



STRUCTURA BIOLOGICĂ.

§2.3. MOTIFS: DEFINIȚIE & METODE DE IDENTIFICARE

Sorana D. BOLBOACĂ

Despre ...

- **MOTIFs:**
 - Definiție
 - Logo
 - Metode de indentificare a secvențelor motifs

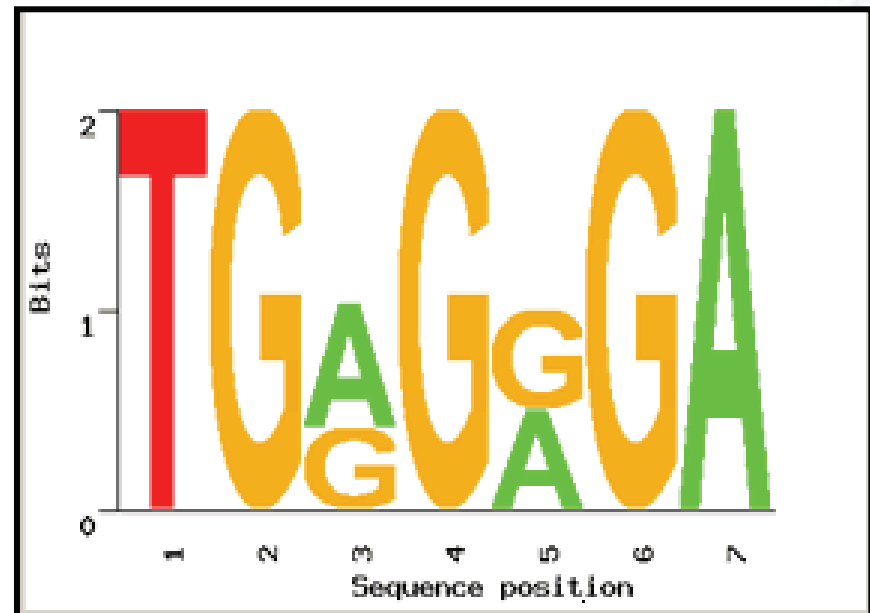
DEFINIȚIE

- Secvență de nucleotide sau amino acizi care apare în mod repetat și care
 - Au semnificație biologică
 - Sunt în relație cu o semnificație biologică
 - Când apare la nivelul exonului unei gene poate codifica motif-ul structural al proteinei
 - Nu se cunosc secvențele “motif”
 - Nu știm unde sunt acestea localizate relativ la punctul de start al genei
 - Motif-ul poate prezenta modificări mici de la o genă la alta
 - Există secvențe motif aleatorii – cum se pot diferenția?

LOGO-ul SECVENȚELOR MOTIF

- LOGO – reprezentarea regiunilor constante și variabile a unui motif
 - Secvențele motif pot suferi mutații

TGGGGGA
TGAGAGA
TGGGGGA
TGAGAGA
TGAGGGA



IDENTIFICAREA SECVENȚELOR MOTIF

- Exprimarea genelor se face prin intermediul proteinelor reglatorii
- Proteinele reglatorii (TF):
 - Acționează asupra regiunii de reglare a genei prin atacarea sau blocarea unei polimeraze ARN
 - Se leagă de o secvență scurtă de ADN denumită motif (TFBS)
- Identificarea aceleași secvențe motif în regiunile de reglare ale mai multor gene sugerează o relație la regiunilor de reglare ale acestor gene

IDENTIFICAREA SECVENȚELOR MOTIF

- Nu știm care sunt secvențele motif
- Nu știm unde este localizată secvența motif relativ la punctul de start
- Secvența motif poate să difere un pic de la o genă la alta
- Cum putem identifica dacă nu este un motif aleator / random?

IDENTIFICAREA SECVENȚELOR MOTIF

- Problemă similară cu: *Gold Bug* story - Edgar Allan Poe (1809 – 1849)

- Mesaj secret:

53++!305))6* ;4826)4+.)4+) ;806* ;48!8`60))85;]8*
: +*8!83(88)5*! ;

46(;88*96*? ;8) *+(;485) ;5*!2: *+(;4956*2(5*-
4)8`8* ; 4069285) ;)6

!8)4++ ;1(+9 ;48081 ;8:8+1 ;48!85 ;4)485!528806*81(
+9 ;48 ;(88 ;4(+?3

4 ;48)4+ ;161 ;:188 ;+? ;

- Descifrați mesajul!

IDENTIFICAREA SECVENȚELOR MOTIF

- Problemă similară cu: *Gold Bug* story - Edgar Allan Poe (1809 – 1849)
- Informații suplimentare:
 - Mesajul este în Engleză
 - Fiecare simbol corespunde unei litere din alfabetul Englezesc
 - Nu au fost codata semnele de punctuație

IDENTIFICAREA SECVENȚELOR MOTIF

9

- Abordarea naivă în rezolvarea problemei:
 - Identifică frecvența absolută a fiecărui simbol din mesajul codat
 - Identifică frecvența literelor din alfabetul Englez într-un text
 - Compară frecvențele și încearcă să corelezi simbolurile cu litere din alfabet.

simbol	8	;	4)	+	*	5	6	(!	1	0	2	9	3	:	?	`	-		.	
f	34	25	19	16	15	14	12	11	9	8	7	6	5	5	4	4	3	2	1	1	1

e t a o i n s r h l d c u m f p g w y b v k x j q z



Mai frecvent

mai puțin frecvent

IDENTIFICAREA SECVENȚELOR MOTIF

10

- Prin înlocuirea simbolurilor cu litere obținem:

```
sfiilfcsoorntaeuroaikoaiotecrntaeleyrcooestvenpinelefheeosn  
lt  
arhteenmrnwteonihtaesotsnlupnihtamsrnuhsnbaoyentacrmuesoto  
rl  
eoaiitdhimtaecedtepeidtaelestaoaeslsueecrnedhimtaetheetahiw  
fa  
taeoaitdrdtpdeetiwt
```

- Dar ... ?

IDENTIFICAREA SECVENȚELOR MOTIF

11

- O abordare mai bună
 - Examinarea frecvenței combinațiilor de simboluri
 - “The” este cea mai frecventă combinație de 3 simboluri în limba Engleză
 - “:48” este cea mai frecventă combinație de 3 simboluri în textul criptat
- ...

IDENTIFICAREA SECVENȚELOR MOTIF

12

- Înlocuim “;48” cu “The”

```
53++!305) ) 6*the26) h+. ) h+) te06*the!e`60) ) e5t]e*:+*e!e3 (ee) 5*  
!t  
h6 (tee*96*?te) *(the5) t5*!2: *(th956*2 (5*h) e`e*th0692e5) t) 6  
!e  
) h++t1 (+9the0e1te:e+1the!e5th) he5!52ee06*e1 (+9thet (eeth (+?3  
ht  
he) h+t161t:leet+?t
```

- “thet(ee” cel mai probabil este “the tree” → “(“ = “r”
- “th(+?3h” devine “thr+?3h”
 - Putem ghici care este valoarea lui “+” și a lui “?”?

IDENTIFICAREA SECVENȚELOR MOTIF

13

- Odată identificate toate corespondențele mesajul final este:

```
AGOODGLASSINTHEBISHOPSHOSTELINTHEDEVILSSEATWENYONEDEGRE  
ESANDTHIRTEENMINUTESNORTHEASTANDBYNORTHMAINBRANCHSEVENT  
HLIMBEASTSIDESHOOTFROMTHELEFTEYEOFTHEDEATHSHEADABEELINE  
FROMTHETREETHROUGHTHESHOTIFYFEETOUT
```

IDENTIFICAREA SECVENȚELOR MOTIF

14

- Semnele de punctuație ne permit însă citirea textului:

A GOOD GLASS IN THE BISHOP'S HOSTEL IN THE DEVIL'S SEA,
TWENY ONE DEGREES AND THIRTEEN MINUTES NORTHEAST AND BY
NORTH,
MAIN BRANCH SEVENTH LIMB, EAST SIDE, SHOOT FROM THE LEFT
EYE OF
THE DEATH'S HEAD A BEE LINE FROM THE TREE THROUGH THE SHOT,
FIFTY FEET OUT.

IDENTIFICAREA SECVENȚELOR MOTIF

15

- Cunoștințe necesare pentru rezolvarea textului codat:
 - Frecvența relativă a fiecărei litere din alfabet
 - Frecvența relativă a combinațiilor de 2 sau mai multe litere în limba Engleză
 - Cunoașterea tuturor cuvintelor din vocabularul limbii Engleze este de dorită pentru a fi capabili să facem inferențe acurate

IDENTIFICAREA SECVENȚELOR MOTIF

16

Similarități

- Nucleotidele din motif codifică un mesaj în limbaj genetic
 - Simbolurile din textul codificat codifică un mesaj în limba Engleză
- Pentru a rezolva problema, analizăm frecvențe ale grupărilor de simboluri în ADN / mesajul codificat
- Cunoașterea regiunilor motifs cu rol în reglare face problema de identificare a regiunilor motifs ușoară.
 - Cunoașterea cuvintelor din dicționarul limbii Engleze ne ajută să găsim soluția în cazul mesajului codificat.

IDENTIFICAREA SECVENȚELOR MOTIF

17

Diferențe

- Căutarea secvențelor motifs
 - Pentru a găsi soluția analizăm frecvența patern-urilor (secvențe de mai multe nucleotide) în secvențele de nucleotide
- Problema textului codificat:
 - Pentru a găsi soluția analizăm frecvențele de pattern-uri (grupuri de simboluri) în textul scris în limba Engleză

Diferențe

- Căutarea secvențelor motifs
 - Cunoașterea secvențelor motifs identificare reduce complexitatea problemei
- Problema textului codificat:
 - Cunoașterea cuvintelor din dicționar este de dorit pentru rezolvarea problemei

IDENTIFICAREA SECVENȚELOR MOTIF

18

- **Problema identificării secvențelor motif este mai grea decât problema textului codificat**
 - Nu avem dicționarul complet al motifs-urilor
 - Limbajul genetic nu are un dicționar standard
 - Doar o fracțiune mică de secvențe nucleotidice codifică motifs-urile; volumul de date este însă enorm

IDENTIFICAREA SECVENȚELOR MOTIF: PROBMELA

19

- Fie o secvență randomizată de ADN:
cctgatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcgt
ttccaacat
agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaacgctcag
aaccagaagtgc
aacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataag
acgaaaattt
agcctccgatgtaagtcatacgtgtaactattacctgccaccctattacatcttacgt
acgtataca
ctgttatacaacgcgctcatggcggggatgcgttttggtcgtcgtacgctcgatcgtt
aacgtacgtc
- Găsiți pattern-ul (motifs-ul) implementat la nivelul
secvențele individuale

IDENTIFICAREA SECVENȚELOR MOTIF: PROBMELA

20

- Informații suplimentare:
 - Secvența motif este de lungime 8
 - Pattern-ul nu este exact la fel în fiecare apariție datorită mutației aleatorii punctiformă

IDENTIFICAREA SECVENȚELOR MOTIF: PROBMELA

21

- **Pattern-ul identificat:**

cctgatagacgctatctggctatcc**acgtacgt**aggtcctctgtgcaatctatgcgtttcca
accat

agtactggtgtacatttgata**acgtacgt**acaccggcaacctgaaacaacgctcagaacca
gaagtgc

aa**acgtacgt**gcaccctctttcttcgtggctctggccaacgagggctgatgtataagacga
aaatfff

agcctccgatgtaagtcatagctgtaactattacctgccaccctattacatctt**acgtacgta**
taca

ctgttatacaacgcgtcatggcgggggtatgcgttttggtcgtcgtacgctcgatcgtta**acgt**
acgtc

IDENTIFICAREA SECVENȚELOR MOTIF: PROBMELA

22

- **Pattern-ul cu 2 mutații punctiforme:**

cctgatagacgctatctggctatcca**GgtacT**taggtcctctgtgpgaattctatgcgttcc
aacat

agtactggtgtacatttgat**CcA**ta**cg**tacaccggcaacctgaaacaacgctcagaacc
agaagtgc

aa**acgtT**A**gt**gcaccctcttcttcgtggctctggccaacgagggtgatgtataagacg
aaaatttt

agcctccgatgtaagtcatagctgtaactattacctgccaccctattacatctt**acgtCcA**t
ataca

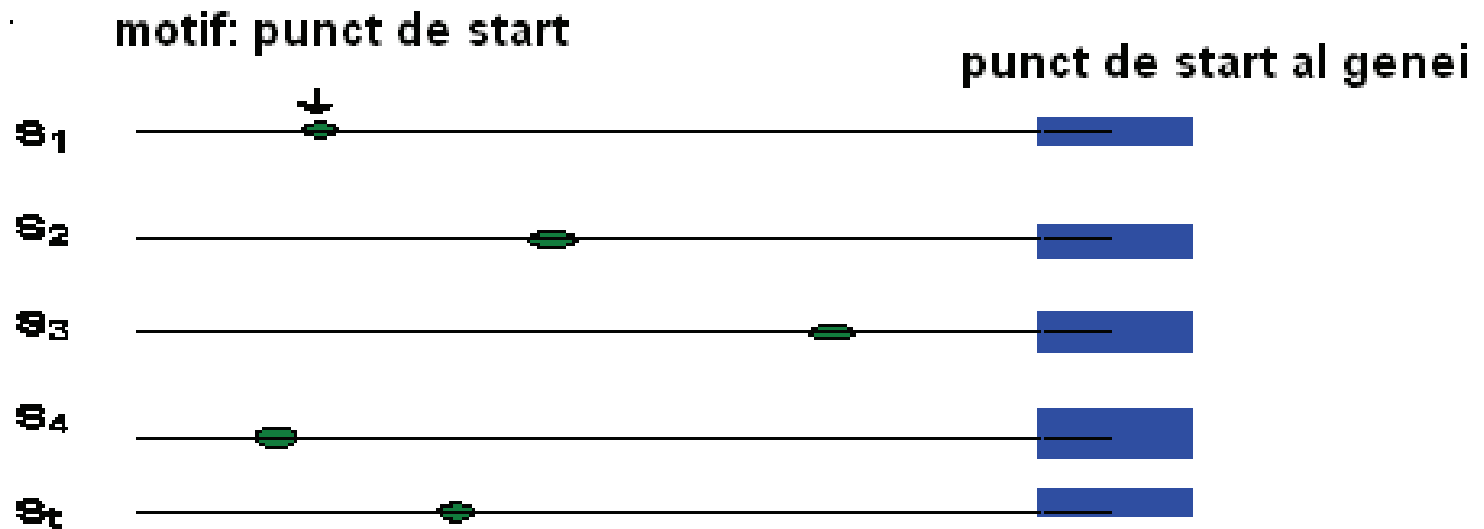
ctgttatacaacgcgtcatggcggggatgctgttttggtcgtcgtacgctcgatcgta**Ccgt**
acgGc

- Secvențele pot fi identificate în cazul existenței mutațiilor?

IDENTIFICAREA SECVENȚELOR MOTIF: PROBMELA

23

- Identificarea pattern-urilor se poate realiza mai ușor dacă cunoaștem punctul de start al secvenței motif
 - $s = (s_1, s_2, s_3, \dots, s_t)$



MOTIFs: PROFILURI ȘI CONSENS

		Secvența MOTIF							
Aliniere		a	G	g	t	a	c	T	t
		C	c	A	t	a	c	g	t
		a	c	g	t	T	A	g	t
		a	c	g	t	C	c	A	t
		C	c	g	t	a	c	g	G
Profil	A	3	0	1	0	3	1	1	0
	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
Consens		A	C	G	T	A	C	G	T

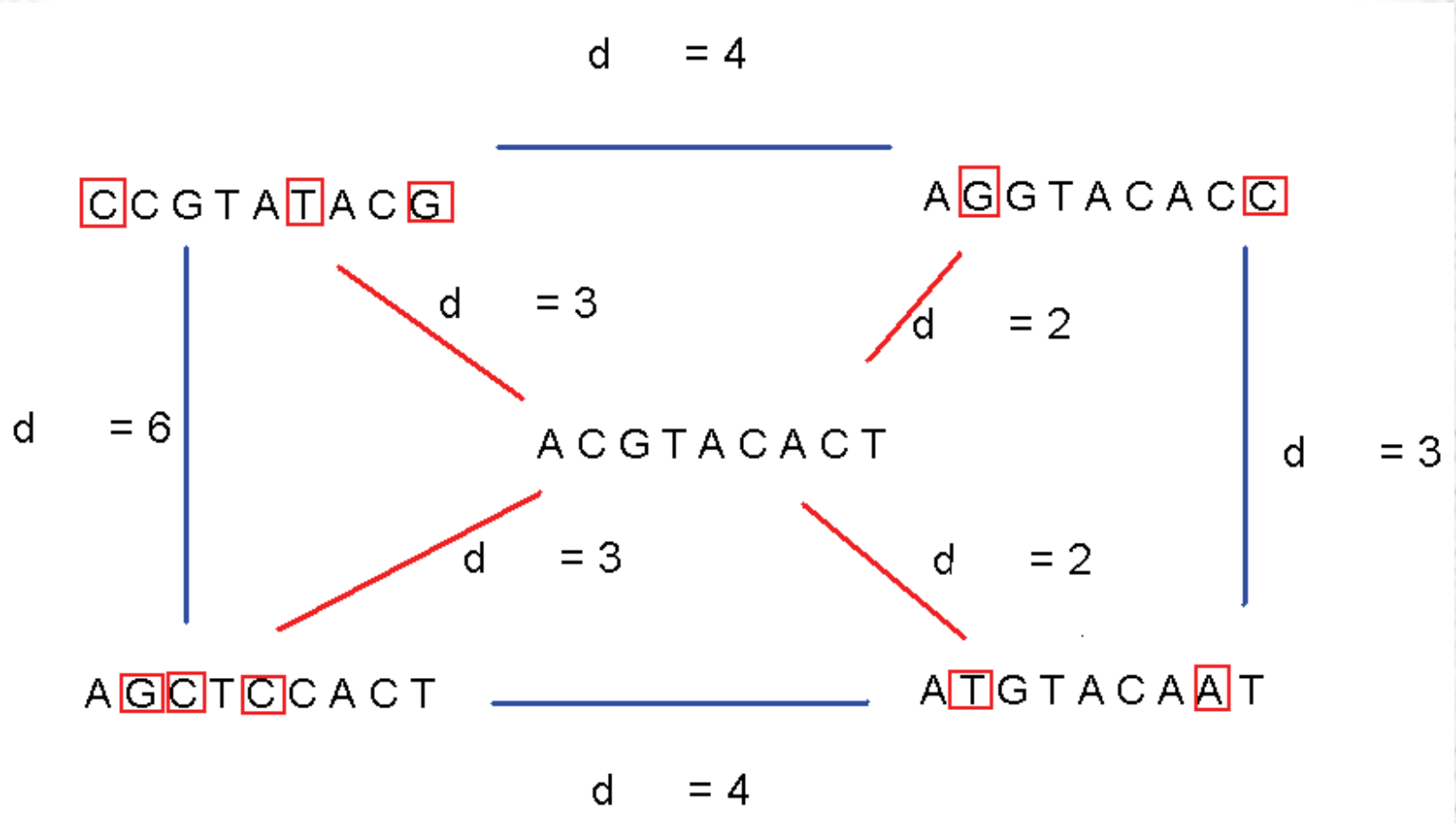
MOTIFs: CONSENS

25

- Un strămoș al secvenței motif
 - Plecând de la acest strămoș emerg pattern-urile care prezintă mutații
- Distanța dintre secvența motif reală și secvența consens este în general mai mică decât aceea pentru două secvențe motif reale

MOTIFs: CONSENS

26



MOTIFs: CONSENS

27

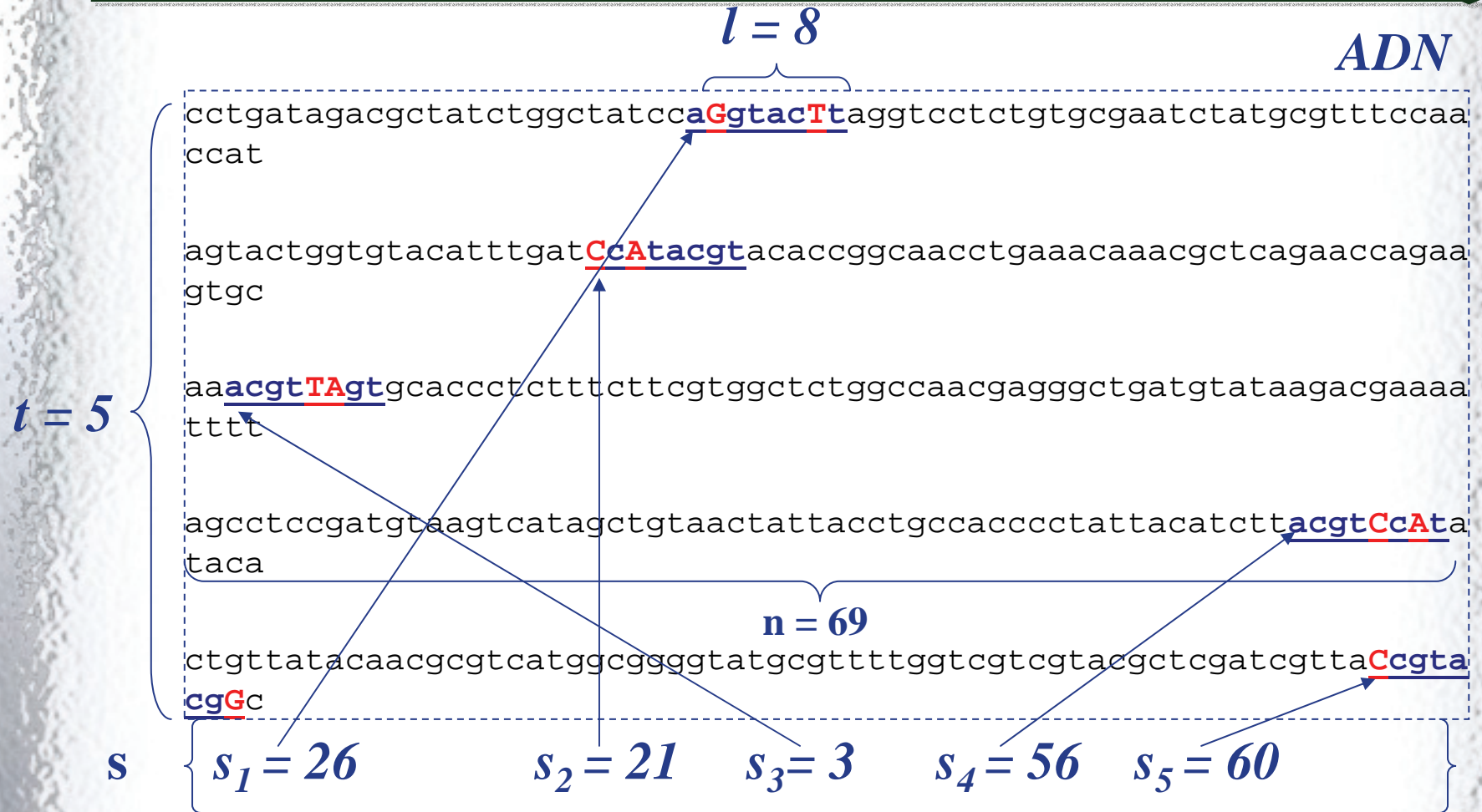
- Avem secvența motifs rezultată din consens
 - Dar ... cât de bun este acest consens?
- Prin introducerea unei funcții de tip scor putem compara diferitele pattern-uri identificate/posibile și putem alege cea mai bună secvență motif

MOTIFs: TERMENI

28

- t = numărul de secvențe ADN
- n = lungimea fiecărei secvențe ADN
- **AND** = eșantionul de secvențe ADN ($t \times n$)
- l = lungimea secvenței motif (l -mer)
- s_i = punctul de start al l -mer în secvența i
- $s = (s_1, s_2, \dots, s_t)$ = mulțimea pozițiilor start al pattern-urilor

MOTIFs: PARAMETRII



MOTIFs: CALCULAREA SCORULUI

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

- $s = 3+4+4+5+3+4+3+4$
- $s = 30$

		Secvența MOTIF							
Aliniere		a	G	g	t	a	c	T	t
		C	c	A	t	a	c	g	t
		a	c	g	t	T	A	g	t
		a	c	g	t	C	c	A	t
		C	c	g	t	a	c	g	G
Profil	A	3	0	1	0	3	1	1	0
	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
Consens		A	C	G	T	A	C	G	T

IDENTIFICAREA SECVENȚELOR MOTIF: PROBLEMA

31

- Dacă cunoaștem punctele de start ale secvențelor motifs identificarea consensului este ușoară chiar și în cazul existenței mutațiilor în interiorul secvenței motifs
- Dar ... pozițiile de start nu se cunosc de obicei. În aceste condiții cum putem identifica cea mai bună matrice de profil?

IDENTIFICAREA SECVENȚELOR MOTIF: PROBLEMA

32

- Formularea problemei:
 - Scop: având un scop de secvențe ADN, identificați pentru fiecare secvență setul *l-mer* care sa maximizeze scorul consensului
 - Date de intrare:
 - matricea $t \times n$ a secvențelor ADN
 - l = lungimea pattern-ului de identificat
 - Date de ieșire: un șir de t poziții start $s = (s_1, s_2, \dots, s_t)$ care maximizează $scor(s, ADN)$

IDENTIFICAREA SECVENȚELOR MOTIF: SOLUȚIA FORȚELOR BRUTE

33

- alculăm scorurile pentru toate combinațiile posibile de puncte de start s
- el mai bun scor identifică profilul cel mai bun și respectiv consensul de pattern în secvențele de ADN
- biectivul global este de a maximiza $scor(s, ADN)$ prin varierea pozițiilor de start s_i (unde $s_i = [1, \dots, n-l+1]$ și $i = [1, \dots, t]$)

IDENTIFICAREA SECVENȚELOR MOTIF: SOLUȚIA FORȚELOR BRUTE

34

- Timpul necesar identificării soluției celei mai bune:
 - Prin varierea pozițiilor $(n-l+1)$ pentru fiecare secvență t vom investiga $(n-l+1)^t$ seturi de poziții start
 - Pentru fiecare set de poziții start, funcția scor va face l operații, astfel încât complexitatea este $l(n-l+1)^t = O(l n^t)$
 - Pentru $t = 8$, $n = 1000$, $l = 10$ trebuie să realizăm $\sim 10^{20}$ operații – va dura foarte mult timp

DISTANȚE TOTALE: EXEMPLU

35

- Fie secvența $v = \text{“acgtacgt”}$ și x secvența de interes “acgtacgt”

$$d_H(v, x) = 0$$

acgtacgt

cctgatagacgctatctggctatccacgtacgttaggtcctctgtgcaatctatgcgtttccaacat

$$d_H(v, x) = 0$$

acgtacgt

agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc

acgtacgt

aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt

$$d_H(v, x) = 0$$

$$d_H(v, x) = 0$$

acgtacgt

agcctccgatgtaagtcatagctgtaactattacctgccaccctattacatcttacgtacgtataca

acgtacgt

ctgttatacaacgcgctcatggcggggtatgcgttttggctcgtcgtacgctcgatcgttaacgtacgtc

$$d_H(v, x) = 0$$

- ***Distanța Totală*** $(v, DNA) = 0$

DISTANȚE TOTALE: EXEMPLU

- Fie secvența $v = \text{“acgtacgt”}$ și x secvența de interes “acgtacgt”

$d_H(v, x) = 1$ → acgtacgt

cctgatagacgctatctggctatccacgtacAtaggctctctgtgcaatctatgcgtttccaacat

$d_H(v, x) = 0$ → acgtacgt

agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc

acgtacgt

aaaAgtCcggtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt

$d_H(v, x) = 2$ → acgtacgt

agcctccgatgtaagtcatagctgtaactattacctgccaccctattacatcttacgtacgtataca

$d_H(v, x) = 0$ → acgtacgt

ctgttatacaacgcgctcatggcggggtatgcgttttggctcgtcgtacgctcgatcgttaacgtaGgtc

acgtacgt

$d_H(v, x) = 1$ → acgtacgt

- Distanța Totală*(v, DNA) = 1+0+2+0+1 = 4**

DISTANȚE TOTALE: DEFINIȚIE

37

Pentru fiecare i secvență de ADN se calculează $d_H(v, x)$, unde x este l -mer-ul cu punctul de start s_i ($1 \leq s_i \leq n - l + 1$)

■

Se identifică valoarea minimă $d_H(v, x)$ printre l -mer în secvența I

■

Distanța Totală(v, ADN) este suma distanțelor minime Hamming pentru fiecare secvență i de AND

- ***Distanța Totală***(v, ADN) = $\min_s d_H(v, s)$
 - Unde s este setul de puncte start s_1, s_2, \dots, s_t

PROBLEMA STRING-ULUI MEDIAN: FORMULARE

38

- Scop: Pentru un set dat de secvențe ADN, identificați string-ul median
- Date de intrare:
 - Matricea $t \times n$ de ADN
 - Lungimea pattern-ului identificat l
- Date de ieșire:
 - Un string v de l nucleotide care să minimizeze $DistanțaTotală(v, ADN)$ din toate stringurile de lungime l

MOTIFS vs PROBLEMA STRINGURILOR MEDIANE

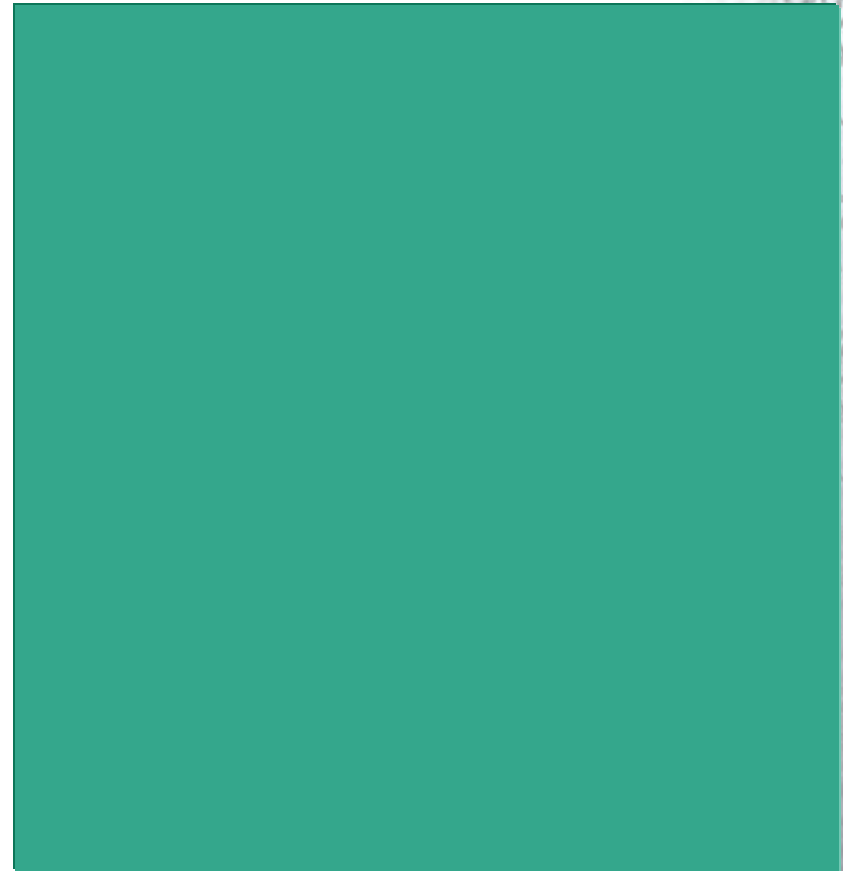
39

- Identificarea secvenței MOTIF
 - Problemă de maximizare
- Stringul median:
 - Problemă de minimizare
- Din punct de vedere computațional sunt echivalente
- Minimizarea *Distanței Totale* este echivalentă cu identificarea *Scorului* maxim

MOTIFS vs PROBLEMA STRINGURILOR MEDIANE

40

- Pentru fiecare coloană i
 $Scor_i + DistanțaTotală_i = t$
- Deoarece avem l coloane
 $Scor + DistanțaTotală = l * t$
- Prin rearanjare:
 $Scor = l * t - DistanțaTotală$
 $l * t$ este constanta de
minimizare a părții drepte
ceea ce e echivalent cu
maximizarea părții drepte



IDENTIFICAREA SECVENȚELOR MOTIF:

PROBLEMA ȘIRULUI MEDIAN

41

- Fie un set de t secvențe *ADN*
- Să se identifice pattern-ul (secvența motif) cu număr minim de mutații care apare în toate secvențele t
- Distanța Hamming:
 - $d_H(\mathbf{v}, \mathbf{w})$ este numărul de perechi de nucleotide care nu se potrivesc când \mathbf{v} și \mathbf{w} sunt aliniate
 - Exemplu: $d_H(\text{AAAAAA}, \text{ACAAAC}) = 2$

IDENTIFICAREA SECVENȚELOR MOTIF VS. PROBLEMA ȘIRULUI MEDIAN

42

- De ce am reformula problema de identificare a secvențelor motifs?
 - Este necesară examinarea tuturor combinațiilor s .
 - Rezultă un număr total de $(n - l + 1)^t$ combinații
 - Problema șirului median necesită examinarea a 4^l combinații – număr relativ mic în comparație cu precedentul

STRUCTURAREA CĂUTĂRII

43

- Pentru problema stringului median trebuie considerate toate 4^l posibilitățile *l-mer-urii*

l
a... aa

a... ac

a... ag

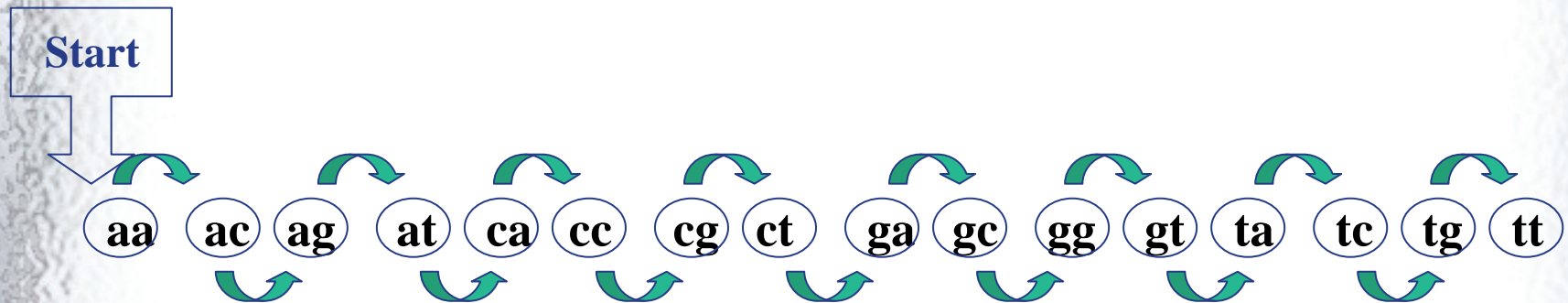
a... at

t... tt

- Cum putem organiza această căutare?

STRUCTURAREA CĂUTĂRII

- Fie $l = 2$



- Este necesară investigarea tuturor predecesorilor unei secvențe pentru a investiga secvența de interes

STRUCTURAREA CĂUTĂRII

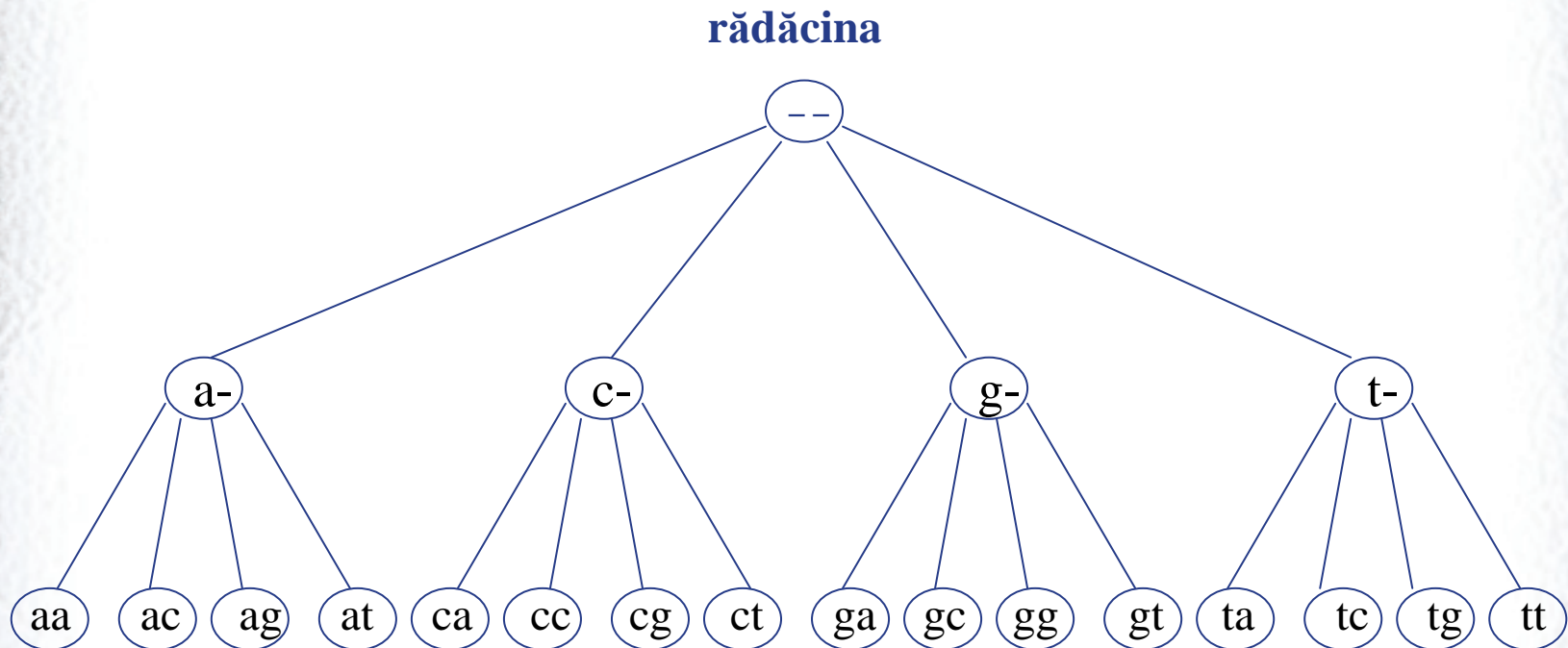
45

- Metoda listei de legături nu este cea mai eficientă metodă de structurare a datelor pentru identificarea secvențelor motif
- Să încercăm gruparea secvențelor după prefixe

aa ac ag at ca cc cg ct ga gc gg gt ta tc tg tt

STRUCTURAREA CĂUTĂRII

46



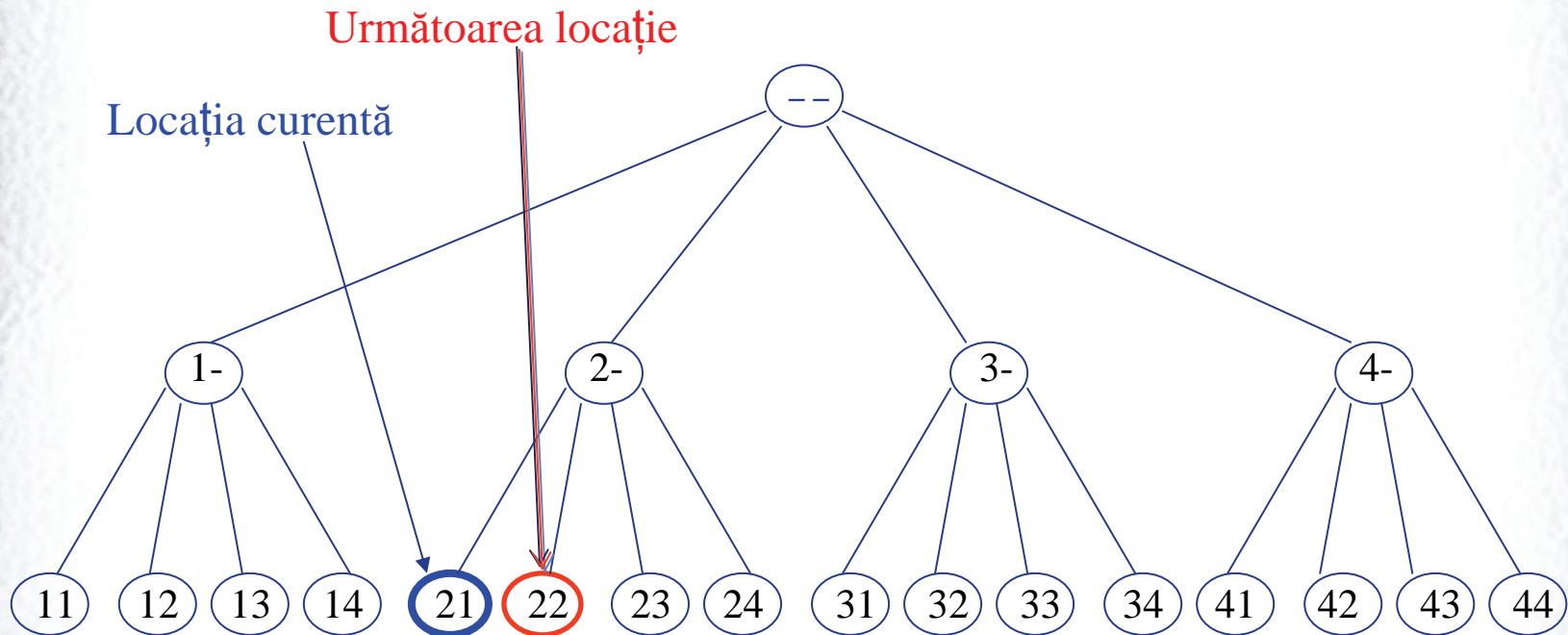
ANALIZA CĂUTĂRII DE TIP ARBORE

47

- Caracteristicile căutării de tip arbore:
 - Secvențele sunt conținute la nivelul frunzelor
 - Părintele unui nod este prefixul copilului
- Care este modalitatea de mișcare în arbore?
 - Următoarea frunză
 - Vizitarea tuturor frunzelor
 - Vizitarea următorului nod
 - Trecerea de la copil la nod (părinte)

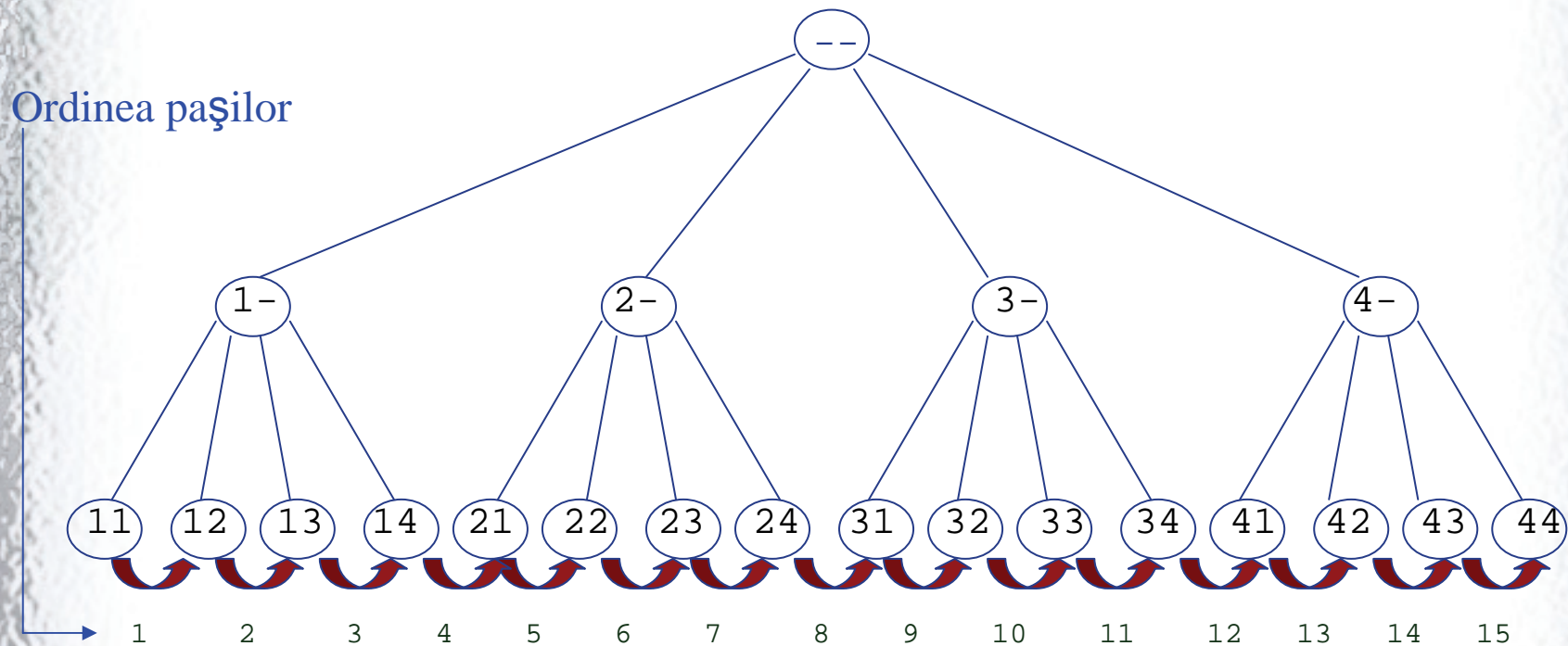
URMĂTOAREA FRUNZĂ: EXEMPLU

- Mișcarea la următoarea frunză:



VIZITAREA TUTUROR FRUNZELOR

- Mișcarea de la o frunză la cealaltă :



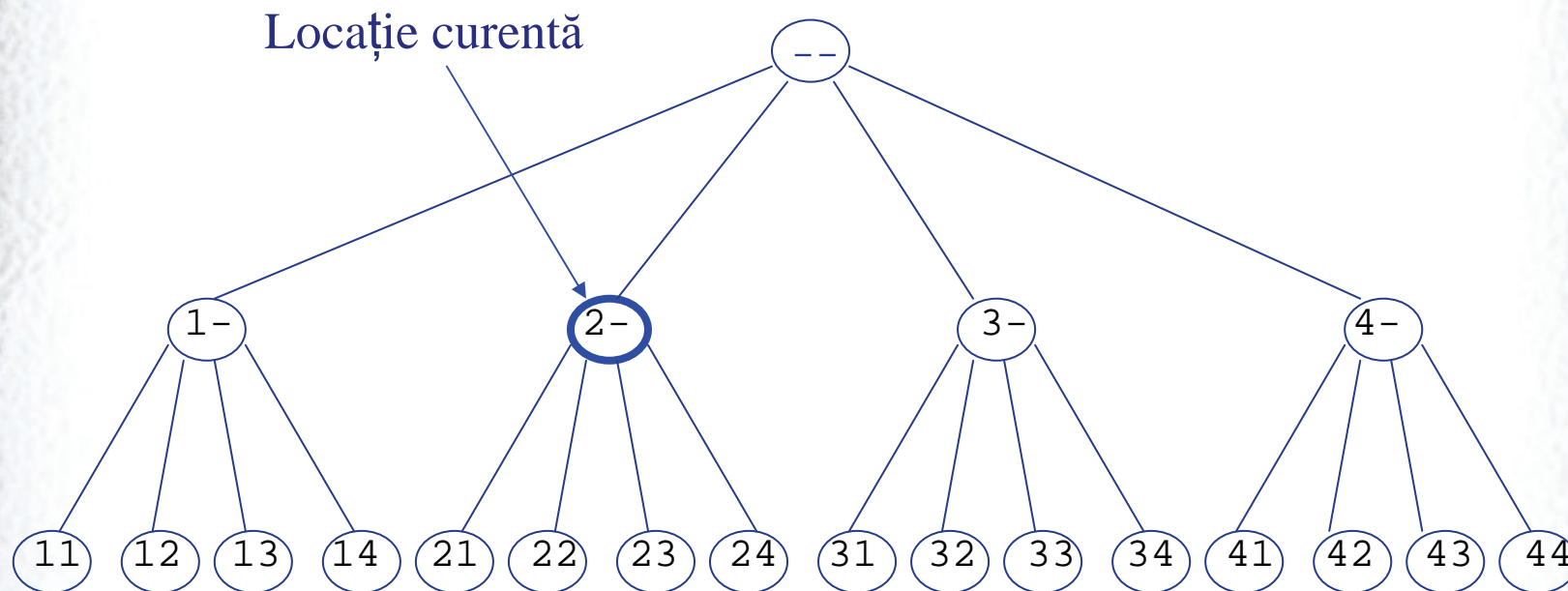
CĂUTAREA ÎN ADÂNCIME

50

- Putem căuta la nivelul frunzelor
- Dar, putem căuta toate vârfurile unui arbore?
 - Putem, prin căutarea primară în adâncime

EXEMPLU

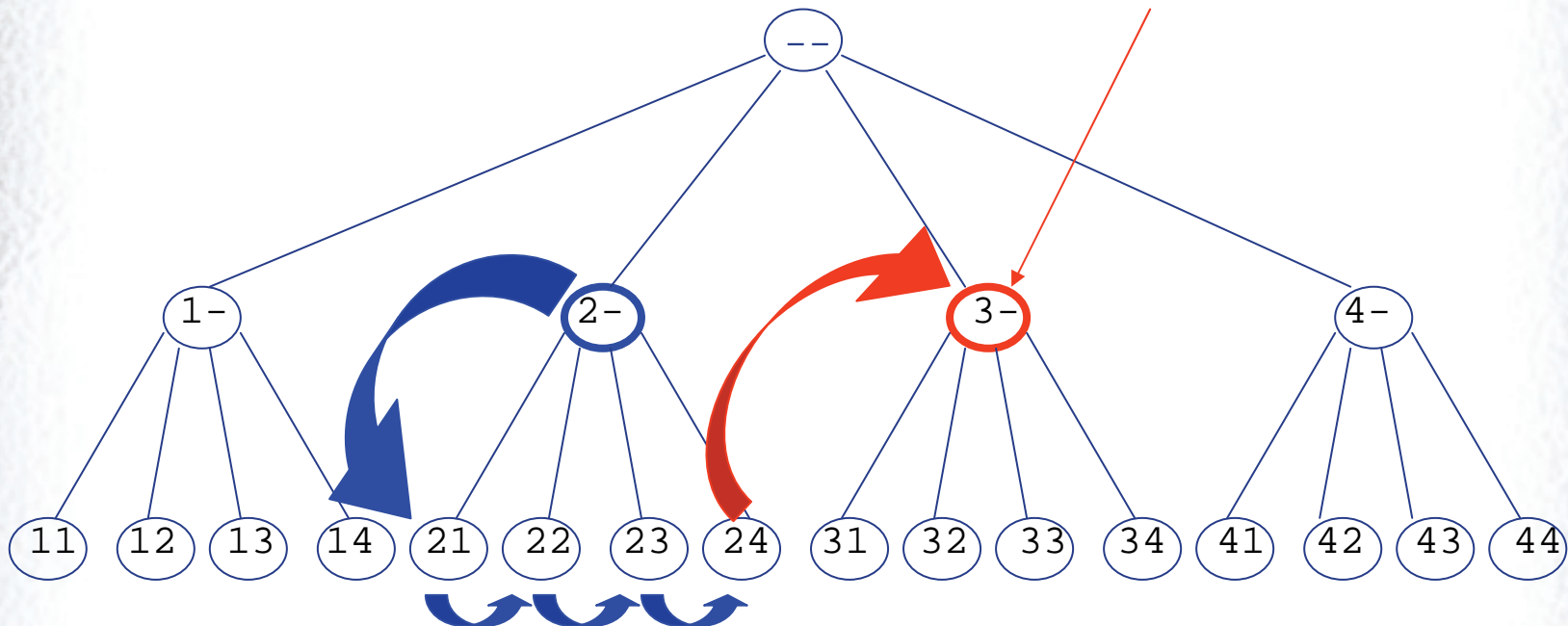
- Mișcarea către următorul vârf:



EXEMPLU

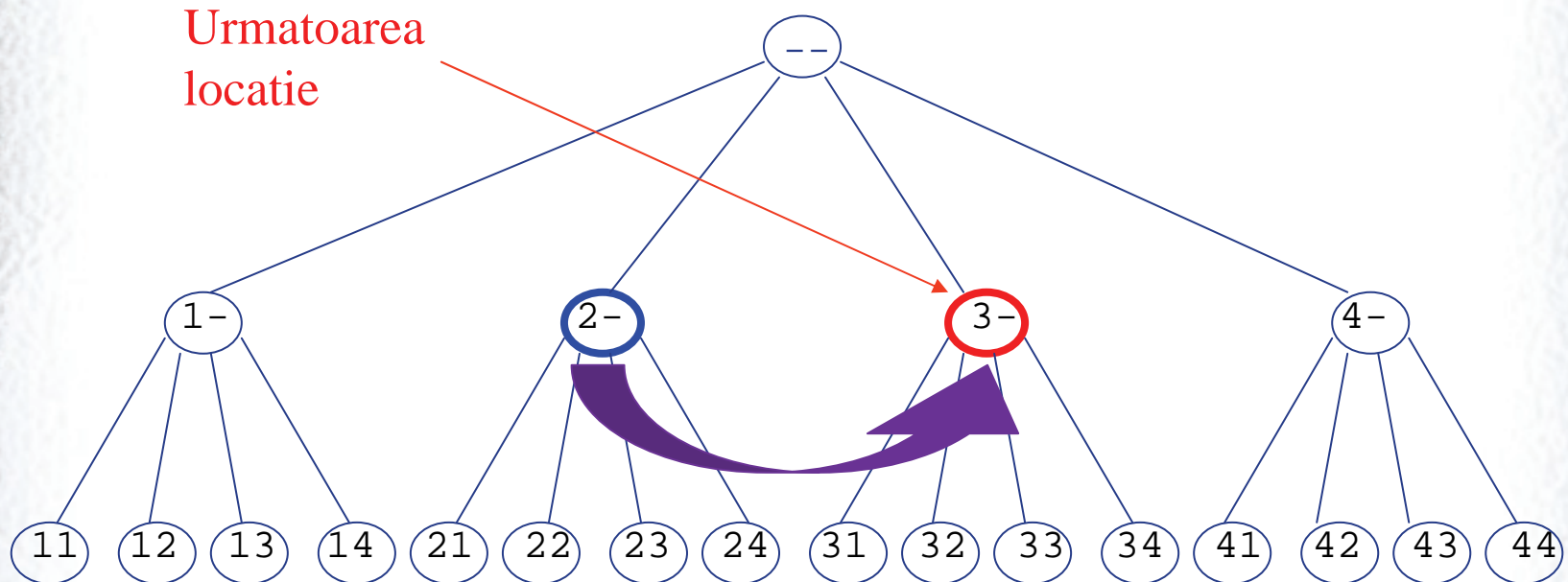
- Mișcarea către următorul vârf:

Locația după mutarea la următoarele 5 vârfuri



EXEMPLU

- Ocolirea descendenților lui “2-”:



MOTIFs: PROGRAME

54

- **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)