



STN is operated in North America
by Chemical Abstracts Service.

STN Database Summary Sheet

PCTGEN (World Patent Application Biosequences) covers nucleotide and amino acid sequence information submitted electronically to the World Intellectual Property Organization (WIPO) by patent applicants.

Records contain sequence and patent information as given by the patent applicant. Each record includes the actual sequence and additional information on the sequence, e.g., molecule type and organism, and patent information.

For direct code match or similarity (homology) sequence searching, FIZ Karlsruhe provides three specialized RUN package options, GETSEQ, GETSIM, and BLAST®.

Subject Coverage

- Nucleotide and amino acid sequence data as submitted electronically by patent applicants to the World Intellectual Property Organization (WIPO)

Sources

- Data submitted electronically by patent applicants

File Data

- August 2001 to the present
- More than 4,509,600 records (6/06)
- More than 3,992,600 nucleic acids (6/06)
- More than 520,700 proteins (6/06)
- Updated weekly
- Automatic current-awareness searches (SDIs) and ALERT sequence searches are run weekly

User Aids

- Online Helps (HELP DIRECTORY lists all help messages available)
- STNGUIDE
- BLAST® from NCBI:
www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

Database Producer

World Intellectual Property Organization
34, Chemin des Colombettes
1211 Geneva
Switzerland

Phone: (+41) 22 338 91 11
Fax: (+41) 22 338 98 20

FIZ Karlsruhe
P.O. Box 2465
D-76012 Karlsruhe
Germany

STNmail: HLPDESKK
Phone: (+49) 7247/808-555
Fax: (+49) 7247/808-259
E-mail: helpdesk@fiz-karlsruhe.de

Database Supplier

FIZ Karlsruhe
P.O. Box 2465
D-76012 Karlsruhe
Germany
STNmail: HLPDESKK
Phone: (+49) 7247/808-555
Fax: (+49) 7247/808-259
E-mail: helpdesk@fiz-karlsruhe.de

In North America
CAS
STN North America
P.O. Box 3012
Columbus, Ohio 43210-0012 U.S.A.

CAS Customer Care:
Phone: 800-753-4227 (North America)
614-447-3700 (worldwide)
Fax: 614-447-3751
E-mail: help@cas.org
Internet: www.cas.org

In Europe
FIZ Karlsruhe
STN Europe
P.O. Box 2465
76012 Karlsruhe
Germany
Phone: +49-7247-808-555
Fax: +49-7247-808-259
E-mail: helpdesk@fiz-karlsruhe.de
Internet: www.stn-international.de

In Japan
JAICI (Japan Association for
International Chemical Information)
STN Japan
Nakai Building
6-25-4 Honkomagome, Bunkyo-ku
Tokyo 113-0021, Japan
Phone: +81-3-5978-3601 (Technical Service)
+81-3-5978-3621 (Customer Service)
Fax: +81-3-5978-3600
E-mail: helpdesk@jaici.or.jp (Technical Service)
cas-stn@jaici.or.jp (Customer Service)
Internet: www.jaici.or.jp

PCTGEN

SEARCH and DISPLAY Fields

The field that allows left truncation (/FEAT) is marked with an asterisk (*).

Search Field Name	Search Code	Search Examples	Display Codes
Basic Index (contains single words from the title (TI), organism species (ORGN), and molecule type (MTY) fields)	None (or /BI)	S ANAPHYLATOXIN S PLANT GENE# AND RNA	TI, ORGN, MTY
Accession Number	/AN	S 2002060924.37/AN	AN
Application Country (code and text)	/AC	S US/AC	AI
Application Date (1)	/AD	S 20011129/AD	AI
Application Number (2)	/AP	S US2001-809003/AP	AI
Application Year (1)	/AY	S 2002/AY	AI
Document Type (code and text)	/DT (or /TC)	S PATENT/DT	DT
Entry Date (1)	/ED	S 20021004/ED	ED
Feature Table* (3)	/FEAT	S (RNA AND BINDING)/FEAT S ?COMBINAT?/FEAT	FEAT
File Segment (code and text)	/FS	S PROTEIN/FS S NS/FS	FS
Molecule Type	/MTY	S RNA/MTY	MTY
Organism	/ORGN	S CRASSOSTREA GIGAS/ORGN	ORGN
Patent Assignee (4)	/PA (or /CS)	S MOLECULAR DYNAMICS/PA	PA
Patent Country (code and text)	/PC	S WO/PC	PI
Patent Number (2)	/PN (or /PATS)	S WO 2002074961/PN	PI
Publication Date (1)	/PD	S 20030130/PD	PI
Publication Year (1)	/PY	S 2003/PY	PI
Related Application Country (code only)	/RLC	S FR/RLC	RLI
Related Application Date (1)	/RLD	S 20020208/RLD	RLI, RLIO
Related Application Number (2)	/RLN (or /RLI)	S EP2001-1102050/RLN	RLI, RLIO
Related Application Year (1)	/RLY	S L1 AND 2000-2001/RLY	RLI, RLIO
Sequence Identity Number (1)	/SEQN	S 337/SEQN	SEQN
Sequence Length (1)	/SQL	S 150-175/SQL	SQL
Title	/TI	S HYBRIDIZATION ASSAY#/TI	TI
Update Date (1)	/UP	S L1 AND UP>=20030200	UP

(1) Numeric search field that may be searched with numeric operators or ranges.

(2) Either STN format or Derwent format may be used.

(3) In addition to right truncation, left and simultaneous left and right truncation are available in this field. At least 4 characters need to be used for the length of the stem.

(4) Search with implied (S) proximity is available in this field.

Super Search Fields

Enter a super search code to execute a search in one or more fields that may contain the desired information.

Super search fields facilitate crossfile and multifile searching. EXPAND may not be used with super search fields.

Use EXPAND with the individual field codes instead.

Search Field Name	Search Code	Fields Searched	Search Examples	Display Codes
Application Number Group (1)	/APPS	/AP, /RLN	S US2001-809003/APPS	AI, RLI

(1) Either STN format or Derwent format may be used.

Sequence Similarity Searching (BLAST/GETSIM)

The BLAST® and GETSIM run packages are available to search for protein and nucleotide sequence data by similarity (homology). BLAST is provided with the permission of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). GETSIM is provided by FIZ Karlsruhe GmbH, and is based upon the FASTA algorithm.

To initiate a BLAST or GETSIM search the following search codes have to be specified:

/SQP - searching peptide sequences (default)
 /SQN - searching nucleotide sequences
 /TSQN - searching a database of peptide sequences translated from PCTGEN nucleotide sequences

The BLAST or GETSIM search can be run in offline BATCH mode or used as the basis of a current-awareness ALERT. When using the SQN or TSQN, it is possible to specify whether single (SIN), complementary (COM), or BOTH strands should be searched. These options can be specified together with the search code, e.g., /SQN COM. SIN is the default for GETSIM; BOTH is the default for BLAST.

Nucleotide and protein sequences can be subjected to a similarity search in various ways. A query can be prepared with the QUERY command and saved beforehand; it can be entered directly on the command line using RUN BLAST or RUN GETSIM, or it may be uploaded from an ASCII file using the UPLOAD command.

A diagram is generated that shows the similarity between the retrieved sequences and the query. The x-axis represents the number of answers with a specific degree of similarity (represented by y-axis). The entire answer set or only the most relevant (at your choice) answers can be kept. The generated L-number contains these answers, but they are sorted by descending accession number. This L-number may be rearranged by descending similarity score. Enter SOR SCORE D and the L-number at an arrow prompt (=>).

It is possible to see the alignment between the retrieved sequence and the query sequence using ALIGN for BLAST or GETSIM. The top line is the query sequence and the bottom line is the hit sequence. BLAST ALIGN follows the standard convention for NCBI alignment displays. GETSIM ALIGN uses two dots to represent identical nucleotides/peptides, a blank if there is no match, and one dot to indicate a chemical "family" match. Gaps inserted in the query or answer sequence for alignment purposes are shown with an underscore.

In addition to the sequence alignment, SEQO display format shows the corresponding original sequence, which might include the nucleotide sequence of a PCTGEN record, together with the protein sequence it expresses as given by the patent applicant.

BLAST/GETSIM Types of Searches

Description	Search Code	Search Example (1)
Peptide Homology	/SQP	RUN BLAST L1/SQP
Nucleotide Homology	/SQN	RUN BLAST L1/SQN
Single Strand (2)		RUN GETSIM L1/SQN SIN
Complementary Strand		RUN GETSIM L1/SQN COM
Both Strands (3)		RUN BLAST L1/SQN BOTH
Translated Peptide Homology	/TSQN	RUN BLAST L1/TSQN
Single Strand (2)		RUN GETSIM L1/TSQN SIN
Complementary Strand		RUN BLAST L1/TSQN COM
Both Strands (3)		RUN BLAST L1/TSQN BOTH

(1) L1 is a sequence query generated using the UPLOAD or QUERY.

(2) GETSIM default.

(3) BLAST default.

To RUN BLAST or GETSIM BATCH searches, include BATCH in the command line, e.g., RUN BLAST L1/TSQN BOTH BATCH; RUN GETSIM L1/TSQN BOTH BATCH.

To RUN BLAST or GETSIM current-awareness ALERT searches, include ALERT in the command line, e.g., RUN BLAST L1/SQN COM ALERT; RUN GETSIM L1/TSQN BOTH ALERT. Homology ALERT searches run every update, i.e., once a week.

PCTGEN

Advanced User Options for BLAST

For the experienced user of BLAST, there are many options available. Altering these parameters will have a profound effect on the outcome of the search. FIZ Karlsruhe strongly recommends that users are completely familiar with NCBI documentation before embarking on customizing any of these settings. For further information see:

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

Advanced user options are specified with a single-letter code preceded by a hyphen and followed by a blank and the required value, e.g., RUN BLAST L1/SQN -E 0.1.

Advanced User Options

Option	Switch	Values
Filter	-f	T (default), F, C If T is set, for peptides the SEG, and for nucleotides the DUST filter is employed. C symbolizes the "coiled coiled" filter.
Expectation Value	-e	Floating point number (default is 10)
Word Size	-w	Nucleotides: 7-23, 11 is the default Peptides: 2-3, 3 is the default
Strand	-s	1 (SIN), 2 (COM), or 3 (BOTH) (default)
Matrix	-m	BLOSUM62 (default), BLOSUM80, BLOSUM45, PAM30, or PAM70
Gap Penalty	-g	Nucleotides: 5 (default) Peptides: 11 (default)
Gap Extension	-x	Nucleotides: 2 (default) Peptides: 1 (default)
Penalty for nucleotide mismatch	-q	-3 (default)
Reward for nucleotide match	-r	1 (default)

BLAST Matrix settings

For a certain matrix, only a restricted set of possible gap and gap extension values is possible. The settings available to each matrix are summarized in the table. Default settings are indicated. Any other combinations will be rejected by the system and a warning message will be issued.

Matrix	Gap	Gap Extension
BLOSUM62	9	2
	8	2
	7	2
	12	1
	11	1 (default)
	10	1
BLOSUM80	8	2
	7	2
	6	2
	11	1
	10	1 (default)
	9	1
BLOSUM45	13	3
	11	3
	12	3
	9	3
	15	2 (default)
	14	2
	13	2
	12	2
	19	1
	18	1
	17	1
16	1	
PAM30	7	2
	6	2
	5	2
	10	1
	8	1
	9	1 (default)
PAM70	8	2
	7	2
	6	2
	11	1
	10	1 (default)
	9	1

PCTGEN

Searching Sequence Data with the GETSEQ RUN Package

Sequence information (protein and nucleic acid sequences) may be retrieved using a variety of search fields available with the GETSEQ package in the RUN command. The query used with GETSEQ may be created with the QUERY command or may be entered directly within RUN GETSEQ. Also, the L-number from a previous sequence search in another STN file with biosequence search capabilities, e.g., REGISTRY,DGENE, may be used as the query.

Sequence Search Terms

Terms	Examples
Single-letter codes for common amino acids (1)	QUE LAGLL/SQSP
Three-letter codes for common amino acids (1) Enclose codes or strings of codes in single quotes. Use dashes to separate codes in strings.	QUE 'THR-SER-GLY-MET-THR'/SQSFP QUE 'GLP'GY/SQSP RUN GETSEQ 'CYS-ASN-THR-ALA'/SQSP
Single-letter codes for nucleic acids (2)	QUE ATGAAN/SQEN RUN GETSEQ ATGAAN/SQSN

(1) Enter HELP AAC at an arrow prompt to display a table of the 1- and 3-letter codes for common amino acids.

(2) Enter HELP NUC at an arrow prompt to display a table of the codes for nucleic acids.

Types of Sequence Searches

Sequence data for nucleic acid and protein sequences are displayed in the SEQ field with 1-letter codes and in the SEQ3 field with 3-letter codes for proteins only.

Type	Definition	Code	Examples
Sequence Exact, Protein	Search for sequences that match the query. The query must be completely defined.	/SQEP	QUE AFFFF/SQEP QUE 'ALA-PHE-PHE-PHE-PHE'/SQEP
Sequence Exact Family, Protein	Search for sequences that match the query and those in which family-equivalent substitution of the query amino acids occurs. (1)	/SQEFP	RUN GETSEQ YGGFL/SQEFP QUE 'TYR-GLY-GLY-PHE-LEU'/SQEFP
Subsequence, Protein	Search for exact answers plus sequences in which the query sequence is embedded. Variability symbols are allowed.	/SQSP	RUN GETSEQ LAGLL/SQSP QUE 'GLP'GY/SQSP
Subsequence Family, Protein	Search for exact sequences, subsequences, and answers in which family-equivalent substitution of the query amino acids occurs. (1)	/SQSFP	RUN GETSEQ ATCXAW/SQSFP QUE 'LEU-ALA-GLY-LEU-LEU'/SQSFP
Sequence Exact, Nucleic Acid	Search for sequences that match the query. Ambiguity codes for nucleic acids are allowed.	/SQEN	QUE ATGAAN/SQEN
Subsequence, Nucleic Acid	Search for exact answers, plus sequences in which the query sequence is embedded. Ambiguity codes for nucleic acids and variability symbols are allowed.	/SQSN	RUN GETSEQ ATGAAN/SQSN

(1) The families of amino acid equivalents retrieved in protein family searches are:

P, A, G, S, T	(neutral, weakly hydrophobic)
Q, N, E, D, B, Z	(hydrophilic, acid amine)
H, K, R	(hydrophilic, basic)
F, Y, W	(hydrophobic, aromatic)
L, I, V, M	(hydrophobic)
C	(cross-link forming)

Variability Symbols for GETSEQ Subsequence Queries

For more information on specifying variability, enter HELP SQQ at an arrow prompt (=>).

Symbol	Function	Examples
[]	To specify alternate residues	RUN GETSEQ LGP[VL]/SQSP QUE LGP["VAL"LEU"LYS']/SQSP
[-]	To exclude a specific residue or alternate residues	QUE LGP[-H]/SQSP QUE LGP[-'HIS']/SQSP RUN GETSEQ LGP[-HL]/SQSP
{m}	To repeat the preceding sequence or sequence query or L-number m times	QUE LGP[-HL]/SQSP RUN GETSEQ L4{2}/SQSP RUN GETSEQ TAA(TAAA){2}/SQSN RUN GETSEQ (CTG){2}/SQSN
{m,u} or {m-u}	To repeat the preceding sequence or sequence query or L-number m to u times	RUN GETSEQ GG(FL){1,2}/SQSP RUN GETSEQ L3{1,3}/SQSP RUN GETSEQ (CTG){1,3}/SQSN
? or {0,1} or {0-1}	To repeat the preceding sequence or sequence query or L-number zero or one time	RUN GETSEQ FLRRI(RP)?K/SQSP RUN GETSEQ L1{0-1}NN/SQSP RUN GETSEQ L1{0,1}NN/SQSP RUN GETSEQ CAT(CGA){0,1}GGAC/SQSN
* or {0,} or {0-}	To repeat the preceding sequence or sequence query or L-number zero or more times	RUN GETSEQ KLK(WD)*N/SQSP RUN GETSEQ L1{0-}NN/SQSP RUN GETSEQ CAT(CTG){0,}TATT/SQSN
+ or {1,} or {1-}	To repeat the preceding sequence or sequence query or L-number one or more times	RUN GETSEQ KLK(DLE){1,}/SQSP RUN GETSEQ KLK(DLE)+/SQSP RUN GETSEQ L2{1-}/SQSP RUN GETSEQ CAT(CTG){1,}TATT/SQSN
&	To join together expressions or L-numbers	RUN GETSEQ L1&L3/SQSFP RUN GETSEQ L2&L5{1,3}/SQSP

In addition, the caret and the vertical bar may be used.

The caret is used at the beginning or at the end of a sequence to search for that sequence at the beginning or end of sequence field.

The vertical bar is the symbol for alternation, i.e., it is used to separate alternate sequence queries.

PCTGEN**Specifying Gaps in GETSEQ Sequence Queries**

Symbol	Function	Examples
.	a gap of one residue	QUE SY.RPG/SQSP RUN GETSEQ SY..RPG/SQSP QUE AAG...TGC/SQSN
.{m} or [m.]	a gap of m residues	QUE SY.{2}RPG/SQSP QUE SY[2.]RPG/SQSP
.{m,u} or .{m-u}	a gap of m to u residues	RUN GETSEQ GFF.{2,10}LSS/SQSP RUN GETSEQ GFF.{2-10}LSS/SQSP QUE AAG.{2,5}TGC/SQSN
: or .? or .{0,1} or .{0-1}	a gap of zero or one residue	RUN GETSEQ AGA:SRI/SQSFP RUN GETSEQ AGA.?SRI/SQSFP QUE AGA.{0,1}SRI/SQSFP QUE AGA.{0-1}SRI/SQSFP
.* or .{0,} or .{0-}	a gap of zero or more residues	RUN GETSEQ HLC.*TYG/SQSP QUE HLC.{0,}TYG/SQSP RUN GETSEQ HLC.{0-}TYG/SQSP QUE AAGGCAGATG.*GCAA/SQSN
.+ or .{1,} or .{1-}	a gap of one or more residues	RUN GETSEQ SY.+TH/SQSFP QUE SY.{1-}TH/SQSFP QUE SY.{1,}TH/SQSP RUN GETSEQ TCCTG.+GTGG/SQSN

DISPLAY and PRINT Formats

Any combination of display fields and formats may be used to display or print answers. Multiple codes must be separated by commas or spaces, e.g., D L1 1-5 TI PA SEQ.

Hit-term highlighting is available in AN, DT, MTY, ORGN, PA, RLI, RLIO, SEQN, SQL, and TI. Highlighting must be ON during SEARCH in order to use the HIT, KWIC, and OCC formats.

Format	Content	Examples
AI (AP) (1) AIO (2) AN DT (TC) ED (3) FEAT FS (2,3) MTY (3) ORGN (3) PA (CS) PI (1) (PATS, PN) RLI (1) RLIO (2) SEQ (5) SEQ3 (5) SEQN (2,3) SEQO (2,5) SQL (3) TI (3) UP (2,3)	Application Information Application Information, Original Accession Number Document Type Entry Date Feature Table File Segment Molecule Type Organism Name Patent Assignee Patent Information Related Application Information Related Application Information, Original Sequence (1-letter codes) Sequence (3-letter codes) Sequence Identity Number Original Sequence (alignment of nucleotide sequence and peptide sequence it expresses when given) Sequence Length Title Update Date	D 1 2 AI D AIO D AN TI D DT D ED D 1 5 10 FEAT D TI FS D L5 1,10 MTY D ORGN D 1-25 PA D 1-15 PA PI D RLI D 2 RLIO D 1-3 TI SEQ D TI SEQ3 D SEQN D SEQO D 1-20 SQL D L7 2, 5-6-TI D TI UP
ALIGN (3,4) ALL (1,5) APPS (1) BIB (1) IALL (1,5) IBIB (1) SCORE (3,4) SQIDE (5) SQ3IDE (5) SCAN (3,6) TRIAL (3) (TRI, SAM, FREE)	Alignment between query and retrieved sequence in a similarity search (RUN BLAST or RUN GETSIM) AN, MTY, TI, PA, PI, AI, RLI, DT, ORGN, SQL, SEQ, FEAT AI, RLI AN, MTY, TI, PA, PI, AI, RLI, DT (BIB is the default) ALL, indented with text labels BIB, indented with text labels Similarity Score TI, SQL, SEQ, FEAT TI, SQL, SEQ3, FEAT TI (random display without answer numbers) TI, MTY, SQL	D ALIGN 1, 10, 100 D ALL D APPS D BIB ALIGN D IALL D IBIB ALIGN D TI SCORE D SQIDE D SQ3IDE 1 5 D SCAN D TRIAL TOTAL
HIT KWIC OCC (3)	Fields containing hit terms Hit terms with 20 words on either side (KeyWord-In-Context) Number of occurrences of hit terms and fields in which they occur	D HIT D KWIC NOH D OCC

(1) By default, patent, application, and priority numbers are displayed in STN Format. To display them in Derwent format, enter SET PATENT DERWENT at an arrow prompt. To reset the display to STN format, enter SET PATENT STN.

(2) Custom display only.

(3) No online display fee for this format.

(4) Use RUN BLAST or RUN GETSIM first.

(5) Sequences in PCTGEN are given according to WST.25 of the WIPO.

(6) SCAN must be specified on the command line, i.e., D SCAN or DISPLAY SCAN.

PCTGEN**SELECT, ANALYZE, and SORT Fields**

The SELECT command is used to create E-numbers containing terms taken from the specified field in an answer set.

The ANALYZE command is used to create an L-number containing terms taken from the specified field in an answer set.

The SORT command is used to rearrange the search results in either alphabetic or numeric order of the specified field(s).

Field Name	Field Code	ANALYZE/ SELECT (1)	SORT
Accession Number	AN	N	Y
Application Country	AC	Y (2)	Y
Application Date	AD	Y (2)	Y
Application Information	AI	Y (2,3)	Y
Application Number	AP	Y (2)	Y
Application Number and Related Application Number	APPS	Y (2)	N
Application Year	AY	Y (2)	Y
Document Type	DT	Y	Y
Entry Date	ED	Y (2)	Y
File Segment	FS	Y (2)	Y
Molecule Type	MTY	Y	Y
Organism Name	ORGN	Y	Y
Patent Assignee	PA	Y	Y
Patent Country	PC	Y	Y
Patent Information	PI	N	Y
Patent Number	PN	Y (2)	Y
	PATS	Y (2)	Y
Publication Date	PD	Y (2)	Y
Publication Year	PY	Y (2)	Y
Related Application Country	RLC	Y	Y
Related Application Date	RLD	Y	Y
Related Application Information	RLI	Y (4)	Y
Related Application Number	RLN	Y	Y
Related Application Year	RLY	Y	Y
Sequence (1-letter codes)	SEQ	Y (2,5)	N
Sequence (3-letter codes)	SEQ3	Y (2,5)	N
Sequence Identity Number	SEQN	Y	Y
Sequence Length	SQL	Y	Y
Similarity Score	SCORE	N	Y (6)
Title	TI	Y (default)	Y
Update Date	UP	Y (2)	Y

(1) HIT may be used to restrict terms extracted to terms that match the search expression used to create the answer set, e.g., SEL HIT PC.

(2) SELECT HIT and ANALYZE HIT are not valid with this field.

(3) Selects or analyzes the application number and appends /AP to the terms created by SELECT.

(4) Selects or analyzes the related application numbers and appends /RLN to the terms created by SELECT.

(5) Appends /SQSP to the terms created by SELECT.

(6) Used with an L-number created with BLAST or GETSIM.

Sample Records

DISPLAY IALL

ACCESSION NUMBER: 2001057272.15599 DNA PCTGEN
 TITLE: HUMAN GENOME-DERIVED SINGLE EXON NUCLEIC ACID PROBES USEFUL
 FOR ANALYSIS OF GENE EXPRESSION IN HUMAN PLACENTA
 PATENT ASSIGNEE: Molecular Dynamics, Inc.Penn, Sharron G.Rank, David R.Hanzel,
 David K.Chen, Wensheng
 PATENT INFO: WO 2001057272 20010809
 REL APPL INFO: US 2000-180312P 20000204; US 2000-207456P 20000526; US
 2000-632366 20000803; GB 2000-24263 20001003; US 2000-236359P
 20000927; US 2000-234687P 20000921; US 2000-608408 20000630
 FILE UPDATE DATE: 20020923
 DOCUMENT TYPE: Patent
 ORGANISM: Homo sapiens
 SEQUENCE LENGTH: 100
 SEQUENCE

1 cccagagatt ctgattctgc aaatcttgag cagcctgaga ttctgcagtt
 51 ctatgaagct tccaggtagt gtcaatgctg gtgctaggct gaccatagta

FEATURE TABLE:

Key	Location
	MAP TO AL035448.28
	EXPRESSED IN PLACENTA, SIGNAL
	= 1.5
	NT HIT: U29185.1, EVALUE
	7.00e-04
	EST_HUMAN HIT: AA047634.1,
	EVALUE 2.20e-01

DISPLAY SQIDE

AN 2002070737.12103 DNA PCTGEN
 TI Compositions and Methods Relating to Osteoarthritis
 SQL 100
 SEQ

1 agttngtgc cgttggaccg naggaaaact catagactca tgggagcgtg
 51 aggcttcgag cgcctaatt ttttaaccct aaatgtcgaaggcttctgg

FEATURE TABLE:

Key	Location
misc_feature	6, 21 n = A,T,C or G

DISPLAY SEQO

SEQO
 cgctcgcagt ctgtgggccc tccgggaggg ggcggaggtc accgcgggga gaggggcggg 60
 cgcagc atg gca gcc tcc tta cgg ctc ctc gga gct gcc tcc ggt ctc 108
 Met Ala Ala Ser Leu Arg Leu Leu Gly Ala Ala Ser Gly Leu
 1 5 10
 cgg tac tgg agc cgg cgg ctg cgg ccg gca gcc ggc agc ttt gca gcg 156
 Arg Tyr Trp Ser Arg Arg Leu Arg Pro Ala Ala Gly Ser Phe Ala Ala
 15 20 25 30
 gtg tgt tct agg tca gtg gct tca aag act cca gtt gga ttc att gga 204

PCTGEN**DISPLAY SEQO (cont'd)**

Val	Cys	Ser	Arg	Ser	Val	Ala	Ser	Lys	Thr	Pro	Val	Gly	Phe	Ile	Gly		
				35					40					45			
ctg	ggc	aac	atg	ggg	aat	cca	atg	gca	aaa	aat	ctc	atg	aaa	cat	ggc		252
Leu	Gly	Asn	Met	Gly	Asn	Pro	Met	Ala	Lys	Asn	Leu	Met	Lys	His	Gly		
			50					55					60				
tat	cca	ctt	att	att	tat	gat	gtg	ttc	cct	gat	gcc	tgc	aaa	gag	ttt		300
Tyr	Pro	Leu	Ile	Ile	Tyr	Asp	Val	Phe	Pro	Asp	Ala	Cys	Lys	Glu	Phe		
		65					70					75					
caa	gat	gca	ggt	gaa	cag	gta	gta	tct	tcc	cca	gca	gat	gtt	gct	gaa		348
Gln	Asp	Ala	Gly	Glu	Gln	Val	Val	Ser	Ser	Pro	Ala	Asp	Val	Ala	Glu		
4			80					85					90				
aaa	gct	gac	aga	att	att	aca	atg	ctg	ccc	acc	agt	atc	aat	gca	ata		396
Lys	Ala	Asp	Arg	Ile	Ile	Thr	Met	Leu	Pro	Thr	Ser	Ile	Asn	Ala	Ile		
	95				100					105					110		
gaa	gct	tat	tcc	gga	gca	aat	ggg	att	cta	aaa	aaa	gtg	aag	aag	ggc		444
Glu	Ala	Tyr	Ser	Gly	Ala	Asn	Gly	Ile	Leu	Lys	Lys	Val	Lys	Lys	Gly		
				115					120					125			
tca	tta	tta	ata	gat	tcc	agc	act	att	gat	cct	gca	gtt	tca	aaa	gaa		492
Ser	Leu	Leu	Ile	Asp	Ser	Ser	Thr	Ile	Asp	Pro	Ala	Val	Ser	Lys	Glu		
			130					135					140				
ttg	gcc	aaa	gaa	ggt	gag	aaa	atg	gga	gca	ggt	ttc	atg	gat	gcc	cct		540
Leu	Ala	Lys	Glu	Val	Glu	Lys	Met	Gly	Ala	Val	Phe	Met	Asp	Ala	Pro		
		145				150						155					
gtt	tct	ggt	ggt	gta	gga	gct	gca	cga	tct	ggg	aac	ctc	acg	ttt	atg		588
Val	Ser	Gly	Gly	Val	Gly	Ala	Ala	Arg	Ser	Gly	Asn	Leu	Thr	Phe	Met		
		160				165					170						
gtg	gga	gga	ggt	gaa	gat	gaa	ttt	gct	gct	gcc	caa	gag	ttg	ctg	ggg		636
Val	Gly	Gly	Val	Glu	Asp	Glu	Phe	Ala	Ala	Ala	Gln	Glu	Leu	Leu	Gly		
				175			180			185					190		
tgc	atg	ggc	tcc	aac	gtg	gtg	tac	tgt	gga	gct	ggt	ggg	act	ggg	cag		684
Cys	Met	Gly	Ser	Asn	Val	Val	Tyr	Cys	Gly	Ala	Val	Gly	Thr	Gly	Gln		
				195					200					205			
gcg	gca	aag	atc	tgc	aac	aac	atg	ctg	tta	gct	att	agt	atg	att	gga		732
Ala	Ala	Lys	Ile	Cys	Asn	Asn	Met	Leu	Leu	Ala	Ile	Ser	Met	Ile	Gly		
			210					215					220				
act	gct	gaa	gct	atg	aat	ctt	gga	atc	agg	tta	ggg	ctt	gac	cca	aaa		780
Thr	Ala	Glu	Ala	Met	Asn	Leu	Gly	Ile	Arg	Leu	Gly	Leu	Asp	Pro	Lys		
			225				230						235				
cta	ctg	gct	aaa	atc	cta	aat	atg	agc	tca	gga	cgg	tgt	tgg	tca	agt		828
Leu	Leu	Ala	Lys	Ile	Leu	Asn	Met	Ser	Ser	Gly	Arg	Cys	Trp	Ser	Ser		
			240				245					250					
gac	act	tat	aat	cct	gta	cct	gga	gtg	atg	gat	ggc	ggt	ccc	tcg	gct		876
Asp	Thr	Tyr	Asn	Pro	Val	Pro	Gly	Val	Met	Asp	Gly	Val	Pro	Ser	Ala		
				255		260				265				270			
aat	aac	tat	cag	ggt	gga	ttt	gga	aca	aca	ctc	atg	gct	aag	gat	ctg		924
Asn	Asn	Tyr	Gln	Gly	Gly	Phe	Gly	Thr	Thr	Leu	Met	Ala	Lys	Asp	Leu		
				275						280				285			
gga	ttg	gca	caa	gac	tct	gct	acc	agc	aca	aag	agc	cca	atc	ctt	ctt		972
Gly	Leu	Ala	Gln	Asp	Ser	Ala	Thr	Ser	Thr	Lys	Ser	Pro	Ile	Leu	Leu		
			290					295					300				
ggc	agt	ctg	gcc	cat	cag	atc	tac	agg	atg	atg	tgt	gca	aag	ggc	tac		1020
Gly	Ser	Leu	Ala	His	Gln	Ile	Tyr	Arg	Met	Met	Cys	Ala	Lys	Gly	Tyr		
			305				310					315					
tca	aag	aaa	gac	ttc	tca	tcc	gtg	ttc	cag	ttc	cta	cga	gag	gag	gag		1068
Ser	Lys	Lys	Asp	Phe	Ser	Ser	Val	Phe	Gln	Phe	Leu	Arg	Glu	Glu	Glu		
		320					325					330					
acc	ttc	tga	gtgtgcc	ctttggccac	ggacactggt	gggaaccaaa	ctctgtcttg										1124

DISPLAY SEQO (cont'd)

```

Thr Phe
335
gagcctcctt ttagctcact ccacaagtaa atggatttaa tcaaaggtca cctatctgct 1184
tttgattgtc taggtcacag taatccctag gatTTTTTc acgcttattct ttttgtcttt 1244
ttaacaaaca tattatccga atTTTTTTtc tgcaagccac tgatagtctc tgctaactag 1304
cttaattgac ctttttacia agtttgatcc ccaagcatcc tcaactaaat cattgaatac 1364
ttcaatcagg atattatctg ctttacttta caaataaaac caaatctttt gtcaacagga 1424
tgaaacccat cttaaaggaa agaaaaggaa ttggtgtgaa gagagaagtt agagaaggga 1484
aatgcagtga attactatct gtgtccatca ggaagtttgt cctgttaacc aaatggttac 1544
tgcactacca gggttactgg tttatTTTtc agggagctga taaagcagga gaactgTTgc 1604
tgcagtTTTT ctatttggac tccgtcacia tatggttagga tatccctcac caactcccga 1664
cactcagcag acttgtTTTT atatTTTTTt ctttcttTgt cattcttact acgtatTTTT 1724
tgacttaaga atgacatctt tagatgcatt tcagagccaa tgatgatatt tgcttttagat 1784
aattattata ttattataaa tatagccata ttatTTTTgaa ttcaaataaa tttctatact 1844
ggtaaaaaaaa aaaaa 1859

```

Search Examples

```

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

```

```

UPLOAD SUCCESSFULLY COMPLETED
L2 GENERATED

```

=> D L2 LQUE

```

L2 ANSWER 1 PCTGEN (C) 2002 WIPO
LQUE MAVMAPRTLL LVLSGVLALT QTWAGSHSMR YFYTSMSRPG RGEPRFFAVGYVDDTQFVRF
DSDAASQRME PRAPWVEQEG PEYWDRETQN MKAQTQNAPVNLRLRGYYN QSEAGSHTLQ
TMHGCDLGPD GRLLRGYYQS AYDGKDYFALNEDLRSWTAA DLAAQNTQRK WEAADVAEQI
RAYLEGRCVE WLRRYLENGKETLQRADPPK THVTHHPVSD HEATLRCWAV GFYPAEITLT
WQRDGEDQTQDTELMETRPA GDGTFQKWA VVPSGKEQR YTCHVQHEGL
PKPLTLRWEPSQSTIPIVG IIAGLVLLGA MVIGAVVA AV MWRKSSDRK
GGSYSQAASSDSAQGSVSL TACKV

```

=> RUN GETSIM L2/SQP

```

RUN GETSIM AT 14:07:03 ON 09 DEC 2002
COPYRIGHT (C) 2002 FIZ KARLSRUHE GMBH

```

```

20000 SEQUENCES PROCESSED
40000 SEQUENCES PROCESSED
.
.
.
260000 SEQUENCES PROCESSED

```

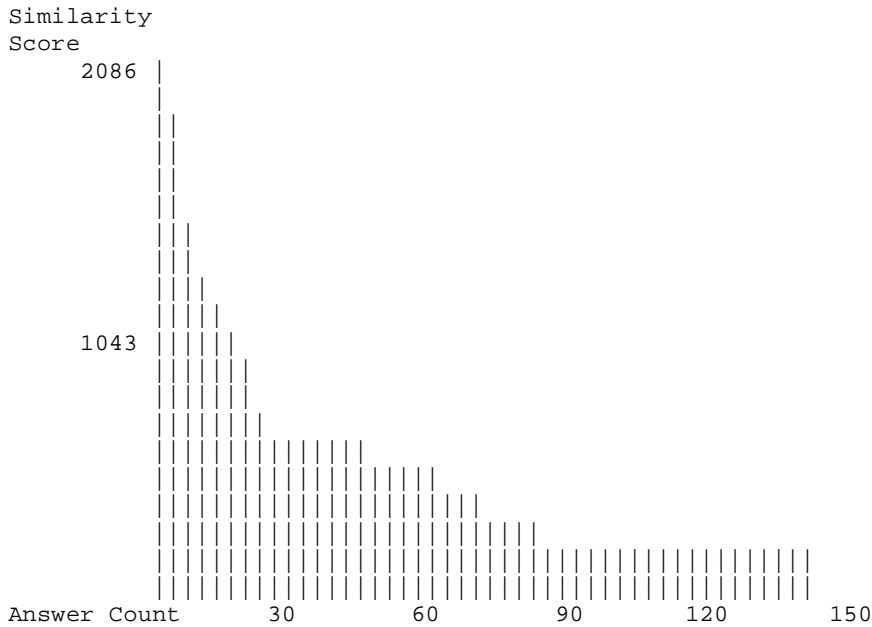
```

138 ANSWERS FOUND ABOVE A THRESHOLD OF 155
QUERY SELF SCORE VALUE IS 2497

```

PCTGEN

DISPLAY SEQO (cont'd)



HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ?:ALL

L3 RUN STATEMENT CREATED

```
L3      138 MAVMAPRTLLLVLSGVLALTQTWAGSHSMRYFYTSMSRPGRGEPRFFAVG
        YVDDTQFVRFSDAASQRMEPRAPWVEQEGPEYWDRETQNMKAQTQNAQPV
        NLRNLRGYYNQSEAGSHTLQTMHGCDLGPDRLLRGYYQSAYDGKDYFAL
        NEDLRSWTAADLAAQNTQRKWEAADVAEQIRAYLEGRVCVEWLRRYLENGK
        ETLQRADPPKTHVTHHPVSDHEATLRCWAVGFYPAEITLTWQRDGEDQTO
        DTELMETRPAGDGTQKWAAVVPSGKEQRYTCHVQHEGLPKPLTLRWEF
        SSQSTIPIVGI IAGLVLLGAMVIGAVVAAMWRRKSSDRKGGSYSQAASS
        DSAQGS DVSLTACKV/SQP
```

```
=> SOR SCORE D
    PROCESSING COMPLETED FOR L3
L4      138 SOR L3 SCORE D
```

=> D TI PA PI ALIGN 1

```
L4      ANSWER 1 OF 138 PCTGEN (C) 2002 WIPO
TI      Novel Nucleic Acids and Polypeptides
PA      HYSEQ, INC
PI      WO 2001064835      20010907
ALIGN   Smith-Waterman score: 2086
        366 aa overlap starting at 9
        mavmaprtllllvlsvglaltqtwagshsmryfytsmsrpggrgeprffavgyvddtqfvrf
        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
        mrvmaprtlillllsgalaltetwacshsmryfytavsrpgrgeprfiavgyvddtqfvrf
        dsdaasqrmeprapwveqegpeywdretqnmkaqtqnapvnlnrnlrgyyynqseagshtlq
        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
        .
        .
        .
```

=> RUN BLAST CATGGTGGTTAACTTACCTCATTAGCAGCATCCCTCTACAAGGTGCATTTAACTATAAGTATACT/SQN -E 100

BLAST Version 2.2

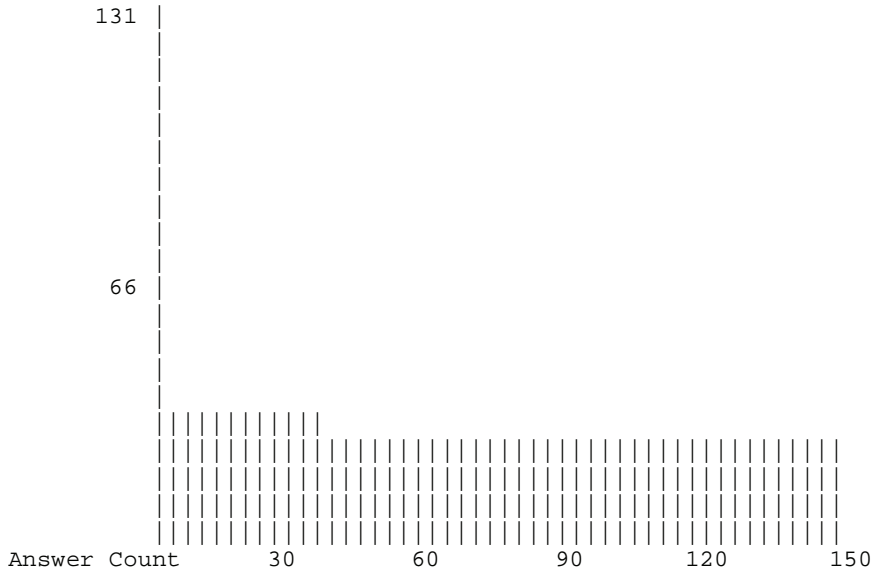
DISPLAY SEQO (cont'd)

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

.
.
.

144 ANSWERS FOUND BELOW EXPECTATION VALUE OF 100.0

Similarity
Score



HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ? :50

L1 RUN STATEMENT CREATED
L1 50 CATGGTGGTTAAACTTACCTCATTAGCAGCATCCCTCTACAAGGTGCATT
TAACTATAAGTATACT/SQN.-E 100

=> SOR SCORE D
PROCESSING COMPLETED FOR L1
L2 50 SOR L1 SCORE D

=> D 1 TI PA PI ALIGN

L2 ANSWER 1 OF 50 PCTGEN (C) 2002 WIPO
TI RNS-sekretierende Bakterien
PA Bachmann, Till; Villatte, Francois
PI WO 2002024904 20020328

BLASTALIGN

Query = 66 letters
Length = 452
Score = 131 bits (66), Expect = 7e-36
Identities = 66/66 (100%)
Strand = Plus / Plus

Query: 1 catggtgggttaaacttacctcattagcagcatccctctacaaggtgcatttaactataag
|||||
Sbjct: 201 catggtgggttaaacttacctcattagcagcatccctctacaaggtgcatttaactataag

Query: 61 tatact 66
|||||
Sbjct: 261 tatact 266