7. vaja: Naloge in odgovori

V07-01

TOČKOVNI DIAGRAM

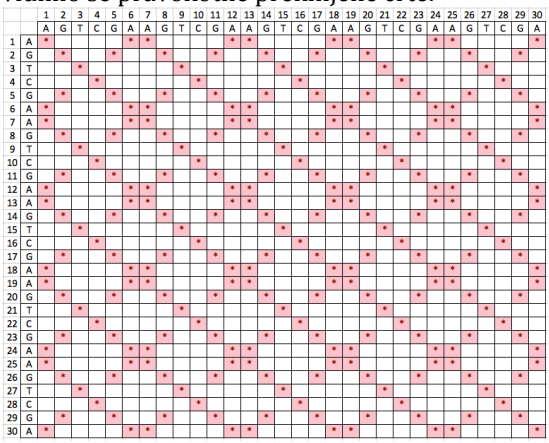
V Excel-u (datoteka "V07 – Naloge.xlsx") izdelajte prileganje dveh zaporedij po metodi točkovnega diagrama.

Kot vhodni podatek imate podane pare zaporedij (npr. par A sestavljata zaporedji A1 in A2, ki ju primerjate med seboj). Tabelo za primerjanje izdelajte tako, da boste v dve celici nad točkovnim diagramom oz. matriko prilepili zaporedji iz posameznega para (npr. A1 in A2, B1 in B2, ...). Uporabite funkciji MID in IF; kot velikost okna upoštevajte vrednost 1 (to pomeni, da direktno primerjate posamezne nukleotide).

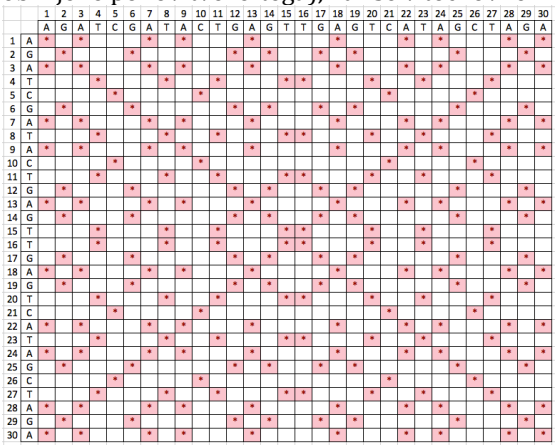
Oglejte si točkovne diagrame, ki pri tem nastanejo, identificirajte insercije/delecije, ponovitve, ... v zaporedjih in si oglejte, kako se odražajo v točkovnem diagramu.

Odgovor

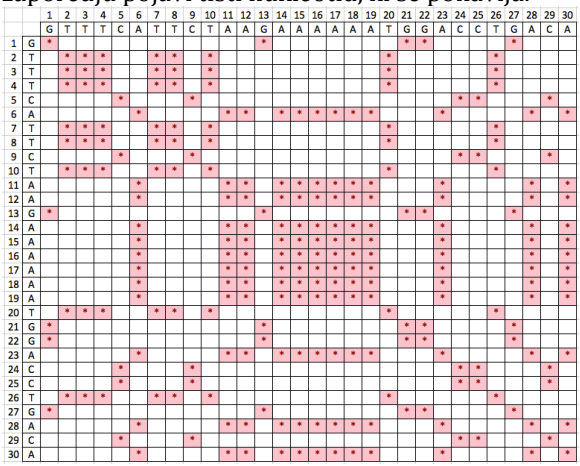
Zaporedje A1 ima interne direktne ponovitve motiva AGTCGA, ki se v točkovnem dirgramu kažejo kot vzporedne diagonalne črte. Zaradi kratke obrnjene ponovitve (AG na začetku oz GA na koncu motiva) vidimo še pravokotne prekinjene črte.



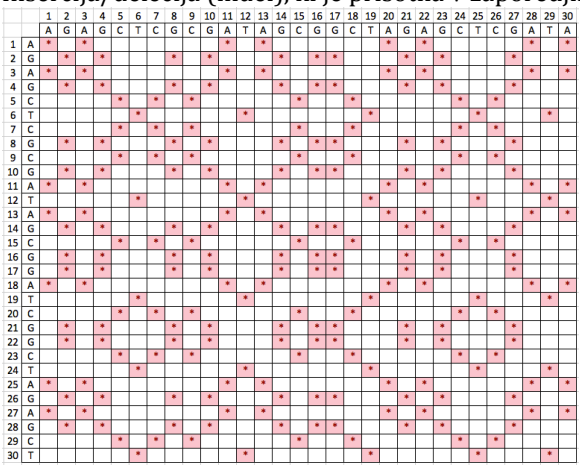
Zaporedje B1 vsebuje obrnjeno ponovitev (sestavljeno je torej iz zaporedja AGATCGATACTGAGT in obrnjene ponovitve le-tega), kar se v točkovnem diagramu kaže kot prekrižani diagonali.



Ponovitev enega nukleotida v zaporedju C1 se kaže kot navpične in vodoravne črte tik ob/na diagonali (ki tvorijo pravokotnik na diagonali) ter še izvendiagonalne vodoravne in navpične črte na mestih, kjer se v zaporedju pojavi tisti nukleotid, ki se ponavlja.



Insercija/delecija (indel), ki je prisotna v zaporedjih skupine D, se kaže kot prelomljena diagonala.



V07-02

TOČKOVNI DIAGRAM

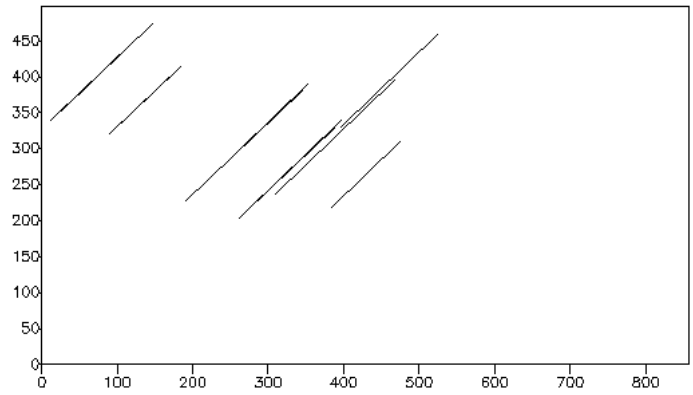
Primerjajte aminokislinski zaporedji človeškega proteina N-CAM-1 (Neural cell adhesion molecule 1) in miotilina (myotilin). Uporabite spletno orodje DotMatcher (http://emboss.bioinformatics.nl/cgi-bin/emboss/dotmatcher), primerjavo zaporedij pa naredite pri različnih vrednostih za velikost okna. Kot matrico zamenjav uporabite EBLOSUM62, kar je privzeta matrika za aminokislinska zaporedja.
Kako se spreminja občutljivost in količina šuma pri različnih vrednostih okna?
Ali v točkovnem diagramu opazite kakšen izrazit vzorec? Razložite!

Odgovor

Zaporedji za omenjena proteina najdemo v UniProt pod kodo:

- P13591 za N-CAM-1: http://www.uniprot.org/uniprot/P13591
- Q9UBF9 za miotilin: http://www.uniprot.org/uniprot/Q9UBF9

Z uporabo privzetih nastavitev (velikost okna – window size 10; mejna vrednost za prikaz - treshold 23) lahko opazimo največ vzporednih diagonal, ki nakazujejo podobnost med N-končno polovico NCAM-1 in C-končno polovico miotilina. Pri ogledu anotacij v bazi UniProt (Sequence Annotations (Features) --> Regions) za ta proteina (povezavi zgoraj) vidimo, da ima v tem delu NCAM-1 5 ponovitev domene imunoglobulinskega tipa (Ig-like C2-type), miotilin pa v omenjenem delu dve taki ponovitvi. Za detekcijo takih podobnosti oz. ponovitev domen ali pa homolognih domen ponavadi nastavimo velikost okna na vrednost, ki je približno enaka velikosti ponovitve. Če torej nastavimo velikost okna na npr. 90, dobimo naslednji diagram (ob privzeti vrednosti za treshold):



Z večanjem velikosti okna tako zmanjšamo šum, a hkrati tudi občutljivost, saj ne detektiramo več krajših podobnih segmentov.

V07-03

ALGORITEM NEEDLEMAN-WUNSCH

V Excel-u "sprogramirajte" globalno prileganje dveh zaporedij po osnovnem algoritmu Needleman-Wunsch. Pomagajte si slikami z uvoda v vajo (pdf).

Odgovor

Rešitev naloge je v datoteki "V07 - Naloge - RESENO.xlsx".

V07-04

GLOBALNO IN LOKALNO PRILEGANJE

Kateri tip prileganja (globalno ali lokalno) je primernejši za primerjavo aminokislinskih zaporedij iz posameznega para?
Par A: izoobliki 1 in 2 človeške karbonske anhidraze
Par B: človeška proteina nidogen-2 in IGFBP-1 (Insulin Growth Factor-Like Binding Protein 1)
Par C: človeški fibronektin in podenota beta mišjega receptorja za interleukin 6
Namig: Zaporedja omenjenih proteinov poiščite v bazi UniProt, oglejte si anotacije v tej bazi (regije oz. domene!), nato pa analizirajte zaporedja še s Prosite.

Za par B izdelajte tako lokalno (Smith-Waterman oz. Water (EMBOSS)) kot tudi globalno prileganje (Needleman-Wunsch oz. Needle (EMBOSS)). Orodja so na tej povezavi: http://www.ebi.ac.uk/Tools/psa/. V ustrezna okenca prilepite zaporedja v formatu FASTA, pri tem pa v naslovno vrstico zaporedja (za znak >) vpišite kratko ime proteina. Oglejte si "izhod" (output) obeh programov – zgoraj so navedeni uporabljeni parametri, spodaj pa na kratko lastnosti prileganja (dolžina, identičnost, podobnost, vrzeli). Mimogrede: EMBOSS pomeni European Molecular Biology Open Software Suite. Oglejte si "izhod" (output) obeh programov – zgoraj so navedeni uporabljeni parametri, spodaj pa na kratko lastnosti prileganja (dolžina, identičnost, podobnost, vrzeli).

Odgovor

Par A: uporabimo **globalno** zaporedje, saj sta zaporedji skoraj enako dolgi (to vidimo v UniProt), prav tako pričakujemo precejšnjo podobnost vzdolž celotnih zaporedij, saj gre za izoobliko istega encima; podobnost preverimo tako, da dejansko naredimo globalno prileganje.
Dolžino zaporedij razberemo, kot omenjeno, iz njunih zapisov v bazo UniProt, ki vsebuje še ostale podatke, ki nakazujejo na veliko podobnost med proteinoma:

- karbonska anhidraza 1: http://www.uniprot.org/uniprot/P00915
- karbonska anhidraza 2: http://www.uniprot.org/uniprot/P00918

Par B:
V UniProt poiščemo zapisa:

- nidogen-2: http://www.uniprot.org/uniprot/Q14112

- IGFBP-1: http://www.uniprot.org/uniprot/P08833

Že iz UniProt zapisov oz. anotacij vidimo, da oba proteina vsebujeta tiroglobulinsko domeno tipa 1 (thyroglobulin type-1), ki se v nidogenu pojavi dvakrat, v IGFBP-1 pa enkrat; ostale domene/regije so različne. V tem primeru je bolj smiselno uporabiti **lokalno** prileganje; zaporedji sta namreč različno dolgi, prav tako pa pričakujemo znatno podobnost zgolj v eni regiji, ki ustreza tiroglobulinski domeni.

Par C:
V UniProt poiščemo zapisa:

- fibronektin: http://www.uniprot.org/uniprot/P02751
- podenota beta mišjega receptorja za interleukin 6: http://www.uniprot.org/uniprot/Q00560

Podobno kot v primeru C gre za proteina z zelo različno dolgimi zaporedji, ki vsebujeta različne domene, nekatere od teh a so podobne – gre za fibronektinsko domeno tipa III, ki se v fibronektinu ponovi 16-krat (ponovitve so oštevilčene), v receptorju za interleukin 6 pa 5-krat, pričakujemo pa lahko tudi oločeno podobnost med fibronektinskimi domenami tipa III receptorja za interleukin 6 ter fibronektinskimi domenami tipa I in II v fibronektinu. Za ta primer je torej prav tako primernejše **lokalno** prileganje.

Prileganje za par B:

LOKALNO

```
#=======================================
#
# Aligned_sequences: 2
# 1: nidogen2
# 2: IGFBP1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 237
# Identity:      54/237 (22.8%)
# Similarity:    73/237 (30.8%)
# Gaps:          69/237 (29.1%)
# Score: 110.0
#
#
#=======================================

nidogen2         812 CGPNSV-CINLPGSYRCECRSGYEFADDRHTCILITPPANPCEDGSHTCA    860
                     |.|.|. |..:..|..|.|            |.:...|...      .|.
IGFBP1            41 CPPVSASCSEVTRSAGCGC------------CPMCALPLGA------ACG     72

nidogen2         861 PAGQARCVHHGGSTFSCACLPGYAGDGHQCTD-----VDECSENRCHPAA    905
                     .| .|||..    ..||..|||.....|..|.     |.|...:..| ||
IGFBP1            73 VA-TARCAR----GLSCRALPGEQQPLHALTRGQGACVQESDASAPH-AA    116

nidogen2         906 TCYNTPGSFSCRCQPGYYGDGFQCIP----------------DSTSSL--    937
                     ...:.....|.........|.|..:.                 |.:.:|
IGFBP1           117 EAGSPESPESTEITEEELLDNFHLMAPSEEDHSILWDAISTYDGSKALHV    166

nidogen2         938 -------TPCE-------QQQRHAQAQYAYPGARFHIPQCDEQGNFLPLQ    973
                           .||.       :....||.......::|::|.|::.|.:...|
IGFBP1           167 TNIKKWKEPCRIELYRVVESLAKAQETSGEEISKFYLPNCNKNGFYHSRQ    216

nidogen2         974 CH----GSTGFCWCVDP-DGHEVPGTQTPPGSTPPHC   1005
                     |.    |..|.||||.| :|..:||:....|.  |:|
IGFBP1           217 CETSMDGEAGLCWCVYPWNGKRIPGSPEIRGD--PNC    251


#---------------------------------------
#---------------------------------------
```

GLOBALNO

```
#=======================================
#
# Aligned_sequences: 2
# 1: nidogen2
# 2: IGFBP1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1414
# Identity:      61/1414 ( 4.3%)
# Similarity:    98/1414 ( 6.9%)
# Gaps:        1194/1414 (84.4%)
# Score: 90.5
#
#
#=======================================

nidogen2           1 MEGDRVAGRPVLSSLPVLLLLPLLMLRAAALHPDELFPHGESWGDQLLQE     50
```

```
IGFBP1            1 --------------------------------------------------      0

nidogen2         51 GDDESSAVVKLANPLHFYEARFSNLYVGTNGIISTQDFPRETQYVDYDFP    100

IGFBP1            1 --------------------------------------------------      0

nidogen2        101 TDFPAIAPFLADIDTSHGRGRVLYREDTSPAVLGLAARYVRAGFPRSARF    150

IGFBP1            1 --------------------------------------------------      0

nidogen2        151 TPTHAFLATWEQVGAYEEVKRGALPSGELNTFQAVLASDGSDSYALFLYP    200

IGFBP1            1 --------------------------------------------------      0

nidogen2        201 ANGLQFLGTRPKESYNVQLQLPARVGFCRGEADDLKSEGPYFSLTSTEQS    250

IGFBP1            1 --------------------------------------------------      0

nidogen2        251 VKNLYQLSNLGIPGVWAFHIGSTSPLDNVRPAAVGDLSAAHSSVPLGRSF    300

IGFBP1            1 --------------------------------------------------      0

nidogen2        301 SHATALESDYNEDNLDYYDVNEEEAEYLPGEPEEALNGHSSIDVSFQSKV    350

IGFBP1            1 --------------------------------------------------      0

nidogen2        351 DTKPLEESSTLDPHTKEGTSLGEVGGPDLKGQVEPWDERETRSPAPPEVD    400

IGFBP1            1 --------------------------------------------------      0

nidogen2        401 RDSLAPSWETPPPYPENGSIQPYPDGGPVPSEMDVPPAHPEEEIVLRSYP    450

IGFBP1            1 --------------------------------------------------      0

nidogen2        451 ASGHTTPLSRGTYEVGLEDNIGSNTEVFTYNAANKETCEHNHRQCSRHAF    500

IGFBP1            1 --------------------------------------------------      0

nidogen2        501 CTDYATGFCCHCQSKFYGNGKHCLPEGAPHRVNGKVSGHLHVGHTPVHFT    550

IGFBP1            1 --------------------------------------------------      0

nidogen2        551 DVDLHAYIVGNDGRAYTAISHIPQPAAQALLPLTPIGGLFGWLFALEKPG    600

IGFBP1            1 --------------------------------------------------      0

nidogen2        601 SENGFSLAGAAFTHDMEVTFYPGEETVRITQTAEGLDPENYLSIKTNIQG    650

IGFBP1            1 --------------------------------------------------      0

nidogen2        651 QVPYVSANFTAHISPYKELYHYSDSTVTSTSSRDYSLTFGAINQTWSYRI    700

IGFBP1            1 --------------------------------------------------      0

nidogen2        701 HQNITYQVCRHAPRHPSFPTTQQLNVDRVFALYNDEERVLRFAVTNQIGP    750
                                 ..::.|.||:        ::...:|.|:|
IGFBP1            1 --------------------MSEVPVARVW--------LVLLLLTVQVG-     21

nidogen2        751 VKEDSDPTPGNPCYDGSHMCDTTARCHPGTGVDYTCECASGYQGDGRNCV    800
                          .|.|.|........:..|.|.|     .:..|:...:..|..|.
IGFBP1           22 ------VTAGAPWQCAPCSAEKLALCPP-----VSASCSEVTRSAGCGCC     60

nidogen2        801 D------ENECATGFHRCGPNSVCINLPGSYRCECRSGYEFADDRHTCIL    844
                .       ...|.....||.....|..|||    |.:..:.....:..|:.
IGFBP1           61 PMCALPLGAACGVATARCARGLSCRALPG----EQQPLHALTRGQGACVQ    106

nidogen2        845 ITPPANP--CEDGS------------------HTCAPAGQARCVHHGG-S    873
                .:..:.|  .|.||                   |..||:.:...:.... |
IGFBP1          107 ESDASAPHAAEAGSPESPESTEITEEELLDNFHLMAPSEEDHSILWDAIS    156

nidogen2        874 TFSCACLPGYAGDGHQCTDVDECSENRCHPAATCYNTPGSFSCRCQPGYY    923
                |:          ||.:...|....:         :..|    ||.:
IGFBP1          157 TY----------DGSKALHVTNIKK---------WKEP----CRIE----    179

nidogen2        924 GDGFQCIPDSTSSLTPCEQQQRHAQAQYAYPGARFHIPQCDEQGNFLPLQ    973
                             |....:....||.......::|::|.|::.|.:...|
IGFBP1          180 -------------LYRVVESLAKAQETSGEEISKFYLPNCNKNGFYHSRQ    216

nidogen2        974 CH----GSTGFCWCVDP-DGHEVPGTQTPPGSTPPHC-------GPSPEP   1011
                |.    |..|.||||.| :|..:||:....|.  |:|       .
IGFBP1          217 CETSMDGEAGLCWCVYPWNGKRIPGSPEIRGD--PNCQIYFNVQN-----    259

nidogen2       1012 TQRPPTICERWRENLLEHYGGTPRDDQYVPQCDDLGHFIPLQCHGKSDFC   1061

IGFBP1          260 --------------------------------------------------    259

nidogen2       1062 WCVDKDGREVQGTRSQPGTTPACIPTVAPPMVRPTPRPDVTPPSVGTFLL   1111
```

```
IGFBP1           260 --------------------------------------------------    259

nidogen2        1112 YTQGQQIGYLPLNGTRLQKDAAKTLLSLHGSIIVGIDYDCRERMVYWTDV   1161

IGFBP1           260 --------------------------------------------------    259

nidogen2        1162 AGRTISRAGLELGAEPETIVNSGLISPEGLAIDHIRRTMYWTDSVLDKIE   1211

IGFBP1           260 --------------------------------------------------    259

nidogen2        1212 SALLDGSERKVLFYTDLVNPRAIAVDPIRGNLYWTDWNREAPKIETSSLD   1261

IGFBP1           260 --------------------------------------------------    259

nidogen2        1262 GENRRILINTDIGLPNGLTFDPFSKLLCWADAGTKKLECTLPDGTGRRVI   1311

IGFBP1           260 --------------------------------------------------    259

nidogen2        1312 QNNLKYPFSIVSYADHFYHTDWRRDGVVSVNKHSGQFTDEYLPEQRSHLY   1361

IGFBP1           260 --------------------------------------------------    259

nidogen2        1362 GITAVYPYCPTGRK   1375

IGFBP1           260 --------------    259

#---------------------------------------
#---------------------------------------
```

Iz primerjave obeh prileganj vidimo, da je delež identičnosti oz. podobnosti pri lokalnem prileganju (22,8 % oz. 30,8 %) večji kot pri globalnem prileganju (4,3 % oz. 6,9 %), kar potrjuje naše predvidevanje, da sta proteina podobna samo lokalno (v konkretnem primeru v eni domeni, to je tiroglobulinski domeni), vzdolž celotne polipeptidne verige pa niti ne. Tako tudi vrzeli (gaps) predstavljajo kar 84,4 % globalnega prileganja, pri lokalnem prileganju, kjer imamo prikazan le najbolj podoben del, pa je ta delež 29,1 %. Odsek zaporedij, prikazan pri lokalnem prileganju, oz. dolžina lokalnega prileganja je odvisna od parametrov, ki jih nastavimo (pod "More options" pod okenci za vnos zaporedij na začetni strani).

V07-05

LOKALNE PODOBNOSTI

S programom LALIGN za aminokislinska zaporedja (http://www.ebi.ac.uk/Tools/psa/lalign) primerjajte zaporedje človeškega kalmodulina s samim seboj. Kaj opazite? Komentirajte/razložite rezultat!

Odgovor

V UniProt poiščemo zapis za človeški kalmodulin; pravilen zadetek iskanja po bazi je ta:
http://www.uniprot.org/uniprot/P62158
(Pri iskanju dombimo še več drugih zadetkov, med drugim tudi kalmodulinu podobne proteine (Calmodulin-like protein), ki pa niso kalmodulin, zato moramo biti pri izbiri zadetka za nadaljnje delo zelo pozorni, saj ni vedno ta prvi zadetek tisti pravi!)
Že pri pregledu anotacij v bazi UniProt ugotovimo, da je kalmodulin sestavljen iz štirih onovitev t.i. rok EF (EF-hand, ki so oštevilčene od 1 do 4):

Slika 1

Anotacije za kalmodulin

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view
Molecule processing					
☐	Initiator methionine	1	1	Removed Ref.13 Ref.14	
☐	Chain	2 – 149	148	Calmodulin	
Regions					
☐	Domain	8 – 43	36	EF-hand 1	
☐	Domain	44 – 79	36	EF-hand 2	
☐	Domain	81 – 116	36	EF-hand 3	
☐	Domain	117 – 149	33	EF-hand 4	
☐	Calcium binding	21 – 32	12	1	
☐	Calcium binding	57 – 68	12	2	
☐	Calcium binding	94 – 105	12	3	
☐	Calcium binding	130 – 141	12	4	

štiri roke EF

relativna razporeditev in obseg rok EF

Hkrati vidimo, da so te štiri roke EF, ki pravzaprav zaobsegajo celoten protein, približno enako dolge (~35 aminokislinskih ostankov).

Pri primerjavi ak-zaporedja kalmodulina s samim seboj kot rezultat dobimo več različnih lokalnih prileganj, ki jih lahko vizualiziramo s klikom na "Visual Output" blizu zgornjega dela strani z rezultati. Ta prikaz nekako spominja na točkovni diagram, čeprav to ni, ampak prikazuje, katere regije enega zaporedja so podobne drugi regiji (tam so narisane črte) ter kolikšna je podobnost (barva črte, ki je povezana z lestvico točk za prileganje). Vidimo naslednje:

- diagonalo, ki ustreza prileganju zaporedja s samim seboj, 100% identičnost --> visok *score* (>200),
- trikotni področji desno pod in levo nad diagonalo sta seveda simetrični, saj zaporedje primerjamo samega s seboj,
- dodatne diagonalne linije predstavljajo primerjavo oz. prileganje posameznih rok EF (oz. dveh ali treh rok EF) s drugimi v istem proteini – glej priloženo sliko.

Omenjene ponovitve rok EF se torej v tem diagramu, ki grafično ponazarja lokalna prileganja, kažejo kot dodatne diagonale (zraven glavne diagonale), podobno, kot se ponovitve kažejo v točkovnem diagramu.

Slika 1

Kalmodulin – lokalno prileganje / Visual Output