19.9.2013

Rautalankamalleja sekvenssihakuihin STN:ssä

VTT, Tietoratkaisut, Riitta Housh

Tässä on esitetty lähinnä rautalankamalleja. Katso esimerkkejä ja lisätietoja sekvenssihauista http://www.stn-international.de/biosequence_searching.html

STN Internationa lwww.stn-international.de
c/o FIZ Karlsruhe
P.O. Box 2465, D-76012 Karlsruhe, Germany
Telephone: +49 7247 808 555
Telefax: +49 7247 808 259
Email: hlpdeskk@fiz-karlsruhe.de

VTT, Tietoratkaisut www.vtt.fi/service/inf/
Vuorimiehentie 5, Espoo
PL 1000, 02044 VTT
Riitta Housh Puh. 020 722 4381
Riitta Metsäkoivu Puh. 020 722 4372
Email: etunimi.sukunimi@vtt.fi

1. Sekvenssitietokannat

Registry/(H)CAplus

- 65 milj. sekvenssiä 63 patenttiviraston julkaisusta (Basic-julkaisusta, mutta 2008- PCT-hakemuksista myös aiemmasta kansallisesta hakemuksesta) ja yli 3000 lehdestä v. 1957-. Myös Genbank 2005 asti. Sekvenssien viive 3 – 4 viikkoa. Päivitys joka päivä.
- Sekvenssit Registryssä, julkaisujen tiedot (H)CAplussassa.
- Sekvenssin sijainti claimed/unclaimed PNTE-kentässä (Patent Annotations)
- BLAST (27 €), Sequence Code Match, SCM (Exact 6 €, Subsequence 27€)
- Osumasekvenssi Registry SQIDE 4 €, (H)CAplus HITSEQ 4 €
- Sekvenssivertailu BLAST-haun jälkeen saadusta BLAST-raportista (osumasekvenssien CAS-numerot, nimet ja vertailu haettuun sekvenssiin). Saadaan mukaan STN Expressin raporttiin Results/BLAST Report with Alignment Data, jos se tallennetaan erikseen, mutta ei näy online.

DGENE

- 9 milj. sekvenssiä 43 patenttivirastosta 1981- (vain Basic-julkaisusta). Viive 1 - 3 kk. Päivitys joka toinen viikko. Patenttivirastojen käyttämä tietokanta
- Julkaisun jokaisella sekvenssilllä oma viite, jossa oma tiivistelmä ja sekvenssin sijainti (PSL)
- Myös indeksoijan lisäämiä sekvenssejä, jotka puuttuvat muodollisista listauksista
- Sekvenssihaku antaa useita viitteitä samasta julkaisusta. Basic-julkaisun kaikki osumasekvenssit voidaan koota yhteen FSORTilla.
- RUN BLAST ja RUN GETSIM (FASTA) kumpikin 24 € tai BATCHinä 9 € + 30 €,
  RUN GETSEQ (Sequence Code Match SCM) 18 €
- Sekvenssivertailu: ALIGN (maksuton) (Ei kuulu ALL-formaattiin!); Osumasekvenssi SEQ ja SQIDE (myös identifiointi) 6,4 €; BIB 3,2 €; ALL 12,8 €

USGENE

- Sekvenssit US-patenttijulkaisuista v. 1981-. Viive 3 pv. Päivitys kerran viikossa. Uusimmat sekvenssit vain täältä. Myös patentin raukeamispäivä.
- Sekvenssin sijainti PSL-kentässä
- Jokaisella sekvenssillä on oma viite, joten sekvenssihaku voi antaa useita viitteitä samasta julkaisusta. US-perheen kaikki osumasekvenssit voidaan koota yhteen FSORTilla.
- RUN BLAST ja RUN FASTA (GETSIM), kumpikin 18 € tai 6 € +21 €,
  RUN GETSEQ (Sequence Code Match SCM) 18 €
- Sekvenssivertailu ALIGN (maksuton); Osumasekvenssi SEQ ja SQIDE (myös identifiointi) 3 €; BIB 1,55 €; ALL 5,7 €

PCTGEN

- 6 milj. sekvenssiä elektronisesti saatavilla olevista PCT-hakemuksista v. 2001-. Viive 1 pv. Päivitys kerran viikossa. Uusimmat sekvenssit vain täältä.
- Myös sekvenssin sijainti
- Jokaisella sekvenssillä on oma viite, joten sekvenssihaku voi antaa useita viitteitä samasta julkaisusta. PCT-julkaisun kaikki osumasekvenssit voidaan koota yhteen FSORTilla.
- RUN BLAST ja RUN GETSIM (FASTA) kumpikin 12 € tai BATCHinä 4 € + 14 €,
  RUN GETSEQ (Sequence Code Match SCM) 10 €
- Sekvenssivertailu: ALIGN (maksuton); Osumasekvenssi SEQ ja SQIDE (myös identifiointi) 1,3 €; BIB 1 €, ALL 2,3 €

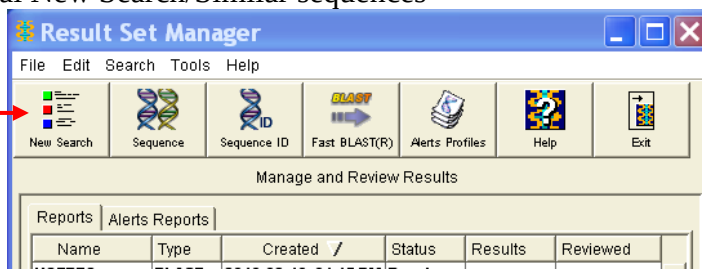
2. BLAST

Registry STN Expressin CAS Registry BLASTilla. Muut tietokannat komennolla RUN BLAST.

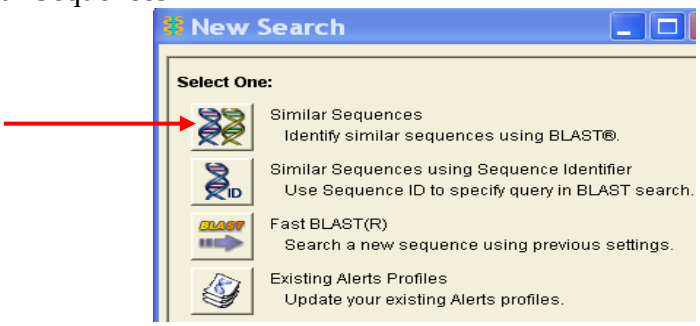
2.1 REGISTRY/(H)CAplus

1) Tee BLAST-haku Registrystä

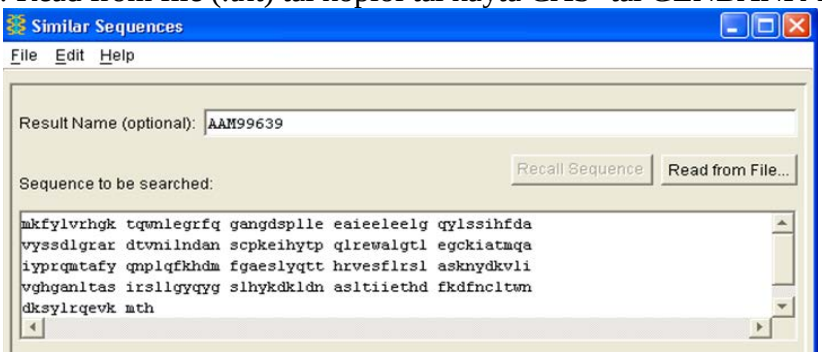
- Valitse Sequence tai New Search/Similar sequences



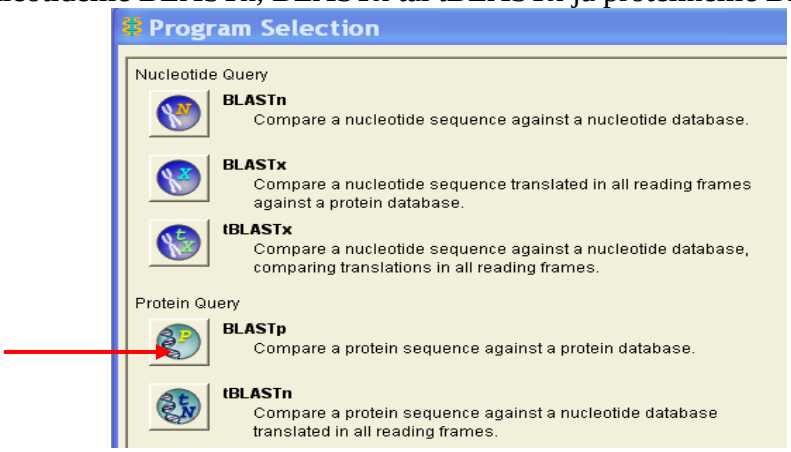
- Valitse Similar Sequences



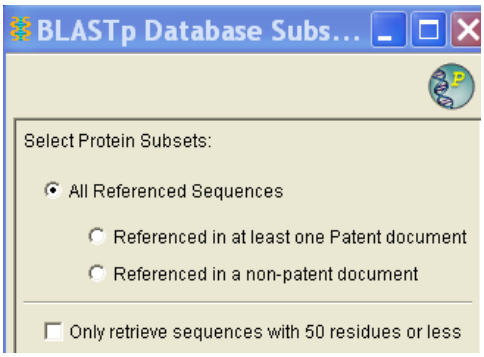
- Hae sekvenssi: Read from file (.txt) tai kopioi tai käytä CAS- tai GENBANK-numeroa.



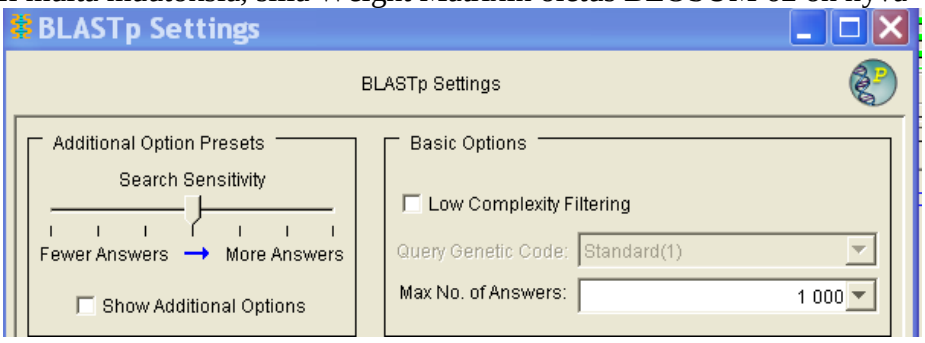
- Valitse nukleotideille BLASTn, BLASTx tai tBLASTx ja proteiineille BLASTp tai tBLASTn



- Valitse All Referenced Sequences

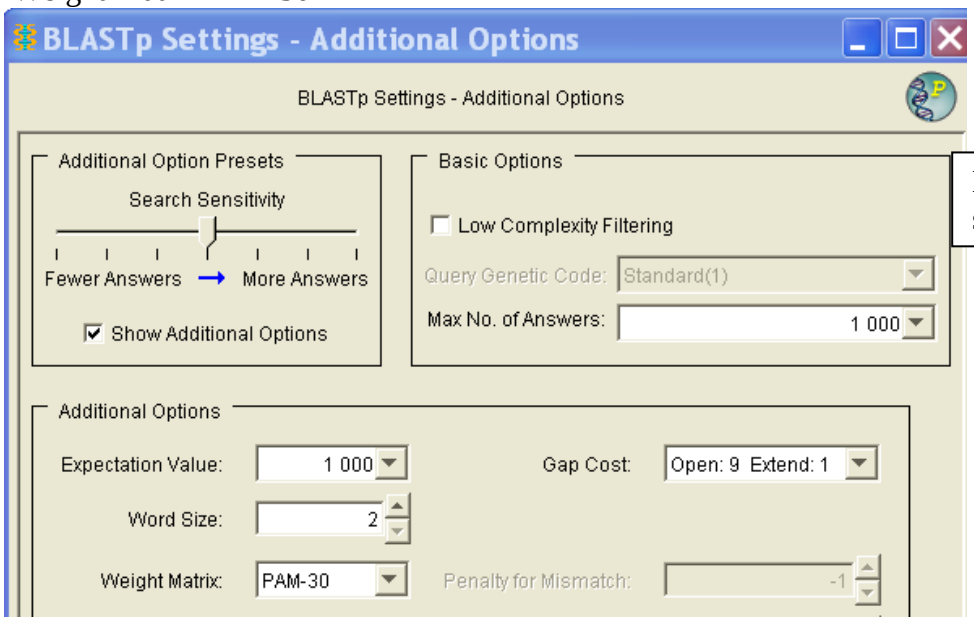


- <u>Pitkille</u> sekvensseille
  - Low Complexity Filteristä ruksi pois patenttihauissa
  - Max Number of Answers Returned => 1000 (ei välilyöntiä 0:n ja 1:n välissä)
  - Ei muita muutoksia, sillä Weight Matrixin oletus BLOSUM 62 on hyvä



Pitkälle sekvenssille Additional Options – oletuksia ei tarvitse muuttaa.

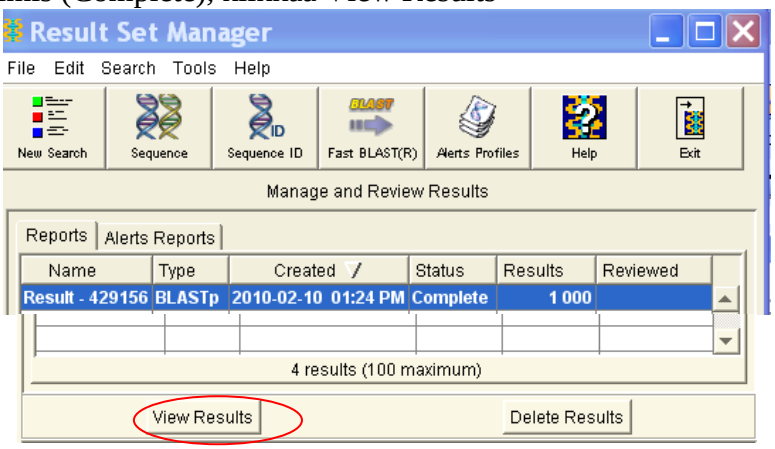
- <u>Lyhyille</u> (alle 35 aminohappoa) sekvensseille
  - Low Complexity Filteristä ruksi pois patenttihauissa
  - Max Number of Answers Returned => 1000 (ei välilyöntiä 0:n ja 1:n välissä)
  - Muuta kohdassa Show Additional Options: Expectation Value 1000, Word size 2, Weight Matrix PAM30



Lyhyelle sekvenssille!

2) Tutki ja valitse vastaukset

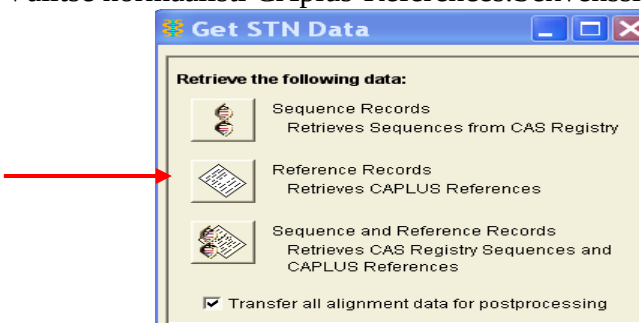
- Kun haku on valmis (Complete), klikkaa View Results



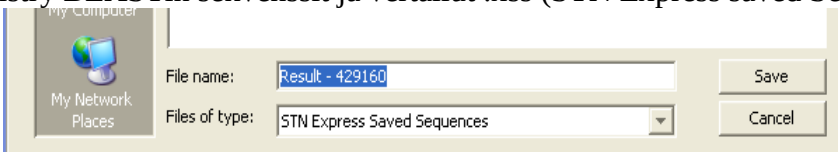
- Vastaukset on järjestetty paremmuusjärjestykseen eri värillisiin osioihin. Klikkaamalla Alignment Score –rivin paksua väripalkkia, saat kaikki siihen kuuluvat sekvenssit. Voit myös ruksata haluamasi sekvenssit. Plus-merkistä näet sekvenssivertailun.

3) Hae STN-viitteet (H)Caplussasta

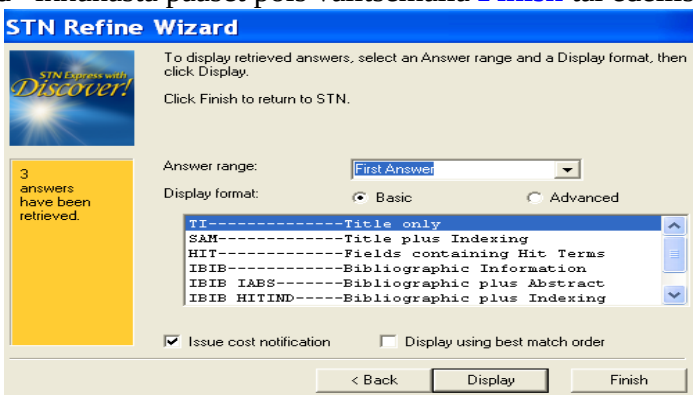
- Get STN Data: Valitse normaalisti CAplus-References.Sekvenssit vain, jos haluat ne Registrystä.



- Tallenna Registry BLASTin sekvenssit ja vertailut .xss (STN Express saved Sequences).



- Ohjelma ottaa automaattisesti yhteyden STN:ään ja hakee ensin sekvenssit Registrystä (ja tulostaa ne, jos valitsit sekvenssin tuloksen) ja hakee sitten niitä vastaavat viitteet CAplussassa.
- **Jos mitään ei tunnu tapahtuvan**, niin **paina end-näppäintä**, niin STN jatkaa.
- STN Refine Wizard –ikkunasta pääset pois valitsemalla **Finish** tai edellisellä sivulla cancel



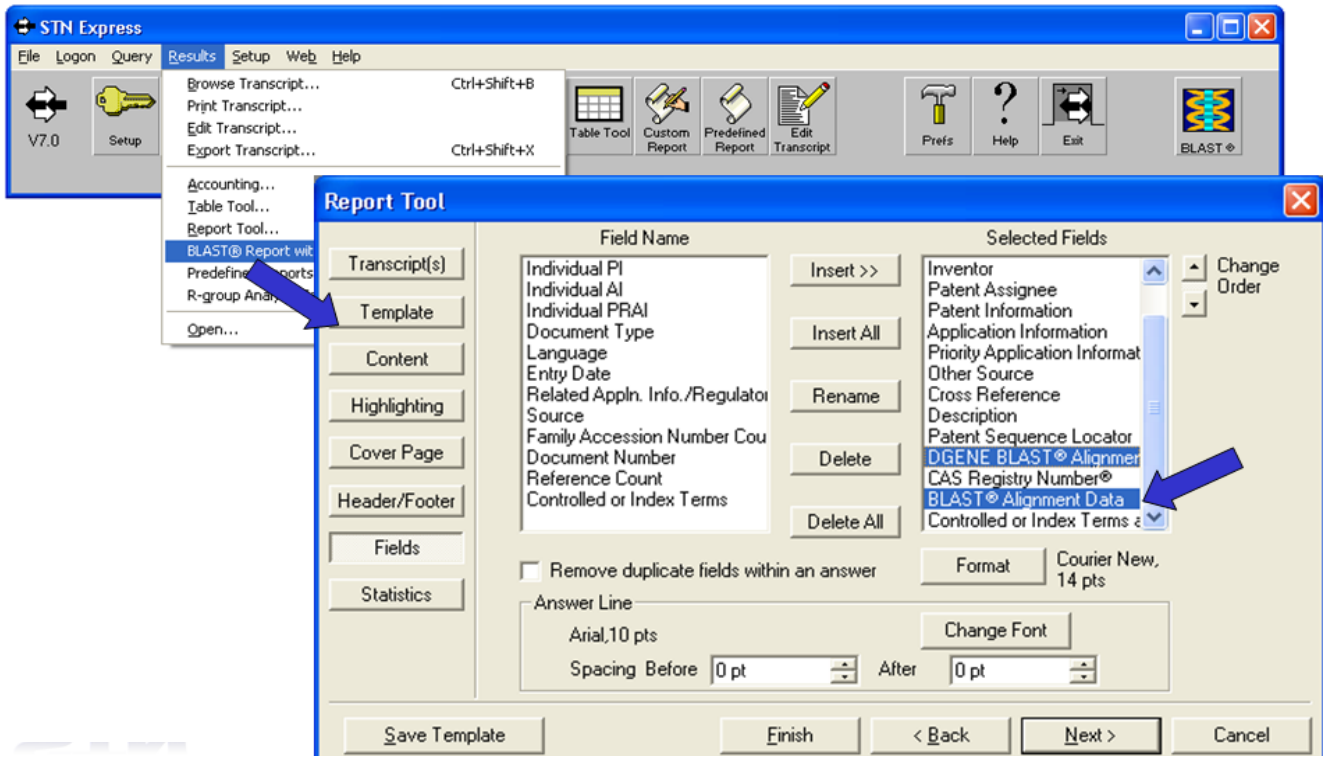
- Sekvenssit ovat lopullisessa joukossa sekaisin. Jos haluat saada CAplus-viitteet erikseen parhaille sekvensseille, niin hae skriptin suorituksen jälkeen Registryn osajoukot S L1, S L2 jne CAplussassa erikseen, sillä ne ovat BLAST-haun mukaisessa paremmuusjärjestyksessä.
- Tee CAplussassa tarvittavat rajaukset, esim. P/DT, roolit tai muu jatkohaku.

4) Tulosta viitteet (H)Caplussasta

- Jotta **sekvenssivertailut** saadaan **mukaan**, niin sekvenssin **CAS-numeron (RN) täytyy olla tulosteessa**, D IBIB AB HITRN tai D IBIB AB HITIND (antaa muutkin osumaindeksitermit). Koko sekvenssi D HITSEQ (ilman vertailua)
- **Jos rajaat sekvenssihakua CAplussassa esim. sanoilla, niin jäljelle jääneet osumasekvenssit saat selville** , kun siirrät osumasekvenssit (HIT RN) REGISTRYyn ja tulostat siellä
  FIL CAPLUS
  SEL HIT RN 1- => E1 THROUGH En ASSIGNED
  FIL REG
  S E1-En
  D SQIDE

5) Tee raportti

- Käytä CAS Registry BLASTin raportin tekoon Results/BLAST Report with Alignment Data, niin saat sekvenssivertailut mukaan.

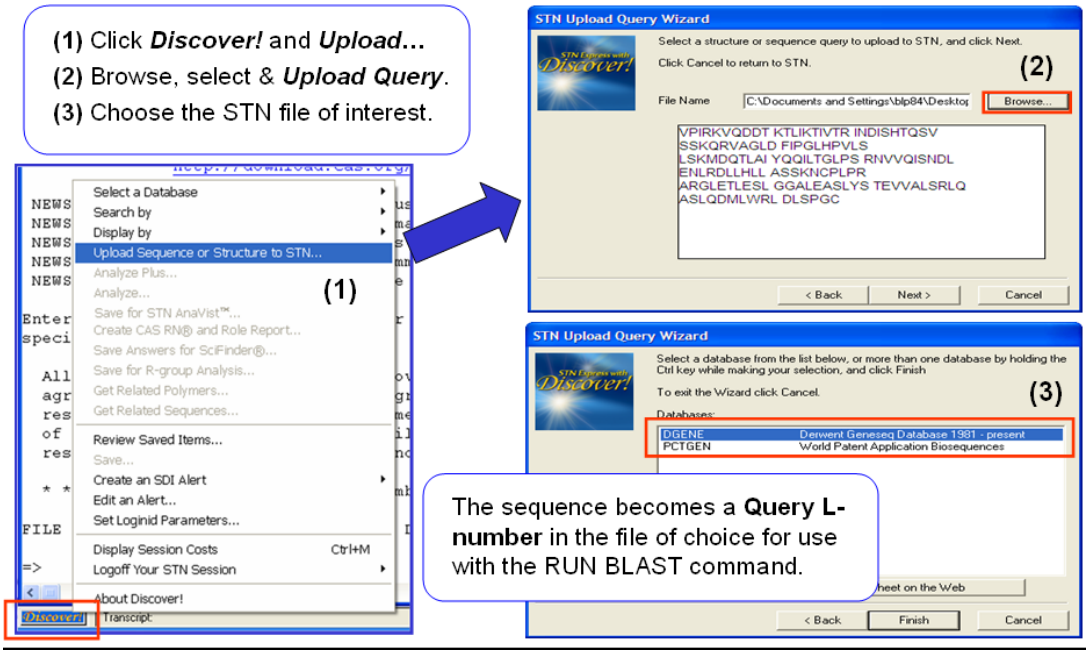



```
L33  ANSWER 2 OF 49  CAPLUS  COPYRIGHT 2004 ACS on STN
AN   2004:176539  CAPLUS
DN   140:176343
TI   Nucleic acid and amino acid sequences relating to Streptococcus . . .
IN   Doucette-stamm, Lynn; Bush, David; Zeng, Qiandong; . . .
PA   Genome Therapeutics Corporation, USA
     PATENT NO.        KIND  DATE          APPLICATION NO.  DATE
     ---------------   ----  --------      ---------------  --------
PI   US 6699703         B1   20040302      US 2000-583110   20000526
PRAI US 1997-51553P     P    19970702
     US 1998-85131P     P    19980512
     US 1998-107433     A2   19980630
IT   660059-83-6 660040-29-9 660049-22-9
     RL: BSU (Biological study, unclassified); DGN (Diagnostic use); . . .
     660059-83-6
     Length = 206     Score = 235     Expect = 4e-61
      Score = 235  Expect = 4e-61
      Identities = 120/204 (58%)  Positives = 141/204 (69%)

      Query:      1 MKFYLVRHGKTQWNLEGRFQGANGDSPLLEEAIEELEELGQYLSSIHFDAVYSSD  55
                    MK Y VRHG+T WN EGRFQGA+GDSPLL E+IE L+ LGQYL  I FD +YSSD
      Subject:    1 MKLYFVRHGRTLWNQEGRFQGASGDSPLLPESIETLKRLGQYLKEIPFDQIYSSD  55
      Query:     56 LGRARDTVNILNDANSCPKEIHYTPQLREWALGTLEGCKIATMQAIYPRQMTAFY 110
                    L RA  +  I+      P  +   P LREW LG LEG KIAT++AIYP+Q+ AF
      Subject:   56 LPRAVKSAEIIQSQLYTPCSLEIVPNLREWQLGKLEGLKIATLEAIYPQQIQAFR 110
      Query:    111 QNPLQFKHDMFGAESLYQTTHRVESFLRSLASKNYDKVLIVGHGANLTASIRSLL 165
                    N  QF   MFGAESLY TT R   F++SL     +++LIVGHGANLTAS+R+LL
      Subject:  111 SNLAQFDTRMFGAESLYSTTQRTIQFIKSLKDSPAERILIVGHGANLTASLRTLL 165
      Query:    166 GYQYGSLHYKDKLDNASLTIIETHDFKDFNCLTWNDKSY 204
                    GY+   L     L NASLTIIETHDF+ F   TWND SY
      Subject:  166 GYKEPLLRKDGGLANASLTIIETHDFETFTLNTWNDTSY 204
```

2.2 DGENE, PCTGEN ja USGENE

Lataa sekvenssi

- Klikkaa STN Expressin Wizardin vasemmassa palkissa olevaa Upload Query => L1
  tai UPLOAD R BLAST tai käytä jo ladattua sekvenssiä. Pitää olla TXT-tiedosto (Notepad)
- Tulosta ladattu sekvenssi D LQUE
- Lyhyen sekvenssin voit kirjoittaa suoraan hakulausekkeeseen.



BLAST-hakulausekkeen määritykset

- Hakukomento RUN BLAST
- Käytä ladatun sekvenssin L-numeroa tai kirjoita sekvenssi (S L2/SQP). Hakukentäksi
  - SQP (polypeptidit/proteiinit)
  - SQN (nukleotidit)
  - TSQN (translated polypeptide) (protein query searched against a nucleotide database translated in all reading frames)
- Pitkille sekvensseille muuta BLASTin oletusasetuksia lisäämällä hakulausekkeeseen
  - -F F (Low Complexity Filter pois päältä). Patenttihauissa hyödyllinen
  - Ei muita muutoksia, sillä Weight Matrixin oletus BLOSUM 62 on hyvä
- Lyhyille (alle 35 aminohappoa) sekvensseille muuta BLASTin oletusasetuksia lisäämällä hakulausekkeeseen
  - -F F (Low Complexity Filter pois päältä). Patenttihauissa hyödyllinen
  - Muuta kohdassa Show Additional Options: Expectation Value 1000, Word size 2, Weight Matrix PAM30
- SQN ja TSQN -hauille voit valita lisäksi
  - Single strand (SIN). Oletus FASTAssa.
  - Complementary strand (COM)
  - Both strands (BOTH). Haku tehdään kumpaankin suuntaan. Oletus BLASTissa.
- BATCH-haku säästää aikaa, koska voit panna kaikki BLAST-haut menemään yhtä aikaa. Tulokset saa komennolla RUN GETBATCH nimi

Esimerkki BLAST-hakulausekkeesta lyhyille proteiinisekvensseille

- RUN BLAST L1/SQP –F F –E 1000 –W 2 –M PAM30 BATCH
- Täydellisessä haussa pitää hakea polypeptidit/proteiinit lisäksi myös TSQN-hakuna ja lisäksi lisätä lausekkeeseen BOTH
  RUN BLAST L1/TSQN BOTH –F F –E 1000 –W 2 –M PAM30 BATCH

Esimerkki BLAST-hakulausekkeista pitkille proteiinisekvensseille

- RUN BLAST L1/SQP –F F BATCH
- RUN BLAST L1/TSQN BOTH –F F BATCH

```
=> RUN BLAST DMGWGSGWRPYYYYGMDV/SQP -F F -E 1000 -W 2 -M PAM30 BATCH

PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS):KOE

BLAST Version 2.2
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM).  See also, Altschul,
Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui
Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein
database search programs." Nucleic Acids Res. 25:3389-3402

BATCH PROCESSING STARTED FOR KOE
```

Valitse, järjestä ja tutki vastauksia

- Ota BATCH-haun tulokset komennolla RUN GETBATCH nimi
- BLAST-haku antaa vastaukseksi kuvan, jossa haun osuvuutta on esitetty viivoilla. ja kysyy sen alla HOW MANY ANSWERS WOULD YOU LIKE TO KEEP?
- Anna prosenttiluvuksi esim. 60% tai vastaa ALL eli valitse kaikki vastaukset säilytettäviksi. Jos tässä vastaat END, niin BLASTia ei veloiteta.
- Järjestä sekvenssit paremmuusjärjestykseen ja toissijaisesti tietuenumeron mukaan järjestykseen SORT SCORE D tai SORT SCORE D AN D
- Tutki sekvenssiosumia maksuttomassa muodossa
  D SCORE TRIAL ALIGN 1-
- **Alignment häviää helposti,** jos teet jatkohakua yhdistämällä joukko teksti- tai aikatermeihin. Jos se häviää, niin **järjestä uudelleen SORT SCORE** D. Jos edelleen poissa, niin yhdistä joukko ANDillä alkuperäiseen joukkoon siten, että alkuperäinen annetaan ensin.
- Sekvenssit voidaan järjestää myös identiteetin mukaan SORT IDENT ja D SCORE D

```
=> run getbatch koe
Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
  Database DGENE AA
    Posted date:  Feb 5, 2010  4:14 PM
  Number of letters in database: 1,146,246,977
  Number of sequences in database:  6,028,597
Lambda     K      H
   0.336    0.272     2.25
Gapped
Lambda     K      H
   0.294    0.110    0.610
```

```
Matrix: PAM30
Gap Penalties: Existence: 9, Extension: 1
Number of Hits to DB: 39,360,653
Number of Sequences: 6028597
Number of extensions: 632767
Number of successful extensions: 13669
Number of sequences better than 1000.0: 5840
Number of HSP's better than 1000.0 without gapping: 5116
Number of HSP's successfully gapped in prelim test: 724
Number of HSP's that attempted gapping in prelim test: 5726
Number of HSP's gapped (non-prelim): 8576
length of query: 18
length of database: 1,146,246,977
effective HSP length: 9
effective length of query: 9
effective length of database: 1,091,989,604
effective search space: 9827906436
effective search space used: 9827906436
T: 16
A: 40
X1: 15 ( 7.3 bits)
X2: 35 (14.8 bits)
X3: 58 (24.6 bits)
S1: 41 (21.7 bits)
S2: 48 (23.5 bits)
 5812 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1000.0
      QUERY SELF SCORE VALUE IS    70
      BEST ANSWER SCORE VALUE IS   70

Similarity
Score
       70 |
          |
          |
          |
          |
          |
          |
          |
          |
          ||
       35 ||||||||||||
          ||||||||||||||
          ||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
Answer Count    1170      2340      3510      4680      5850

ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
 OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
 (BEST ANSWER PERCENTAGE OF SELF SCORE IS 100%)
ENTER (ALL) OR ? :60
L1    RUN STATEMENT CREATED
L1          60 DMGWGSGWRPYYYYGMDV/SQP.-F F -E 1000 -W 2 -M PAM30
Answer set arranged by accession number; to sort by descending
similarity score, enter at an arrow prompt (=>) "sor score d".
```

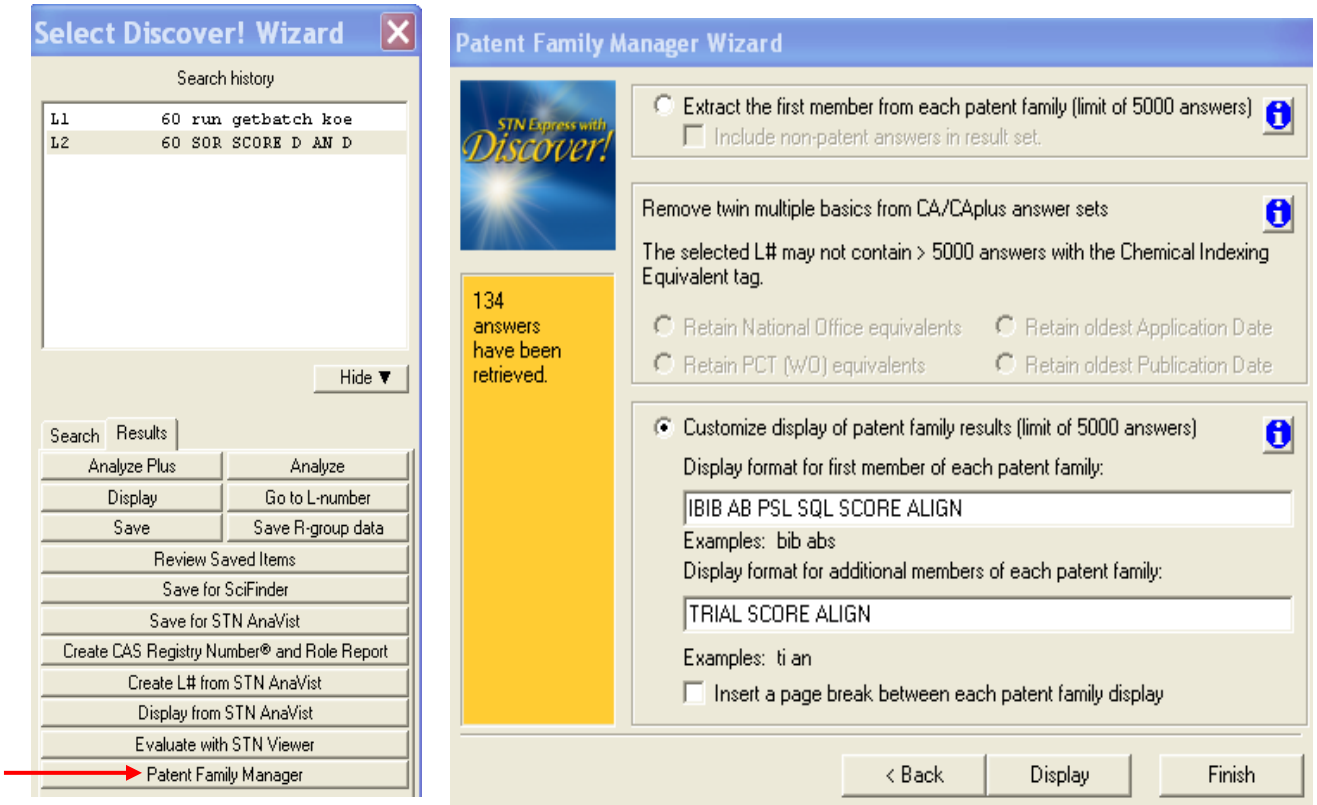
```
=> SOR SCORE D AN D
PROCESSING COMPLETED FOR L1
L2            60 SOR L1 SCORE D AN D
```

Tulostus

- Yhdistä DGENE, PCTGEN ja USGEN-joukot ennen tulostusta DUP IDEllä
- Järjestä viitteet ennen lopullista tulostusta järjestykseen ensin parhaiden sekvenssien mukaan (SORT SCORE D) sitten patenttiperheittäin (FSORT).

- Tulostustapa 1. (automaattinen)
  - Käytä Patent Family Managerin tulostusosaa “Customize display of patent family results”
    Display format for first member of each family => IBIB AB PSL SQL SCORE ALIGN
    Display format for additional members of each family => TRIAL SCORE ALIGN
  - Patent Family Manager järjestää vastaukset ensin automaattisesti perheittäin FSORTilla ja tulostaa sitten jokaisesta perheestä ensimmäiselle sekvenssille bibliografiset tiedot ja tiivistelmät ja muille sekvensseille vain maksuttomat tiedot, mm. sekvenssivertailut
  - Saat vastaukset myös erilliseen ikkunaan. Poistu Patent Family Managerista Finishillä

- Tulostustapa 2. (manuaalinen)
  - Julkaisutiedot kullekin perheelle vain kerran ja lisäksi osuvin sekvenssi ja vertailu
    D PFAM=1- IBIB AB PSL SQL SCORE ALIGN tai DPAM=1- IALL SCORE ALIGN
  - Loput haussa löytyneet hakemuksen sisältämät sekvenssit ilman bibliografisia tietoja.
    D PFAM=1- SCORE TRIAL ALIGN 2-

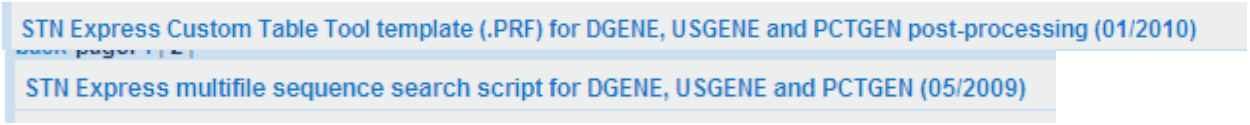


```
=> DIS L9 PFAM=1 1 IBIB,AB,PSL,SQL,SCORE,ALIGN

L9    ANSWER 1 OF 60  DGENE  COPYRIGHT 2010 THOMSON REUTERS on STN FAMILY    1
ACCESSION NUMBER: ATS30301  protein       DGENE  Full-text
TITLE:            New polypeptide comprises a binding domain capable of binding
                  to an  epitope of human and non-chimpanzee primate CD3
                  epsilon chain, useful for preventing, treating, or
                  ameliorating a proliferative, tumor, or  immunological disorder.
INVENTOR:         Ebert E; Meier P; Sriskandarajah M; Burghart E; Wissing S;
                  Klinger M; Bluemel C; Raum T; Rau D; Mangold S; Kvesic M;
PATENT ASSIGNEE:  (MICR-N)MICROMET AG.
PATENT INFO:      WO 2008119565   A2 20081009               397
APPLICATION INFO: WO 2008-EP2662       20080403
PRIORITY INFO:    EP 2007-6988         20070403
PAT. SEQ. LOC:    Claim 15; SEQ ID NO 592
DOCUMENT TYPE:    Patent
LANGUAGE:         English
OTHER SOURCE:     2008-N22814 [77]
CROSS REFERENCES: N-PSDB: ATS30302
DESCRIPTION:      Anti-CD3/anti-EpCAM cross-species single chain Ab protein,
                  SEQ: 592.
AB     The present invention relates to a novel polypeptide comprising a first
       human binding domain capable of binding to an epitope of human and non-
        ..... jne .....
       monkey CD3 epsilon as well as a domain with a binding specficity cross-
       species specific for human and cynomolgus EpCAM of the present invention.
PSL   Claim 15; SEQ ID NO 592
SQL   504
SCORE 70      100% of query self score 70
BLASTALIGN
   Query  = 18 letters
   Length = 504
   Score  = 69.8 bits (157), Expect = 5e-18
   Identities = 18/18 (100%), Positives = 18/18 (100%)
  Query: 1   DMGWGSGWRPYYYYGMDV 18
              DMGWGSGWRPYYYYGMDV
  Sbjct: 99  DMGWGSGWRPYYYYGMDV 116


=> DIS L9 PFAM=1 2-TOT TRIAL,SCORE,ALIGN

L9    ANSWER 2 OF 60  DGENE  COPYRIGHT 2010 THOMSON REUTERS on STN FAMILY    1
AN    ATS30299  protein       DGENE
TI    New polypeptide comprises a binding domain capable of binding to an
      epitope of human and non-chimpanzee primate CD3 epsilon chain, useful for
      preventing, treating, or ameliorating a proliferative, tumor, or
      immunological disorder.
DESC  Anti-CD3/anti-EpCAM cross-species single chain Ab protein, SEQ: 590.
KW    single chain antibody; CD3E; T-cell CD3 glycoprotein epsilon chain;
      TACSTD1; EpCAM; protein production; protein therapy; therapeutic;
      prophylactic to disease; protein detection; immune disorder;
      immunomodulator; cancer; cytostatic; hyperproliferation; Fusion protein.
SQL   504
SCORE 70      100% of query self score 70
BLASTALIGN
   Query  = 18 letters
   Length = 504
   Score  = 69.8 bits (157), Expect = 5e-18
   Identities = 18/18 (100%), Positives = 18/18 (100%)
  Query: 1   DMGWGSGWRPYYYYGMDV 18
              DMGWGSGWRPYYYYGMDV
```

```
  Sbjct: 99  DMGWGSGWRPYYYYGMDV 116
……..…….. jne ……..……..
```

Raportin teko

- STN-sivuilla on kaksi valmista templaattia. Kopioi ne STN Expressin Trnscrpt-kansioon. Toinen niistä tekee Word-raportin ja toinen Excel-taulukon. Voit muokata templaatteja, jos haluat



2.3 Koko BLAST-hakuproseduuri

- Tee ensin CAS Registry BLAST.
- Pistä sitten BATCH-haut menemään peräjälkeen tietokannoissa USGENE, DGENE ja PCTGEN
- Käy tulokset läpi jokaisessa tietokannassa erikseen
- Yhdistä eri tietokantojen lopulliset hakujoukot DUP REM tai DUP IDE –komennoilla
- Tulosta viitteet Patent Family Managerin tulostusosan avulla
  Display format for first member of each family => BIB PSL SQL SCORE ALIGN
  Display format for additional members of each family => TRIAL SCORE ALIGN
- Tee raportti esimerkiksi käyttäen hyväksi STN-sivulla http://www.stn-international.de/biosequence_searching.html olevia valmiita templaatteja
  - STN Express Custom Report Tool template (.PRF) for sequence search post-processing
  - STN Express Custom Table Tool template (.PRF) for DGENE, USGENE and PCTGEN post-processing

3. GETSIM (=FASTA)

- GETSIM on käytettävissä vain DGENE-, PCTGEN- ja USGENE-tietokannoissa.
- FASTA on täydellisempi, mutta hitaampi hakutapa kuin BLAST, mutta jompikumpi riittää.
- FASTA katsoo enemmän kokonaisuutta, BLAST vertailee lyhyempiä pätkiä.
- GETSIM-parametrejä ei voida muuttaa.

- Hakukomento on RUN GETSIM.

- Tee aina eräajona lisäämällä hakulauseeseen termi BATCH, sillä GETSIM-haku kestää yleensä melko kauan vähintään 0,5 h.
    - RUN GETBATCH tarkistaa, onko haku päässyt loppuun asti (Completed)

- Lisää SQN- ja TSQN-hakuihin BOTH (Both Strands)
    - sillä GETSIMissä on oletuksena SIN (Single Srand).

- RUN GETSIM L1/SQP BATCH
- RUN GETSIM L1/SQN BOTH BATCH
- RUN GETSIM L1/TSQN BOTH BATCH

- Koko muu proseduuri on samanlainen kuin BLAST-haussa.

4. Sequence Code Match (SCM) = GETSEQ

- Sopii primeereille ja muille lyhyille sekvenssipätkille.
- Käytä Notepadiä tai muuta plain text editoria
- Hakukenttävaihtoehdot

	Polypeptidit/proteiinit	Nukleotidit
Exact	/SQEP	/SQEN
Exact family	/SQEFP	
Subsequence	/SQSP	/SQSN
Subsequence Family	/SQSFP	

- Subsequence-hauissa voit käyttää erityisiä symboleja kuvaamaan sallittuja ”motifs, patterns and gaps”. Katso esim. DGENE Workshop material, Liite 6
  http://www.stn-international.de/training_center/bioseq/dgene_wm.pdf
- Sekvenssivertailu näytetään viivoina a.o. kohdassa sekvenssin alla. Osumasekvenssit näytetään aina kokonaan. Ei ole maksutonta muotoa vertailun näyttämiseksi.

4.1 REGISTRY/(H)CAplus (S …./SQ..)

Tee sekvenssihaku Registryssä

- FIL REG
  S ……/SQ.. => L1
- Voit rajoittaa hakua esim, sekvenssin pituudella SQL-kentässä S L1 AND SQL>=40
- Voit tulostaa sekvenssin tiedot ja osumasekvenssin. Sekvenssivertailu mukaan automaattisesti.
  D SEQ (pelkkä sekvenssi) tai D SQID (sekvenssi + identifiointitiedot)

Hae ja tulosta julkaisut (H)CAplussasta

- FIL HCAPLUS tai FIL CAPLUS
  S L1 => L2
- Tulosta
  D BIB AB / D IALL /
- Jos haluat myös osumasekvenssin, niin lisää komentoon HITSEQ. Ei anna sekvenssivertailua. D BIB AB HITSEQ
- Voit myös jatkaa hakua, esim.
  S L2 AND P/DT (vain patentit)
  S L2 AND ANTIBODY?

- **Jos haluat haun jälkeen tietää, mitkä osumasekvenssit jäivät rajoituksen jälkeen jäljelle**
  Siirrä osumasekvenssit (HIT RN) REGISTRYyn ja tulosta siellä.
  SEL HIT RN 1- => E1 THROUGH En ASSIGNED
  FIL REG
  S E1-En

4.2 DGENE, PCTGEN ja USGENE (RUN GETSEQ)

Tee sekvenssihaku

- FIL DGENE tai FIL PCTGEN tai FIL USGENE
  RUN GETSEQ …../SQ.. => L1

Mahdollisia rajoituksia sekvenssihaun jälkeen

- Sekvenssin pituus numeerinen SQL-kenttä, esim. S L1 AND SQL>=8
- Annotation NTE-kenttä esim. kokoproteiinin laaja luokitus (cyclic ym.) tai kemiallinen modifikaatio (metal complex, bridge ym.) S L1 AND CYCLIC/NTE

Voit jatkaa hakua muilla termeillä. Jotta sekvenssivertailu säilyy tulosteissa,

- yhdistä nämä uusi joukko ANDillä alkuperäisen GETSEQ-haussa saatuun joukkoon siten, että GETSEQillä saatu alkuperäinen joukko annetaan ensin.

Tulostus

Kuten BLAST-haun jälkeen. Katso edellisistä kohdista

Multifile-haut

Kuten BLAST-haussa. Katso edellisistä kohdista