots \pi_n$ . *Reversal distance*  $d(\pi)$  is defined as the minimum number of reversals  $\rho_1, \dots, \rho_t$  to transform  $\pi$  into the *identity* permutation [HP96].

Let  $i \sim j$  if  $|i - j| = 1$ . Extend a permutation  $\pi = \pi_1 \dots \pi_n$  by adding  $\pi_0 = 0$  and  $\pi_{n+1} = n + 1$ . We call a pair of consecutive elements  $\pi_i$  and  $\pi_{i+1}$ ,  $0 \leq i \leq n$ , of  $\pi$  an *adjacency* if  $\pi_i \sim \pi_{i+1}$ , and a *breakpoint* if  $\pi_i \not\sim \pi_{i+1}$ . Define a *block* of  $\pi$  as an interval  $\pi_i \dots \pi_j$  containing no breakpoints. Define a *strip* of  $\pi$  as a maximal block. A strip of one element is called a *singleton* [HP96].

*Instance:* Given a permutation  $\pi$  of  $\{1, 2, \dots, n\}$  with  $k$  singletons.

*Parameter:*  $k$

*Question:* Does there exist at most  $k$  reversals needed to transform  $\pi$  into the *identity* permutation?

**Complexity:**

NP-hard [KS93].

**Parameterized Complexity:**

FPT,  $O(2^k n^3 + n^4)$  time [HP96].

## 2.3 SYNTENIC DISTANCE

**Biological Motivation:**

Computational models measuring the genetic distance between two species can be used in the construction of tree of evolutionary history, or—if such a tree is known through other means—in estimating the rate of genomic evolution. These measures are generally based on a hypothesized set of *transformations* that can alter a genome; the *distance* between the genomes of two species is then the minimum number of these steps necessary to transform one into the other. Two genes are *syntenic* if they appear in the same chromosome [LN02].

When comparing genomes containing multiple chromosomes, one must consider transformations acting between chromosomes in addition to those acting within a single chromosome. These transformations include *fissions*, in which one chromosome splits into two, *fusions*, in which two chromosomes merge into one, and *translocations*, in which two chromosomes exchange contiguous blocks (usually prefixes or suffixes of genes) [LN02].

**Problem Definition:**

In this model, a genome is given by  $k$  subsets of a set of  $n$  characters (genes). These subsets represent the chromosomes and the characters in a set represent the genes located on the chromosome. The mutation events in this model are the union of two chromosomes sets, the splitting of a chromosome set into two sets, and the exchange of genes between two sets. [AGN01].

A genome can be transformed by any of the following operations:

- a *fusion*  $(S, T) \longrightarrow U$ , in which two chromosomes  $S$  and  $T$  merge into a single chromosome  $U$ , where  $U = S \cup T$ .
- a *fission*  $U \longrightarrow (S, T)$ , in which a chromosome  $U$  splits into two chromosomes  $S$  and  $T$ , where  $U = S \cup T$ .
- a *translocation*  $(S, T) \longrightarrow (S', T')$ , in which two chromosomes  $S$  and  $T$  exchange arbitrary subsets of their genes, producing two new chromosomes  $S'$  and  $T'$ , where  $S \cup T = S' \cup T'$ .

The *syntenic distance*  $d(\mathcal{S}_1, \mathcal{S}_2)$  between two genomes  $\mathcal{S}_1$  and  $\mathcal{S}_2$  is the minimum number of fusions, fissions, and translocations required to transform  $\mathcal{S}_1$  into  $\mathcal{S}_2$ , ignoring all genes that appear in only one of the two genomes [LN02].

Herein,  $d(\mathcal{S}_1, \mathcal{S}_2)$  denotes the *syntenic distance* between  $\mathcal{S}_1$  and  $\mathcal{S}_2$ .

*Instance:* Given two genomes  $\mathcal{S}_1 = S_{1_1}, \dots, S_{1_n}$  and  $\mathcal{S}_2 = S_{2_1}, \dots, S_{2_m}$ .

*Parameter:*  $k$

*Question:* Does there exist  $d(\mathcal{S}_1, \mathcal{S}_2) \leq k$ ?

Complexity:

NP-complete [LN02, DJK<sup>+</sup>97].

Parameterized Complexity:

FPT,  $O(nm + 2^{O(k \log k)})$  time [LN02].

## 2.4 VERTEX BIPARTIZATION

Biological Motivation:

In SNP haplotype assembly problems, the goal is to *assign* a given haplotype fragment, represented by its sequence of SNP states, to one of two possible haplotypes. In the reconstruction of haplotype structure, the goal is to *divide* the given genotype fragments, represented by their sequence of not necessarily unique SNP states, into two haplotype fragments each. The commonality of both problems is that we require a bipartition of haplotype fragments into two sets such that haplotype fragments with differences in their SNP states belong to different sets.

In VERTEX BIPARTIZATION we ask, given a graph  $G$  and a non-negative integer  $k$ , whether we can transform the graph into a bipartite graph by deleting at most  $k$  vertices [Gra03].

Problem Definition:

*Instance:* Given a graph  $G = (V, E)$ ; a non-negative integer  $k$ .

*Parameter:*  $k$

*Question:* Can we transform the graph into a bipartite graph by deleting at most  $k$  vertices?

Complexity:

NP-complete [GJ79], (Problem number GT25).

Parameterized Complexity:

FPT,  $O(4^k kmn)$  time [RSV04].

## 2.5 Open

### 2.5.1 EDGE BIPARTIZATION

Biological Motivation:

See problem 2.4 for biological motivation.

In EDGE BIPARTIZATION we ask, given a graph  $G$  and a non-negative integer  $k$ , whether we can transform the graph into a bipartite graph by deleting at most  $k$  edges [Gra03].

Problem Definition:

*Instance:* Given a graph  $G = (V, E)$ ; a non-negative integer  $k$ .

*Parameter:*  $k$

*Question:* Can we transform the graph into a bipartite graph by deleting at most  $k$  edges?

Complexity:

NP-complete [GJ79], (Problem number #).

Parameterized Complexity:

Is EDGE BIPARTIZATION fixed-parameter tractable with respect to the number of allowed edge deletions? [Gra03].

## 3 Sequence Alignment and Evolution

### 3.1 BINARY CLADISTIC CHARACTER COMPATIBILITY

Biological Motivation:

In systematics, parsimony methods construct phylogenies, or evolutionary trees, in which characters evolve with the least evolutionary change. The Camin-Sokal and Dollo parsimony criteria are used to construct phylogenies from discrete characters [DJS86].

Major strategies for inferring phylogenies have been developed from basic concepts of compatibility and parsimony. For a given set of objects (e.g., terminal taxa, operational taxonomic units), compatibility criteria are used to seek phylogenies on which a largest set of characters is perfectly compatible, whereas parsimony criteria are used to seek phylogenies on which characters evolve with the least evolutionary change [DJS86].

Problem Definition:

There exist finite nonempty sets of objects (e.g., terminal taxa) and of characters that describe the objects.

Each character has two states and so is called *binary*.

A binary character is called *cladistic* if its states are ordered so that one is *ancestral* and the other *derived*.

The  $n$  character states of an object  $x$  are described by a vector  $v(x) = \langle v_1, \dots, v_n \rangle$ , in which  $v_i$  is the state of character  $i$  for object  $x$ .

Two binary characters are compatible with each other if three or fewer of the four possible combinations of their states 00, 01, 10, and 11 are present in the objects being examined.

*Instance:* A set  $C$  of  $n$  binary cladistic characters over  $m$  objects; a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a subset  $C' \subseteq C$ ,  $|C'| = k$ , such that all pairs of characters in  $C'$  are compatible?

Complexity:

NP-complete [DS86].

Parameterized Complexity:

$W[1]$ -complete, hardness: reduction from CLIQUE [DJS86].

Comments:

The unconstrained-character version of this problem is also  $W[1]$ -complete [DJS86].

If  $k = |C|$ , one obtains the PERFECT PHYLOGENY problem [DJS86].

### 3.2 BINARY QUALITATIVE CHARACTER COMPATIBILITY

Biological Motivation:

See problem 3.1 for biological motivation.

Major strategies for inferring phylogenies have been developed from basic concepts of compatibility and parsimony. For a given set of objects (e.g., terminal taxa, operational taxonomic units), compatibility criteria are used to seek phylogenies on which a largest set of characters is perfectly compatible, whereas parsimony criteria are used to seek phylogenies on which characters evolve with the least evolutionary change [DJS86].

Problem Definition:

A binary character is called *qualitative* if its states are an unordered set on which no further structure is imposed.

*Instance:* A set  $C$  of  $n$  binary qualitative characters over  $m$  objects; a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a subset  $C' \subseteq C$ ,  $|C'| = k$ , such that all pairs of characters in  $C'$  are compatible?

Complexity:

NP-complete, by reduction from BINARY CLADISTIC CHARACTER COMPATIBILITY [DS86].

Parameterized Complexity:

$W[1]$ -complete, hardness: reduction from BINARY CLADISTIC CHARACTER COMPATIBILITY [DJS86].

Comments:

The unconstrained-character version of this problem is also  $W[1]$ -hard [DJS86].

### 3.3 BOUNDED DUPLICATION SHORTEST COMMON SUPERSEQUENCE (SCS) FOR COMPLETE P-SEQUENCES

Biological Motivation:

When trying to resolve the species tree for a set of  $n$  taxa, one typically creates a set of  $k$  gene trees. It is not always the case that the gene trees agree. One such reason is due to paralogous duplications of genes followed by subsequent loss of genes. This model implicitly makes use of trees with repeated leaf labels. For problems about sequences, it is usually assumed that the sequences of interest will contain occurrences of the same symbol many times. But there are some applications where attention may be restricted to sequences  $x$  where any symbol occurring in  $x$  occurs at most once [FHKS98b].

Problem Definition:

**p-sequence:** Say that a string of symbols (or sequence)  $x \in \Sigma^*$  is a *p-string* (*p-sequence*) if no symbol of  $\Sigma$  occurs more than once in  $x$  [FHKS98b].

**complete p-sequence:**  $x$  is a *complete* p-sequence if each symbol of the alphabet occurs exactly once in  $x$  [FHKS98b].

A string  $x$  is a *supersequence* of a string  $s$  if we can delete some characters in  $x$  such that the remaining string is equal to  $s$ .

A sequence  $S$  contains  $r$  *duplication events* if  $S$  is not a p-sequence but the exactly  $r$  symbols need to be removed from  $S$  to result in a p-sequence [FHKS98b].

*Instance:* Complete p-sequences  $x_i$  over an alphabet  $\Sigma$  of size  $n$ , a positive integer  $r$ , and a cost function  $c : \Sigma \rightarrow \mathbb{Z}^+$

*Parameter:*  $r$

*Question:* Is there a common supersequence  $x$  of duplication cost  $\|x\|_c \leq r$  where the duplication cost is defined as  $\|x\|_c = \sum_{a \in \Sigma} (n_a(x) - 1)c(a)n_a(x)$ ,  $a \in \Sigma$ , denotes the number of occurrences of symbol  $a$  in  $x$ .

Complexity:

NP-complete [FHKS98b].

Parameterized Complexity:

FPT [FHKS98b].

### 3.4 BOUNDED DUPLICATION SCS FOR P-SEQUENCES

Biological Motivation:

See problem 3.3 for biological motivation.

Problem Definition:

**rl-sequence:**  $x$  is an *rl-string* or *rl-sequence* if it is a string in the usual sense, where symbols of  $\Sigma$  may be repeated [FHKS98b].

*Instance:* A family of  $k$  p-sequences  $x_i \in \Sigma^*$  for  $i = 1, \dots, k$  and a positive integer  $r$  representing the number of duplication events (for the definition of p-sequence and duplication event see problem 3.3). Assume that  $|\Sigma| = n$  and that each symbol of  $\Sigma$  occurs in at least one of the input sequences.

*Parameter:*  $k$  and  $r$

*Question:* Is there a common rl-supersequence  $x$  of length at most  $n + r$ ?

Complexity:

NP-complete [FHKS98b].

Parameterized Complexity:

FPT,  $O(k^r \cdot n)$  [FHKS98b].

### 3.5 BOUNDED DUPLICATION SMALLEST COMMON SUPERTREE (SCT) FOR BINARY P-TREES

Biological Motivation:

This definition is reasonable for applications in the study of gene duplication events in the sense that both  $k$  and  $r$  may be small and the input trees complete when complete sequence data is available for all of the species under consideration [FHKS98b].

Problem Definition:

**p-tree:** *Phylogenetic tree* or *p-tree* is a rooted tree where the leaves are labeled from an alphabet  $\Sigma$ , and where no symbol in  $\Sigma$  is used more than once as a label [FHKS98b].

A tree  $T$  contains  $r$  *duplication events* if  $T$  is not a p-tree but the exactly  $r$  leaves must be removed which result in a tree homeomorphic to a p-tree [FHKS98b].

*Instance:* A family of  $k$  complete binary p-trees  $T_i$  with leaf label set  $\Sigma$ ,  $|\Sigma| = n$ , and a positive integer  $r$  representing the number of duplication events.

*Parameter:*  $k$  and  $r$

*Question:* Is there a common binary supertree  $T$  of the  $T_i$  of size at most  $n + r$ ?

Complexity:

NP-complete [FHKS98b].

Parameterized Complexity:

FPT, [FHKS98b].

### 3.6 CLOSEST STRING

Biological Motivation:

**Primer Design:** Primers are short sequences of nucleotides which are designed such that the primer hybridizes to a given DNA sequence (or to all of a given set of DNA sequences) in order to provide a start point for DNA strand synthesis by PCR (polymerase chain reaction). The hybridization of primers depends on complex thermodynamic rules, but is largely determined by the number of “mismatching” positions which should be as small as possible. Designing candidates for primers is a task often done by biological experts using the output of multiple alignment programs which is evaluated by hand [Gra03].



**Motif Search:** A motif is a string that occurs approximately preserved, i.e., with changes in at most  $d$  positions for a fixed integer  $d$ , as a substring in several DNA sequences. Motifs are candidates for substrings of non-coding parts of the DNA sequence that have functions related to, e.g., gene expression [Gra03].

**Consensus Sequences:** Given a collection of related sequences, a consensus sequence is a single sequence that best represents the collection. A challenge associated with creating consensus sequences is sample bias. For example, given a dataset of sequences of orthologous genes from many closely related species and a few more distantly related ones, the resulting consensus sequence could be biased towards sequences from the over-represented species group. One proposed approach to deal with the bias is to create a consensus sequence by minimizing the maximum distance from any sequence rather than minimizing the total distance [BDLPR97] and this task is carried out by CLOSEST STRING problem [LLM<sup>+</sup>03].

Problem Definition:

The Hamming distance between two strings  $s_i$  and  $s_j$ , both of length  $l$ , is given by  $d_H(s_i, s_j) = \{1 \leq p \leq l \mid s_i[p] \neq s_j[p]\}$ .

Herein,  $d_H(s, s_i)$  denotes the *Hamming distance* between strings  $s$  and  $s_i$ .

*Instance:* Strings  $s_1, s_2, \dots, s_k$  over alphabet  $\Sigma$  of length  $L$  each, and non-negative integers  $d$  and  $k$ .

*Parameters:*

1.  $k$
2.  $d$

*Question:* Is there a string  $s$  of length  $L$  such that  $d_H(s, s_i) \leq d$  for all  $i = 1, \dots, k$ ?

Complexity:

NP-complete even for binary alphabet [FL97, LLM<sup>+</sup>03].

Parameterized Complexity:

FPT, when parameterized by  $d$ ,  $O(kL + kd \cdot d^d)$  time [Gra03].

FPT, when parameterized by  $k$  [Gra03].

### 3.7 CLOSEST SUBSTRING

Biological Motivation:

A formal definition of the motif search problem leads to the CLOSEST SUBSTRING problem. These problems are of central importance for sequence analysis in computational molecular biology. These problems have applications in fields such as genetic drug target identification or signal finding [GHN02].

Problem Definition:

*Instance:* Strings  $s_1, s_2, \dots, s_k$  over alphabet  $\Sigma$ , and non-negative integers  $d$  and  $L$ .

*Parameters:*

1.  $L, d$  and  $k$

## 2. $k$

*Question:* Is there a string  $s$  of length  $L$  such that, for every  $i = 1, \dots, k$ , there is a length- $L$  substring  $s'_i$  of  $s_i$  with  $d_H(s, s'_i) \leq d$ ? (by  $d_H(s, s'_i)$  we denote the *Hamming distance* between strings  $s$  and  $s'_i$ , for definition see problem 3.6).

Complexity:

NP-complete [FGN02].

Parameterized Complexity:

$W[1]$ -hard, when parameterized by the combination of  $L$ ,  $d$ , and  $k$ , in case of unbounded alphabet size, by reduction from CLIQUE [FGN02].

$W[1]$ -hard, when parameterized by the number  $k$  of input strings (even over a binary alphabet) [FGN02].

## 3.8 $k$ -CLUSTER EDITING

Biological Motivation:

Novel DNA microarray technologies enable the monitoring of expression levels of thousands of genes simultaneously. This allows a global view on the transcription levels of many (or all) genes when the cell undergoes specific conditions or processes. Analyzing gene expression data requires the clustering of gene into groups with similar expression patterns [SS00].

A key step in the analysis of gene expression data is the identification of groups of genes that manifest similar expression patterns over several conditions. The corresponding algorithmic problem is to cluster multicondition gene expression patterns. The grouping of genes with similar expression patterns into clusters helps in unraveling relations between genes, deducing the function of genes and revealing the underlying gene regulatory network [SS00].

Problem Definition:

*Instance:* An undirected graph  $G = (V, E)$ , and a nonnegative integer  $k$ .

*Parameter:*  $k$

*Question:* Can we transform  $G$ , by deleting and adding at most  $k$  edges, into a graph that consists of a disjoint union of cliques?

Complexity:

NP-complete [Hüf03].

Parameterized Complexity:

FPT,  $O(1.92^k + |V|^3)$  time [Hüf03].

## 3.9 DISTINGUISHING STRING SELECTION (DSS)

Biological Motivation:

DSS problems have the potential to help out in drug target selection. Given a dataset of sequences of orthologous genes from a group of closely related pathogens, and a host (such as humans or

livestock), the goal would be to find an essential sequence that is more conserved in all or most of the pathogens but not as conserved in the hosts. The protein encoded by this fragment could become a target for novel antibiotic development [LLM<sup>+</sup>03].

Another application of DSS problems is with consensus sequences. Given a collection of related sequences, a consensus sequence is a single sequence that best represents the collection. A challenge associated with creating consensus sequences is sample bias. For example, given a dataset of sequences of orthologous genes from many closely related species and a few more distantly related ones, the resulting consensus sequence could be biased towards sequences from the over-represented species group [LLM<sup>+</sup>03].

Finally, DSS problems may also find applications in creating diagnostic probes for bacterial infection and creating universal PCR primers [LLM<sup>+</sup>03].

**Problem Definition:**

Deng *et al.* [DLL<sup>+</sup>02] let all good strings be of same length  $L$ . The terminology “good” and “bad” has its motivation in the application [LLM<sup>+</sup>99] of designing genetic markers to distinguish the sequences of harmful germs (to which the markers should bind) from human sequences (to which the markers should not bind) [GGN03].

*Instance:* Given “good” strings  $s_1, \dots, s_{k_1}$  “bad” strings  $s'_1, \dots, s'_{k_2}$ , positive integers  $d_1$  and  $d_2$ .

*Parameter:*  $d_1$  and  $d_2$

*Question:* Is there an  $s$  “close” to the good strings, i.e.,

$$\max_{i=1, \dots, k_1} d_H(s, s'_i) \leq d_1$$

and “far away” from the bad ones, i.e.,

$$\min_{j=1, \dots, k_2} d_H(s, s'_j) \geq L - d_2?$$

By  $d_H$  we denote the *Hamming distance* for definition see problem 3.6.

**Complexity:**

NP-hard [FL97, LLM<sup>+</sup>99].

**Parameterized Complexity:**

FPT, for fixed alphabet,  $O((k_1 + k_2)L \cdot (\max\{d_1 + 1, (d_2 + 1)(|\Sigma| - 1)\})^{d_1})$  time [Gra03].

### 3.10 FIXED ALPHABET LONGEST COMMON SUBSEQUENCE

**Biological Motivation:**

The computational problem of finding the *longest common subsequence* (LCS) of a set of  $k$  strings has been studied extensively over the last twenty years. This problem has many applications. When  $k = 2$ , the *longest common subsequence* is a measure of the similarity of two strings and is thus useful in molecular biology, pattern recognition, and text compression. The version of *longest common subsequence* in which the number of strings is unrestricted is also useful in text compression, and is a special case of the multiple sequence alignment and consensus subsequence discovery problems in molecular biology [DF99].

**Problem Definition:**

A string  $s$  is a *subsequence* of a string  $r$  if we can delete some characters in  $r$  such that the remaining string is equal to  $s$ .

*Instance:* An alphabet  $\Sigma$  having fixed size; a set of  $k$  strings  $r_1, \dots, r_k$  over the alphabet  $\Sigma$  a positive integer  $m$ .

*Parameters:*

1.  $k$
2.  $k$  and  $m$

*Question:* Is there a string  $s \in \Sigma^*$  of length at least  $m$  that is a subsequence of each  $r_i$ , for  $i = 1, \dots, k$ ?

Complexity:

NP-complete [Mai].

Parameterized Complexity:

$W[1]$ -hard, when parameterized by  $k$ , by reduction from PARTITIONED CLIQUE [Pie03].

FPT, when parameterized by  $k$  and  $m$  (by the trivial algorithm that generates all  $|\Sigma|^m$  possible subsequence strings and checks them against each  $r_i$ ) [Ces04].

Comments:

See also 3.15 problem, LONGEST COMMON SUBSEQUENCE (LCS).

### 3.11 FIXED ALPHABET SHORTEST COMMON SUPERSEQUENCE

Biological Motivation:

Current technology allows only relatively short regions of DNA or protein to be sequenced; hence, the base sequences of longer regions must be determined by breaking such regions into fragments that can be sequenced and then reconstructing the region from these fragments. In much the same way as the LCS problem underlies various versions of multiple sequence alignments and consensus. This problem underlies sequence reconstruction [BDF<sup>+</sup>].

Problem Definition:

A string  $s$  is a *supersequence* of a string  $r$  if we can delete some characters in  $s$  such that the remaining string is equal to  $r$  [Ces04].

*Instance:* An alphabet  $\Sigma$  having fixed size; a set of strings  $\{r_1, \dots, r_k\}$  formed over alphabet  $\Sigma$ ; a positive integer  $\lambda$ .

*Parameters:*

1.  $k$
2.  $\lambda$

*Question:* Does there exist a string  $s \in \Sigma^*$  of length at most  $\lambda$  such that  $s$  is a supersequence of each string  $r_i$ ,  $1 \leq i \leq k$ ?

Complexity:

NP-complete when  $|\Sigma| \geq 2$  [GMS80].

Parameterized Complexity:

$W[1]$ -hard, when parameterized by  $k$ , by reduction from PARTITIONED CLIQUE [Pie03].

FPT, when parameterized by  $\lambda$  [FHK].

Comments:

See also 3.27 problem, SHORTEST COMMON SUPERSEQUENCE (SCS).

### 3.12 GENE DUPLICATION

Biological Motivation:

When trying to resolve the *tree of life* one usually wants to compute the phylogenetic relationships between the organisms based on the data provided by the DNA or protein sequences of families of homologous genes [Ste99].

The problem is the determination of the correct species tree for a set of taxa given a set of possibly contradictory gene trees. Several models for attacking the problem have appeared in the literature including the famous MAXIMUM AGREEMENT SUBTREE (MAST) [Ste99].

The GENE DUPLICATION is the problem of computing the optimal species tree for a given set of gene trees under the *Gene-Duplication Model* [Ste99].

A *species tree* or *evolutionary tree* for a given set of taxa is a complete rooted binary tree built over the set of taxa representing the phylogenetic relationships between the taxa [Ste99].

*Taxa* is a taxonomic group of any rank, including all the subordinate groups. Any group of organisms, populations, or taxa considered to be sufficiently distinct from other such groups to be treated as a separate unit.

*Gene trees* and *species trees* are rooted, binary and leaf labeled [Ste99].

Problem Definition:

*Gene trees* and *species trees* are rooted, binary, and leaf labeled.

$T = (V, E, L)$  is a tree where  $V$  is the vertex set,  $E$  is the edge set, and  $L \subseteq V$  is the leaf-label set (in short, *leafset*).

The vertex  $u$  is such that  $u \in V - L$  and  $T_u$  is the subtree of  $T$  rooted by vertex  $u$ . The root of each tree  $T$  has a left and right subtree, rooted by the two kids of the root  $root(T)$  and denoted by  $T_l$  and  $T_r$ .

The leafset  $L$  of  $T$  is denoted by  $L(T)$ , and a node  $u \in V$  is denoted by  $L(u)$  instead of  $L(T_u)$ .

For trees  $T_1 = (V_1, E_1, L)$  and  $T_2 = (V_2, E_2, L)$  and a vertex  $u \in V_1$  let  $lca_{T_2}(L(v))$  be the least common ancestor of all the leaves in  $L(v)$  in tree  $T_2$ .

Let  $G = (V_G, E_G, L)$  be a gene tree and  $S = (V_S, E_S, L')$ ,  $L \subseteq L'$  be a species tree.

The function  $loc_{G,S} : V_G \mapsto V_S$  associate each vertex in  $G$  with a vertex in  $S$ .

The function  $event_{G,S} : V_G \mapsto \{dup, spec\}$  indicate whether the event in  $G$  corresponds to a duplication or speciation event.

The function  $M$  maps a gene tree  $G$  into a species tree  $S$  by defining functions  $loc_{G,S}$  and  $event_{G,S}$ . The quantity  $cost(G, S) = |\{u \in V_G - L, event_{G,S}(u) = dup\}|$  is the minimum number of gene-duplication events necessary to rectify the gene tree  $G$  with the species tree  $S$ ,

$M(G, S)$ : for each  $u \in V_G - L$ ,  $loc(u) = lca_S(L(u))$  and

$$event(u) = \begin{cases} spec & \text{if } loc_{G,S}(u') \neq loc_{G,S}(u), \text{ for all } u' \text{ where } u' \text{ is a kid of } u \text{ in } G, \\ dup & \text{otherwise.} \end{cases}$$

For given  $G_1, G_2, \dots, G_k$ , and  $S$  let  $cost(G_1, G_2, \dots, G_k, S) = \sum_{i=1}^k cost(G_i, S)$ .

Let  $|L| = n$ .

*Instance*: Gene trees  $G_1, \dots, G_k$  over leaf set  $L$ , and a positive integer  $\mathcal{C}$ .

*Parameter*:  $\mathcal{C}$

*Question*: Does there exist a species tree  $S$  with  $cost(G_1, \dots, G_k, S) \leq \mathcal{C}$ ?

Complexity:

NP-complete [Ste99].

Parameterized Complexity:

FPT,  $O(4^k \cdot n^3 \cdot m^2)$  time [Ste99].

### 3.13 3-HITTING SET

Biological Motivation:

In computational biology the 3-HITTING SET has several applications that go from helping to combine different phylogenetic trees [GW02, NR99] to help into gene regulatory networks [PH].

In phylogenetic when trying to combine different trees, the idea is to model the structure in triples and delete a minimum number of species in order to avoid all conflicts in the tree structures [PH].

Problem Definition:

*Instance*: Collection  $C$  of subsets of size three of a finite set  $S$ , and a positive integer  $k$ .

*Parameter*:  $k$

*Question*: Is there a subset  $S' \subseteq S$  with  $|S'| \leq k$  which allows  $S'$  contain at least one element from each subset in  $C$ ?

Complexity:

NP-complete [NR99].

Parameterized Complexity:

FPT,  $O(2.270^k + n)$  time [NR99].

### 3.14 IMPERFECT PHYLOGENY RECONSTRUCTION

Biological Motivation:

Perfect phylogeny (PP) is a fundamental structure in computational biology, as it describes evolutionary histories in the case that every position is affected by a mutation at most once. The

positions can be pieces of DNA, but also features of phenotypes. The notion of PP can be generalized to more than two characters. Then the condition is that every mutation creates a new character (that never occurred before) at the affected position [Dam].

**Problem Definition:**

A well-known characterization of binary matrices that have a perfect phylogeny (PP matrices for short). A pair of columns is called *complete* if each of the couples 00, 01, 10, 11 appears as a row in the submatrix induced by these two positions [Dam].

A matrix is a PP matrix if and only if it does not contain complete pairs [Dam].

*Instance:* An  $n \times m$  binary matrix and an integer  $k$ .

*Parameter:*  $k$

*Question:* Changing a minimal set of at most  $k$  bits in such matrix so that it becomes a PP matrix.

**Complexity:**

NP-hard if the number of characters is part of the input [BFW92, Ste92].

**Parameterized Complexity:**

FPT,  $O(k6^k nm)$  time [Dam].

### 3.15 LONGEST COMMON SUBSEQUENCE

**Biological Motivation:**

See problem 3.10 for biological motivation.

**Problem Definition:**

*Instance:* An alphabet  $\Sigma$  a set of  $k$  strings  $X_1, \dots, X_k$  over the alphabet  $\Sigma$  a positive integer  $m$ .

*Parameters:*

1. A positive integer  $k$  (LCS-1).
2. A positive integer  $m$  (LCS-2).
3. Positive integers  $k, m$  (LCS-3).
4. Positive integers  $k, |\Sigma|$  (LCS-4).

*Question:* Is there a string  $X \in \Sigma^*$  of length at least  $m$  that is a subsequence (for definition see problem 3.10) of  $X_i$ , for  $i = 1, \dots, k$ ?

**Complexity:**

NP-complete [Mai].

**Parameterized Complexity:**

$W[t]$ -hard for all  $t$  for LCS-1 [BDFW95], by reduction from MONOTONE WEIGHTED T-NORMALIZED SATISFIABILITY [BDFW95, BDFW94, DF99]).

$W[2]$ —hard for LCS-2 in  $W[P]$  [BDFW95], membership is easy; hardness: by reduction from DOMINATING SET [BDFW95, DF99]; in FPT if  $|\Sigma|$  is parameter, by the trivial algorithm that generates all  $|\Sigma|^m$  possible subsequence strings and checks them against each  $r_i$ .

$W[1]$ —complete for LCS-3 [BDFW95], membership: by reduction to WEIGHTED Q-CNF SATISFIABILITY by [BDFW95, BDFW94, DF99]; hardness: by reduction from CLIQUE [BDFW95, BDFW94, DF99].

$W[t]$ —hard for all  $t$  for LCS-4, by reduction from LCS-1 to LCS-4 [BDF<sup>+</sup>, BDFW94]. The reduction required the size of  $\Sigma$  to grow as a function of the parameter.

Comments:

See also 3.10 problem, FIXED ALPHABET LONGEST COMMON SUBSEQUENCE.

### 3.16 $k$ -MAXIMUM AGREEMENT SUBTREE (MAST)

Biological Motivation:

The MAST problem arises naturally in biology and linguistics as a measure of consistency between two evolutionary trees over species and languages, respectively. It is often difficult to determine the true phylogeny for a set of taxa, and one way to gain confidence in a particular tree is to have different lines of evidence supporting that tree. In the biological taxa case, one may construct trees from different parts of the DNA of species. These are known as *gene trees*. For many reasons, these trees need not entirely agree, and so one is left with the task of finding a consensus of the various gene trees. The MAXIMUM AGREEMENT SUBTREE is one method of arriving at such a consensus [CCH<sup>+</sup>].

Therefore, the parameter  $k$  is the number of species to exclude from analysis [AGN01].

Problem Definition:

*Instance:* A set of rooted trees  $T_1, \dots, T_r$  ( $r \geq 3$ ) with the leaf set of each  $T_i$  labeled 1 : 1 with a set of species  $X$ , and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a subset  $S \subseteq X$  of size at most  $k$  such that  $T_i$  restricted to the leaf set  $X' = X - S$  is the same (up to label-preserving isomorphism and ignoring vertices of degree 2) for  $i = 1, \dots, r$ ?

Complexity:

NP-complete [DFS99].

Parameterized Complexity:

FPT,  $O(2.270^k + rn^3)$  time [AGN01].

### 3.17 $k$ -MINIMUM QUARTET INCONSISTENCY (MQI)

Biological Motivation:

An application of MINIMUM QUARTET INCONSISTENCY problem in biology is the reconstruction of evolutionary tree from biological data between quartet paradigm [VJLW02].



*Quartet methods* infer the evolutionary tree only for four taxa, called a *quartet*. Once having determined the evolutionary tree for every quartet of taxa, they try to combine these evolutionary trees involving four taxa, called *quartet topologies*, in order to obtain a tree containing all taxa [Gra03].

Problem Definition:

*Phylogeny* or *phylogenetics* is the classification of species and organisms according to their evolutionary relationships. In *molecular phylogenetics*, this classification is based on genomic data. The single units being compared, usually species, are referred to as *taxa*. Given a set of taxa, a commonly used model for their evolutionary relationship is a tree called *phylogenetic tree* in which the leaves are in one-to-one correspondence to the taxa and in which inner nodes correspond to (unknown) ancestors of these taxa [Gra03].

Herein, we consider an *evolutionary tree* to be an unrooted binary tree  $T$  in which the leaves are bijectively labeled by a set of taxa  $S$  [GN01].

A quartet, then, is a size four subset  $\{a, b, c, d\}$  of  $S$ , and the topology for  $\{a, b, c, d\}$  induced by  $T$  simply is the four leaf subtree of  $T$  induced by  $\{a, b, c, d\}$ . The three possible quartet topologies for  $\{a, b, c, d\}$  are  $[ab|cd]$ ,  $[ac|bd]$ , and  $[ad|bc]$ , the fourth possible topology would be the star topology, which is not considered here because it is not binary [GN01].

*Instance:* A set  $S$  of  $n$  taxa and a set  $Q_S$  of  $\binom{n}{4}$  quartet topologies such that there is exactly one topology for *every* quartet set corresponding to  $S$ , and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there an evolutionary tree  $T$  where the leaves are bijectively labeled by the elements from  $S$  such that the set of quartet topologies induced by  $T$  differs from  $Q_S$  in at most  $k$  quartet topologies?

Complexity:

NP-complete [GN01].

Parameterized Complexity:

FPT,  $O(4^k \cdot n + n^4)$  time [GN01].

### 3.18 $k$ -MISMATCH

Biological Motivation:

See problem 3.6 for biological motivation.

Problem Definition:

Let  $s_{i,p,L}$  denote the length- $L$  substring of a given string  $s_i$  at position  $p$ .

*Instance:* Given strings  $s_1, s_2, \dots, s_m$  of length  $n$ , and integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a string  $s$  of length  $L$  and a position  $p$  with  $1 \leq p \leq n - L + 1$ , such that  $d_H(s, s_{i,p,L}) \leq k$  for all  $i = 1, \dots, m$ ? (by  $d_H$  we denote the *Hamming Distance* for definition see problem 3.6).

Complexity:

NP-hard [FL97, LLM<sup>+</sup>99].

Parameterized Complexity:

FPT,  $O(mL + (n - L)mk \cdot k^k)$  time [Gra03].

### 3.19 MODIFIED DISTINGUISHING SUBSTRING SELECTION (MDSSS)

Biological Motivation:

See problem 3.9 for biological motivation.

Problem Definition:

*Instance:* Given an alphabet  $\Sigma = \{0, 1\}$ , two sets of strings over  $\Sigma$ ,

- $S_g = \{s_1, \dots, s_{k_g}\}$ , each string of length at least  $L$  (the “good” strings),
- $S_b = \{s'_1, \dots, s'_{k_b}\}$ , each string of length at least  $L$  (the “bad” strings), and two non-negative integers  $d_g$  and  $d_b$

*Parameter:*  $d_g$  and  $d_b$

*Question:* Is there a length- $L$  string  $s$  over  $\Sigma$  such that,

- in every  $s_i \in S_g$  for *every* length- $L$  substring  $t_i$ ,  $d_H(s, t_i) \geq d_g$  and
- every  $s'_i \in S_b$  has *at least one* length- $L$  substrings  $t'_i$  with  $d_H(s, t'_i) \leq d_b$ ?

By  $d_H$  we denote the *Hamming distance* for definition see problem 3.6.

Complexity:

DISTINGUISHING SUBSTRING SELECTION problem is NP-complete [GGN03].

Parameterized Complexity:

FPT,  $O(L \cdot k_g + ((d'_g)^2 k_g + N \sqrt{L \log L}) \cdot (d'_g)^{d'_g})$  time where  $N = \sum_{s'_i \in S_b} |s'_i|$  is the total size of the bad strings [GGN03].

### 3.20 MULTIPLE GENE DUPLICATION

Biological Motivation:

A fundamental problem in computational biology is the determination of the correct *species tree* for a set of taxa given a set of (possibly contradictory) *gene trees*. In recent literature, the *Duplication Loss* model has received considerable attention. Here one measures the similarity/dissimilarity between a set of gene trees by counting the number of *paralogous gene duplications* and subsequent *gene losses* which need to be postulated in order to explain (in an evolutionarily meaningful way) how the gene trees could have arisen with respect to the species tree. Here we count the number of *multiple gene duplication events* (duplication events in the genome of the organism involving one or more genes) without regard to gene losses [FHS98].

The MULTIPLE GENE DUPLICATION asks to find the species tree  $S$  which requires the fewest number of multiple gene duplication events to be postulated in order to explain a set of gene trees  $G_1, G_2, \dots, G_k$  [FHS98].

A *duplication event* in the genome of an organism involves a stretch of DNA where one or more genes may reside. [FHS98].

Problem Definition:

*Instance:* Set of gene trees  $G_1, \dots, G_k$ , a species tree  $S$  (for definitions see problem 3.12), and integer  $\mathcal{C}$ .

*Parameter:*  $\mathcal{C}$

*Question:* Do there exist functions  $loc_{G_i, S}, event_{G_i, S}$  (for definitions see problem 3.12), for  $1 \leq i \leq k$ , s.t.  $S$  receives  $G_1, \dots, G_k$  with at most  $\mathcal{C}$  multiple gene duplications?

Complexity:

NP-complete [FHS98].

Parameter Complexity:

$W[1]$ -hard, by reduction from combinatorial model called the BALL AND TRAP GAME [FHS98].

### 3.21 $k$ -PATHWIDTH

Biological Motivation:

In order to study a genome, several copies of it are cut or broken down, and some of the resulting shorter segments (called *clones*) are preserved for further analysis. Depending on the technique used, the preserved clones may have variable length, or they may all have essentially the same length. In the process of producing the clones, all information on their relative position along the DNA chain is lost. The goal of physical mapping of DNA is to reconstruct that order, based on experimental data on the overlaps between pairs of clones [KS96].

An important feature of real biological data is that the “width” of the map is consistently very small: The largest number of mutually overlapping clones is typically between 5 and 15, compared to a total number of clones in the thousands [KS96].

Problem Definition

A *path decomposition* of a given graph  $G = (V, E)$  is a sequence of subsets of  $V$ ,  $X = (X_1, \dots, X_l)$  such that:

1.  $V = \cup_i X_i$
2. For each edge  $(u, v) \in E$ , there exists some  $i \in \{1, \dots, l\}$  so that both  $u$  and  $v$  belong to  $X_i$ .
3. For each  $v \in V$  there exist some  $s(v), e(v) \in \{1, \dots, l\}$  so that  $s(v) \leq e(v)$ , and  $v \in V_j$  if and only if  $j \in \{s(v), s(v) + 1, \dots, e(v)\}$ .

The *width* of  $X$  is defined by  $pw_X(G) = \max\{|X_i| | i = 1, \dots, l\} - 1$ . The *pathwidth* of  $G$ , denoted  $pw(G)$ , is the minimum value of  $pw_X(G)$  over all path decompositions, i.e.,

$pw(G) = \min\{pw_X(G) | X \text{ is a path decomposition of } G\}$

The *pathwidth* problem is to decide for a given graph  $G$  and a given integer  $k$  if  $pw \leq k$  [KS96].

*Instance:* A graph  $G = (V, E)$ , and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is the pathwidth of  $G$  no more than  $k$ ?

Complexity:

NP-complete [ACP87].

Parameterized Complexity:

FPT,  $O(2^{k^2}n)$  time [BK96, Bod96, BT98].

### 3.22 PERFECT PHYLOGENY

Biological Motivation:

Infer the evolutionary history of a set of species is a fundamental problem in biology. Each of such that set of species is specified by the set of *traits of characters* that exhibits. All information about evolutionary history can be conveniently represented by an *evolutionary tree* or *phylogenetic tree*, and often referred as a *phylogeny* [AFB96, VLM].

Problem Definition:

*Instance:* A set  $C = \{1, \dots, m\}$  of characters; for each  $c \in C$ , a set  $A_c = \{1, \dots, r_c\}$  of states; and a set  $S \subseteq A_1 \times \dots \times A_m$  where  $|S| = n$  ( $S$  represents a set of  $n$  species).

*Parameters:*

1.  $r = \max_{c \in C} r_c$
2.  $r = \max_{c \in C} r_c, m$

*Question:* Is there a tree  $T$  with the properties:

1.  $S \subseteq V(T) \subseteq A_1 \times \dots \times A_m$ .
2. Every leaf in  $T$  is in  $S$ .
3. For each  $c \in C$  and each  $j \in A_c$ , the set of vectors  $v \in V(T)$  such that  $v_c = j$  induces a subtree of  $T$ ?

Complexity:

NP-complete [BFW92, Ste92].

Parameterized Complexity:

FPT, when parameterized by  $r$ ,  $O(2^{3r}(nm^3 + m^4))$  time [AFB94].

FPT, when parameterized by  $r$  and  $m$ ,  $O((r - n/m)^m rnm)$  time [AFB96].

Comments:

This problem is also known as the CHARACTER COMPATIBILITY PROBLEM.

This problem is also related with problem 4.11, TRIANGULATING  $k$ -COLORED GRAPHS.

### 3.23 REMOVING EXTRA COLUMNS IN IMPERFECT PHYLOGENY RECONSTRUCTION

Problem Definition:

*Instance:* An  $n \times m$  binary matrix and an integer  $k$ .

*Parameter:*  $k$

*Question:* Deleting at most most  $k$  columns such that the rest matrix has a PP (for definition see problem 3.14).

Complexity:

?

Parameterized Complexity:

FPT,  $O(k^2nm + k^22^k)$  time [Dam].

### 3.24 REMOVING EXTRA ROWS IN IMPERFECT PHYLOGENY RECONSTRUCTION

Problem Definition:

*Instance:* An  $n \times m$  binary matrix and an integer  $k$ .

*Parameter:*  $k$

*Question:* Finding all minimal combinations of at most  $k$  rows whose deletion leaves a PP matrix (for definition see problem 3.14).

Complexity:

?

Parameterized Complexity:

FPT,  $O(3^knm)$  time [Dam].

### 3.25 SCSS FOR P-SEQUENCES

Biological Motivation:

See problem 3.11 for biological motivation.

Problem Definition:

*Instance:* p-sequences (for definition see problem 3.3)  $x_1, \dots, x_k$  and a positive integer  $M$ .

*Parameter:*  $k$

*Question:* Is there a rl-sequence (for definition see problem 3.4)  $x$ , with  $|x| \leq M$  and  $x_i$  is a subsequence (for definition see problem 3.10) of  $x_i$  for  $i = 1, \dots, k$ ?

Complexity:

NP-complete [FHKS98b].

Parameterized Complexity:

$W[1]$ -hard [FHKS98a], by reduction from the CLIQUE problem [FHKS98b].

### 3.26 SCT FOR P-SEQUENCES

Biological Motivation:

In computational biology the question arises how to resolve the species tree for a given set of trees such that the number of paralogous duplications is minimized [FHK98b].

Problem Definition:

**rl-tree:** It is a rooted tree with leaves labeled from  $\Sigma$ , where labels may be repeated [FHK98b].

*Instance:* Binary p-trees (for definition see problem 3.5)  $T_1, \dots, T_k$  and a positive integer  $m$ .

*Parameter:*  $k$

*Question:* Is there an rl-tree  $T$ , with  $|T| \leq m$  and  $T_i$  is contained in  $T$  by topological containment that respects ancestry with label isomorphism at the leaves for  $i = 1, \dots, k$ ?

Complexity:

NP-complete [FHK98b].

Parameterized Complexity:

$W[1]$ -hard, by reduction from the CLIQUE problem [FHK98b].

### 3.27 SHORTEST COMMON SUPERSEQUENCE (SCS)

Biological Motivation:

See problem 3.11 for biological motivation.

Problem Definition:

*Instance:* An alphabet  $\Sigma$ ; a set of strings  $\{r_1, \dots, r_k\}$  formed over alphabet  $\Sigma$ ; a positive integer  $\lambda$ .

*Parameters:*

1.  $k, |\Sigma|$
2.  $\lambda$

*Question:* Does there exist a string  $s \in \Sigma^*$  of length at most  $\lambda$  such that  $s$  is a supersequence (for definition see problem 3.3) of each string  $r_i$ ,  $1 \leq i \leq k$ ?

Complexity:

NP-complete [Mai].

Parameterized Complexity:

$W[t]$ -hard for all  $t$ , when parameterized by  $k, |\Sigma|$  [FHK, Hal].

FPT, when parameterized by  $\lambda$  [Hal96].

Comments:

See also 3.11 problem, FIXED ALPHABET SHORTEST COMMON SUPERSEQUENCE.

### 3.28 STEINER TREE

Biological Motivation:

Phylogeny construction from molecular sequence data is a prominent application of the notion of a minimal Steiner Tree [HRW92, FHP79]. This is due to the use of the notion of a most parsimonious tree to formalize the biological problem of reconstructing the evolutionary history of a set of sequences. A most parsimonious tree is a tree whose leaves are labeled with the given sequences and where sequences are assigned to the inner nodes in such a way that the overall number of mutations along the tree edges is minimized [SV97].

Problem Definition:

*Instance:* A Graph  $G = (V, E)$ , a set  $S$  of at most  $k$  vertices in  $V$ , an integer  $m$ .

*Parameters:*

1.  $k$
2.  $m$

*Question:* Is there a set of vertices  $T \subseteq V - S$  such that  $|T| \leq m$  and  $G[S \cup T]$  is connected?

Complexity:

NP-complete by a reduction from EXACT COVER [GKR, GJ79], (Problem number ND12).

Parameterized Complexity:

FPT, when parameterized by  $k$ ,  $O(3^k n + 2^k n^2 + n^3)$  time [DW71]

$W[2]$ -hard, when parameterized by  $m$ , by a reduction from DOMINATING SET( $k$ ) in [DF95].

### 3.29 STEINER TREE IN HYPERCUBES

Biological Motivation:

The STEINER PROBLEM FOR HYPERCUBES is of interest to biologists in the computation of phylogenetic trees under the criterion of minimum evolution/maximum parsimony. The set  $S$  corresponds to a set of species, and the binary vectors correspond to information about the species, each component recording the answer to some question (as 0 or 1), such as: “Does it have wings?” or “Is there a thymine at a certain position in the DNA sequence?” [DFS99].

Problem Definition:

**$q$ -dimensional binary hypercube:** all binary sequence of length  $q$ .

**Edges:** two nodes are adjacent if the  $d_H(X_1, X_2) = 1$  (by  $d_H(X_1, X_2)$  we denote the *Hamming distance* between strings  $X_1$  and  $X_2$ , for definition see problem 3.6).

*Instance:* Binary sequences  $X_1, \dots, X_k$ , where each  $X_i$  has length  $q$ ; a positive integer  $M$  encoded in binary.

*Parameter:*  $k$

*Question:* Is there a subgraph  $S$  of the  $q$ -dimensional binary hypercube that includes the vertices  $X_1, \dots, X_k$ , such that  $S$  has at most  $M$  edges?

Complexity:

NP-complete [DFS99].

Parameterized Complexity:

FPT, by the reduction to problem kernel method [Ces04, DF99].

### 3.30 SUBSTRING PARSIMONY (SP)

Biological Motivation:

The SUBSTRING PARSIMONY problem (SPP) is a formalization of the phylogenetic footprinting, which is a technique that identifies regulatory elements by finding unusually well conserved regions in a set of orthologous non-coding DNA sequences from multiple species. Most of these regulatory elements are relatively short stretches of DNA (5 to 25 nucleotide-long), located in the non-coding sequence surrounding a gene. Most known transcription factor binding sites are located 5' of the coding region, but some are also found in the 3' sequence, and even in introns. In all these cases, REs are located in otherwise non-functional sequences [BST02].

This technique uses this functional/non-functional sequence dichotomy to identify regulatory elements. Functional sequences tend to evolve much slower than non-functional sequences, as they are subject to selective pressure. It is the difference in mutation rates that phylogenetic footprinting exploits. To identify regulatory elements associated with a given gene, one will consider a set of orthologous non-coding sequences from a group of related species. If these sequences contain unusually well conserved regions, it is a good conjecture that these regions have some regulatory function [BST02].

Problem Definition:

Parsimony score of a set of sequences is the minimum total number of substitutions over the tree  $T$  needed to explain the observed sequences [BST02, BST00].

*Instance:* A set of orthologous sequences (same string in different species)  $S_1, \dots, S_n$ , one from each of  $n$  related species, the phylogenetic tree  $T = (V, E)$  relating these species, the size  $k$  of the motifs (substrings), and an integer  $d$ .

*Parameters:*

1.  $k$
2.  $d$

*Question:* Find all sets of substrings  $s_1, \dots, s_n$  of  $S_1, \dots, S_n$  respectively, each of size  $k$ , such that the parsimony score of  $s_1, \dots, s_n$  on  $T$  is at most  $d$ .

Complexity:

NP-hard [Aku98, Bla00]

Parameterized Complexity:

FPT,  $O(n \cdot \min\{l \cdot (3k)^{d/2}, N \cdot (4^k + l)\})$  time [BST02].

If each string has  $N$  neighbors,  $O(n \cdot \min\{l \cdot N^{d/2}, N \cdot (4^k + l)\})$  time [BST02].



### 3.31 $k$ -VERTEX COVER

Biological Motivation:

It is naturally that in computational biology, the data sets are often incomplete or faulty. It is frequently, to formulate the corresponding problem of cleaning up data as a covering problem [NR99].

Given a set of experimental data points, some of which are in conflict. Is possible to determine a minimum size set of data points such that, if “deleted” from the experimental data, this would remove or explain all inconsistencies? [NR99].

Problem Definition:

A vertex cover is a subset  $V' \subseteq V$  such that  $\forall (v, w) \in E, v \in V'$  or  $w \in V'$ .

*Instance:* A graph  $G = (V, E)$ , and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Does  $G$  have a vertex cover of size at most  $k$ ?

Complexity:

NP-complete [GJ79], (Problem number GT1).

Parameterized Complexity:

FPT,  $O(1.271^k + kn)$  time [CKJ99].

### 3.32 Open

#### 3.32.1 CLOSEST STRING

Biological Motivation:

See problem 3.6 for biological motivation.

Problem Definition:

*Instance:* Strings  $s_1, s_2, \dots, s_k$  over alphabet  $\Sigma$  of length  $L$  each, and a non-negative integer  $d$ .

*Parameters:*

1.  $d$  and  $k$
2.  $d$

*Question:* Is there a string  $s$  of length  $L$  such that  $d_H(s, s_i) \leq d$  for all  $i = 1, \dots, k$ ? (by  $d_H$  we denote the *Hamming distance* for definition see problem 3.6).

Complexity:

NP-complete [dlHC00].

Parameterized Complexity:

1. Algorithm in [Gra03], however, suffers from huge constant factors in the running time, even for moderate values of  $k$ , that seem to make it impossible to find exact solutions with this algorithm for  $k > 4$ . It is possible to give a fixed-parameter algorithm for parameter  $k$  that is usable for larger values of  $k$  and arbitrary values of  $L$  and  $d$ ? [Gra03].
2. CLOSEST STRING is considered with respect to Hamming distance. What is, for constant alphabet size, the parameterized complexity of CLOSEST STRING with respect to parameter  $d$  when using edit distance instead, i.e., allowing insertions, deletions, and substitutions? [Gra03].

### 3.32.2 CLOSEST SUBSTRING

Biological Motivation:

A formal definition of the motif search problem leads to the CLOSEST SUBSTRING problem. These problems are of central importance for sequence analysis in computational molecular biology. These problems have applications in fields such as genetic drug target identification or signal finding [GHN02].

Problem Definition:

*Instance:* Strings  $s_1, s_2, \dots, s_k$  over alphabet  $\Sigma$ , and non-negative integers  $d$  and  $L$ .

*Parameters:*

1.  $d$  and  $k$
2.  $d$

*Question:* Is there a string  $s$  of length  $L$  such that, for every  $i = 1, \dots, k$ , there is a length- $L$  substring  $s'_i$  of  $s_i$  with  $d_H(s, s'_i) \leq d$ ? (by  $d_H$  we denote the *Hamming distance* for definition see problem 3.6).

Complexity:

NP-complete [FGN02].

Parameterized Complexity:

In the case of constant alphabet size, the complexity of the problem remains open when parameterized by  $d$  and  $k$  together, or by  $d$  alone [FGN02].

### 3.32.3 GENE DUPLICATION AND LOSS

Biological Motivation:

See problem 3.12 for biological motivation.

A biological cost model which has received considerable attention is the *Gene Duplication and Loss* model. The basic idea is to measure the similarity/dissimilarity between a set of gene trees by counting the number of postulated paralogous gene duplications and subsequent gene losses required to explain (in evolutionary meaningful way) how the gene trees could have arising with respect to the species tree [Ste99].

Problem Definition:

See problem 3.12 for definition of *species tree*, *gene trees* and *cost* model.

*Instance:* Gene trees  $T_1, \dots, T_k$ .

*Parameters:*

1.  $m$  and  $k$
2.  $m$

*Question:* Does there exist a species tree  $S$  with  $\text{cost}(T_1, \dots, T_k, S) \leq m$ ?

Complexity:

NP-complete [Ste99].

Open problem:

In [Ste99] suspect the problem to be in FPT when parameterized by both the *number of duplication and loss events* ( $m$ ) and the *number of gene trees* ( $k$ ).

In [Ste99] conjecture the DUPLICATION AND LOSS problem to be  $W[1]$ -hard when parameterized by the *number of duplications and losses* ( $m$ ) only.

## 4 Sequencing and Mapping

### 4.1 ARC PRESERVING LONGEST COMMON SUBSEQUENCE (LAPCS)

Biological Motivation:

Molecular biologists use algorithms that compare and otherwise sequences that represent genetic and protein molecules. However, most of these algorithms, operate on the basic sequence and do not incorporate the additional information that is often known about the molecule and its pieces. The annotation schemes include adding colors and arcs to the sequence, and these arcs can be used to link sequence symbols or colored substrings to indicate molecular bonds or other relationships. Adding these annotations to sequence analysis problems such as sequence alignment or finding the longest common subsequence can make the problem more complex, often depending on the complexity of the annotation scheme [Eva99].

The arcs represent a few types of information that go naturally with these restrictions, and produce five different levels of allowed arc structure for the problem [Eva99].

The term “*plain*” refers to sequences without arcs, “*crossing*” denotes arc structures where no two arcs share an endpoint, and “*unlimited*” refers to a completely unrestricted arc structure. With these terms, it is possible to define various versions of LAPCS where LAPCS(TYPE1, TYPE2) refers to the case in which input sequence  $S_1$  has an arc structure of TYPE1 and  $S_2$  has an arc structure of TYPE2 [Gra03].

Problem Definition:

**Annotation:** The descriptive text that accompanies a sequence in a database record [Eva99].

An **arc** is a directed edge  $(p_1, p_2) \in P \times P$ , where  $P$  is the set of positions in the sequence. If  $n$  is the length of the sequence,  $P = \{1, \dots, n\}$ . An arc can be viewed as a link that connects two symbols that are part of the same sequence. The order of the pair  $(p_1, p_2)$  should be consistent with the sequence order, so  $p_1 < p_2$  [Eva99].

A sequence  $y$  is a **common subsequence** of sequences  $S_1$  and  $S_2$  if  $y$  is a subsequence of  $S_1$  and  $y$  is a subsequence of  $S_2$  [Eva99].

An **annotation scheme** is a system of representing additional information (beyond that found in the basic sequence) in a way that relates it to the basic sequence [Eva99]. An individual *annotation* for a specific sequence is its associated additional information, as represented according to the chosen annotation scheme (pag "2"-11)

A **basic sequence** is the sequence of base symbols that form the fundamental, unannotated sequence. Mathematically, an *alphabet* is a set of symbols, generally represented by  $\Sigma$  [Eva99].

For a sequence  $S$  of length  $|S| = n$ , an **arc annotation (or arc set)**  $A$  of  $S$  is a set of pairs of numbers from  $\{1, 2, \dots, n\}$ . Each pair  $(i, j)$  connects the two *bases*  $S[i]$  and  $S[j]$  at positions  $i$  and  $j$  in  $S$  by an arc [Gra03].

**Arc annotation (or arc set):**  $A$  of  $S$  is a set of unordered pairs of number from  $\{1, 2, \dots, n\}$ . Each pair  $(i, j)$  connects the two *bases*  $S[i]$  and  $S[j]$  at positions  $i$  and  $j$  in  $S$  by an arc [Eva99].

*Instance:* The target length  $k$ , and the pair of annotated sequences  $(S_1, P_1)$  and  $(S_2, P_2)$ . These annotated pairs consist of the sequences  $S_1$  and  $S_2$  over some fixed alphabet  $\Sigma$ , with arc annotations  $P_1 \subset \{1, \dots, |S_1|\}^2$  and  $P_2 \subset \{1, \dots, |S_2|\}^2$ . The length of  $S_1$  and  $S_2$  are  $n$  and  $m$  respectively.

*Parameters:*

1.  $l$ , length of desired subsequence.
2.  $s$ , levels of nested arcs (for non-crossing arcs).
3.  $k$ , cutwidth of arc structure.
4.  $d$ , bandwidth of arc structure (so  $(i_2 - i_1) \leq d$  for any arc  $(i_1, i_2)$ ).
5. The length of desired subsequence  $l$  is independent of the other parameters. The others are related;  $s = k$ , and  $k \leq d$  for all restriction levels except *unlim*.

*Question:* Finding a common subsequence of length  $k$  which preserves induced arcs.

Complexity:

NP-complete [Gra03].

Parameterized Complexity:

FPT, LAPCS(CROSS, CROSS), when parameterized by cutwidth  $k$ ,  $O(9^k nm)$  time [Eva99].

FPT, LAPCS(CROSS, CROSS), when parameterized by bandwidth  $d$ ,  $O(9^d nm)$  time [Eva99].

FPT, LAPCS(NESTED, NESTED), when parameterized by nesting depth  $s$ , with modifications to take advantage of non-crossing arcs,  $O(s^2 4^s nm)$  time [Eva99].

FPT, LAPCS(NEST, NEST), when parameterized by bandwidth  $d$ ,  $O(d^2 4^d nm)$  time [Eva99].

## 4.2 COLORED PROPER INTERVAL GRAPH COMPLETION

Biological Motivation:

Suppose a set of clones is obtained by complete digestion of the genome by one or more restriction enzymes. Since the digestion is complete, in such a set, no two clones will overlap. Consider a Physical Mapping project in which the set of clones consists of equal length clones, and it is composed of several subsets of clones, where each subset is obtained by a complete digest with a

different set of enzymes. One would like to reconstruct the map from clone overlap data, in the presence of “false negative” errors, i.e., some overlaps which are not detected experimentally. One wishes to construct a map which is as close as possible to our input data, i.e., it assumes as few errors as possible [GGK<sup>+</sup>95].

A biologically motivated restriction of RESTRICTED COMPLETION TO A PROPER INTERVAL GRAPH COMPLETION WITH BOUNDED CLIQUE SIZE is defined by the graph and a  $k$ -coloring  $c$  of it, and the requirement that the set of added edges must not violate the coloring [KST94].

**Problem Definition:**

A graph  $G = (V, E)$  is a *supergraph* of the graph  $G' = (V', E')$  if  $V = V'$  and  $E \supseteq E'$ .

*Instance:* A graph  $G = (V, E)$ , and a  $k$ -coloring of  $G$ .

*Parameter:*  $k$

*Question:* Is there a supergraph  $G'$  of  $G$  which is a proper interval graph and has clique size at most  $k$ , and no edge in  $G'$  connects two vertices in  $G$  with the same color?

**Complexity:**

NP-complete [AS99, GGK<sup>+</sup>95].

NP-complete for colored caterpillars of hair length 2 and in  $P$  for caterpillars of hair length 1 or 0, by reduction from the MULTIPROCESSOR SCHEDULING problem [AS99].

**Parameterized Complexity:**

$W[1]$ -hard, by a parameterized reduction from INDEPENDENT SET [KS96].

**Comments:**

This problem is equivalent to COLORED UNIT INTERVAL GRAPH COMPLETION, as the class of unit interval graphs and proper interval graphs are equivalent [Ces01].

See also 4.9 problem, RESTRICTED COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE [Ces01].

### 4.3 COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE

**Biological Motivation:**

Most work on Physical Mapping with errors has involved heuristics. Imposing an additional constraint, prevalent in real biological data, leads to a polynomial-time problem: The “width” of a map (or of a set of interval on the line) is the largest number of mutually overlapping clones. In the corresponding interval graph  $G$ , this is its clique size, denoted  $\omega(G)$ . Typical biological maps have width between 5 and 15, even when the total number of clones is in the thousands [KST94].

This problem is motivated by the situation where overlap information for pairs of clones (intervals) may be definite yes, definite no or undetermined [KST94].

**Problem Definition:**

*Instance:* Given a graph  $G = (V, E)$  and a constant  $k$ .

*Parameter:*  $k$

*Question:* Does there exist a supergraph (for definition see problem 4.2)  $G'$  of  $G$  which is a proper interval graph and has clique size at most  $k$ ?

Complexity:

NP-hard [KST94].

Parameterized Complexity:

$W[t]$ -hard for any  $t > 0$  [KST94].

This problem is equivalent to BANDWIDTH problem [KS96], by reduction from UNIFORM EMULATION ON A PATH problem.

Comments:

This problem is a completion problem, but instead of bounding the number of added edges, we bound the clique size of the map. Here even the existence of a polynomial algorithm for fixed  $k$  is not obvious [KST94].

This problem is equivalent to decide whether the proper pathwidth of  $G$  is not greater than  $k - 1$  [KST94].

#### 4.4 CONSENSUS PATTERN

Biological Motivation:

Applications for the consensus word analysis of DNA, RNA, or protein sequences include locating binding sites and finding conserved regions in unaligned sequences for genetic drug target identification, for designing genetic probes, and for universal PCR primer design. These problems can be regarded as various generalizations of the common substring problem, allowing errors. This leads to CLOSEST SUBSTRING and CONSENSUS PATTERN, where errors are modeled by the (Hamming) distance parameter  $d$  [Gra03].

Problem Definition:

*Instance:* Strings  $s_1, s_2, \dots, s_k$  over alphabet  $\Sigma$ , and a non-negative integer  $d$  and  $L$ .

*Parameters:*

1.  $k$
2.  $d$  and  $L$

*Question:* Is there a string  $s$  of length  $L$ , and, for every  $i = 1, \dots, k$ , a length- $L$  substring  $s'_i$  of  $s_i$  such that  $\sum_{i=1}^k d_H(s, s'_i) \leq d$ ? (by  $d_H$  we denote the *Hamming distance* for definition see problem 3.6).

Complexity:

NP-complete [FGN02].

Parameterized Complexity:

$W[1]$ -hardness, by reduction from CLIQUE, results as for CLOSEST SUBSTRING given unbounded alphabet size [FGN02].

$W[1]$ -hard, when parameterized by the number  $k$  of strings, for a binary alphabet [FGN02].

## 4.5 INTERVALIZING COLORED GRAPHS OR DNA PHYSICAL MAPPING

Biological Motivation:

This problem models a problem arising in sequence reconstruction, which appears in some investigations in molecular biology (such as protein sequencing, nucleotide sequencing and gene sequencing). A sequence  $X$  (usually a large piece of DNA) is fragmented (or  $k$  copies of the sequence  $X$  are fragmented) such that the fragments can be further analyzed. The information about the order of the fragments in the original sequence is lost during the fragmentation process. The objective of DNA physical mapping is to reconstruct this order. To this end, a set of characteristics is determined for each fragment (list ‘fingerprint’ or ‘signature’), and based on respective fingerprints, an ‘overlap’ measure is computed. Using this overlap information, the fragments are assembled into islands of contiguous fragments (contigs) [BdF95].

Instances of ICG model the situation where  $k$  copies of  $X$  are fragmented, and some fragments (clones) are known to overlap. Fragments of the same copy of  $X$  will not overlap. Now each vertex in  $V$  represents one fragment; the color of a vertex represents to which copy of  $X$  the fragment belongs. It can be seen that ICG (and specially the constructive version of ICG, which also outputs an interval model of the interval graph  $G'$ ) helps here to predict other overlaps and to work towards reconstruction of the sequence  $X$  [BdF95].

Problem Definition:

*Instance:* A graph  $G = (V, E)$  and a coloring  $c : V \rightarrow \{1, \dots, k\}$ ; and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a supergraph (for definition see problem 4.2)  $G' = (V, E')$  of  $G$  which is an interval graph and has clique size at most  $k$ , and no edge in  $G'$  connects two vertices in  $G$  with the same color?

Complexity:

NP-complete for four or more colors (for any fixed number of colors  $\geq 4$ ) even when the graph is a caterpillar tree, colored with  $k \geq 4$  colors [BdF95, ADS01, BFH<sup>+</sup>00].

Parameterized Complexity:

W[t]-hard for all  $t \in \mathbb{N}$ , by reduction from COLORED CUTWIDTH (CC-1) [BFH94, BFH<sup>+</sup>00].

Comments:

1. ICG is closely related to TRIANGULATING COLORED GRAPH (TCG) [BdF95].
2. The PROPER PATH DECOMPOSITION (PPD) is equivalent to INTERVALIZING COLORED GRAPHS (ICG) [BdF95].

## 4.6 MINIMUM FILL-IN

Biological Motivation:

The MINIMUM FILL-IN problem is very important in the area of computational biology called perfect phylogeny [DF99].

This problem is to decide if a graph can be triangulated by adding at most  $k$  edges. Is to find a minimum triangulating (fill-in) of a given graph [KST99]. The importance of this problem lies

mainly in the fact that it is equivalent to finding an order of Gaussian elimination steps of a (usually sparse) symmetric matrix, minimizing the number of generated non-zero entries [BKMM].

This problem is also known as CHORDAL COMPLETION problem [KST99], and there are studied variants of the completion problem, motivated by DNA mapping, in which the input graph is pre-colored and the required supergraph also obeys the coloring [NSS01].

**Problem Definition:**

*Instance:* A graph  $G = (V, E)$  and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Can we add no more than  $k$  edges to  $G$  and cause  $G$  to become chordal?

**Complexity:**

NP-complete [Yan81].

**Parameter Complexity:**

FPT,  $O(k^2 mn + k^6 2^{4k})$  time [KST99].

**Comments:**

This problem is also known as CHORDAL GRAPH COMPLETION problem [KST99].

## 4.7 $k$ -INTERVAL POSITIONAL SEQUENCING BY HYBRIDIZATION (INTERVAL PSBH)

**Biological Motivation:**

In SEQUENCING BY HYBRIDIZATION (SBH), one has to reconstruct a sequence from its  $l$ -long substring. SBH was proposed as an alternative to gel-based DNA sequencing approaches, but in its original form the method is not competitive. POSITIONAL SBH (PSBH) is a recently proposed enhancement of SBH in which one has additional information about the possible positions of each substring along the target sequence [Pe'02, BDPSS01].

In PSBH additional information is gathered concerning the position of the  $l$ -mers in the target sequence. More precisely, for each  $l$ -mer in the spectrum its allowed positions along the target are registered [Pe'02, BDPSS01].

**Problem Definition:**

**$p$ -spectrum** The  $p$ -spectrum of a string  $X \in \Sigma^*$  is the multi-set of all  $p$ -long substrings of  $X$  [Pe'02, BDPSS01].

**$p$ -long:** It is a substring of length  $p$  [Pe'02, BDPSS01].

*Instance:* A multi-set  $S$  of  $p$ -long strings. For each  $s \in S$ , a set  $P(s)$  which is a sub-interval of  $\{0, |S| - 1\}$ .

*Parameter:*  $k$

*Question:* Is  $S$  the  $p$ -spectrum of some string  $X$ , such that for each  $s \in S$  its positions along  $X$  is in  $P(s)$ ?

**Complexity:**

NP-complete, even if all sets of allowed positions are intervals of equal length, by reduction from INTERVAL POSITIONAL EULERIAN PATH (PEP) problem [Pe'02, BDPSS01].



Parameterized Complexity:

FTP,  $O(mk^{1.5}4^k)$  time [Pe'02].

Comments:

The parameter  $k$  is an upper bound on the sizes of the intervals of allowed positions for each edge [Pe'02, BDPSS01].

## 4.8 PROPER INTERVAL GRAPH COMPLETION (PIGC)

Biological Motivation:

This problem is to find the smallest edge set whose addition to the input graph will form a proper interval graph. Interval completion problems arise in molecular biology and in the Human Genome Project: In physical mapping of DNA, a set of long contiguous intervals of the DNA chain (called *clones*) is given together with experimental on their pairwise overlaps. The goal is to build a map describing the relative position of the clones [KST94].

The biologically important case is where all clones have equal length. In the presence of “false negative” errors (unidentified overlaps) the problem of building a map with fewest errors is equivalent to PROPER INTERVAL GRAPH COMPLETION (PIGC) [KST94].

Problem Definition:

*Instance:* A graph  $G = (V, E)$ , and a positive integer  $k$ .

*Parameter:*  $k$ .

*Question:* Does there exist a set of no more than  $k$  edges, whose addition to the input graph will form a proper interval graph?

Complexity:

NP-hard [GKS94].

Parameterized Complexity:

FPT, when  $k$  is all minimal triangulations of a graph  $G$  and  $m$  is the edge set,  $O(2^{4k}m)$  time [KST94].

## 4.9 RESTRICTED COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE

Biological Motivation:

See problem 4.3 for biological motivation.

Problem Definition:

*Instance:* A graph  $G = (V, E)$ ; a set  $E' \subseteq V \times V$  of forbidden edges; a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Is there a  $G' \supset G$  which is a proper interval graph, has clique size at most  $k$ , and  $G'$  has no edges from  $E'$ ?

Complexity:

NP-Complete [KST94].

Parameterized Complexity:

$W[t]$ -hard for all  $t$  [KST94, KS96].

It remains  $W[t]$ -hard even when  $E' = \emptyset$  [KS96].

Comments:

This problem is a generalization of COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE [KST94].

#### 4.10 STRONGLY CHORDAL COMPLETION

Biological Motivation:

See problem 4.6 for biological motivation.

Problem Definition:

A graph is *strongly chordal* if it admits a strong elimination ordering [KST99].

*Instance:* Given a graph  $G = (V, E)$  and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Does there exist an edge set  $A$  such that  $|A| \leq k$  and  $G = (V \cup A)$  is strongly chordal graph?

Complexity:

NP-hard [KST99].

Parameterized Complexity:

FPT,  $O(8^{2k} m \log n)$  time [KST99].

#### 4.11 TRIANGULATING $k$ -COLORED GRAPHS

Biological Motivation:

Infer the evolutionary history of a set of species is a fundamental problem in biology. Each of such that set of species is specified by the set of *traits of characters* that exhibits. All information about evolutionary history can be conveniently represented by an *evolutionary tree* or *phylogenetic tree*, and often referred as a *phylogeny* [AFB96, VLM].

Problem Definition:

*Instance:* A graph  $G = (V, E)$ , a vertex coloring  $c : V \rightarrow \{1, \dots, k\}$ , and a positive integer  $k$ .

*Parameter:*  $k$

*Question:* Does there exist a supergraph (for definition see problem 4.2)  $G' = (V', E')$  where  $E \subseteq E'$ ,  $G'$  is properly colored by  $c$ , and  $G'$  is triangulating?

Complexity:

NP-complete [BFH<sup>+</sup>00].

Parameterized Complexity:

$W[t]$ -hard for all  $t$ , the perfect phylogeny algorithm leads to an  $O((2e/k)^k e^2 k)$  algorithm for triangulating a  $k$ -colored graph [AFB96], by reduction from LONGEST COMMON SUBSEQUENCE when parameterized by  $k$  [BFH94].

Comments:

This problem is related with 3.22 problem, PERFECT PHYLOGENY.

## 4.12 Open

### 4.12.1 CONSENSUS PATTERN

Biological Motivation:

Applications for the consensus word analysis of DNA, RNA, or protein sequences include locating binding sites and finding conserved regions in unaligned sequences for genetic drug target identification, for designing genetic probes, and for universal PCR primer design. These problems can be regarded as various generalizations of the common substring problem, allowing errors. This leads to CLOSEST SUBSTRING and CONSENSUS PATTERN, where errors are modeled by the (Hamming) distance parameter  $d$  [Gra03].

Problem Definition:

*Instance:* Strings  $s_1, s_2, \dots, s_k$  over alphabet  $\Sigma$ , and a non-negative integer  $d$  and  $L$ .

*Parameter:*  $d$

*Question:* Is there a string  $s$  of length  $L$ , and, for every  $i = 1, \dots, k$ , a length- $L$  substring  $s'_i$  of  $s_i$  such that  $\sum_{i=1}^k d_H(s, s'_i) \leq d$ ? (by  $d_H$  we denote the *Hamming distance* for definition see problem 3.6).

Complexity:

NP-complete [FGN02].

Parameterized Complexity:

Parameterized by “distance parameter”  $d$ , the complexity remains open for alphabets of constant size [FGN02].

## 5 ?

### 5.1 Multicut in Trees

Biological Motivation:

Problem Definition:

*Instance:* An undirected Tree  $T = (V, E)$ ,  $n = |V|$ , a collection  $H$  of  $m$  pairs of nodes in  $V$ ,  $H = \{(u_i, v_i) | u_i, v_i \in V, 1 \leq i \leq m\}$ , and an integer  $k$ .

*Parameter:*  $k$

*Question:* Does there exist a subset  $E' \leq k$  of  $E$  such that the removal of edges in  $E'$  separates each pair of nodes in  $H$ ?

Complexity:

NP-complete [GVY97].

Parameterized Complexity:

? [Nie04].

## 6 Acknowledgements

## Glossary

**Alignment** A one-to-one matching of two sequences, so that each character in a pair of sequences is associated with a single character of the other sequence or with a gap. Alignments are often displayed as two rows with a third row in between indicating levels of similarity.

**Chromosome** The self-replicating genetic structures of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins.

**Clone** Contiguous chain of DNA.

**Consensus** A single sequence that represents, at each subsequent position, the variation found within corresponding columns of a multiple sequence alignment.

**Contig** A set of overlapping sequence fragments that represent a large piece of DNA, usually a genomic region from a particular chromosome.

**DNA** The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus the base sequence of each single strand can be deduced from that of its partner.

**DNA sequencing** Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

**Dichotomy** Successive division and subdivision, as of a stem of a plant or a vein of the body, into two parts as it proceeds from its origin; successive bifurcation.

**Enzyme** Proteins that act as catalysts, speeding the rate at which biochemical reactions proceed but not altering the direction or nature of the reactions.

**Evolution** A change in the genetic composition of a population over time.

**Evolutionary Tree** It is a two-dimensional graph showing evolutionary relationships among organisms, or in the case of sequences, in certain genes from separate organisms. The separate sequences are referred to as taxa (singular taxon), defined as phylogenetically distinct units on the tree. The tree is composed of outer branches (or leaves) represented as sequences.

**False Negative** A negative data point collected in a data set that was incorrectly reported due to a failure of the test in avoiding negative results.

**False positive** A positive data point collected in a data set that was incorrectly reported due to a failure of the test. If the test had correctly measured the data point, the data would have been recorded as negative.

**Fingerprint** A set of characteristics for each fragment.

**Fission** One chromosome splits into two.

**Fusion** Two chromosomes merge into one.

**Gap** Mismatch in the alignment of two sequences caused by either an insertion in one sequence or a deletion in the other.

**Gene** A segment of DNA (a locus on a chromosome) that serves as the basic unit of biological inheritance. It includes a region that is transcribed into RNA as well as flanking regulatory sequences. A Discrete subunit of the DNA molecule.

**Gene Expression** Biochemical process which genes are read.

**Gene Tree** A tree based on different parts of the DNA of species.

**Genome** All of the genetic material in a cell or an organism.

**Genotype** The genetic constitution of an organism. Compare phenotype.

**Haplotype** A combination of alleles (for different genes) which are located closely together on the same chromosome and which tend to be inherited together.

**Hybridization** The process of joining two complementary strands of DNA or one each of DNA and RNA to form a double- stranded molecule.

**Homologous Genes** Two genes with a common ancestor. A pair of genes from different but related species which correspond to each other and which are identical or very similar to each other.

**Human Genome Project** Collective name for several projects begun in 1986 by the Department of Energy (DOE) to create an ordered set of DNA segments from known chromosomal locations, develop new computational methods for analyzing genetic map and DNA sequence data, and develop new techniques and instruments for detecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The national effort, led by DOE and National Institute of Health (NIH), is known as the Human Genome Project.

**Indel** An insertion or deletion in a sequence alignment.

**Intron (intervening sequence)** A segment of DNA that is transcribed, but removed from the mRNA by a splicing reaction before translation into protein occurs.

**Maximum Parsimony** The minimum number of evolutionary steps required to generate the observed variation in a set of sequences, as found by comparison of the number of steps in all possible phylogenetic trees.

**Mismatch** In an alignment, two corresponding symbols that are not the same.

**Motif** A region within a group of related protein or DNA sequences that is evolutionary conserved- presumably due to its functional importance.

**Mutation** A heritable change in DNA sequence resulting from mutagens. Various types of mutations include frame-shift mutations, missense mutations, and nonsense mutations.

**Nucleotide** A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule.

**Orthologous Genes** A gene from one species which corresponds to a gene in another species that is related via a common ancestral species (a homologous gene), but which has evolved to become different from the gene of the other species.

**Overlap** *Buscar*

**Pattern** *Buscar*

**Pathogen** Organism which can cause disease in another organism.

**Pattern Recognition** It aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space.

**Parsimony** The principle that the hypothesis that requires the fewest assumptions is the most likely to be true (i.e., the most defensible hypothesis).

**PCR** (Polymerase Chain Reaction). A method of repeatedly copying segments of DNA using short oligonucleotide primers (10-30 bases long) and heat stable polymerase enzymes in a cycle of heating and cooling so as to produce an exponential increase in the number of target fragments.

**Phenotype** The physical appearance/observable characteristics of an organism. See genotype.

**Phylogenetic** The field of biology that deals with the relationships between organisms. It includes the discovery of these relationships and the study of the causes behind these patterns.

**Phylogeny** The evolutionary history of an organism as it is traced back, connecting through shared ancestors to lineages of other organisms.

**Physical Map** A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest-resolution physical map is the banding patterns on the 24 different chromosomes; the highest-resolution map would be the complete nucleotide sequence of the chromosomes.

**Primer** A short DNA (or RNA) fragment that can anneal to a single-stranded template DNA to form a starting point for DNA polymerase to extend a new DNA strand complementary to the template, forming a duplex DNA molecule.

**Protein** A large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs, and each protein has unique functions. Examples are hormones, enzymes, and antibodies.

**Protein sequencing** Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

**Quartet** A quadruple of taxa, with an associated topology —a partition of the four taxa into two pairs of taxa. This subdivision expresses the most likely topology induced by the underlying  $n$  taxa phylogeny.

**RNA** (Ribonucleic Acid) A chemical found in the nucleus and cytoplasm of cells; it plays an important role in protein synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including messenger RNA, transfer RNA, ribosomal RNA, and other small RNAs, each serving a different purpose.

**Sequence** The order in which subunits appear in a chain, such as amino acids in a polypeptide or nucleotide bases in a DNA or RNA molecule.

**Sequence Alignment** It is the procedure of comparing two (pair-wise alignment) or more (multiple sequence alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.

**Signature** A set of characteristics for each fragment.

**Single Nucleotide Polymorphism (SNP)** DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered.

**Species** Groups of populations (which are groups of individuals living together that are separated from other such groups) which can potentially interbreed or are actually interbreeding, that can successfully produce viable, fertile offspring (without the help of human technology). The species is the most fundamental unit of evolution and is the most specific taxonomic level.

**Syntenic** Two genes appearing in the same chromosome.

**Synteny** The presence of a set of homologous genes in the same order on two genomes.

**Systematics** The process of classification of organisms into a formal hierarchical system of groups (taxa).

**Taxa** A named group of related organisms identified by systematics. The single units being compared, usually species.

**Translocation** Two chromosomes exchange contiguous blocks (usually prefixes or suffixes) of genes.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Genome Rearrangement</b>	<b>1</b>
2.1	$k$ -BREAKPOINT MEDIAN . . . . .	1
2.2	SORTING BY REVERSALS . . . . .	2
2.3	SYNTENIC DISTANCE . . . . .	3
2.4	VERTEX BIPARTIZATION . . . . .	4
2.5	Open . . . . .	5
2.5.1	EDGE BIPARTIZATION . . . . .	5
<b>3</b>	<b>Sequence Alignment and Evolution</b>	<b>5</b>
3.1	BINARY CLADISTIC CHARACTER COMPATIBILITY . . . . .	5
3.2	BINARY QUALITATIVE CHARACTER COMPATIBILITY . . . . .	6
3.3	BOUNDED DUPLICATION SHORTEST COMMON SUPERSEQUENCE (SCS) FOR COMPLETE P-SEQUENCES . . . . .	7
3.4	BOUNDED DUPLICATION SCS FOR P-SEQUENCES . . . . .	7
3.5	BOUNDED DUPLICATION SMALLEST COMMON SUPERTREE (SCT) FOR BINARY P-TREES . . . . .	8
3.6	CLOSEST STRING . . . . .	8
3.7	CLOSEST SUBSTRING . . . . .	9
3.8	$k$ -CLUSTER EDITING . . . . .	10
3.9	DISTINGUISHING STRING SELECTION (DSS) . . . . .	10
3.10	FIXED ALPHABET LONGEST COMMON SUBSEQUENCE . . . . .	11
3.11	FIXED ALPHABET SHORTEST COMMON SUPERSEQUENCE . . . . .	12
3.12	GENE DUPLICATION . . . . .	13
3.13	3-HITTING SET . . . . .	14
3.14	IMPERFECT PHYLOGENY RECONSTRUCTION . . . . .	14
3.15	LONGEST COMMON SUBSEQUENCE . . . . .	15
3.16	$k$ -MAXIMUM AGREEMENT SUBTREE (MAST) . . . . .	16
3.17	$k$ -MINIMUM QUARTET INCONSISTENCY (MQI) . . . . .	16
3.18	$k$ -MISMATCH . . . . .	17
3.19	MODIFIED DISTINGUISHING SUBSTRING SELECTION (MDSSS) . . . . .	18
3.20	MULTIPLE GENE DUPLICATION . . . . .	18
3.21	$k$ -PATHWIDTH . . . . .	19
3.22	PERFECT PHYLOGENY . . . . .	20

3.23	REMOVING EXTRA COLUMNS IN IMPERFECT PHYLOGENY RECONSTRUCTION . .	20
3.24	REMOVING EXTRA ROWS IN IMPERFECT PHYLOGENY RECONSTRUCTION . . . .	21
3.25	SCSs FOR P-SEQUENCES . . . . .	21
3.26	SCT FOR P-SEQUENCES . . . . .	22
3.27	SHORTEST COMMON SUPERSEQUENCE (SCS) . . . . .	22
3.28	STEINER TREE . . . . .	23
3.29	STEINER TREE IN HYPERCUBES . . . . .	23
3.30	SUBSTRING PARSIMONY (SP) . . . . .	24
3.31	$k$ -VERTEX COVER . . . . .	25
3.32	Open . . . . .	25
3.32.1	CLOSEST STRING . . . . .	25
3.32.2	CLOSEST SUBSTRING . . . . .	26
3.32.3	GENE DUPLICATION AND LOSS . . . . .	26
<b>4</b>	<b>Sequencing and Mapping</b>	<b>27</b>
4.1	ARC PRESERVING LONGEST COMMON SUBSEQUENCE (LAPCS) . . . . .	27
4.2	COLORED PROPER INTERVAL GRAPH COMPLETION . . . . .	28
4.3	COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE . .	29
4.4	CONSENSUS PATTERN . . . . .	30
4.5	INTERVALIZING COLORED GRAPHS OR DNA PHYSICAL MAPPING . . . . .	31
4.6	MINIMUM FILL-IN . . . . .	31
4.7	$k$ -INTERVAL POSITIONAL SEQUENCING BY HYBRIDIZATION (INTERVAL PSBH) . .	32
4.8	PROPER INTERVAL GRAPH COMPLETION (PIGC) . . . . .	33
4.9	RESTRICTED COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE . . . . .	33
4.10	STRONGLY CHORDAL COMPLETION . . . . .	34
4.11	TRIANGULATING $k$ -COLORED GRAPHS . . . . .	34
4.12	Open . . . . .	35
4.12.1	CONSENSUS PATTERN . . . . .	35
<b>5</b>	<b>?</b>	<b>35</b>
5.1	Multicut in Trees . . . . .	35
<b>6</b>	<b>Acknowledgements</b>	<b>36</b>

## Alphabetical Index

ARC PRESERVING LONGEST COMMON SUBSEQUENCE (LAPCS) .....	27
BINARY CLADISTIC CHARACTER COMPATIBILITY .....	5
BINARY QUALITATIVE CHARACTER COMPATIBILITY .....	6
BOUNDED DUPLICATION SHORTEST COMMON SUPERSEQUENCE (SCS) FOR COMPLETE P-SEQUENCES 7	
BOUNDED DUPLICATION SCS FOR P-SEQUENCES .....	7
BOUNDED DUPLICATION SMALLEST COMMON SUPERTREE (SCT) FOR BINARY P-TREES ....	8
$k$ -BREAKPOINT MEDIAN .....	1
CLOSEST STRING .....	8
CLOSEST SUBSTRING .....	9
$k$ -CLUSTER EDITING .....	10
COLORLED PROPER INTERVAL GRAPH COMPLETION .....	28
COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE .....	29
CONSENSUS PATTERN .....	30
DISTINGUISHING STRING SELECTION .....	10
FIXED ALPHABET LONGEST COMMON SUBSEQUENCE .....	11
FIXED ALPHABET SHORTEST COMMON SUPERSEQUENCE .....	12
GENE DUPLICATION .....	13
3-HITTING SET .....	14
IMPERFECT PHYLOGENY RECONSTRUCTION .....	14
$k$ -INTERVAL POSITIONAL SEQUENCING BY HYBRIDIZATION (INTERVAL PSBH) .....	32
INTERVALIZING COLORED GRAPHS OR DNA PHYSICAL MAPPING .....	31
LONGEST COMMON SUBSEQUENCE (LCS) .....	15
$k$ -MAXIMUM AGREEMENT SUBTREE (MAST) .....	16
MINIMUM FILL-IN .....	31
$k$ -MINIMUM QUARTET INCONSISTENCY (MQI) .....	16
$k$ -MISMATCH .....	17
MODIFIED DISTINGUISHING SUBSTRING SELECTION .....	18
MULTIPLE GENE DUPLICATION .....	18
$k$ -PATHWIDTH .....	19
PERFECT PHYLOGENY .....	20
PROPER INTERVAL GRAPH COMPLETION (PIGC) .....	33
REMOVING EXTRA COLUMNS IN IMPERFECT PHYLOGENY RECONSTRUCTION .....	20
REMOVING EXTRA ROWS IN IMPERFECT PHYLOGENY RECONSTRUCTION .....	21

RESTRICTED COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE	33
SCSSs FOR P-SEQUENCES .....	21
SCT FOR P-SEQUENCES .....	22
SHORTEST COMMON SUPERSEQUENCE (SCS) .....	22
SORTING BY REVERSALS .....	2
STEINER TREE .....	23
STEINER TREE IN HYPERCUBES .....	23
STRONGLY CHORDAL COMPLETION .....	34
SUBSTRING PARSIMONY .....	24
SYNTENIC DISTANCE .....	3
TRIANGULATING $k$ -COLORED GRAPHS .....	34
VERTEX BIPARTIZATION .....	4
$k$ -VERTEX COVER .....	25

# Hierarchical Index

## FPT

ARC PRESERVING LONGEST COMMON SUBSEQUENCE (LAPCS) .....	27
BOUNDED DUPLICATION SHORTEST COMMON SUPERSEQUENCE (SCS) FOR COMPLETE P-SEQUENCES	
7	
BOUNDED DUPLICATION SCS FOR P-SEQUENCES .....	7
BOUNDED DUPLICATION SMALLEST COMMON SUPERTREE (SCT) FOR BINARY P-TREES ....	8
$k$ -BREAKPOINT MEDIAN .....	1
CLOSEST STRING .....	8
$k$ -CLUSTER EDITING .....	10
DISTINGUISHING STRING SELECTION .....	10
FIXED ALPHABET LONGEST COMMON SUBSEQUENCE $(k, m)$ .....	11
FIXED ALPHABET SHORTEST COMMON SUPERSEQUENCE $(m)$ .....	12
GENE DUPLICATION .....	13
3-HITTING SET .....	14
IMPERFECT PHYLOGENY RECONSTRUCTION .....	14
$k$ -INTERVAL POSITIONAL SEQUENCING BY HYBRIDIZATION (INTERVAL PSBH) .....	32
$k$ -MAXIMUM AGREEMENT SUBTREE (MAST) .....	16
$k$ -MINIMUM QUARTET INCONSISTENCY (MQI) .....	16
MINIMUM FILL-IN .....	31
$k$ -MISMATCH .....	17
MODIFIED DISTINGUISHING SUBSTRING SELECTION .....	18
$k$ -PATHWIDTH .....	19
PERFECT PHYLOGENY .....	20
REMOVING EXTRA COLUMNS IN IMPERFECT PHYLOGENY RECONSTRUCTION .....	20
REMOVING EXTRA ROWS IN IMPERFECT PHYLOGENY RECONSTRUCTION .....	21
SHORTEST COMMON SUPERSEQUENCE (SCS) $(\lambda)$ .....	22
SORTING BY REVERSALS .....	2
STEINER TREE $(k)$ .....	23
STEINER TREE IN HYPERCUBES .....	23
STRONGLY CHORDAL COMPLETION .....	34
SUBSTRING PARSIMONY .....	24
SYNTENIC DISTANCE .....	3
VERTEX BIPARTIZATION .....	4
$k$ -VERTEX COVER .....	25

## $W[1]$

INTERVALIZING COLORED GRAPHS OR DNA PHYSICAL MAPPING .....	31
--	----

## $W[1]$ -complete

BINARY CLADISTIC CHARACTER COMPATIBILITY .....	5
BINARY QUALITATIVE CHARACTER COMPATIBILITY .....	6
LONGEST COMMON SUBSEQUENCE (LCS) $(k, m)$ .....	15

## $W[1]$ -hard

CLOSEST SUBSTRING $(k)$ .....	9
CLOSEST SUBSTRING $(L, d, k)$ .....	9
COLORED PROPER INTERVAL GRAPH COMPLETION .....	28
CONSENSUS PATTERN .....	30
FIXED ALPHABET LONGEST COMMON SUBSEQUENCE $(k)$ .....	11
FIXED ALPHABET SHORTEST COMMON SUPERSEQUENCE $(k)$ .....	12
MULTIPLE GENE DUPLICATION .....	18
SCSS FOR P-SEQUENCES .....	21
SCT FOR P-SEQUENCES .....	22

## $W[2]$ -hard

LONGEST COMMON SUBSEQUENCE (LCS) $(m)$ .....	15
STEINER TREE $(m)$ .....	23

## $W[t]$ -hard for all $t$

LONGEST COMMON SUBSEQUENCE (LCS) $(k)$ .....	15
LONGEST COMMON SUBSEQUENCE (LCS) $(k,  \Sigma )$ .....	15
SHORTEST COMMON SUPERSEQUENCE (SCS) $(k)$ .....	22
PROPER INTERVAL GRAPH COMPLETION (PIGC) .....	33
RESTRICTED COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE	33
TRIANGULATING $k$ -COLORED GRAPHS .....	34

**$W[t]$ -hard for all  $t \in \mathbb{N}$**

INTERVALIZING COLORED GRAPHS OR DNA PHYSICAL MAPPING ..... 31

**$W[t]$ -hard for any  $t > 0$**

COMPLETION TO A PROPER INTERVAL GRAPH WITH BOUNDED CLIQUE SIZE ..... 29

## **Open**

CLOSEST STRING  $(L, d, k)$  FOR  $k > 4$  ..... 25

CLOSEST SUBSTRING  $(d)$  ..... 26

CLOSEST SUBSTRING  $(d, k)$  OR  $(d)$  ..... 26

CONSENSUS PATTERN  $(d)$  FOR FIXED ALPHABET ..... 35

EDGE BIPARTIZATION ..... 5

GENE DUPLICATION AND LOSS ..... 26

## References

- [ACP87] S. Arnborg, D. J. Cornil, and A. Proskurowski. Complexity of finding embedding in a  $k$ -tree. *SIAM J. Alg. Disc. Meth.*, 8:227–284, 1987.
- [ADS01] C. Alvarez, J. Diaz, and M. Serna. The Hardness of Intervalizing Four Colored Caterpillars. *Discrete Math.*, 25(235):19–27, 2001.
- [AFB94] R. Agarwala and D. Fernández-Baca. A polynomial time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM Journal on Computing*, 23:1216–1224, 1994.
- [AFB96] R. Agarwala and D. Fernández-Baca. Fast and simple algorithms for perfect phylogeny and triangulating colored graphs. *J. Foundations of Comp. Sc.*, 7(1):11–22, 1996.
- [AGN01] Jochen Alber, Jens Gramm, and Rolf Niedermeier. Faster exact algorithms for hard problems: a parameterized point of view. *Discrete Math.*, 229(1-3):3–27, 2001. Combinatorics, graph theory, algorithms and applications.
- [AK94] A. Amir and D. Keselman. Maximum Agreement Subtree in a Set of Evolutionary Trees - Metrics and Efficient Algorithms. In *In: 35th Annual Symposium on Foundations of Computer Science*, pages 758–769, 1994.
- [Aku98] T. Akutsu. Hardness results on gapless local multiple sequence alignment. Technical Report 98-MPS-24-2, Information Processing Society of Japan, 1998.
- [AS99] C. Alvarez and M. Serna. The Proper Interval Colored Graph problem for caterpillar trees. Technical report, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica Catalunya, 1999.
- [BDF<sup>+</sup>] Hans Bodlaender, Rodney G. Downey, Michael R. Fellows, Michael T. Hallett, and H. Todd Wareham. Parameterized Complexity Analysis in Computational Biology. *Computer Applications in the Biosciences*, 11(1995):49–57.
- [BdF95] Hans L. Bodlaender and Babette de Fluiter. Intervalizing  $k$ -colored graphs. In *Automata, languages and programming (Szeged, 1995)*, volume 944 of *Lecture Notes in Comput. Sci.*, pages 87–98. Springer, Berlin, 1995.
- [BDFW94] H. L. Bodlaender, R. G. Downey, M. R. Fellows, and H. T. Wareham. The parameterized complexity of sequence alignment and consensus (extended abstract). In *Proceedings of the Fourth Conference on Combinatorial Pattern Matching (CPM'94)*, 1994.
- [BDFW95] Hans L. Bodlaender, Rodney G. Downey, Michael R. Fellows, and Harold T. Wareham. The parameterized complexity of sequence alignment and consensus. *Theoret. Comput. Sci.*, 147(1-2):31–54, 1995.
- [BDLPR97] A. Ben-Dor, G. Lancia, J. Perone, and R. Ravi. Banishing bias from consensus sequences, Combinatorial Pattern Matching. *8th Annual Symposium*, pages 247–261, 1997.
- [BDPSS01] A. Ben-Dor, I. Pe'er, R. Shamir, and R. Sarna. On the complexity of positional sequencing by Hybridization. Technical Report TR01-054, 2001.
- [BFH94] H. L. Bodlaender, M. R. Fellows, and M. T. Hallett. Beyond NP-completeness for problems of bounded width: Hardness for the W hierarchy (extended abstract). In *Proc. 26th Annual ACM Symposium on Theory of Computing*, pages 449–458. Association of Computing Machinery, Academic Press, May 1994.



- [BFH<sup>+</sup>00] Hans L. Bodlaender, Michael R. Fellows, Michael T. Hallett, H. Todd Wareham, and Tandy J. Warnow. The hardness of perfect phylogeny, feasible register assignment and other problems on thin colored graphs. *Theoret. Comput. Sci.*, 244(1-2):167–188, 2000.
- [BFW92] Hans L. Bodlaender, Mike R. Fellows, and Tandy J. Warnow. Two strikes against perfect phylogeny. In *Automata, languages and programming (Vienna, 1992)*, volume 623 of *Lecture Notes in Comput. Sci.*, pages 273–283. Springer, Berlin, 1992.
- [BK96] Hans L. Bodlaender and Ton Kloks. Efficient and Constructive Algorithms for the Pathwidth and Treewidth of Graphs. *Journal of Algorithms*, 21(2):358–402, 1996.
- [BKKM] H. L. Bodlaender, T. Kloks, D. Kratsch, and H. Müller. Treewidth and minimum fill-in on  $d$ -trapezoid graphs. *Journal on Graph Algorithms and Applications*, 2(1998):1–23.
- [Bla00] M. Blanchette. An exact algorithm to identify motifs in orthologous sequences from multiple species. Technical Report Qualification project, University of Washington, 2000.
- [Bod96] Hans L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM J. Comput.*, 25(6):1305–1317, 1996.
- [BST00] M. Blanchette, B. Schwikowski, and M. Tompa. An exact algorithm to identify motifs in orthologous sequences from multiple species. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 37–45, 2000.
- [BST02] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *Journal of Computational Biology*, 9(2):211–224, 2002.
- [BT98] Hans L. Bodlaender and Dimitrios M. Thilikos. Computing Small Search Numbers in Linear Time. Technical Report UU-CS-1998-05, Dept. of Computer Science, Utrecht University, 1998.
- [CCH<sup>+</sup>] Richard Cole, Martin Farach Colton, Ramesh Hariharan, Teresa Przytycka, and Mikkel Thorup. A  $O(N \log N)$  Algorithm for the Maximum Agreement Subtree Problem for Binary Trees. *SIAM Journal of Computing*, 30(5):1385–1404. (2000).
- [Ces01] Marco Cesati. Compendium of Parameterized Problems. *Department of Computer Science, Systems, and Industrial Engineering, University of Rome “Tor Vergata”*, 22 February 2001.
- [Ces04] Marco Cesati. Compendium of Parameterized Problems. *Department of Computer Science, Systems, and Industrial Engineering, University of Rome “Tor Vergata”*, 8 January 2004.
- [CKJ99] J. Chen, I.A. Kanj, and W. Jia. Vertex Cover: Further Observations and Further Improvements. In *Proceedings of the 25th International Workshop on Graph Theoretic Concepts in Computer Science (WG’99)*, volume 1665 of *Lecture Notes in Computer Science*, pages 313–324, 1999.
- [Dam] Peter Damaschke. Full Fixed-Parameter Tractability and Phylogeny Reconstruction.
- [DF95] Rod G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness I. Basic results. *SIAM J. Comput.*, 24(4):873–921, 1995.
- [DF99] R. G. Downey and M. R. Fellows. *Parameterized complexity*. Monographs in Computer Science. Springer-Verlag, New York, 1999.

- [DFS99] Rodney G. Downey, Michael R. Fellows, and Ulrike Stege. Computational tractability: the view from Mars. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, (69):73–97, 1999.
- [DJK<sup>+</sup>97] B. DasGupta, T. Jiang, S. Kannan, M. Li, and Z. Sweedyk. On the complexity and approximation of syntenic distance. In *Proceedings of the first annual international conference on Computational molecular biology*, pages 99–108. ACM Press, 1997.
- [DJS86] William H. E. Day, David S. Johnson, and David Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.*, 81(1):33–42, 1986.
- [dHHC00] C. de la Higuera and F. Cascuberta. Topology of strings: Median String is NP-complete. *Theoretical Computer Science*, 230(1–2):39–48, 2000.
- [DLL<sup>+</sup>02] X. Deng, G. Li, Z. Li, B. Ma, and L. Wang. A PTAS for Distinguishing (Sub)string Selection. In *Proc. of the 29th ICALP*, number 2380, pages 740–751, 2002.
- [DS86] W. Day and D. Sankoff. Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 35(2):224–229, 1986.
- [DW71] S. Dreyfus and R. Wagner. The Steiner Problem in Graphs. *NETWORKS*, (1):195–207, 1971.
- [Eva99] Patricia Anne Evans. *Algorithms and Complexity for Annotated Sequence Analysis*. PhD thesis, Department of Computer Science, University of Victoria, Canada, 1999.
- [FGN02] M. R. Fellows, J. Gramm, and R. Niedermeier. On the parameterized intractability of Closest Substring and related problems. In *Proc. Of 19th STACS*, number 2285 in LNCS, pages 262–273. Springer, 2002.
- [FHK] M.R. Fellows, M.T. Hallett, and D. Kirby. The parameterized complexity of shortest common supersequences.
- [FHKS98a] M. R. Fellows, M. T. Hallett, C. Korostensky, and U. Stege. The complexity of problems on sequences and trees. Technical report, ETH-Zurich, 1998.
- [FHKS98b] Michael Fellows, Michael Hallett, Chantal Korostensky, and Ulrike Stege. Analogs and duals of the MAST problem for sequences and trees. In *Algorithms—ESA '98 (Venice)*, volume 1461 of *Lecture Notes in Comput. Sci.*, pages 103–114. Springer, Berlin, 1998.
- [FHP79] L. R. Foulds, M. D. Hendy, and D. Penny. A graph theoretic approach to the development of minimal phylogenetic trees. *Journal of Molecular Evolution*, 13:127–149, 1979.
- [FHS98] Michael Fellows, Michael Hallett, and Ulrike Stege. On the multiple gene duplication problem. In *Algorithms and computation (Taejon, 1998)*, volume 1533 of *Lecture Notes in Comput. Sci.*, pages 347–356. Springer, Berlin, 1998.
- [FL97] M. Frances and A. Litman. On covering problems of codes. *Theory of Computing Systems*, 30:113–119, 1997.
- [GGK<sup>+</sup>95] P. W. Goldberg, M. C. Golumbic, H. Kaplan, , and R. Shamir. Four strikes against physical mapping of DNA. Technical report, Computer Science Dept., Tel Aviv University, 1995. *Journal of Computational Biology*.
- [GGN03] Jens Gramm, Jiong Guo, and Rolf Niedermeier. On exact and approximation algorithms for Distinguishing Substring Selection. In *Proceedings of the 14th International Symposium on Fundamentals of Computation Theory (FCT)*, number 2751 in LNCS, pages 195–209. Springer, 2003.

- [GHN02] J. Gramm, F. Hüffner, and R. Niedermeier. Closest strings, primer design, and motif search. In L. Florea, B. Walenz, and S. Hannenhalli, editors, *Currents in Computational Molecular Biology 2002*, pages 74–75, 2002.
- [GJ79] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W.H. Freeman and Co, San Francisco, 1979.
- [GKR] W. Gasarch, M. Krentel, and K. Rappoport. OptP as the Normal Behavior of NP-Complete Problems. to appear in *Mathematical Systems Theory*.
- [GKS94] Martin Charles Golumbic, Haim Kaplan, and Ron Shamir. On the complexity of DNA physical mapping. *Adv. in Appl. Math.*, 15(3):251–261, 1994.
- [GMS80] J. Gallant, D. Maier, and J. Storer. On finding minimal length superstrings. *JCSS*, 20(1):50–58, 1980.
- [GN01] Jens Gramm and Rolf Niedermeier. Minimum quartet inconsistency is fixed parameter tractable. In *Combinatorial pattern matching (Jerusalem, 2001)*, volume 2089 of *Lecture Notes in Comput. Sci.*, pages 241–256. Springer, Berlin, 2001.
- [GN02] Jens Gramm and Rolf Niedermeier. Breakpoint Medians and Breakpoint Phylogenies: a Fixed Parameter Approach. *Bioinformatics*, 18(90002):128S–139, 2002.
- [Gra03] Jens Gramm. *Fixed-Parameter Algorithms for the Consensus Analysis of Genomic Data*. Doktors der naturwissenschaften, der Fakultät für Informations- und Kognitionswissenschaften der Eberhard-Karls-Universität Tübingen zur Erlangung des Grades eines, 2003.
- [GVY97] G. Garg, V. V. Vazirani, and M. Yannakakis. Primal-Dual Approximation Algorithms for Integral Flow and Multicul in Trees. *Algorithmica*, 18(1):3–20, May 1997.
- [GW02] Ganashkumar Ganapathy and Tandy Warnow. Approximating the Complement of the Maximum Compatible Subset of Leaves of  $k$  Trees. In *Proceedings of the Fifth International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 122–134, 2002.
- [Hal] M.T. Hallet. Shortest Common Supersequence is hard for  $w[t]$ , for all  $t$ .
- [Hal96] M.T. Hallet. *An integrated complexity analysis of problems form computational biology*. Phd thesis, Department of Computer Science, University of Victoria, Victoria, B.C., Canada, 1996.
- [HP96] S. Hannenhalli and P.A. Pevzner. To cut... or not to cut: applications of comparative physical maps in molecular evolution. In *Proc. of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 96)*, pages 304–313. Atlanta, Georgia, 1996.
- [HRW92] Frank K. Hwang, Dana S. Richards, and Pawel Winter. *The Steiner tree problem*, volume 53 of *Annals of Discrete Mathematics*. North-Holland Publishing Co., Amsterdam, 1992.
- [Hüf03] Falk Hüffner. Graph Modification Problems and Automated Search Tree Generation. *NWG Theoretische Informatik/Parametrisierte Algorithmen Wilhelm-Schickard-Institut für Informatik Universität Tübingen*, October 2003.

- [KS93] J. Kececioğlu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. In *Combinatorial Pattern Matching, Proc. 4th Annual Symposium (CPM'93)*, volume 684 of *Lecture Notes in Computer Science*, pages 87–105. Springer-Verlag, Berlin, 1993.
- [KS96] Haim Kaplan and Ron Shamir. Pathwidth, bandwidth, and completion problems to proper interval graphs with small cliques. *SIAM J. Comput.*, 25(3):540–561, 1996.
- [KST94] H. Kaplan, R. Shamir, and R. E. Tarjan. Tractability of parameterized completion problems on chordal and interval graphs: Minimum Fill-in and physical mapping. In *35th Annual Symposium on Foundations of Computer Science (FOCS' 94)*, pages 780–791. IEEE Computer Society Press, 1994.
- [KST99] Haim Kaplan, Ron Shamir, and Robert E. Tarjan. Tractability of parameterized completion problems on chordal, strongly chordal, and proper interval graphs. *SIAM J. Comput.*, 28(5):1906–1922 (electronic), 1999.
- [LLM<sup>+</sup>99] J. K. Lanctot, M. Li, B. Ma, S. Wang, and L. Zhang. Distinguishing String Selection Problems. In *Proc. of 10th ACM-SIAM SODA*, pages 633–642, 1999.
- [LLM<sup>+</sup>03] J. K. Lanctot, M. Li, B. Ma, S. Wang, and L. Zhang. Distinguishing String Selection Problems. *Information and Computation*, 185(1):41–55, 2003.
- [LN02] David Liben-Nowell. Gossip is synteny: incomplete gossip and the syntenic distance between genomes. *J. Algorithms*, 43(2):264–283, 2002.
- [Mai] D. Maier. The Complexity of Some Problems on Subsequences and Supersequences. *Journal of the ACM*, 25,2(1978):322–336.
- [Nie04] Rolf Niedermeier. Ubiquitous Parameterization—Invitation to Fixed-Parameter Algorithms. In *Proceedings of the 29th International Symposium on Mathematical Foundations of Computer Science (MFCS'04)*, Lecture Notes in Computer Science, Prague, Czech Republic, pages 84–103. Springer-Verlag Berlin Heidelberg, 2004.
- [NR99] Rolf Neidermeier and Peter Rossmanith. An Efficient Fixed Parameter Algorithm for 3-Hitting Set. Technical Report WSI-99-18, Universität Tübingen, Wilhelm-Schickard-Institut für Informatik, October 1999. Revised version in *Journal of Discrete Algorithms*.
- [NSS01] Assaf Natanzon, Ron Shamir, and Roded Sharan. Complexity classification of some edge modification problems. *Discrete Appl. Math.*, 113(1):109–128, 2001. 25th International Workshop on Graph-Theoretic Concepts in Computer Science (WG'99) (Ascona).
- [Pe'02] Itsik Pe'er. *Algorithmic Methods for Reconstruction of Biological Sequences, Gene Orders and Maps*. Ph. D., Tel-Aviv University, 2002.
- [PH] T.J. Perkins and M. T. Hallett. On the Computational Complexity of Finding Small Sets of Explanatory Variables. *NIPS 2002 Workshop on Machine Learning Techniques for Bioinformatics*.
- [Pie03] K. Pietrzak. On the parameterized complexity of the fixed alphabet shortest common supersequence and longest common subsequence problems. *Journal of Computer and System Sciences*, 67(4):757–771, 2003.

- [PS] Itsik Pe’er and Ron Shamir. Approximation Algorithms for the Median Problem in the Breakpoint Model. *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*. (D. Sankoff and J. H. Nadeau, editors), Kluwer Academic Press (Dordrecht) 2000.
- [RSV04] Bruce Reed, Kaleigh Smith, and Adrian Vetta. Finding odd cycle transversals. *Oper. Res. Lett.*, 32(4):299–301, 2004.
- [SFR<sup>+</sup>99] N. Stojanovic, L. Florea, C. Riemer, D. Gumucio, J. Slightom, M. Goodman and W. Miller, and R. Hardison. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research*, 27(19):3899–3910, 1999.
- [SS00] Roded Sharan and Ron Shamir. CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis. In *Proceedings: ISMB*, pages 307–316, 2000.
- [Ste92] M. A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.
- [Ste99] Ulrike Stege. Gene trees and species trees: the gene-duplication problem is fixed-parameter tractable. In *Algorithms and data structures (Vancouver, BC, 1999)*, volume 1663 of *Lecture Notes in Comput. Sci.*, pages 288–293. Springer, Berlin, 1999.
- [SV97] Benno Schiwickoski and Martin Vingron. The Deferred Path Heuristic for the Generalized Tree Alignment Problem. In *Proceedings of the First Annual International Conference on Computational Molecular Biology*, 1997.
- [VJLW02] Gianluca Della Vedova, Tao Jiang, Jing Liz, and Jianjun Wen. Approximating Minimum Quartet Inconsistency. *13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA2002)*, pages 894–895, 2002.
- [VLM] Martin Vingron, Hans Peter Lenhof, and Petra Mutzel. Computational Molecular Biology. Chapter written for Annotated Bibliography in Combinatorial Optimization.
- [Yan81] M. Yannakakis. Computing the Minimum Fill-in is NP-complete. *SIAM J. Alg. Disc. Meth.*, 2, 1981.