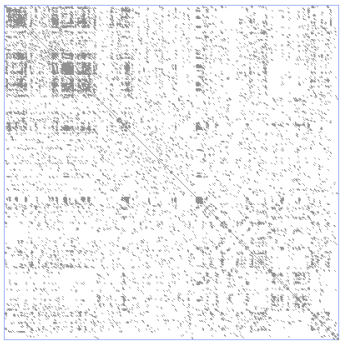
Praktične vaje pri predmetu
Biokemijska informatika

7. vaja

Prileganje zaporedij



		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Miha Pavšič
april 2014

Metode za prileganje zaporedij
Metode

Prileganje dveh zaporedij (pairwise sequence alignment):

- točkovni diagram
- dinamično programiranje
- metode besed oz. nizov

Prileganje večih zaporedij (*multiple sequence alignment*):

- dinamično programiranje
- progresivne metode
- iteracijske metode
- iskanje motivov
- druge metode (HMM, genetski algoritmi, ...)

Pregled vaje

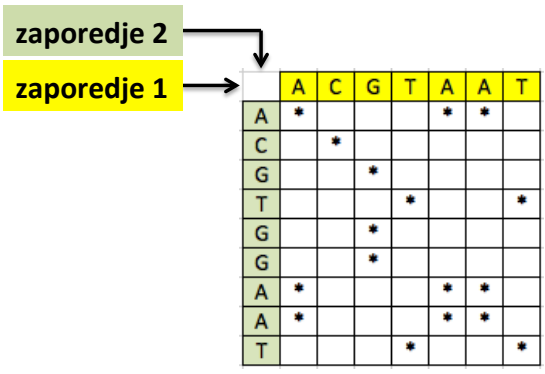
Prileganje dveh zaporedij:

- točkovni diagram (*dot plot*) → Excel in spletna orodja
- lokalno in globalno prileganje:
  - algoritmi (dinamično programiranje):
    - Needleman-Wunsch (globalno) → Excel
  - praktični primeri uporabe → spletna orodja

Navodila za vse naloge so v obliki kviza v spletni učilnici.

Točkovni diagram (*dot plot, dot-matrix plot*)

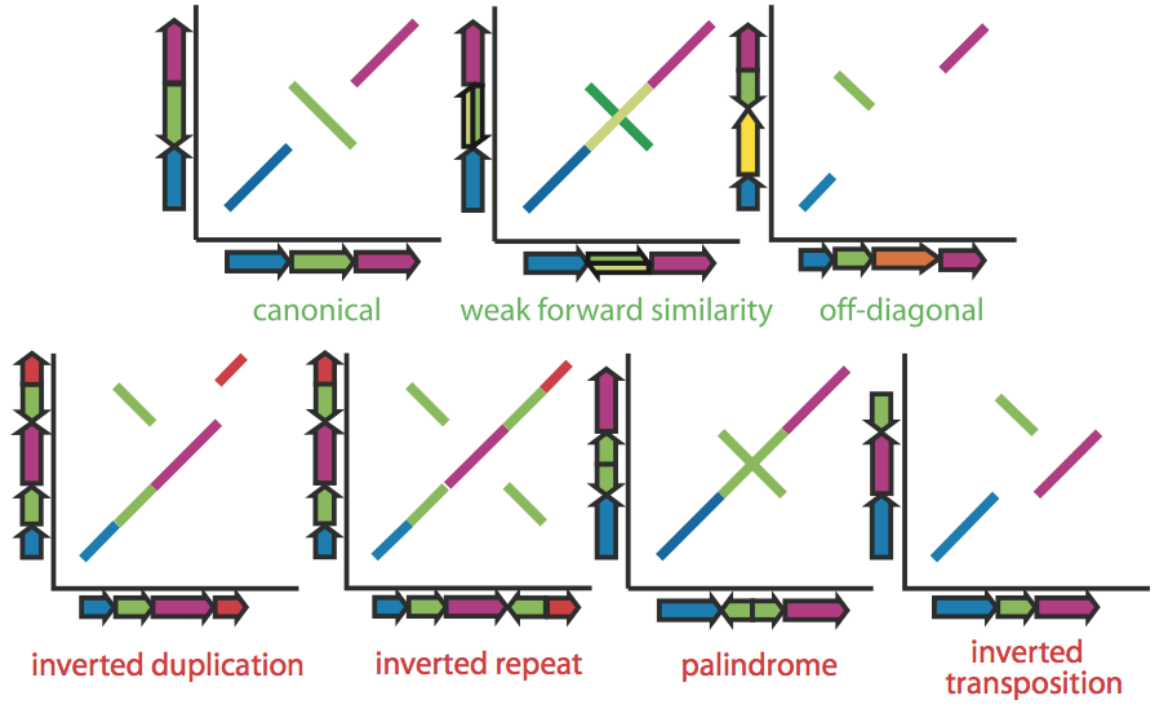
- najstarejša metoda primerjave dveh zaporedij:
  - koncept je enostaven
  - kvalitativne narave
  - analiza večjih vzorcev / v večjem merilu je zamudna
- dobimo celo družino prileganj posameznih regij zaporedij, ki ju primerjamo
- vizualna identifikacija insercij, delecij, ponovitev in obratnih ponovitev
- ponovitve na enostaven način identificiramo tudi, če zaporedje primerjamo s samim seboj



* pomeni identičnost

- najbolj enostaven primer: točka za identičnost, prazno polje za razliko (primer zgoraj)
- kompleksnejši primer: velikost/barva točke je povezana s stopnjo podobnosti (npr. E in D sta si bolj podobna kot E in K) – upoštevamo matriko zamenjav (*substitution matrix*)
- precej šuma → zmanjšamo ga z uporabo okna (primerjamo dva niza določene velikosti)

Točkovni diagram – primeri nt zaporedij



Vir: 10.1073/pnas.0603984103

Globalno in lokalno prileganje

- pri **globalnem** prileganju primerjamo dve zaporedji po njuni **celotni dolžini** → globalno prileganje je primerno za primerjavo dveh zaporedij, ki:
  - sta podobni po njuni celotni dolžini
  - sta približno enako dolgi
- **lokalno** prileganje je za razliko od globalnega prileganja primernejše za **identifikacijo podobnih regij** v dveh zaporedij
  - za podobne regije ni nujno, da si sledijo v obeh zaporedjih v enakem vrstnem redu

```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
  |  || |  ||  | | | |||     || |  | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

globalno prileganje

```
tccCAGTTATGTCAGgggacacgagcatgcagagac
   ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

lokalno prileganje

Algoritem Needleman-Wunsch

za globalno prileganje dveh zaporedij

Osnovni koraki:

1. priprava inicializacijske matrike
2. priprava seštevalne matrike
3. sledenje v seštevalni matriki in izpis prileganja

		A	C	T	T	A	T	C	A
	C	0	1	0	0	0	0	1	0
	T	0	0	1	1	0	1	0	0
	T	0	0	1	1	0	1	0	0
	G	0	0	0	0	0	0	0	0
	A	1	0	0	0	1	0	0	1
	T	0	0	1	1	0	1	0	0
	C	0	1	0	0	0	0	1	0
	A	1	0	0	0	1	0	0	1

inicializacija

		A	C	T	T	A	T	C	A
	C								
	T								
	T								
	G								
	A					4	2	1	1
	T					2	3	1	0
	C					1	1	2	0
	A					1	0	0	1

seštevanje

		A	C	T	T	A	T	C	A
	C	6	7	5	4	3	2	2	0
	T	5	5	6	5	3	3	1	0
	T	4	4	5	5	3	3	1	0
	G	4	4	4	4	3	2	1	0
	A	4	3	3	3	4	2	1	1
	T	2	2	3	3	2	3	1	0
	C	1	2	1	1	1	1	2	0
	A	1	0	0	0	1	0	0	1

sledenje

V nadaljevanju si bomo pogledali osnoven algoritem,
ki ne upošteva neke posebne matrike zamenjav.

Algoritem Needleman-Wunsch

KORAK 1: Priprava inicializacijske matrike

Pripravimo matriko, ki je podobna točkovnemu diagramu.
Za enakost v našem primeru napišemo 1, za neenakost pa 0.

Zaporedje 1: **`ACTTATCA`**
Zaporedje 2: **`CTTGATCA`**

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

		A	C	T	T	A	T	C	A
	C	0	1	0	0	0	0	1	0
	T	0	0	1	1	0	1	0	0
	T	0	0	1	1	0	1	0	0
	G	0	0	0	0	0	0	0	0
	A	1	0	0	0	1	0	0	1
	T	0	0	1	1	0	1	0	0
	C	0	1	0	0	0	0	1	0
	A	1	0	0	0	1	0	0	1

Pri izdelavi inicializacijske matrike bi v kompleksnejšem primeru uporabili matriko zamenjav (*substitution matrix*), prav tako pa bi upoštevali različne vrednost za vrzeli (za slednje v naslednjem koraku – seštevanje).

Algoritem Needleman-Wunsch

KORAK 2: Seštevalna matrika

Na **posameznih mestih** seštevalne matrike:

- vrednost na **ekvivalentem mestu v inicializacijski matriki** in
- največjo vrednost v **stolpcu nižje in desno ter vrstici nižje in desno v seštevalni matriki**.

	A	C	T	T	A	T	C	A
C	0	1	0	0	0	0	1	0
T	0	0	1	1	0	1	0	0
T	0	0	1	1	0	1	0	0
G	0	0	0	0	0	0	0	0
A	1	0	0	0	1	0	0	1
T	0	0	1	1	0	1	0	0
C	0	1	0	0	0	0	1	0
A	1	0	0	0	1	0	0	1

inicializacijska matrika

	A	C	T	T	A	T	C	A
C								
T								
T								
G								
A					4	2	1	1
T					2	3	1	0
C					1	1	2	0
A					1	0	0	1

seštevalna matrika

vrednost v celici seštevalne matrike: $M_{ij} = M_{ij}^{*} + \max(M_{k, j+1}, M_{i+1, l})$

M_{ij}^{*} je celica v inicializacijski matriki $k > i, l > j$

V prikazanem primeru je vrednost za vrzel enaka 0. V primeu ne-ničelne vrednosti (je pa <0) bi slednjo prišteli pri pomiku navzdol ali navzgor.

Algoritem Needleman-Wunsch

KORAK 3: Sledenje v seštevalni matriki

	A	C	T	T	A	T	C	A
C	6	7	5	4	3	2	2	0
T	5	5	6	5	3	3	1	0
T	4	4	5	5	3	3	1	0
G	4	4	4	4	3	2	1	0
A	4	3	3	3	4	2	1	1
T	2	2	3	3	2	3	1	0
C	1	2	1	1	1	1	2	0
A	1	0	0	0	1	0	0	1

tu začnemo

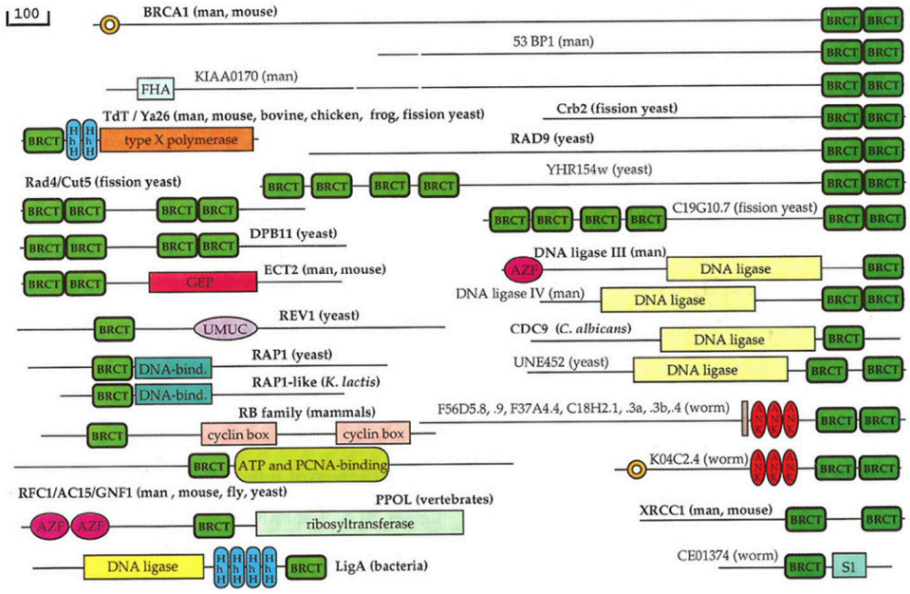
→

```
ACTT-ATCA
-CTTGATCA
```

- premiki po **diagonali** pomenijo **enakost**
- premiki **gor ali dol** pomenijo **vrzel**

Lokalno prileganje

Lokalno prileganje pride v poštev npr. pri primerjavi zaporedij proteinov, ki vsebujejo homologne domene, ki pa niso nujno enako razvrščene, se ne ponovijo v enakem številu, lahko pa so prisotne tudi druge (sicer različne) domene.



Matrike zamenjav (substitucijske matrike, substitution matrices)

- Z njimi opišemo frekvenco zamenjav nukleotidov ali aminokislinskih ostankov.
- Za proteine se pogosto uporablja matrika **BLOSUM62**.

NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 3 MARCH 2008

BLOSUM62 miscalculations improve search performance

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

BLOSUMXX

XX = za pripravo matrice so bila vsa zaporedja, ki so >80% identična, združena v eno, s čimer se zmanjša prispevek zelo podobnih zaporedij; **manjši XX → primerno za primerjavo manj podobnih zaporedij**