Finding Regulatory Motifs in DNA Sequences

Bioinformatics Algorithms part 6

František Mráz, KSVI

Based on slides from http://bix.ucsd.edu/bioalgorithms/slides.php
and other sources

Regulatory Motifs

Gene Regulation

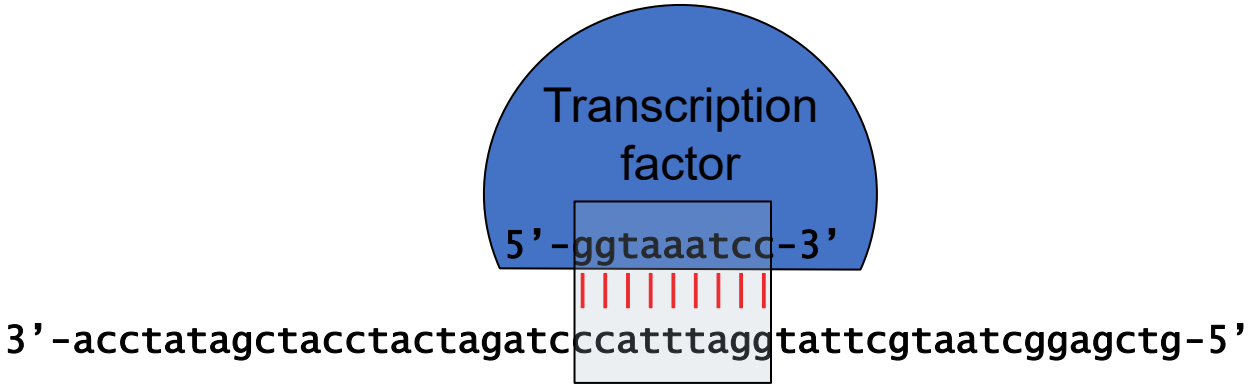
- *A microarray experiment showed that when gene X is knocked out, 20 other genes are not expressed.*
  - ***How can one gene have such drastic effects?***
- As a precursor to transcription (the reading of DNA to construct RNAs that eventually leads to protein synthesis) special proteins bind to the DNA, separate it to enable its reading.
  - How do these proteins know where the coding genes are in order to bind?
  - Genes are relatively rare
    - $O(1{,}000{,}000{,}000)$ bases/genome
    - $O(10{,}000)$ genes/genome
    - $O(1000)$ bases/gene
  - Approximately 1% of DNA codes for genes ($10^3 10^4 / 10^9$)

Gene Regulation

- RNA polymerases seek out *regulatory* or *promoting* region located 100-1000 bp upstream from the coding region
  - They work in conjunction with special proteins called *transcription factors* whose presence enables gene expression
  - Within these regions are the *Transcription Factor Binding Sites* (TFBS), special DNA sequence patterns known as ***motifs*** that are specific to a given transcription factor
- A gene X encodes regulatory protein, also known as transcription factor (TF)
  - The 20 unexpressed genes rely on gene X's TF to induce transcription
  - A single TF may regulate multiple genes

Regulatory Proteins

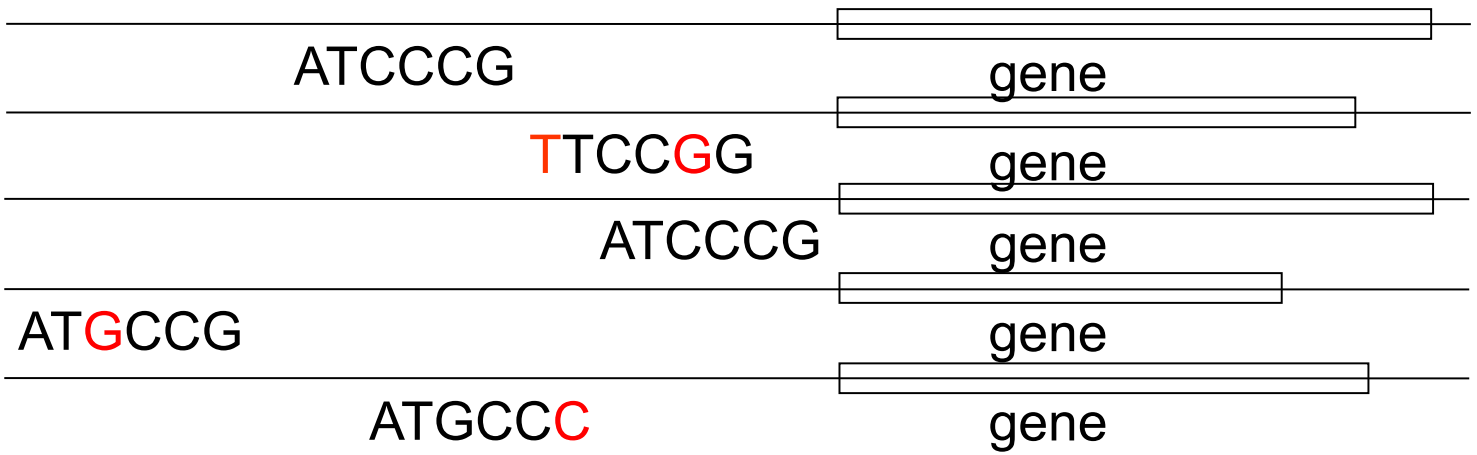
- A TFBS can be located anywhere within the regulatory region.
- TFBS *may vary* slightly across different regulatory regions since non-essential bases could mutate
- Transcription factors are robust (they will still bind) in the presence of small changes in a few bases



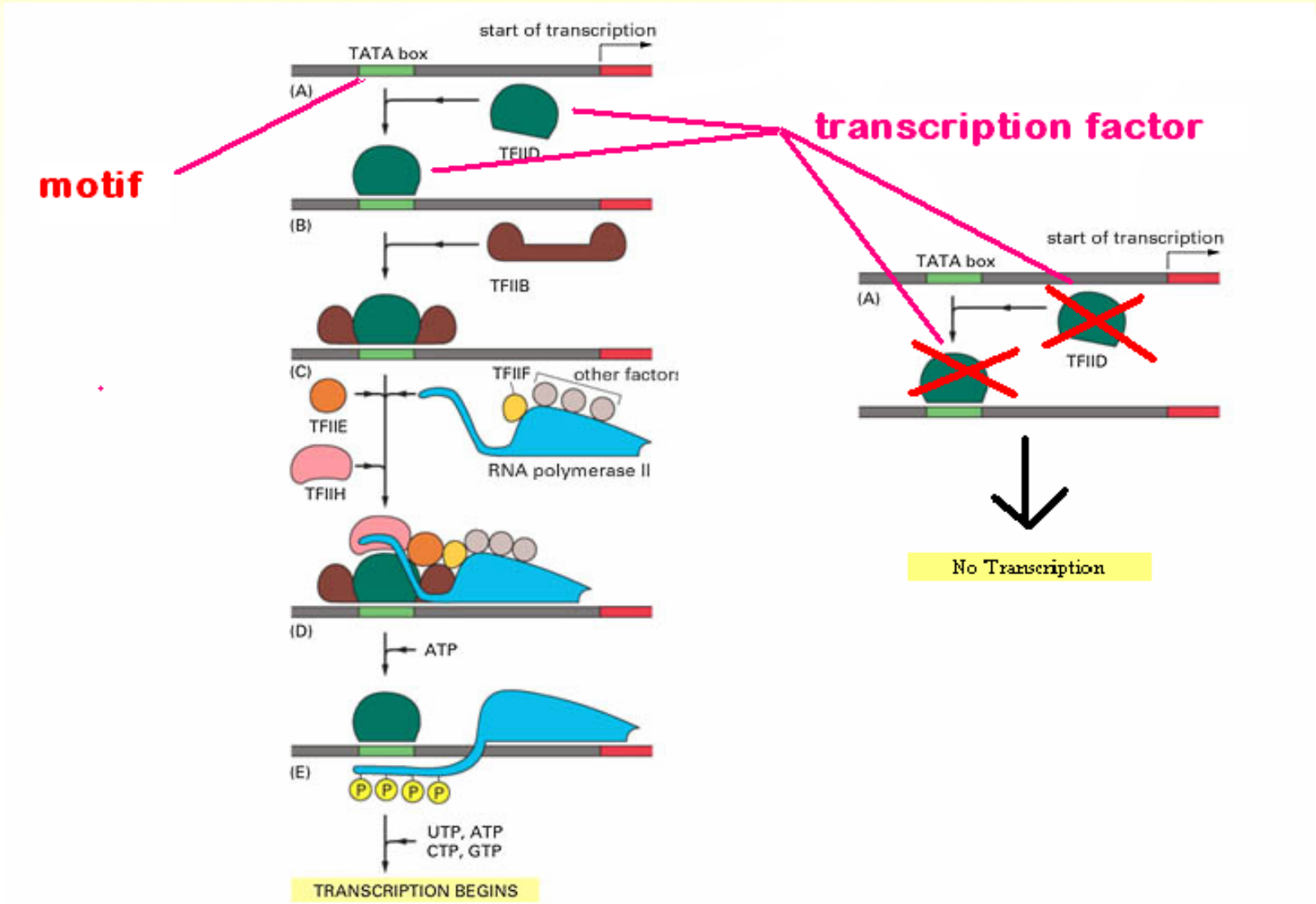
Motifs and Transcriptional Start Sites

- A TFBS can be located anywhere within the Regulatory Region.
- TFBS may vary slightly across different regulatory regions since non-essential bases could mutate
- Transcription factors are robust (they will still bind) in the presence of small changes in a few bases

- So finding the same motif in multiple genes' regulatory regions suggests a regulatory relationship amongst those genes



Transcription Factors and Motifs



Challenge Problem (Pevzner and Sze)

- Find a motif in a sample of
  - 20 "random" sequences (e.g. 600 bases long)
  - each sequence containing an implanted pattern of length 15
  - each pattern appearing with 4 mismatches as $(15,4)$-motif.

Random Sample

```
atgaccgggatactgataccgtatttggcctaggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaatactgggcataaggtaca
tgagtatccctgggatgacttttgggaacactatagtgctctcccgattttgaatatgtaggatcattcgccagggtccga
gctgagaattggatgaccttgtaagtgttttccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatggcccacttagtccacttatag
gtcaatcatgttcttgtgaatggattttaactgagggcatagaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
cggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcat
aacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttctgggtactgatagca
```

- 10 random sequences of length 82 each

Implanting Motif AAAAAAAAGGGGGGG

atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa

Implanting Motif AAAAAAAAGGGGGGG

atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag

CtggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa

Where is the Implanted Motif?

atgaccgggatactgataaaaaaaaggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaactttccgaataaaaaaaaggggggga
tgagtatccctgggatgacttaaaaaaaaggggggtgctctcccgattttgaatatgtaggatcattcgccagggtccga
gctgagaattggatgaaaaaaaaggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggcttatag
gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
cggttttggcccttgttagaggcccccgtaaaaaaaaggggggcaattatgagagagctaatctatcgcgtgcgtgttcat
aacttgagttaaaaaaaaggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggggaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga

- Probability of a given 15-mer in an infinite sequence is $\frac{1}{4^{15}} \approx 9.3 \times 10^{-10}$ (1 every billion bases)
- Assuming 10 strings of length 82, there are $10 \cdot (82 - 15 + 1) = 680$ distinct 15-mers
- Probability of any one 15-mer is $680/4^{15} \approx 0.000\,000\,63$
- So any repeat is rare – solution is easy

Implanting Motif AAAAAAAAAGGGGGGGG with Four Mutations

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

Where is the Motif???

atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacaataaaacggcggga

tgagtatccctgggatgacttaaaataatggagtggtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag

gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

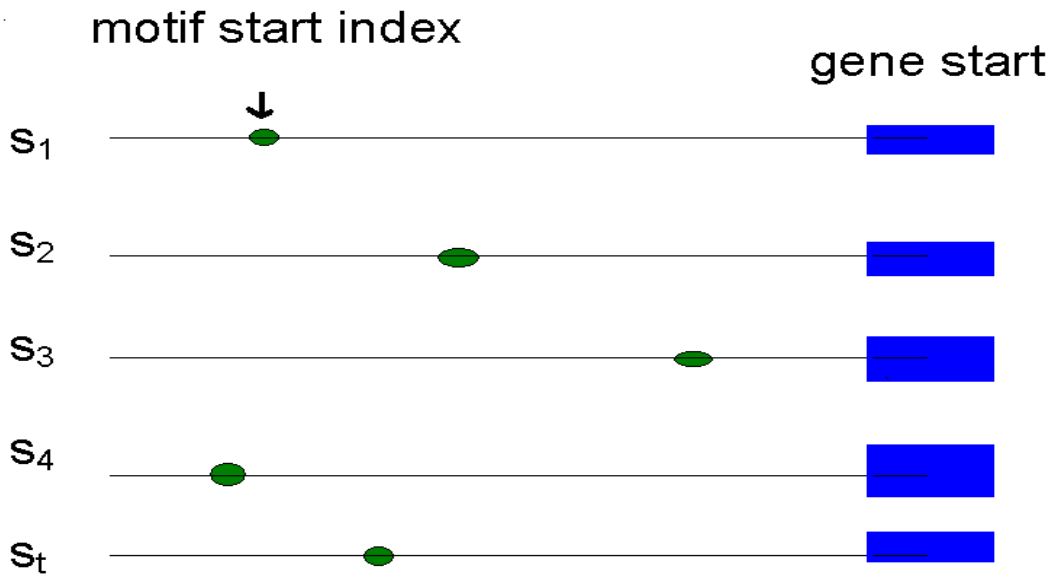
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga

Why Finding (15,4) Motif is Difficult?

Defining Motifs

- To define a motif, let us say we know where the motif starts in the sequence
- The motif start positions in their sequences can be represented as $s = (s_1, s_2, s_3, \ldots, s_t)$



Motifs: Profiles and Consensus

		A	G	G	T	A	C	T	T
		C	C	A	T	A	C	G	T
Alignment		A	C	G	T	T	A	G	T
		A	C	G	T	C	C	A	T
		C	C	G	T	A	C	G	G
	A	3	0	1	0	3	1	1	0
Profile	C	2	4	0	0	1	4	0	0
	G	0	1	4	0	0	0	3	1
	T	0	0	0	5	1	0	1	4
Consensus		A	C	G	T	A	C	G	T

1. Line up the patterns by their start indexes

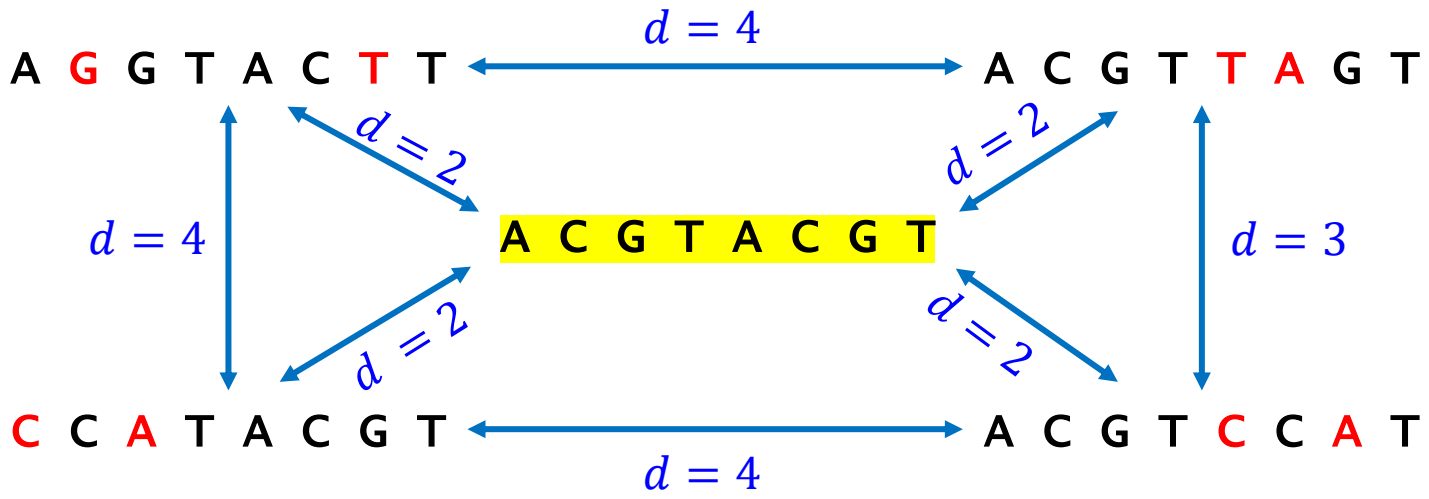
   $s = (s_1, s_2, \ldots, s_t)$

2. Construct matrix profile with frequencies of each nucleotide in columns

3. Consensus nucleotide in each position has the highest score in the column

Consensus

- Think of consensus as an “ancestor” motif, from which mutated motifs emerged
- The distance = Hamming distance between a real motif and the consensus sequence; it is generally less than that for two real motifs

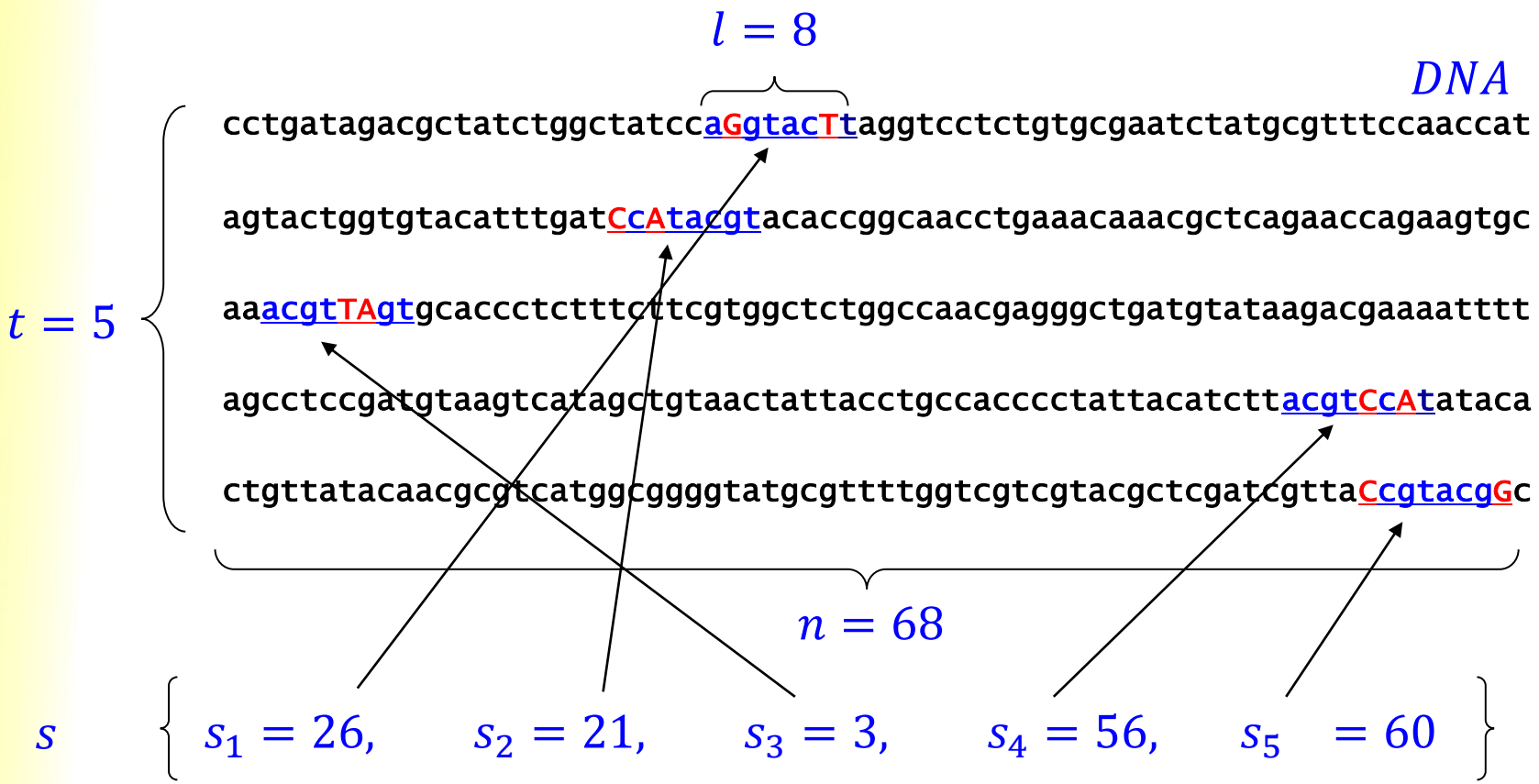


Evaluating Motifs

- We have a guess about the consensus sequence, but how "good" is this consensus?
- Need to introduce a scoring function to compare different guesses and choose the "best" one.

Parameters

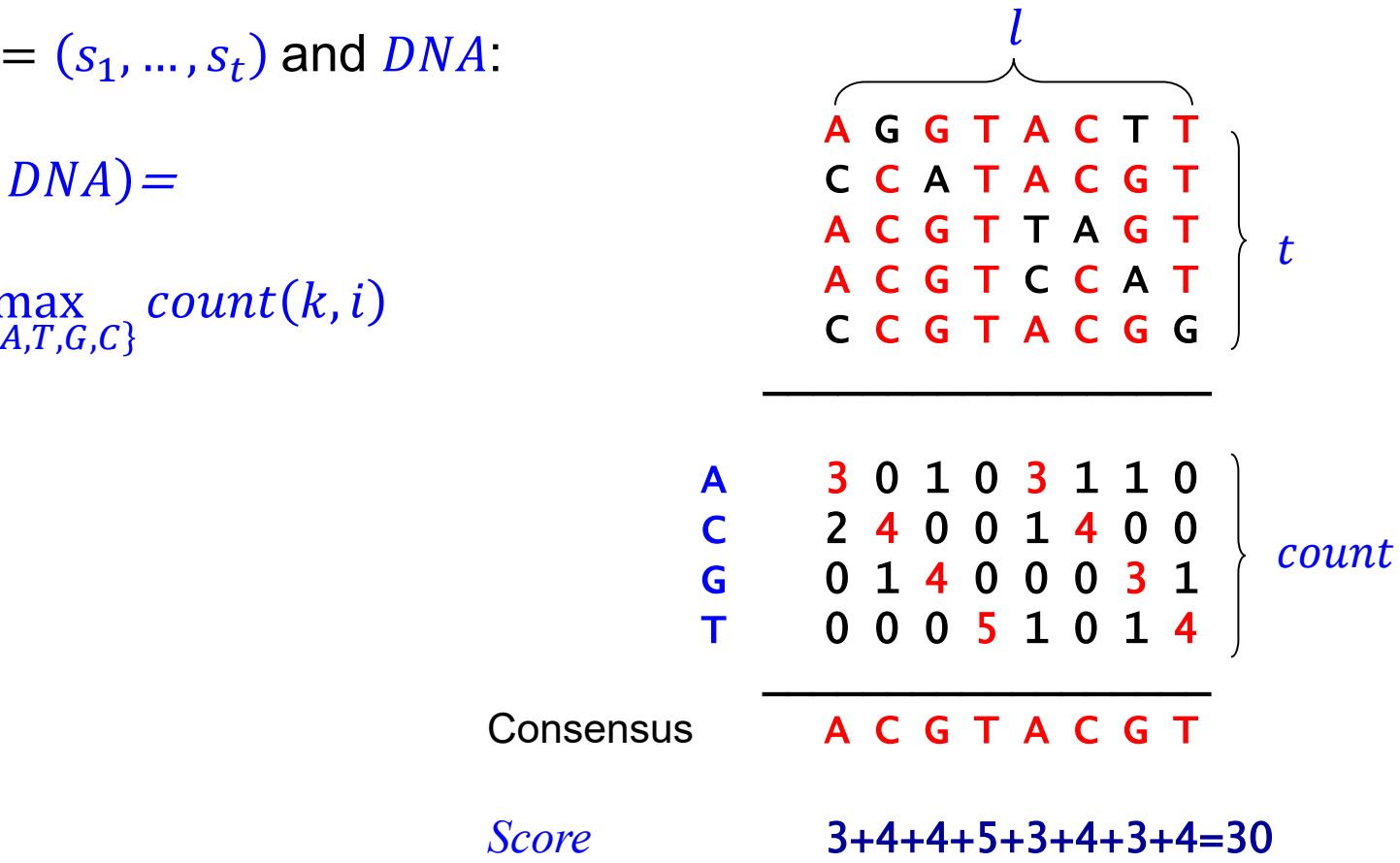
- t – number of sample DNA sequences
- n – length of each DNA sequence
- DNA – sample of DNA sequences ($t \times n$ array)
- l – length of the motif (l-mer)
- s_i – starting position of an l-mer in sequence i
- $s = (s_1, s_2, \ldots, s_t)$ – array of motif's starting positions



Scoring Motifs

- Given $s = (s_1, \ldots, s_t)$ and DNA:

$$Score(s, DNA) = \sum_{i=1}^{l} \max_{k \in \{A,T,G,C\}} count(k, i)$$



Exact Algorithms for Motif Finding Problem

The Motif Finding Problem

- **Goal:** Given a set of DNA sequences, find a set of l-mers, one from each sequence, that maximizes the consensus score
- **Input:** A $t \times n$ matrix of DNA, and l, the length of the pattern to find
- **Output:** An array of t starting positions $s = (s_1, \ldots, s_t)$ maximizing $Score(s, DNA)$

A: Searching the Space of Possible Positions

- Compute the scores for each possible combination of starting positions s
- The best score will determine the best profile and the consensus pattern in DNA
- The goal is to maximize $Score(s, DNA)$ by varying the starting positions s_i, where:

$$s_i = [1, \ldots, n - l + 1]$$
$$i = [1, \ldots, t]$$

BruteForceMotifSearch

BruteForceMotifSearch(DNA, t, n, l)

$bestScore \leftarrow 0$

for each $s = (s_1, \ldots, s_t)$ from $(1,1 \ldots 1)$ to $(n - l + 1, \ldots, n - l + 1)$

if ($Score(s, DNA) > bestScore$)

$bestScore \leftarrow Score(s, DNA)$

$bestMotif \leftarrow s = (s_1, \ldots, s_t)$

return $bestMotif$

Running Time of BruteForceMotifSearch

- Varying $(n - l + 1)$ positions in each of t sequences, we are looking at $(n - l + 1)^t$ sets of starting positions

- For each set of starting positions, the scoring function makes lt operations, so complexity is

$$lt(n - l + 1)^t = O(ltn^t)$$

- That means that for $t = 8$, $n = 1000$, $l = 10$ we must perform approximately 10^{25} computations – it will take thousands of years assuming 10^9 operations/second

- **!!! NOT USABLE !!! we must use a different approach**

B: Searching the Space of Motifs
The Median String Problem

- Given a set of t DNA sequences find a pattern that appears in all t sequences with the minimum number of mutations

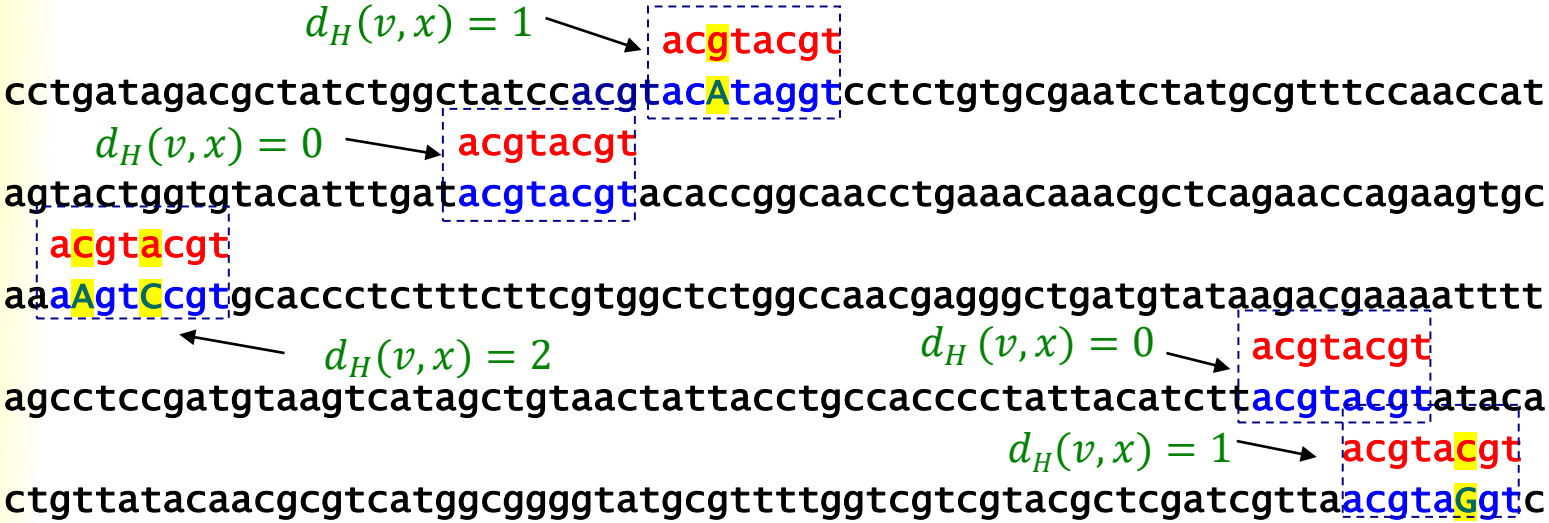
- This pattern will be the motif

- The minimal number of mutations = Hamming distance:
  - $d_H(v, w)$ is the number of nucleotide pairs that do not match when v and w are aligned. For example:

$$d_H(AAAAAA, ACAAAC) = 2$$

Total Distance: An Example

- Given v = "**acgtacgt**" and $s = (s_1, \ldots, s_t)$



v is the sequence in red, x is the sequence in blue

$$TotalDistance(v, DNA) = 1 + 0 + 2 + 0 + 1 = 4$$

Total Distance: Definition

- For each DNA sequence DNA_i , compute all $d_H(v, x)$, where x is an l-mer with starting position s_i $(1 < s_i < n - l + 1)$
- Find minimum of $d_H(v, x)$ among all l-mers x in sequence DNA_i

$$TotalDistance(v, DNA_i) = \min_{x \in DNA_i} d_H(v, x)$$

- $TotalDistance(v, DNA)$ is the sum of the minimum Hamming distances for each DNA sequence DNA_i

$$TotalDistance(v, DNA) = \sum_{i=1}^{t} \min_{x \in DNA_i} d_H(v, x)$$

The Median String Problem: Formulation

- **Goal:** Given a set of DNA sequences, find a median string
- **Input:** A $t \times n$ matrix DNA, and l, the length of the pattern to find
- **Output:** A string v of l nucleotides that minimizes $TotalDistance(v, DNA)$ over all strings of that length

Median String Search Algorithm

MedianStringSearch(DNA, t, n, l)

$bestWord$ ← AAA…A

$bestDistance$ ← ∞

for each l-mer s from AAA…A to TTT…T

if $TotalDistance(s, DNA) < bestDistance$

$bestDistance$ ←$TotalDistance(s, DNA)$

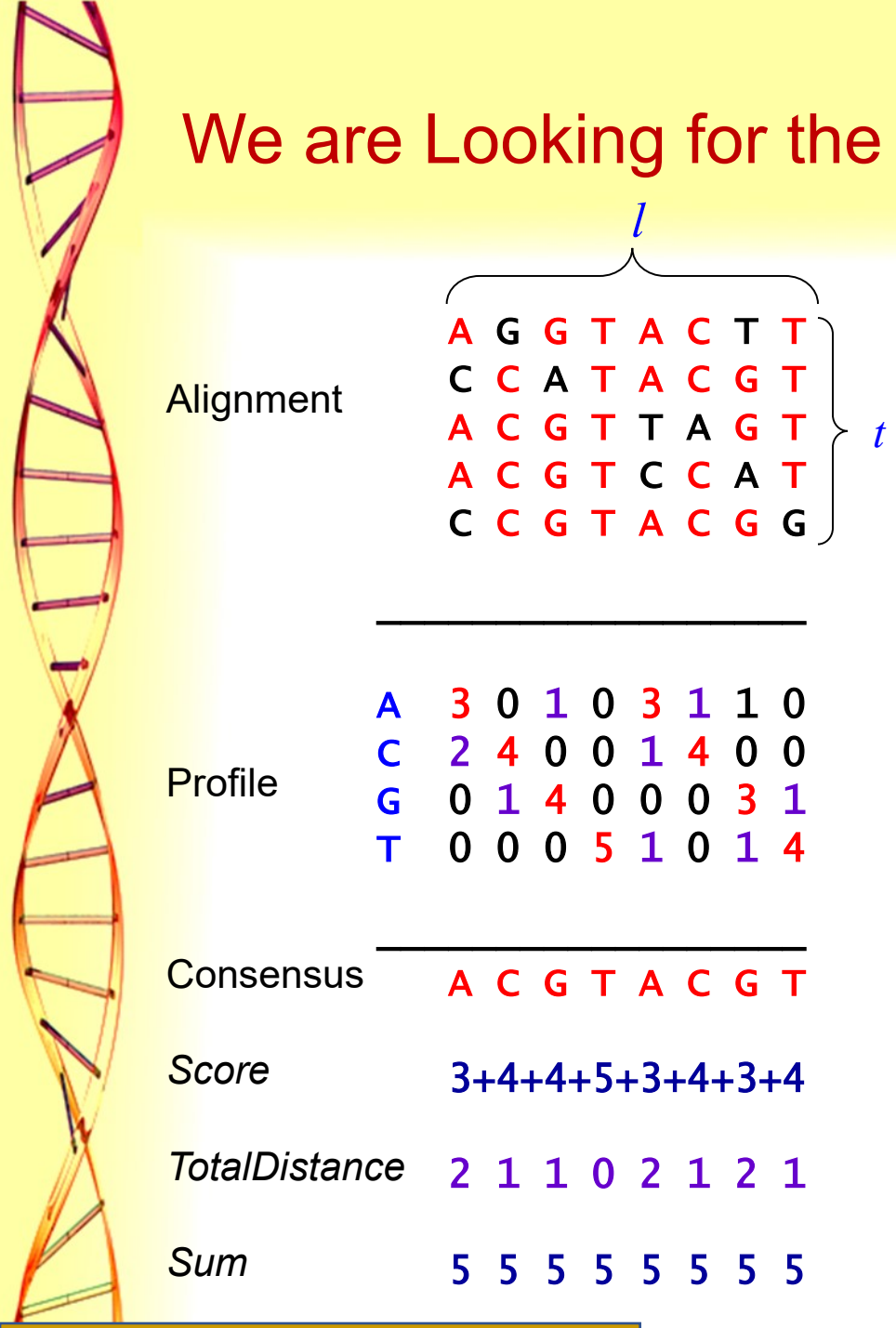
$bestWord$ ← s

return $bestWord$

Motif Finding Problem = Median String Problem

- The *Motif Finding* is a maximization problem while *Median String* is a minimization problem
- However, the *Motif Finding* problem and *Median String* problem are computationally equivalent
- We will show that minimizing $TotalDistance$ is equivalent to maximizing $Score$

We are Looking for the Same Thing



- At any column i
  $Score_i + TotalDistance_i = t$
- Because there are l columns
  $Score + TotalDistance = l \cdot t$
- Rearranging:

  $Score = l \cdot t - TotalDistance$
- $l \cdot t$ is constant; the minimization of $TotalDistance$ is equivalent to the maximization of $Score$

Motif Finding Problem vs. Median String Problem

- Why bother reformulating the Motif Finding problem into the Median String problem?
  - The Motif Finding Problem needs to examine all the combinations for s. That is $(n - l + 1)^t$ combinations!!!
  - The Median String Problem needs to examine all 4^l combinations for v. This number is relatively smaller

$n = 1000, l = 10, t = 8$

$$l \cdot t \cdot (n - l + 1)^t = 10 \cdot 8 \cdot (1000 - 10 + 1)^8 \approx 7.4 \times 10^{25}$$

$$l \cdot t \cdot (n - l + 1) \cdot 4^l = 10 \cdot 8 \cdot (1000 - 10 + 1) \cdot 4^{10} \approx 8.3 \times 10^{10}$$

Motif Finding: Improving the Running Time

- Recall the BruteForceMotifSearch:

BruteForceMotifSearch(DNA, t, n, l)

$bestScore \leftarrow 0$

for each $s = (s_1, \ldots, s_t)$ from $(1,1 \ldots 1)$ to $(n - l + 1, \ldots, n - l + 1)$

if ($Score(s, DNA) > bestScore$)

$bestScore \leftarrow Score\ (s, DNA)$

$bestMotif \leftarrow s = (s_1, \ldots, s_t)$

return $bestMotif$

Branch-and-bound searching

Motif Finding: Improving the Running Time

BFMotifSearch(DNA, t, n, l)

    $s \leftarrow (1,1 \ldots 1)$
    $bestMotif \leftarrow$ FindInSeq(s, $1, DNA$, t, n, l)
    return $bestMotif$

FindInSeq(s, i, DNA, t, n, l)

    $bestScore \leftarrow 0$
    for j from 1 to $n - l + 1$
        $s_i \leftarrow j$
        if $i < t$
            $s \leftarrow$ FindInSeq(s, $i + 1, DNA$, t, n, l)
        if ($Score(s, DNA) > bestScore$)
            $bestScore \leftarrow Score\ (s, DNA)$
            $bestMotif \leftarrow s$
    return $bestMotif$

Can We Do Better?
Yes! Branch and Bound

- Sets of $s = (s_1, s_2, \ldots, s_t)$ may have a weak profile for the first i positions of $(s_1, s_2, \ldots, s_i)$
- Every row of alignment may add at most l to $Score$
- New notation: $Score(s, i, DNA)$ denotes a partial consensus score of the $i \times l$ alignment matrix involving only the first i rows of DNA with starting positions $(s_1, s_2, \ldots, s_i)$
- Optimism: if all subsequent $(t - i)$ positions $(s_{i+1}, \ldots, s_t)$ add

$$(t - i) \cdot l \text{ to } Score(s, i, DNA)$$

- If $Score(s, i, DNA) + (t - i) \cdot l < bestScore$, it makes no sense to search in vertices of the current subtree

Motif Finding: Improving the Running Time

BFMotifSearch(DNA, t, n, l)

    $s \leftarrow (1,1 \ldots 1)$
    $bestMotif \leftarrow$ FindInSeq(s, $1, DNA$, t, n, l)
    return $bestMotif$

FindInSeq(s, i, DNA, t, n, l)

    $bestScore \leftarrow 0$
    for j from 1 to $n - l + 1$
        $s_i \leftarrow j$
        $bestPossibleScore \leftarrow Score(s, i, DNA) + (t - i) \cdot l$
        if $bestPossibleScore > bestScore$
            if $i < t$
                $s \leftarrow$ FindInSeq(s, $i + 1, DNA$, t, n, l)
            if ($Score(s, DNA) > bestScore$)
                $bestScore \leftarrow Score\ (s, DNA\ )$
                $bestMotif \leftarrow s$
    return $bestMotif$

Branch and Bound Applied to Median String Search

- Note that if the total distance for a prefix is greater than that for the best word so far:

$$TotalDistance\,(prefix, DNA\,) > BestDistance$$

- There is no use to explore the remaining part of the word
- Similarly to BFMotifSearch() we can skip extending the prefix and rather try a different prefix

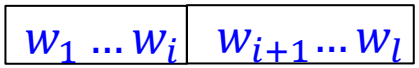
Improving the Bounds

- By using the condition

$$TotalDistance\ (prefix, DNA)\ >\ BestDistance$$

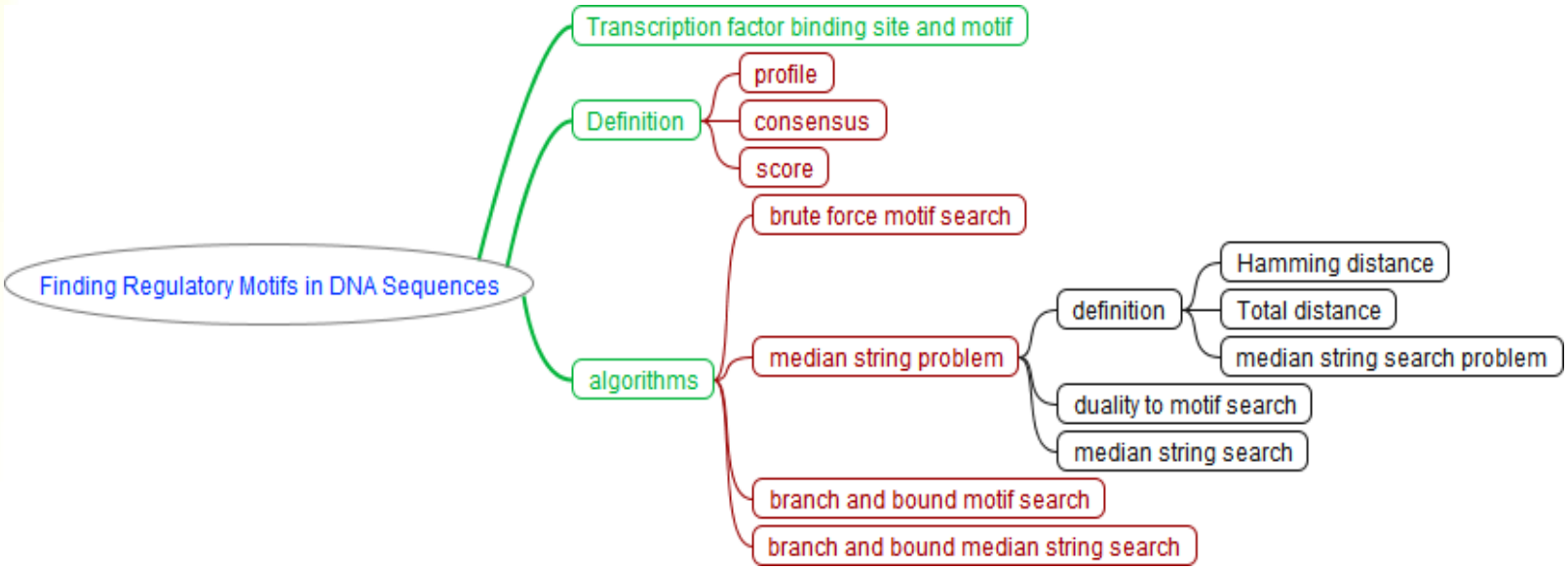
we suppose optimistic 0 distance of the remaining suffix of the motif

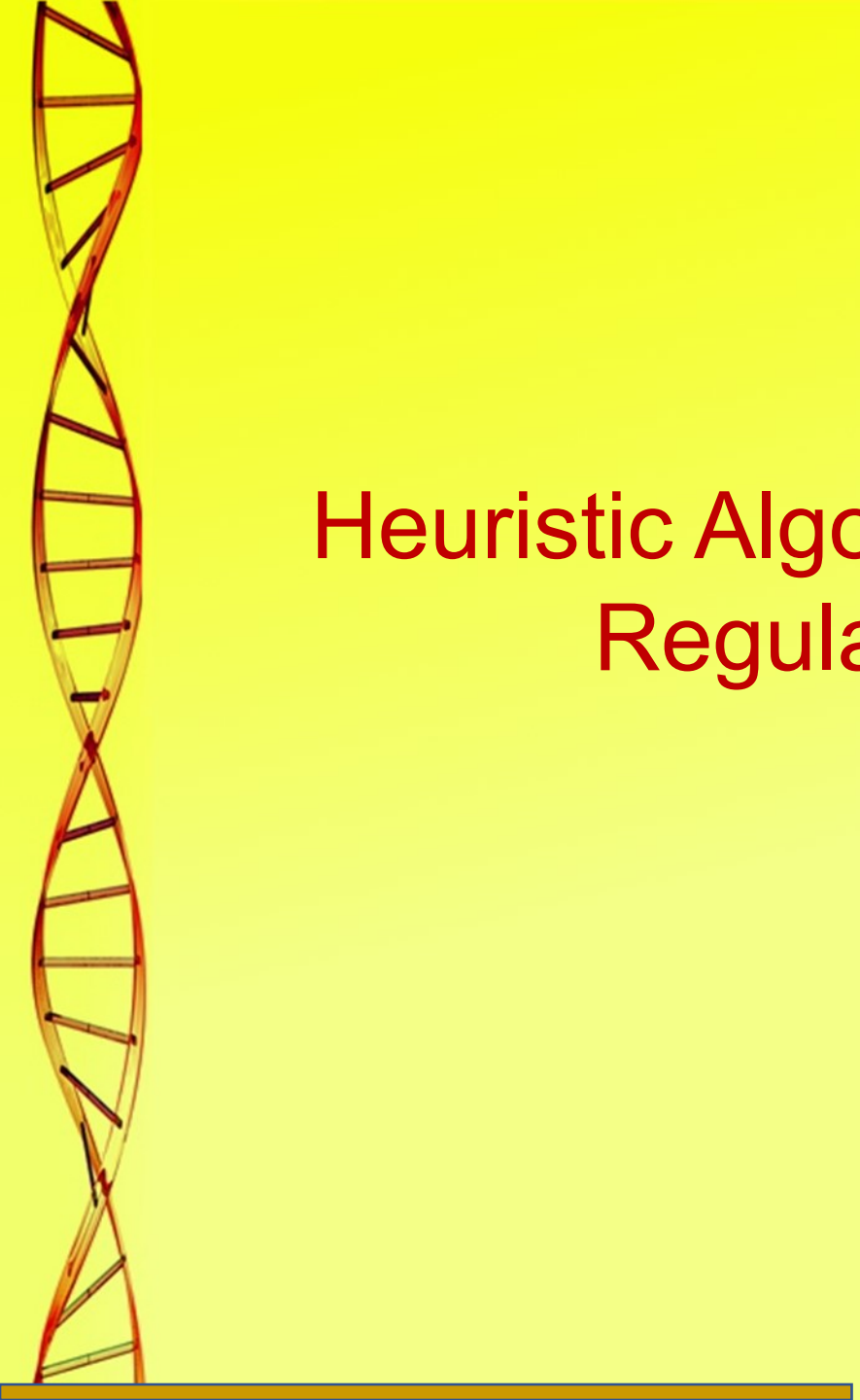
- Given an l-mer w, divided into two parts at point i
  - u : prefix $w_1, \dots, w_i$,
  - v : suffix $w_{i+1}, \dots, w_l$



- No instances of u in the sequence have distance less than the minimum distance
- Note this does not tell us anything about whether u is part of any motif. We only get a minimum distance for prefix u
- Since u and v are two substrings of w, and included in motif w, we can assume that the minimum distance of u plus minimum distance of v can only be less than the minimum distance for w
- If $TotalDistance\ (u, DNA) + \mathrm{TotalDistance}(v, DNA)\ > BestDistance$, motif uv cannot give a better (lower) score than $TotalDistance\ (u, DNA) + TotalDistance(v, DNA)$

What we Have Learned





Heuristic Algorithms for Finding Regulatory Motifs

More on the Motif Problem

- Exhaustive Search and Median String are both exact algorithms
- They always find the optimal solution, though they may be too slow to perform practical tasks
- Many algorithms sacrifice optimal solution for speed

CONSENSUS: Greedy Motif Search

- Find two closest l-mers in sequences 1 and 2 and forms $2 \times l$ alignment matrix with $Score(s, 2, DNA)$

AGGTACTT
CCATACGT

- At each of the following $t - 2$ iterations CONSENSUS finds a "best" l-mer in sequence i from the perspective of the already constructed $(i - 1) \times l$ alignment matrix for the first $(i - 1)$ sequences

i

AGGTACTT
CCATACGT
ACGTTAGT
ACGTCCAT
CCGTACGG

- In other words, it finds an l-mer in sequence i maximizing

$$Score(s, i, DNA)$$

under the assumption that the first $(i - 1)$ l-mers have been already chosen

- CONSENSUS sacrifices optimal solution for speed: in fact the bulk of the time is actually spent locating the first two l-mers

Some Motif Finding Programs

- CONSENSUS
  Hertz, Stromo (1989)
- GibbsDNA
  Lawrence et al (1993)
- MEME
  Bailey, Elkan (1995)
- RandomProjections
  Buhler, Tompa (2002)
- MULTIPROFILER
  Keich, Pevzner (2002)
- MITRA
  Eskin, Pevzner (2002)
- Pattern Branching
  Price, Pevzner (2003)

Some Motif Finding Programs

- MEME (1995)
- EXTREME (2014)
- YAMDA (2018) – GPU and Deep Learning
- AlignAce, Amadeus, CisModule, FIRE, Gibbs Motif Sampler, PhyloGibbs, SeSiMCMC, ChIPMunk and Weeder.
- SCOPE, MotifVoter, and MProfiler – ensemble motif finders – uses several algorithms simultaneously
- benchmark: Weirauch *et al*.: **Evaluation of methods for modeling transcription factor sequence specificity,** Nature Biotechnology 31, 126–134 (2013), doi:10.1038/nbt.2486
- Hashim, F. A., Mabrouk, M. S., & Al-Atabany, W.: Review of Different Sequence Motif Finding Algorithms. *Avicenna journal of medical biotechnology*, *11*(2), 2019, 130–148.

Planted Motif Challenge

- **Input:**
- Input
  - t sequences of length n each
  - l, d
- **Output:**
  - Motif M, of length l
  - Variants of interest have a hamming distance at most d from M

When is the Problem Solvable?

- Assume that the background sequences are **independent and identically-distributed** (**i.i.d.**)
- the probability that a given l-mer C occurs with up to d substitutions at a given position of a random sequence is:

$$p_{(l,d)} = \sum_{i=0}^{d} \binom{l}{i} \left(\frac{3}{4}\right)^{i} \left(\frac{1}{4}\right)^{l-i}$$

- the expected number of length l motifs that occur with up to d substitutions at least once in each of the t random length n sequences is:

$$E(l,d,t,n) = 4^{l}\left(1-\left(1-p_{(l,d)}\right)^{n-l+1}\right)^{t}$$

- Very rough estimate – overlapping motifs not modelled, and the assumption of i.i.d. background distribution is usually incorrect.

When is the Problem Solvable?

- the expected number of length l motifs that occur with up to d substitutions at least once in each of the random t length n sequences is:

Probability that a sequence of length n contains an l-mer with Hamming distance at most d from a given l-mer

$$E(l,d,t,n) = 4^l \left(1 - \left(1 - p_{(l,d)}\right)^{n-l+1}\right)^t$$

Probability that an l-mer at a given position has Hamming distance $> d$ from a given l-mer

When is the Problem Solvable?

- 20 random sequences of length 600 are expected to contain more than one $(9, 2)$-motif by chance ($E(9,2,20,600) > 1$), whereas the chances of finding a random $(10, 2)$-motif are less than 10^{-7}.
- So, the $(9, 2)$ problem is impossible to solve, because “random motifs” are as likely as the planted motif. However, for the $(10,2)$ the probability of a random motif occurring is very small.

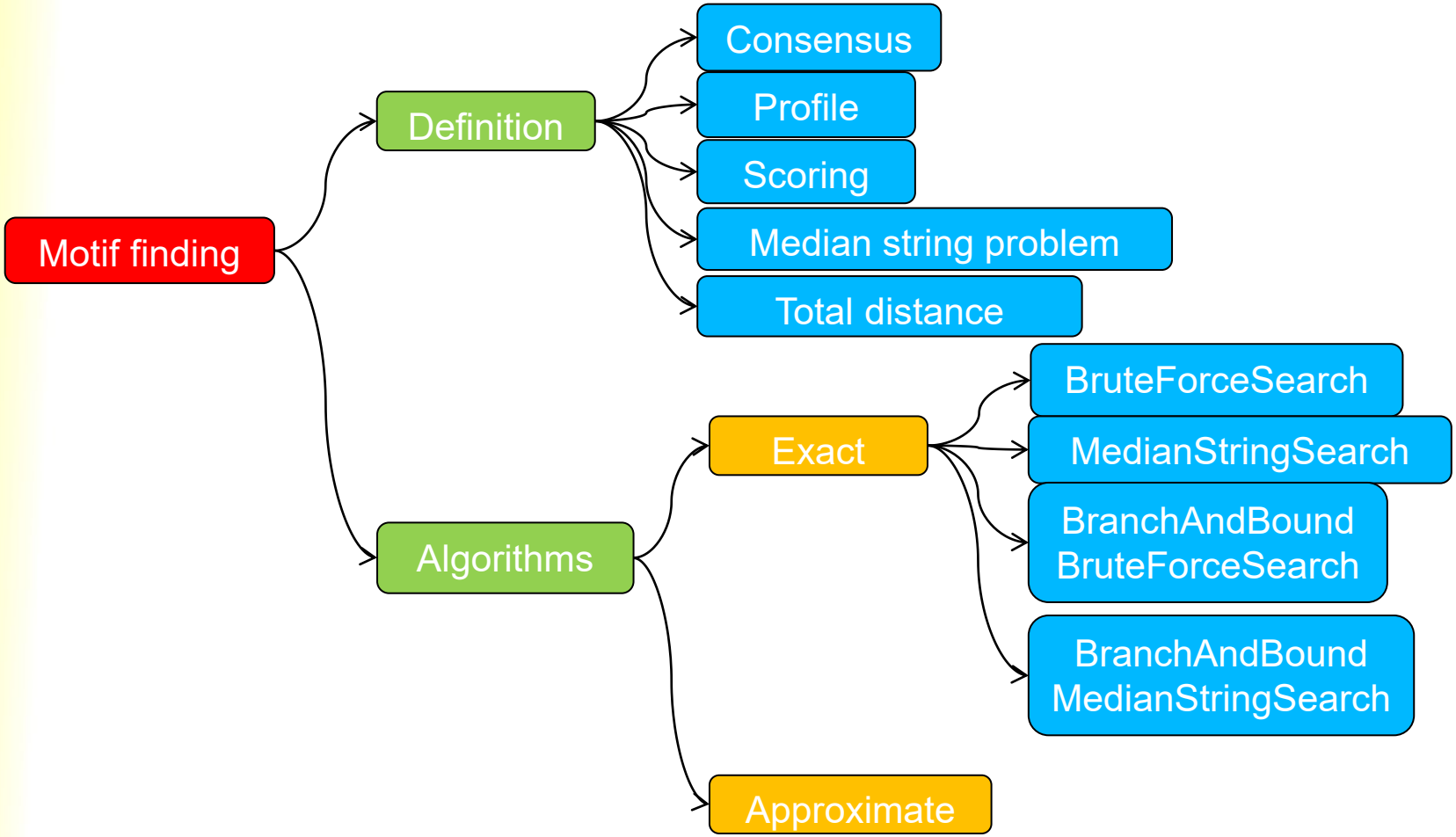
How to Proceed?

- Exhaustive search?

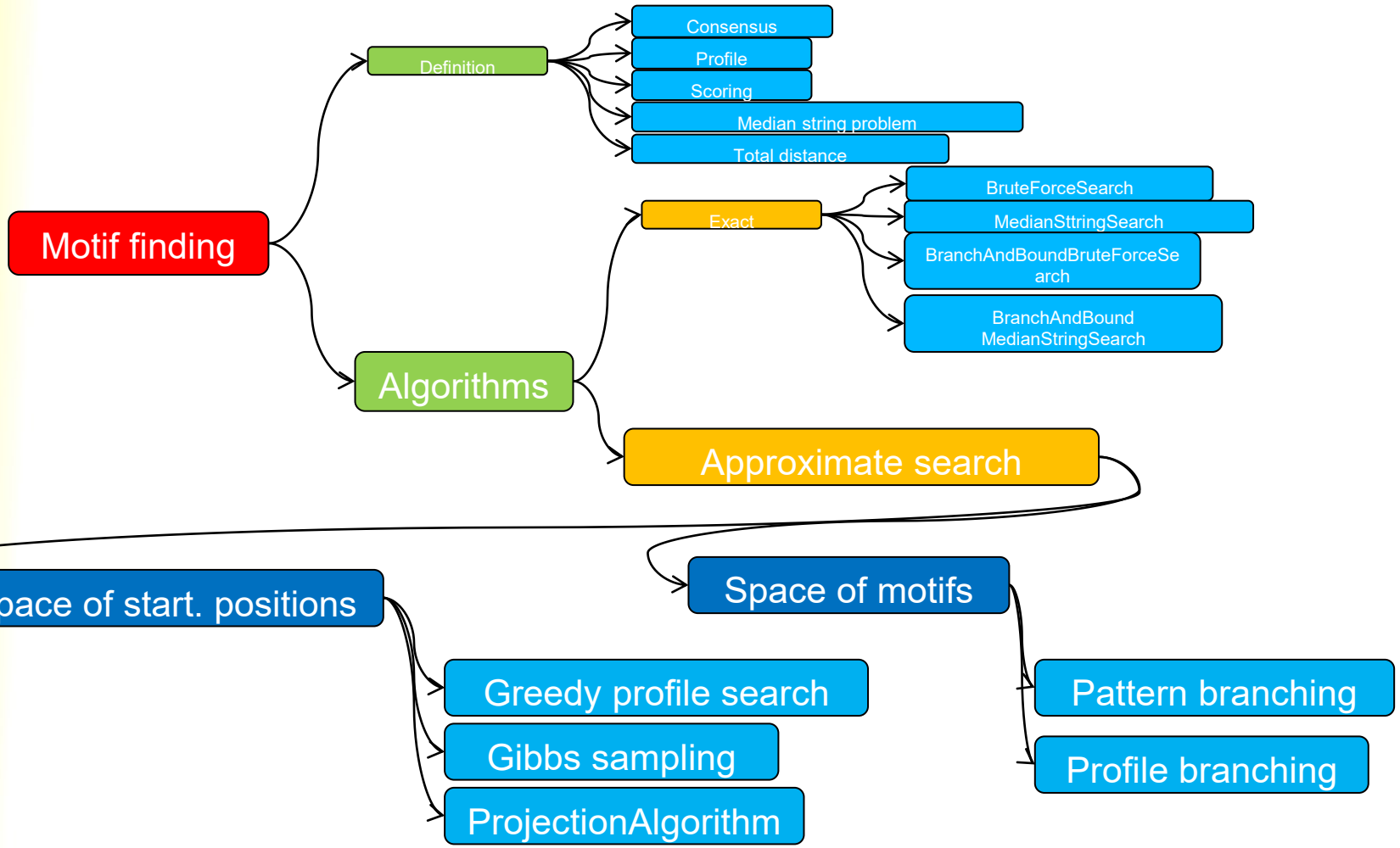
ACGTCCTGATCAATGTCCGTAATGCTAGCTACTGATCCTAGAGA
GATCTATGAGACTATCAGCTTCTATCATCTGGTTGGCATGCAAT
GCTACGTGACCGTAGCATGCATCGACATGCACTGACGATGCATG
CTTCGATCGATCTAGCCTAGCTGACTGACGTAGCATCGATCGTA
CGACGTACGTGACATCTCACTGCTACTTATCGCGATCGCTATCT
GGTCTCGGCCTTGAGCATGGCCACATGGATGCTAGCATGGCATA

- Run time is high

Motif Finding



Motif Finding



Heuristic Search

- Searching the space of starting positions
  - Gibbs sampling
  - The Projection Algorithm
- Searching the space of motifs
  - Pattern Branching
  - Profile Branching

Notation

- t sequences $DNA_1, \ldots, DNA_t$, each of length n
- $l > 0$ integer; the goal is to find an l-mer in each of the sequences such that the „similarity“ between these l-mers is maximized
- Let $(a_1, \ldots, a_t)$ be a list of l-mers contained in $DNA_1, \ldots, DNA_t$. These form a $t \times l$ alignment matrix.
- Let $X(a) = (x_{ij})$ denote the corresponding $4 \times l$ profile, where x_{ij} denotes the frequency with which we observe nucleotide i at position j. Usually, we add pseudo counts to ensure that X does not contain any zeros (*Laplace correction*)
  - Let $n_{a,j}$ denote the number of occurrences of nucleotide a_i at position j; p_{a_i} denote the probability of the occurrence of nucleotide a_i in all $DNA_1, \ldots, DNA_t$
  - β is a weight of the correction

$$x_{a_i,j} = \frac{n_{a_i,j} + \beta p_{a_i}}{t + \beta}$$

Greedy Profile Search

- the probability that a given l-mer $z = z_1 \cdots z_l$ was generated by a given profile X

$$P(\mathrm{z} \mid X) = \prod_{i=1}^{l} x_{z_i,i}$$

- Any l-mer that is similar to the consensus string of X will have a "high" probability, while dissimilar ones will have "low" probabilities.

Greedy Profile Search

$$P(z \mid X) = \prod_{i=1}^{l} x_{z_i,i}$$

	1	2	3	4	5	6	7	8	9
A	.33	.60	.08	0	0	.49	.71	.06	.15
C	.37	.13	.04	0	0	.03	.07	.05	.19
G	.18	.14	.81	1	0	.45	.12	.84	.20
T	.12	.13	.07	0	1	.03	.09	.05	.46

profile without the Lapace correction

- $P(\text{CAGGTAAGT} \mid X) = 0.02417294365920$ and
- $P(\text{TCCGTCCCA} \mid X) = 0.00000000982800$

Greedy Profile Search

- For a given profile X and a sequence s we can find the X-most probable l-mer in s

$$z = \arg \max_{z\ l\text{-}mer\ in\ s} P(z \mid X)$$

- **Algorithm:** *"start with a random **seed** profile and then attempt to improve on it using a greedy strategy"*
  - Given sequences $DNA_1, \ldots, DNA_t$ of length n, randomly select one l-mer a_i from each sequence DNA_i and construct an initial profile X. For each sequence DNA_i, determine the X-most probable l-mer a_i'. *Set X equal to the profile obtained from $a_1', \ldots, a_t'$ and repeat.*
- Does not work well
  - the number of possible seeds is huge
  - In each iteration, the greedy profile search method can change any or all t of the profile l-mers and thus will jump around in the search space.

Gibbs Sampling

- *"start with a random seed profile, then change one l-mer per iteration."*
- Randomly select an l-mer a_i in each input sequence DNA_i .
- Randomly select one input sequence DNA_h.
- Build a $4 \times l$ profile X from $a_1, \ldots, a_{h-1}, a_{h+1}, \ldots, a_t$.
- Compute background frequencies Q from input sequences

$$DNA_1, \ldots, DNA_{h-1}, DNA_{h+1}, \ldots, DNA_t .$$

- For each l-mer $a \in DNA_h$, compute

$$w(a) = \frac{P(a \mid X)}{P(a \mid Q)}$$

- Set $a_h = a$, for some $a \in DNA_h$ chosen randomly with probability

$$\frac{w(a)}{\sum_{a' \in DNA_h} w(a')}$$

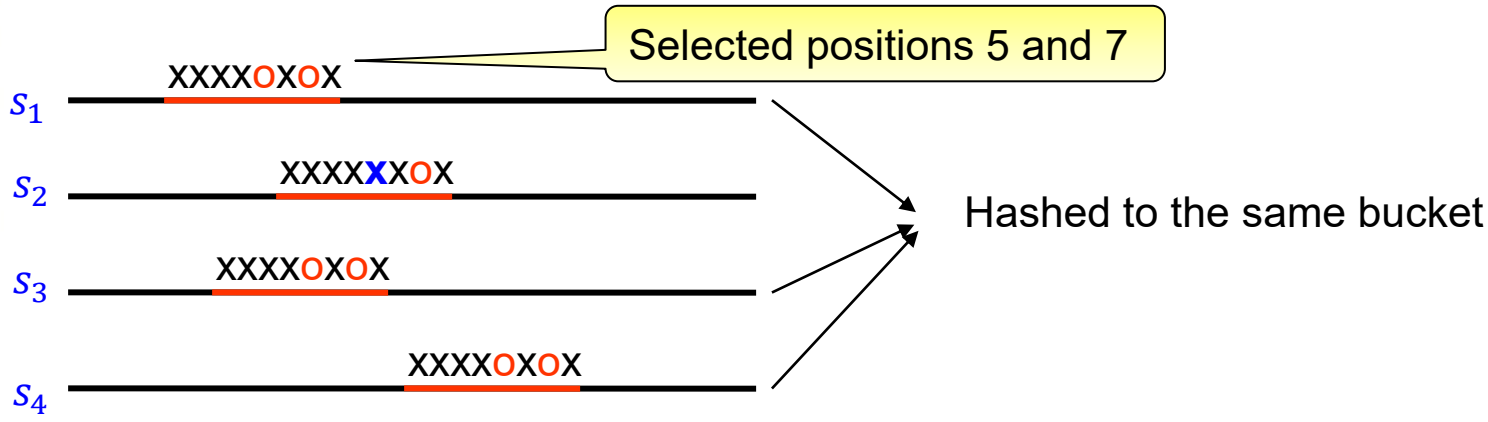
- Repeat until "converged"

Gibbs Sampling

- often works well in practice
- difficulties:
  - finding subtle motifs
  - its performance degrades if the input sequences are skewed, that is, if some nucleotides occur much more often than others. The algorithm may be attracted to low complexity regions like AAAAAA....
- modifications:
  - use “relative entropies” rather than frequencies
  - Another modification is the use of “phase shifts”: The algorithm can get trapped in local minima (of total distance between motif and the whole DNA) that are shifted up or down a few positions from the strongest pattern. To address this, in every M-th iteration the algorithm tries shifting some a_i up or down a few positions

The Projection Algorithm

- *"choose k of l positions at random, then use the k selected positions of each l-mer x as a hash function $h(x)$. When a sufficient number of l-mers hash to the same bucket, it is likely to be enriched for the planted motif"*



- Viewing x as a point in l-dimensional Hamming space, $h(x)$ is the projection of x onto a k-dimensional subspace.

The Projection Algorithm

- Choose distinct k of the l positions at random. For an l-mer x, the hash function $h(x)$ is obtained by concatenating the selected k residues of x.
- If M is the (unknown) motif, then we call the bucket with hash value $h(M)$ the planted bucket.
- The key idea is that, if $k < l - d$, then there is a good chance that some of the t planted instances of M will be hashed to the planted bucket, namely all planted instances for which the k hash positions and d substituted positions are disjoint.
- So, there is a good chance that the planted bucket will be enriched for the planted motif and it contains more entries than an average bucket.

The Projection Algorithm – an Example

- s_1 cagtaat
- s_2 ggaactt
- s_3 aagcaca

- and the (unknown) $(3, 1)$-motif $M =$ aaa, hashing with $k = 2$ using the first 2 of $l = 3$ positions produces the following hash table:

$h(x)$	positions	$h(x)$	positions	$h(x)$	positions
aa	(1,5), (2,3), (3,1)	cg	–	ta	(1,4)
ac	(2,4), (3,5)	ct	(2,5)	tc	–
ag	(1,2), (3,2)	ga	(2,2)	tg	–
at	(1,6)	gc	(3,3)	tt	(2,6)
ca	(1,1), (3,4), (3,6)	gg	(2,1)		
cc	–	gt	(1,2)		

- The motif M is planted at positions $(1, 5)$, $(2, 3)$ and $(3, 1)$ in this example, all three instances hash to the planted bucket $h(M) =$ aa.

The Projection Algorithm – Finding the Planted Bucket

- the algorithm does not know which bucket is the planted bucket.
- it attempts to recover the motif from every bucket that contains at least s elements, where s is a threshold that is set to identify buckets that look as if they may be the planted bucket.
- In other words, the first part of the Projection algorithm is a heuristic for finding promising sets of l-mers in the sequence. It must be followed by a refinement step that attempts to generate a motif from each such set.
- The algorithm has *three main parameters*:

A. the projection size k,

B. the bucket (inspection) threshold s, and

C. and the number of independent trails m.

The Projection Algorithm
A. Projection Size

- the algorithm should hash a significant number of instances of the motif into the planted bucket, while avoiding contamination of the planted bucket by random background l-mers.
- What k to choose so that the average bucket will contain less than 1 random l-mer?
- Since we are hashing $t(n-l+1)$ l-mers into 4^k buckets, if we choose k such that $4^k > t(n-l+1)$, then the average bucket will contain less than one random l-mer.
- For example, in the Challenge $(15,4)$-problem, with $t = 20$ and $n = 600$, we must choose k to satisfy $k < l - d = 15 - 4 = 11$ and

$$k > \frac{\log\big(t(n-l+1)\big)}{\log(4)} = \frac{\log(20(600-15+1))}{\log(4)} \approx 6.76$$

The Projection Algorithm
B. Bucket Threshold

- In the Challenge Problem, a bucket size of $s = 3$ or 4 is practical, as we should not expect too many instances to hash to the same bucket in a reasonable number of trials.
- If the total amount of sequence is very large, then it may be that one cannot choose k to satisfy both

$$k < l - d \qquad \text{and} \qquad 4^k > t(n - l + 1).$$

- In this case, set $k = l - d - 1$, as large as possible, and set the bucket threshold s to twice the average bucket size

$$2 \cdot \underbrace{\frac{t(n - l + 1)}{4^k}}_{\text{Average bucket size}}$$

The Projection Algorithm
C. Number of Independent Trials

- Our goal: to choose m so that with the probability at least $q = 0.95$, the planted bucket contains s or more planted motif instances in at least one of the m trials.
- let $p'(l, d, k)$ be the probability that a given planted motif instance hashes to the planted bucket, that is:

$$p'(l, d, k) = \frac{\binom{l-d}{k}}{\binom{l}{k}}$$

- Then the probability that fewer than s planted instances hash to the planted bucket in a given trial is $B_{t,p'(l,d,k),s}$. Here, $B_{t,p,s}$ is the probability that there are fewer than s successes in t independent Bernoulli trials, each trial having probability p of success (binomial probability distribution function).

The Projection Algorithm – Number of Independent Trials

- Binomial distribution

$$B_{t,p,s} = \sum_{i=0}^{s-1} \binom{t}{i} p^i (1-p)^{t-i}$$

- If the algorithm is run for m trails, the probability that s or more planted instances hash to the planted bucket in at least one trial is:

$$1 - \left(B_{t,p'(l,d,k),s}\right)^m \geq q.$$

- To satisfy this equation, choose:

$$m = \left\lceil \frac{\log(1-q)}{\log\left(B_{t,p'(l,d,k),s}\right)} \right\rceil$$

- Using this criterion for m, the choices for k and s above require at most thousands of trails, and usually many fewer, to produce a bucket containing sufficiently many instances of the planted motif.

Projection Algorithm

1. Choose k of the l positions at random
2. Hash all l-mers of the given sequences into buckets
3. Inspect all buckets with s or more positions and refine the found motifs
4. Repeat m times, return the motif with the best score

Motif Refinement

- we have already found k of the planted motif residues. These, together with the remaining $(l-k)$ residues, should provide a strong signal that makes it easy to obtain the motif in only a few iterations of refinement.
- We will process each bucket of size at least s to obtain a candidate motif. Each of these candidates will be "refined" and the best refinement will be returned as the final solution.
- Candidate motifs are refined using the expectation maximization (EM) algorithm. This is based on the following probabilistic model:
  - An instance of some length-l motif occurs exactly once per input sequence.
  - Instances are generated from a $4 \times l$ weight matrix model W, whose (i,j)-th entry gives the probability that base i occurs in position j of an instance, independent of its other positions.
  - The remaining $n-l$ residues in each sequence are chosen randomly and independently according to some background distribution.

Motif Refinement - Expectation Maximization

- Based on the following probabilistic model:
  - An instance of some length-l motif occurs exactly once per input sequence.
  - Instances are generated from a $4 \times l$ weight matrix model W, whose (i, j)-th entry gives the probability that base i occurs in position j of an instance, independent of its other positions.
  - The remaining $n - l$ residues in each sequence are chosen randomly and independently according to some background distribution.
- Let S be a set of t input sequences and let P be the background distribution. EM-based refinement seeks a weight matrix model W^* that maximizes the likelihood ratio

$$\frac{\Pr(S|W^*, P)}{\Pr(S|P)} = \max_W \frac{\Pr(S|W, P)}{\Pr(S|P)}$$

that is, a motif model that explains the input sequences much better than P alone.

Motif Refinement

- The position at which the motif occurs in each sequence is not fixed a priori, making the computation of W^* difficult, because $\Pr(S|W,P)$ must be summed over all possible locations of the instances.
- To address this, the EM algorithm uses an iterative calculation that, given an initial guess W_0 at the motif model, converges linearly to a locally maximum-likelihood model in the neighbourhood of W_0.
- An initial guess W_h for a bucket h is formed as follows: set $W_h(i,j)$ to the frequency of base i among the j-th positions of all l-mers in h.
- This guess forms a centroid for h, in the sense that positions that are well conserved in h are strongly biased in W_h, while poorly conserved positions are less biased. To avoid zero entries in W_h, add a Laplace correction of b_i, to $W_h\,(i,j)$, where b_i is the background probability of residue i in the input.

Motif Refinement

- Once we have used the EM algorithm to obtain a refinement W_h^* of W_h, the final step is to identify the planted motif from W_h^*. *(Details of EM skipped.)*
- To do so, we select from each input sequence the l-mer x with the largest likelihood ratio:

$$\frac{\Pr(x|W_h^*)}{\Pr(x|P)}$$

- The resulting multiset T of l-mers represents the motif in the input that is most consistent with W_h^*.
- Let C_T be the consensus of T and let $s(T)$ be the number of elements of T whose Hamming distance to C_T is $\leq d$. The algorithm returns the sequence C_T^* that maximizes $s(T)$ (and has the maximal score), over all considered buckets h and over all trials.

Summary of Projection Algorithm

Input: sequences $s_1, \ldots, s_t$, parameters k, s and m

Output: best guess motif

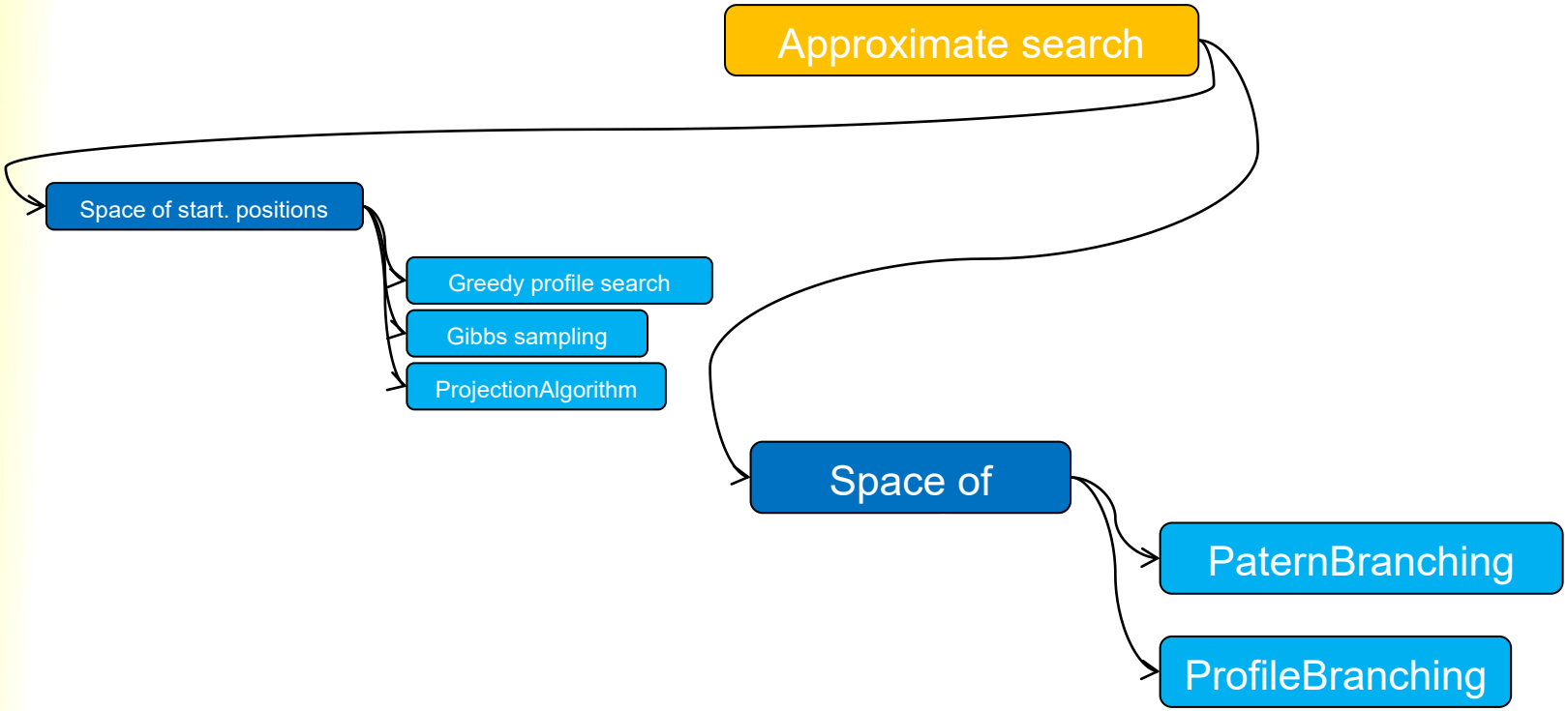
- for $i = 1$ to m do
  - choose k different positions $I_k \subset \{1, 2, \ldots, l\}$
  - for each l-mer $x \in s_1, \ldots, s_t$ do
    - compute hash value $h_{I_k}(x)$
    - Store x in hash bucket
  - for each bucket with $\geq s$ elements do
    - refine bucket using EM algorithm
- return consensus pattern of the best refined bucket

Performance of Projection Algorithm

- The performance of PROJECTION compared to other motif finders on the (l, d)-motif problem. The measure is the average performance defined as $|K \cap P| \setminus |K \cup P|$ where K is the set of the lt residue positions of the planted motif instances, and P is the corresponding set of positions predicted by the algorithm.

l	d	Gibbs	WINNOWER	SP-STAR	PROJECTION
10	2	0.20	0.78	0.56	0.82
11	2	0.68	0.90	0.94	0.91
12	3	0.03	0.75	0.33	0.81
13	3	0.60	0.92	0.92	0.92
14	4	0.02	0.02	0.20	0.77
15	4	0.19	0.92	0.73	0.93
16	5	0.02	0.03	0.04	0.70
17	5	0.28	0.03	0.69	0.93
18	6	0.03	0.03	0.03	0.74
19	6	0.05	0.03	0.40	0.96

Motif Finding



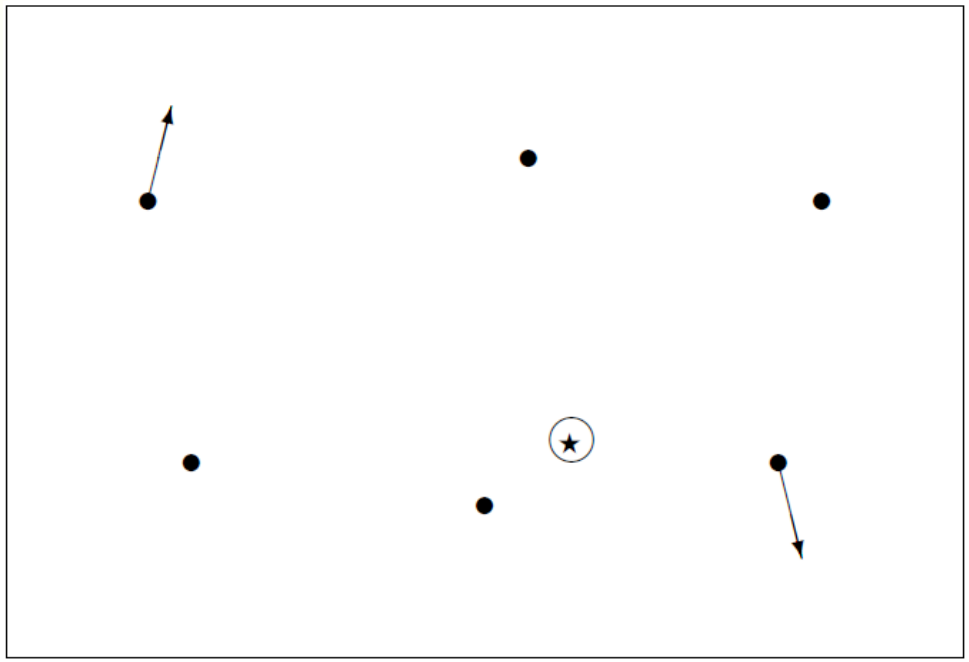
Pattern Branching Algorithm

- let M be an unknown motif of length l, and let A_0 be an occurrence of M in the sample with exactly k substitutions.
- Given A_0, how do we determine M? Since the Hamming distance $d(M, A_0) = k$, we have $M \in D_{=k}(A_0)$, defined as the set of patterns of distance exactly k from A_0.
- We could look at all

$$\binom{l}{k} 3^k$$

- elements of $D_{=k}(A_0)$ and score each pattern as a guess of M. However, as this must be applied to all sample strings A_0 of length l, it would be too slow.

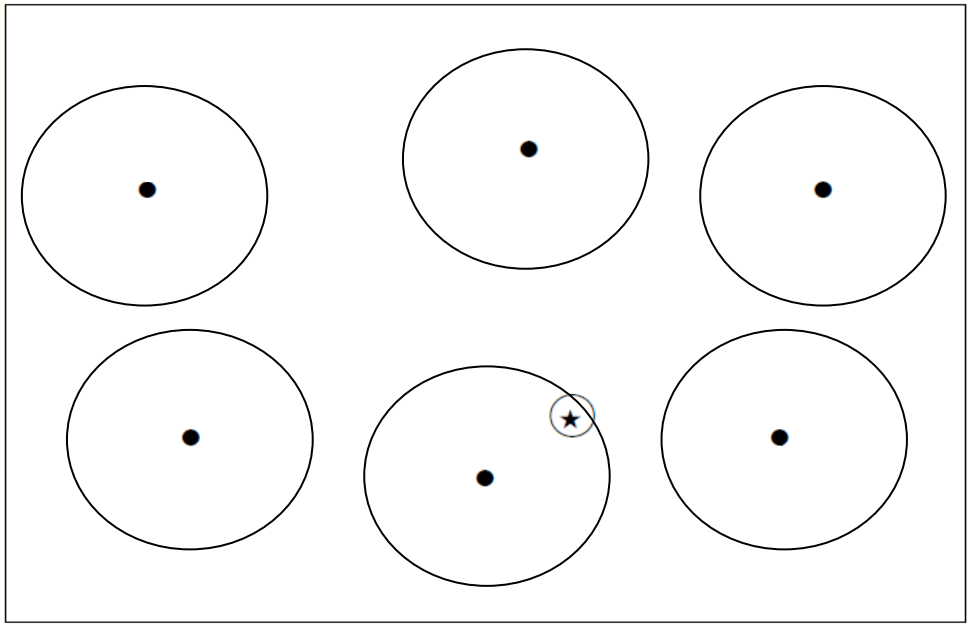
How to Search Motif Space?



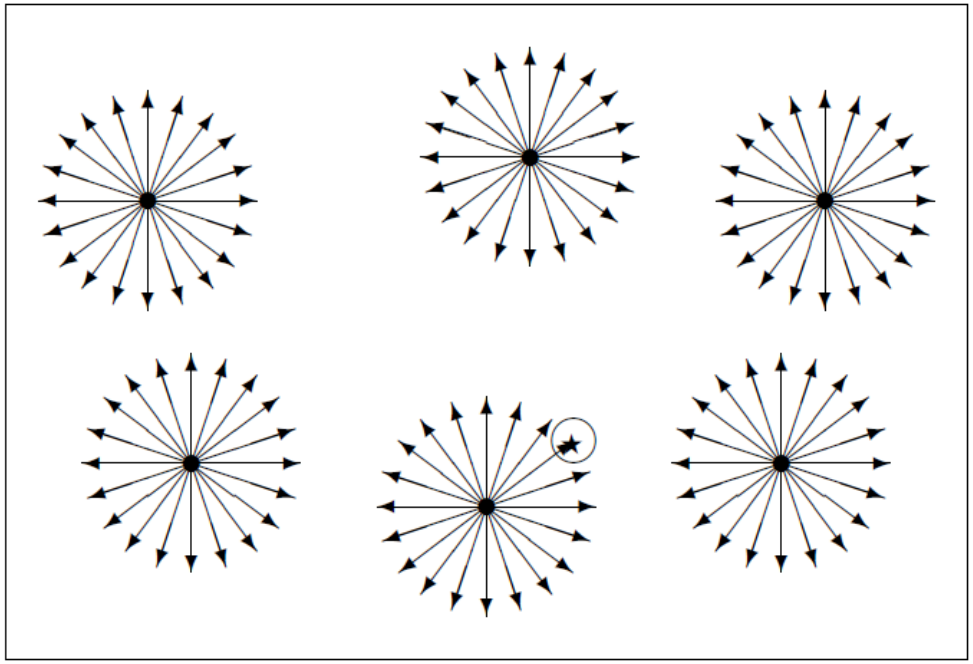
Start from random sample strings (A_0)

Search motif space for the star

Search Small Neighborhoods

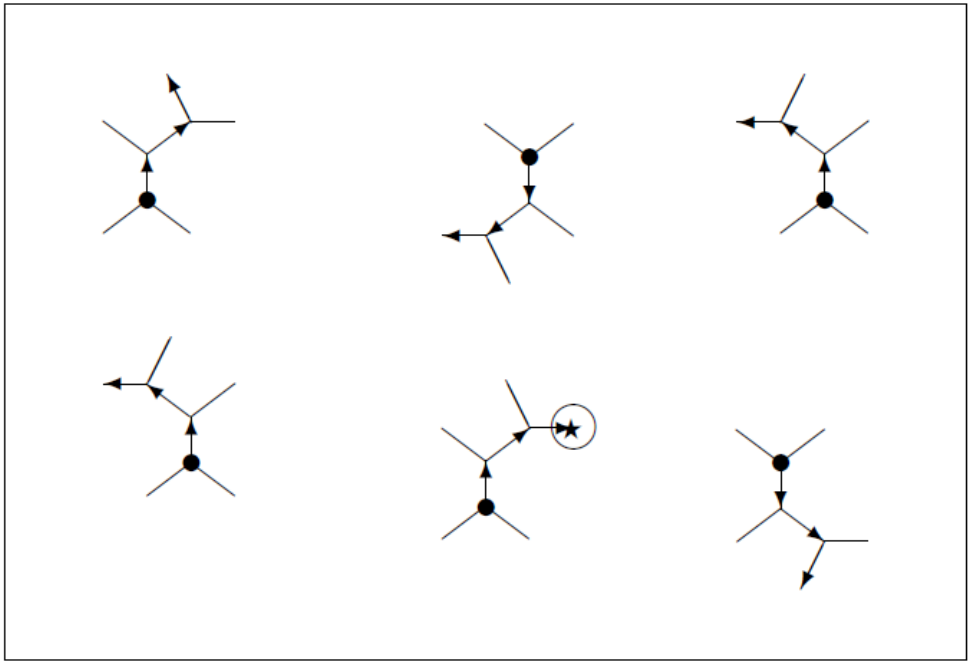


Exhaustive local search



A lot of work, most of it unnecessary

Best Neighbor



Branch from the seed strings

Find best neighbor – highest score

Don't consider branches where the upper bound is not as good as best score so far – in each step move to the "best neighbor" in $D_{=1}(A_i)$

Scoring

- **PatternBranching** uses total distance score:
  - For each sequence DNA_i in the sample $DNA = \{DNA_1, \ldots, DNA_n\}$, let
    $$d(A, DNA_i) = \min\{d(A, P) | P \in DNA_i\}.$$
  - Then the total distance of pattern A from the sample is
    $$d(A, DNA) = \sum_{DNA_i \in DNA} d(A, DNA_i)$$
  - For a pattern A, let $D = Neighbor(A)$ be the set of patterns which differ from A in exactly 1 position.
  - We define $BestNeighbor(A)$ as the pattern $B \in D = Neighbor(A)$ with lowest total distance $d(B, DNA)$.

PatternBranching Algorithm

PatternBranching (DNA, l, k):

$Motif =$ arbitrary motif pattern

For each l-mer A_0 **in** DNA

For $j = 0$ **to** k

If $d(A_j, DNA) < d(Motif, DNA)$

$Motif = A_j$

$A_{j+1} = BestNeighbor(j)$

Output $Motif$

- *More thorough search*: instead of single pattern we can keep r patterns $B \in D_{=1}(A_j)$ with the lowest total distance $d(B, DNA)$
- *Instead of starting from all l-mers in all sequences we can start from l-mers from any single sequence containing the motif*

PatternBranching Performance

- **PatternBranching** is faster than other pattern-based algorithms
- Motif Challenge Problem:
  - sample of $n = 20$ sequences
  - $N = 600$ nucleotides long
  - implanted pattern of length $l = 15$
  - $k = 4$ mutations

Algorithm	Success Rate	Running Time
PROJECTION	about 100%	2 minutes
MITRA	100%	5 minutes
MULTIPROFILER	99.7%	1 minute
PatternBranching	99.7%	3 seconds

Profile Branching

- Profile Branching algorithm is similar to the Pattern Branching Algorithm, however, it searches in the space of motif profiles, instead of motif patterns. The algorithm is obtained from the Pattern Branching Algorithm by making the following changes:
  - convert each sample string A_0 to a profile $X(A_0)$,
  - generalize the scoring method to score profiles,
  - modify the branching method to apply to profiles, and
  - use the top-scoring profile found as a seed for the EM algorithm.
- Details omitted

Profile Branching

- Profile Branching is about 5 times slower than the Pattern Branching algorithm
- The Pattern Branching Algorithm clearly outperforms the Profile Branching Algorithm on Challenge-like problems. However, pattern-based algorithms have difficulty finding motifs with many degenerate positions.

PMS (Planted Motif Search)
Combinatorial approach

- Generate all possible l-mers out of the input sequence DNA_i. Let C_i be the collection of these l-mers.
- Example:

  AAGTCAGGAGT

  $C_i = 3$-mers:

  AAG AGT GTC TCA CAG AGG GGA GAG AGT

All Patterns at Hamming Distance $d = 1$

AAGTCAGGAGT

AAG	**AGT**	**GTC**	**TCA**	**CAG**	**AGG**	**GGA**	**GAG**	**AGT**
CAG	CGT	ATC	ACA	AAG	CGG	AGA	AAG	CGT
GAG	GGT	CTC	CCA	GAG	TGG	CGA	CAG	GGT
TAG	TGT	TTC	GCA	TAG	GGG	TGA	TAG	TGT
ACG	ACT	GAC	TAA	CCG	ACG	GAA	GCG	ACT
AGG	ATT	GCC	TGA	CGG	ATG	GCA	GGG	ATT
ATG	AAT	GGC	TTA	CTG	AAG	GTA	GTG	AAT
AAC	AGA	GTA	TCC	CAA	AGA	GGC	GAA	AGA
AAA	AGC	GTG	TCG	CAC	AGT	GGG	GAC	AGC
AAT	AGG	GTT	TCT	CAT	AGC	GGT	GAT	AGG

Sort the Lists

AAG	AGT	GTC	TCA	CAG	AGG	GGA	GAG	AGT
AAA	AAT	ATC	ACA	AAG	AAG	AGA	AAG	AAT
AAC	ACT	CTC	CCA	CAA	ACG	CGA	CAG	ACT
AAT	AGA	GAC	GCA	CAC	AGA	GAA	GAA	AGA
ACG	AGC	GCC	TAA	CAT	AGC	GCA	GAC	AGC
AGG	AGG	GGC	TCC	CCG	AGT	GGC	GAT	AGG
ATG	ATT	GTA	TCG	CGG	ATG	GGG	GCG	ATT
CAG	CGT	GTG	TCT	CTG	CGG	GGT	GGG	CGT
GAG	GGT	GTT	TGA	GAG	GGG	GTA	GTG	GGT
TAG	TGT	TTC	TTA	TAG	TGG	TGA	TAG	TGT

Eliminate Duplicates

AAG	AGT	GTC	TCA	CAG	AGG	GGA	GAG	AGT
AAA	~~AAT~~	ATC	ACA	AAG	~~AAG~~	~~AGA~~	~~AAG~~	~~AAT~~
AAC	ACT	CTC	CCA	CAA	~~ACG~~	CGA	~~CAG~~	~~ACT~~
AAT	AGA	GAC	GCA	CAC	~~ACA~~	GAA	~~GAA~~	~~ACA~~
ACG	AGC	GCC	TAA	CAT	AGC	~~GCA~~	~~GAC~~	~~AGC~~
AGG	~~AGG~~	GGC	TCC	CCG	AGT	~~GGC~~	GAT	~~AGG~~
ATG	ATT	GTA	TCG	CGG	~~ATG~~	~~GGG~~	GCG	~~ATT~~
CAG	CGT	GTG	TCT	CTG	~~CGG~~	~~GGT~~	~~GGG~~	~~CGT~~
GAG	GGT	GTT	TGA	~~GAG~~	GGG	~~GTA~~	~~GTG~~	~~GGT~~
TAG	TGT	TTC	TTA	~~TAG~~	TGG	~~TGA~~	~~TAG~~	~~TGT~~

Let L denote the obtained list of l-mers

Find Motif Common to All Lists

- Follow this procedure for all sequences
- Find the motif common to all L_i (once duplicates have been eliminated)
- This is the planted motif

PMS Running Time

- It takes time to
  - Generate variants
  - Sort lists
  - Find and eliminate duplicates

$$O\left(n\binom{l}{d}3^d\right)$$

($O(n)$ is the number of different l-mers which are in the first DNA sequence) a check if it is a valid (l,d)-motif or not can be made in $O(tnl)$ time

- Running time of this algorithm:

$$O\left(n\binom{l}{d}3^d\, tnl\right)$$

- For $n = 1000$, $l = 10$, $d = 2$ and $t = 8$

$$1000 \cdot \binom{10}{2} 3^2 \cdot 8 \cdot 1000 \cdot 10 = 3.2 \times 10^9$$