# A List of Parameterized Problems in Bioinformatics

Liliana Félix Ávila  Alina García  Maria Serna  Dimitrios M. Thilikos

**Abstract**

In this report we present a list of problems that originated in Bionformatics. Our aim is to collect information on such problems that have been analyzed from the point of view of Parameterized Complexity. For every problem we give its definition and biological motivation together with known complexity results.

## 1 Introduction

The theory of NP-completeness seems to be one of the greatest achievements in Theoretical Computer Science during the last 35 years. In particular, it offered a solid background for characterising and investigating the "hardness" of combinatorial problems [GJ79]. However, for practical purposes, such a theory seems to introduce a rather pessimistic viewpoint as the majority of natural non-trivial combinatorial problems seems to be NP-hard and thus one cannot expect that they admit an efficient (i.e. polynomial time) algorithmic solution. However, a most optimistic point of view can be adopted if we take in mind that the NP-completeness concerns only the *worst case complexity* of a combinatorial problem. In many real applications, the inputs of a generally tractable problem may be structured or restricted in a way that makes them tractable in practice. This motivated the idea of splitting the input of a combinatorial problem into two parts: the *main part* and the *parameterized* part. The splitting should be done in a way that the size of the parameterized part is "small" in the majority of the "real world" applications while the main part is the one that includes elements of the problem that can be really big. The hope in this splitting is that we may be able to design algorithms with complexities whose super-polynomial part is *exclusively* depending on the "small" parameterized part. In other words, when we fix the parameterized part to be of constant size then the problem becomes tractable. If this is possible, then we may consider such a problem "tractable in practice" as it is easy to solve it in most of the cases where we may require a solution. This idea motivated what is now called *Parameterized Complexity*, a theory that during the last 16 years offered a solid and attractive alternative for investigating the hardness of combinatorial problems for from both algorithmic and complexity point of view.

Formally, a parameterized problem has as instances pairs $(I, K)$ where $I$ is the main part and $K$ is the parameterized part. We use the notation $n = |I|$ and $k = |K|$ for the sizes of $I$ and $K$ respectively. The Parameterized Complexity settles the question of whether the problem is solvable by an algorithm (we call it FPT-*algorithm*) of time complexity

$$f(k) \cdot n^{O(1)}$$

where $f(k)$ is a (super-polynomial) function that does not depend on $n$. If such an algorithm exists, we say that the parameterized problem belongs in the class FPT. In a series of fundamental papers (see [DF95a, ADF95, DF95b, DF93, DF92]), Downey and Fellows defined a series of

complexity classes, namely the classes

$$\mathsf{W}[1] \subseteq \mathsf{W}[2] \subseteq \cdots \subseteq \mathsf{W}[SAT] \subseteq \mathsf{W}[P]$$

and proposed special types of reductions such that hardness for some of the above classes makes it rather impossible that a problem belongs in FPT (we stress that $\mathsf{FPT} \subseteq \mathsf{W}[1]$). The above theoretical framework initiated the classification of several parameterized problems according to their parameterized complexity. As it is expected in such a project, special attention has been given to parameterizations of problems that emerge from practical applications where *fast* (or "as fast as possible") solutions where really wanted. Parameterized complexity offered insightful results in a great variety of research areas like VLSI-design [FL92, FL88a, FL88b], Robot Motion Planning [CW95], Data Bases [GSS02, DFT97, PY99], Logical Programming, [LP97] and others (see also [Fel01, DFS99, DF99b, DF99a]). So far, the most complete list of parameterized problems along with their classification according to their parameterized complexity is the compendium of Marco Cesati [Ces01] including more than 200 problems reflecting the huge amount of work that has been devoted on this theory during the last two decades.

Perhaps, the topic where Parameterized Complexity has been more extensively applied is bioinformatics. One of the first important steps in this direction was done 10 years ago in [BFH94] where it was noticed that several computational problems in bioinformatics involve parameters that in most of applications do not obtain big values. This provided a hope that their general NP-hardness may not be an obstacle for an efficient solution when these parameters are small. Currently, parameterized complexity is able to present results in a big variety of topics in biology such as genome sequence alignment and reconstruction, biopolymer folding, and problems in phylogeny and evolution. While a lot of problems where classified to be hard for some of the classes of the parameterized complexity hierarchy, there were also considerably many problems that have been classified in FPT. It is interesting to notice that there were cases where known techniques in bioinformatics were proved to be nothing more than FPT-algorithms [DFS99]. Moreover, it seems that there are several algorithmic results on bioinformatics that may fit into the framework of Parameterized Complexity while they where never presented as such. This makes us believe that the encounter between bioinformatics and parameterized complexity has more future than history. With this problem list we attempt a first classification of the existing results. Our aim is double: first to make the parameterized complexity community more familiar to the results and the challenges arising from bioinformatics and second, to invite researchers from bioinformatics to adopt the tool of parameterization as a way to cope with the structural hardness of the problems they deal with.

One of the most difficult tasks in the compilation of our problem list was to bridge the difference in the way problems are presented in each of the two communities. In Parameterized Complexity problems and solutions should be presented using str ict mathematical formalism. However, in bionformatics, problems and solutions escape from the "ideal" world of theoreticians and obtain more attributes of the real world. There were papers where it was hard to "transcript" the given problem into a formal definition while it is apparent that what is presented is an FPT-algorithm.

We present our list in five main sections. The first four are related to a bioinformatics concept, however also the data and the computational problems hold similarity. The fifth one is devoted to classical graph problems that also have motivations in Bioinformatics. We finish with a small list of open problems.

The list of problems is accompanied with a glossary on terms on bioinformatics and two indices of the listed problems, one alphabetic and one hierarchical (with respect to the levels of the parameterized complexity hierarchy).

# 2 Sequence Alignment

Biological relationships are obtained by different tools and models from sequence analysis. *Sequence alignment* is the assignment of residue-residue correspondence between biological sequences. *Similarity* is the observation or measurement of resemblance an difference, independently of the source of resemblance. This is done through different quantitative measures of string similarity.

We will use the following terminology: A string $x$ is a *substring* of a string $s$ if $x$ appears in contiguous positions in $s$. Conversely $s$ is a *supersting* of $x$. A string $x$ is a *supersequence* of a string $s$ if we can delete some characters in $x$ in such a way that the remaining string is equal to $s$. Conversely $s$ is a *subsequence* of $x$.

## 2.1 Bounded Duplication Shortest Common Supersequence for Complete p-Sequences

Problem Definition [FHKS98b]:

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 $x$.

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

A string $x$ contains $r$ *duplication events* if $x$ is not a $p$-sequence, but removing exactly $r$ symbols from $x$ result in a $p$-sequence

The *duplication cost* of a sequence $x$ is defined as $\|x\|_c = \sum_{a\in\Sigma}(n_a(x)-1)c(a)$, where, for $a \in \Sigma$, $n_a(x)$ denotes the number of occurrences of symbol $a$ in $x$.

> *Instance:* Given a set of complete $p$-sequences $x_i$ over an alphabet $\Sigma$ of size $n$, a positive integer $r$, and a cost function $c : \Sigma \to Z^+$
>
> *Parameter:* $r$
>
> *Question:* Is there a common supersequence $x$ of duplication cost $\|x\|_c \leq r$?

Biological Motivation [FHKS98b]:

The problem is the sequence analog of the same problem on trees, see problem 3.1

Complexity: NP-complete [FHKS98b].

Parameterized Complexity: FPT [FHKS98b].

## 2.2 Bounded Duplication Shortest Common Supersequence for p-Sequences

Problem Definition [FHKS98b]:

For the definition of $p$-sequence and duplication event see Problem 2.1.

> *Instance:* Given an alphabet $\Sigma$ with $k$-symbols, a family $x_1, \ldots, x_k$ of $p$-sequences over $\Sigma$, such that each symbol of $\Sigma$ occurs in at least one of the input sequences, and a positive integer $r$.

*Parameter:* $k$ and $r$

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

Biological Motivation: See problem 2.1 for biological motivation.

Complexity: NP-complete [FHKS98b].

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

## 2.3 Closest String

Problem Definition :

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

*Instance:* Given a collection of $k$ strings $s_1, s_2, \ldots, s_k$, all of them with length $L$, over an alphabet $\Sigma$ and two 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 $i = 1, \ldots, k$?

Biological Motivation [Gra03]:

*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 polymerase chain reaction (PCR). The hybridization of primers depends on complex thermodynamic rules, but it 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.

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.

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 form 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+03].

Complexity: NP-complete even for binary alphabet [FL97, LLM+03].

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

## 2.4 Closest Substring

Problem Definition [FGN02]:

By $d_H(s, s'_i)$ we denote the *Hamming distance* between strings $s$ and $s'_i$, for a definition see problem 2.3.

> *Instance:* Given a collection of $k$ strings $s_1, s_2, \ldots, s_k$ over alphabet $\Sigma$, and two 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 $i = 1, \ldots, k$, there is a length-$L$ substring $s'_i$ of $s_i$ with $d_H(s, s'_i) \leq d$? .

Biological Motivation [GHN02]:

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.

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].

## 2.5 Consensus Pattern

Problem Definition:

By $d_H$ we denote the *Hamming distance* for definition see problem 2.3.

> *Instance:* Strings $s_1, s_2, \ldots, s_k$ 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 $i = 1, \ldots, k$, a length-$L$ substring $s'_i$ of $s_i$ such that $\sum_{i=1}^{k} d_H(s, s'_i) \leq d$? (.

Biological Motivation [Gra03]:

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 $d$.

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].

## 2.6 Distinguishing String Selection (DSS)

Problem Definition [GGN03]:

By $d_H(s, s'_i)$ we denote the *Hamming distance* between strings $s$ and $s'_i$, for a definition see problem 2.3.

*Instance:* Given a collection of $k_1$ *good* strings $s_1, \ldots, s_{k_1}$, a collection of $k_2$ *bad* strings $s'_1, \ldots, s'_{k_2}$, all of them with $L$ characters, and two positive integers $d_1$ and $d_2$.

*Parameter:* $d_1$ and $d_2$

*Question:*

Is there a string $s$ of length $L$ for which

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

and

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

Biological Motivation [LLM+03]: 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.

Deng *et al.* [DLL+02] let all good strings be of same length $L$. The terminology "good" and "bad" has its motivation in the application [LLM+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].

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.

Finally, DSS problems have also applications in creating diagnostic probes for bacterial infection and creating universal PCR primers.

Complexity: NP-hard [FL97, LLM+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].

## 2.7 Fixed Alphabet LongestCommon Subsequence

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 $r_1,\ldots,r_k$ over the alphabet $\Sigma$, and a positive integer $m$.
>
> *Parameters:*
>
> 1. $k$
> 2. $k$ and $m$
>
> *Question:* Is there a string $s\in\Sigma^*$, with length at least $m$, that is a subsequence of each $r_i$, for $i=1,\ldots,k$?

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 [DF99a].

Complexity: NP-complete [Mai].

Parameterized Complexity:

W[1]-hard, when parameterized by $k$, by a 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 Problem 2.10, Longest Common Subsequence (LCS).

## 2.8 Fixed Alphabet Shortest Common Supersequence

Problem Definition [Ces04]:

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 $r$ .

> *Instance:* An alphabet $\Sigma$ having fixed size, a set of strings $\{r_1, \ldots, r_k\}$ over the alphabet $\Sigma$, and 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 $r_i$, $1 \leq i \leq k$?

Biological Motivation [BDF$^+$]:

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$^+$].

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

Parameterized Complexity:

W[1]-hard, when parameterized by $k$, by a reduction from Partitioned Clique [Pie03].

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

Comments: See also Problem 2.13, Shortest Common Supersequence (SCS).

## 2.9 Longest Arc Preserving Common Subsequence (LAPCS)

Problem Definition [Eva99, Gra03]:

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 $\sum$.

Given a sequence $S$, let $P$ be the set of positions in the sequence. If $S$ has length $n$ then $P = \{1, \ldots, n\}$. An *arc* is a directed edge $(p_1, p_2) \in P \times P$. 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 $p_1 < p_2$.

A sequence $y$ is a *common subsequence* of the sequences $S_1$ and $S_2$ if $y$ is a subsequence of both $S_1$ and $S_2$.

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

Given an arc annotation $A$ of a sequence $S$, we can consider several measures

- $s$, the maximum number of levels of nested arcs.
- $k$, the crossing or cutwidth of the arc structure, that is the maximum number of arcs across a cut between consecutive positions.
- $d$, the bandwidth of arc structure, that is the maximum $(j - i)$ for any arc $(i, j)$ in $A$.

All those measures can used as parameters restricting the input. We use the abbreviations CROSS and NEST for arc annotations with bounded crossings or nested arcs.

> *Instance:* Given a target length $l$, and a pair of arc annotated sequences $(S_1, A_1)$ and $(S_2, A_2)$.
>
> *Parameters:*
>
> 1. $l$, length of desired subsequence.
> 2. $s$, levels of nested arcs.
> 3. $k$, cutwidth of arc structure.
> 4. $d$, bandwidth of arc structure.
>
> *Question:* Finding a common subsequence of length $l$ which preserves induced arcs.

Observe that the length of desired subsequence $l$ is independent of the other parameters. However the others are related; $s = k$, and $k \leq d$ for all restriction levels except when unlimited.

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, various versions of the LAPCS problem 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.

Biological Motivation [Eva99]:

Molecular biologists use algorithms that compare 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 descriptive text that accompanies a sequence in a database record is called an *annotation*. 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. An individual *annotation* for a specific sequence is its associated additional information, as represented according to the chosen annotation scheme.

The basic 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.

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

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].

## 2.10 Longest Common Subsequence (LCS)

Problem Definition:

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

*Parameters:*

1. $k$ (LCS-1).
2. $m$ (LCS-2).
3. $k$, $m$ (LCS-3).
4. $k$, $|\Sigma|$ (LCS-4).

*Question:* Is there a string $X \in \Sigma^*$ of length at least $m$ such that it is a subsequence of $X_i$, for $i = 1, \ldots, k$?

Biological Motivation: See problem 2.7 for biological motivation.

Complexity: NP-complete [Mai].

Parameterized Complexity:

W[$t$]−hard for all $t$ for LCS-1 [BDFW95], by a reduction from Monotone Weighted t-Normalized Satisfiability [BDFW95, BDFW94, DF99a]).

W[2]−hard for LCS-2 but in W[$P$] [BDFW95], membership is easy; hardness: by reduction from Dominating Set [BDFW95, DF99a]; in FPT if $|\Sigma|$ is a 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, DF99a]; hardness: by reduction from Clique [BDFW95, BDFW94, DF99a].

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

Comments: See also 2.7 problem, Fixed Alphabet Longest Common Subsequence.

### 2.11 $k$ -Mismatch

Problem Definition [Gra03]:

By $d_H(s, s'_i)$ we denote the *Hamming distance* between strings $s$ and $s'_i$, for a definition see problem 2.3.

Let $s(p, L)$ denote the length-$L$ substring of a given string $s$ starting at position $p$.

> *Instance:* Given $m$ strings $s_1, s_2, \ldots, s_m$ of length $n$, and two integers $L$ and $k$.
>
> *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 $i = 1, \ldots, m$?.

Biological Motivation:

See problem 2.3 for biological motivation.

Complexity: NP-hard [FL97, LLM$^+$99].

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

### 2.12 Modified Distinguishing Substring Selection (MDSSS)

Problem Definition:

By $d_H$ we denote the *Hamming distance* for definition see problem 2.3.

> *Instance:* Given two sets of strings of length at least $L$ over $\Sigma = \{0, 1\}$, $S_g = \{s_1, \ldots, s_{k_g}\}$ and $S_b = \{s'_1, \ldots ..., s'_{k_b}\}$, 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 $t_i$ of $s_1$, $d_H(s, t_i) \geq d_g$; and that, every $s'_i \in S_b$ has *at least one* length-$L$ substrings $t'_i$ with $d_H(s, t'_i) \leq d_b$?

Biological Motivation:

See problem 2.6 for biological motivation. Recall that $S_g$ represent a set of god strings while $S_b$ represents a set of bad strings.

Complexity: NP-hard [FL97, LLM$^+$99].

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].

Comments: See problem 2.6 (see also [GGN03]).

## 2.13 Shortest Common Supersequence (SCS)

Problem Definition:

*Instance*: An alphabet $\Sigma$, a set of $k$ strings $\{r_1, \ldots, r_k\}$ formed over alphabet $\Sigma$, and 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 of each string $r_i$, $1 \leq i \leq k$?

Biological Motivation:

See problem 2.8 for biological motivation.

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 Problem 2.8, Fixed Alphabet Shortest Common Supersequence.

## 2.14 Shortest Common Supersequence for p-Sequences

Problem Definition:

For the definition of $p$-sequence see problem 2.1.

*Instance:* A collection of $k$ $p$-sequences, $x_1, \ldots, x_k$, and a positive integer $M$.

*Parameter:* $k$

*Question:* Is there a sequence $x$, with $|x| \leq M$ such that $x_i$ is a subsequence of $x$, for $i = 1, \ldots, k$?

Biological Motivation:

See problem 2.8 for biological motivation.

Complexity: NP-complete [FHKS98b].

Parameterized Complexity: W[1]-hard [FHKS98a], by a reduction from the Clique problem [FHKS98b].

# 3 Evolution and Phylogeny

Biological nomenclature is based on the idea that living things are divided into units called *species*, groups of similar organisms with a common gene pool. With the discovery of evolution it emerged that the system largely reflects biological ancestry, so that the question of which similarities truly reflect common ancestry must be faced. Sequence analysis gives the most unambiguous evidence for the relationships among species.

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.

*Homology* is the observation or measurement of resemblance an difference, due to evolution.

*Phylogeny* or *phylogenetics* is the classification of species and organisms according to their evolutionary relationships. In *molecular* phylogenetics, this classifications 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

An *$X$-tree* is a tree (rooted or not) in which the leaves are labeled from $X$.

A *Phylogenetic tree* or *p-tree* is a rooted tree where the leaves are labeled from a set of labels $X$ and where no symbol in $X$ is used more than once as a label

We say that a $p$-tree $Z$ is a *supertree* of a $p$-tree $T$ if $X$ is contained in $Y$ by a topological containment that respects ancestry with label isomorphisms at the leaves.

## 3.1 Bounded Duplication Smallest Common Supertree for Binary p-Trees

Problem Definition [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.

> *Instance:* A family of $k$ complete binary $p$-trees, $T_1, \ldots, T_k$ 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 binary $\Sigma$-tree $T$, such that $T$ is a supertree of all the $T_i$, and contains at most $r$ duplication events?

Biological Motivation [FHKS98b]:

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.

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.

Complexity: NP-complete [FHKS98b].

Parameterized Complexity: FPT, [FHKS98b].

## 3.2 Compatibility of Unrooted Phylogenetic Trees

Problem Definition [SS03]:

An *unrooted phylogenetic tree* $T$ is an $X$-tree having no vertices of degree two.

A tree $T$ displays another tree $T'$ if

> *Instance:* Given a collection of unrooted phylogenetic trees $T_1, T_2, \ldots, T_k$.
>
> *Parameter:* $k$.
>
> *Question:* Does there exist an unrooted phylogenetic tree $T$ that simultaneously displays each tree $T_i$, for $1 \leq i \leq k$?

Biological Motivation [BL04]:

This problem was first discussed by Gordon in [Gor86], who introduced the notion of subtrees as *samples* of the true evolutionary tree, i.e., given a collection of phylogenetic trees for different sets of species, can we find a 'super' tree such that all the input trees are restrictions, or samples, of the larger tree.

Complexity: NP-hard [Ste92].

Parameterized Complexity: FPT, $O(nf(k))$ time [BL04].

## 3.3 Gene Duplication

Problem Definition:

*Gene trees* and *species trees* are rooted, binary, and leaf labeled. $T = (V, E, L)$ denotes a gene or species tree where $V$ is the vertex set, $E$ is the edge set, and $L \subseteq V$ is the leaf-labeling. For a vertex $u \in V$, $T_u$ denotes the subtree of $T$ rooted at $u$. The *leafset* $L$ of $T$ is denoted by $L(T)$, and the leafset of a subtree $T_u$ is denoted by $L(u)$ instead of $L(T_u)$.

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 $T_r$.

For trees $T = (V_1, E_1, L_1)$ and $S = (V_2, E_2, L_2)$, given a vertex $u \in V_1$ let $lca_{T_2}(L(v))$ be the *least common ancestor* of all the labels 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. A *mapping* $M$ of gene tree $G$ into a species tree $S$ specifies two functions $loc_{G,S}$ and $event_{G,S}$:

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

- The function $event_{G,S} : V_G \mapsto \{dup, spec\}$ associates to each vertex in $G$ an event, indicating whether the event in $G$ corresponds to a duplication or a speciation event.

Those function as defined as follows: for each $u \in V_G - L$,

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

The quantity $cost(G, S) = |\{u | u \in V_G - L,\ event_{G,S}(u) = dup\}|$ is the minimum number of gene-duplication events necessary to rectify the gene tree $G$ with the species tree $S$,

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

Let $|L| = n$.

> *Instance:* A collection of $k$ gene trees $G_1, \ldots, G_k$ over the leaf set $L$, and a positive integer $c$.
>
> *Parameter:* $c$
>
> *Question:* Does there exist a species tree $S$ with $cost(G_1, \ldots, G_k, S) \leq c$?

Biological Motivation [Ste99]:

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.

*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.

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.

The Gene Duplication is the problem of computing the optimal species tree for a given set of gene trees under the *Gene-Duplication Model*. That is for a set of taxa given by a set of possibly contradictory gene trees. Several models for attacking the problem have appeared in the literature including Problem 3.4, the Maximum Agreement Subtree (MAST).

Complexity: NP-complete [Ste99].

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

## 3.4 $k$ -Maximum Agreement Subtree (MAST)

Problem Definition:

> *Instance:* A set of rooted trees $T_1, \ldots, T_r$ ($r \geq 3$) 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 $i = 1, \ldots, r$?

Biological Motivation [CCH+]:

The MAST problem arises naturally in biology and linguistics as a measure of consistency between two evolutionary trees over species or 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.

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

Complexity: NP-complete [DFS99].

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

## 3.5 Minimum Quartet Inconsistency (MQI)

Problem Definition [GN01]:

Let $S$ be a set of labels an $T$ and $p$-tree on $S$. A *quartet* is a size four subset $\{a, b, c, d\}$ of $S$, and the topology for $\{a, b, c, d\}$ induced by $T$ 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 of $\binom{n}{4}$ quartet topologies, such that there is exactly one topology for every quartet set in $Q_S$, 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?

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].

Complexity: NP-complete [GN01].

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

### 3.6 Perfect Path Phylogeny Haplotyping with Missing Data ($\{0,1,2,?\}$-PPPH)

Problem Definition:

A *genotype* is a string over an alphabet $\{0,1,2\}$. A *genotype matrix* is a matrix whose rows are genotypes. We say that a genotype matrix $A$ *admits a perfect phylogeny* or is *pp-realizable* if there exists a labeled rooted tree $T$, called *perfect phylogeny* such that:

1. Every edge of $T$ is labeled by *at least* one column of $A$.
2. Each column of $A$ labels *exactly* one edge of $T$.
3. For each row $r$ of $A$ there are two nodes in $T$ labeled $r'$ and $r''$
4. For every row $r$ of $A$ the set of columns with value 2 in this row forms a path $p$ in $T$ connecting $r'$ and $r''$ and the set of columns with value 1 in row $r$ forms a path from the root of $T$ to the top-most node on the path $p$.

A *genotype matrix with ?-entries* is a genotype matrix in which we substitute some of its entries with question marks (the *?-entries*). A genotype matrix $A$ with ?-entries is *pp-realizable* if we can replace all its question marks with values from $\{0,1,2\}$ so that the resulting genotype matrix admits a perfect phylogeny. If additionally this perfect phylogeny tree is a path, then we say that $A$ is *perfect path phylogeny realizable*, *ppp-realizable* for short.

*Instance:* A $n \times m$ genotype matrix $A$ with $k$ ?-entries.

*Parameter:* $k$

*Question:* Is $A$ *ppp-realizable*?

Biological Motivation [GNST04, Gus02]:

Haplotyping via *perfect phylogeny* is a method for haplotype inference where it is assumed that the (unknown) haplotypes underlying the (observed) genotype data can be arranged in a genetic tree in which each haplotype results from an ancestor haplotype via mutations. The perfect phylogeny approach is popular due to its applicability to real haplotype inference problems and its theoretical elegance. It was introduced by Gusfield [Gus02] and received considerable attention which resulted, among others, in quadratic-time algorithms for the case of complete and error-free-input data [VDGS03, EHK02]. In the special case where perfect *path* phylogenies are sought, even a linear time algorithm is known [GNST04].

A detailed biological justification for considering perfect phylogenies and for requiring the above properties for the tree $T$ can be found in [GNST04, Gus02].

Complexity: NP-complete [GNST04].

Parameterized Complexity: FPT, in $O(4^k m^3 n + 3^{O(k^3 6^k k!)} n^2)$ time [GNT04].

## 3.7 Perfect Phylogeny

Problem Definition:

*Instance:* A set $C = \{1, \ldots, m\}$ of characters; for each $c \in C$, a set $A_c = \{1, \ldots, r_c\}$ of states; and a set $S \subseteq A_1 \times \ldots \times A_m$ 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. $S \subseteq V(T) \subseteq A_1 \times \ldots \times A_m$.
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 $T$?

Biological Motivation [AFB96, VLM]:

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*.

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 and it is also related with problem 5.9, Triangulating $k$-Colored Graphs.

## 3.8 Perfect Phylogeny plus $k$ columns

Problem Definition:

For the definitions of *genotype matrix* and *perfect phylogeny (PP)* see Problem 3.6

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

*Parameter:* $k$

*Question:* Is it possible to delete at most $k$ columns from $M$ in such a way that the resulting matrix has a perfect phylogeny?

Biological Motivation [Dam04]:

Perfect phylogeny 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 .

Complexity: NP-complete [BFW92].

Parameterized Complexity: FPT, $O(k^2nm + k^2 2^k)$ time [Dam04].

## 3.9 Perfect Phylogeny plus $k$ rows

Problem Definition:

For the definitions of *genotype matrix* and *perfect phylogeny (PP)* see Problem 3.6

> *Instance:* An $n \times m$ genotype matrix $M$ and an integer $k$.
>
> *Parameter:* $k$
>
> *Question:* Is it possible to delete at most $k$ rows from $M$ in such a way that the resulting matrix has a perfect phylogeny?

Biological Motivation:

See Problem 3.8 for biological motivation.

Complexity: NP-complete [BFW92].

Parameterized Complexity: FPT, $O(3^k nm)$ time [Dam04].

## 3.10 Perfect Phylogeny with $k$ errors

Problem Definition:

For the definitions of *genotype matrix* and *perfect phylogeny (PP)* see Problem 3.6

> *Instance:* An $n \times m$ binary genotype matrix $M$ and an integer $k$.
>
> *Parameter:* $k$
>
> *Question:* Is it possible to change a set of at most $k$ bits in $M$, so that it has a perfect phylogeny.

Biological Motivation:

See problem 3.8 for biological motivation.

Complexity: NP-complete [BFW92].

Parameterized Complexity: FPT, $O(k6^k nm)$ time [Dam04].

### 3.11 Smallest Common Supertree for p-trees

Problem Definition [FHKS98b]:

An *rl-tree:* It is a rooted tree with leaves labeled from a set $X$, where labels may be repeated.

> *Instance:* A collection of $k$ binary $p$-trees $T_1, \ldots, T_k$ and a positive integer $m$.
>
> *Parameter:* $k$
>
> *Question:* Is there an $rl$-tree $T$, with $|T| \leq m$ such that any $T_i$ is contained in $T$ by a topological containment that respects ancestry with label isomorphism at the leaves, for any $i = 1, \ldots, k$?

Biological Motivation [FHKS98b]:

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.

Complexity: NP-complete [FHKS98b].

Parameterized Complexity: W[1]-hard, by a reduction from the Clique problem [FHKS98b].

## 4 Genome rearrangement

The genome rearrangement field provides some of the currently most popular measures in phylogenetic studies for related species.

### 4.1 Breakpoint Median

Problem Definition [GN02]:

Given a set $S = \{1, \ldots, n\}$, an *ordering* $\pi$ on $S$ is a permutation on $S$. Every ordering is extended by two special elements namely $s$, marking the start, and $t$, marking the end, and the ordering $\pi$ is written as $\langle s\ \pi(1)\ \pi(2)\ \ldots\ \pi(n)\ t\rangle$. Then $S_s$ is $S \cup \{s\}$ ($S_t$ and $S_{s,t}$, analogously) .

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 $\{-1, 1, -2, 2, \ldots, -n, n\}$ and $S_s^{\pm}$ for $S^{\pm} \cup \{s\}$ ($S_t^{\pm}$ and $S_{s,t}^{\pm}$ analogously).

We use $succ_\pi(x)$, for signed ordering $\pi$ and $x \in S_s$, to denote the *succesor* $y \in S_{s,t}^{\pm}$ 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 $x$.

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. $succ_{\pi_1}(x) = y$ and $succ_{\pi_2}(x) \neq y$

Finally, define the *breakpoint distance* $d_{bp}$ between two signed orderings as follows: $d_{bp}(\pi_1, \pi_2) = |\{(x, y)|x, y \in S^{\pm}_{s,t}, x, y \text{ is breakpoint of } \pi_1 \text{ w.r.t. } \pi_2\}|$.

Observe that, due to symmetry, $d_{bp}(\pi_1, \pi_2) = d_{bp}(\pi_2, \pi_1)$.

> *Instance:* Given $m$ signed orderings $\pi_1, \pi_2, ..., \pi_m$ on the set $S = \{1, \ldots, n\}$, and a positive integer $k$.
>
> *Parameter:* $k$
>
> *Question:* Is there a signed ordering $\pi$ such that $\sum_{i=1}^{m} d_{bp}(\pi_i, \pi) \leq k$?

Biological Motivation [GN02]:

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.

Complexity: NP-complete [PS].

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

## 4.2 Sorting by Reversals

Problem Definition [HP96]:

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

Let $i \sim j$ if $|i - j| = 1$. Extend a permutation $\pi = \pi_1 \ldots \pi_n$ 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 $\pi_i \nsim \pi_{i+1}$. Define a *block* of $\pi$ as an interval $\pi_i \ldots \pi_j$ containing no breakpoints. Define a *strip* of $\pi$ as a maximal block. A strip of one element is called a *singleton*.

> *Instance:* Given a permutation $\pi$ of $\{1, 2, \ldots, n\}$ with $k$ singletons.
>
> *Parameter:* $k$
>
> *Question:* Does there exist at most $k$ reversals needed to transform $\pi$ into the *identity* permutation?

Biological Motivation [HP96]:

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.

$O(\log n)$ 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.

Complexity: NP-hard [KS93].

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

### 4.3 Syntenic Distance

Problem Definition [LN02]:

In this model, a gene is represented by a subset of a set of $n$ characters and a genome is given by $k$ such subsets.

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

- a *fusion* of $S$ and $T$ is the replacement of $(s, T)$ by $U$, where $U = S \cup T$.
- in a *fission* a set $U$ is replace by to non-empty sets $(S, T)$ where $U = S \cup T$.
- a *translocation* of a pair of sets $(S, T)$ replaces the pair for another pair $(S', T')$, 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.

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

*Parameter:* $k$

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

Biological Motivation [AGN01] [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].

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.

Two genes are *syntenic* if they appear in the same chromosome. The *syntenic distance* between two genomes is the minimum number of fusions, fissions, and translocations required to transform the common genes in one chromosome to the genes in the other.

Complexity: NP-complete [LN02, DJK+97].

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

# 5 Assembling DNA fragments

The problem of DNA assembly became very important for sequencing very large genomes such as the human genome. The methods consists sequencing relatively small fragments and then seek for a suitable method to assemble those fragments. The problem of assembly becomes complex because that include, *orientation*, *repeats*, *overlaps*, and sequencing errors.

It is possible to construct information based in overlaps and model those overlaps by means of graphs and completion problems.

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

A *k-coloring* of a graph is an assignment of vertices to a set of $k$-colors such that the endpoints of an edge always get different colors.

## 5.1 Colored Proper Interval Graph Completion

Problem Definition:

A graph $G = (V, E)$ is an *interval graph* if ...

> *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?

Biological Motivation [GGK+95]:

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.

Complexity: NP-complete [AS99, GGK+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 Problem 5.7, Restricted Completion to a Proper Interval Graph with Bounded Clique Size, 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].

## 5.2 Completion to a Proper Interval Graph with Bounded Clique Size

Problem Definition:

A *clique* is a complete bipartite graph. The *clique size* of a graph is the size of the largest clique contained in it.

> *Instance:* Given a graph $G = (V, E)$ and a constant $k$.
>
> *Parameter:* $k$
>
> *Question:* Does there exist a supergraph (for definition see problem 5.1) $G'$ of $G$ which is a proper interval graph and has clique size at most $k$?

Biological Motivation [KST94]:

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.

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

Complexity: NP-hard [KST94].

Parameterized Complexity: W[$t$]-hard for any $t > 0$ [KST94], by reduction from Uniform Emulation on a Path problem.

Comments [KST94]:

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.

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

## 5.3 $k$-Interval Positional Sequencing by Hybridization (Interval PSBH)

Problem Definition:

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

> *Instance:* A multi-set $S$ of strings with length $p$ and, for each $s \in S$, a set $P(s)$ which is a sub-interval of $\{0, |S|-1\}$.
>
> *Parameter:* $p$
>
> *Question:* Is $S$ the $p$-spectrum of some string $X$, such that for each $s \in S$ its positions along $X$ is in $P(s)$?

Biological Motivation:

In Sequencing by Hybridization (SBH), one has to reconstruct a sequence from its $s$-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 $l$-mears in the target sequence. More precisely, for each $l$-mer in the spectrum its allowed positions along the target are registered [Pe'02, BDPSS01].

Complexity: NP-complete, even if all sets of allowed positions are intervals of equal length, by a reduction from Interval Positional Eulerian Path (PEP) problem [Pe'02, BDPSS01].

Parameterized Complexity: FPT, $O(mk^{1.5}4^k)$ 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].

## 5.4 Intervalizing Colored Graphs or DNA Physical Mapping

Problem Definition:

> *Instance:* A graph $G = (V, E)$ and a coloring $c : V \to \{1, \ldots, k\}$; and a positive integer $k$.
>
> *Parameter:* $k$
>
> *Question:* Is there a supergraph (for definition see problem 5.1) $G' = (V, E')$ 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?

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].

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$^+$00].

Parameterized Complexity: W[t]-hard for all $t \in \mathbb{N}$, by reduction from Colored Cutwidth (CC-1) [BFH94, BFH$^+$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].

## 5.5 Minimum Fill-in

Problem Definition:

A *chordal* graph is ...

*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?

Biological Motivation:

The Minimum Fill-in problem is very important in the area of computational biology called perfect phylogeny [DF99a].

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 [BKKM].

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].

Complexity: NP-complete [Yan81].

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

Comments:

This problem is also known as Chordal Graph Completion problem [KST99].

## 5.6 Proper Interval Graph Completion (PIGC)

Problem Definition:

For the definition of proper interval graph see Problem 5.1.

> *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?

Biological Motivation [KST94]:

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 .

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).

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].

## 5.7 Restricted Completion to a Proper Interval Graph with Bounded Clique Size

Problem Definition:

For the definition of proper interval graph see Problem 5.1.

> *Instance:* A graph $G = (V, E)$, a set $E' \subseteq V \times V$ of forbidden edges, and 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 $E'$?

Biological Motivation:

See Problem 5.2 for biological motivation.

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].

## 5.8 Strongly Chordal Completion

Problem Definition:

A graph is *strongly chordal* if it is chordal (every cycle of length 4 or more has a chord) and if every even cycle of length 6 or more contains a chord splitting the cycle into two odd length paths.

> *Instance:* Given a graph $G = (V, E)$ and a positive integer $k$.
>
> *Parameter:* $k$
>
> *Question:* Does there exist an edge set $A$ such that $|A| \leqslant k$ and $G = (V \cup A)$ is strongly chordal graph?

Biological Motivation:

See problem 5.5 for biological motivation.

Complexity: NP-hard [KST99].

Parameterized Complexity: FPT, $O(8^{2k}mlogn)$ time [KST99].

## 5.9 Triangulating $k$-Colored Graphs

Problem Definition:

> *Instance:* A graph $G = (V, E)$, a vertex coloring $c : V \to \{1, \ldots, k\}$, and a positive integer $k$.
>
> *Parameter:* $k$
>
> *Question:* Does there exist a supergraph (for definition see problem 5.1) $G' = (V', E')$ where $E \subseteq E'$, $G'$ is properly colored by $c$, and $G'$ is triangulating?

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].

Complexity: NP-complete [BFH$^+$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 3.7 problem, Perfect Phylogeny.

# 6 Graph problems

In this section we present some graph problems that also have applications in bioinformatics.

## 6.1 $k$ -Cluster Editing

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?

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].

Complexity: NP-complete [Hüf03].

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

## 6.2 Vertex Bipartization

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?

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 [Gra03].

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].

Complexity: NP-complete [GJ79], (Problem number GT21).

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

## 6.3 3-Hitting Set

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 $C$?

Biological Motivation [PH]:

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.

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.

Complexity: NP-complete [NR99].

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

## 6.4 $k$ -Pathwidth

Problem Definition [KS96]:

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

1. $V = \cup_i X_i$
2. For each edge $(u, v) \in E$, there exists some $i \in \{1, \ldots, l\}$ so that both $u$ and $v$ belong to $X_i$.
3. $\forall\, i, j, h$, if $i \leq j \leq h$, then $X_i \cap X_h \subseteq X_j$.

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

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

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

*Parameter:* $k$

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

Biological Motivation [KS96]:

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.

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.

Complexity: NP-complete [ACP87].

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

## 6.5 Steiner Tree

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?

Biological Motivation [SV97]:

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.

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 [DF95a].

## 6.6 Steiner Tree in HyperCubes

Problem Definition:

A *$q$-dimensional hypercube:* is a graph whose vertex are labelled by all binary sequence of length $q$. Two nodes with labels $x$ and $y$ are adjacent if the $d_H(x, y) = 1$ (where $d_H(x, y)$ is the Hamming distance, for a definition see Problem 2.3).

*Instance:* Binary sequences $X_1, \ldots, X_k$, 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 $X_1, \ldots, X_k$, such that $S$ has at most $M$ edges?

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].

Complexity: NP-complete [DFS99].

Parameterized Complexity: FPT, by the reduction to problem kernel method [Ces04, DF99a].

### 6.7 Vertex Cover

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 $k$?

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].

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

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

## 7 Open

### 7.1 Closest String

Problem Definition:

$d_H$ denotes the Hamming distance, for a definition see Problem 2.3.

> *Instance:* Strings $s_1, s_2, \ldots, s_k$ 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 $i = 1, \ldots, k$?

Biological Motivation:

See problem 2.3 for biological motivation.

Complexity: NP-complete [dlHC00].

Parameterized Complexity:

The algorithm proposed in [Gra03] 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 $k > 4$. Is it 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].

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].

## 7.2 Closest Substring

Problem Definition:

$d_H$ denotes the Hamming distance, for a definition see problem 2.3.

> *Instance:* A collection of $k$ strings, $s_1, s_2, \ldots, s_k$, over an alphabet $\Sigma$, and two 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 $i = 1, \ldots, k$, there is a length-$L$ substring $s'_i$ of $s_i$ with $d_H(s, s'_i) \leq d$?.

Biological Motivation [GHN02]:

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.

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].

## 7.3 Consensus Pattern

Problem Definition:

$d_H$ denotes the Hamming distance, for a definition see problem 2.3.

> *Instance:* Given a collection of $k$ strings, $s_1, s_2, \ldots, s_k$, over an alphabet $\Sigma$, and two non-negative integers $d$ and $L$.
>
> *Parameter:* $d$
>
> *Question:* Is there a string $s$ of length $L$, and, for every $i = 1, \ldots, k$, a length-$L$ substring $s'_i$ of $s_i$ such that $\sum_{i=1}^{k} d_H(s, s'_i) \leq d$? .

Biological Motivation [Gra03]:

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 $d$.

Complexity: NP-complete [FGN02].

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

## 7.4 Gene Duplication and Loss

Problem Definition:

See problem 3.3 for definition of *species tree*, *gene trees* and *cost* model.

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

*Parameters:*

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

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

Biological Motivation [Ste99]:

The *Gene Duplication and Loss* is a biological cost model which has received considerable attention. 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 .

See Problem 3.3 for further biological motivation.

Complexity: NP-complete [Ste99].

Parameterized Complexity:

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.

## 7.5 Edge Bipartization

Problem Definition:

A graph $G$ is bipartite if its vertex set can be partitioned into parts, $X$ and $Y$, in such a way that all the edges in $G$ have one endpoint in $X$ and the other in $Y$.

> *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?

Biological Motivation:

See problem 6.2 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].

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

Parameterized Complexity: Is Edge Bipartization fixed-parameter tractable with respect to the number of allowed edge deletions? [Gra03].

## 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.

**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 an the study of the causes behind this 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 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

## Alphabetical Index

# Hierarchical Index

## FPT

$W[1]$



### $W[1]$-complete



### $W[1]$-hard



### $W[2]$-hard



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

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



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



### Open

## References


[ACP87] S. Arnborg, D. J. Corncil, and A. Proskurowski. Complexity of finding embedding in a $k$-tree. *SIAM J. Alg. Disc. Meth.*, 8:227–284, 1987.

[ADF95] Karl A. Abrahamson, Rodney G. Downey, and Michael R. Fellows. Fixed-parameter tractability and completeness. IV. On completeness for W[P] and PSPACE analogues. *Ann. Pure Appl. Logic*, 73(3):235–276, 1995.

[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.

[AS99] C. Alvarez and M. Serna. The Proper Interval Colored Graph problem for caterpillar trees. Technical report, Departament de Llenguatges i Sistemes Informátics, Universitat Politècnica Catalunya, 1999.

[BDF$^+$] 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 Hibridization. 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. Associaton of Computing Machinery, Academic Press, May 1994.

[BFH+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.

[BL04] David Bryant and Jens Lagergren. Compatibility of Unrooted Phylogenetic Trees is FPT. September 2004. International Workshop on Parameterized and Exact Computation.

[Bod96] Hans L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM J. Comput.*, 25(6):1305–1317, 1996.

[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+] 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 (WG99)*, volume 1665 of *Lecture Notes in Computer Science*, pages 313–324, 1999.

[CW95] M. Cesati and H.T. Wareham. Parameterized complexity analysis in robot motion planning. In *In Proceedings of the 25th IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, Los Alamitos, CA, 1995. IEEE Press.

[Dam04] Peter Damaschke. Parameterized Enumeration, Transversals, and Imperfect Phylogeny Reconstruction. In *1st International Workshop on Parameterized and Exact Computation IWPEC'2004 (part of ALGO'2004)*, volume 3162 of *LNCS,Bergen, Norway*, pages 1–12, 2004.

[DF92] Rod G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness. In *Proceedings of the Twenty-first Manitoba Conference on Numerical Mathematics and Computing (Winnipeg, MB, 1991)*, volume 87, pages 161–178, 1992.

[DF93] Rod Downey and Michael Fellows. Fixed-parameter tractability and completeness. III. Some structural aspects of the $W$ hierarchy. In *Complexity theory*, pages 191–225. Cambridge Univ. Press, Cambridge, 1993.

[DF95a] Rod G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness I. Basic results. *SIAM J. Comput.*, 24(4):873–921, 1995.

[DF95b] Rod G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness. II. On completeness for $W[1]$. *Theoret. Comput. Sci.*, 141(1-2):109–131, 1995.

[DF99a] R. G. Downey and M. R. Fellows. *Parameterized complexity.* Monographs in Computer Science. Springer-Verlag, New York, 1999.

[DF99b] Rodney G. Downey and Michael R. Fellows. Parameterized complexity after (almost) ten years: review and open questions. In *Combinatorics, computation & logic, DMTCS 1999, CATS 1999*, pages 1–33. Springer Verlag, Discrete Mathematics and Theoretical Computer Science, Singapore, 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.

[DFT97] Rod G. Downey, Michael R. Fellows, and Udayan Taylor. The parameterized complexity of relational database queries and an improved characterization of $W[1]$. In *Combinatorics, complexity, & logic (Auckland, 1996)*, Springer Ser. Discrete Math. Theor. Comput. Sci., pages 194–213. Springer, Singapore, 1997.

[DJK+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.

[dlHC00] 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+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.

[DW71] S. Dreyfus and R. Wagner. The Steiner Problem in Graphs. *NETWORKS*, (1):195–207, 1971.

[EHK02] Eleazar Eskin, Eran Halperin, and Richard M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. Technical report, 2002.

[Eva99] Patricia Anne Evans. *Algorithms and Complexity for Annotated Sequence Analysis*. PhD thesis, Department of Computer Science, University of Victoria, Canada, 1999.

[Fel01] Michael R. Fellows. Parameterized complexity: The main ideas and some research frontiers. In *ISAAC*, pages 291–307, 2001.

[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. Hallet, 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.

[FL88a] Michael R. Fellows and Michael A. Langston. Fast self-reduction algorithms for combinatorial problems of VLSI design. In *VLSI algorithms and architectures (Corfu, 1988)*, volume 319 of *Lecture Notes in Comput. Sci.*, pages 278–287. Springer, New York, 1988.

[FL88b] Michael R. Fellows and Michael A. Langston. Layout permutation problems and well-partially-ordered sets. In *Advanced research in VLSI (Cambridge, MA, 1988)*, pages 315–327. MIT Press, Cambridge, MA, 1988.

[FL92] Michael R. Fellows and Michael A. Langston. On well-partial-order theory and its application to combinatorial problems of VLSI design. *SIAM J. Discrete Math.*, 5(1):117–126, 1992.

[FL97] M. Frances and A. Litman. On covering problems of codes. *Theory of Computing Systems*, 30:113–119, 1997.

[GGK+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.

[GNST04] Jens Gramm, Till Nierhoff, Roded Sharan, and Till Tantau. On the complexity of haplotyping via perfect phylogeny. In *Proceedings of Second RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotypes*, Lecture Notes in Bioinformatics. Springer-Verlag, 2004. To appear.

[GNT04] Jens Gramm, Till Nierhoff, and Till Tantau. Perfect path phylogeny haplotyping with missing data is fixed-parameter tractable. In *Proceedings of the 2004 International Workshop on Parameterized and Exact Computation*, Lecture Notes in Computer Science, pages 174–186. Springer-Verlag, 2004.

[Gor86] A. D. Gordon. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *J. Classification*, 3(2):335–348, 1986.

[Gra03] Jens Gramm. *Fixed-Parameter Algorithms for the Consensus Analysis of Genomic Data*. Doktors der naturwissenchaften, der Fakultät für Informations- und Kognitionswissenschaften der Eberhard-Karls-Universität Tübingen zur Erlangung des Grades eines, 2003.

[GSS02] Georg Gottlob, Francesco Scarcello, and Martha Sideri. Fixed-parameter complexity in AI and nonmonotonic reasoning. *Artificial Intelligence*, 138(1-2):55–86, 2002. Knowledge representation and logic programming (El Paso, TX, 1999).

[Gus02] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proceedings of the 6th RECOMB*, pages 166–75. ACM Press, 2002.

[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. Kececioglu 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+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+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.

[LP97] O. Lichtenstein and A. Pneuli. Chacking that the finite state concurrents programs satisfy their linear specification. In *Proceedings of the 12th ACM Symposioum of Principles of Programming Languages*, pages 97–107, 1997.

[LP98] Michael A. Langston and Barbara C. Plaut. On algorithmic applications of the immersion order. *Discrete Math.*, 182(1-3):191–196, 1998. An overview of ongoing work presented at the Third Slovenian International Conference on Graph Theory, Graph theory (Lake Bled, 1995).

[Mai] D. Maier. The Complexity of Some Problems on Subsequences and Supersequences. *Journal of the ACM*, 25,2(1978):322–336.

[NR99] Rolf Neidermeir 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.

[PY99] Christos H. Papadimitriou and Mihalis Yannakakis. On the complexity of database queries. *J. Comput. System Sci.*, 58(3):407–427, 1999. Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (Tucson, AZ, 1997).

[RSV04] Bruce Reed, Kaleigh Smith, and Adrian Vetta. Finding odd cycle transversals. *Oper. Res. Lett.*, 32(4):299–301, 2004.

[SS00] Roded Sharan and Ron Shamir. CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis. In *Proceedings: ISMB*, pages 307–316, 2000.

[SS03] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[Ste92] Michael Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *J. Classification*, 9(1):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 Schiwikoski 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.

[VDGS03] Bafna V., Gusfield D., Lancia G., and Yooseph S. Haplotyping as perfect phylogeny: a direct approach. *Journal of Computational Biology*, 10((3–4)):323–340, 2003.

[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.