Mustrite avastamine joondamata bioloogilistest järjestustest

Age Tats
GT IV
26.11.2003

Milleks mustrit otsida?

- Geeni- ja valgujärjestuste primaarstruktuuri sarnasus esineb tavaliselt järjestuste vahel, mis on evolutsiooniliselt konserveerunud, kuna neil on oluline funktsionaalne või strukturaalne roll.

- Diskreetsed/deterministlikud mustrid

- Tõenäosuslikud mustrid
- Kaalumaatriks (*Position weight matrix*)

- Tähestik Σ
- Mitmetähenduslikud sümbolid

R=[AG], Y=[CT], W=[AT], S=[GC], B=[CGT], D=[AGT], H=[ACT], V=[ACG], N=[ACGT]

- *Wild-card/don't care*
- Paindlik gäp x(i,j) (*flexible gap*)
- *Mismatch'*id (asendused, insertsioonid, deletsioonid)

Gäpi mudelid

- Gäppe ei lubata üldse

- Lineaarse karistuse mudel

Penalty = b*X

- Afiinse gäpi mudel

Penalty = a + b*X

Joondatud või joondamata?

1. Bioloogilised teadmised (järjestuse globaalsed omadused, fülogeneetilised seosed, sekundaar-/tertsiaarstruktuur)

2. Täiesti uus muster

   Nõrgad mustrid

Kõikvõimalike mustrite loetlemine (*enumerating all patterns*)

Tahame leida kõige olulisemat mustrit pikkusega 10 lubades maksimaalselt 2 mismatchi:

Tähestik Σ={A,C,G,T}

Võimalikke stringe ehk mustreid

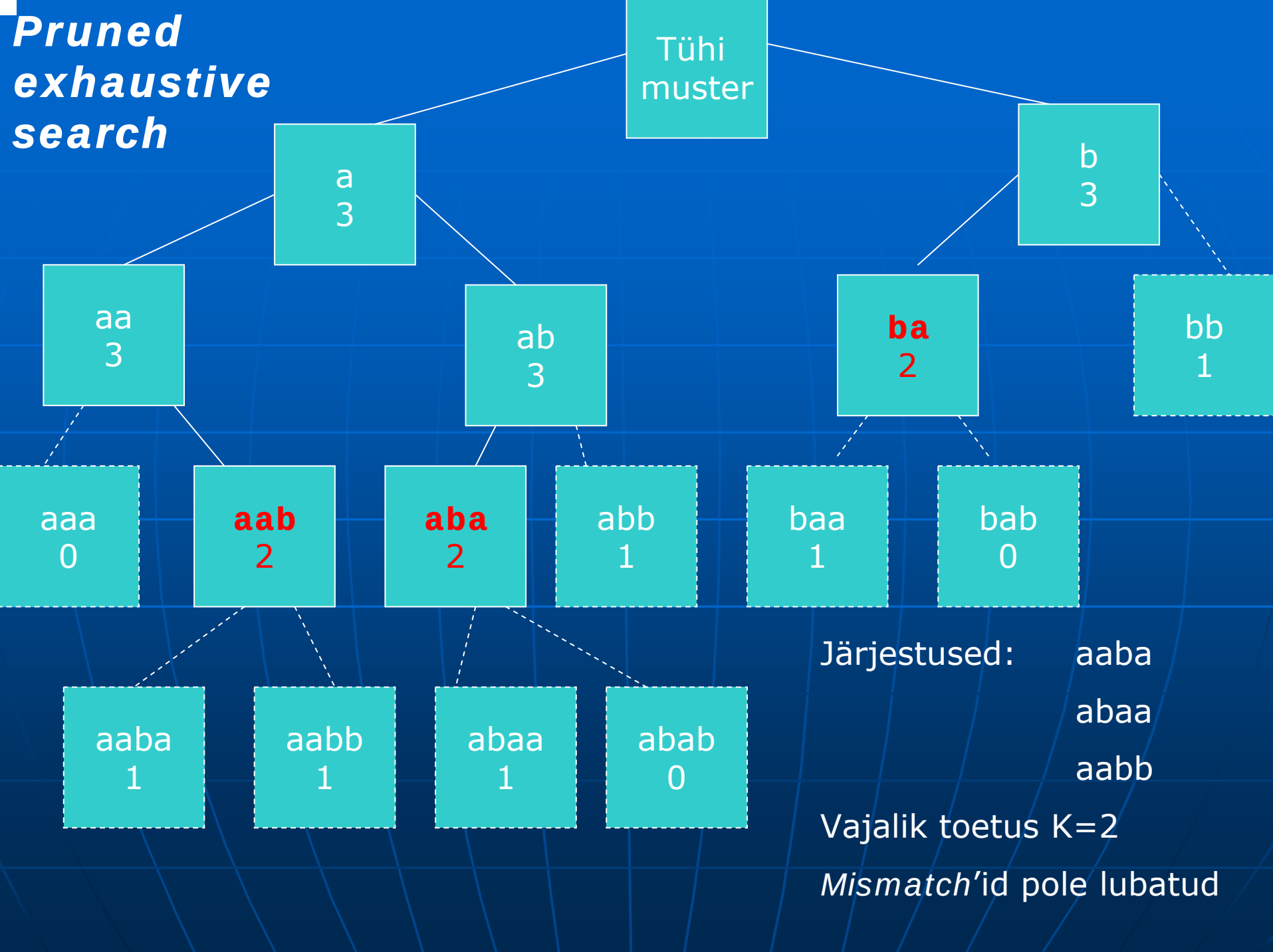
4e10= 1 048 576

Tööaeg kasvab eksponentsiaalselt mustri pikkusega.

Tööaeg kasvab lineaarselt sisestatud järjestuste arvuga.

Sobib lühikeste mustrite leidmiseks suurest hulgast järjestustest.

Pruned exhaustive search



Järjestused: aaba
abaa
aabb

Vajalik toetus K=2

Mismatch'id pole lubatud

Pratt
(Jonassen,1997)

- *Pruned exhaustive search*
- Lubatud paindlikud gäpid ja mitmetähenduslikud sümbolid
- Kasutaja määrab mustri kogupikkuse, gäppide maksimaalse arvu, paindlike gäppide maksimaalse arvu, lubatavad mitmetähenduslikud sümbolid.
- Vähemspetsiifilised eemaldatakse

http://www.ebi.ac.uk/pratt/

ftp://ftp.ii.uib.no/pub/bio/Pratt

TEIRESIAS
(Rigoutsos and Floratos, 1998)

- Leiab kõik mustrid, mis esinevad vähemalt K-s kasutaja poolt ette antud arvus järjestustes, joondamata järjestuste hulgast
- Võimalus otsida leitud mustrite esinemisi SWISS-PROT'ist
- Põhineb lühikeste mustrite põhjalikult otsingul ning lühemate mustrite kombineerimisel pikemateks'
- Sisend limiteeritud – 30Kb.

http://cbcsrv.watson.ibm.com/Tspd.html
http://cbcsrv.watson.ibm.com/download.phtml.html

TEIRESIAS

- Definitsioon:

Muster P on (L,W) muster, kui ta vastab järgmistele nõuetele:

- P on hulga Σ sümbolitest ja *wild-card*'idest '.' koosnev string.
- P algab ja lõpeb sümboliga hulgast Σ.
- Iga P alammuster (alamjärjestus, mis algab ja lõpeb sümboliga hulgast Σ) sisaldab täpselt L mitte-*wild-card*'i ja on maksimaalselt W pikkune.

TEIRESIAS

- Kui muster P on (L,W) muster, mis esineb vähemalt K-s järjestuses, siis on ka tema alammustrid (L,W) mustrid, mis esinevad vähemalt K-s järjestuses

Maksimaalseid mustreid saab koostada väiksematest alammustritest

TEIRESIAS

- Sama esinemisarvuga mustritest väljastatakse ainult kõige spetsiifilisem.

AB.CD.E AB..D

- Kui vähemspetsiifilisel on suurem toetus, väljastatakse ka vähemspetsiifiline.

TEIRESIAS

- Skaneerimise faas

Leitakse kõik (L,W) mustrid, mis esinevad vähemalt K-s järjestuses ja sisaldavad täpselt L mitte-*wild-card*'i.

(*pruned exhaustive search*)

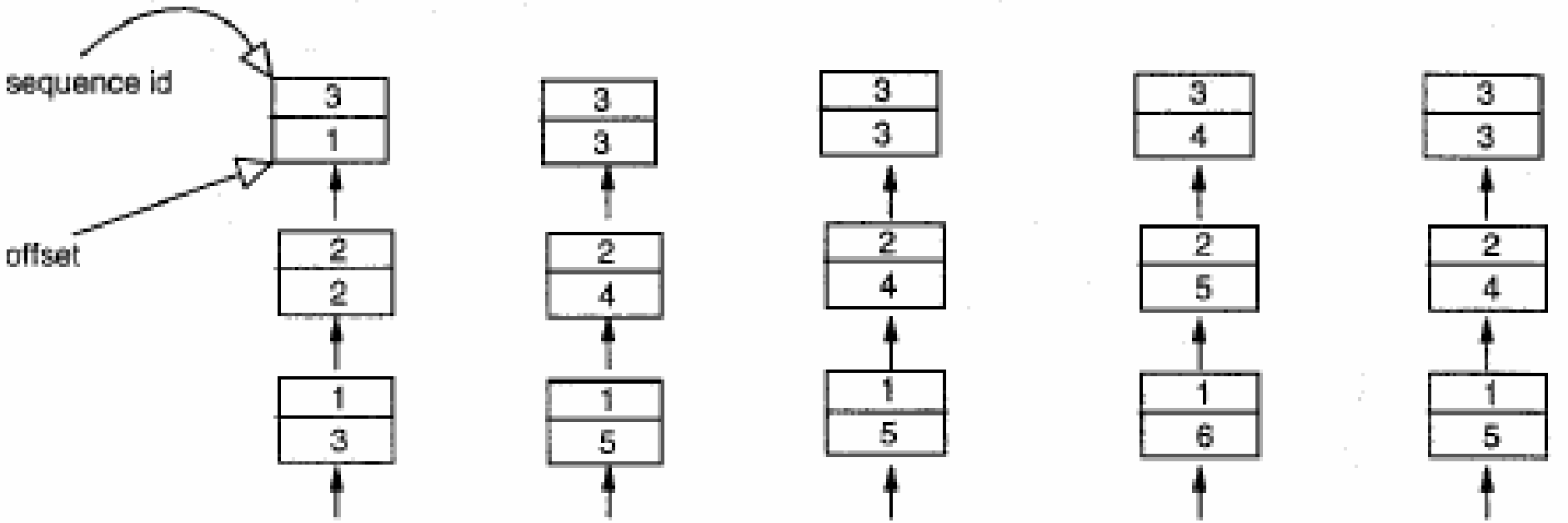
- Konvolutsiooni faas

Püütakse saadud mustreid pikendada neid omavahel kokku liimides. Saadud mustri esinemine arvutatakse alammustrite esinemiste põhjal.

TEIRESIAS

L = 3, W = 4, K = 3

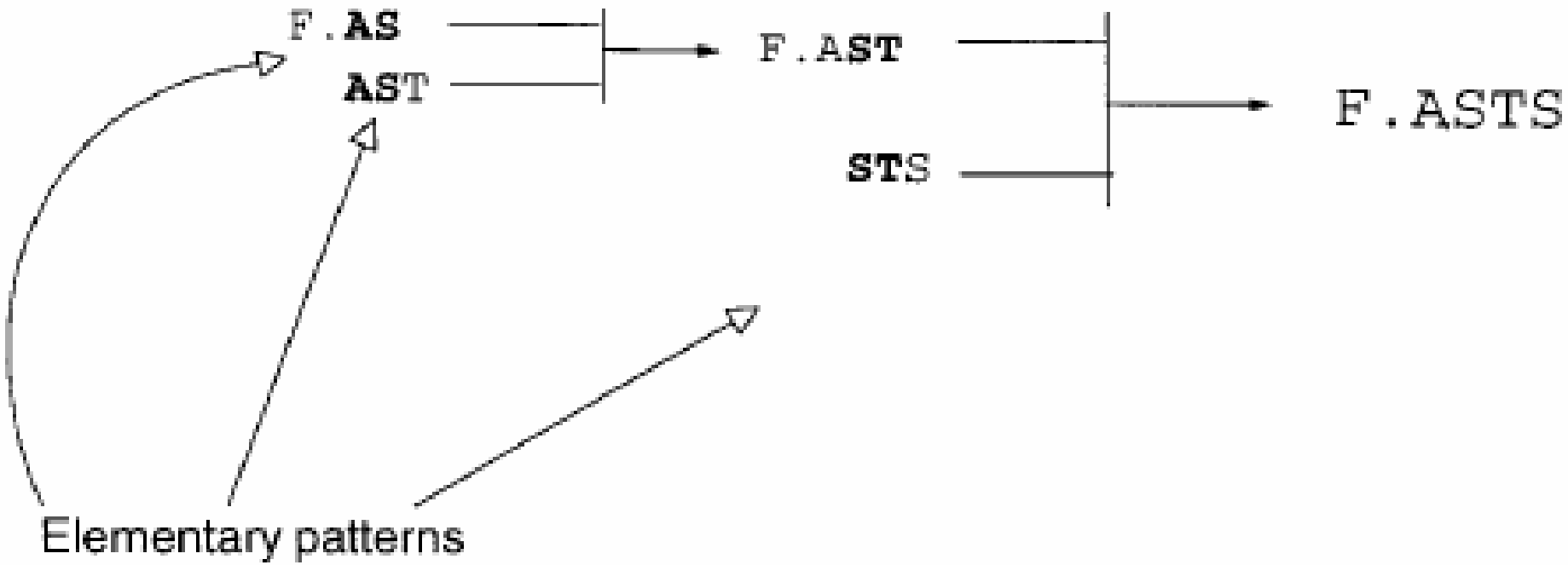
S = {s_1 = "SDFBASTS", s_2 = "LFCASTS", s_3 = "FDASTSNP"}



Elementary Patterns = { F.AS AST AS.S STS A.TS }

s_1:	SDFBASTS	SDFBASTS	SDFBASTS	SDFBASTS	SDFBASTS
s_2:	LFCASTS	LFCASTS	LFCASTS	LFCASTS	LFCASTS
s_3:	FDASTSNP	FDASTSNP	FDASTSNP	FDASTSNP	FDASTSNP

TEIRESIAS



TEIRESIAS

- Algoritm annab kõik maksimaalsed (L,W) mustrid.
- Tööaeg lineaarne väljundi suurusega.

SPEXS (Sequence Pattern EXhaustive Search) (Vilo, 2000)

- Kiire põhjaliku otsingu algoritm
- Määratavad parameetrid näiteks: motiivi maksimaalne pikkus, mitmetähenduslike sümbolite arv, *wild-card*'ide pikkus ja arv, sobivuse lävi.
- Erinevad otsingustrateegiad:

-lühematest pikemateni (breadth first)

-tähestikulises järjekorras (depth first)

-kõige sagedasemad motiivid enne

-laiendades kõige ´lootustandvamat' mustrit esimesena

SPEXS

- Mustreid laiendatakse alati paremale.
- Vahetult enne väljastamist laiendatakse ´huvitavaid’ mustreid mõlemale poole ilma toetuse kaotuseta.
- Väljastatakse mustrid koos esinemise arvuga, vajadusel ka positsiooniga.
- G valkude retseptorite seostumine G valkudega

http://ep.ebi.ac.uk/EP/SPEXS/

Expectation maximization
(Lawrence and Reilly, 1990)

- Sisend: joondamata järjestused ja motiivi pikkus (W)
- Väljund: ühise motiivi tõenäosuslik mudel (PWM)
- Eeldatakse, et motiiv esineb igas järjestuses ainult 1 korra.

-Etapp E:

Iga positsiooni jaoks igas järjestuses s arvutatakse tõenäosus, et muster esineb s-is alates sellest positsioonist.

-Etapp M:

Mustri iga positsiooni jaoks arvutatakse uued sümbolite tõenäosused selles positsioonis.

MEME (Multiple EM for Model Elucidation)
(Bailey and Elkan, 1995)

- EM algoritmi modifikatsioon.
- Esitab motiivid PWM-idena.
- Motiivid on lühikesed ja gäppideta.
- (Varieeruva pikkusega gäppe sisaldavad mustrid jagatakse 2 või enamasse eraldi motiivi)
- Sisend limiteeritud – 60 000 märki.

http://meme.sdsc.edu/meme/website/meme.html
ftp://ftp.sdsc.edu/pub/sdsc/biology/meme/

MEME

- Igale andmehulga alamjärjestusele moodustatakse algne mudel – PWM, kus igal sümbolil igas alamjärjestuse positsioonis on tõenäosus p.
- Iga sellise mudeli peal teostatakse EM algoritm. Saadud mudelite jaoks arvutatakse tõenäosuse skoor.
- Valitakse parima skooriga mudel, mis saab järgmiseks algseks mudeliks EM algoritmile.

MEME

- MEME – avastab üleesindatud motiivid teie järjestustes
- MAST – otsib avastatud motiividega järjestuste andmebaasi
- MetaMEME – kombineerib mitu MEME motiivi (HMM mudelina) ja otsib nendega.

Olulisus

- Suhe mustri esinemise ja oodatava esinemise vahel
- Statistilised mudelid

- Z-skoor

$$z_s = \frac{N_s - E(X_s)}{\sigma(X_s)}$$

Tundlikkus = TP/(TP+FN)

Spetsiifilisus = TN/(TN+FP)

Katse

- Vähemalt 3 sümbolit
- Max. gäppe 2
- Paindlikud gäpid (0,2)
- Toetus 3
- Sisend 20 järjestust (pikkus 20 ah)

Pratt

	fitness	hits(seqs)	Pattern
1:	16.6802	2(2)	E-R-x-L-R
2:	16.6802	2(2)	S-A-x-L-A
3:	16.6802	2(2)	L-A-x-S-L
4:	16.1802	2(2)	S-D-x(0,1)-A-S
5:	12.5102	2(2)	I-L-I
6:	12.5102	2(2)	I-I-x-I
7:	12.5102	2(2)	G-I-x-L
8:	12.5102	2(2)	G-x-I-I
9:	12.5102	2(2)	A-S-x-G
10:	12.5102	2(2)	A-S-x(2)-I
11:	12.5102	2(2)	S-x(2)-A-S
12:	12.5102	2(2)	F-F-S
13:	12.5102	2(2)	H-L-x(2)-F
14:	12.5102	2(2)	P-L-T
15:	12.5102	2(2)	N-T-x-R
16:	12.5102	2(2)	M-x-N-T
17:	12.5102	2(2)	L-S-L
18:	12.5102	2(2)	Y-L-N
19:	12.5102	2(2)	R-x-L-R
20:	12.5102	2(2)	E-R-E

20:	12.5102	2(2)	E-R-E
21:	12.5102	2(2)	A-x-L-A
22:	12.5102	2(2)	S-H-x-N
23:	12.5102	2(2)	N-Y-x-H
24:	12.5102	2(2)	N-Y-x(2)-D
25:	12.5102	2(2)	M-x(2)-S-H
26:	12.5102	2(2)	I-x(2)-V-S
27:	12.5102	2(2)	V-A-x-M
28:	12.5102	2(2)	V-x-I-x-D
29:	12.5102	2(2)	L-x-T-T
30:	12.5102	2(2)	S-L-x-L
31:	12.5102	2(2)	I-x-L-x-L
32:	12.5102	2(2)	A-x-S-L
33:	12.5102	2(2)	A-S-x-A
34:	12.5102	2(2)	P-A-P
35:	12.5102	2(2)	S-P-A
36:	12.5102	3(2)	P-x-S-x-A
37:	12.5102	2(2)	A-P-x-G
38:	12.5102	2(2)	A-x-T-G
39:	12.5102	2(2)	S-x-A-x-T
40:	12.5102	2(2)	L-T-x-L

TEIRESIAS

```
000001: 3 3 -7.609923 S..LA
000002: 3 2 -8.292739 P.S.A
000003: 2 2 -7.459193 LSL
000004: 2 2 -8.353868 ERE
000005: 2 2 -8.728578 FFS
000006: 2 2 -8.588881 YLN
000007: 2 2 -7.955048 PLT
000008: 2 2 -8.187221 ILI
000009: 2 2 -8.383854 PAP
000010: 2 2 -13.654320 ER.LR
000011: 2 2 -8.305680 KL.N
000012: 2 2 -8.335152 AP.G
000013: 2 2 -16.246918 M..SH.N
000014: 2 2 -8.835074 NT.R
000015: 2 2 -9.058732 VA.M
000016: 2 2 -8.757319 FG.T
000017: 2 2 -7.881617 GI.L
000018: 2 2 -7.398973 LT.L
000019: 2 2 -13.191239 SA.LA
000020: 2 2 -23.969492 S..LA.SL.L
```

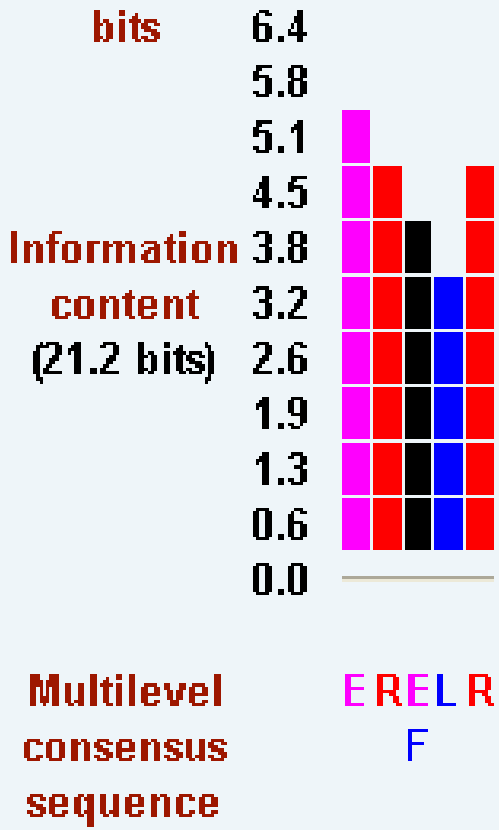
```
000021: 2 2 -13.945793 G.II.I
000022: 2 2 -13.433773 S..AS.G
000023: 2 2 -7.898198 AS.A
000024: 2 2 -10.179206 NY.H
000025: 2 2 -8.194163 AS..I
000026: 2 2 -7.914737 ASS
000027: 2 2 -7.779013 VL..S
000028: 2 2 -8.878059 R..TT
000029: 2 2 -9.321776 NY..D
000030: 2 2 -9.077474 HL..F
000031: 2 2 -7.661864 KL..L
000032: 2 2 -8.008338 A.TG
000033: 2 2 -8.549232 S.AN
000034: 2 2 -8.183880 T.VA
000035: 2 2 -8.224718 L.TT
000036: 2 2 -8.563931 S.IN
000037: 2 2 -9.649038 M.NT
000038: 2 2 -7.703485 I.L.L
000039: 2 2 -7.824456 L.I.L
000040: 2 2 -8.059701 SPA
```

SPEXS

1.	S..{0,2}I.L	3/3
2.	A.{0,2}A..G	3/3
3.	S..{0,2}S.G	3/3
4.	S..LA	3/3
5.	V.{0,2}L..S	3/3
6.	S..{0,2}A.T	3/3
7.	S..{0,2}A.S	3/3
8.	S.{0,2}L..S	3/3
9.	S.{0,2}A..G	3/3
10.	A..{0,2}L.L	3/3
11.	L..{0,2}L.L	3/4
12.	L.{0,2}L..F	3/3
13.	G.{0,2}I..I	2/3
14.	P..{0,2}S.A	2/3
15.	S..{0,2}T.T	2/3
16.	T..{0,2}S.A	2/3
17.	I..{0,2}L.L	2/3
18.	LA.{0,2}L.L	2/3

Submit all patterns to PATMATCH

MEME



NAME	START	P-VALUE	SITES		
B0017.SEQ=m52(16177	3	1.94e-07	VP	ERELR	LFYYLNCLSL
B0002.SEQ=m52(337>2799);	15	4.14e-07	KFGGTSVANA	ERFLR	V

MEME

NAME	START	P-VALUE	SITES
B0020.SEQ=m52(18715>1962	6	1.35e-07	MSMSH I NYNH LYYFWHVYKE
B0016.SEQ=m52(15445>1655	1	9.37e-07	MNYSH DNWSAILAHI

NAME	START	P-VALUE	SITES
B0004.SEQ=m52(3734>5020)	7	3.47e-07	MKLYNL KDHNE QVSFAQAVT
B0015.SEQ=m52(14168>1529	14	1.89e-06	QDYYEILGVS KTAEE RE

NAME	START	P-VALUE	SITES
B0016.SEQ=m52(15445>1655	16	2.95e-07	DNWSAILAHI GKPEE
B0018.SEQ=m52(16960	12	6.39e-06	LNTCRVPLTD RKVKE KRAM
B0011.SEQ=m52(11356	16	9.26e-06	LNDSDLDFLQ HCSEE

Kokkuvõte

- TEIRESIAS tundus andvat täpseima tulemuse
- MEME kulutas palju aega, samas väljund informatiivsem. NB! Ei luba gäppe
- Pratt annab palju võimalusi parameetrite valikuks, kuid veebivariant ei tundu neid kõiki toetavat.

Kasutatud materjalid

- Brejovà *et al*. (2000) Finding patterns in biological sequences.
- Rigoutsos I and Floratos A (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics 14: 55- 67.
- I.Jonassen, J.F.Collins, D.G.Higgins. (1995) Finding flexible patterns in unaligned protein sequences. Protein Science 4, 1587 -1595
- Timothy L. Bailey and Charles Elkan, The value of prior knowledge in discovering motifs with MEME, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pp. 21-29, AAAI Press, Menlo Park, California, 1995
- Jaak Vilo. (2002) Pattern Discovery from Biosequences PhD Thesis, Department of Computer Science, University of Helsinki, Finland