STRUCTURA BIOLOGICĂ.
§2.3. MOTIFs

Sorana D. BOLBOACĂ

Despre …

- **MOTIFs:**
  - Definiţie
  - Logo
  - Metode de indentificare a secvenţelor motifs

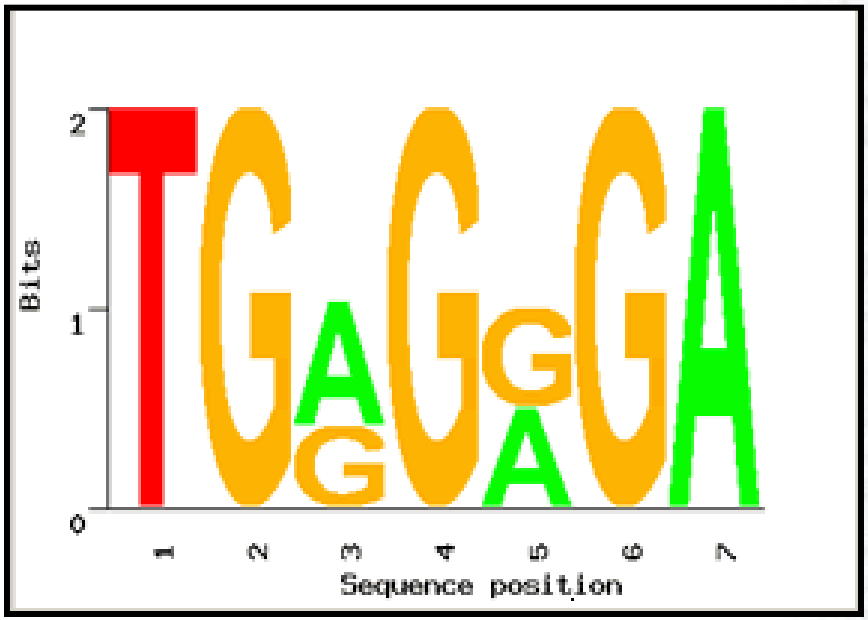
DEFINIŢIE

- Secvenţă de nucleotide sau amino acizi care apare în mod repetat şi care
  - Au semnificaţie biologică
  - Sunt în relaţie cu o semnificaţie biologică
    - Când apare la nivelul exonului unei gene poate codifica motif-ul structural al proteinei
  - Nu se cunosc secvenţele “motif”
  - Nu ştim unde sunt acestea localizate relativ la punctul de start al genei
    - Motif-ul poate prezenta modificări mici de la o genă la alta
  - Există secvenţe motif aleatorii – cum se pot diferenţia?

LOGO-ul Secvenţelor Motif

- LOGO – reprezentarea regiunilor constante şi variabile a unui motif
  - Secvenţele motif pot suferi mutaţii

TGGGGGA
TGAGAGA
TGGGGGA
TGAGAGA
TGAGGGA



IDENTIFICAREA SECVENŢELOR MOTIF

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

IDENTIFICAREA SECVENŢELOR MOTIF

- Nu ştim care sunt secvenţele motif
- Nu ştim unde este localizată secvenţa motif relativ la punctul de start
- Secvenţa motif poate să difere un pic de la o genă la alta
- Cum putem identifica dacă nu este un motif aleator / random?

IDENTIFICAREA SECVENŢELOR 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;+?;
```

- Descifraţi mesajul!

IDENTIFICAREA SECVENȚELOR MOTIF

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

IDENTIFICAREA SECVENŢELOR MOTIF

- Abordarea naivă în rezolvarea problemei:
  - Identifică frecvenţa absolută a fiecărui simbol din mesajul codat
  - Identifică frecvenţa literelor din alfabetul Englez într-un text
  - Compară frecvenţele şi 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 puţin frecvent

IDENTIFICAREA SECVENŢELOR MOTIF

- Prin înlocuirea simbolurilor cu litere obţinem:

```
sfiilfcsoorntaeuroaikoaiotecrntaeleyrcooestvenpinelefheeosn
lt
arhteenmrnwteonihtaesotsnlupnihtamsrnuhsnbaoeyentacrmuesoto
rl
eoaiitdhimtaecedtepeidtaelestaoaeslsueecrnedhimtaetheetahiw
fa
taeoaitdrdtpdeetiwt
```

- Dar ... ?

IDENTIFICAREA SECVENŢELOR MOTIF

- O abordare mai bună
  - Examinarea frecvenţei combinaţiilor de simboluri
  - “The” este cea mai frecventă combinaţie de 3 simboluri în limba Engleză
    - “;48” este cea mai frecventă combinaţie de 3 simboluri în textul criptat
  - …

IDENTIFICAREA SECVENȚELOR MOTIF

- 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:1eet+?t
```

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

IDENTIFICAREA SECVENŢELOR MOTIF

- Odată identificate toate corespondenţele mesajul final este:

```
AGOODGLASSINTHEBISHOPSHOSTELINTHEDEVILSSEATWENYONEDEGRE
ESANDTHIRTEENMINUTESNORTHEASTANDBYNORTHMAINBRANCHSEVENT
HLIMBEASTSIDESHOOTFROMTHELEFTEYEOFTHEDEATHSHEADABEELINE
FROMTHETREETHROUGHTHESHOTFIFTYFEETOUT
```

IDENTIFICAREA SECVENŢELOR MOTIF

- Semnele de punctuaţie 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 SECVENŢELOR MOTIF

- Cunoştinţe necesare pentru rezolvarea textului codat:
  - Frecvenţa relativă a fiecărei litere din alfabet
  - Frecvenţa relativă a combinaţiilor de 2 sau mai multe litere în limba Engleză
  - Cunoaşterea tuturor cuvintelor din vocabularul limbii Engleze este de dorită pentru a fi capabili să facem inferenţe acurate

IDENTIFICAREA SECVENŢELOR MOTIF

Similarităţi

- Nucleotidele din motif codifică un mesaj în limbaj genetic
  - Simbolurile din textul codificat codifică un mesaj în limba Engleză
- Pentru a rezolva problema, analizăm frecvenţe ale grupărilor de simboluri în ADN / mesajul codificat
- Cunoaşterea regiunilor motifs cu rol în reglare face problema de identificare a regiunilor motifs uşoară.
  - Cunoaşterea cuvintelor din dicţionarul limbii Engleze ne ajută să gasim soluţia în cazul mesajului codificat.

IDENTIFICAREA SECVENŢELOR MOTIF

Diferenţe

- Căutarea secvenţelor motifs
  - Pentru a găsi soluţia analizăm frecvenţa patern-urilor (secvenţe de mai multe nucleotide) în secvenţele de nucleotide
- Problema textului codificat:
  - Pentru a găsi soluţia analizăm frecvenţele de pattern-uri (grupuri de simboluri) în textul scris în limba Engleză

Diferenţe

- Căutarea secvenţelor motifs
  - Cunoaşterea secvenţelor motifs identificare reduce complexitatea problemei
- Problema textului codificat:
  - Cunoaşterea cuvintelor din dicţionar este de dorit pentru rezolvarea problemei

IDENTIFICAREA SECVENŢELOR MOTIF

- **Problema identificării secvenţelor motif este mai grea decât problema textului codificat**
  - Nu avem dicţionarul complet al motifs-urilor
  - Limbajul genetic nu are un dicţionar standard
  - Doar o fracţiune mică de secvenţe nucleotidice codifică motifs-urile; volumul de date este însă enorm

IDENTIFICAREA SECVENŢELOR MOTIF: PROBMELA

- Fie o serie de secvenţe randomizate de ADN:

  cctgatagacgctatctggctatccacgtacgtaggtcctctgtgcgaatctatgcgt
  ttccaaccat

  agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcag
  aaccagaagtgc

  aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataag
  acgaaaatttt

  agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatcttacgt
  acgtataca

  ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgtt
  aacgtacgtc

- Găsiţi pattern-ul (motifs-ul) implementat la nivelul secvenţele individuale

IDENTIFICAREA SECVENŢELOR MOTIF: PROBMELA

- Informaţii suplimentare:
  - Secvenţa motif este de lungime 8
  - Pattern-ul nu este exact la fel în fiecare apariţie datorită mutaţiei aleatorii punctiformă

IDENTIFICAREA SECVENŢELOR MOTIF: PROBMELA

- **Pattern-ul identificat:**

cctgatagacgctatctggctatcc**acgtacgt**aggtcctctgtgcgaatctatgcgtttcca
accat

agtactggtgtacatttgat**acgtacgt**acaccggcaacctgaaacaaacgctcagaacca
gaagtgc

aa**acgtacgt**gcaccctctttcttcgtggctctggccaacgagggctgatgtataagacga
aaatttt

agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatctt**acgtacgt**a
taca

ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgtta**acgt
acgt**c

IDENTIFICAREA SECVENŢELOR MOTIF: PROBMELA

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

cctgatagacgctatctggctatcc**aGgtacTt**aggtcctctgtgcgaatctatgcgtttcc aaccat

agtactggtgtacatttgat**CcAtacgt**acaccggcaacctgaaacaaacgctcagaacc agaagtgc

aa**acgtTAgt**gcaccctctttcttcgtggctctggccaacgagggctgatgtataagacg aaaatttt

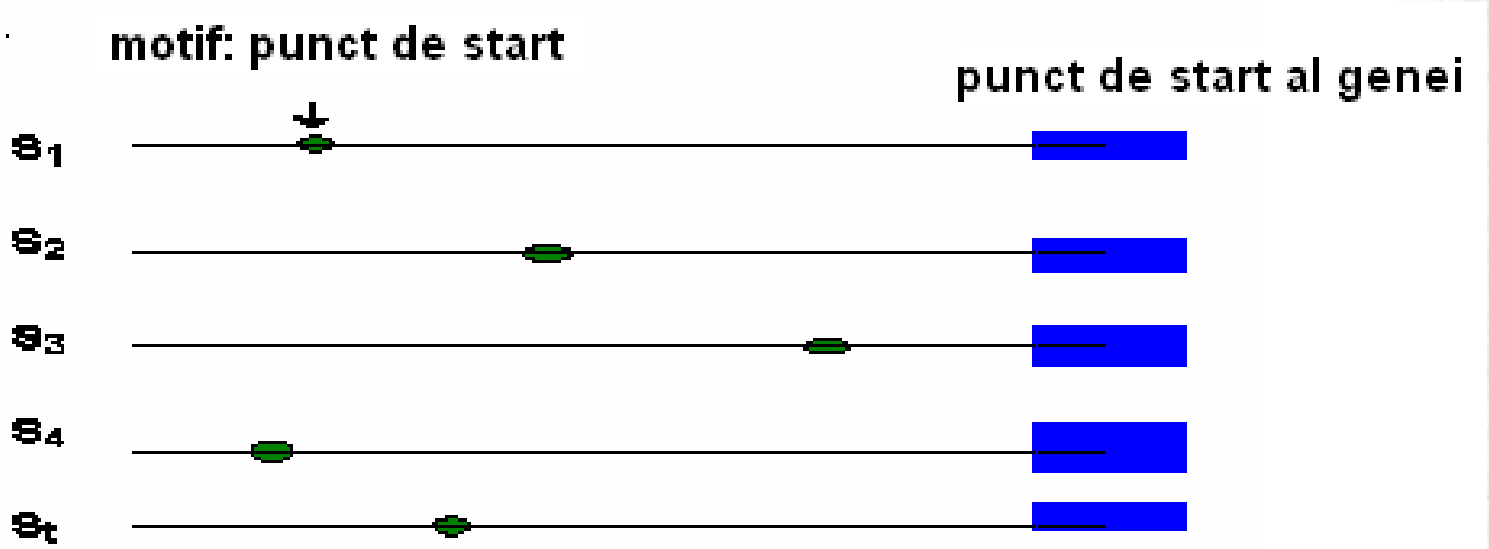
agcctccgatgtaagtcatagctgtaactattacctgccacccctattacatctt**acgtCcAt** ataca

ctgttatacaacgcgtcatggcggggtatgcgttttggtcgtcgtacgctcgatcgtta**Ccgt acgG**c

- Secvenţele pot fi identificate în cazul existenţei mutaţiilor?

IDENTIFICAREA SECVENŢELOR MOTIF: PROBMELA

- Identificarea pattern-urilor se poate realiza mai uşor dacă cunoaştem punctul de start al secvenţei motif
  - $\boldsymbol{s} = (s_1, s_2, s_3, \ldots, s_t)$



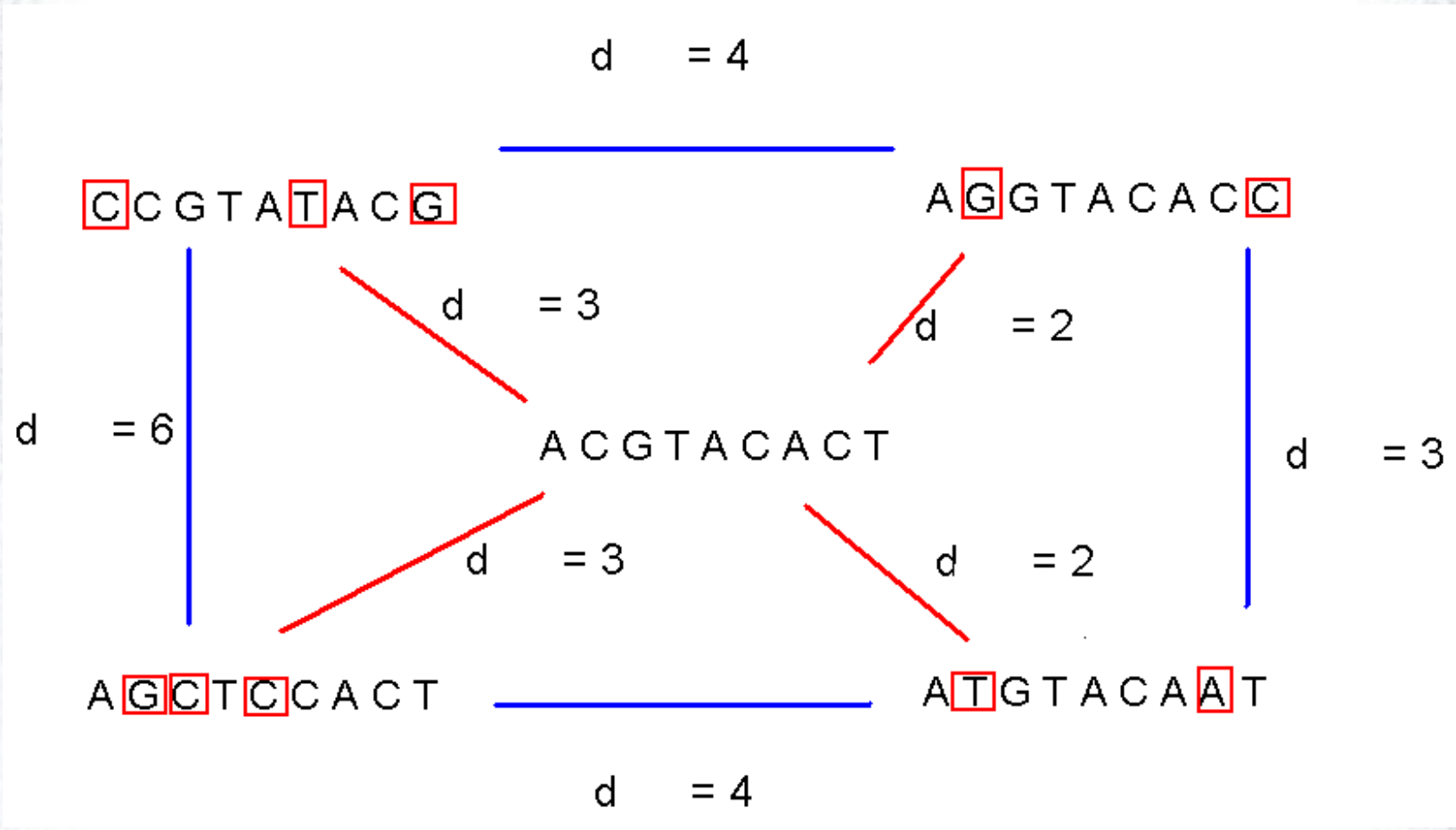
MOTIFs: PROFILURI ŞI CONSENS

		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

- Un strămoş al secvenţei motif
  - Plecând de la acest strămoş emerg pattern-urile care prezintă mutaţii
- Distanţa dintre secvenţa motif reală şi secvenţa consens este în general mai mică decât aceea pentru două secvenţe motif reale

MOTIFs: CONSENS



MOTIFs: CONSENS

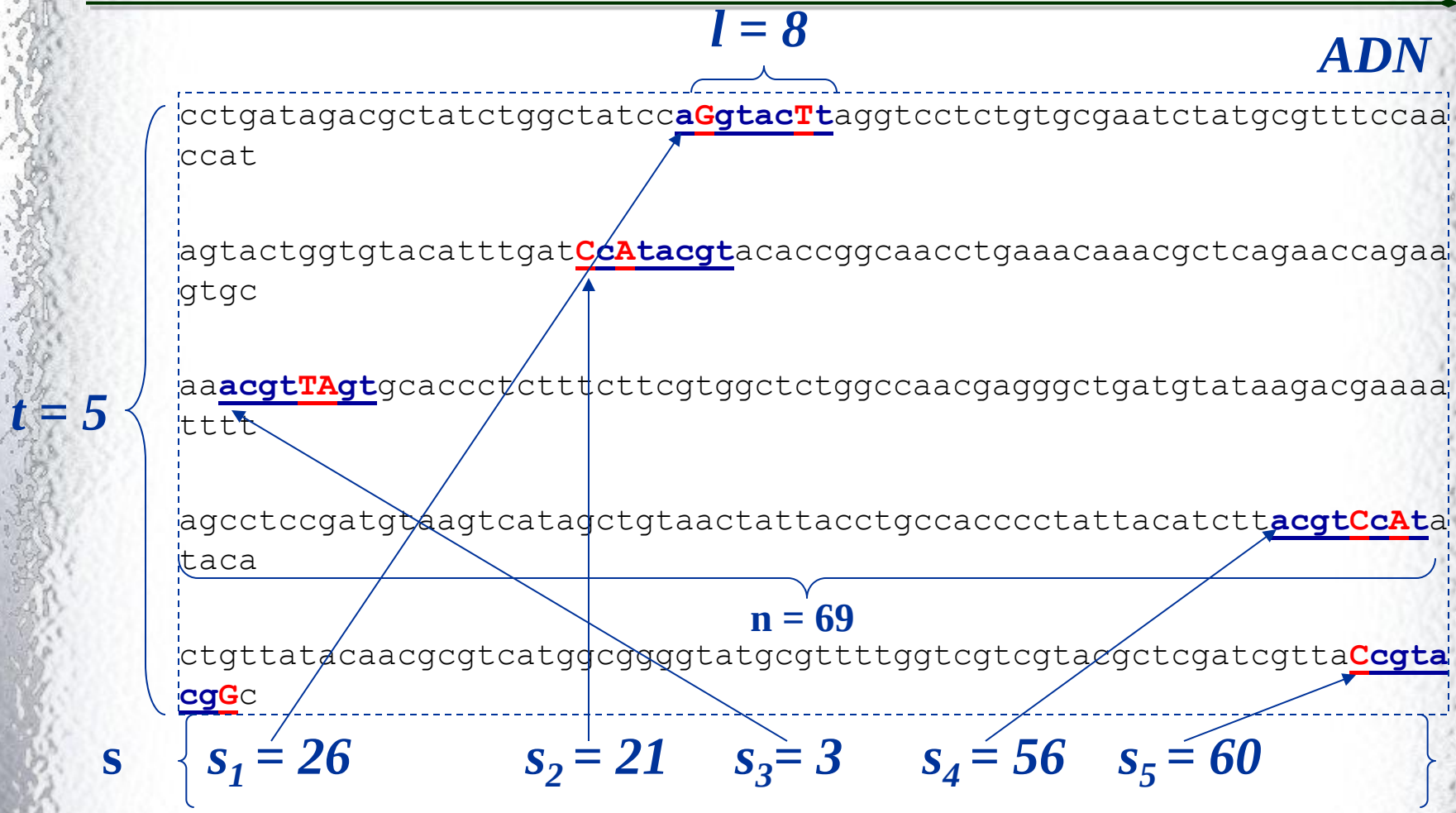
- Avem secvenţa motifs rezultată din consens
  - Dar ... cât de bun este acest consens?

- Prin introducerea unei funcţii de tip scor putem compara diferitele pattern-uri identificate/posibile şi putem alege cea mai bună secvenţă motif

MOTIFs: TERMENI

- **t =** numărul de secvenţe ADN
- **n =** lungimea fiecărei secvenţe ADN
- **AND =** eşantionul de secvenţe ADN (t×n)
- ℓ = lungimea secvenţei motif (ℓ-mer)
- s_i = punctul de start al *l-mer* în secvenţa i
- $s=(s_1, s_2,\ldots s_t)$ = mulţimea poziţiilor start al pattern-urilor

MOTIFs: PARAMETRII



MOTIFs: Calculareа Scorului

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

- 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 SECVENŢELOR MOTIF: PROBMELA

- Dacă cunoaştem punctele de start ale secvenţelor motifs identificarea consensului este uşoară chiar şi în cazul existenţei mutaţiilor în interiorul secvenţei motifs

- Dar … poziţiile de start nu se cunosc deobicei. În aceste condiţii cum putem identifica cea mai bună matrice de profil?

IDENTIFICAREA SECVENŢELOR MOTIF: PROBMELA

- Formularea problemei:
  - Scop: având un scop de secvenţe ADN, identificaţi pentru fiecare secvenţă setul ***l-mer*** care sa maximizeze scorul consensului
  - Date de intrare:
    - matricea t×n a secvenţelor ADN
    - l = lungimea pattern-ului de identificat
  - Date de ieşire: un şir de ***t*** poziţii start $\boldsymbol{s = (s_1, s_2, \ldots, s_t)}$ care maximizează ***scor(s,ADN)***

IDENTIFICAREA SECVENŢELOR MOTIF: SOLUŢIA FORŢELOR BRUTE

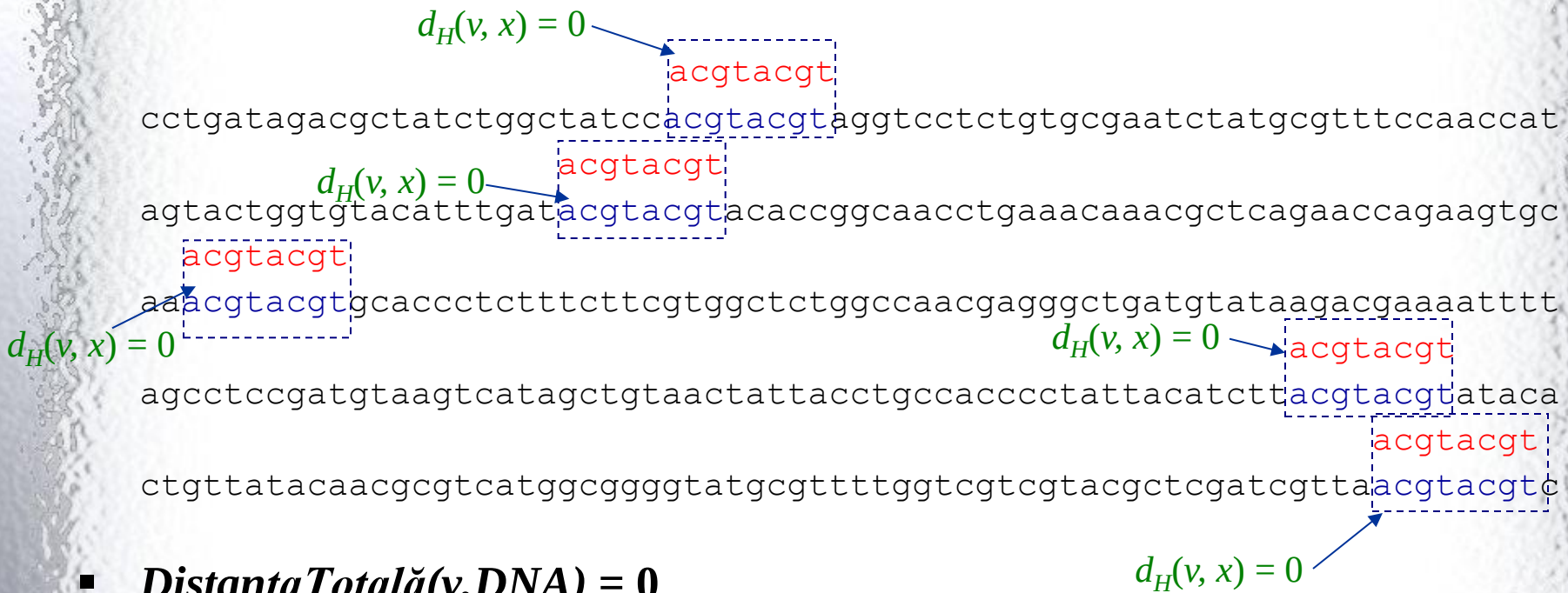
- Calculăm scorurile pentru toate combinaţiile posibile de puncte de start ***s***
- Cel mai bun scor identifică profilul cel mai bun şi respectiv consensul de pattern în secvenţele de ADN
- Obiectivul global este de a maximiza *scor(s,ADN)* prin varierea poziţiilor de start s_i (unde $s_i = [1, \ldots, n-l+1]$ şi $i = [1, \ldots, t]$)

IDENTIFICAREA SECVENŢELOR MOTIF: SOLUŢIA FORŢELOR BRUTE

- Timpul necesar identificării soluţiei celei mai bune:
  - Prin varierea poziţiilor *(n-l+1)* pentru fiecare secvenţă *t* vom investiga $(n-l+1)^t$ seturi de poziţii start
  - Pentru fiecare set de poziţii start, funcţia scor va face *l* operaţii, astfel încât complexitatea este $\boldsymbol{l}\,(n - \boldsymbol{l} + 1)^t = O(\boldsymbol{l}\, n^t)$
  - Pentru $\boldsymbol{t} = 8$, $\boldsymbol{n} = 1000$, $\boldsymbol{l} = 10$ trebuie să realizăm ~ 10^{20} operaţii – va dura foarte mult timp

DISTANȚE TOTALE: EXEMPLU

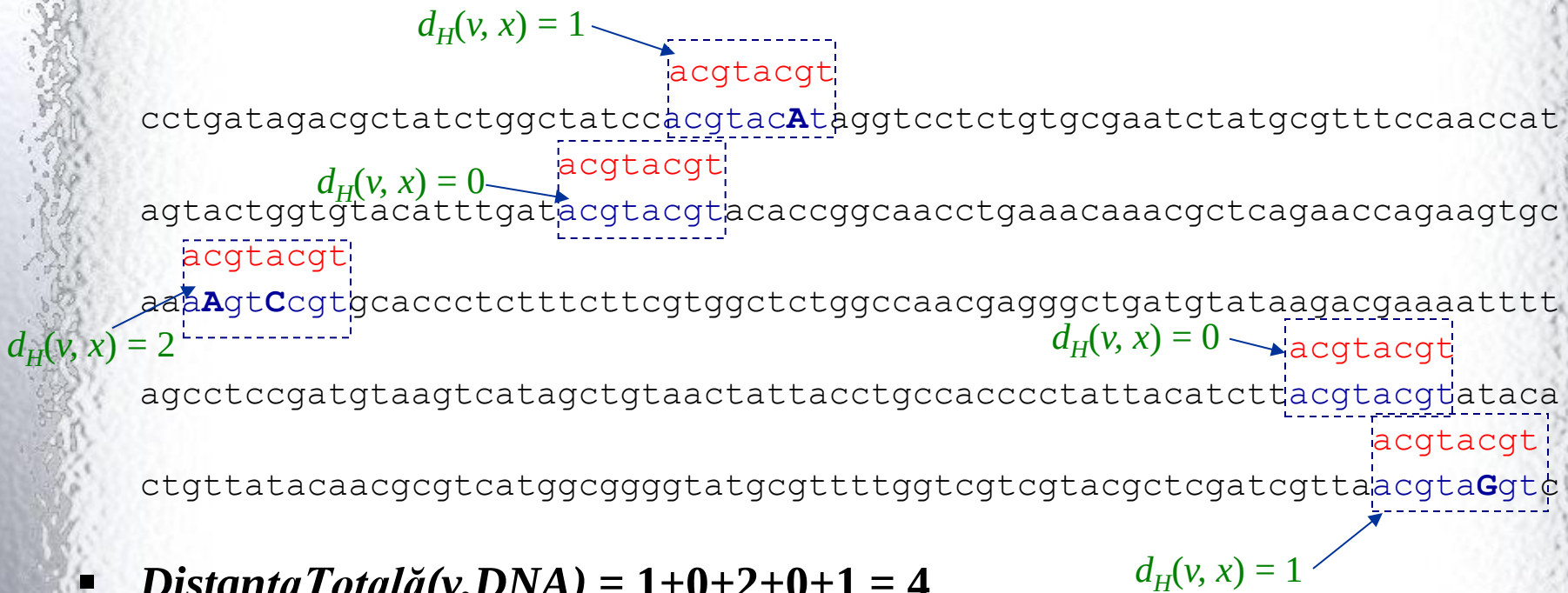
- Fie secvenţa ***v*** = "acgtacgt" şi x secvenţa de interes "acgtacgt"



- ***DistanţaTotală(v,DNA)*** **= 0**

DISTANŢE TOTALE: EXEMPLU

- Fie secvenţa ***v*** = "acgtacgt" şi x secvenţa de interes "acgtacgt"



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

DISTANȚE TOTALE: DEFINIȚIE

- Pentru fiecare *i* secvenţă de ADN se calculează $d_H(v, x)$, unde x etse l-mer-ul cu punctul de start s_i ($1 \leq s_i \leq n - l + 1$)
- SSe identifică valoarea minimă $d_H(v, x)$ printre ***l-mer*** în secvenţa ***I***
- ***DDistanţaTotală(v,ADN)*** este suma distanţelor minime Hamming pentru fiecare secvenţă ***i*** de ***AND***
  - $DistanţaTotală(v,ADN) = min_s\ d_H(v, s)$
    - Unde ***s*** este setul de puncte start $s_1, s_2, \ldots s_t$

PROBLEMA STRING-ULUI MEDIAN: FORMULARE

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

MOTIFS vs PROBLEMA STRINGURILOR MEDIANE

- Identificarea secvenţei MOTIF
  - Problemă de maximizare
- Stringul median:
  - Problemă de minimizare
- Din punct de vedere computaţional sunt echivalente
- Minimizarea ***DistanţeiTotale*** este echivalentă cu identificarea ***Scorului*** maxim

IDENTIFICAREA SECVENŢELOR MOTIF: PROBLEMA ŞIRULUI MEDIAN

- Fie un set de ***t*** secvenţe ***ADN***
- Să se identifice pattern-ul (secvenţa motif) cu număr minim de mutaţii care apare în toate secvenţele ***t***
- Distanţa Hamming:
  - $d_H(\boldsymbol{v},\boldsymbol{w})$ este numărul de perechi de nucleotide care nu se potrivesc când ***v*** şi ***w*** sunt aliniate
  - Exemplu: d_H(AAAAAA,ACAAAC) = 2

Identificarea Secvenţelor Motif vs. Problema Şirului Median

- De ce am reformula problema de identificare a secvenţelor motifs?
  - Este necesară examinarea tuturor combinaţiilor ***s***.
    - Rezultă un număr total de $(n - l + 1)^t$ combinaţii
  - Problema şirului median necesită examinarea a 4^l combinaţii – număr relativ mic în comparaţie cu precedentul

STRUCTURAREA CĂUTĂRII

- Pentru problema stringului median trebuie considerate toate 4^l posibililele ***l-mer***-uri

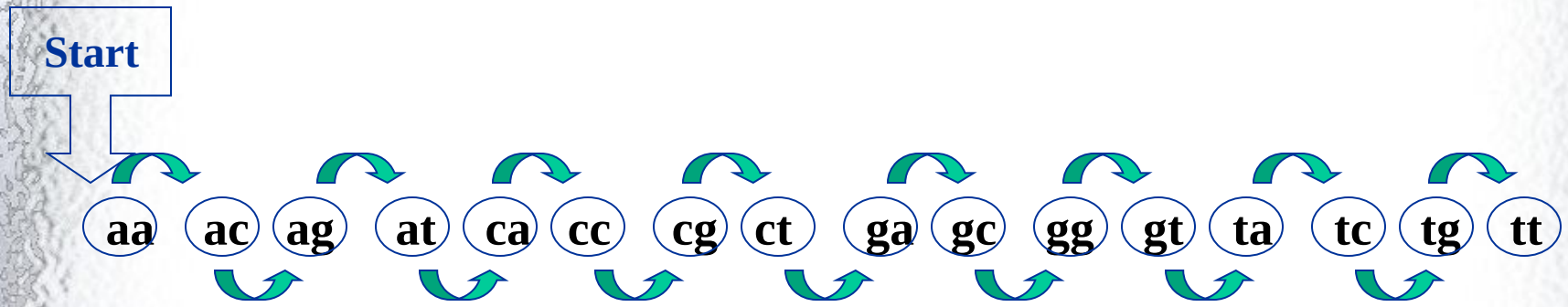
aa… aa
aa… ac
aa… ag
aa… at
.
.
tt… tt

l

- Cum putem organiza această căutare?

STRUCTURAREA CĂUTĂRII

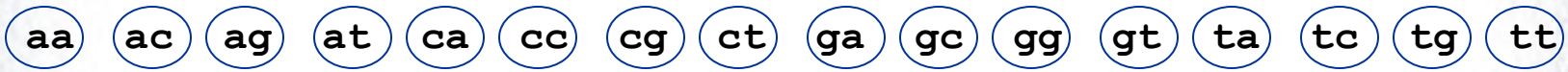
- Fie ***l*** = 2



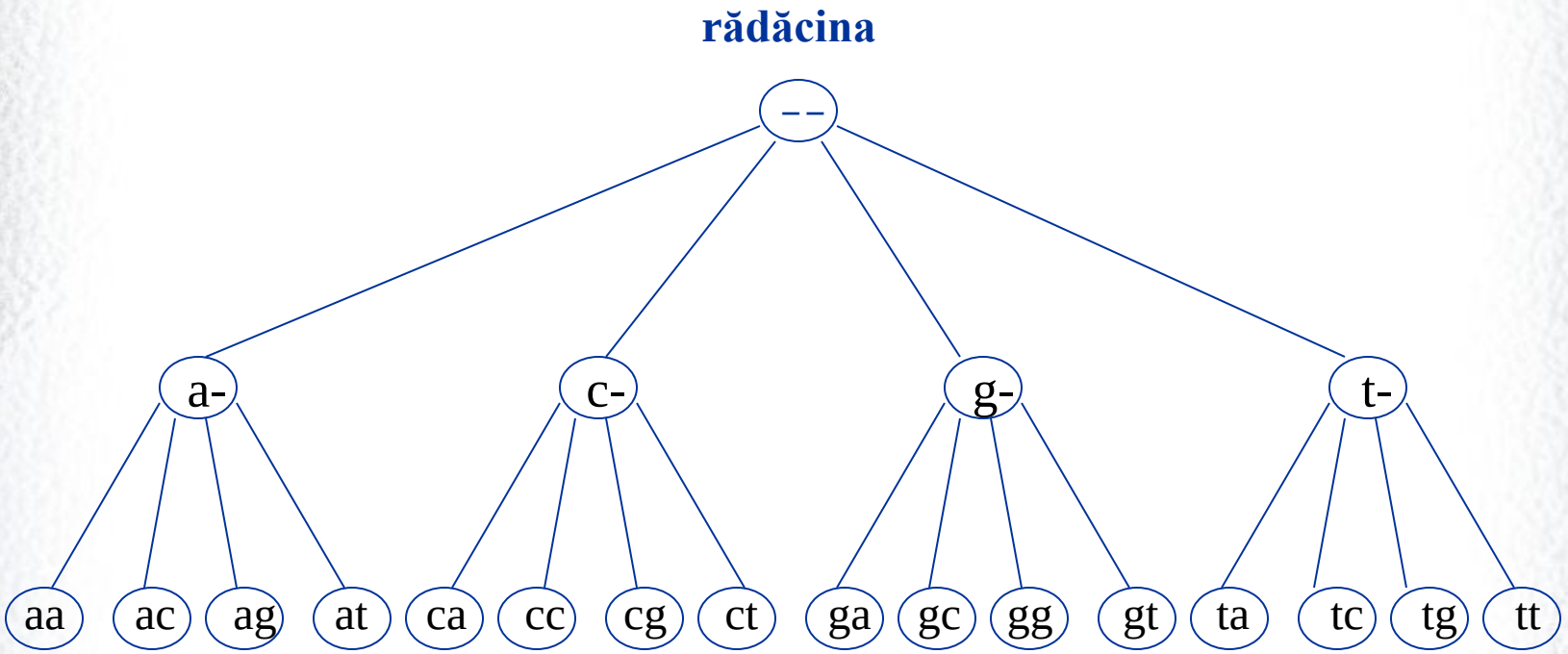
- Este necesară investigarea tuturor predecesorilor unei secvenţe pentru a investiga secvenţa de interes

STRUCTURAREA CĂUTĂRII

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



STRUCTURAREA CĂUTĂRII



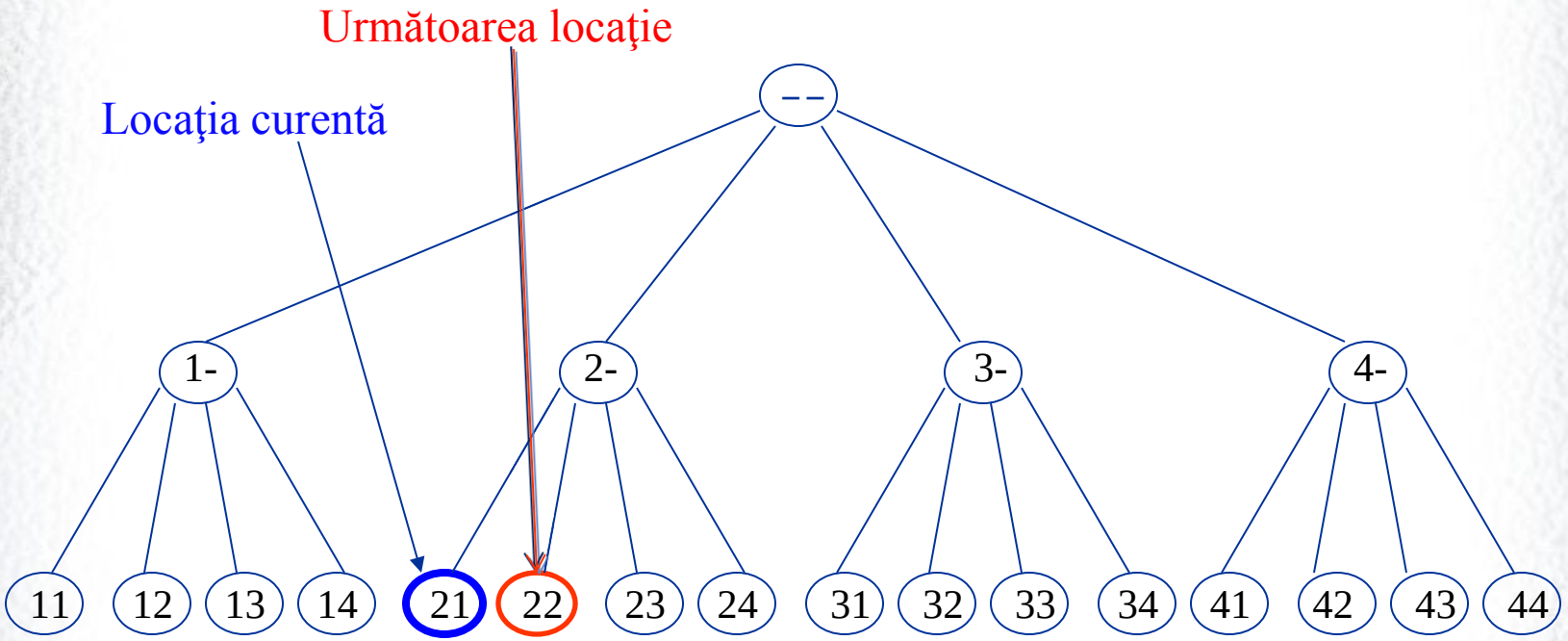
Analiza Căutării de Tip Arbore

- Caracteristicile căutării de tip arbore:
  - Secvenţele sunt conţinute la nivelul frunzelor
  - Părintele unui nod este prefixul copilului

- Care este modalitatea de mişcare în arbore?
  - Următoarea frunză
  - Vizitarea tuturor frunzelor
  - Vizitarea următorului nod
  - Trecerea de la copil la nod (părinte)

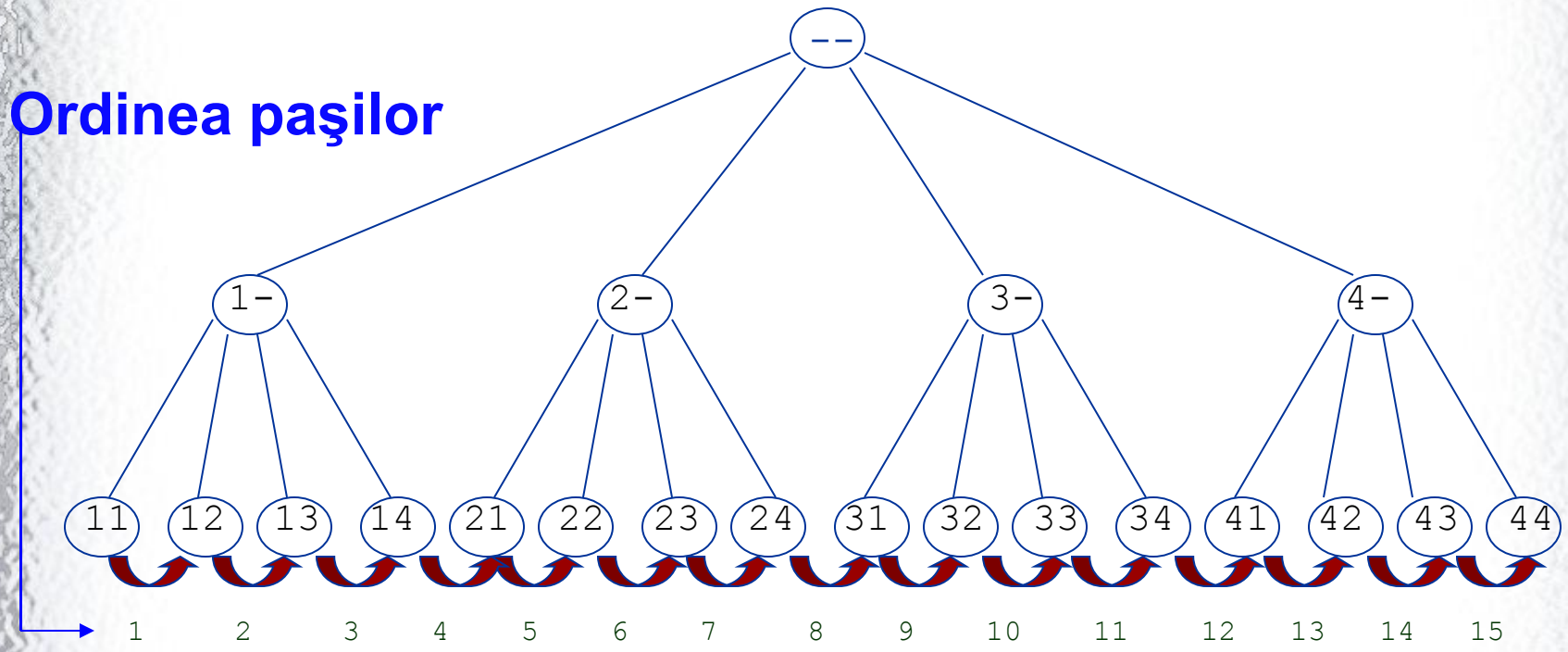
URMĂTOAREA FRUNZĂ: EXEMPLU

- Mişcarea la următoarea frunză:



VIZITAREA TUTUTOR FRUNZELOR

- Mişcarea de la o frunză la cealaltă :

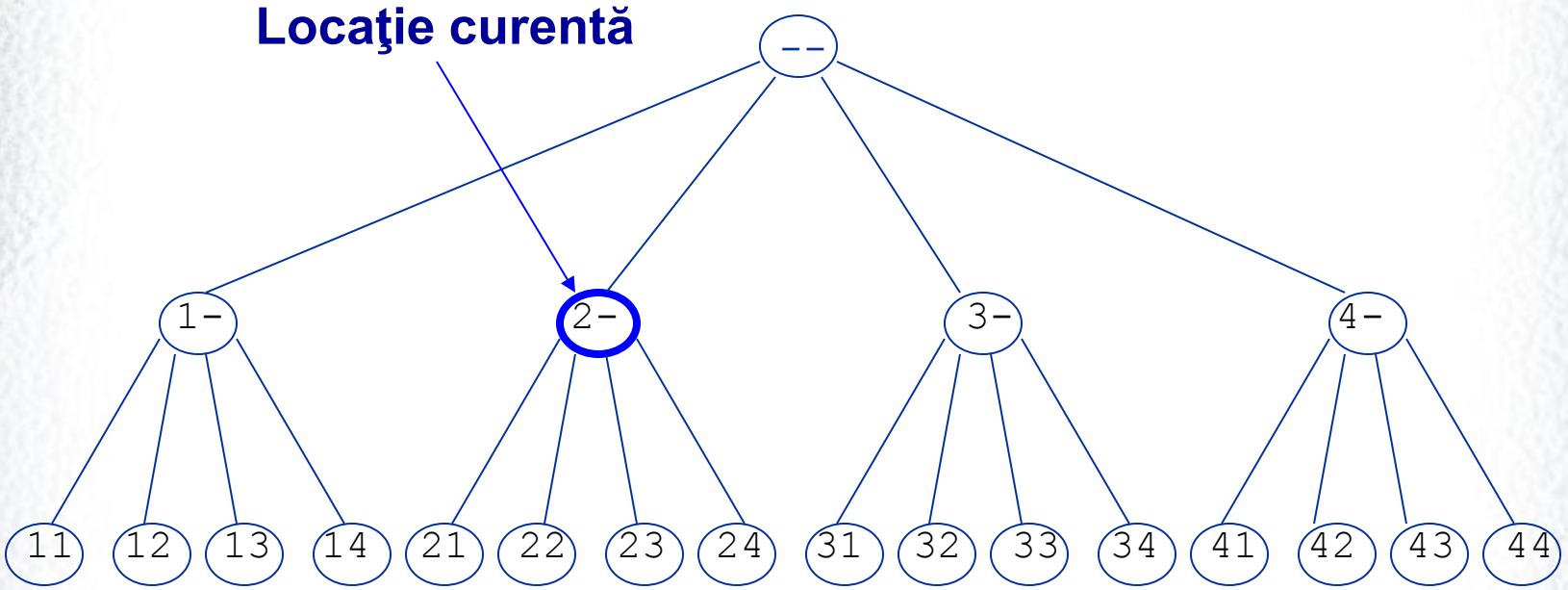


CĂUTAREA ÎN ADÂNCIME

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

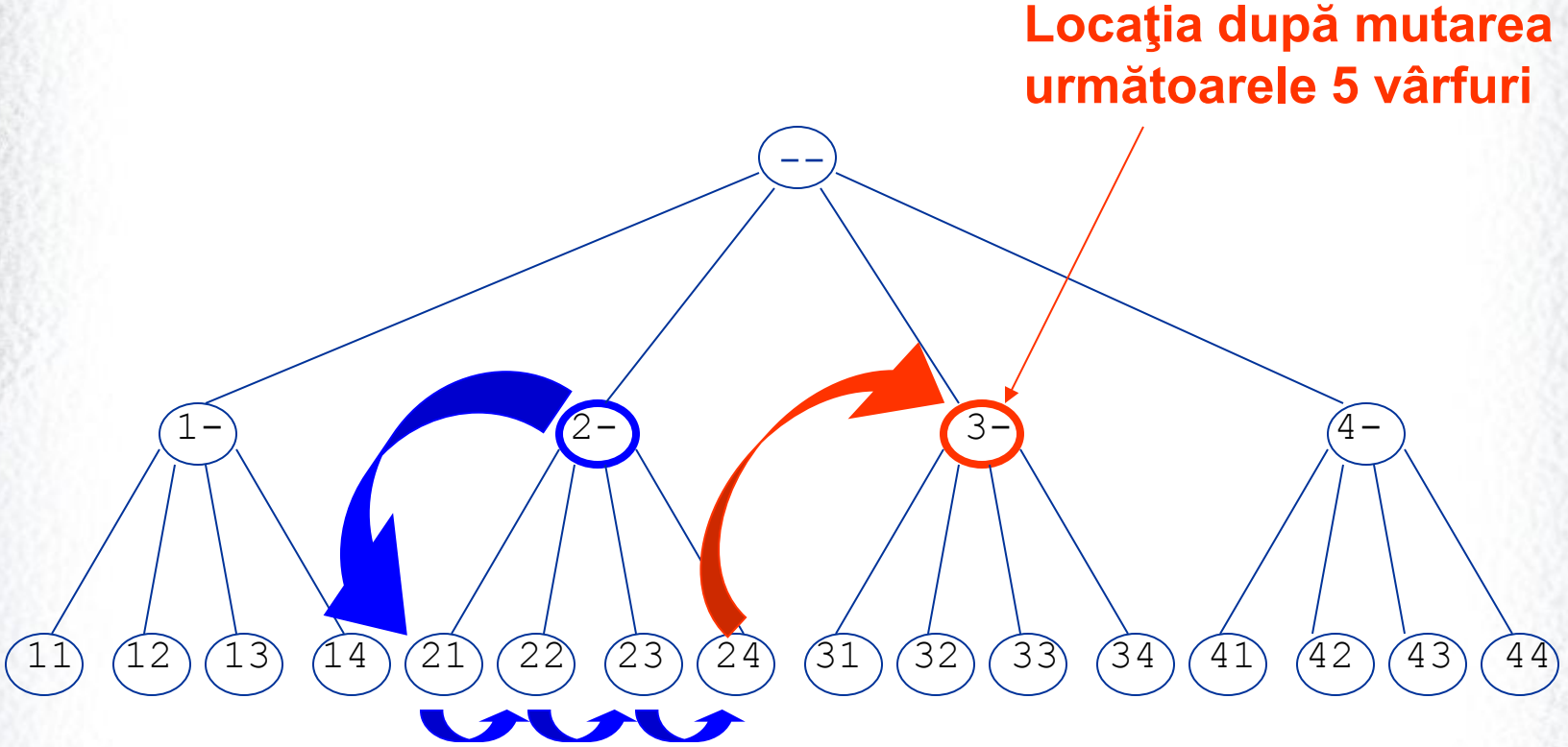
EXEMPLU

- Mişcarea către următorul vârf:



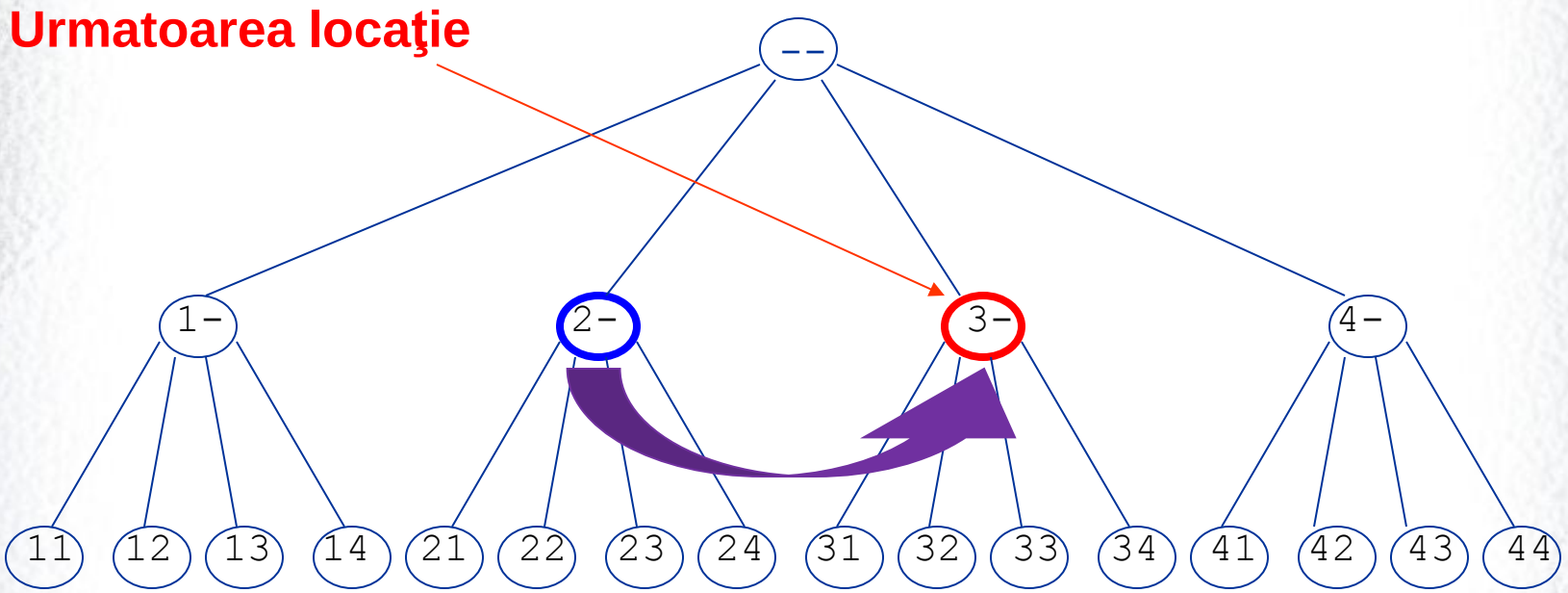
EXEMPLU

- Mişcarea către următorul vârf:



EXEMPLU

- Ocolirea descendenţilor lui “2-”:



MOTIFs: PROGRAME

- **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)
- …