Interactive Search for Needles in the Haystack

inspired by Planted Motif Problem and Bioinformatics

Meelis Kull

Hendrik Nigul

CS Theory Days, Feb 2006, Kokõ, Estonia

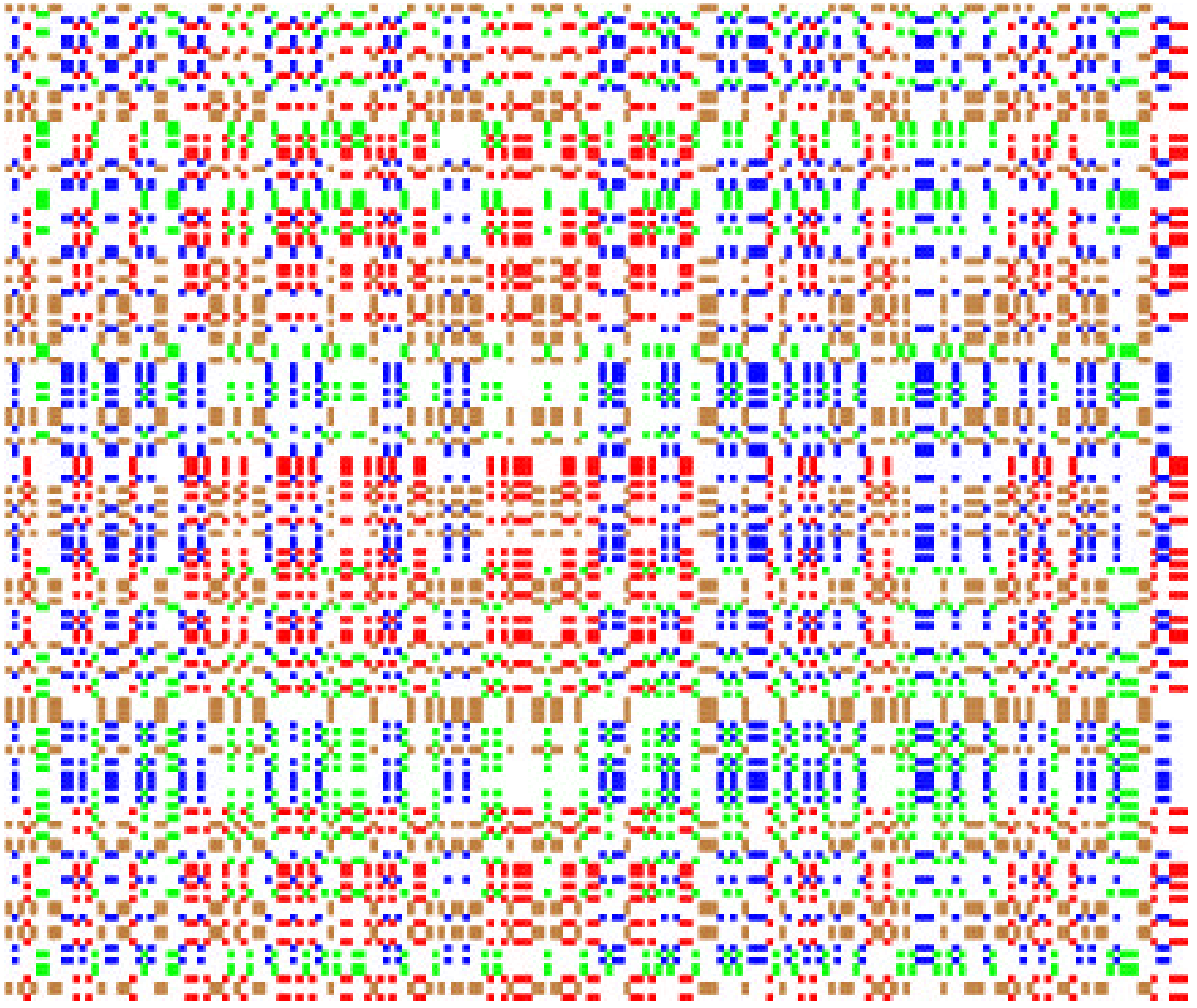
- If you enjoy this interactive seminar:

This is because Bioinformatics is truly fun!!

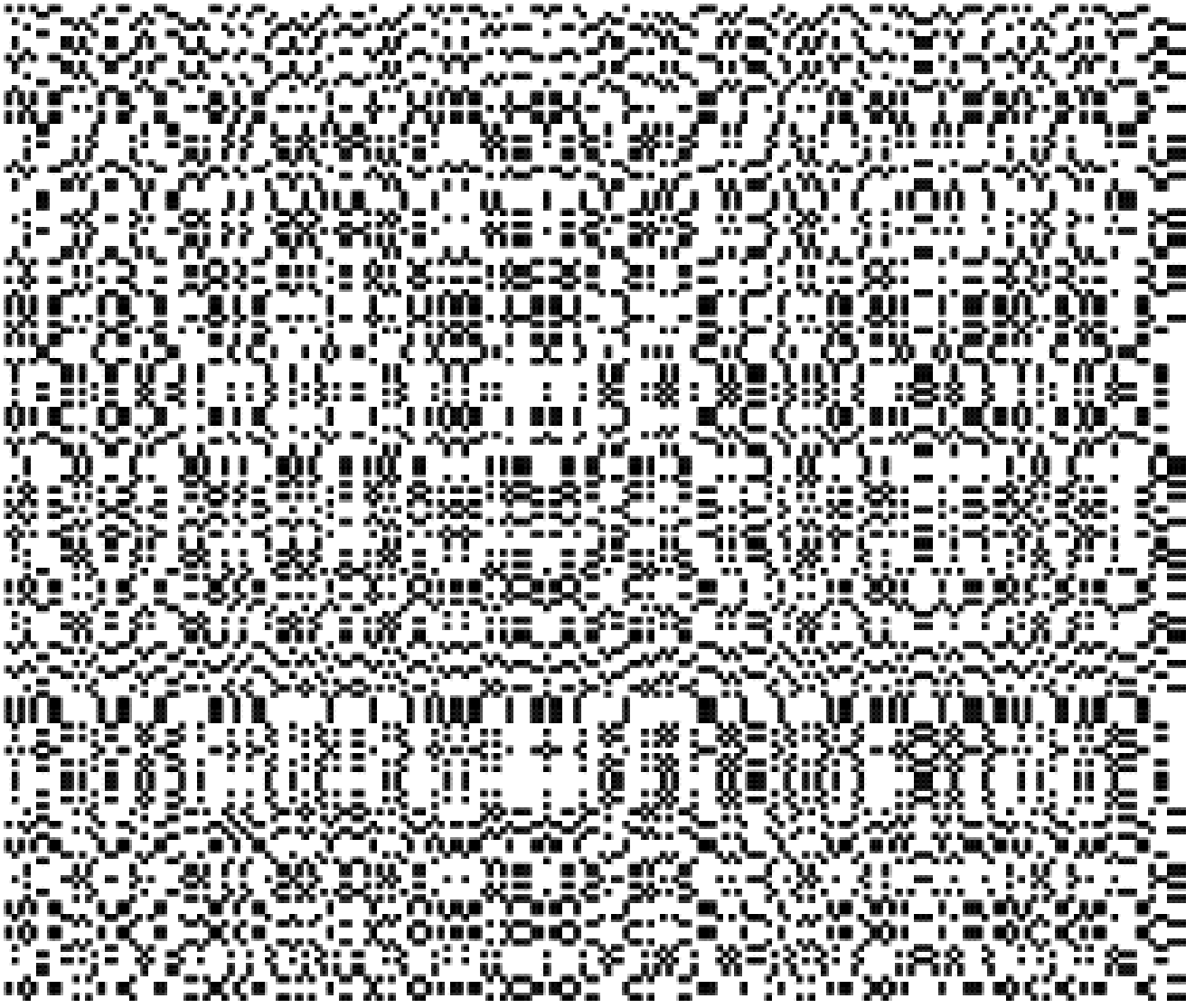
- In case you find it boring:

This is just because I was not good enough.

A Needle in the Haystack



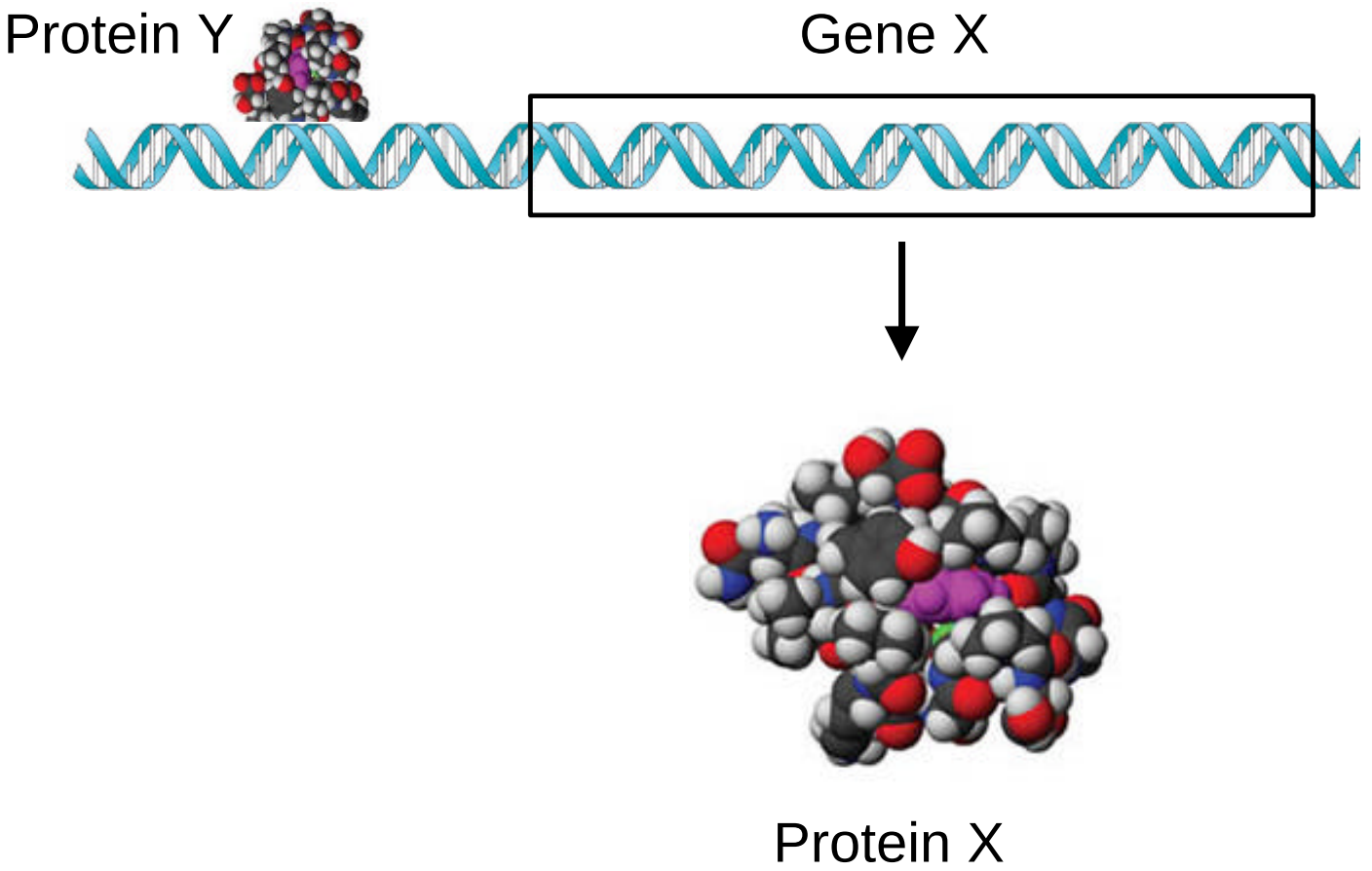
A Needle in the Haystack



A Biological problem: Protein binding sites

- Proteins are able to bind to the DNA if there is a specific pattern of nucleotides
- Example of a pattern: GCGATGAGAT
- Binding occurs also in the case of some mismatches, for example: GCGACGAGAT
- The simplest mathematical model: binding occurs iff there are at most *d* mismatches.

Binding affects “productivity”



How to find binding sites

- Look for genes with similar productivity pattern through different conditions
- Find a substring that occurs in front of all of these genes

How to find binding sites: Back to computer science

CACTTCGTCGTTAGCCACGCAGTGCCGGCTCACAGTCCTGT

Gene X

TGAGAAAGAGCGTCGTTAGCCACGCT

Gene Y

How to find binding sites: Back to computer science

CACTT**CGTCGTTAGCCACGC**AGTGCCGGCTCACAGTCCTGT

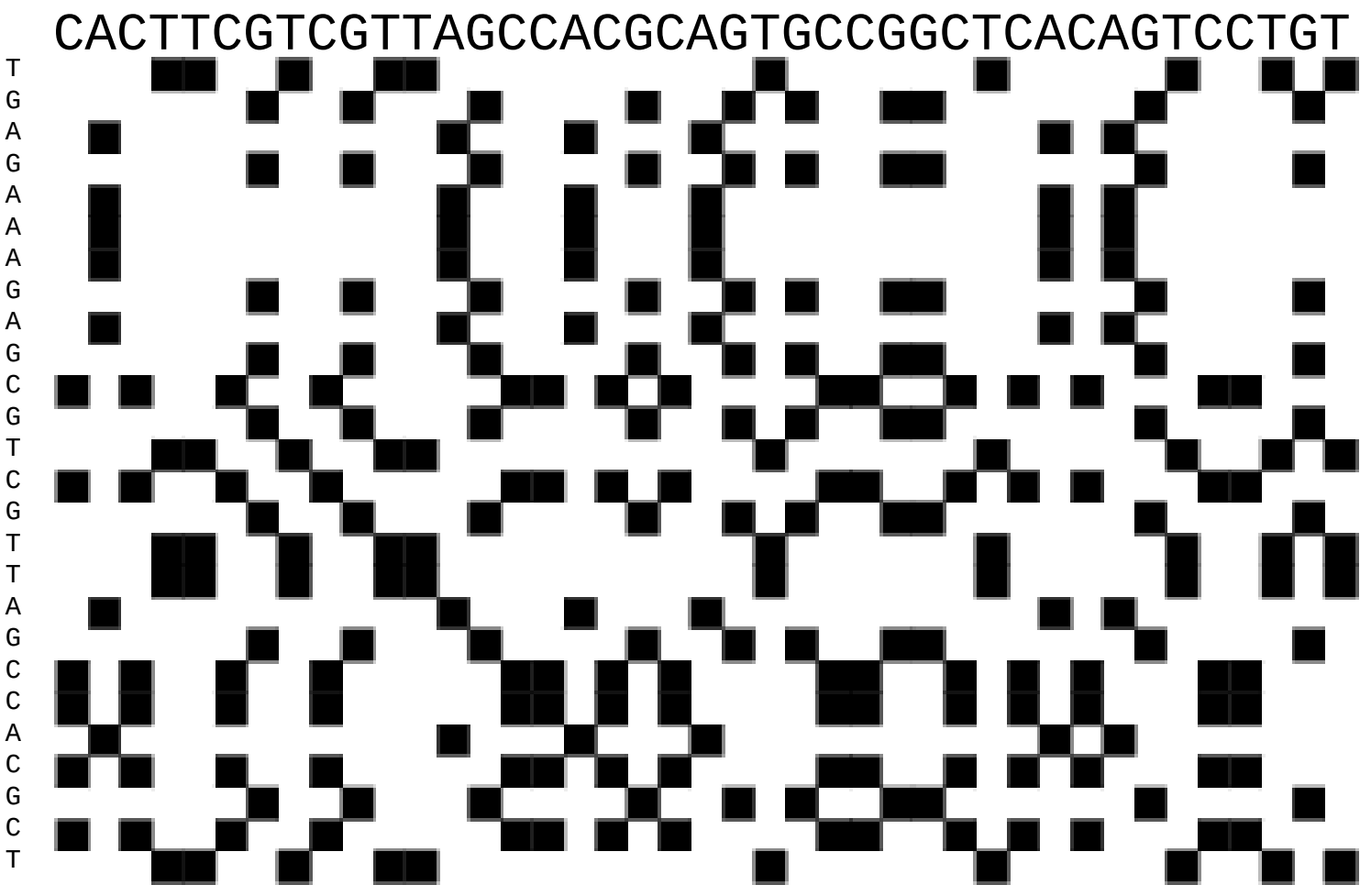
Gene X

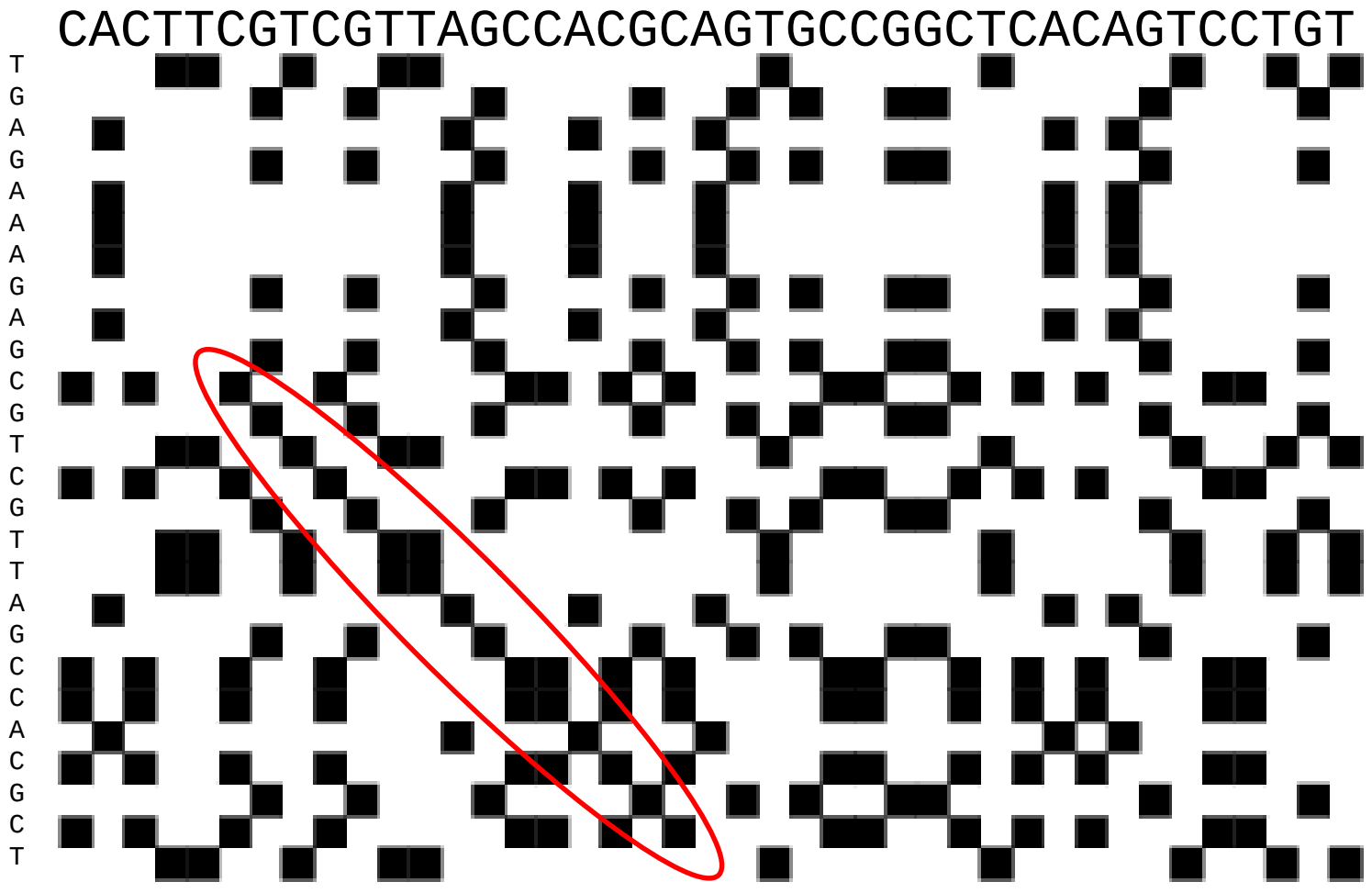
TGAGAAAGAG**CGTCGTTAGCCACGC**T

Gene Y

CACTTCGTCGTTAGCCACGCAGTGCCGGCTCACAGTCCTGT

T
G
A
G
A
A
A
G
A
G
C
G
T
C
G
T
T
A
G
C
C
A
C
G
C
T


CACTTCGTCGTTAGCCACGCAGTGCCGGCTCACAGTCCTGT
TGAGAAAAGAGCGTCGTTAGCCACACGCT


CACTTCGTCGTTAGCCACGCAGTGCCGGCTCACAGTCCTGT
TGAGAAAAGAGCGTCGTTAGCCACGCACT

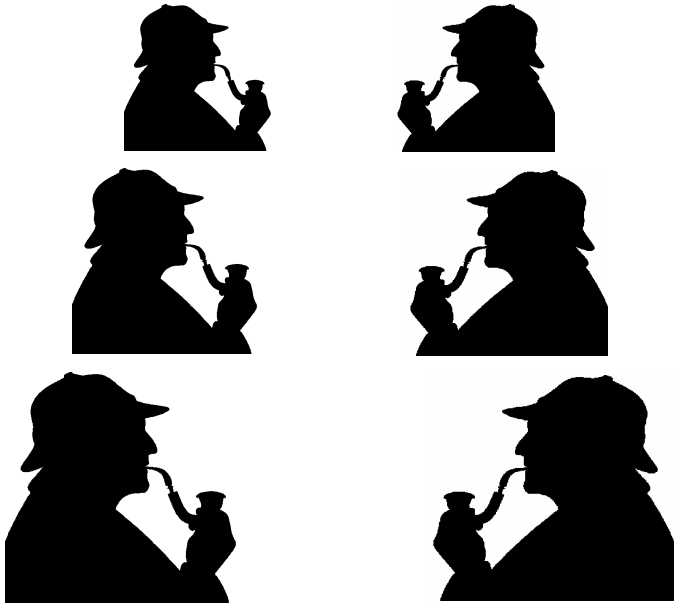
Planted Motif Problem (PMP)

PMP (l,d):

- **Given:** t sequences of length n
- **Find:** motif of length l which has a variant occurring in each sequence. The variant differs from the original by at most d characters

Your Turn! Solve the Problems

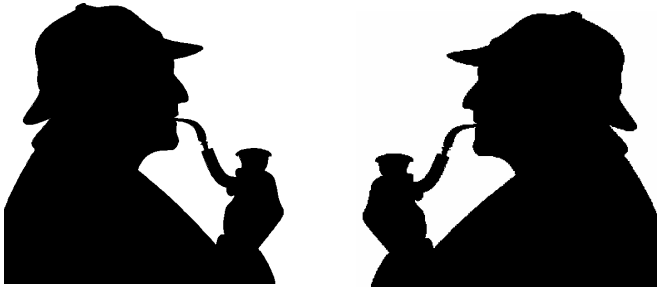
John Nash breaking Soviet codes in the film “A Beautiful Mind”



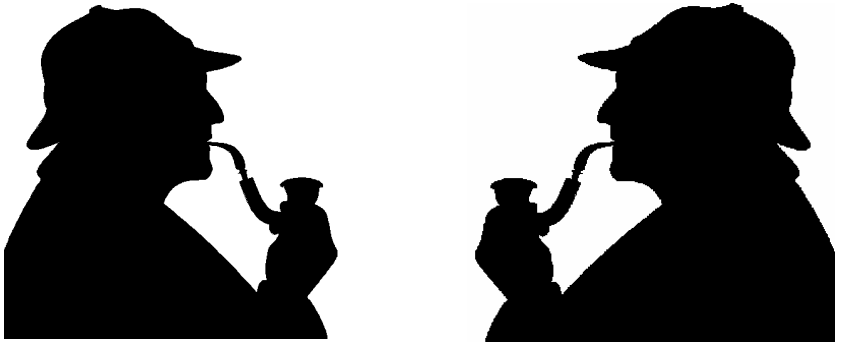
Interactive Search

Please form groups of
4—6 people

15 minutes
for preparation with
an example problem



up to 15 minutes
for solving the challenge



References

S. Rajasekaran, S.Balla, and C.-H. Huang
Exact algorithms for planted motif problems
Journal of Computational Biology, Vol.12,No.8, **2005**

References

S. Rajasekaran, S.Balla, and C.-H. Huang

Exact algorithms for planted motif problems

Journal of Computational Biology, Vol.12,No.8, **2005**

"The development of efficient parallel algorithms for the planted motif problem

is an interesting open problem."