



New sequence search interaction to refine answer sets in DGENE, USGENE and PCTGEN

A new sequence search interaction to refine answer sets in DGENE, USGENE and PCTGEN is available since May 25, 2008. This new interaction allows you to define a specific subset of the original complete answer set of a homology sequence search. This feature now enables you to immediately focus results to the most relevant answers and reduce review time. It will also explicitly help to refine large answer sets.

To meet customer requirements, the homology sequence searching features, RUN BLAST and RUN GETSIM in DGENE, USGENE and PCTGEN, have been enhanced with new capabilities for processing sequence search results. For the first time, customers will be able to process and refine homology sequence search answer sets in these databases by selecting a minimum percentage of the query self score value as a cut-off point. This new interaction is available for all BLAST and GETSIM online search options as well as for BLAST and GETSIM BATCH and ALERT searches in all three databases.

It has already been possible to define the number of answers from a sequence search to be kept for further processing. Now, also a minimum percentage of the query self score value may be applied as a cut-off resulting in an answer set that contains only those answers with a higher percentage. For

this purpose, the sequence query self score value will be calculated for each search. Two values are displayed within the graphic representation of the sequence search result, the query self score value defining the maximum score value possible when the query is aligned to itself and the score value of the best answer. You have three possibilities to select a result answer set.

You can keep:

- the complete answer set (ALL)
- a subset of the complete answer set by specifying a smaller number of just the top scoring answers, or
- you can specify the minimum percentage of the self score value, to keep a subset of the complete answer set, where the answers have a better score than your chosen minimum percentage of the query self score value.

Also, the SCORE field has generally been enhanced to show the percentage of the query self score value in addition to the actual bit score value. Until now, this had been the case only after RUN GETSIM sequence searches and is now also introduced after RUN BLAST sequence searches. See an example of the new homology sequence search interaction below.

Example of the new homology sequence search feature

See below a search example with the ErbB2 protein, or Herceptin2 receptor, showing the use of the new query self score feature in DGENE and USGENE. The search strategy reflects a protein (/SQP) sequence homology search as a starting point for further refinements and processing. The new query self score feature is used to focus results to highly similar hits and the STN Patent Family Sort (FSORT) command is employed to reduce the selected sequence hits to the relevant underlying patent publications.

=> FIL DGENE

FILE 'DGENE' ENTERED AT 17:21:24 ON 27 MAY 2008
COPYRIGHT (C) 2008 THOMSON REUTERS

=> UPLOAD

IS THIS DATA A QUERY, OR FOR A RUN PACKAGE? Q/R/(END):R
ENTER NAME OF RUN PACKAGE, END OR (?):BLAST
START LOCAL KERMIT TRANSMIT PROCESS

UPLOAD SUCCESSFULLY COMPLETED

L1 GENERATED

=> D L1 LQUE

```
L1 ANSWER 1 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
LQUE melaalcrwg lllal1ppga astqvctgtd mklrlpaspe thldmlrhly
      qgcqvvgqnl eltylptnas lsflqdiqev qyvliahnq vrqvp1qrlr
      ivrgtqlfed nyalavldng dplnnttpvt gaspgglrel qlrslteilk
      ggvliqrnpq lcyqdt1lwk difhknnqla ltlidtnrsr achpcspmk
      gsrcwgesse dcqsltrtvc aggcacrckgp lptdccheqc aagctgpkhs
      dclaclhfnh sgi celhcpa lvtyntdtfe smpnpegryt fgascvtacp
      ynylstdvgs ct1vcplhnq evtaedgtqr cekcskpcar vcyglgmehl
      revravtsan iqefagckki fgslaf1pes fdgdpasnta plqpeqlqvf
      etleeitgyl yisawpds1p d1svfqnlqv irgrilhnqa ys1tlqglgi
      swlglrslre lgs1galihh nthlcfvhtv pwdqlfrnph qallhtanrp
      edecvgegla chqlcarghc wpggptqcvn csqflrgqec veecrvlqgl
      preyvnrhc lpchpecqpq ngsvtcf1gpe adqcvacahy kdppfcvarc
      psgvkpd1sy mpiwkfpdee gacqpcpinc thscvd1ddk gcpaeqrasp
      ltsivsavvg illvvvl1gvv fgil1krrqq kirkytmr1l lqetel1vepl
      tpsgampnqa qmrilketel rk1kv1lgsga fgtvyk1giwi pdgenvkipv
      aikvlrents pkankeilde ayvmagv1gsp yvsr1l1gic1 tstvqlvtq1
      mpygc1ldhv renrgr1gsq d1lnwcmqia kgmsy1edvr lvhrdlaarn
      vlvkspnhvk itdfglar1l dideteyhad g1kvpi1k1ma lesilrrrft
      hqsdvwsygv tvwelmtf1ga kpydgipare ipd1lekger lpqppictid
      vymimvkcwm idsecrprfr elvsefsrma rdpqr1fv1iq nedl1gaspl
      dstfyr1s1le dddmgdlvda ee1ylv1qqgf fcpdpap1gag gmvh1hr1rss
      strsgggd1t lg1epse1e1a prs1lap1seg agsdv1fd1dl gmgaak1g1qs
      lpthdps1lq r1sedptv1p psetdgyv1p ltcspqpey1v nqpdvrp1qpp
      spregplpaa rpagat1era kt1lspgkngv vkdvf1f1gga venpey1ltpq
      ggaapqpphp pafspaf1dnl yywdqdp1per gappst1f1kgt ptaenpey1l
      ldvpv
```

The UPLOAD command generates an L-number representing the query sequence. This L-number may be used in sequence searches in DGENE, USGENE and PCTGEN.

=> FSORTENTER (L3) or L#:.
.....

```

L4          139 FSO L3

                26 Multi-record Families           Answers 1-74
                  Family 1                       Answers 1-2
                  Family 2                       Answers 3-4
                  Family 3                       Answers 5-7
                  Family 4                       Answers 8-10
                  Family 5                       Answers 11-22
                  Family 6                       Answers 23-24
                  Family 7                       Answers 25-26
                  Family 8                       Answers 27-28
                  Family 9                       Answers 29-30
                  Family 10                      Answers 31-32
                  Family 11                      Answers 33-36
                  Family 12
                  Family 13
                  Family 14
                  Family 15
                  Family 16
                  Family 17
                  Family 18
                  Family 19
                  Family 20
                  Family 21
                  Family 22
                  Family 23                       Answers 65-68
                  Family 24                       Answers 69-70
                  Family 25                       Answers 71-72
                  Family 26                       Answers 73-74
                65 Individual Records             Answers 75-139
                 0 Non-patent Records

```

The FSORT command groups all answers into patent families. In this case, all sequence records belonging to one publication will be grouped together.

139 hit sequence records are grouped into 26 multi-record families (comprising several sequence records per publication) and 65 individual records (with only one sequence record per publication). Each family reflects one underlying patent publication.

The subgroup of the multi-record families as well as the subgroup of the individual records retains the sorting according to score. In addition, within each family the first answer will be the one with the highest score value.

=> DISPLAY PFAMENTER (L4) OR L#:.
ENTER PATENT FAMILY NUMBER OR RANGE (1):1-
ENTER ANSWER NUMBER OR RANGE (1):1
ENTER DISPLAY FORMAT (BIB):TRIAL SCORE ALIGN

```

L4 ANSWER 1 OF 139 DGENE COPYRIGHT 2008 THOMSON
AN ARL13164 protein DGENE
TI Predicting, diagnosing, or prognosing malignant
cancer, by detecting markers that are genes and
located in a chromosomal region that is altered
DESC Human ERBB2 protein, SEQ ID 36.
KW diagnosis; prognosis; therapeutic; diagnostic;
marker; neoplasia; cytostatic; breast tumor; ovary tumor; gastrinoma;
gastrointestinal-gen.; colon tumor; esophagus tumor; bladder tumor;
non-small-cell lung cancer; ERBB2; BOND_PC; HER2 receptor; GO166;
GO4714; GO4715; GO4716; GO4872; GO5006; GO5524; GO5576; GO5737; GO5886;
GO6468; GO7169; GO7422; GO7507; GO8283; GO16020; GO16021; GO16324;
GO16740; GO30879; GO42552; GO42802; GO43125; GO43406; GO45765; GO46982;
GO48015; GO50679.
SQL 1255
SCORE 2628 100% of query self score 2628
BLASTALIGN
Query = 1255 letters
Length = 1255
Score = 2628 bits (6812), Expect = 0.0
Identities = 1255/1255 (100%), Positives = 1255/1255 (100%)
Query: 1 MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDMRLRHLYQGCQVVGQNL
MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDMRLRHLYQGCQVVGQNL
Sbjct: 1 MELAALCRWGLLLALLPPGAASTQVCTGTDMKLRLPASPETHLDMRLRHLYQGCQVVGQNL
Query: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLQFEDNYALAVLDNG
ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLQFEDNYALAVLDNG
Sbjct: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLQFEDNYALAVLDNG
.
.
Query: 1141 NQPDVVRPQPPSPREGPLPAARPAGATLERAKTSLPGKNGVVKDVFVAFGGAVENPEYLTPO
NQPDVVRPQPPSPREGPLPAARPAGATLERAKTSLPGKNGVVKDVFVAFGGAVENPEYLTPO
Sbjct: 1141 NQPDVVRPQPPSPREGPLPAARPAGATLERAKTSLPGKNGVVKDVFVAFGGAVENPEYLTPO
Query: 1201 GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVPV
GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVPV
Sbjct: 1201 GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVPV

```

Displaying the first (=best) hit from each family in the DISPLAY FAM mode allows for a precise review of relevant answers. In DGENE, the title (TI), description (DESC) and keyword (KW) fields in the free-of-charge format TRIAL help evaluating the relevance of answers.

```

L4 ANSWER 3 OF 139 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN FAMILY 2
AN AQY87162 protein DGENE
TI Characterizing state of neoplastic disease, e.g. breast cancer, by
determining pattern of expression levels of marker genes differentially
expressed in the neoplastic tissue to characterize state of the
neoplastic disease.
DESC Human erythroblastic leukemia viral oncogene homolog 2, SEQ ID 251.
KW gene expression; diagnostic; diagnosis; screening; prognosis; neoplasm;
breast tumor; cytostatic; ERBB2; erythroblastic leukemia viral oncogene
homolog 2; BOND_PC; HER2 receptor; GO166; GO4714; GO4715; GO4716;
GO4872; GO5006; GO5524; GO5576; GO5737; GO5886; GO6468; GO7169; GO7422;
GO7507; GO8283; GO16020; GO16021; GO16324; GO16740; GO30879; GO42552;
GO42802; GO43125; GO43406; GO45765; GO46982; GO48015; GO50679.
SQL 1255
SCORE 2628 100% of query self score 2628
BLASTALIGN
Query = 1255 letters
Length = 1255
Score = 2628 bits (6812), Expect = 0.0
Identities = 1255/1255 (100%), Positives = 1255/1255 (100%)
Query: 1 MELAALCRWGLLLALPPGAASTQVCTGTMKLRASPETHLDMLRHLYQGCQVVGNNL
MELAALCRWGLLLALPPGAASTQVCTGTMKLRASPETHLDMLRHLYQGCQVVGNNL
Sbjct: 1 MELAALCRWGLLLALPPGAASTQVCTGTMKLRASPETHLDMLRHLYQGCQVVGNNL
Query: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNG
ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNG
Sbjct: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNG
.
.
.
Query: 1141 NQPDVVRQPPSPREGPLPAARFAGATLERAKTSLSPGKNGVVKDVFAFGGAVENPEYLTPQ
NQPDVVRQPPSPREGPLPAARFAGATLERAKTSLSPGKNGVVKDVFAFGGAVENPEYLTPQ
Sbjct: 1141 NQPDVVRQPPSPREGPLPAARFAGATLERAKTSLSPGKNGVVKDVFAFGGAVENPEYLTPQ
Query: 1201 GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVP
GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVP
Sbjct: 1201 GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVP
.
.
.
L4 ANSWER 139 OF 139 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
AN AAG62860 Protein DGENE
TI Novel isolated murine homolog of human Her-2/neu useful for inhibiting
development of cancer, preferably breast cancer in a patient -
DESC Amino acid sequence of a murine neu polypeptide from C57B16 mice.
KW Neu polypeptide; Her-2; cancer; breast cancer; T cell expansion;
vaccine.
SQL 1256
SCORE 2305 87% of query self score 2628
BLASTALIGN
Query = 1255 letters
Length = 1256
Score = 2305 bits (5972), Expect = 0.0
Identities = 1099/1256 (87%), Positives = 1155/1256 (91%), Gaps = 1/125
Query: 1 MELAALCRWGLLLALPPGAASTQVCTGTMKLRASPETHLDMLRHLYQGCQVVGNNL
MELAA CRWG LLALL PGAA TVQVCTGTMKLRASPETHLDMLRHLYQGCQVVGNNL
Sbjct: 1 MELAAWCRWGFLLALLSPGAAGTQVCTGTMKLRASPETHLDMLRHLYQGCQVVGNNL
Query: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTQLFEDNYALAVLDNG
ELTYLP NASLSFLQDIQEVQGY+LIAHN+V+ VPLQRLRIVRGTQLFED YALAVLDN
Sbjct: 61 ELTYLPANASLSFLQDIQEVQGYMLIAHNRVHKHVPLQRLRIVRGTQLFEDKYALAVLDNR
.
.
.
Query: 1080 GAGSDVFDGDLGMGAAGLQSLPTHDPSPQRYSPTVPLPSETDGYVAPLTCSPQPEY
GAGSDVFDGDL +G KGLQSL HD SPLQRYSPTVPLPSETDGYVAPL CSPQPEY
Sbjct: 1081 GAGSDVFDGDLAVGVTKGLQSLSPHDLSPQRYSPTVPLPSETDGYVAPLACSPQPEY
Query: 1140 VNQP+VRPQ P EGP P RPAGATLER KTLSPGKNGVVKDVFAFGGAVENPEYL P
VNQPEVRPQSPLTPEGPPPIRPAATLERPKTSLSPGKNGVVKDVFAFGGAVENPEYLAP
Sbjct: 1141 VNQPEVRPQSPLTPEGPPPIRPAATLERPKTSLSPGKNGVVKDVFAFGGAVENPEYLAP
Query: 1200 QGGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVVP
+ G A QPHP PAFSPAFDNLYYWDQ+ E+G PPSTF+GTPTAENPEYLGLDVVP
Sbjct: 1201 RAGTASQPHPPPAFSPAFDNLYYWDQNSSEQPPSTFEGTPTAENPEYLGLDVVP

```

=> **FIL USGENE**

FILE 'USGENE' ENTERED AT 17:31:22 ON 27 MAY 2008
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

=> **RUN BLAST L1/SQP -F F**

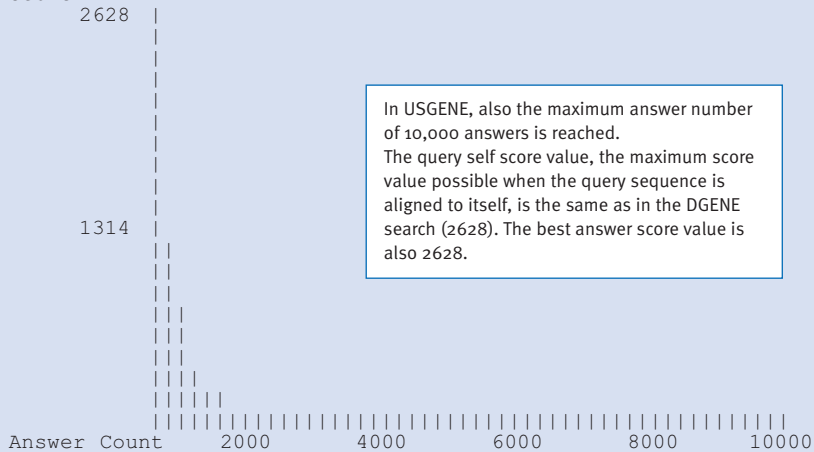
BLAST Version 2.2

The BLAST software is used herein with permission of
National Center for Biotechnology Information (NCBI)
the National Library of Medicine (NLM).....

10000 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

QUERY SELF SCORE VALUE IS 2628
BEST ANSWER SCORE VALUE IS 2628

Similarity
Score



In USGENE, also the maximum answer number of 10,000 answers is reached. The query self score value, the maximum score value possible when the query sequence is aligned to itself, is the same as in the DGENE search (2628). The best answer score value is also 2628.

ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
(BEST ANSWER PERCENTAGE IS 100%)

ENTER (ALL) OR ? :85%

Maximum answer limit of 10,000 reached

L5 RUN STATEMENT CREATED

L5 127 MELAALCRWGLLLALLPPGAASTQVCTGTDMLRLE
QGCQVVQGNLELTYLPTNASLSFLQDIQEVQGYVL
.....
GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLG
LDVPV/SQP.-F F

Selecting 85% of the query self score as the cut-off point leads to a subset of 127 answers in USGENE with all answers having a higher score value than 85% of the query self score.

Answer set arranged by accession number; to sort by descending similarity score, enter at an arrow prompt (=>) "sor score d".

```

=> SOR SCORE D
PROCESSING COMPLETED FOR L5
L6          127 SOR L5 SCORE D
    
```

The score values of the 127 hit sequences range from 100% to 85% of the query self score value, just as specified after the graphic representation of the result set.

```

=> D 1 50 100 127 SCORE
    
```

```

L6      ANSWER 1 OF 127  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 2628          100% of query self score 2628

L6      ANSWER 50 OF 127  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 2626          99% of query self score 2628

L6      ANSWER 100 OF 127 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 2578          98% of query self score 2628

L6      ANSWER 127 OF 127 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 2242          85% of query self score 2628
    
```

```

=> FSORT
ENTER (L6) or L#:..
.....
L7          127 FSO L6
    
```

```

          21 Multi-record Families      Answers 1-62
            Family 1                   Answers 1-2
            Family 2                   Answers 3-4
            Family 3                   Answers 5-6
            Family 4
            Family 5
            Family 6
            Family 7
            Family 8
            Family 9
            Family 10
            Family 11                  Answers 26-40
            Family 12                  Answers 41-42
            Family 13                  Answers 43-44
            Family 14                  Answers 45-46
            Family 15                  Answers 47-48
            Family 16                  Answers 49-50
            Family 17                  Answers 51-52
            Family 18                  Answers 53-54
            Family 19                  Answers 55-56
            Family 20                  Answers 57-58
            Family 21                  Answers 59-62
          65 Individual Records        Answers 63-127
           0 Non-patent Records
    
```

The 127 hit sequence records are grouped into 21 multi-record families (comprising several sequence records per publication) and 65 individual records (with only one sequence record per publication).

```

=> DISPLAY PFAM
ENTER (L7) OR L#:..
ENTER PATENT FAMILY NUMBER OR RANGE (1):1-
ENTER ANSWER NUMBER OR RANGE (1):1
ENTER DISPLAY FORMAT (BIB):TRIAL SCORE ALIGN
    
```

```

L7      ANSWER 1 OF 127  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY1
TI      Methods and materials for characterizing and modulating interaction
        between heregulin and HER3 (Patent)
MTY     Protein
SQL     1255
SCORE 2628          100% of query self score 2628
BLASTALIGN
  Query = 1255 letters
  Length = 1255
  Score = 2628 bits (6812), Expect = 0.0
  Identities = 1255/1255 (100%), Positives = 1255/1255 (100%)
  Query: 1      MELAALCRWGLLLALPPGAASTQVCTGTDMLRRLPASPETHLDMLRHLYQGCQVVQGNL
                MELAALCRWGLLLALPPGAASTQVCTGTDMLRRLPASPETHLDMLRHLYQGCQVVQGNL
  Sbjct: 1      MELAALCRWGLLLALPPGAASTQVCTGTDMLRRLPASPETHLDMLRHLYQGCQVVQGNL
  Query: 61     ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIRVGTQLFEDNYALAVLDNG
                ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIRVGTQLFEDNYALAVLDNG
  Sbjct: 61     ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIRVGTQLFEDNYALAVLDNG
  .
  .
  Query: 1141   NQPDVVRQPPSPREGPLPAARPAGATLERAKTILSPGKNGVVKDVFAFGGAVENPEYLTPO
                NQPDVVRQPPSPREGPLPAARPAGATLERAKTILSPGKNGVVKDVFAFGGAVENPEYLTPO
  Sbjct: 1141   NQPDVVRQPPSPREGPLPAARPAGATLERAKTILSPGKNGVVKDVFAFGGAVENPEYLTPO
  Query: 1201   GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVPV
                GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVPV
  Sbjct: 1201   GGAAPQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVPV
    
```

Displaying the first (=best) hit from each family in the DISPLAY FAM mode allows for a precise review of relevant answers.

```

L7 ANSWER 3 OF 127 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY2
TI Binding peptides specific for the extracellular domain of ErbB2 and uses
therefor (Patent)
MTY Protein
SQL 1255
SCORE 2628 100% of query self score 2628
BLASTALIGN
  Query = 1255 letters
  Length = 1255
  Score = 2628 bits (6812), Expect = 0.0
  Identities = 1255/1255 (100%), Positives = 1255/1255 (100%)
Query: 1 MELAALCRWGLLALLPPGAASTQVCTGTMKRLRPASPETHLDMLRHLYQGCQVQGNL
MELAALCRWGLLALLPPGAASTQVCTGTMKRLRPASPETHLDMLRHLYQGCQVQGNL
Sbjct: 1 MELAALCRWGLLALLPPGAASTQVCTGTMKRLRPASPETHLDMLRHLYQGCQVQGNL
Query: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLFEDNYALAVLDNG
ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLFEDNYALAVLDNG
Sbjct: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLFEDNYALAVLDNG
.
.
.
Query: 1081 AGSDVFDGDLGMGAAGKLSLPHDPSPLQRYSEDPTVPLPSETDGYVAPLTCSPQPEYV
AGSDVFDGDLGMGAAGKLSLPHDPSPLQRYSEDPTVPLPSETDGYVAPLTCSPQPEYV
Sbjct: 1081 AGSDVFDGDLGMGAAGKLSLPHDPSPLQRYSEDPTVPLPSETDGYVAPLTCSPQPEYV
Query: 1141 NQPDVVRPQPPSPREGPLPAARPAGATLERAKTSLPGKNGVVKDVFAGGAVENPEYLTPO
NQPDVVRPQPPSPREGPLPAARPAGATLERAKTSLPGKNGVVKDVFAGGAVENPEYLTPO
Sbjct: 1141 NQPDVVRPQPPSPREGPLPAARPAGATLERAKTSLPGKNGVVKDVFAGGAVENPEYLTPO
Query: 1201 GGAAQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVPV
GGAAQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVPV
Sbjct: 1201 GGAAQPHPPPAFSPAFDNLYYWDQDPPERGAPPSTFKGTPTAENPEYLGLDVPV
.
.
.
L7 ANSWER 127 OF 127 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI Compositions and methods for treatment of neoplastic disease
(PublishedApplication)
MTY Protein
SQL 1260
SCORE 2310 87% of query self score 2628
BLASTALIGN
  Query = 1255 letters
  Length = 1260
  Score = 2310 bits (5985), Expect = 0.0
  Identities = 1103/1257 (87%), Positives = 1153/1257 (90%), Gaps = 2/125
Query: 1 MELAALCRWGLLALLPPGAASTQVCTGTMKRLRPASPETHLDMLRHLYQGCQVQGNL
MELAA CRWG LLALLPPG A TQVCTGTMKRLRPASPETHLDMLRHLYQGCQVQGNL
Sbjct: 4 MELAAWCRWGFLLALLPPGIAGTQVCTGTMKRLRPASPETHLDMLRHLYQGCQVQGNL
Query: 61 ELTYLPTNASLSFLQDIQEVQGYVLIAHNQVRQVPLQRLRIVRGTLFEDNYALAVLDNG
ELTY+P NASLSFLQDIQEVQGY+LIAHNQV+ +VPLQRLRIVRGTLFED YALAVLDN
Sbjct: 64 ELTYVPANASLSFLQDIQEVQGYMLIAHNQVQVPLQRLRIVRGTLFEDKYALAVLDNR
.
.
.
Query: 1139 YVNQPDVVRPQPPSPREGPLPAARPAGATLE
YVNQ +V+PQPP EGPLP RPAGATLE
Sbjct: 1144 YVNQSEVQPQPLTPEGPLPPVRPAGATLE
Query: 1199 PQGGAAPQHPPPAFSPAFDNLYYWDQDPP
P+ G A PHP PAFSPAFDNLYYWDQ+
Sbjct: 1204 PREGTASPPHPSPAFSPAFDNLYYWDQNSS

```

The presented search strategy started with two sequence homology searches generating 10,000 answers each.

The new percent of self score feature and the STN FSORT command supported the refinement of the answer sets to focus on 177 patent publications with highly similar and highly relevant sequence data.

Likewise, the same search strategy could also be used within the PCTGEN file.

FIZ Karlsruhe
STN Europe
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
Germany

E-mail: helpdesk@fiz-karlsruhe.de
Phone: +49 7247 808 555
Fax: +49 7247 808 259
www.fiz-karlsruhe.de

