8. vaja: Naloge z odgovori oz. postopki reševanja

V08-01

BLAST: IDENTIFIKACIJA HOMOLOGOV

Poiščite človeški ortolog mišjega inzulinu podobnega rastnega faktorja I (ang. *insulin-like growth factor I*, IGF-1, izooblika A). Za iskanje uporabite dve različni obliki programa BLAST ter primerjajte hitrost iskanja in rezultate – komentirajte.

Potek:

1. V bazi UniProt poiščite aminokislinsko zaporedje za omenjeni protein.
2. Iskalno zaporedje prilepite v ustrezno okence ustreznega programa ter izberite ustrezen organizem.

Odgovor

V bazi UniProt lahko preko iskalnega pojma insulin-like growth factor I mus musculus (lahko bi uporabili tudi npr. IGF-1 mus musculus) dobimo pravi zadetek: http://www.uniprot.org/uniprot/P05017
Bodite pozorni, da slučajno ne izberete receptorja za omenjeni protein (angl. *Insulin-like growth factor 1 receptor*)! Za iskanje bomo v nadaljevanju uporabili kar celotno zaporedje tega hormona, torej vseh 153 aminokislinskih ostankov, in sicer izoobliko 1 (poimenovana kot isoform *IGF-IA*).
Ker so aminokislinska zaporedja ponavadi bolj ohranjena kot nukleotidna bomo za iskanje uporabili naslednja programa BLAST: **blastp** in **tblastn**. (s pomočjo slednjega lahko dejansko najdemo proteinske homologe v bazah še ne anotiranih nukleotidnih zaporedij). Pri obeh bomo pri iskanju kot organizem izbrali Homo sapiens (taxid:9606), za iskanje pa lahko uporabimo bazo nr oz. nr/nt (non-redundant). Z omejitvijo npr. na bazo UniProt bi sicer tvegali, da zgrešimo ortolog (recimo, da še le-ta ni anotiran), kar je pa v konkretnem primeru zelo malo verjetno, saj je ta protein že pri miši dobro raziskan).

Iskanje z blastp je hitrejše, saj je dejansko v bazi manj zaporedij. V bazi tblastn več zaporedij, saj je vsako nukleotidno zaporedje v bazi nt zaporedij prevedeno v aminokislinsko zaporedje (faktor 6), je pa res, da se ob prevejanju število znakov zmanjša za faktor 3.

Z **blastp** najdemo več zadetkov, ki so vsi zelo podobni, trije za pokrivajo naše iskalno zaporedje v celoti (100% coverage):

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
insulin-like growth factor I precursor	297	297	100%	1e-102	92%	1001199A
insulin-like growth factor I isoform 4 preproprotein [Homo sapiens] >emb|CAA40092.1| IGF-1a [Homo sapien	296	296	100%	1e-102	92%	NP_000609.1
insulin-like growth factor 1 (somatomedin C); insulin-like growth factor 1 (somatomedia C) variant [Homo sap	295	295	100%	1e-101	92%	BAD92421.1
insulin-like growth factor I isoform 1 preproprotein [Homo sapiens]	261	261	87%	1e-88	92%	NP_001104753.1

Pravzaprav gre za enake zadetke, identičnost ter identičnost+podobnost je za vse enaka (140/153 oz. 144/153), gre le za to, da so to različni vnosi v bazo in imajo nekoliko drugačno številčenje aminokislinskih ostankov. Vzamemo kar prvi zadetek, zaporedje človeškega ortologa je pod kodo 1001199A v bazi NCBI Protein. Če iskanje ponovimo, le da kot bazo izberemo UniProtKB, dobimo kot najboljši zadetek zapis v bazo UniProt: P05019 (http://www.uniprot.org/uniprot/P05019). Pri tej bazi dobimo manj zadetkov (saj je manjša!), drugi zadetek je IGF-II, 3. in 4. zadetek sta dve obliki inzulina (a je podobnost bistveno manjša!), naslednji zadetki pa več niso relevantni (glede na vrednost E).

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Insulin-like growth factor I; Short=IGF-I; AltName: Full=Mechano growth factor; Short=MGF; AltNam	262	262	87%	2e-89	92%	P05019.1
RecName: Full=Insulin-like growth factor II; Short=IGF-II; AltName: Full=Somatomedin-A; Contains: RecName: Full=	84.3	84.3	55%	2e-20	55%	P01344.1
RecName: Full=Insulin; Contains: RecName: Full=Insulin B chain; Contains: RecName: Full=Insulin A chain; Flags: I	43.5	43.5	38%	3e-06	35%	P01308.1
RecName: Full=Insulin, isoform 2; AltName: Full=INS-IGF2 readthrough transcript protein	34.7	34.7	16%	0.012	56%	F8WCM5.1
RecName: Full=Mediator of RNA polymerase II transcription subunit 13-like; AltName: Full=Mediator complex subun	29.6	29.6	64%	0.92	21%	Q71F56.1
RecName: Full=E3 ubiquitin-protein ligase LRSAM1; AltName: Full=Leucine-rich repeat and sterile alpha motif-conta	28.1	28.1	23%	2.3	33%	Q6UWE0.1
RecName: Full=Multidrug resistance-associated protein 4; AltName: Full=ATP-binding cassette sub-family C membe	27.3	27.3	16%	4.6	42%	O15439.3
RecName: Full=Sacsin; AltName: Full=DnaJ homolog subfamily C member 29; Short=DNAJC29	26.6	26.6	25%	8.6	37%	Q9NZJ4.2

S **tblastn** in iskanjem po človeških zaporedjih v bazi nr/nt traja iskanje dlje časa (večja baza), pa tudi identičnih zadetkov je več, saj so v bazi nr vnešena ista zaporedja večkrat.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
Synthetic construct DNA, clone: pF1KB3912, Homo sapiens IGF1 gene for insulin-like growth factor IA precursor	297	297	100%	3e-103	92%	AB384874.1
Homo sapiens full open reading frame cDNA clone RZPDo834H0732D for gene IGF1, insulin-like growth factor 1	296	296	100%	3e-103	92%	CR541861.1
Synthetic construct Homo sapiens clone IMAGE:100064018, MGC:193197 insulin-like growth factor 1 (somatom	296	296	100%	1e-102	92%	BC160082.1
H.sapiens mRNA for IGF-1a	296	296	100%	2e-102	92%	X56773.1
Homo sapiens cDNA, FLJ92522, Homo sapiens insulin-like growth factor 1 (somatomedin C) (IGF1),mRNA	296	296	100%	2e-102	92%	AK312231.1
Homo sapiens mRNA for insulin-like growth factor 1A precursor, complete CDS	297	297	100%	4e-102	92%	X00173.1
Human insulin-like growth factor I mRNA, complete cds	297	297	100%	5e-102	92%	M29644.1
Human insulin-like growth factor mRNA, complete cds	296	296	100%	3e-100	92%	M27544.1
Homo sapiens insulin-like growth factor 1 (somatomedin C), mRNA (cDNA clone MGC:157712 IMAGE:4012926	296	296	100%	3e-93	92%	BC152321.1
Homo sapiens mRNA for insulin-like growth factor 1 (somatomedin C); insulin-like growth factor 1 (somatomedia	296	296	100%	5e-93	92%	AB209184.1
Human IGF-I mRNA for insulin-like growth factor I	296	296	100%	3e-92	92%	X57025.1
Homo sapiens insulin-like growth factor 1 (somatomedin C) (IGF1), transcript variant 4, mRNA	296	296	100%	3e-92	92%	NM_000618.3
Homo sapiens insulin-like growth factor 1 (somatomedin C), mRNA (cDNA clone MGC:166952 IMAGE:8860206)	262	262	87%	9e-89	92%	BC148266.1
H.sapiens mRNA for IGF-1b	262	262	87%	2e-88	92%	X56774.1
Human insulin-like growth factor I (IGF-I) mRNA, complete cds	259	259	86%	9e-88	92%	M37484.1
Homo sapiens insulin-like growth factor 1 (somatomedin C) (IGF1), transcript variant 3, mRNA	262	262	87%	2e-87	92%	NM_001111285.1
Human insulin-like growth factor IB (IGF-IB) cDNA to mRNA	262	262	87%	9e-87	92%	M11568.1
Homo sapiens insulin-like growth factor 1 (somatomedin C) (IGF1), transcript variant 1, mRNA	259	301	100%	3e-79	92%	NM_001111283.1
Homo sapiens insulin-like growth factor 1 (somatomedin C) (IGF1), transcript variant 2, mRNA	259	259	86%	3e-79	92%	NM_001111284.1
Human alternatively spliced human insulin-like growth factor-I (IGF-I) mRNA, partial cds	225	259	85%	3e-75	92%	U40870.1
PREDICTED: Homo sapiens insulin-like growth factor 1 (somatomedin C) (IGF1), transcript variant X1, mRNA	225	225	73%	4e-73	92%	XM_005268835.1

V08-02

BLAST: IDENTIFIKACIJA FUNKCIJE

Iz nekega še neopisanega mikroorganizma smo izolirali genomsko DNA, pomnožili en odsek z uporabo degeneriranih začetnih oligonukleotidov (njihovo tarčno zaporedje obsega osrednjo regijo neke skupine encimov), ter določili zaporedje tega pomnoženega fragmenta (GenBank zapis spodaj). Ugotovite, za kateri tip encima gre.

Razmislite:

1. Kodira določeno zaporedje za protein ali ne?
2. Je smiselno iskati ORF v tem zaporedju?
3. Katerega od programov BLAST je najbolj smiselno uporabiti?

```
LOCUS       CLONE12.DNA     609 BP DS-DNA             UPDATED   06/14/98
DEFINITION  UWGCG file capture
ACCESSION   -
KEYWORDS    -
SOURCE      -
COMMENT     Non-sequence data from original file:
BASE COUNT      174 A    116 C    162 G    157 T      0 OTHER
ORIGIN      ?
    clone12.dna Length: 609   Jun 13, 1998 - 03:39 PM   Check: 6014 ..
        1 AACGGGCACG GGACGCATGT AGCTGGAACA GTGGCAGCCG TAAATAATAA TGGTATCGGA
       61 GTTGCCGGGG TTGCAGGAGG AAACGGCTCT ACCAATAGTG GAGCAAGGTT AATGTCCACA
      121 CAAATTTTTA ATAGTGATGG GGATTATACA AATAGCGAAA CTCTTGTGTA CAGAGCCATT
      181 GTTTATGGTG CAGATAACGG AGCTGTGATC TCGCAAAATA GCTGGGGTAG TCAGTCTCTG
      241 ACTATTAAGG AGTTGCAGAA AGCTGCGATC GACTATTTCA TTGATTATGC AGGAATGGAC
      301 GAAACAGGAG AAATACAGAC AGGCCCTATG AGGGGAGGTA TATTTATAGC TGCCGCCGGA
      361 AACGATAACG TTTCCACTCC AAATATGCCT TCAGCTTATG AACGGGTTTT AGCTGTGGCC
      421 TCAATGGGAC CAGATTTTAC TAAGGCAAGC TATAGCACTT TTGGAACATG GACTGATATT
      481 ACTGCTCCTG GCGGAGATAT TGACAAATTT GATTTGTCAG AATACGGAGT TCTCAGCACT
      541 TATGCCGATA ATTATTATGC TTATGGAGAG GGAACATCCA TGGCTTGTCC ACATGTCGCC
```

```
      601 GGCGCCGCC
//
```

Odgovor

Za iskanje je najbolj smiselno uporabiti nukleotidno zaporedje direktno brez predhodnega iskanja ORF, saj ne vemo, kateri ORF naj vzamemo (ni nujno, da je najdaljši tisti ta pravi), prav tako pa ne moremo vedeti, ali imamo celotno zaporedje mRNA oz. le-to vsebuje celoten ORF. Za iskanje pa uporabimo orodje **blastx** (lahko bi načeloma uporabili tblastx, a tu dobimo veliko nerelevantnih zadetkov – prevedenih zaporedij, za katera ne vemo, če dejansko kodirajo nek protein).
Od zadetkov gre pri večini primerov za "hipotetične" proteine, pri edem izmed njih (2. po vrsti) pa piše, da gre za inhibitor serinskih proteaz. Ker imajo ti zadetki zelo podobno max score in vrednost E ter % identičnosti upravičeno sklepamo, da gre pri vseh primerih za inhibitor proteaz, kar pomeni, da tudi zaporedje, ki ga imamo, nosi zapis za tovrstni inhibitor.

V08-03

BLAST: ISKANJE PROTEINOV Z ZNANIMI 3D STRUKTURAMI

Je znana struktura človeškega opsina-2? Če ni – je znana struktura homolognega proteina? Katerega in iz katerega organizma? Kolikšen je odstotek identičnosti? Moramo za primerjavo našega zaporedja z najboljšim zadetkom v celoti narediti globalno prileganje ali ne – zakaj?

Kaj pa za človeški protein klavdin-1? Je morda za ta protein znana struktura (deloma ali v celoti)? Če ni, je znana struktura homolognega proteina (deloma ali v celoti)?
Potek:

1. V bazi UniProt poiščite aminokislinsko zaporedje za omenjeni protein.
2. Iskalno zaporedje prilepite v ustrezno okence programa ustreznega programa ter izberite ustrezno bazo (pdb). Polje "Organism" pustite prazno!

Odgovor

Opsin-2

Poiščemo zaporedje človeškega opsina-2 v UniProt (http://www.uniprot.org/uniprot/P08100) --> s tem zaporedjem iščemo z blastp (protein blast) po bazi pdb (organizma ni potrebno izbirati). Kot najboljši zadetek je goveji opsin (torej homolog), struktura človeškega opsina-2 pa torej ni znana, saj bi bil tak zadetek uvrščen najvišje.
Odstotek identičnosti med človeškim in govejim opsinom je 93%; v tem primeru lahko to preberemo kar iz zadetkov blastp, saj imamo coverage 100%. V nasprotnem primeru (da bi prileganje zajemalo le krajši del zaporedja) bi morali narediti globalno prileganje celotnih zaporedij.

Klavdin-1

Naredimo podobno kot v primeru človešjega opsina-1. Kot zadetek dobimo mišji klavdin-15, ki je homolog človeškega klavdina, torej je znana struktura homologa, ni pa znana struktura človeškega klavdina-1.

V08-04

BLAST: IDENTIFIKACIJA PARALOGOV

Identificirajte paraloge mišjega proteina EpCAM. Koliko jih je?

Odgovor

Poiščemo zaporedje mišjega proteina EpCAM (http://www.uniprot.org/uniprot/Q99JW5) --> s tem zaporedjem iščemo po bazi proteinskih zaporedij – orodje blastp. Iščemo po mišjih zaporedjih.
Najdemo identičen zadetek (torej mišji EpCAM) ter pa paralog Trop2.
Najbolj zanesljivo bi sicer bilo iskanje po bazi prevedenih nukleotdinoh zaporedij (tblastn), saj ni nujno, da je zaporedje paraloga (oz. vseh paralogov) v bazi proteinskih zaporedij, a v slednjem primeru dobimo veliko "podvojenih" zadetkov.

V08-05

BLAST: ZA KATERI PROTEIN KODIRA DANO NUKLEOTIDNO ZAPOREDJE?

V laboratoriju smo našli epico, na pokrovčku pa je pisalo samo "pET-22 + insert". Želimo ugotoviti, za kateri protein vsebuje omenjeni plazmidni konstrukt zapis. Vemo, da je pET-22 plazmidni vektor za izražanje proteinov v *E. coli*, to pa nam je omogočilo sekvenciranje inserta (vključka) z uporabo oligonukleotida, ki se veže na vektor tik pred multiplo klonirno regijo. Pri sekvenciranju smo dobili zaporedje, v katerem smo identificirali najdaljši ORF (zaporedje spodaj).

Odgovorite na vprašanja:

1. Kateri protein bi torej lahko izrazili z uporabo najdenega plazmida?
2. Kateri program BLAST ste uporabili in zakaj? Kaj pa, če bi uporabili katerega drugega – kakšen je rezultat in zakaj (komentirajte)?

```
>sequencing_result-ORF
ATGGGTATGACACGTATGCTGCTGGAGTGCTCCCTGAGCGATAAGCTTTGCGTGATCCAGGAGAAACAATATGAG
GTGATTATCGTGCCGACCTTGTTGGTGACCATCTTTCTTATTCTGCTGGGCGTGATTCTGTGGCTGTTTATTCGC
GAGCAACGCACGCAACAGCAGCGCTCTGGCCCACAGGGAATCGCGCCCGTGCCTCCGCCACGCGATCTCTCCTGG
GAAGCGGGCCACGGGGGAAACGTAGCCCTCCCCCTGAAAGAAACGTCGGTGGAAAATTTCCTGGGGGCGACTACC
CCAGCACTCGCGAAACTGCAGGTTCCGCGCGAACAGTTGTCTGAAGTATTGGAGCAGATTTGCAGCGGGTCATGT
GGGCCTATTTTTCGTGCGAACATGAACACAGGCGATCCTAGTAAACCCAAATCTGTCATCCTGAAAGCCCTGAAG
GAACCGGCCGGATTGCATGAAGTTCAGGACTTTCTTGGCCGTATTCAGTTTCACCAGTATCTTGGCAAACATAAA
AACTTGGTGCAACTGGAAGGCTGCTGCACCGAGAAACTGCCGCTGTACATGGTATTGGAAGACGTCGCCCAGGGG
GATCTGCTGTCATTCCTGTGGACATGTCGCCGCGACGTGATGACTATGGATGGCCTGCTGTACGATCTGACTGAA
AAACAGGTGTACCATATCGGTAAACAGGTGTTACTCGCGCTTGAATTTCTGCAGGAAAAGCACCTGTTCCATGGT
GACGTGGCGGCCCGCAATATTTTAATGCAGTCTGATTTGACCGCGAAACTGTGCGGGTTGGGTTTAGCGTATGAA
GTTTATACGCGCGGCGCTATTTCTTCCACCCAGACAATCCCGCTCAAATGGTTAGCACCCGAGCGCCTGCTTCTT
CGCCCAGCATCGATCCGCGCGGATGTTTGGTCTTTCGGCATCTTACTGTACGAGATGGTCACTCTTGGCGCGCCA
CCCTATCCGGAAGTCCCGCCCACCAGTATCCTGGAGCATCTGCAGCGCCGCAAAATTATGAAGCGCCCAAGTAGC
TGCACGCATACAATGTACTCTATCATGAAAAGCTGCTGGCGTTGGCGCGAAGCGGACCGCCCGTCCCCGCGCGAA
CTGCGCCTTCGTCTTGAAGCAGCCATTAAAACGGCCGATGATGAGGCCGTCCTGCAGGTCCCGGAATTGGTTGTT
CCGGAATTATACGCTGCGGTGGCCGGGATTCGCGTGGAAAGCCTGTTTTATAACTACAGTATGCTGTAA
```

Odgovor

Glede na to, da imamo popolen bralni okvir, ga lahko z orodjem ORF Finder ali pa z Expasy Translate prevedemo v aminokislinsko zaporedje. Z iskanjem po proteinskih zaporedij z blast (blastp) ugotovimo, da gre za tirozin-protein kinazo STYK1 iz človeka (100% ujemanje, vrednost E je 0).

V primeru, da na ta način ne bi našli zadetka, bi lahko iskanje ponovili tako, da bi s proteinskim zaporedjem iskali po bazi prevedenih nukleotidnih zaporedij (za vsak primer, če zapisa ni v bazi proteinskih zaporedij) – orodje tblastn. V našem primeru dobimo enak zadetek.

Če z danim nukleotidnim zaporedjem iščemo direktno v bazi nukleotidnih zaporedij sicer najdemo enak zadetek, a ujemanje ni 100%! To pa zato, ker je v danem zapisu raba kodonov optimirana, a samo zaporedje še vedno kodira za isto proteinsko zaporedje kot pred optimizacijo rabe kodonov.