PRIMERJAVE ZAPOREDIJ

- Primerjave zaporedij; homologije, podobnosti, identičnost; poravnava, prileganje;
- Homologi, ortologi paralogi;
- Ocenjevanje poravnav; ocenjevalne matrike, rezultat poravnave, teža vrzeli;
- Globalne, lokalne poravnave;
- Točkovni diagrami.

PRIMERJAVE ZAPOREDIJ

Primerjamo med seboj dve zaporedji: proteini, DNA, RNA

Osnoven proces. Iskanje po bazah.
querry sequence (*probe*) *vs* ***podatkovna zbirka*** (*database, subject*)

Kako podobne so si zaporedja

malo počakati na analizo bioloških podatkov ali filogenetske analize

zelo lahko govoriš o družinah proteinov
npr. v neki družini proteinov je znanih 10 predstavnikov pri podgani, a le 7 pri človeku. Verjetno bodo odkriti še trije:
Farmakološko stališče: poznamo biološki efekt, a ne poznamo proteina zanj
Molekularno-biološko stališče: priložnost kloniranja novih človeških proteinov (*in silico* kloniranje).

PRIMERJAVE ZAPOREDIJ

Homologija (*homology*) Kvalitativna oznaka

Poudarja evolucijsko povezanost dveh zaporedij. Homologni proteini so se razvili iz skupnega prednika. Podobnost je med različnimi družinami homolognih proteinov različno ohranjena. Homologni proteini imajo vedno podobno 3D zgradbo. Pri sklepanju na homologije je potrebna previdnost in statistična analiza rezultatov (eksperimentiranje, ki je podobno eksperimentiranju v laboratorijih, "*wet labs*")

Podobnost (*similarity*) Kvantitativna oznaka

Zaporedji sta si lahko zelo podobni, ni pa nujno tudi evolucijske povezanosti.

Identičnost (identity) Kvantitativna oznaka

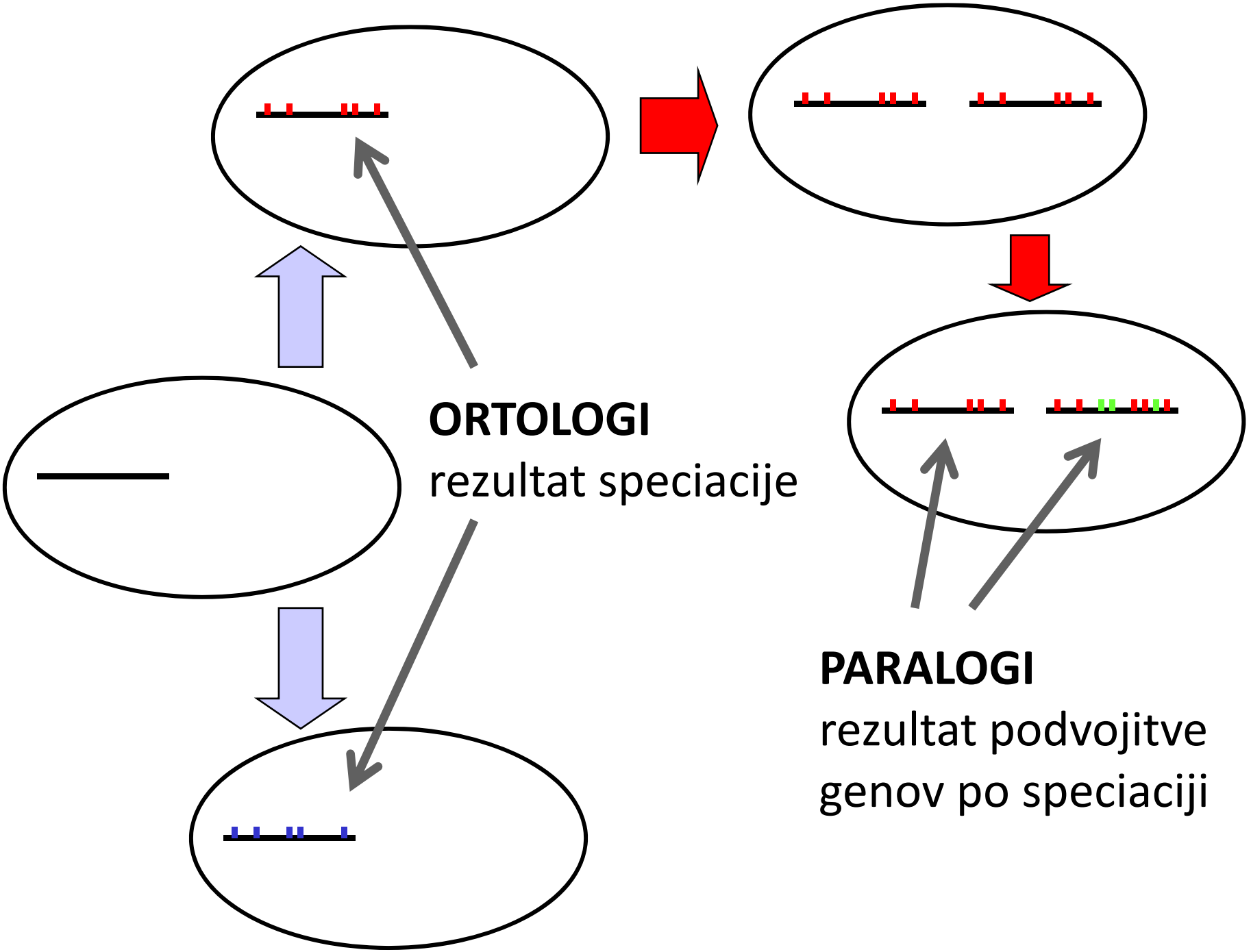
Pove koliko je identičnih amino kislin ali nukleotidov med dvema zaporedjima.

PORAVNAVANJE, PRILEGANJE

PORAVNAVA
Poseben zapis dveh ali več zaporedij. Znake izpišemo v stolpcih.

```
A G G V L I I Q V G N M R T P
: : :   :   : :     : :
A G G D L V I Q – - N M K S N
```

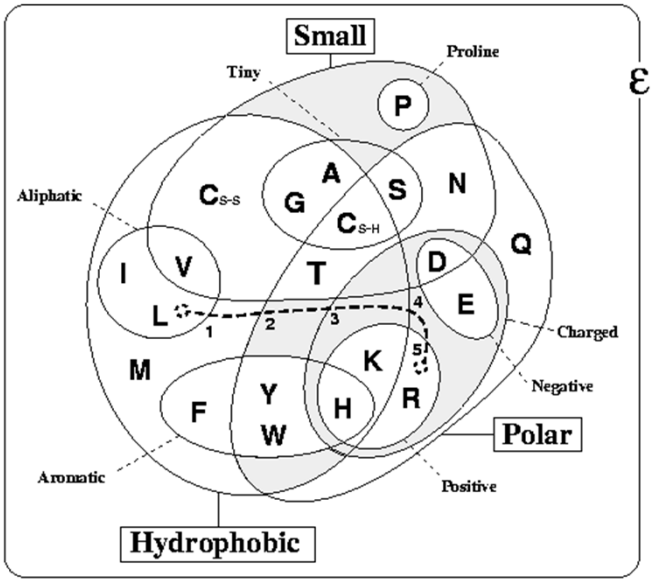
Bistvo poravnav zaporedij je poiskati takšno, ki pove nekaj o evolucijskih dogodkih, ki so privedli do današnjega stanja (zamenjave, insercije, delecije).

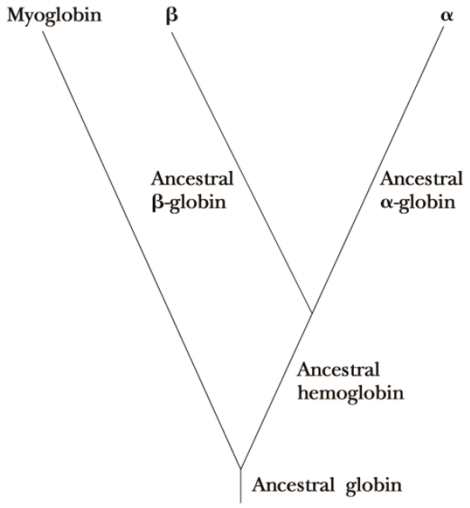

ORTOLOGI
rezultat speciacije
PARALOGI
rezultat podvojitve
genov po speciaciji

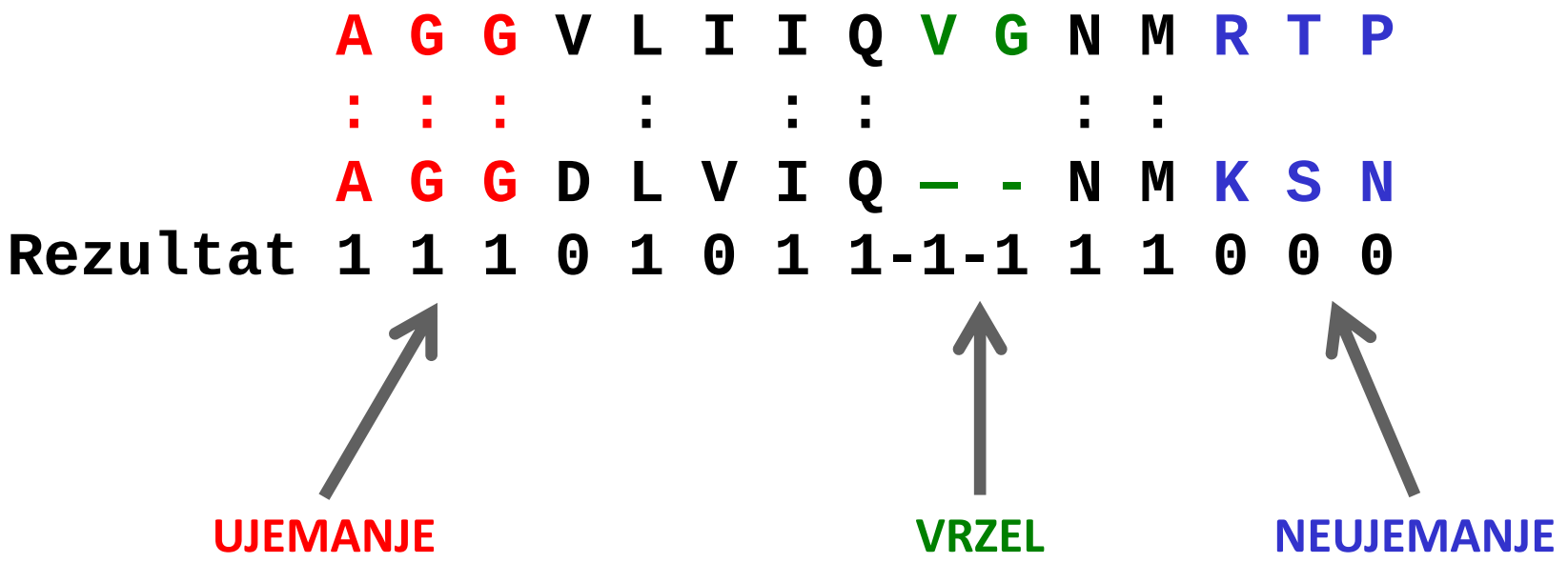
PORAVNAVA, PRILEGANJE

```
                    10        20        30        40        50        60        70        80
ALPHA1      -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH-----G-SAQVKGHGKKVADALTNAVAHVDDMPNA   80
ALPHA2      -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH-----G-SAQVKGHGKKVADALTNAVAHVDDMPNA   80
THETA       -MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAFPATKTYFSHLDLSP-----G-SSQVRAHGQKVADALSLAVERLDDLPHA   80
GAMMA       MGHFTEEDKATITSLWGKVN--VEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKHLDDLKGT   85
BETA        MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGT   85
EPSILON     MVHFTAEEKAAVTSLWSKMN--VEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKVKAHGKKVLTSFGDAIKNMDNLKPA   85
DELTA       MVHLTPEEKTAVNALWGKVN--VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGT   85
MYOGLOBIN   -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAE   86

                    90        100       110       120       130       140       150
ALPHA1      LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------ 142
ALPHA2      LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------ 142
THETA       LSALSHLHACQLRVDPASFQLLGHCLLVTLARHYPGDFSPALQASLDKFLSHVISALVSEYR------ 142
GAMMA       FAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVASALSSRYH------ 147
BETA        FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
EPSILON     FAKLSELHCDKLHVDPENFKLLGNVMVIILATHFGKEFTPEVQAAWQKLVSAVAIALAHKYH------ 147
DELTA       FSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFGKEFTPQMQAAYQKVVAGVANALAHKYH------ 147
MYOGLOBIN   IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
```







Poravnavi lahko pripišemo ***rezultat*** poravnave, ki je odraz:

- ujemanj (števila identičnih elementov med obema zaporedjima) 1
- neujemanj (števila različnih elementov med obema zaporedjima) 0
- vrzeli (insercij ali delecij)
- -1

1+1+1+0+1+0+1+1-1-1+1+1+0+0+0=6

Uteži za ujemanja in neujemanja, ki jih uporabljamo za določitev rezultata poravnav imenujemo ***ocenjevalna matrika***, uteži za vrzeli imenujemo ***teža vrzeli***.

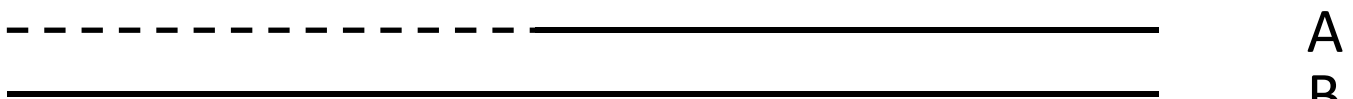
S kvantifikacijo ujemanj in neujemanj poskušamo poiskati poravnavo, ki da najvišji rezultat.
Kako poravnati?
“Na oko” *vs* računalnik z algoritmi

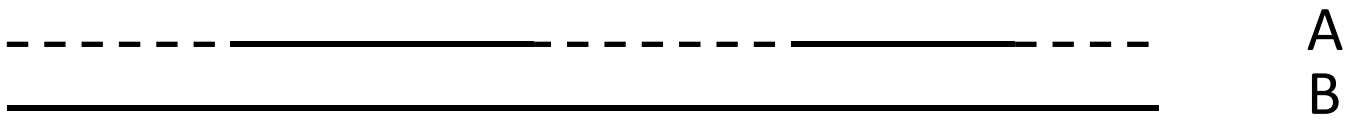
Primer:

```
AGGVLIIQVG      6          AGGVLIIQVG      9
******                     ****** ***

AGGVLIQVG                  AGGVLI-QVG
```

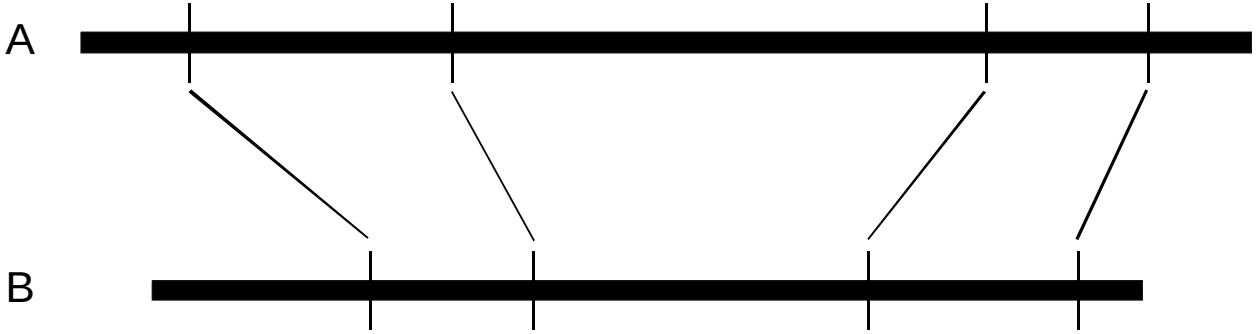
Zaporedja bioloških makromolekul daleč od idealnih primerov:
dolga zaporedja (večina proteinov ima 200-500 amino kislin)
različne dolžine
možne različne poravnave





Kadar pričakujemo podobnost celih ali večine zaporedij

Globalna poravnava



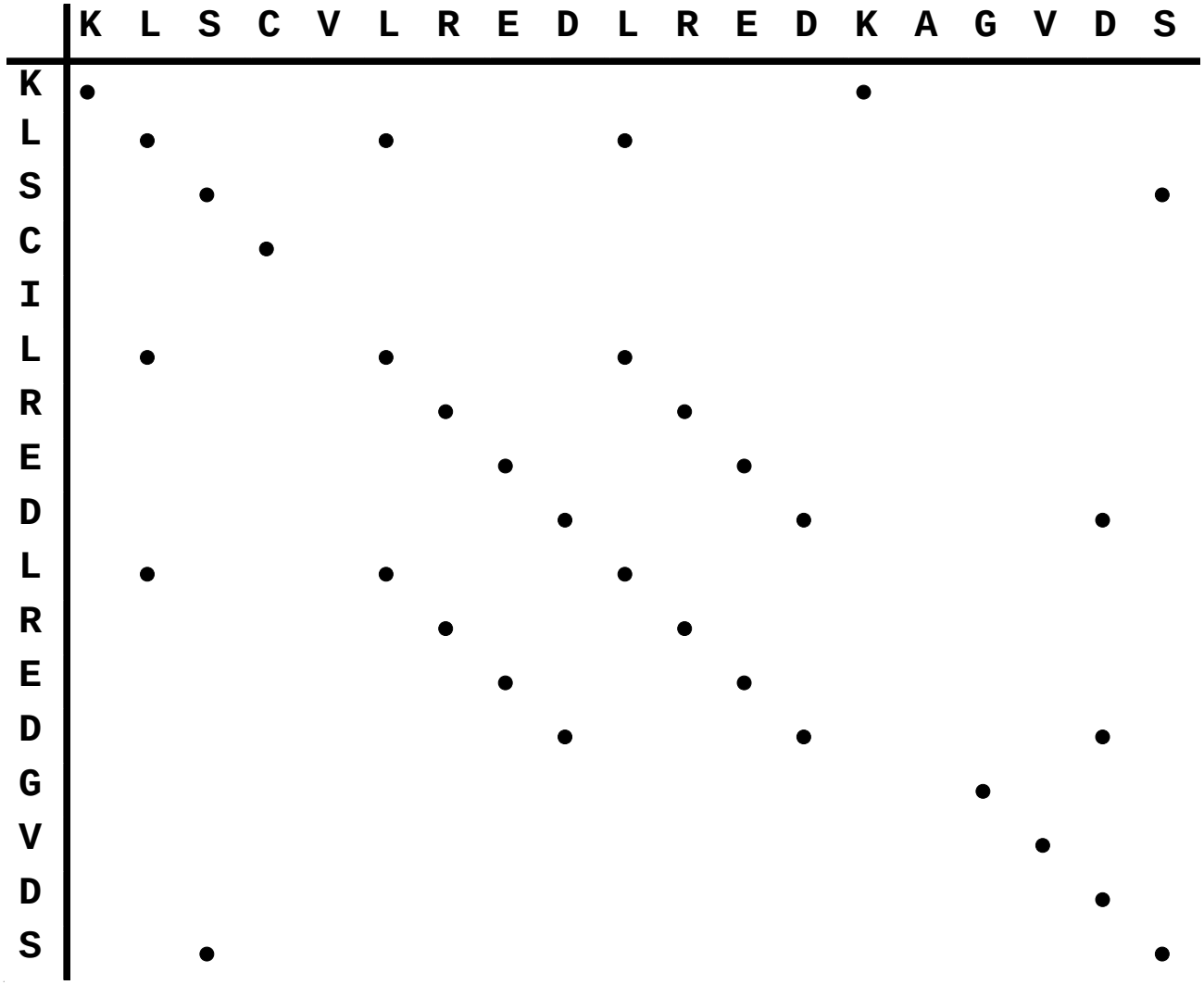
Kadar so zaporedja podobna v nekem krajšem segmentu

Lokalna poravnava

TOČKOVNI DIAGRAMI

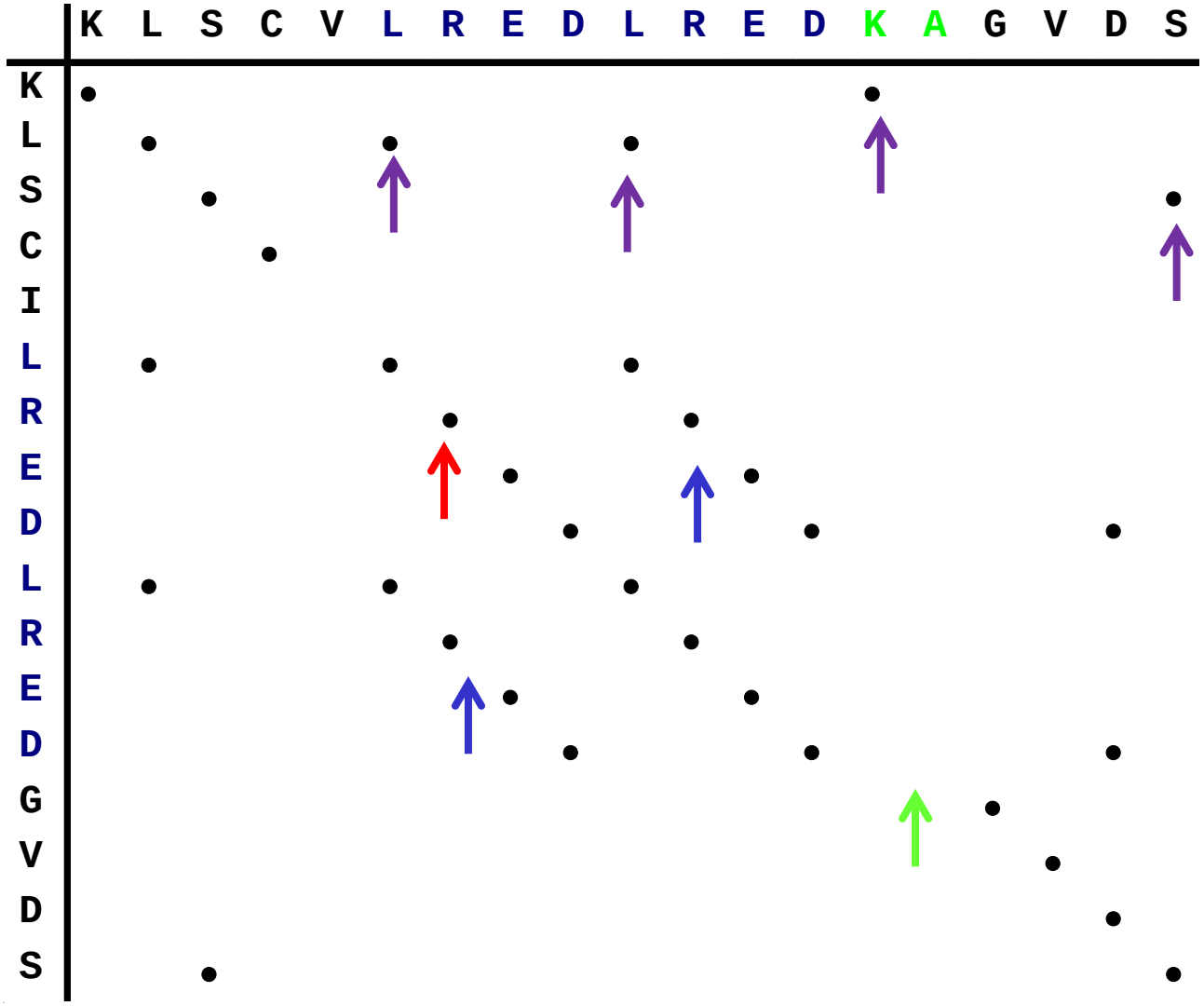
DOT PLOT

Gibbs AJ and McIntyre GA (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *European Journal of Biochemistry* 16:1-11.



Najbolj osnovna metoda. Dve zaporedji primerjamo med seboj vizuelno, “na oko”

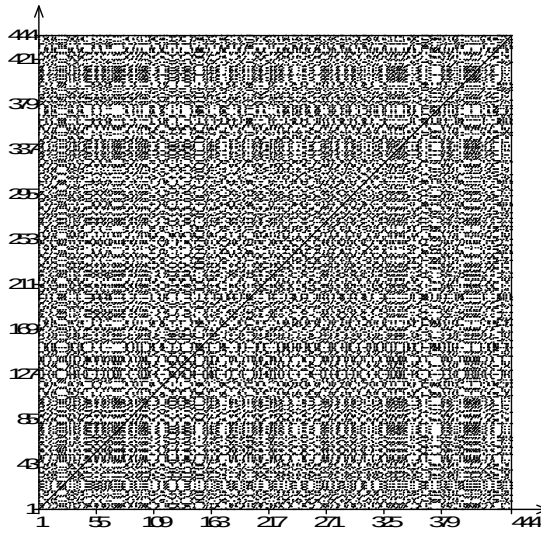
Pripravimo matriko tako, da na eno os nanesemo elemente enega zaporedja in na drugo elemente drugega zaporedja. Položaj z identičnimi elementi označimo.



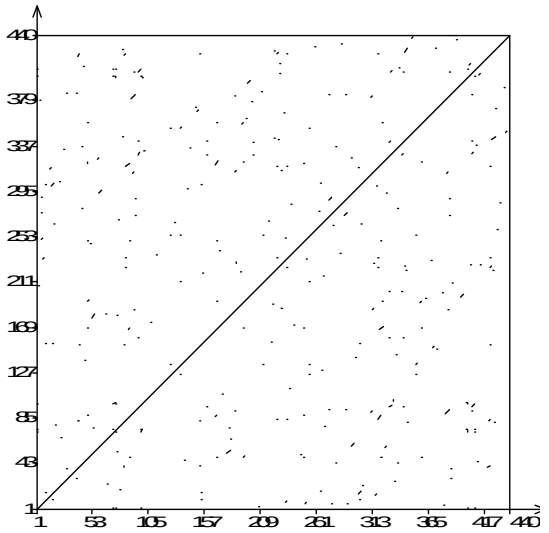
Identične aminokisline predstavlja diagonalna črta.

Lepo so vidne **ponovitve** znotraj zaporedja in **delecije**.

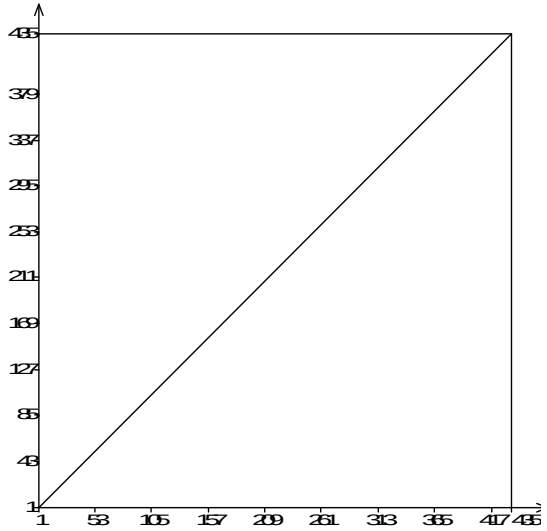
Šum zaradi zadetkov, ki se pojavijo naključno



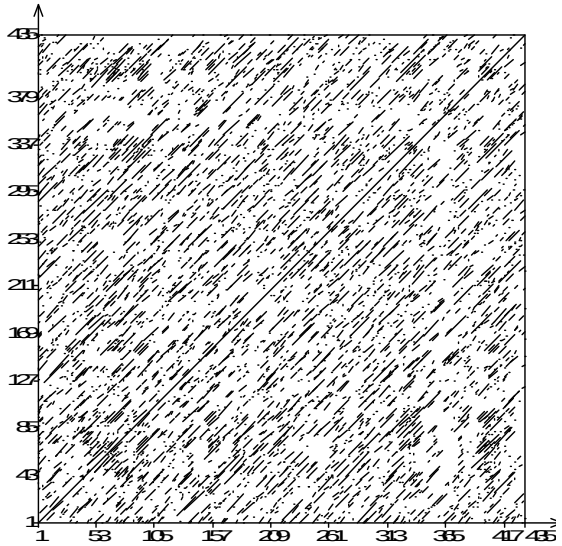
Ujemanje 100%
Okno 1



Ujemanje 100%
Okno 5



Ujemanje 100%
Okno 10



Ujemanje 50%
Okno 10

Uporaba filtrov za zmanjšanje šuma. Okna, ujemanje.

Maizel JV Jr and Lenk RP (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proceedings of National Academy of Sciences USA* 78:7665-7669.

Uporaba matrik za računanje rezultata v posameznem oknu, npr. kemijska podobnost.

Staden R. (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Research* 10:2951-2961.

>gi|122615|sp|P02023|HBB_HUMAN HEMOGLOBIN BETA CHAIN
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPVKAHGKKVLGAF
SDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALA
HKYH

>gi|7428621|pir||HBCZ hemoglobin beta chain - chimpanzee
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL
AHKYH

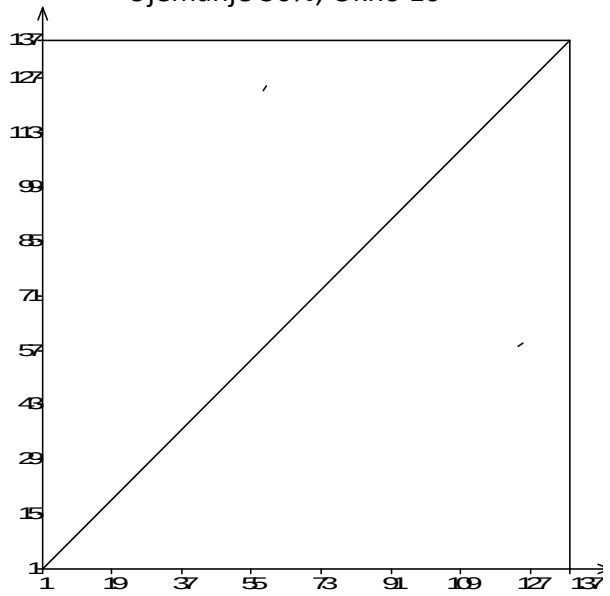
>gi|2144722|pir||HBRB hemoglobin beta chain - rabbit
MVHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSSANAVMNNPKVKAHGKKVLAA
FSEGLSHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVIVLSHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

>gi|70509|pir||HBGY hemoglobin beta chain - goldfish
VEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNPKVAAHGRTVMGGL
ERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFGPSGFNADVQEAWQKFLSVVVSAL
CRQYH

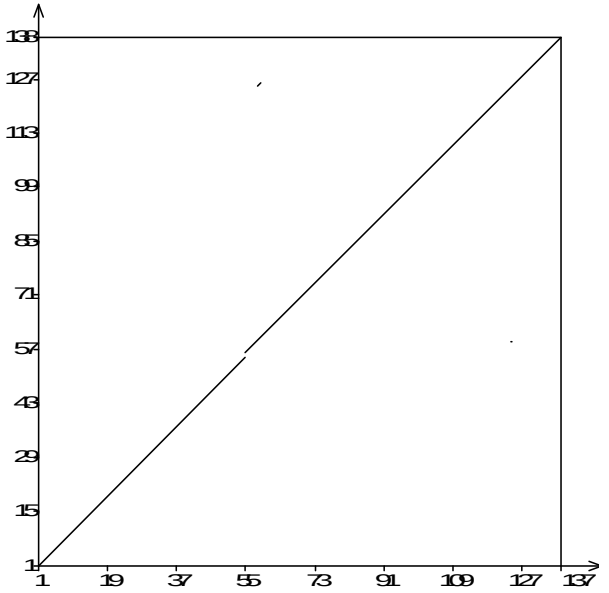
>gi|86549|pir||B24625 hemoglobin beta chain - Eurasian tree sparrow
VQWTAEEKQLITGLWGKVNVAECGGEALARLLIVYPWTQRFFASFGNLSSPTAVLGNPKVQAHGKKVLTSF
GEAVKNLDSIKNTFSQLSELHCDKLHVDPENFRLLGDILVVVLAAHFGKDFTPDCQAAWQKLVRVVAHALA
RKYH

>gi|70497|pir||HBAK hemoglobin beta chain - Nile crocodile
ASFDPHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKVQAHGKKVLASF
GEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYPKDFGLECHAAYQKLVRQVAAALA
AEYH

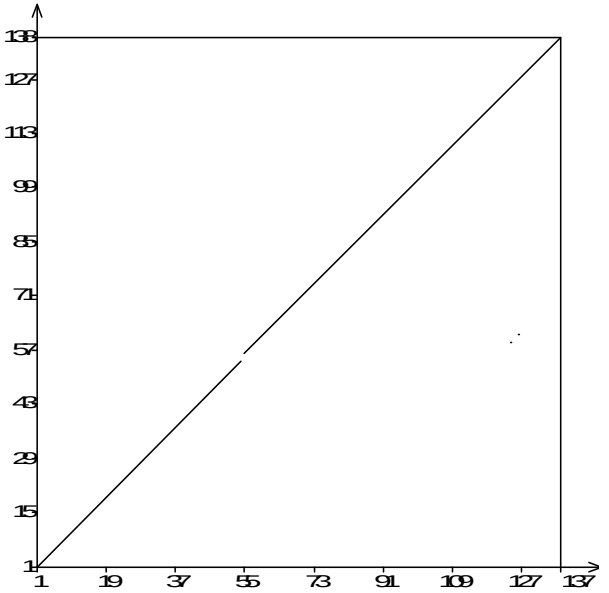
Ujemanje 50%; Okno 10



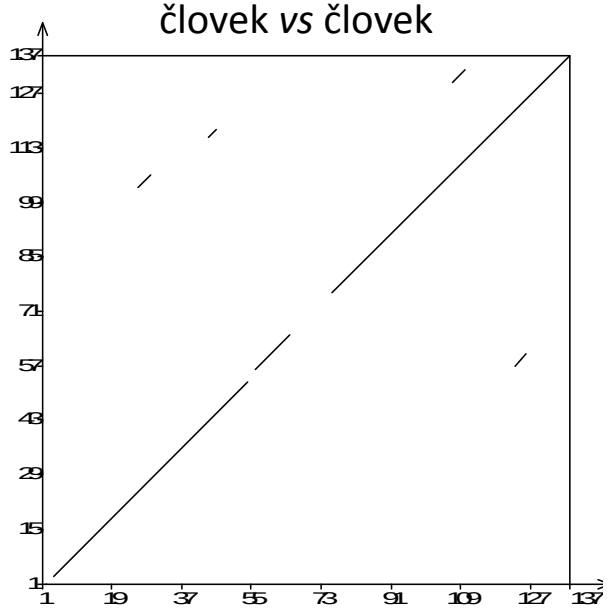
človek *vs* človek



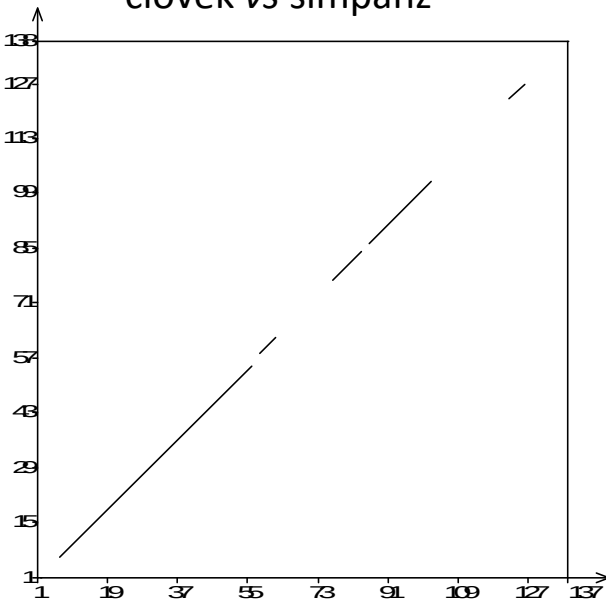
človek *vs* šimpanz



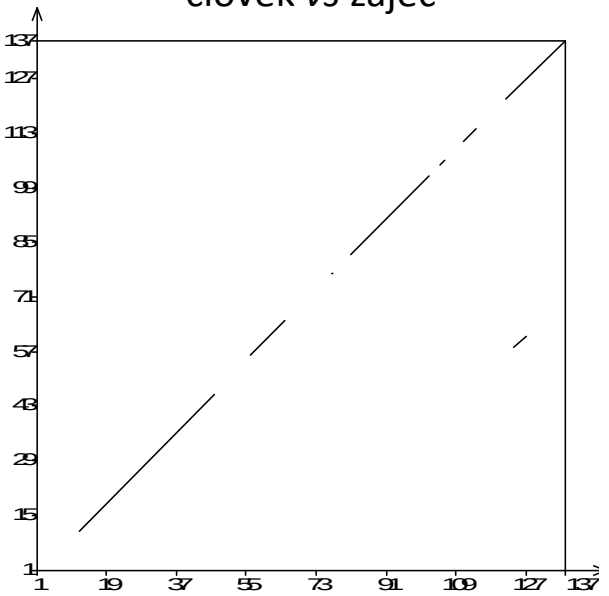
človek *vs* zajec



človek *vs* lastovka



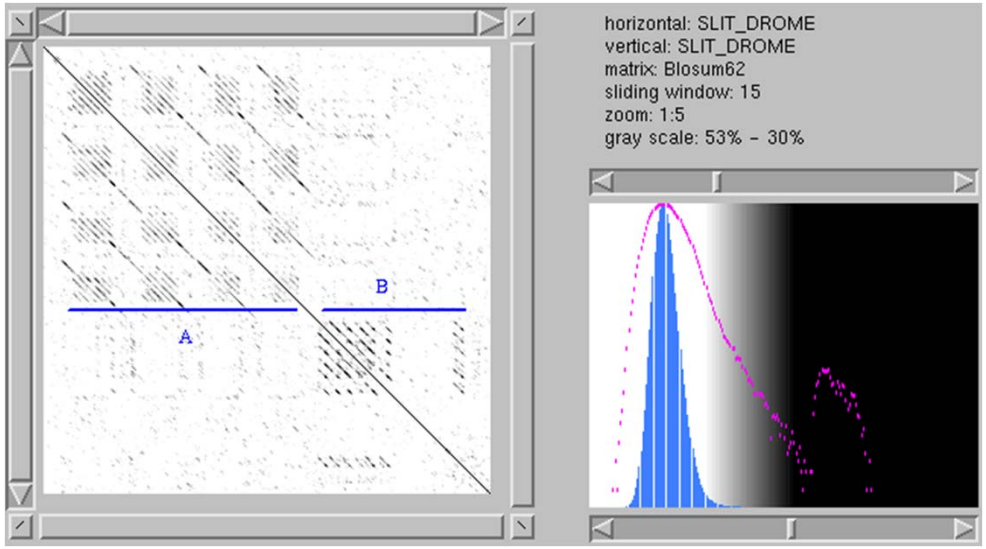
človek *vs* zlata ribica

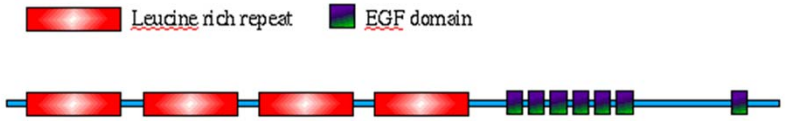


človek *vs* krokodil

Ponavljajoča zaporedja

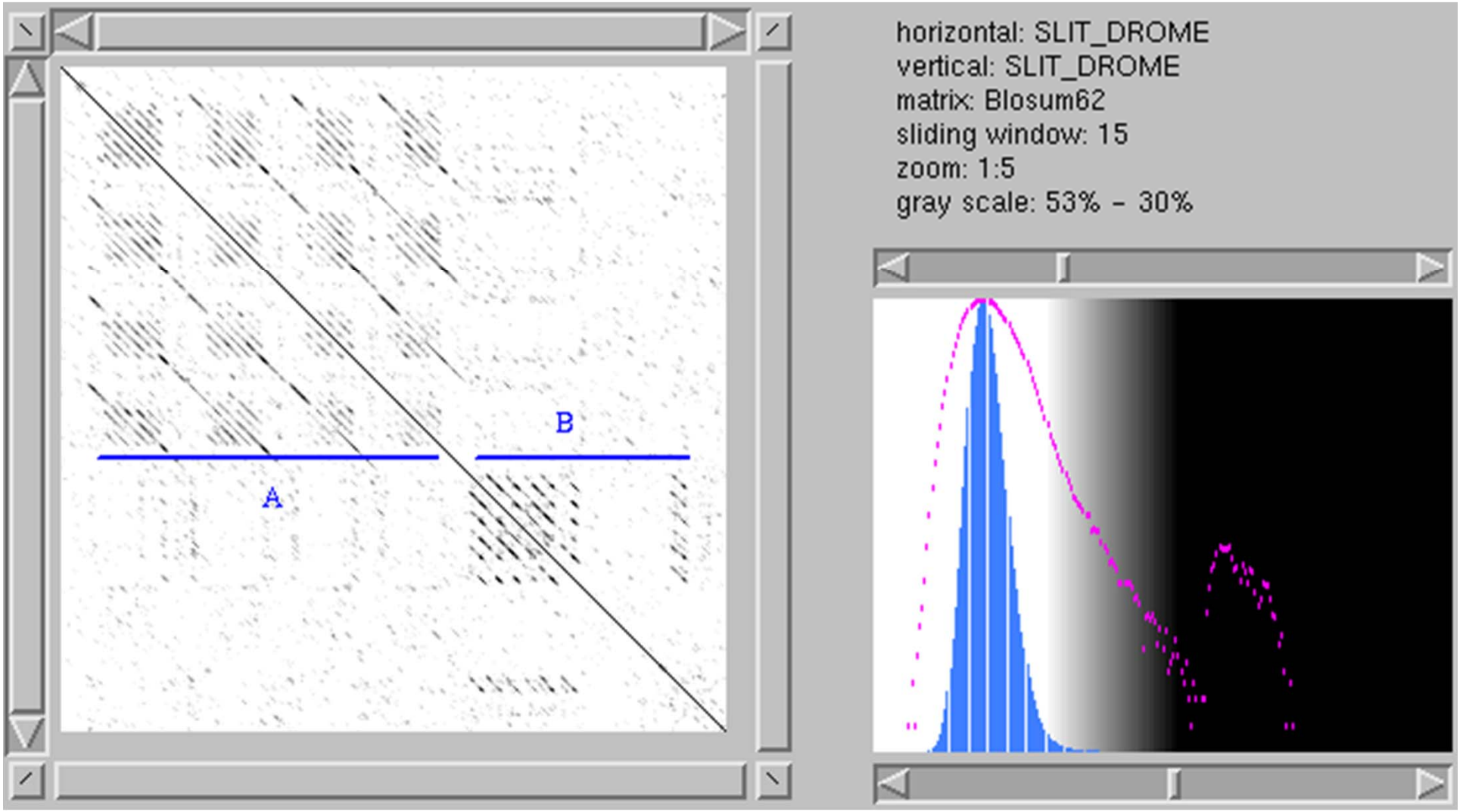
Drosophila melanogaster
SLIT protein





http://www.isrec.isb-sib.ch/java/dotlet/dotlet_examples.html

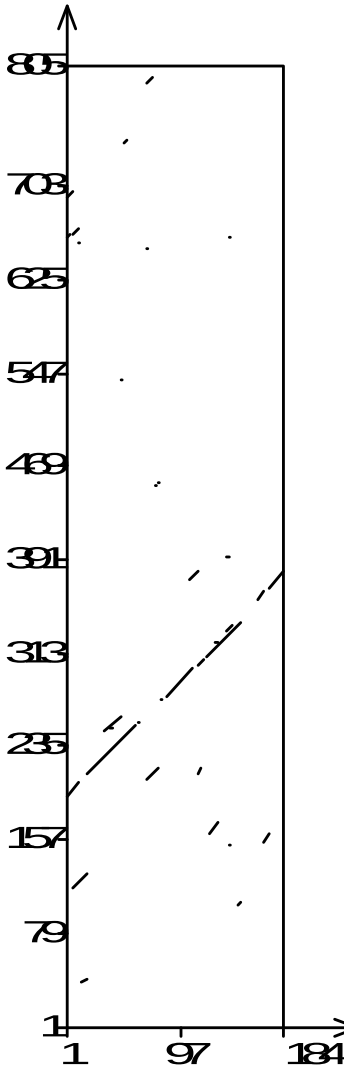
>SLIT_DROME (P24014):
MAAPSRTTLMPPPFRLQLRLLILPILLLLRHDAVHAEPYSGGFGSSAVS
SGGLGSVGIHIPGGGVGVITEARCPRVCSCTGLNVD**CSHRGLTSVPRKI**
SADVERLELQGNNLTVIYETDFQRLTKLRMLQLTDNQIHTIERNSFQDL
VSLERLDISNNVITTVGRRVFKGAQSLRSLQLDNNQITCLDEHAFKGLV
ELEILTLNNNNLTSLPHNIFGGLGRLRALRLSDNPFACDCHLSWLSRFL
RSATRLAPYTRCQSPSQLKGQNVADLHDQEFKCSGLTEHAPMECGAENS
CPHPCRCADGIVDCREKSLTSVPVTLPDDTTDV**RLEQNFITELPPKSFS**
SFRRLRRIDLSNNNISRIAHDALSGLKQLTTLVLYGNKIKDLPSGVFKG
LGSLRLLLLNANEISCIRKDAFRDLHSLSLLSLYDNNIQSLANGTFDAM
KSMKTVHLAKNPFICDCNLRWLADYLHKNPIETSGARCESPKRMHRRRI
ESLREEKFKCSWGELRMKLSGECRMDSDCPAMCHCEGTTVDCTGRRLKE
IPRDIPLHTTELLLNDNELGRISSDGLFGRLPHLVKLE**LKRNQLTGIEP**
NAFEGASHIQELQLGENKIKEISNKMFLGLHQLKTLNLYDNQISCVMPG
SFEHLNSLTSLNLASNPFNCNCHLAWFAECVRKKSLNGGAARCGAPSKV
RDVQIKDLPHSEFKCSSENSEGCLGDGYCPPSCTCTGTVVACSRNQLKE
IPRGIPAETSELYLESNEIEQIHYERIRHLRSLTRLDLSNNQITILSNY
TFANLTKLSTLIISYNKLQCLQRHALSGLNNLRVVSLHGNRISMLPEGS
FEDLKSLTHIALGSNPLYCDCGLKWFSDWIKLDYVEPGIARCAEPEQMK
DKLILSTPSSSFVCRGRVRNDILAKCNACFEQPCQNQAQCVALPQREYQ
CLCQPGYHGKHCEFMIDACYGNPCRNNATCTVLEEGRFS**CQCAPGYTGA**
RCETNIDDCLGEIKCQNNATCIDGVESYK**CECQPGFSGEFC**DTKIQFCS
PEFNPCANGAKCMDHFTHYS**CDCQAGFHGTNCT**DNIDDCQNHMCQNGGT
CVDGINDYQ**CRCPDDYTGKYC**EGHNMISMMYPQTSPCQNHECKHGVCFQ
PNAQGSDY**LCRCHPGYTGKWC**EYLTSISFVHNNSFVELEPLRTRPEANV
TIVFSSAEQNGILMYDGQDAHLAVELFNGRIRVSYDVGNHPVSTMYSFE
MVADGKYHAVELLAIKKNFTLRVDRGLARSIINEGSNDYLKLTTPMFLG
GLPVDPAQQAYKNWQIRNLTSFKGCMKEVWINHKLVDFGNAQRQQKITP
GCALLEGEQQEEEDDEQDFMDETPHIKEEPVDPCLENKCRRGSRCVPNS
NARDGYQ**CKCKHGQRGRYCD**QGEGSTEPPTVTAASTCRKEQVREYYTEN
DCRSRQPLKYAKCVGGCGNQCCAAKIVRRRKVRMVCSNNRKYIKNLDIV
RKCGCTKKCY


horizontal: SLIT_DROME
vertical: SLIT_DROME
matrix: Blosum62
sliding window: 15
zoom: 1:5
gray scale: 53% - 30%
A
B

Ohranjene domene

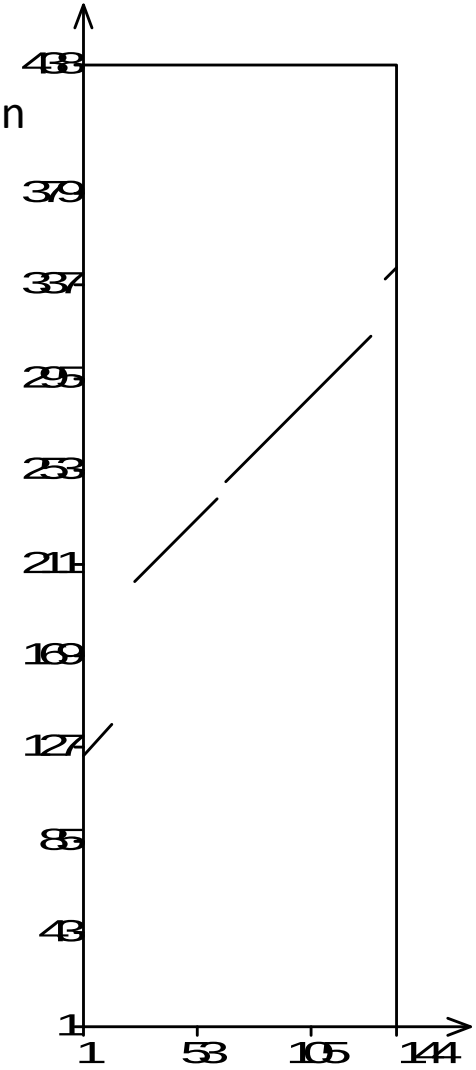
MS2 površinski celični antigen *vs* adamalysin II, metaloproteinaza iz strupa *Crotalus adamanteus* (Eastern diamondback rattlesnake). Oba vsebujeta Zn-proteazno domeno

Ujemanje 100%
Okno 5



>MS2_HUMAN (P78325)
MRGLGLWLLGAMMLPAIAPSRPWALMEQYEVVLPRRLPG
PRVRRALPSHLGLHPERVSYVLGATGHNFTLHLRKNRDL
LGSGYTETYTAANGSEVTEQPRGQDHCLYQGHVEGYPDS
AASLSTCAGLRGFFQVGSDLHLIEPLDEGGEGGRHAVYQ
AEHLLQTAGTCGVSDDSLGSLLGPRTAAVFRPRPGDSLP
SRETRYVELYVVVDNAEFQMLGSEAAVRHRVLEVVNHVD
KLYQKLNFRVVLVGLEIWNSQDRFHVSPDPSVTLENLLT
WQARQRTRRHLHDNVQLITGVDFTGTTVGFARVSAMCSH
SSGAVNQDHSKNPVGVACTMAHEMGHNLGMDHDENVQGC
RCQERFEAGRCIMAGSIGSSFPRMFSDCSQAYLESFLER
PQSVCLANAPDLSHLVGGPVCGNLFVERGEQCDCGPPED
CRNRCCNSTTCQLAEGAQCAHGTCCQECKVKPAGELCRP
KKDMCDLEEFCDGRHPECPEDAFQENGTPCSGGYCYNGA
CPTLAQQCQAFWGPGGQAAEESCFSYDILPGCKASRYRA
DMCGVLQCKGGQQPLGRAICIVDVCHALTTEDGTAYEPV
PEGTRCGPEKVCWKGRCQDLHVYRSSNCSAQCHNHGVCN
HKQECHCHAGWAPPHCAKLLTEVHAASGSLPVLVVVVLV
LLAVVLVTLAGIIVYRKARSRILSRNVAPKTTMGRSNPL
FHQAASRVPAKGGAPAPSRGPQELVPTTHPGQPARHPAS
SVALKRPPPAPPVTVSSPPFPVPVYTRQAPKQVIKPTFA
PPVPPVKPGAGAANPGPAEGAVGPKVALKPPIQRKQGAG
APTAP
>ADAM_CROAD (P34179)
QQNLPQRYIELVVVADRRVFMKYNSDLNIIRTRVHEIVN
IINGFYRSLNIDVSLVNLEIWSGQDPLTIQSSSSNTLNS
EGLWREKVLLNKKKKDNAQLLTAIEFKCETLGKAYLNSM
CNPRSSVGIVKDHSPINLLVAVTMAHELGHNLGMEHDGK
DCLRGASLCIMRPGLTPGRSYEFSDDSMGYYQKFLNQYK
PQCILNKP

Ujemanje 30%
Okno 20

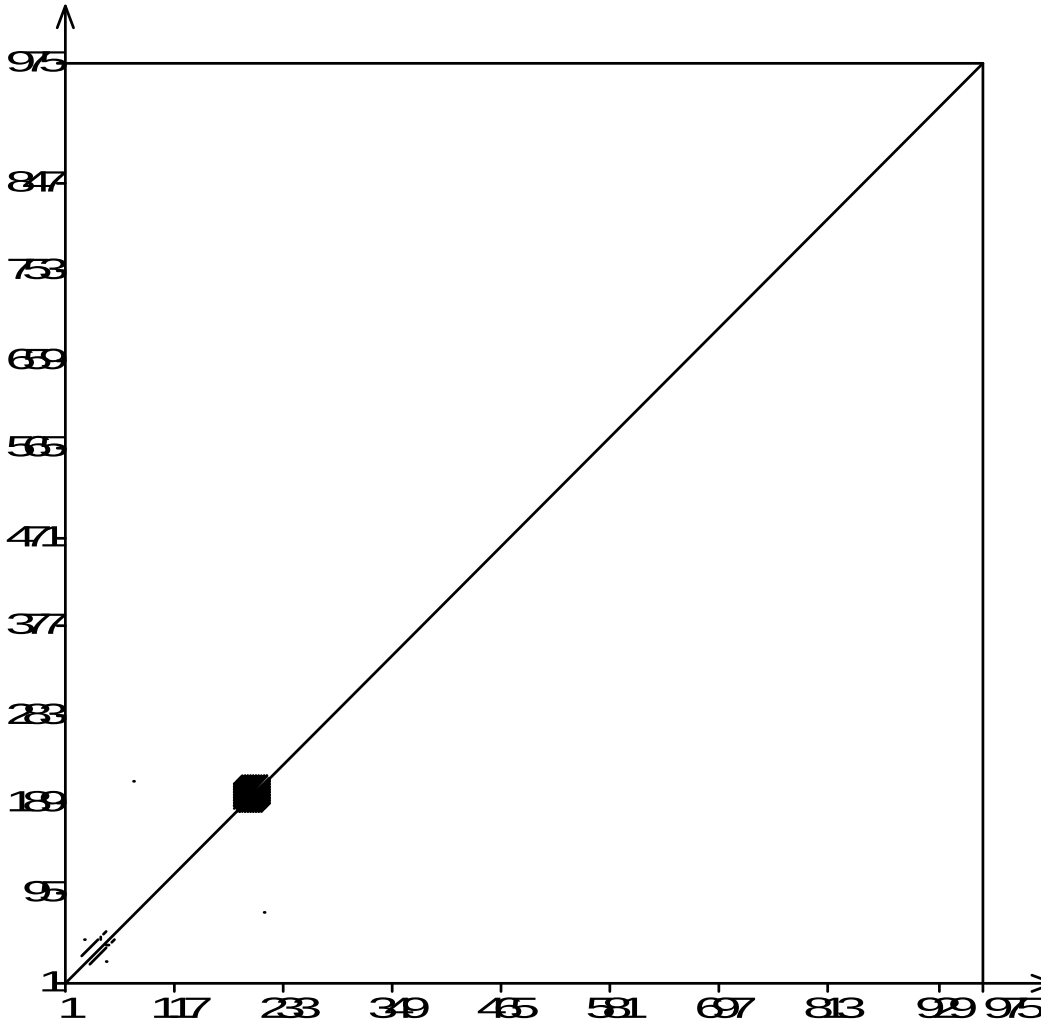


Eksoni, introni

Preveden *Emericella (Aspergillus) nidulans* calmodulin gen *vs* protein

http://www.isrec.isb-sib.ch/java/dotlet/dotlet_examples.html

Ujemanje 50%
Okno 15



http://www.isrec.isb-sib.ch/java/dotlet/dotlet_examples.html

Regije z nizko kompleksnostjo

Plasmodium falciparum proteinski prekurzor- antigen s serinskimi ponovitvami

>SERA_PLAFG (P13823)
MKSYISLFFILCVIFNKNVIKCTGESQTGNTGGGQAG
NTVGDQAGSTGGSPQGSTGASQPGSSEPSNPVSSGHS
VSTVSVSQTSTSSEKQDTIQVKSALLKDYMGLKVTGP
CNENFIMFLVPHIYIDVDTEDTNIELRTTLKETNNAI
SFESNSGSLEKKKYVKLPSNGTTGEQGSSTGTVRGDT
EPISD**SSSSSSSSSSSSSSSSSSSSSSSSSSSSSS**
SSSESLPANGPDSPTVKPPRNLQNICETGKNFKLVVY
IKENTLIIKWKVYGETKDTTENNKVDVRKYLINEKET
PFTSILIHAYKEHNGTNLIESKNYALGSDIPEKCDTL
ASNCFLSGNFNIEKCFQCALLVEKENKNDVCYKYLSE
DIVSNFKEIKAETEDDDEDDYTEYKLTESIDNILVKM
FKTNENNDKSELIKLEEVDDSLKLELMNYCSLLKDVD
TTGTLDNYGMGNEMDIFNNLKRLLIYHSEENINTLKN
KFRNAAVCLKNVDDWIVNKRGLVLPELNYDLEYFNEH
LYNDKNSPEDKDNKGKGVVHVDTTLEKEDTLSYDNSD
NMFCNKEYCNRLKDENNCISNLQVEDQGNCDTSWIFA
SKYHLETIRCMKGYEPTKISALYVANCYKGEHKDRCD
EGSSPMEFLQIIEDYGFLPAESNYPYNYVKVGEQCPK
VEDHWMNLWDNGKILHNKNEPNSLDGKGYTAYESERF
HDNMDAFVKIIKTEVMNKGSVIAYIKAENVMGYEFSG
KKVQNLCGDDTADHAVNIVGYGNYVNSEGEKKSYWIV
RNSWGPYWGDEGYFKVDMYGPTHCHFNFIHSVVIFNV
DLPMNNKTTKKESKIYDYYLKASPEFYHNLYFKNFNV
GKKNLFSEKEDNENNKKLGNNYIIFGQDTAGSGQSGK
ESNTALESAGTSNEVSERVHVYHILKHIKDGKIRMGM
RKYIDTQDVNKKHSCTRSYAFNPENYEKCVNLCNVNW
KTCEEKTSPGLCLSKLDTNNECYFCYV

TOČKOVNI GRAFI NA INTERNETU

DOTLET

http://myhits.isb-sib.ch/cgi-bin/dotlet

Pagni M and Junier T. Swiss Institute of Bioinformatics, Epalinges, Switzerland.
Točkovni graf dveh zaporedij, določiš okno, matriko.

DNADOT

http://arbl.cvmbs.colostate.edu/molkit/dnadot/

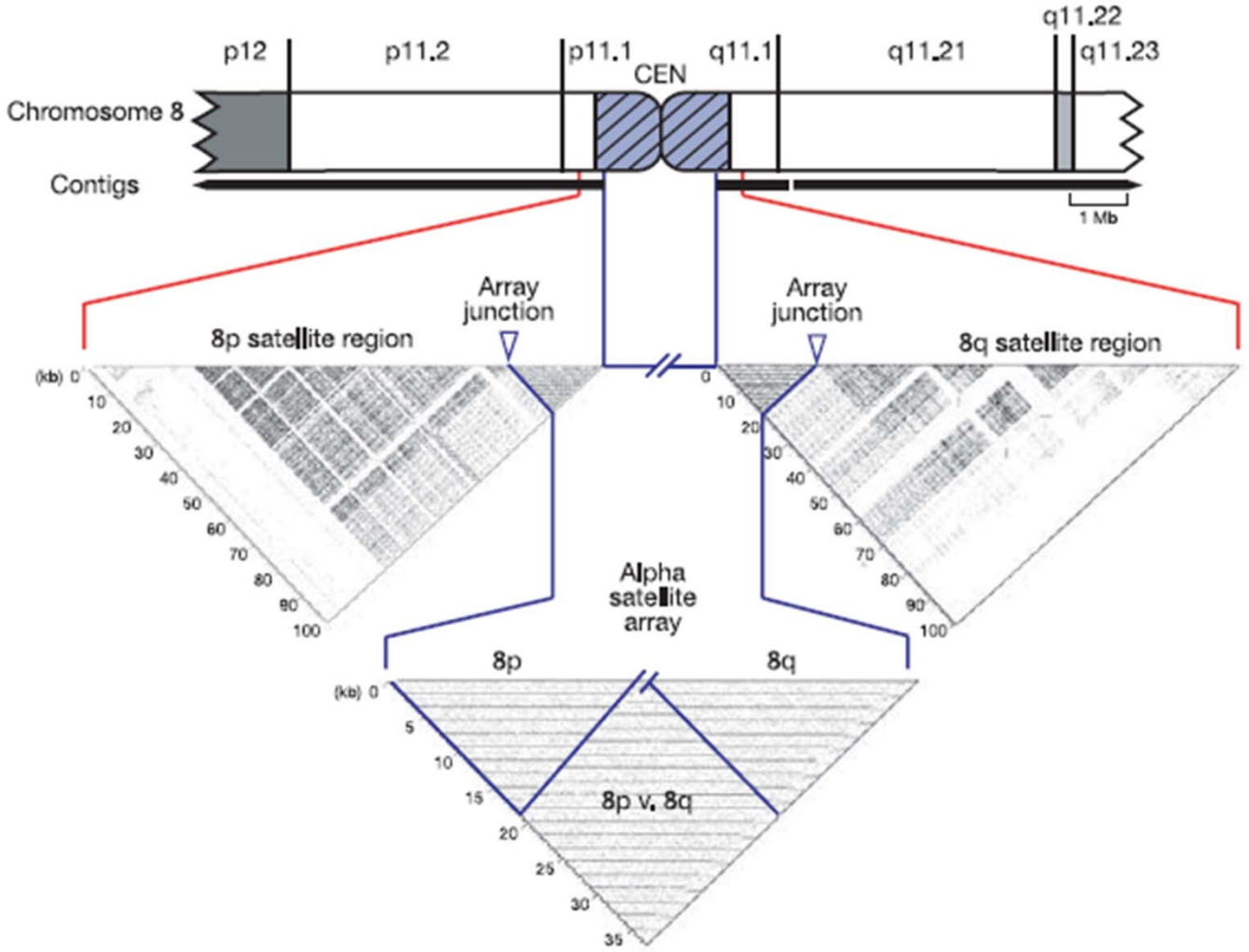
V Molecular Tool Kit. Na Colorado State University, USA.
Za primerjavo DNA zaporedij. Določiš okno in ujemanje.

DOTTER

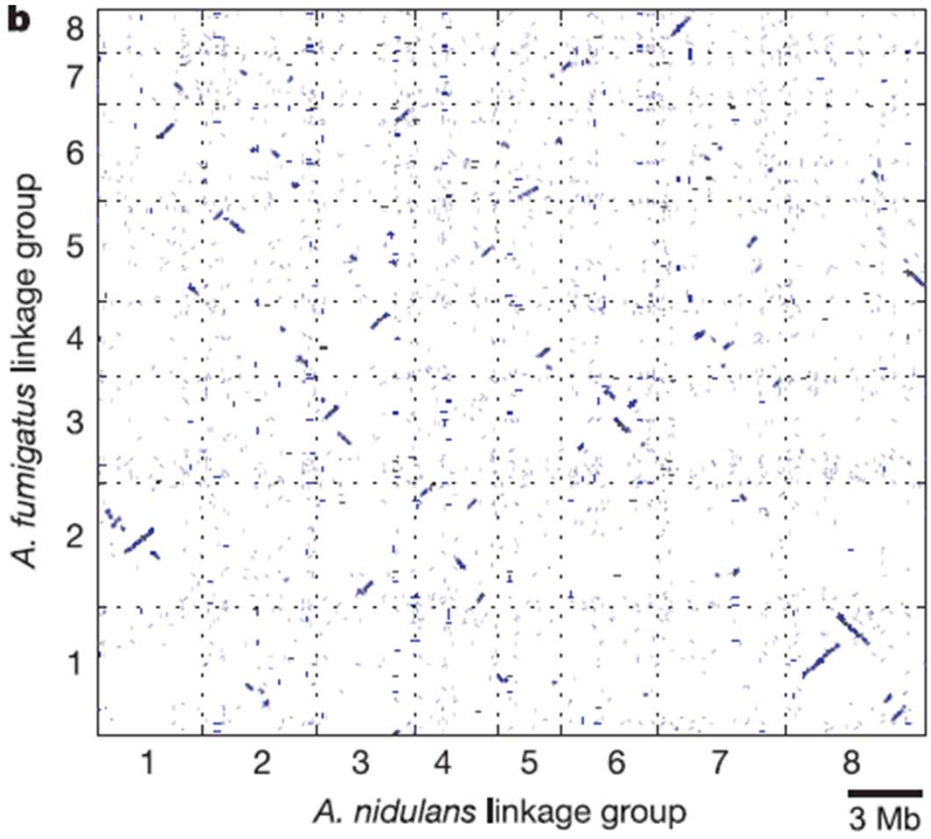
http://sonnhammer.sbc.su.se/Dotter.html

Karolinska Institutet, Center for Genomic Research, Švedska
Točkovni graf dveh zaporedij, določiš okno, matriko. Zaporedji primerja med seboj v dveh dimenzijah, v tretji dimenziji poda velikost rezultata kot vrh.

TOČKOVNI GRAFI V GENOMIKI



Nusbaum C et al. (2006) DNA sequence and analysis of human chromosome 8. Nature 439, 331-335.



Galagan JE et al. (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 438, 1105-1115.