Bioinformatika és genomanalízis az orvostudományban

Szekvenciaillesztés

Cserző Miklós

2020

https://semmelweis.zoom.us/j/96102872458?pwd=Rk1PL2tqS21sdlUwc3B4eDFCZkNKQT09

A mai előadás

- A szekvencia illesztés elméleti alapja
- Az AAindex adatbázis, aminósavak hasonlósága
- Pontozómátrixok
- Az illesztőfelület
- Needlemann-Wunsch algoritmus
- Smith–Waterman algoritmus
- Résbüntetési sémák

Az evolúció molekuláris modellje

- Sorozatos másolási hibák sejtosztódáskor
  - Pontmutációk
  - Beszúrás
  - Törlés
- Minden köztes állapot életképes
- A végső állapotban a hibák felhalmozva jelennek meg

```
ANCIENTVARIANT
ANCIENTVARYANT
ANCIENTVARYAN-
ANCIENTVATYAN-
ANCIENT-ATYAN-
ENCIENT-ATYAN-
-ECIENT-ATYAN-
-ECIENT--TYAN-
-ECIENT--TYAE-
-ECIENT--TYPE-
RECIENT--TYPE-
REC-ENT--TYPE-
```

A probléma bionformatikai szempontból

- Csak az evolúciós folyamat végeredményét látjuk – a köztes lépéseket nem:
- Két vagy több, közös őstől származó modern szekvencia
- A kihalt változatok nem érhetők el
- Mi annak a valószínűsége, hogy két adott szekvencia közös őstől származik?

Két professzor játszik

- Mondok egy szekvencia részletet, találd ki melyikről van szó.
- Inkább kezdd el mondani a részletet betűnként és megállítalak, ha kitaláltam.
- Az nyer, aki rövidebb részlet alapján kitalálja.
- Csak igazi szekvenciát ér mondani!

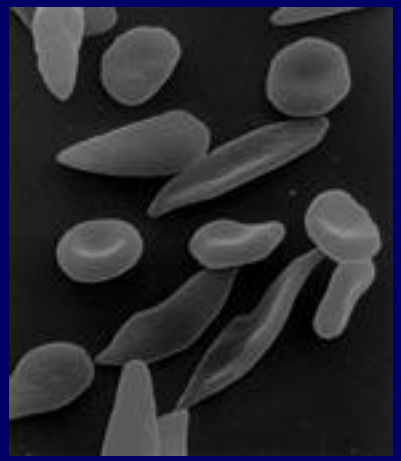
Mi legyen a játék stratégiája?

Megfontolások DNS (RNS) esetén:

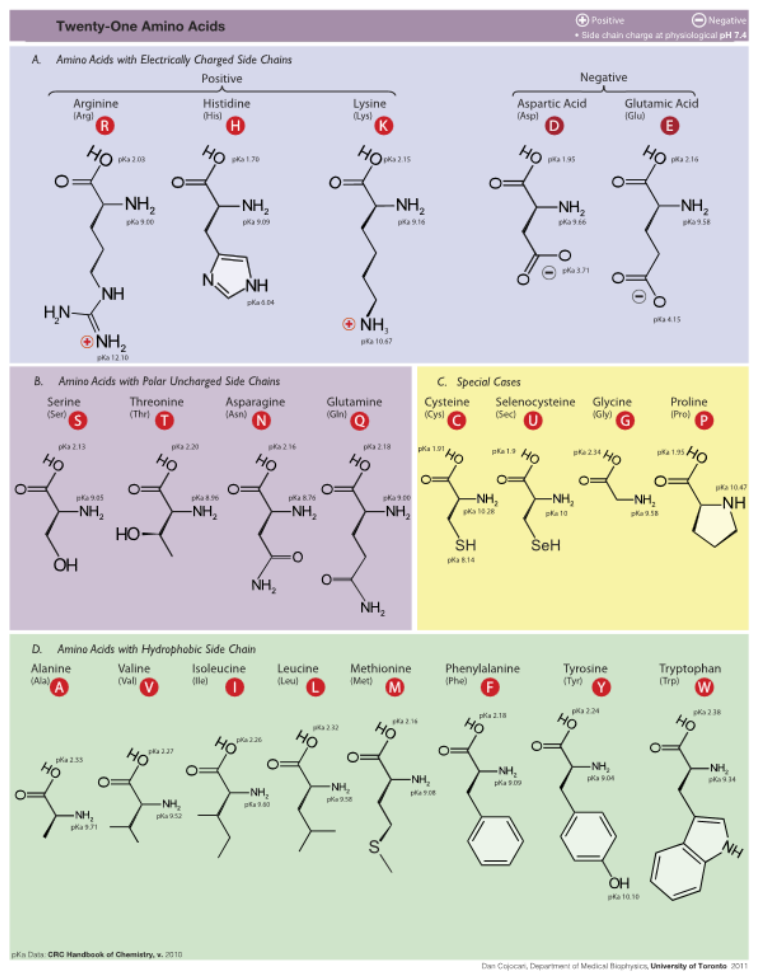
- 4 betűs ABC (2011, bin: 11111011011)
- Egy *n* hosszú fragmensből 4^n változat lehetséges
- DNS: A pontmutáció nem változtatja meg lényegesen a szerkezetet, stabilitást – minden helyettesítés ugyan olyan jó
- Ez RNS-re nem áll!

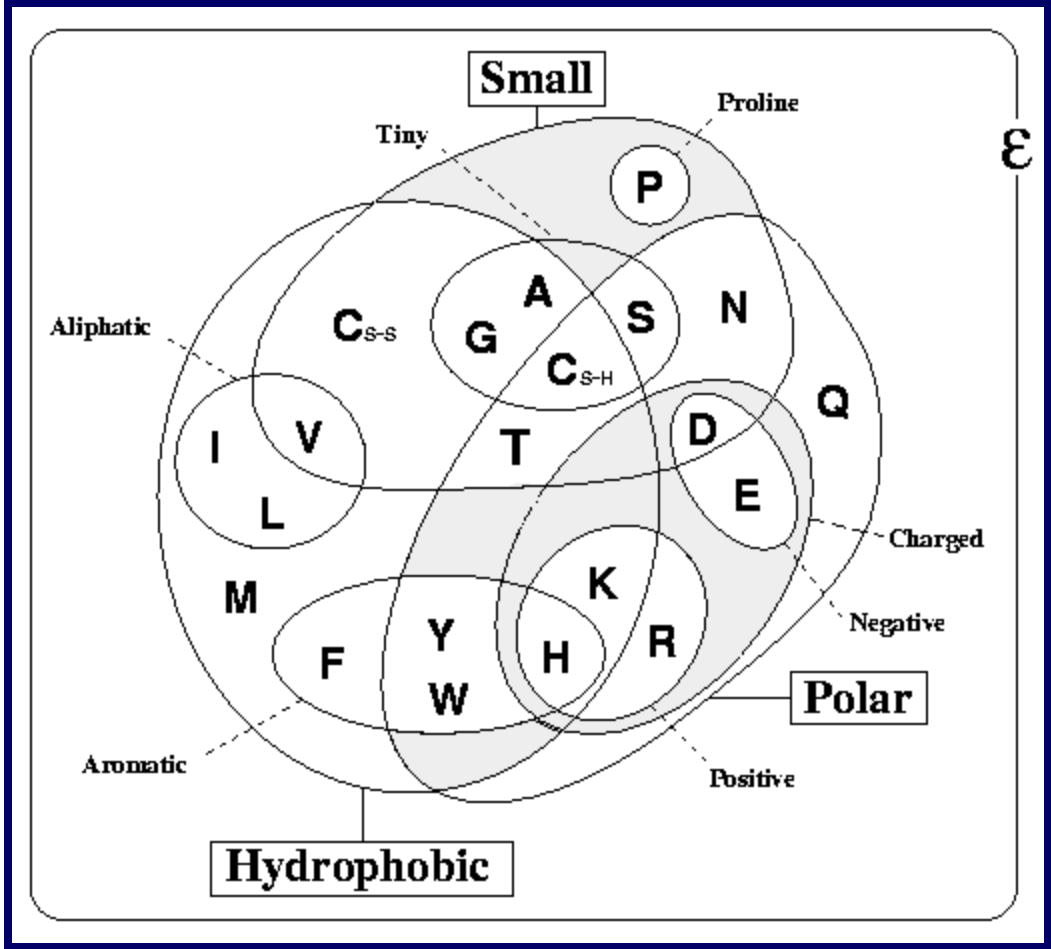
Fehérje esetén

- 20 betűs ABC (2011, hex: 7DB)
- 20^n számú változat
- A térszerkezet kitüntetetten fontos
- Az aminósavak eltérő mértékben képesek helyettesíteni egymást
- Egyetlen rossz csere tönkreteszi az egész fehérjét



Aminósavak tulajdonságai



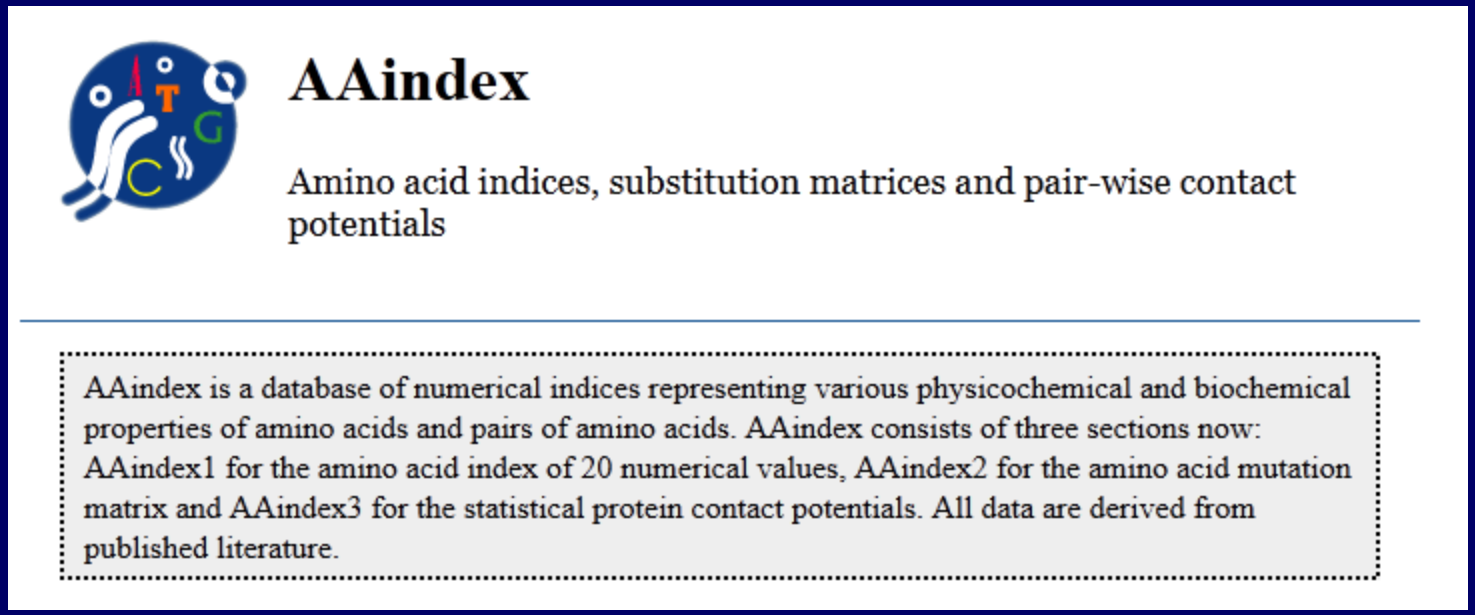


A hidrofóbicitás

- A fehérje térszerkezet kialakulásának hajtóereje, stabilitásának kulcsa
- Komplex jelenség: rengeteg mikroszkópikus erő makroszkópikus eredője
- Lehet mérni: megoszlási hányados, reverz fázisú kromatográfia
- Viszont csak mesterséges körülmények között

Az AAindex adatbázis

- Az aminósavak tulajdonságai egybegyűjtve: http://www.genome.jp/aaindex/



AAindex

Amino acid indices, substitution matrices and pair-wise contact potentials

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. AAindex consists of three sections now: AAindex1 for the amino acid index of 20 numerical values, AAindex2 for the amino acid mutation matrix and AAindex3 for the statistical protein contact potentials. All data are derived from published literature.

```
List of 544 Amino Acid Indices in AAindex ver.9.1

The columns correspond to the AAindex accession number and the description of
each index.

ANDN920101 alpha-CH chemical shifts (Andersen et al., 1992)
ARGP820101 Hydrophobicity index (Argos et al., 1982)
ARGP820102 Signal sequence helical potential (Argos et al., 1982)
ARGP820103 Membrane-buried preference parameters (Argos et al., 1982)
BEGF750101 Conformational parameter of inner helix (Beghin-Dirkx, 1975)
BEGF750102 Conformational parameter of beta-structure (Beghin-Dirkx, 1975)
BEGF750103 Conformational parameter of beta-turn (Beghin-Dirkx, 1975)
BHAR880101 Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)
BIGC670101 Residue volume (Bigelow, 1967)
BIOV880101 Information value for accessibility; average fraction 35% (Biou et al., 1988)
BIOV880102 Information value for accessibility; average fraction 23% (Biou et al., 1988)
BROC820101 Retention coefficient in TFA (Browne et al., 1982)
BROC820102 Retention coefficient in HFBA (Browne et al., 1982)
BULH740101 Transfer free energy to surface (Bull-Breese, 1974)
BULH740102 Apparent partial specific volume (Bull-Breese, 1974)
BUNA790101 alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
BUNA790102 alpha-CH chemical shifts (Bundi-Wuthrich, 1979)
BUNA790103 Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
BURA740101 Normalized frequency of alpha-helix (Burgess et al., 1974)
BURA740102 Normalized frequency of extended structure (Burgess et al., 1974)
CHAM810101 Steric parameter (Charton, 1981)
CHAM820101 Polarizability parameter (Charton-Charton, 1982)
CHAM820102 Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
CHAM830101 The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)
CHAM830102 A parameter defined from the residuals obtained from the best correlation of
CHAM830103 The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)
CHAM830104 The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)
CHAM830105 The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)
CHAM830106 The number of bonds in the longest chain (Charton-Charton, 1983)
CHAM830107 A parameter of charge transfer capability (Charton-Charton, 1983)
CHAM830108 A parameter of charge transfer donor capability (Charton-Charton, 1983)
CHOC750101 Average volume of buried residue (Chothia, 1975)
```

GenomeNet

Database: AAindex
Entry: CHAM810101
LinkDB: CHAM810101

```
H CHAM810101
D Steric parameter (Charton, 1981)
R LIT:2004112b PMID:7300379
A Charton, M.
T Protein folding and the genetic code: An alternative quantitative model
J J. Theor. Biol. 91, 115-123 (1981) (Pro !)
C FAUJ880102    0.881  LEVM760104   -0.818  KIMC930101   -0.848
I    A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
    0.52    0.68    0.76    0.76    0.62    0.68    0.68    0.00    0.70    1.02
    0.98    0.68    0.78    0.70    0.36    0.53    0.50    0.70    0.70    0.76
//
```

DBGET integrated database retrieval system

GenomeNet

Database: AAindex
Entry: BIGC670101
LinkDB: BIGC670101

```
H BIGC670101
D Residue volume (Bigelow, 1967)
R LIT:2004108b PMID:6048539
A Bigelow, C.C.
T On the average hydrophobicity of proteins and the relation between it and
  protein structure
J J. Theor. Biol. 16, 187-211 (1967) (Asn Gln 5.0)
C GOLD730102    1.000  KRIW790103    0.993  TSAJ990101    0.993
  TSAJ990102    0.992  CHOC750101    0.990  GRAR740103    0.984
  FAUJ880103    0.972  CHAM820101    0.966  HARY940101    0.960
  CHOC760101    0.960  PONJ960101    0.950  FASG760101    0.919
  LEVM760105    0.913  ROSG850101    0.910  DAWD720101    0.903
  LEVM760102    0.896  ZHOH040102    0.884  LEVM760106    0.876
  CHAM830106    0.870  LEVM760107    0.863  FAUJ880106    0.860
  RADA880106    0.856  MCMT640101    0.814  RADA880103   -0.865
I    A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
    52.6   109.1    75.7    68.4    68.3    89.7    84.7    36.3    91.9   102.0
   102.0   105.1    97.7   113.9    73.6    54.9    71.2   135.4   116.2    85.1
//
```

DBGET integrated database retrieval system

```
List of 94 Amino Acid Matrices in AAindex ver.9.1

The columns correspond to the AAindex accession number and the description of
each matrix.

ALTS910101 The PAM-120 matrix (Altschul, 1991)
BENS940101 Log-odds scoring matrix collected in 6.4-8.7 PAM (Benner et al., 1994)
BENS940102 Log-odds scoring matrix collected in 22-29 PAM (Benner et al., 1994)
BENS940103 Log-odds scoring matrix collected in 74-100 PAM (Benner et al., 1994)
BENS940104 Genetic code matrix (Benner et al., 1994)
CSEM940101 Residue replace ability matrix (Cserzo et al., 1994)
DAYM780301 Log odds matrix for 250 PAMs (Dayhoff et al., 1978)
FEND850101 Structure-Genetic matrix (Feng et al., 1985)
FITW660101 Mutation values for the interconversion of amino acid pairs (Fitch, 1966)
GEOD900101 Hydrophobicity scoring matrix (George et al., 1990)
GONG920101 The mutation matrix for initially aligning (Gonnet et al., 1992)
GRAR740104 Chemical distance (Grantham, 1974)
HENS920101 BLOSUM45 substitution matrix (Henikoff-Henikoff, 1992)
HENS920102 BLOSUM62 substitution matrix (Henikoff-Henikoff, 1992)
HENS920103 BLOSUM80 substitution matrix (Henikoff-Henikoff, 1992)
JOHM930101 Structure-based amino acid scoring table (Johnson-Overington, 1993)
JOND920103 The 250 PAM PET91 matrix (Jones et al., 1992)
JOND940101 The 250 PAM transmembrane protein exchange matrix (Jones et al., 1994)
KOLA920101 Conformational similarity weight matrix (Kolaskar-Kulkarni-Kale, 1992)
LEVJ860101 The secondary structure similarity matrix (Levin et al., 1986)
LUTR910101 Structure-based comparison table for outside other class (Luthy et al., 1991)
LUTR910102 Structure-based comparison table for inside other class (Luthy et al., 1991)
LUTR910103 Structure-based comparison table for outside alpha class (Luthy et al., 1991)
LUTR910104 Structure-based comparison table for inside alpha class (Luthy et al., 1991)
LUTR910105 Structure-based comparison table for outside beta class (Luthy et al., 1991)
LUTR910106 Structure-based comparison table for inside beta class (Luthy et al., 1991)
LUTR910107 Structure-based comparison table for other class (Luthy et al., 1991)
LUTR910108 Structure-based comparison table for alpha helix class (Luthy et al., 1991)
LUTR910109 Structure-based comparison table for beta strand class (Luthy et al., 1991)
MCLA710101 The similarity of pairs of amino acids (McLachlan, 1971)
MCLA720101 Chemical similarity scores (McLachlan, 1972)
MIYS930101 Base-substitution-protein-stability matrix (Miyazawa-Jernigan, 1993)
```

GenomeNet

Database: AAindex
Entry: GRAR740104
LinkDB: GRAR740104

```
H GRAR740104
D Chemical distance (Grantham, 1974)
R LIT:2004143 PMID:4843792
A Grantham, R.
T Amino acid difference formula to help explain protein evolution
J Science 185, 862-864 (1974)
M rows = ARNDCQEGHILKMFPSTWYV, cols = ARNDCQEGHILKMFPSTWYV
      0.
    112.      0.
    111.     86.      0.
    126.     96.     23.      0.
    195.    180.    139.    154.      0.
     91.     43.     46.     61.    154.      0.
    107.     54.     42.     45.    170.     29.      0.
     60.    125.     80.     94.    159.     87.     98.      0.
     86.     29.     68.     81.    174.     24.     40.     98.      0.
     94.     97.    149.    168.    198.    109.    134.    135.     94.      0.
     96.    102.    153.    172.    198.    113.    138.    138.     99.      5.      0.
    106.     26.     94.    101.    202.     53.     56.    127.     32.    102.    107.      0.
     84.     91.    142.    160.    196.    101.    126.    127.     87.     10.     15.     95.      0.
    113.     97.    158.    177.    205.    116.    140.    153.    100.     21.     22.    102.     28.      0.
     27.    103.     91.    108.    169.     76.     93.     42.     77.     95.     98.    103.     87.    114.      0.
     99.    110.     46.     65.    112.     68.     80.     56.     89.    142.    145.    121.    135.    155.     74.      0.
     58.     71.     65.     85.    149.     42.     65.     59.     47.     89.     92.     78.     81.    103.     38.     58.      0.
    148.    101.    174.    181.    215.    130.    152.    184.    115.     61.     61.    110.     67.     40.    147.    177.    128.      0.
    112.     77.    143.    160.    194.     99.    122.    147.     83.     33.     36.     85.     36.     22.    110.    144.     92.     37.      0.
     64.     96.    133.    152.    192.     96.    121.    109.     84.     29.     32.     97.     21.     50.     68.    124.     69.     88.     55.      0.
//
```

DBGET integrated database retrieval system

A PAM matrix sorozat

- PAM: Percentage Accepted Mutations
- Dayhoff (1978), 1572 megfigyelt pontnutáció 71 közeli rokonságban álló fehérjecsaládon
- PAM1: 1%-os mutációs hatás éri a szekvenciát
- PAM2: egymás után kétszer 1-1%-os mutációs hatás – mátrix szorzás
- Stb.. Egészen PAM250-ig

GenomeNet

Database: AAindex
Entry: DAYM780301
LinkDB: DAYM780301

```
H DAYM780301
D Log odds matrix for 250 PAMs (Dayhoff et al., 1978)
R
A Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C.
T A model of evolutionary change in proteins
J In "Atlas of Protein Sequence and Structure", Vol.5, Suppl.3 (Dayhoff,
  M.O., ed.), National Biomedical Research Foundation, Washington, D.C.,
  p.352 (1978)
M rows = ARNDCQEGHILKMFPSTWYV, cols = ARNDCQEGHILKMFPSTWYV
      2.
     -2.      6.
      0.      0.      2.
      0.     -1.      2.      4.
     -2.     -4.     -4.     -5.     12.
      0.      1.      1.      2.     -5.      4.
      0.     -1.      1.      3.     -5.      2.      4.
      1.     -3.      0.      1.     -3.     -1.      0.      5.
     -1.      2.      2.      1.     -3.      3.      1.     -2.      6.
     -1.     -2.     -2.     -2.     -2.     -2.     -2.     -3.     -2.      5.
     -2.     -3.     -3.     -4.     -6.     -2.     -3.     -4.     -2.      2.      6.
     -1.      3.      1.      0.     -5.      1.      0.     -2.      0.     -2.     -3.      5.
     -1.      0.     -2.     -3.     -5.     -1.     -2.     -3.     -2.      2.      4.      0.      6.
     -4.     -4.     -4.     -6.     -4.     -5.     -5.     -5.     -2.      1.      2.     -5.      0.      9.
      1.      0.     -1.     -1.     -3.      0.     -1.     -1.      0.     -2.     -3.     -1.     -2.     -5.      6.
      1.      0.      1.      0.      0.     -1.      0.      1.     -1.     -1.     -3.      0.     -2.     -3.      1.      2.
      1.     -1.      0.      0.     -2.     -1.      0.      0.     -1.      0.     -2.      0.     -1.     -3.      0.      1.      3.
     -6.      2.     -4.     -7.     -8.     -5.     -7.     -7.     -3.     -5.     -2.     -3.     -4.      0.     -6.     -2.     -5.     17.
     -3.     -4.     -2.     -4.      0.     -4.     -4.     -5.      0.     -1.     -1.     -4.     -2.      7.     -5.     -3.     -3.      0.     10.
      0.     -2.     -2.     -2.     -2.     -2.     -2.     -1.     -2.      4.      2.     -2.      2.     -1.     -1.     -1.      0.     -6.     -2.      4.
//
```

DBGET integrated database retrieval system

A BLOSUM mátrix sorozat

- **BLO**cks of Amino Acid **SU**bstitution **M**atrix
- Block: rokon fehérjék toldás nélkül illeszett szakaszai
- Több mint 2000 block-ból számoltak
- Több mint 500 fehérjecsaládból indultak ki
- A sorozatban minden fehérjét bevettek, amelyek rokonsági foka meghaladt egy bizonyos mértéket

GenomeNet

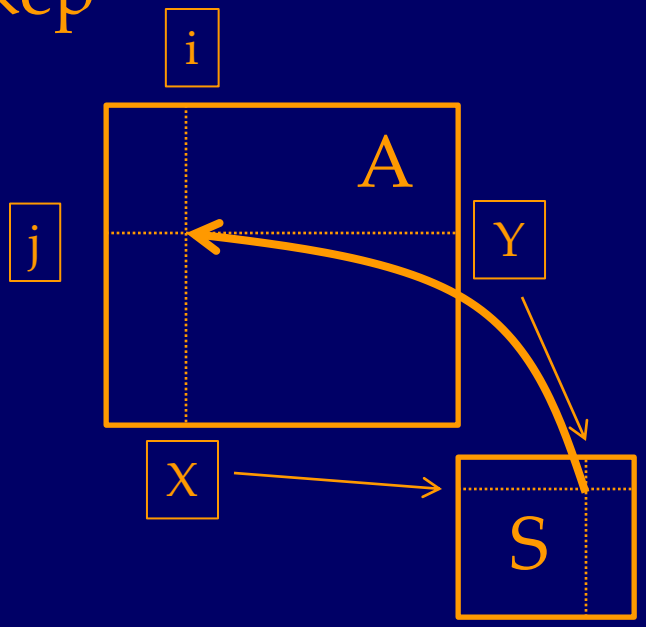
Database: AAindex
Entry: HENS920101
LinkDB: HENS920101

```
H HENS920101
D BLOSUM45 substitution matrix (Henikoff-Henikoff, 1992)
R LIT:1902106 PMID:1438297
A Henikoff, S. and Henikoff, J.G.
T Amino acid substitution matrices from protein blocks
J Proc. Natl. Acad. Sci. USA 89, 10915-10919 (1992)
* matrix in 1/3 Bit Units
M rows = ARNDCQEGHILKMFPSTWYV, cols = ARNDCQEGHILKMFPSTWYV
      5.
     -2.      7.
     -1.      0.      6.
     -2.     -1.      2.      7.
     -1.     -3.     -2.     -3.     12.
     -1.      1.      0.      0.     -3.      6.
     -1.      0.      0.      2.     -3.      2.      6.
      0.     -2.      0.     -1.     -3.     -2.     -2.      7.
     -2.      0.      1.      0.     -3.      1.      0.     -2.     10.
     -1.     -3.     -2.     -4.     -3.     -2.     -3.     -4.     -3.      5.
     -1.     -2.     -3.     -3.     -2.     -2.     -2.     -3.     -2.      2.      5.
     -1.      3.      0.      0.     -3.      1.      1.     -2.     -1.     -3.     -3.      5.
     -1.     -1.     -2.     -3.     -2.      0.     -2.     -2.      0.      2.      2.     -1.      6.
     -2.     -2.     -2.     -4.     -2.     -4.     -3.     -3.     -2.      0.      1.     -3.      0.      8.
     -1.     -2.     -2.     -1.     -4.     -1.      0.     -2.     -2.     -2.     -3.     -1.     -2.     -3.      9.
      1.     -1.      1.      0.     -1.      0.      0.      0.     -1.     -2.     -3.     -1.     -2.     -2.     -1.      4.
      0.     -1.      0.     -1.     -1.     -1.     -1.     -2.     -2.     -1.     -1.     -1.     -1.     -1.     -1.      2.      5.
     -2.     -2.     -4.     -4.     -5.     -2.     -3.     -2.     -3.     -2.     -2.     -2.     -2.      1.     -3.     -4.     -3.     15.
     -2.     -1.     -2.     -2.     -3.     -1.     -2.     -3.      2.      0.      0.     -1.      0.      3.     -3.     -2.     -1.      3.      8.
      0.     -2.     -3.     -3.     -1.     -3.     -3.     -3.     -3.      3.      1.     -2.      1.      0.     -3.     -1.      0.     -3.     -1.      5.
//
```

DBGET integrated database retrieval system

Az illesztőfelület

- A két szekvencia minden részletét összevtjük a pontozómátrix alapján
- Az illesztőfelület egy mátrix – „domborzati térkép”



Példák:

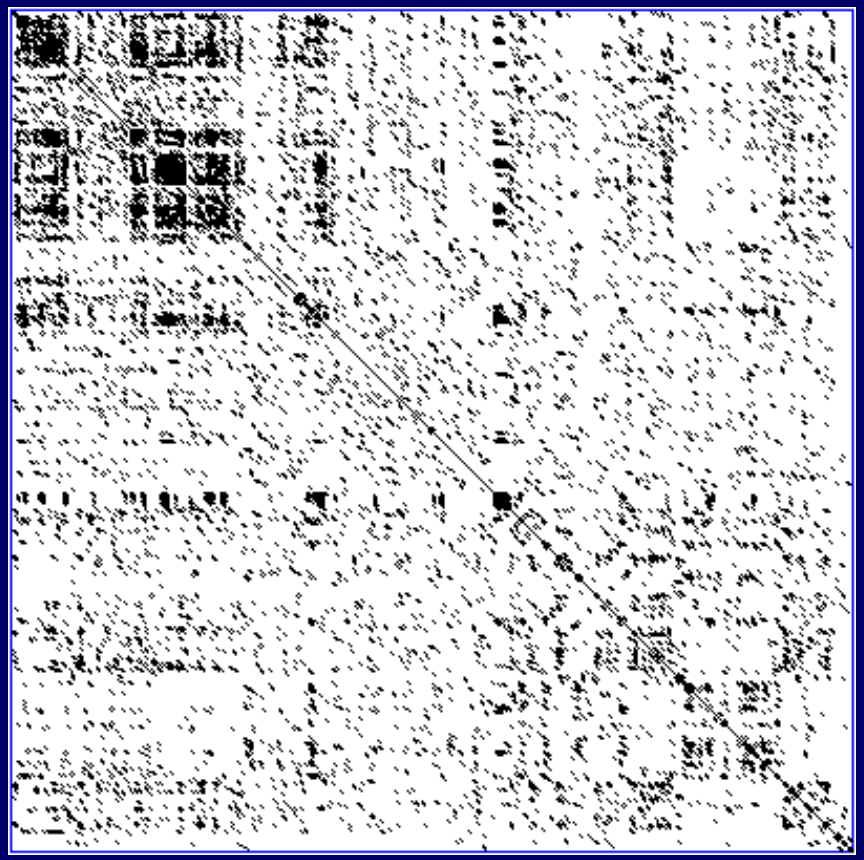
PAM10:

	A	N	C	I	E	N	T
R	-10	-9	-11	-8	-15	-9	-10
E	-5	-5	-20	-8	8	-5	-9
C	-10	-17	10	-9	-20	-17	-11
E	-5	-5	-20	-8	8	-5	-9
N	-7	9	-17	-8	-5	9	-5
T	-3	-5	-11	-5	-9	-5	8

Blosum30:

	A	N	C	I	E	N	T
R	-1	-2	-2	-3	-1	-2	-3
E	0	-1	1	-3	6	-1	-2
C	-3	-1	17	-2	1	-1	-2
E	0	-1	1	-3	6	-1	-2
N	0	8	-1	0	-1	8	1
T	1	1	-2	0	-2	1	5

A DotPlot

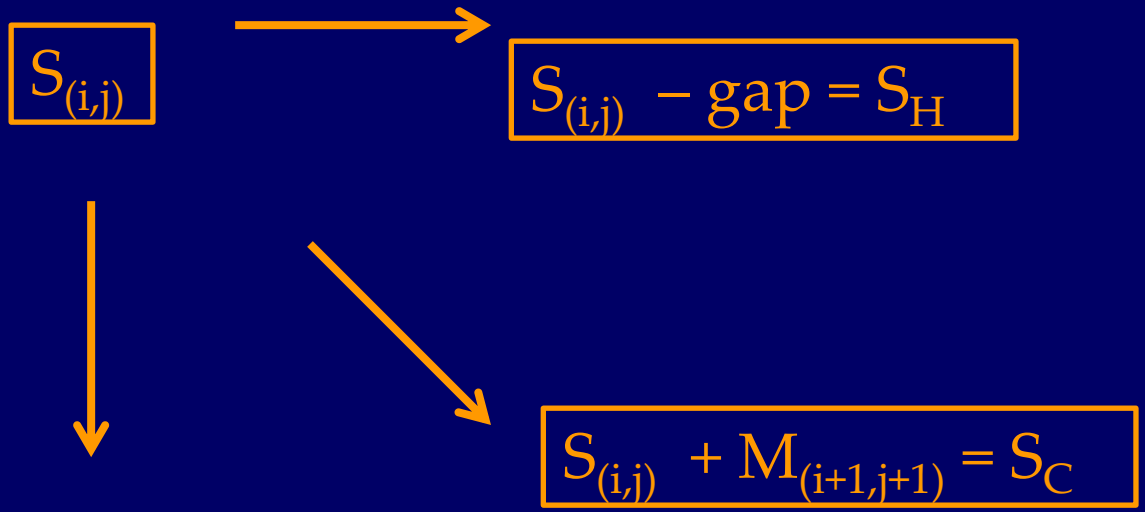


- Az illesztőfelület elemeit pontok jelölik
- Ha a pontszám nagyobb, mint egy megadott érték – a pont fekete
- A hasonló szakaszokat a diagonálissal párhuzamos pontsorok jelzik

Needlemann-Wunsch algoritmus

- A felület minden pontja uj illesztés kiindulópontja
- A széleket nem számítva minden pontbol 3 lehetséges irányba lehet lépni
- Minden pontba 3 irányból lehet odaérni (kivéve a széleken)
- Diagonális irány: az illesztés folytatása
- A szélekkel párhuzamos irány: betoldás valamelyik szekvenciába
- Melyik útvonal adja a legnagyobb pontszámot összesítve

Az algoritmus 3 elemi lépése



A legmagasabb pontszám nyer

$S_{(i,j-1)} - gap = S_V$

$S_{(i-1,j-1)} + M_{(i,j)} = S_C$

$S_{(i-1,j)} - gap = S_H$

$S_{(i,j)} = \max\{S_C, S_V, S_H\}$

	A	N	C	I	E	N	T
R	-1	-2	-2	-3	-1	-2	-3
E	0	-1	1	-3	6	-1	-2
C	-3	-1	17	-2	1	-1	-2
E	0	-1	1	-3	6	-1	-2
N	0	8	-1	0	-1	8	1
T	1	1	-2	0	-2	1	5

0	0	0	0	0	0	0	0
0	-1			-3	-1	-2	-3
0				-3	6	-1	-2
			17	-2	1	-1	-2
0	0	-1	1	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Ver: 0 -0 = 0

Cont: 0 -1 = -1

Hor: 0 -0 = 0

0	0	0	0	0	0	0	0
0	0	-2				-2	-3
0	0				6	-1	-2
0				-2	1	-1	-2
0	0	-1	1	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Ver: 0 -0 = 0

Cont: 0 -2 = -2

Hor: 0 -0 = 0

0	0	0	0	0	0	0	0
0	0	0	-2			2	-3
0	0	-1	1			-1	-2
0	-3	-1			1	-1	-2
0	0	-1	1	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Ver: 0 -0 = 0

Cont: 0 -2 = -2

Hor: 0 -0 = 0

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0				6	-1	-2
0	-3			-2	1	-1	-2
0			1	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Ver: 0 -0 = 0

Cont: 0 +0 = 0

Hor: 0 -0 = 0

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	-1	1			-1	-2
0	-3	-1			1	-1	-2
0	0			-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Ver: 0 -0 = 0

Cont: 0 -1 = -1

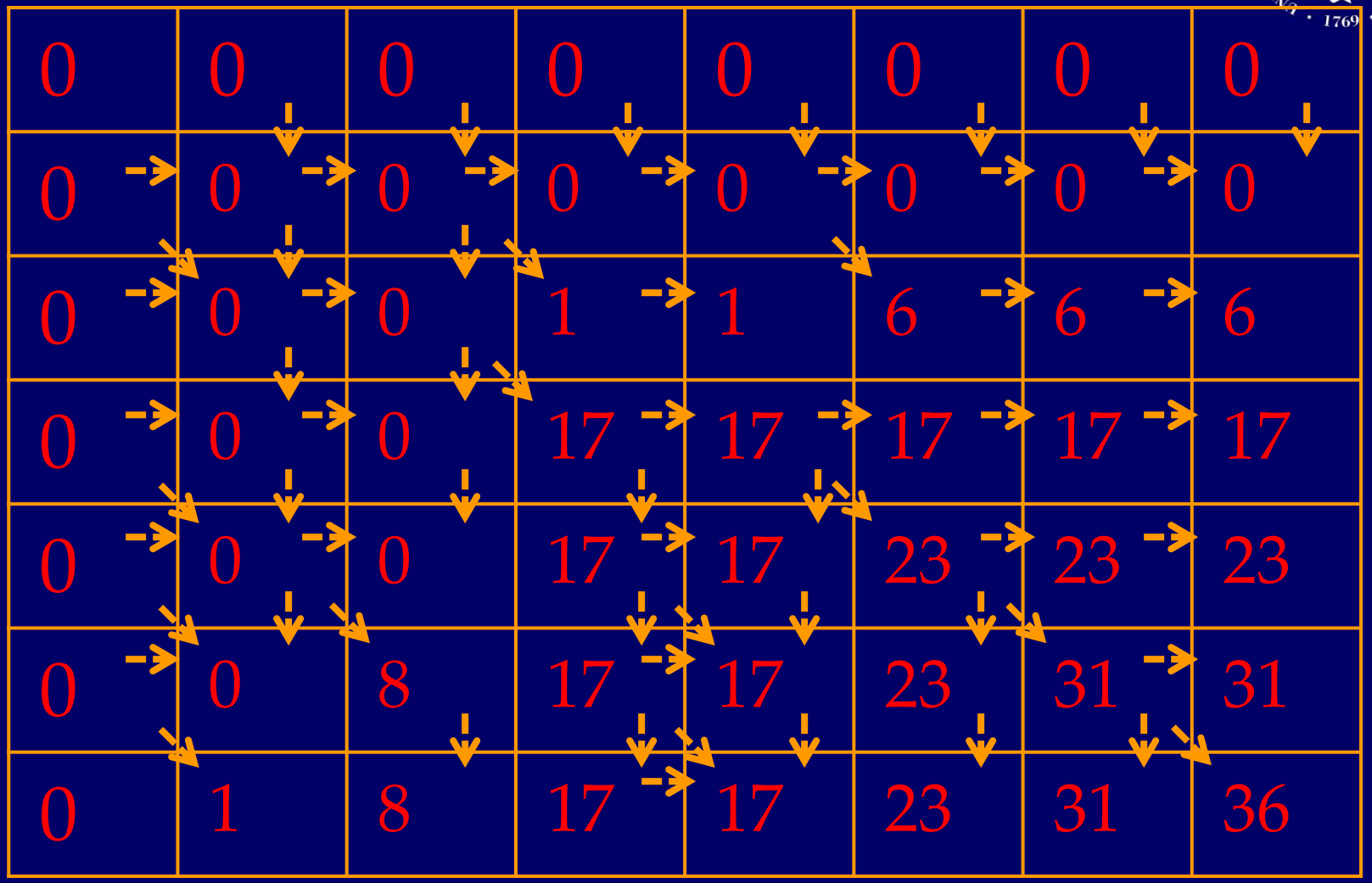
Hor: 0 -0 = 0

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	1	-1			-2
0	-3	-1	17				-2
0	0	-1				-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Ver: 0 -0 = 0

Cont: 0 +1 = 1

Hor: 0 -0 = 0


0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 1 1 6 6 6
0 0 0 17 17 17 17 17
0 0 0 17 17 23 23 23
0 0 8 17 17 23 31 31
0 1 8 17 17 23 31 36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N	0	8	17	17	23	31	31
T	1	8	17	17	23	31	36

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	1	6	6	6
C	0	0	17	17	17	17	17
E	0	0	17	17	23	23	23
N				17	23	31	31
T				17	23	31	36

ANCIENT

REC-ENT

Az algoritmus jellemzői

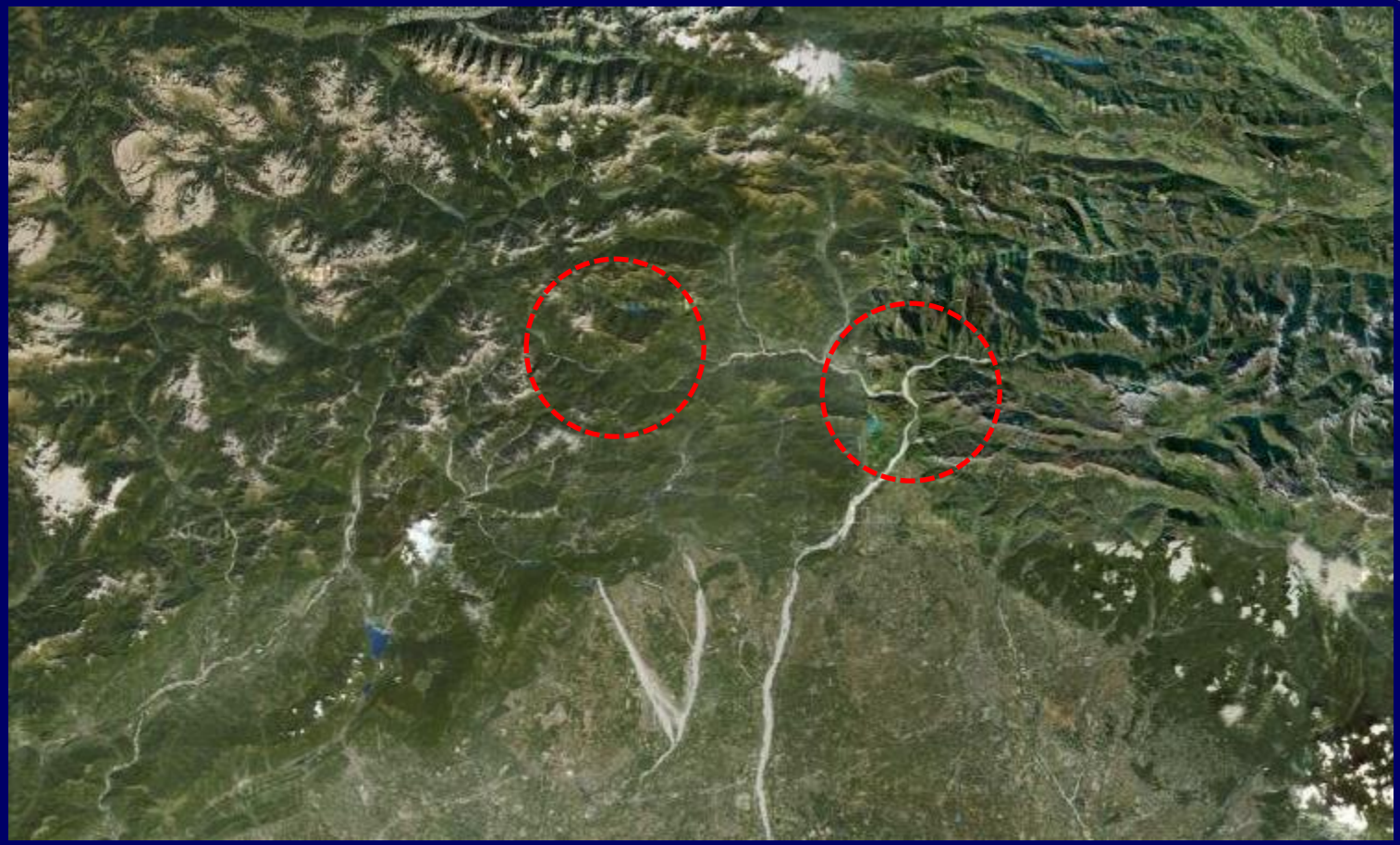
- A feladat mindig megoldható az összpontszámra nézve
- Az optimális közös ösvényre nem feltétlenül van egyértelmű megoldás
- A megoldás függ a használt pontozómátrixtól két adott szekvenciára
- A memóriaigény a két szekvencia hosszának szorzatával arányos

Mit jelent ez fehérjék esetén?

- A jellemző hossz kisebb 1000 AA-nál
- A mátrix mérete: 1 000 000 elem
- Memóriaigényben ez kb. 1 Mb szűkösen, 5 Mb kényelmesen
- Vagy sokkal kisebb memória, de több idő
- Az új számítógépeket legalabb 2 Gb memóriával árulják, de gyakran még többel
- Fehérjék N-W illesztése nem probléma

Mit jelent ez nukleinsavak esetén?

- A szekvenciák hossza lényegesen nagyobb – akár 200 000 000 bázis is lehet
- A pontozó mátrix az egységmátrix: 1 a diagonálisban, 0 máshol
- Ezért sokkal kevésbé differenciál
- Sokkal több lesz a nem visszakövethető ösvény
- Nukleinsav szekvenciák illesztése nehezebb




UNIVERSITAS BUDAPESTINENSIS DE SEMMELWEIS NOMINATA
1769

Szünet

Smith–Waterman algoritmus

- Lényegileg az N-W algoritmus változata két apró eltéréssel:
  - A betoldásokat (vizszintes vagy függőleges lépéseket) bünteti az algoritmus
  - A negatív pontszámokat felülírja és 0-ra állítja be
  - Így az illeszkedő szegmensek nem feltétlenül érik el a matrix szélét
  - A legnagyobb mátrixelemtől kezdve kezdjük visszafejteni az illesztést, 0-nál megállunk

	A	N	C	I	E	N	T
R	-1	-2	-2	-3	-1	-2	-3
E	0	-1	1	-3	6	-1	-2
C	-3	-1	17	-2	1	-1	-2
E	0	-1	1	-3	6	-1	-2
N	0	8	-1	0	-1	8	1
T	1	1	-2	0	-2	1	5

Set: E = 0 *

Ver: 0 -1 = -1

Cont: 0 -1 = -1

Hor: 0 -1 = -1

0	0	0				0	0
0	-1				1	-2	-3
0					6	-1	-2
				-2	1	-1	-2
0	0	-1	1	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

0	0	0	0				0
0	0	-2				2	-3
0	0	-				-1	-2
0	-				1	-1	-2
0	0	-1	1	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Set: E = 0 *

Ver: 0 -1 = -1

Cont: 0 -2 = -2

Hor: 0 -1 = -1

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	-1	-1	0	0	-1	-2
0	-3	-2	-2	-2	1	-1	-2
0	-3	-2	-2	-3	6	-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Set: E = 0 *

Ver: 0 -1 = -1

Cont: 0 +0 = 0 *

Hor: 0 -1 = -1

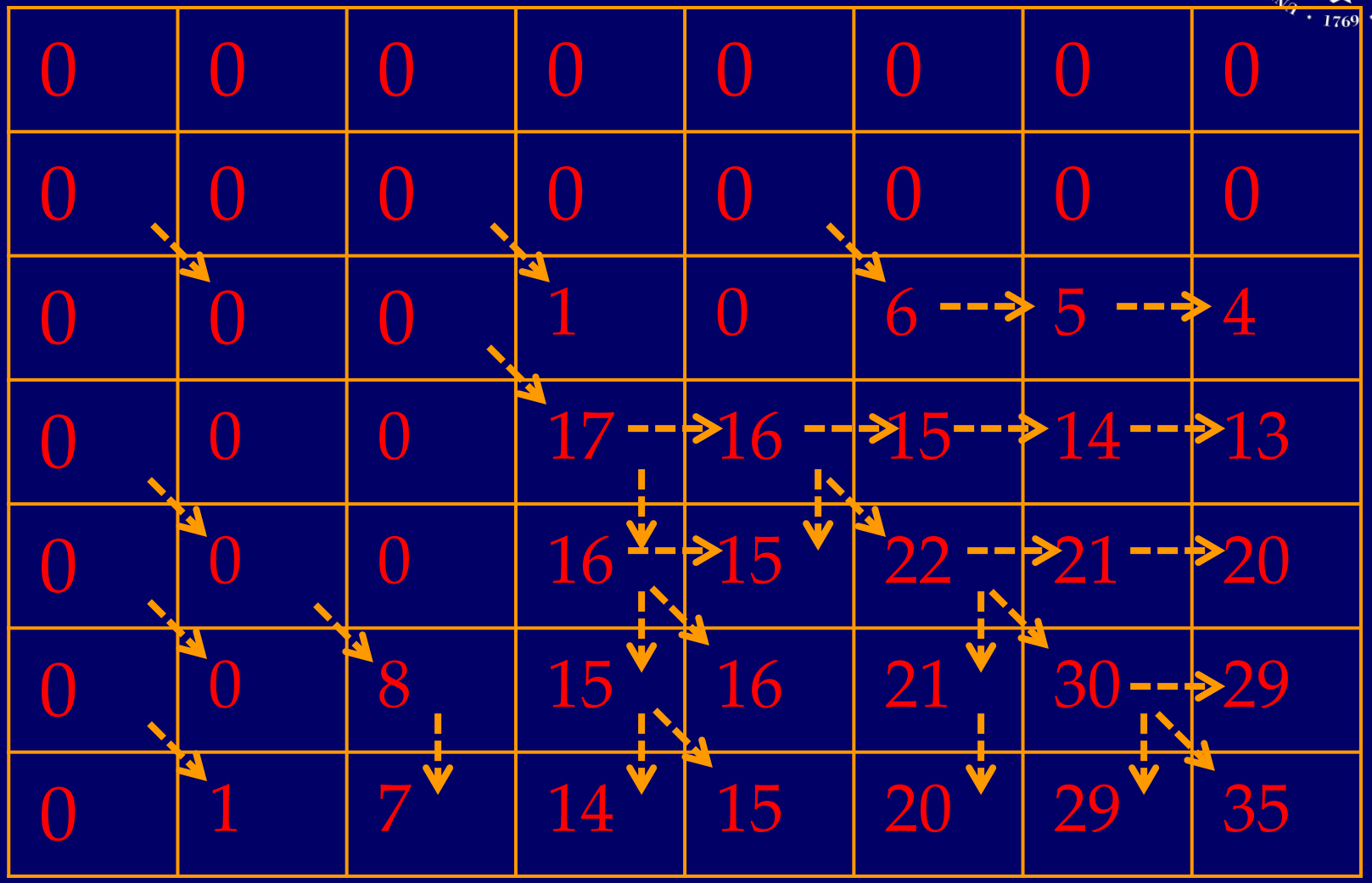
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	1				2
0	-3	-1					-2
0	0					-1	-2
0	0	8	-1	0	-1	8	1
0	1	1	-2	0	-2	1	5

Set: E = 0

Ver: 0 -1 = -1

Cont: 0 +1 = 1 *

Hor: 0 -1 = -1


0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 1 0 6 5 4
0 0 0 17 16 15 14 13
0 0 0 16 15 22 21 20
0 0 8 15 16 21 30 29
0 1 7 14 15 20 29 35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	0	8	15	16	21	30	29
T	1	7	14	15	20	29	35

	A	N	C	I	E	N	T
R	0	0	0	0	0	0	0
E	0	0	1	0	6	5	4
C	0	0	17	16	15	14	13
E	0	0	16	15	22	21	20
N	**CIENT**		15	16	21	30	29
T	**C-ENT**		14	15	20	29	35

Résbüntetési sémák

- Az N-W algoritmus nem alkalmaz résbüntetést
- Az S-W algoritmus érzékeny a résbüntetés mértékére
- Nem egyenletes résbüntetési sémák:
  - A szekvenciák két végén nincs büntetés
  - A toldás megnyitását jobban büntetik, mint a toldás szélesítését
  - Helytől függő büntetőpontszám

Mire jó végül is ez az egész?

- Honnan tudni, hogy volt-e közös őse két szekvenciának?
- Mindkét algoritmus ad valamilyen megoldást
- Mindig, még két véletlen szekvenciára is
- Az igazi kérdés: két szekvencia jobban illeszkedik-e, mint az véletlen esetben várható? Ha jobb, milyen mértékben jobb?
- Valószínűségi válasz a közös eredetre

Mit tanultunk ma?

- Két illesztendő szekvencia egy illesztő felületet feszít ki
- A felület pontjait egy helyettesítési mátrix megfelelő elemei adják
- Az illesztő felület optimális bejárásával kapjuk magát a szekvencia-illesztést
- Megkülönböztetük lokális és globális illesztést

Feladat 3

- Készítsd el az illesztő felületet két tetszőleges szekvencia (pl. a neved, és az Egyetem neve) közt és ez alapján keressed meg az optimális illesztést.

- Internetes forrás:
  - http://baba.sourceforge.net/