

# DERWENT GENESEQ

## **An introduction to Derwent GENESEQ**

© 2000 Thomson Derwent. All rights reserved

Edition 5  
ISBN: 0 901157 09 9

Copyright 2000 Derwent Information Limited  
Published by Derwent Information  
14 Great Queen Street, London WC2B 5DF,  
United Kingdom

Visit the Derwent web site at <http://www.derwent.com/>

Edition 1 published April 1996  
Edition 2 published May 1997  
Edition 3 published March 1998  
Edition 4 published March 2000  
Edition 5 published September 2000

ISBN: 0 901157 70 8 (Edition 3)  
ISBN: 0 901157 87 2 (Edition 4)  
ISBN: 0 901157 09 9 (Edition 5)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means – electronic, mechanical, recording, photocopying or otherwise – without express written permission from the copyright owner.

# Contents

An introduction to Derwent Information .....	1
An introduction to patents .....	2
An introduction to Derwent's GENESEQ Sequence Databases .....	3
How do GENESEQ and GENESEQ FASTA/ert compare? .....	5
Analysis of the GENESEQ and GENESEQ FASTA/ert databases .....	6
Availability and updating frequency .....	7
Why use Derwent GENESEQ and GENESEQ FASTA/ert? .....	8
Using Derwent GENESEQ and GENESEQ FASTA/ert .....	10
How does the information in Derwent GENESEQ compare with that in the Derwent World Patents Index? .....	12
The growth of biomolecular sequences in patents .....	13
Sample Derwent GENESEQ FASTA/ert records .....	14
Sample Derwent GENESEQ flat file records .....	15
Summary of Derwent GENESEQ flat file fields .....	17
Sample Derwent GENESEQ records on STN .....	19
Summary of Derwent GENESEQ fields on STN .....	21
Sequence searching on STN (GETSEQ and GETSIM) .....	23
Appendix I: Feature keys used in Feature Tables in Derwent GENESEQ .....	24
Appendix II: IUPAC codes used for nucleic acid and amino acid sequences in Derwent GENESEQ and GENESEQ FASTA/ert .....	32
Appendix III: Molecular type codes in GENESEQ .....	36
Appendix IV: Sample searches in GENESEQ on STN (file DGENE) .....	37
Appendix V: Countries covered by Derwent GENESEQ and GENESEQ FASTA/ert with start date for GENESEQ* .....	48
Appendix VI: Genetic Code Table .....	49
Appendix VII: Nucleic Acid and Amino Acid Sequences .....	50
Frequently asked questions .....	52
Customer support .....	55



# An introduction to Derwent Information

Derwent, the leading specialist in scientific and patent information, has for over 50 years provided vital information to companies and research institutes across the world.

Derwent World Patents Index (DWPI) is unrivalled in its comprehensive, enhanced patent information covering more than 8 million separate inventions from 40 patent-issuing authorities including the USPTO, WIPO, EPO, Japanese and German patent offices.

Derwent's products are designed to meet the needs of not only the major multinationals, but equally importantly, to fulfil the information demands of smaller, more specialised organisations.

Used by a global audience, Derwent's information products give a comprehensive picture of technological innovations worldwide – providing critical advantage by highlighting new opportunities, identifying competitors and assisting R&D.

As part of The Thomson Corporation, Derwent works closely with global leaders in the information industry to guarantee customers access to unequalled business and technological intelligence.

# An introduction to patents

The nature of patents and patent law makes these documents a unique data source. Under patent law the details of the invention must normally be kept secret until the patent has been filed. Studies have shown that 70% of inventions are only ever disclosed in patents. Monitoring patents therefore provides much earlier, and often unique, information than other sources. In addition patents have to contain enough detail for an expert specialising in the field to recreate the invention. Therefore, they give a more detailed description than is normally disclosed in a research paper. Each year one million patents are published, adding to the 30 million plus patent documents already filed.

- can provide details of an innovative new method, new research results, and fresh conclusions
- helps you track your competitors' work, giving you an insight into their product development strategy
- reveals possible infringements on your own patents
- adds a new dimension to your knowledge base

## How can patent information help you?

Patent information...

- is unique – 70% of new inventions are only ever disclosed in patents. Access to this data ensures that your knowledge of new technological advances is complete
- will help you identify potential products of the future, and new research areas

# An introduction to Derwent's GENESEQ sequence databases

Derwent's GENESEQ Sequence Databases consist of Derwent GENESEQ FASTA*Alert* and Derwent GENESEQ. These databases are produced by Derwent's qualified molecular biology experts, who assess each patent before producing individual records for each sequence within that patent. This team of qualified geneticists and molecular biologists receive patents from Derwent's Patent Administration team following patent publication, and then check through each page to find nucleic acid and amino acid sequences in claims, examples and other parts of the disclosure. The team includes individuals with foreign language skills who translate the required information to produce English language records.

appear in Derwent GENESEQ.

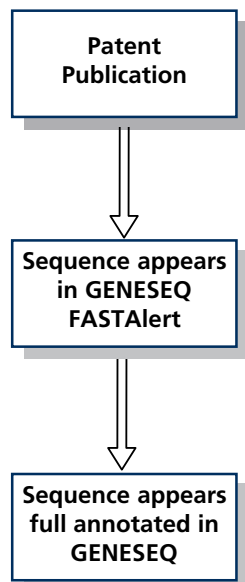
Derwent GENESEQ FASTA*Alert* records contain fully searchable sequences in FASTA format obtained from printed patent documents, which contain the legally recognised versions of the sequences.

## Derwent GENESEQ

Established in 1990, Derwent GENESEQ has rapidly grown to become the world's largest database dedicated to collection and annotation of sequence information from patents, dating back to 1980.

Derwent scans patents from 40 patent-issuing authorities for those which contain nucleic acid and amino acid sequences. These sequences are annotated and then added to the database which is updated every two weeks.

Derwent's molecular biology experts assess every patent to ensure that Derwent GENESEQ contains the most comprehensive, global coverage of sequence information.



## Derwent GENESEQ FASTA*Alert*

Launched early in 2000, Derwent GENESEQ FASTA*Alert* is the fast-alerting companion to Derwent GENESEQ. FASTA*Alert* covers biomolecular sequences in patents from 40 patent issuing authorities that are processed and updated into the database every week.

Derwent GENESEQ FASTA*Alert* is a rolling database that is updated with records within two weeks of patent publication. Records remain in the database until they

Every GENESEQ record is prepared by Derwent's team of experts to bring you:

- specially-written titles to cover the main developments and content of the patent
- full bibliographic data (see pages 17-18 for the fields included in each record)
- concise summaries detailing the pertinent points about the sequence
- specially written English language abstracts of patents originally in one of thirteen different languages (20% of the database records are translations from foreign language patents)
- Feature Tables which highlight specific areas within the sequence such as promoter regions, CAAT box, coding regions etc.
- searchable sequences

# How do GENESEQ and GENESEQ FASTA/ert compare?

Derwent GENESEQ FASTA/ert is designed as a companion database to GENESEQ.

- It enables swift searching of business critical sequences to ensure exclusivity and intellectual property rights.
- Records contain the legally recognised version of the nucleic acid or amino acid sequence, taken from the full patent specification not electronic patent filings
- The database is published as quickly as possible, within two weeks of the patent publication date.

On the other hand, Derwent GENESEQ is the fully annotated version.

- It contains detailed and comprehensive records, published after in depth analysis of patents containing sequences.

- The records contain easy to use and informative data on all aspects of the sequence, including patentee, inventor and priority information, details on the specific sequence and its features, and a unique abstract that includes the use of the invention.

- Each GENESEQ record also includes the Derwent Primary Accession Number to allow easy searching for corresponding records in other Derwent databases.

# Analysis of the **GENESEQ** and **GENESEQ FASTA/ert** databases

## **Derwent GENESEQ FASTA/ert**

All new nucleic acids (10 or more bases in length) and amino acids (4 or more residues in length) plus all PCR primers and probes (of any length) from patents are included in the database.

Every week about 2,000 - 3,000 new records are added to the database.

## **Derwent GENESEQ**

All new nucleic acids (10 or more bases in length) and amino acids (4 or more residues in length) plus all PCR primers and probes (of any length) from patents are included in the database.

As of January 2000, Derwent GENESEQ contained in excess of 590,000 records. Of the total records, 20% are foreign language patents which have been translated into English. The foreign language patents which are included in the database include patents from Japan, Germany, France, Latin America, Russia, Korea, China, Hungary and many more countries.

Every two weeks about 3,000 - 4,000 new records are added to the database.

# Availability and updating frequency

Derwent GENESEQ FASTA/*ert* is available in-house, directly from Derwent in FASTA format (ASCII text), with updates every week (ftp delivery).

Derwent GENESEQ is available:

- In-house – directly from Derwent in flat file EMBL format (ASCII text) or GCG format

The in-house database is currently updated every two weeks (ftp delivery) or quarterly (magnetic tape).

- Online – via STN (file DGENE) featuring the GETSIM/GETSEQ software

The online database is updated every two weeks.

- From October 2000, GENESEQ will be available on the DoubleTwist web portal. Customers will be able to use DoubleTwist's Agents to search and monitor the database and to analyse their results. Access to GENESEQ through DoubleTwist will require the purchase of a GENESEQ User Licence from Derwent. For more information on the web interface and DoubleTwist's computational tools, visit their web-site at <http://www.doubletwist.com>

---

**Note:** GETSEQ and GETSIM software are available on STN, in the DGENE file, and do not have to be bought as separate packages.

---

# Why use Derwent GENESEQ and GENESEQ FASTA/ert?

## Derwent GENESEQ FASTA/ert

- Save time and money in your research. Use GENESEQ FASTA/ert as a novelty filter to screen your proprietary sequences against those appearing in patents as soon as they appear.

Sequences in GENESEQ FASTA/ert appear within 2 weeks of publication in patents.

- To keep yourself informed of all the latest sequence discoveries.

## Derwent GENESEQ

- Save time by searching a focused database:

Derwent GENESEQ is the only database which focuses specifically on nucleic acid and amino acid sequences (including primers and probes) from patents.

- Get the complete picture by searching unique sequence information:

A significant proportion of the data available in Derwent GENESEQ is not

available in any other sequence database, so it is an invaluable source of new sequence information.

- Search patent sequence information from patent coverage that is wider than that found in other databases:

Although some other sequence and chemical databases include sequences from patents, none of them cover as many patent issuing authorities as Derwent GENESEQ does – this makes much of the Derwent GENESEQ information unique.

- Access sequence information years before it appears in other sequence databases:

Most Patent Offices, excluding the USPTO, publish unexamined patents within 18 months of application. The US still publish patents only once they have been granted. However, since most patents are filed in more than one country, this means they will usually be published outside of the USA first. Derwent's coverage of 40 patent issuing authorities allows access to sequence information from patents which may not yet have been released by the USPTO, or entered into other sequence databases.

**Typically, US priority patents appear in Derwent GENESEQ two to four years before they appear in other sequence databases.**

- Keep one step ahead of your competitors:

Not only will Derwent GENESEQ allow you to search your competitors' sequences so that you get an insight into their R & D strategies, it will also give you the chance of finding licensing opportunities before your competitors do. Derwent GENESEQ on STN will also allow easy cross-reference into the Derwent World Patents Index (DWPI) on STN and the whole patent family information from DWPI can be displayed with the Derwent GENESEQ record.

- Derwent GENESEQ brings user-friendly patent information into the library or research department.

Patent jargon is removed so that end users unfamiliar with patents can easily understand the scientific and technical content of the database.

- Save money by doing your own prior art searches:

Check the prior art situation before submitting your own patent applications, and also monitor infringements of your own patents.

- For the same reasons that you search other sequence databases:

To keep yourself well informed of all the latest sequence discoveries.

# Using Derwent GENESEQ and GENESEQ FASTA/ert

## Information Professionals

Every organisation that intends to apply for a patent containing sequence information, or which is embarking on a research project with the intention of developing a commercial product, must be able to investigate whether the sequence has already appeared in the patent literature to avoid risking millions of dollars in wasted effort.

Derwent GENESEQ and GENESEQ FASTA/ert enable professional searchers to identify patents containing a given sequence, or a sequence substantially homologous with it, in a single operation.

Each sequence record in GENESEQ FASTA/ert contains a country code and patent number, for cross-reference to full text patents.

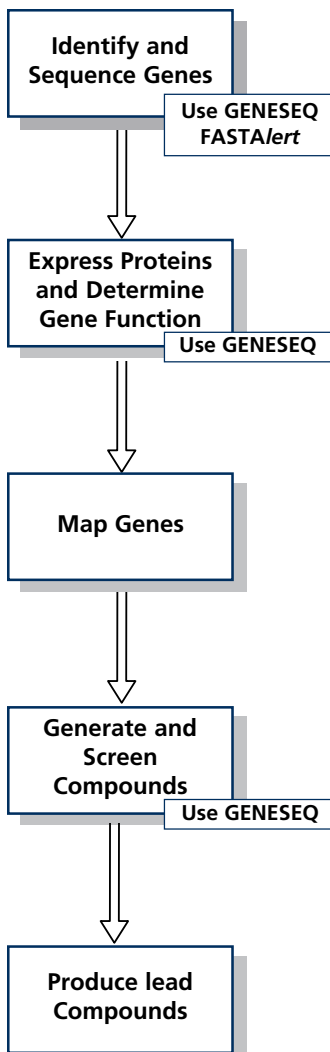
Each sequence record in GENESEQ contains patent bibliographic information such as the patent number, filing date, priority, patent assignee and inventors. Details are also provided of where the sequence appears in the patent specification so that it is easy to locate the retrieved sequence in the original document.

Derwent GENESEQ records include keyword indexing for fast retrieval of concepts, the enhanced Derwent title and a description of the sequence features and applications. Also included is the accession number for cross-reference to the Derwent World Patents Index, Derwent Biotechnology Abstracts and the Derwent Gene Therapy Database in which you can obtain patent family information plus a detailed abstract of the invention.

## Research Scientists

Derwent GENESEQ is an information tool for the research scientist. Patent documents contain a complete description of how the invention can be carried out, including detailed examples of the procedures involved. This information is often available long before an inventor is permitted to submit a paper to a journal and much of the information in patents never appears elsewhere in the literature.

GENESEQ and GENESEQ FASTA/ert together will help research scientists to make decisions by providing them with information on the intellectual property associated with any sequence, as soon as it is available (within two weeks of appearing in the patent document).



Derwent GENESEQ records are supplied in the familiar style of EMBL. The records are annotated to a high level of consistency by bioinformaticists working within Derwent.

Derwent GENESEQ FASTA/*ert* records are supplied in the familiar FASTA format.

By using Derwent GENESEQ in conjunction with the publicly available sequence databases, you can be assured that your sequence search covers all the published sequence material available world-wide.

# How does the information in Derwent GENESEQ compare with that in the Derwent World Patents Index?

Derwent World Patents Index (DWPI) is the most comprehensive, enhanced online patent database in the world, containing over 8 million separate inventions from 40 patent issuing authorities, across all technologies.

Derwent GENESEQ abstracts are written by bioinformatics experts and have a different emphasis to the corresponding DWPI abstract: Derwent GENESEQ abstracts focus on the bioinformatics and IP aspects of the sequence in the patent whilst the DWPI abstracts focus on the legal and IP aspects of the whole invention.

Derwent GENESEQ records display the sequence which can be searched using the appropriate searching software: DWPI does not generally display sequences.

Derwent GENESEQ records state the position in the patent where the sequence appears i.e. the claim or disclosure page number: DWPI does not state sequence positions.

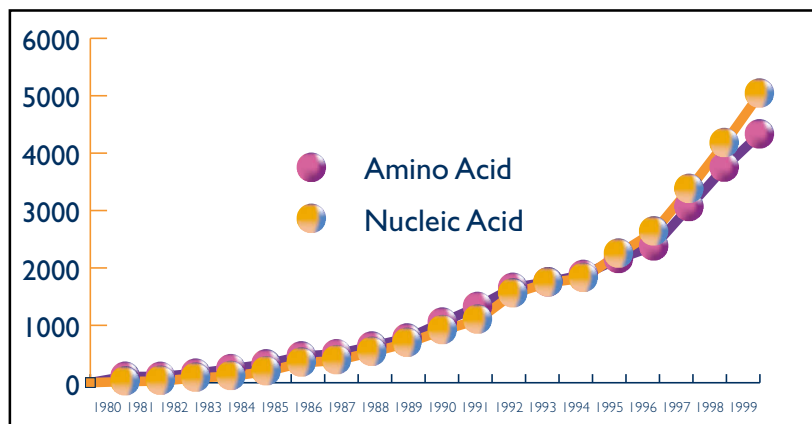
Derwent GENESEQ records contain Feature Tables which provide more information about specific features of the sequence: DWPI does not cover sequence features.

Derwent GENESEQ provides a new record for *each* sequence i.e. some patents contain many sequences and each sequence will be given its own record: DWPI provides a record for each new patent.

Both Derwent GENESEQ and DWPI records contain the same Enhanced Title, for easy identification and comparison of relevant search results.

Both Derwent GENESEQ and DWPI records (as well as other Derwent databases) contain the same Primary Accession Number to simplify searching for the same record in multiple databases.

# The growth of biomolecular sequences in patents



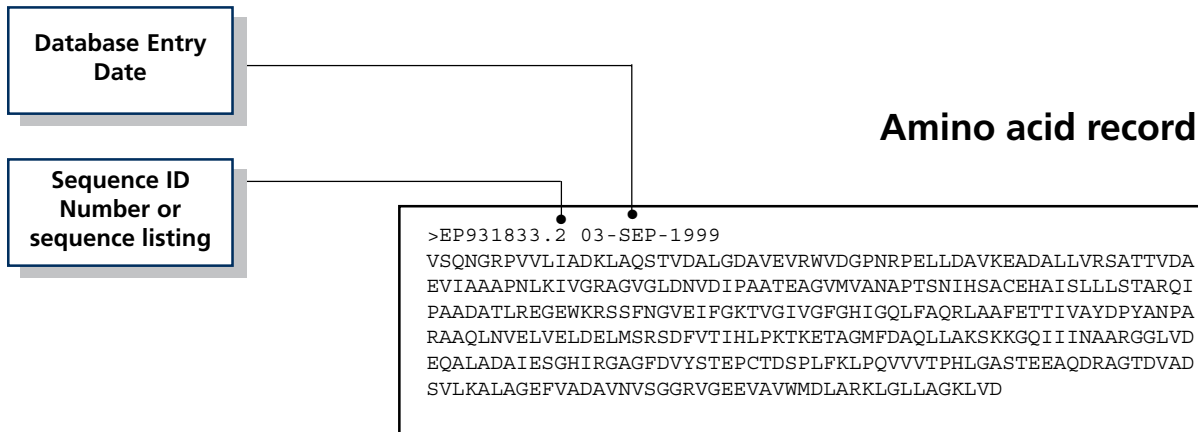
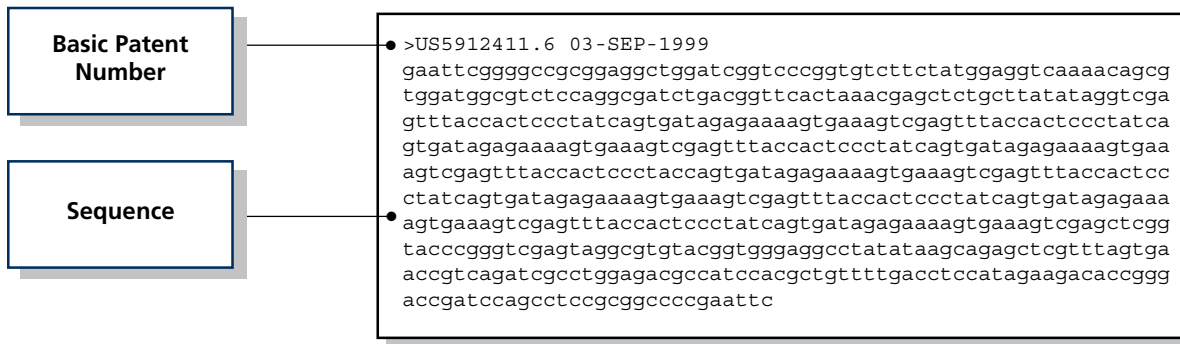
The number of patents containing sequences has grown rapidly in the last few years. The graph shows the volume of patents added to Derwent GENESSEQ over the period 1980-1999.

**Graph showing the number of patents, containing sequences, which are added to Derwent GENESSEQ each year**

# Sample Derwent GENESEQ FASTA/ert records

Derwent GENESEQ FASTA/ert flat file format  
is based on the FASTA format

## Nucleic acid record



# Sample Derwent GENESEQ flat file records

Derwent GENESEQ flat file format is based on the format of EMBL sequence databases.

## Nucleic acid record

```

ID X81839 standard; DNA; 1432 BP.
XX
AC X81839;
XX
DT 02-SEP-1999 (first entry)
XX
DE DNA encoding 3-PGDH protein, also known as serA.
XX
KW Coryneform bacterium; L-serine production; phosphoserine phosphatase; phosphoserine
KW transaminase; large-scale fermentation; 3-PGDH; serA; ss.
XX
OS Corynebacterium glutamicum.
XX
PN EP931833-A2.
XX
PD 28-JUL-1999.
XX
PF 12-JAN-1999; 99EP-0100325.
XX
PR 11-DEC-1998; 98JP-0353521.
PR 12-JAN-1998; 98JP-0003751.
XX
PA (AJIN ) AJINOMOTO CO INC.
XX
PI Hibino W, Ito M, Nakamatsu T, Osumi T, Suga M, Sugimoto M;
XX
DR WPI; 1999-397161/34.
DR P-PSDB; Y22645.
XX
PT New coryneform bacterial strain useful for producing L-serine
XX
PS Example 5; Page 13-15; 33pp; English.
XX
CC The specification describes a coryneform bacterium which is able to produce L-serine. The
CC activity of at least one phosphoserine phosphatase and phosphoserine transaminase is
CC enhanced in the organism. The organism is used for large-scale fermentation of L-serine
CC for amino acid mixtures which are used in pharmaceuticals, chemicals and cosmetics. The
CC present sequence encodes 3-PGDH protein, also known as serA, and is used in the course
CC of the invention.
XX
SQ Sequence 1432 BP; 274 A; 361 C; 416 G; 381 T; 0 other;
ggatccggac acacgtgaca aaattgtaga aaattggatg atttgtcac gcctgtctgg 60
tttagctctg gttcgggacg ggcgtggaat ggaggtagcg caccgagacc ttgaccgcgg 120
gcccgaacaag ccaaaagtcc ccaaaacaaa cccactctgc cggagacgtg aataaaattc 180
gcagctcatt ccatcagcgt aaacgcagct ttttgcatgg tgagacacct ttgggggtaa 240
atctcacagc atgaatctct gggttagatg actttctggg tgggggaggg ttagaatgt 300
ttctagtcgc acgcaaaaac ccggcgtgga cacgtctgca gccgacgcgg tcgtgcctgt 360
tgtaggcgga cattcctagt tttccagga gtaacttggt agccagaatg gccgtccggt 420
agtccctcgc gccgataaag ttgpcagtc cactgttgac gcgcttggag atgocagtaga 480
agtccgttgg gttgacggac ctaaccgccc agaactgctt gatgcagtta aggaagcgga 540
cgcactgctc gtgcgttctg ctaccactgt cgatgctgaa gtcacgcccg ctgccccctaa 600
cttgaagatc gtcggtcgtg ccggcgtggg cttggacaac gttgacatcc ctgctgccac 660
tgaagctggc gtcattggtt ctaacgcacc gacctctaac attcactctg cttgtgagca 720
cgcaatttct ttgctgctgt ctaactgctg ccagatccct gctgctgatg cgaagctggc 780
tgagggcgag tggaaagcgt cttcttcaa cgggtgggaa atttcggaa aaactgctgg 840
tatcgtcgtt tttggccaca ttggtcagtt gtttgcctcag cgtcttctg cgtttgagac 900
caccattgtt gttacgatac cttacgocaa ccctgctcgt gcaagtcagc tgaacgttga 960
gttggttgag ttggatgagc tgatgagcgc tctgacttt gtcaccattc accttctcaa 1020
gaccaaggaa actgctggca tgtttgatgc gcaagctcctt gctaagtcca agaagggcca 1080
gatcatcacc aacgctgctc gtggtggcct tgttgatgag caggctttgg ctgatgcgat 1140
tgagtcggtg cacattcgtg gcgctggttt cgatgtgtac tccaccgagc cttgcactga 1200
ttctccttgg ttcaagttgc ctcaggttgt tgtgactcct cacttgggtg cttctactga 1260
agaggtccag gatcgtgcgg gtaactgact tgcgtattct gtgctcaagg cgtggtctgg 1320
cgagttcgtg gcggatgctg tgaacgttcc cgggtgctgc gtgggcgaag aggttgcgtg 1380
gtggatggat ctggtctgca agcttggctt ccttctgctg aagcttctgc ac 1432
//
  
```

## Amino acid record

```

ID Y22647 standard; Protein; 530 AA.
XX
AC Y22647;
XX
DT 02-SEP-1999 (first entry)
XX
DE Mutant 3-PGDH protein, also known as serA.
XX
KW Coryneform bacterium; L-serine production; phosphoserine phosphatase;
KW phosphoserine transaminase; large-scale fermentation; 3-PGDH; serA.
XX
OS Brevibacterium flavum.
XX
PN EP931833-A2.
XX
PD 28-JUL-1999.
XX
PF 12-JAN-1999; 99EP-0100325.
XX
PR 11-DEC-1998; 98JP-0353521.
PR 12-JAN-1998; 98JP-0003751.
XX
PA (AJIN ) AJINOMOTO CO INC.
XX
PI Hibino W, Ito M, Nakamatsu T, Osumi T, Suga M, Sugimoto M;
XX
DR WPI; 1999-397161/34.
DR N-PSDB; X81849.
XX
PT New coryneform bacterial strain useful for producing L-serine
XX
PS Example 5; Page 25-27; 33pp; English.
XX
CC The specification describes a coryneform bacterium which is able to produce L-
CC serine. The activity of at least one phosphoserine phosphatase and phosphoserine
CC transaminase is enhanced in the organism. The organism is used for large-scale
CC fermentation of L-serine for amino acid mixtures which are used in
CC pharmaceuticals, chemicals and cosmetics. The present sequence represents a
CC mutant 3-PGDH protein, also known as serA, and is used in the course of the
CC invention.
XX
SQ Sequence 530 AA;
SQ 75 A; 26 R; 17 N; 31 D; 0 B; 2 C; 16 Q; 41 E; 0 Z; 49 G; 7 H;
SQ 29 I; 55 L; 20 K; 4 M; 14 F; 17 P; 33 S; 31 T; 3 W; 4 Y; 56 V;
SQ 0 Others;
vsgngrpvvl iadkqlagstv dalgdavevr wvdgpnrpel ldtvkeadal lvrsattvda 60
eviaaapnlk ivragvgld nvdipaatea gvmvanapts nihsacehai slllstarqi 120
paadatlreg ewkrssfngv eifgktvgiv gfghigqlfa qrlaafetti vaydpyanpa 180
raaqlnvelv eldelmsrsd fvtihlpkpk etagmfdaql lakskkgqii inaargglvd 240
equaladaies ghirgagfdv ystepctdsp lfklpqvvvt phlgasteea qdragtdvad 300
svlkalagef vadavnvsgg rvgekvaavm dlarklglla gklvdaapvs ieveargels 360
segvdalgls avrglfgsii eesvtfvnap riaaergldi svktnsesvt hrsvlqvkvi 420
tgsgsaatvv galtglerve kitringrgl dlraeglmlf lqytdapgal gtvgtklgaa 480
ginieaaalt qaekgdgavl ilrvesavse eleaeinael gatsfqvdlid 530

```

//

# Summary of Derwent GENESEQ flat file fields

Field	Full Name	Description
ID	Identifier	contains the type of molecule, (cDNA, mRNA, polypeptide, etc.), and its length
AC	Accession Number	each sequence is given its own unique number in the database
DT	Date	the date the record was added or corrected and re-entered into the database
DE	Description	a brief description of the sequence
KW	Keywords	keywords or phrases describing the sequence and patent content (not controlled language)
OS	Organism	the Latin genus and species of the organism from which the sequence was obtained
FH	Feature Header	a header for the Feature Table
FT	Feature Table	a table indicating the exact positions of known features of the sequence (which are given in the patent), i.e. coding region, promoter region, ribosomal binding site, etc. For nucleotides the format is the same as in EMBL, but for polypeptides the format has been borrowed from PIR. (See Appendix I for definitions of features and qualifiers)
PN	Patent Number	the publication number of the source patent
PD	Publication Date	the date of publication of the patent
PF	Patent Filing	the local patent filing date and number
PR	Priority	priority details of the patent
PA	Patent Assignee	the company or individual taking out the patent, includes the Derwent standard Company Code
PI	Patent Inventors	the inventors of the product, process, etc. being claimed in the patent

Field	Full Name	Description
DR	Database Cross-Reference	contains the accession number for Derwent World Patents Index plus the Derwent GENESEQ accession number for corresponding Derwent GENESEQ records
PT	Patent Title	contains the Derwent World Patents Index enhanced title for the patent
PS	Patent Sequence Location	contains the exact location of the sequence within the patent document, i.e. claim or disclosure (example/figure), the total number of pages, and the language of the patent document
CC	Comments	unique Derwent GENESEQ abstract which highlights important aspects of the sequence in relation to the invention
SQ	Sequence	the sequence in its entirety (which can be searched with the appropriate software), headed by a breakdown of its components. IUPAC single-letter amino acid and nucleotide codes used throughout
XX	Field Terminator	denotes the end of the current Derwent GENESEQ field
//	Record Terminator	denotes the end of the current Derwent GENESEQ record

# Sample Derwent GENESQ records on STN

The name of the Derwent GENESQ file on STN is DGENE.

## Nucleic acid record

```
AN      1999N-X81839 DNA          DGENE
TI      New coryneform bacterial strain useful for producing L-serine
IN      Hibino W; Ito M; Nakamatsu T; Osumi T; Suga M; Sugimoto M
PA      (AJIN)          AJINOMOTO CO INC
PI      EP--931833 A2 19990728          33p
AI      1999EP-0100325 19990112
PRAI    1998JP-0353521 19981211
        1998JP-0003751 19980112
PSL     Example 5; Page 13-15
DED     02 SEP 1999 (first entry)
DT      Patent
LA      English
OS      1999-397161 [34]
CR      P-PSDB: 1999P-Y22645
DESC    DNA encoding 3-PGDH polypeptide, also known as serA
KW      Coryneform bacterium; L-serine production; phosphoserine phosphatase; phosphoserine
        transaminase; large-scale fermentation; 3-PGDH; serA; ss
ORGN    Corynebacterium glutamicum
AB      The specification describes a coryneform bacterium which is able to produce L-serine.
        The activity of at least one phosphoserine phosphatase and phosphoserine transaminase
        is enhanced in the organism. The organism is used for large-scale fermentation of L-
        serine for amino acid mixtures which are used in pharmaceuticals, chemicals and
        cosmetics. The present sequence encodes 3-PGDH polypeptide, also known as serA, and
is
NA      used in the course of the invention
        274 A; 361 C; 416 G; 381 T;
SQL     1432
SEQ     1 ggatccggac acacgtgaca aaattgtaga aaattggatg atttgtcac
        51 gcctgtctgg ttagctctg gttcgggacg ggcgtggaat ggaggtagcg
        101 caccgagacc ttgaccgcg gcccgacaag ccaaaagtcc ccaaaacaaa
        151 cccacctcgc cggagacgtg aataaaattc gcagctcatt ccatcagcgt
        201 aaacgcagct ttttgcatgg tgagacacct ttgggggtaa atctcacagc
        251 atgaatcctt gggttagatg actttctggg tgggggaggg tttagaatgt
        301 ttctagtcgc acgccccaaa cggcgtgga cagctctgca gccgacgcgg
        351 tcgtgcctgt ttagggcgga cattcctagt tttccaagga gtaactttgt
        401 agccagaatg gccgtccggt agtccctcgc gccgataagc ttgcgcagtc
        451 cactgttgac gcgcttggag atgcagtaga agtccgttgg gttgacggac
        501 ctaaccgccc agaactgctt gatgcagtta aggaagcgga cgcactgctc
        551 gtgcgtttctg ctaccaactgt cgatgctgaa gtcctgcgcy ctgcccctaa
        601 cttgaagatc gtccgtctgt cggcgtgggg ctggacaacc gttgacatcc
        651 ctgctgccac tgaagctggc gtcctggttg ctaacgcacc gacctctaac
        701 atcaactctg cttgtgagca cgcaatttct ttgctgctgt ctaactgctg
        751 ccagatccct gctgctgatg cgacgctgcy tgagggcgag tggaaagcgt
        801 cttcttcoa cgggtggaa attttcggaa aaactgtcgg tatcgtcgtt
        851 tttggccaca ttggtcagtt gtttgcctag cgtcttgcgt cgtttgagac
        901 caccattggt gcttacgac cttacgccc cctgctcgt gcagctcagc
        951 tgaacgttga gttggttag ttggatgagc tgatgagccg tttgacttt
        1001 gtcaccatc accttctaa gaccaaggaa actgctggca tgttgatgc
        1051 gcagctcctt gctaagtcca agaaggcca gatcatcacc aacgctgctc
        1101 gttggtggcct tttgatgag caggcttttg ctgatgcgat tgagtccggt
        1151 cacattcgtg gcgctggttt cgatgtgtac tccaccgagc cttgactga
        1201 ttctccttga tcaagttag ctcaggttgt tgtgactcct cactggggtg
        1251 cttctactga agaggtcag gatcgtgctg gtactgacgt tgctgattct
        1301 gtgctcaagg cgtggtctg cgagttcgtg gcggatgctg tgaacgttcc
        1351 cgggtgctgc gttggcgaag aggttctgtg gtggatggat ctggctcgca
        1401 agcttggtct tcttctgctg aagcttctgc ac
```

## Amino acid record

```

AN      1999P-Y22645 Polypeptide          DGENE
TI      New coryneform bacterial strain useful for producing L-serine
IN      Hibino W; Ito M; Nakamatsu T; Osumi T; Suga M; Sugimoto M
PA      (AJIN)      AJINOMOTO CO INC
PI      EP-931833 A2 19990728          33p
AI      1999EP-0100325 19990112
PRAI    1998JP-0353521 19981211
        1998JP-0003751 19980112
PSL     Example 5; Page 15-16
DED     02 SEP 1999 (first entry)
DT      Patent
LA      English
OS      1999-397161 [34]
CR      N-PSDB: 1999N-X81839
DESC    3-PGDH polypeptide, also known as serA
KW      Coryneform bacterium; L-serine production; phosphoserine phosphatase; phosphoserine
        transaminase; large-scale fermentation; 3-PGDH; serA
ORGN    Corynebacterium glutamicum
AB      The specification describes a coryneform bacterium which is able to produce L-serine.
        The activity of at least one phosphoserine phosphatase and phosphoserine transaminase
        is enhanced in the organism. The organism is used for large-scale fermentation of L-
        serine for amino acid mixtures which are used in pharmaceuticals, chemicals and
        cosmetics. The present sequence represents 3-PGDH polypeptide, also known as serA,
and is
        used in the course of the invention
AA      52 A; 16 R; 11 N; 24 D; 0 B; 2 C; 11 Q; 22 E; 0 Z; 29 G; 6 H; 17 I; 34 L; 14 K;
        4 M; 10 F; 14 P; 18 S; 18 T; 3 W; 3 Y; 37 V;
SQL     345
SEQ     1 vsqngprpvl iadklaqstv dalgdavevr wvdgpnrpel ldavkeadal
        51 lvsattvda eviaaapnlk ivgragvgld nvdipaatea gvmvanapts
        101 nihsacehai s111starqi paadat1reg ewkrssfngv eifgktvgiv
        151 gfgihgqlfa qrlaafetti vaydpyanpa raaqlnvelv eldelmsrsd
        201 fvtihlpktk etagmfdaql laskkkgqii inaargglvd egaladaies
        251 ghirgagfdv ystepctdsp lfk1pqvvvt phlgasteea qdragtdvad
        301 svlkalagef vadavnvsqg rvqeevavwm dlarklqla gklvd

```

# Summary of Derwent GENESEQ fields on STN

Field	Full Name	Description
AN	Accession Number	each sequence is given its own unique number in the database
TI	Patent Title	contains the Derwent World Patents Index enhanced title for the patent
IN	Patent Inventors	the inventors of the product, process, etc. being claimed in the patent
PA	Patent Assignee	the company or individual taking out the patent, includes the Derwent standard Company Code
PI	Patent Information	patent number, publication date and the number of pages
AI	Application Information	local filing number, date of filing
PRAI	Priority Information	priority application number(s), date of filing
PSL	Patent Sequence	contains the exact location of the sequence Location within the patent document, i.e. claim or disclosure (example/figure), the total number of pages, and the language of the application
DED	Data Entry Date	the date the record was added or corrected and re-entered into the database
DT	Document Type	patent
LA	Language	language of original patent
OS	Other Source	contains the accession number for Derwent World Patent Index to enable cross-referencing into the corresponding DWPI record
CR	Cross Reference	the Derwent GENESEQ accession number for corresponding Derwent GENESEQ records
DESC	Description	a brief description of the sequence
KW	Keywords	keywords or phrases describing the sequence and patent content (not controlled language)
ORGN	Organism Name	the Latin genus and species of the organism from which the sequence was obtained

<b>Field</b>	<b>Full Name</b>	<b>Description</b>
AB	Abstract	unique Derwent GENESEQ, abstract which highlights important aspects of the sequence in relation to the invention
AA	Amino Acid	break down of amino acid sequence into individual residues with the number of times the residues appear in the sequence
NA	Nucleic Acid	break down of nucleic acid sequence into individual bases with the number of times the residues appear in the sequence
SQL	Sequence Length	length of the sequence
SEQ	Sequence	the sequence in its entirety (which can be searched with the appropriate software), headed by a breakdown of its components. Nucleotide sequences are in IUPAC format, polypeptide sequences are in one-letter format
FEAT	Features Table	a table indicating the exact positions of known features of the sequence (which are given in the patent), i.e. coding region, promoter region, ribosomal binding site, etc. (see Appendix I for definitions of features and qualifiers)

---

# Sequence searching on STN (GETSEQ and GETSIM)

Two run packages are available for sequence searching in DGENE.

- GETSEQ – for nucleic acid sequence and polypeptide sequence searching
- GETSIM – for similarity (homology) searching of nucleic acid and polypeptide sequences

## GETSEQ

Polypeptide and nucleic acid sequence information may be easily retrieved from the DGENE file using a variety of search fields available with the RUN GETSEQ package. The database may be searched for polypeptide sequences that exactly match the query, sequences in which family-equivalent substitution of the query amino acids occur, and/or exact answers plus sequences in which the query sequence is embedded. Similar types of searches may be carried out on nucleic acid sequences.

Ambiguity codes for nucleic acids and variability symbols for amino acids and nucleic acids are allowed. Gaps within amino acid and polypeptide subsequence queries may be specified. See page 37 for a sample search using GETSEQ. More advanced search functions are available with GETSIM.

## GETSIM

Similarity (homology) searching, based on the algorithm developed by Pearson and Lipman, is available for peptide and nucleotide sequences with the RUN GETSIM package. The module provides polypeptide and nucleic homology searching and translated polypeptide to nucleic acid homology searching. It retrieves sequences which include the exact or a similar sequence query and assigns a similarity score. A diagram is generated that shows the similarity between the retrieved sequences and the query. The answers, sorted by descending accession number, may be re-arranged by descending similarity score. The alignment between the retrieved polypeptide and nucleic acid sequence and the query sequence may be viewed with the display format ALIGN. See page 45 for a sample search using GETSIM.

The STN help message in DGENE, HELPQLIMITS, gives details on the maximum length of sequence queries which can be used with the RUN commands GETSIM and GETSEQ.

# Appendix I: Feature Keys used in Feature Tables in Derwent GENESQ

## Nucleic Acids

Nucleic acid sequences have a large choice of feature keys (see list which follows):

- each selected feature is tagged (e.g. tag a, tag b, tag c...)
- each selected feature can be qualified by one or more "qualifiers" which will be followed by additional information about the feature

e.g.

Feature	: repeat unit
Qualifier	: repeat type = DIRECT
Feature	: modified_base
Qualifier	: mod_base = i
Feature	: CDS
Qualifier	: product = human_insulin transl except = pos: 22..24, aa: Trp

Each selected feature can be further described by a "note". Notes are used when there is not an appropriate "qualifier" for a feature or when information given in a "qualifier" requires further explanation

e.g.

Feature	: modified_base
Note	: "5'-deoxy-5'-S-(4,4'-dimethoxy tritylthymidine)"

Feature	Definition
allele	a related individual or strain contains stable, alternative forms of the same gene which differs from the presented sequence at this location (and perhaps others)
attenuator	1) regions of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons; 2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription
CAAT_signal	CAAT box; part of a conserved sequence located about 75 bp upstream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus = GG(C or T)CAATCT[1,2]
CDS	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a polypeptide (location includes stop codon)
cellular	cellular genomic sequences that have recombined with a foreign sequence (from the organism specified in /ORGANISM qualifier)
conflict	independent determinations of the "same" sequence differ at this site or region
D_loop	displacement loop; a region within mitochondrial DNA in which a short length of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region; also used to describe the displacement of a region of one strand of duplex DNA by a single invader in the reaction catalysed by RecA polypeptide
enhancer	a cis-acting sequence that increases the utilisation of (some) eukaryotic promoters, and can function in either orientation in any location (upstream or downstream) relative to the promoter
exon	region of genome that codes for portion of spliced mRNA; may contain 5'UTR, all CDSs, and 3'UTR
GC_signal	GC box; a conserved GC-rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGGCGG
iDNA	intervening DNA; DNA which is eliminated through any of several kinds of recombination

---

Feature	Definition
insertion_seq	insertion sequence; IS; a small transposon that carries only the genes needed for its own transposition
intron	a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it
LTR	long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses
mat_peptide	mature peptide or polypeptide coding sequence; coding sequence for the mature or final peptide or polypeptide product following post-translational modification. The location does not include the stop codon (unlike the corresponding CDS)
misc_binding	site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other Binding key (primer_bind or polypeptide_bind)
misc_difference	feature sequence is different from that presented in the entry and cannot be described by any other Difference key (conflict, unsure, old_sequence, mutation, variation, allele, or modified_base)
misc_feature	region of biological interest which cannot be described by any other feature key; a new or rare feature
misc_recomb	site of any generalized, site-specific or replicative recombination event where there is a breakage and reunion of duplex DNA that cannot be described by other Recombination keys (cellular, iDNA, insertion_seq, transposon, provirus, and virion)
misc_RNA	any transcript or RNA product that cannot be defined by other RNA keys (prim_transcript, precursor_RNA, mRNA, 5'clip, 3'clip, 5'UTR, 3'UTR, exon, CDS, sig_peptide, transit_peptide, mat_peptide, intron, polyA_site, rRNA, tRNA, scRNA, and snRNA)
misc_signal	any region containing a signal controlling or altering gene function or expression that cannot be described by other Signal keys (promoter, CAAT_signal, TATA_signal, -35_signal, -10_signal, GC_signal, RBS, polyA_signal, enhancer, attenuator, terminator, and rep_origin)
misc_structure	any secondary or tertiary structure or conformation that cannot be described by other Structure keys (stem_loop and D-loop)

Feature	Definition
modified_base	the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)
mRNA	messenger RNA; includes 5' untranslated region (5'UTR), coding sequences (CDS, exon) and 3' untranslated region (3'UTR)
mutation	a related strain has an abrupt, inheritable change in the sequence at this location
old_sequence	the presented sequence revises a previous version of the sequence at this location
polyA_signal	recognition region necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA
polyA_site	site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation
precursor_RNA	any RNA species that is not yet the mature RNA product; may include 5' clipped region (5'clip), 5' untranslated region (5'UTR), coding sequences (CDS, exon), intervening sequences (intron), 3' untranslated region (3'UTR), and 3' clipped region (3'clip)
primer_bind	non-covalent primer binding site for initiation of replication, transcription, or reverse transcription
prim_transcript	primary (initial, unprocessed) transcript; includes 5' clipped region (5'clip), 5' untranslated region (5'UTR), coding sequences (CDS, exon) intervening sequences (intron), 3' untranslated region (3'UTR), and 3' clipped region (3'clip)
promoter	region on a DNA molecule involved in RNA polymerase binding to initiate transcription
protein_bind	non-covalent polypeptide binding site on nucleic acid
provirus	proviral sequence specified by organism
RBS	ribosome binding site
repeat_region	region of genome containing repeating units
repeat_unit	single repeat element

Feature	Definition
rep_origin	origin of replication; starting site for duplication of nucleic acid to give two identical copies
rRNA	mature ribosomal RNA; the RNA component of the ribonucleoprotein particle (ribosome) which assembles acids into polypeptides
satellite	many tandem repeats (identical or related) of a short basic repeating unit; many have a base composition or other property different from the genome average that allows them to be separated from the bulk (main band) genomic DNA
scRNA	small cytoplasmic RNA; any one of several small cytoplasmic RNA molecules present in the cytoplasm and (sometimes) nucleus of a eukaryote
sig_peptide	signal peptide coding sequence; coding sequence for an N-terminal domain of a secreted polypeptide; this domain is involved in attaching nascent polypeptide to the membrane; leader sequence
snRNA	small nuclear RNA; any one of many small RNA species confined to the nucleus; several of the snRNAs are involved in splicing or other RNA processing reactions
stem_loop	hairpin; a double-helical region formed by base-pairing between adjacent (inverted) complementary sequences in a single strand of RNA or DNA
TATA_signal	TATA box; Goldberg-Hogness box; a conserved AT-rich septamer found about 25 bp before the start point of each eukaryotic RNA polymerase II transcript unit which may be involved in positioning the enzyme for correct initiation; consensus=TATA(A or T)A(A or T)
terminator	sequence of DNA located either at the end of the transcript or adjacent to a promoter region that causes RNA polymerase to terminate transcription; may also be site of binding of repressor polypeptide
transit_peptide	transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar polypeptide; this domain is involved in post-translational import of the polypeptide into the organelle
transposon	transposable element; TN; a DNA sequence able to replicate and insert one copy at (or, without replication, to move itself to) a new location in the genome

Feature	Definition
tRNA	mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence
unsure	patentee is unsure of exact sequence in this region
variation	a related strain contains stable mutations from the same gene (e.g. RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)
virion	viral genomic sequence as it is encapsidated, as distinguished from its proviral form (integrated in a host cell's chromosome)
3'clip	3'-most region of a precursor transcript that is clipped off during processing
5'clip	5'-most region of a precursor transcript that is clipped off during processing
3'UTR	region near or at the 3' end of a mature transcript (usually following the stop codon) that is not translated into a polypeptide; trailer
5'UTR	region near or at the 5' end of a mature transcript (usually preceding the initiation codon) that is not translated into a polypeptide; leader
-10_signal	Pribnow box; a conserved region about 10 bp upstream of the start point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TAtAaT
-35_signal	a conserved hexamer about 35 bp upstream of the start of bacterial transcription units; consensus=TTGACa or TGTTGACA

## Amino Acids

Amino acid sequences have a limited choice of possible features (see list which follows):

- features are not tagged
- each selected feature can be qualified by a "label"

e.g.

Feature	: Misc-difference
Qualifier	: label = Ala, Val, Ile, Leu
Feature	: Domain
Qualifier	: label = transmembrane
Feature	: Region
Qualifier	: label = complementarity_ determining

Each selected feature can be further described by a "note". Notes are used when a label would be inappropriate or when information given as a label requires further explanation.

e.g.

Feature	: Modified-site
Note	: "cyclohexylalanine"
Feature	: Disulfide-bond
Note	: "optional"

Feature	Definition
Active-site	Denotes the residues constituting an active site of a polypeptide
Binding-site	Site of molecule- or ion- or non-covalent binding to a polypeptide
Cleavage-site	Polypeptide or peptide cleavage site
Cross-links	Bonds linking the sequence to an adjacent polypeptide chain
Disulfide-bond	Identifies residues connected by disulphide bonds
Domain	Distinct polypeptide functional regions
Duplication	Regions evolved by sequence duplication
Inhibitory-site	Residues constituting a polypeptide's inhibitory site
Misc-difference	Sequence differs from that presented in the sequence part of the record in a way that cannot be described using any other feature key  Note: Used mainly to indicate a generic site i.e. x in the sequence is not unknown and not simply a modified site, in which case the "Modified-site" term is used. If x is generic and modified, Misc-difference is used, x is defined as a label and the modification is defined using a note
Modified-site	Site of modified amino acid in polypeptide sequence
Peptide	Peptides derived from the displayed sequence
Polypeptide	Mature polypeptide derived from the displayed sequence
Region	Biologically significant polypeptide sequence regions not covered by other polypeptide keys
Thiolester-bond	Identifies residues connected by thiolester bonds

## Appendix II: IUPAC codes used for nucleic acid and amino acid sequences in Derwent GENESEQ and GENESEQ FASTA/ert

Permitted entries in a nucleic acid sequence are as follows:

A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
M	A or C
R	A or G
W	A or T/U
S	C or G
Y	C or T/U
K	G or T/U
V	A, C or G; not T/U
H	A, C or T/U; not G
D	A, G or T/U; not C
B	C, G or T/U; not A
N	Unknown or Other

The codes for non-standard (atypical) nucleic acids and amino acids are shown on the following page, and are the same as those proposed by the US Patent and Trademark Office (US Official Gazette May 16th 1989).

Atypical nucleic acids are represented in the sequence by the letters shown on the following page and the Feature Table will define the exact composition of the appropriate letter.

Permitted entries in an amino acid sequence are as follows:

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid
X	Xaa	Undetermined or atypical

IUPAC amino acid codes are used for:

- unmodified L-amino acids which have IUPAC codes
- the corresponding D-amino acids
- modified\* versions of the above amino acids 'X' is used for all other cases.

\*Modified amino acids include those which have:

- acylated or alkylated amino groups
- amidated or esterified C-terminal or side chain groups
- protected or otherwise derivatised side chain functional groups (see page 30 for specific examples)

## Atypical nucleic acids

Abbrev.	Base Description	Shown in sequence as
ac4c	4-acetylcytidine	c
chm5u	5-(carboxyhydroxymethyl)uridine	u
cm	2'-O-methylcytidine	c
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine	u
cmnm5u	5-carboxymethylaminomethyl-uridine	u
d	dihydrouridine	u
fm	2'-O-methylpseudouridine	u
gal q	beta,D-galactosylqueosine	n
gm	2'-O-methylguanosine	g
i	inosine	n
i6a	N6-isopentenyladenosine	a
m1a	1-methyladenosine	a
m1f	1-methylpseudouridine	u
m1g	1-methylguanosine	g
m1i	1-methylinosine	n
m22g	2,2-dimethylguanosine	g
m2a	2-methyladenosine	a
m2g	2-methylguanosine	g
m3c	3-methylcytidine	c
m5c	5-methylcytidine	c
m6a	N6-methyladenosine	a
m7g	7-methylguanosine	g
mam5u	5-methylaminomethyluridine	u
mam5s2u	5-methoxyaminomethyl-2-thiouridine	u
man q	beta,D-mannosylqueosine	n
mcm5s2u	5-methoxycarbonylmethyl-2-thiouridine	u
mcm5u	5-methoxycarbonylmethyluridine	u
mo5u	5-methoxyuridine	u
ms2i6a	2-methylthio-N6-isopentenyladenosine	a
ms2t6a	N-((9-beta-D-ribofuranosyl-2-methylthiopurin-6-yl)carbonyl)-threonine	t
mt6a	N-((9-beta-D-ribofuranosylpurin-6-yl)N-methylcarbonyl)threonine	t
mv	uridine-5-oxyacetic acid methyl ester	u
o5u	uridine-5-oxyacetic acid	u
osyw	wybutoxosine	n
p	pseudouridine	u
q	queosine	n
s2c	2-thiocytidine	c
s2t	5-methyl-2-thiouridine	u
s2u	2-thiouridine	u
s4u	4-thiouridine	u
t	5-methyluridine	u
t6a	N-((9-beta-D-ribofuranosylpurin-6-yl)carbonyl)threonine	t
tm	2'-O-methyl-5-methyluridine	u
um	2'-O-methyluridine	u
yw	wybutosine	n
x	3-(3-amino-3-carboxypropyl)-uridine, (acp3)u	u
OTHER		n

For OTHER, the feature will be labelled with the name of the base.

Atypical amino acids are represented as an "x" in the sequence and the Feature Table will define the exact composition of the "x" residue.

### Atypical amino acids

Aad	2-aminoadipic acid
bAad	3-aminoadipic acid
bAla	beta alanine
Abu	2-aminobutyric acid
4Abu	4-aminobutyric acid
Acp	6-aminocaproic acid
Ahe	2-aminoheptanoic acid
Aib	2-aminoisobutyric acid
bAib	3-aminoisobutyric acid
Apm	2-aminopimelic acid
Dbu	2,4-diaminobutyric acid
Des	desmosine
Dpm	2,2'-diaminopimelic acid
Hyl	hydroxy lysine
Dpr	2,3-diaminopropionic acid
Ide	isodesmosine
aHyl	allo-hydroxy lysine
Nle	norleucine
alle	allo-isoleucine
Nva	norvaline
Orn	Ornithine

EtAsn	N-ethyl asparagine (represented as n in sequence)
4Hyp	4-hydroxy proline (represented as p in sequence)
Melle	N-methyl isoleucine (represented as i in sequence)
EtGly	N-ethyl glycine (represented as g in sequence)
3Hyp	3-hydroxy proline (represented as p in sequence)
MeGly	N-methyl glycine (represented as g in sequence)
MeLys	N-methyl lysine (represented as k in sequence)
pGlu	pyroglutamic acid (represented as e in sequence)

\* This same rule applies to D-configured amino acid

Any amino acid not in the above list is defined by a note, e.g.:

FT	Modified-site 23
FT	/note= "2-mercaptopimelic acid"

The following modified (atypical) amino acids are represented as the parent amino acid in the sequence, with the exact modification being shown in the Feature Table.\*

## Appendix III : Molecular type codes in GENESEQ

<b>Code</b>	<b>Definition</b>
DNA	includes unspecified DNA
RNA	includes unspecified RNA
mRNA	messenger RNA
rRNA	ribosomal RNA
tRNA	transfer RNA
scRNA	small cytoplasmic RNA
preRNA	precursor RNA
cDNA to mRNA	
cDNA to rRNA	
cDNA to tRNA	complementary DNA to specific types of RNA
cDNA to scRNA	
cDNA to snRNA	
cDNA	general or unspecified cDNA
polypeptide specification	as described in the specification
peptide specification	as described in the specification

# Appendix IV: Sample Searches in GENESEQ on STN (file DGENE)

## GETSEQ

Find references on how polymorphism is used to detect heart disease in humans – find both the DNA sequences and their related amino acid sequences, plus the patent family information from DWPI.

```
=> file dgene
FILE 'DGENE' ENTERED AT 12:03:26 ON 15 AUG 96
COPYRIGHT (C) 1996 DERWENT INFORMATION LTD

FILE LAST UPDATED: 11 AUG 96<960811/UP>

=> s polymorphi?
L1      4435 POLYMORPHI?

=> s l1 and (heart or coronary)
        2046 HEART
        751 CORONARY

L2      27 L1 AND (HEART OR CORONARY)
```

Once you have carried out your initial search, you may want to check to see if any of the records are particularly relevant. A good way to do this is to select the **trial** format, which is free for display on STN. This will display the Accession Number, Title, Descriptor, Keywords and Number of sequences.

```
=> d l2 tri 1

L2      ANSWER 1 OF 27  DGENE COPYRIGHT 1996 DERWENT INFORMATION LTD

AN      88P-P81655  polypeptide          DGENE
TI      ***Polymorphisms***  in human genes - used for genetic identification of
individuals and for predicting development  of diseases, esp. hypertension
DESC   Sequence encoded by probe "clone 6" used for detection of
***polymorphisms***  in the ANP gene
KW      Hypertension; hybridisation

SQL    152
```

If the record is one that you are interested in, you can then download all the fields

```
=> d l2 all 1

L2      ANSWER 1 OF 27  DGENE COPYRIGHT 1996 DERWENT INFORMATION LTD
AN      88P-P81655  polypeptide          DGENE
TI      ***Polymorphisms***  in human genes - used for genetic identification
        of individuals and for predicting development of diseases, esp.
        hypertension
IN      Frossard P M
PA      (BIOT-N)      Biotechn Res Ptnrs
PI      WO8808457  A   881103          42 pp
AI      88WO-US01402  880428
PRAI    87US-0093347  870904
        87US-0043437  870401
PSL     Example; Fig 3
DED     31 DEC 1990  (first entry)
DT      Patent
LA      English
OS      88-322776 [45]
CR      N-PSDB: 88N-N80193
DESC    Sequence encoded by probe "clone 6" used for detection of
        ***polymorphisms*** in the ANP gene
KW      Hypertension; hybridisation
ORGN    Homo sapiens
AB      ***Polymorphisms***  in the ANP gene and in the apoAI/apoCIII complex
        were shown to correlate with hypertension. Several sources for probes
        useful in the detection of ***polymorphisms*** of the ANP gene are
        known in the art. Probe designated "clone 6" was isolated from a human
        ***heart*** cDNA library in lambda-gt10 by probing with the rat
        ANP-encoding cDNA
AA      12 A; 10 R; 5 N; 9 D; 0 B; 1 C; 4 Q; 8 E; 0 Z; 15 G; 2 H; 4 I; 18 L; 5K;
        4 M; 8 F; 12 P; 20 S; 5 T; 2 W; 2 Y; 6 V; SQL      152
SEQ     1  mgsfsitkgf flflafwlpq higanpvysa vsntdlmdfk nlldhleekm
        51  pvedevmppq alseqtdeap aalsslsevp pwtgevnpqsq rdggalgrgp
        101 sdpsdrsall kskltallag prslrssssf ggridrigaq sglgcnsfry
        151  rr
```

You can then use the information found in the OS to search DGENE for all other records with the same cross reference. The OS (Other Source) is the DWPI record accession number for the patent in which the sequences appear.

```
=> s l2 <os 1>
*** SmartSELECT INITIATED ***
SET SMARTSELECT ON
SET COMMAND COMPLETED

SEL L2 OS 1
L3      SEL
L2 1 OS : 1 TERM

SET SMARTSELECT OFF
SET COMMAND COMPLETED

S L3

L4 2    L3
```

Download the full record for one of the hits

```
=> d l4 all 2

L4      ANSWER 2 OF 2  DGENE COPYRIGHT 1996 DERWENT INFORMATION LTD
AN      88N-N80193  DNA      DGENE
TI      Polymorphisms in human genes - used for genetic identification of
        individuals and for predicting development of diseases, esp.
        hypertension
IN      Frossard P M
PA      (BIOT-N)      Biotechn Res Pttrs
PI      WO8808457  A  881103      42 pp
AI      88WO-US01402  880428
PRAI    87US-0093347  870904
        87US-0043437  870430
PSL     Example; Fig 3
DED     31 DEC 1990  (first entry)
DT      Patent
LA      English
OS      88-322776 [45]
CR      P-PSDB: 88P-P81655
DESC    Sequence of probe "clone 6" used for detection of polymorphisms in the
        ANP gene
KW      Hypertension; hybridisation; ss
```

```

ORGN      Homo sapiens
AB        Polymorphisms in the ANP gene and in the apoAI/apoCIII complex were
          shown to correlate with hypertension. Several sources for probes
          useful in the detection of polymorphisms of the ANP gene are known in
          the art. Probe designated "clone 6" was isolated from a human heart
          cDNA library in lambda-gt10 by probing with the rat ANP-encoding cDNA
NA        191 A; 228 C; 210 G; 162 T;
SQL       791
SEQ
1 ggctgagaga gaaaccagag agtgagccga gacagcaaac atcagatcgt
51 gccccgacc acgccagcat gggctccttc tccatcacca agggcttctt
101 cctcttctcg gccttttggc tcccaggcca tattggagca aatcccgtat
151 acagtgcggt gtccaacaca gatctgatgg atttcaagaa cctgctagac
201 cacctggagg agaagatgcc ggtagaagat gaggtcatgc ctccgcaggc
251 cctgagcgag cagaccgatg aagcgccggc ggcacttagc tcctctctg
301 aggtgcctcc ctggactggg gaagtcaacc cgtctcagag agatggaggt
351 gctctcgggc gcggcccctc ggaccctcc gatagatctg ccctcttgaa
401 aagcaaaactg acggctctgc tcgctggccc tcggagcctg cgaagctcaa
451 gctcgttcgg gggtaggatt gacaggattg gagcccagag cggactaggc
501 tgcaacagct tccggtaccg aagataacag ccaaatctgc tcgagcagat
551 cgcaaaagat cccaagcctt gccgtgtgtc acacagcttg gtcgcattgc
601 cactgagagg tgggtgaatac cctcctggag ctgcagcttc ctgtcttcat
651 ctatcacgat cgatgttaag tgtagatgag tggtttagtg aggccttacc
701 tctcccactc tgcataataa ggtagatcct caccctttc agaaagcagt
751 tggaaaaaaa taaatccgaa taaacttcag caccacggac a

FEATURE TABLE:
Key      |Location |Qualifier |
=====+=====+=====+=====
CDS69..527          *tag= a |
polyA_signal      |759..764 |*tag= b |
polyA_signal      |769..774 |*tag= c |
polyA_site        |791      |*tag= d |

```

You can also download the DWPI patent family data directly in DGENE.

```

=> d 14 fam 2

L4      ANSWER 2 OF 2  DGENE COPYRIGHT 1996 DERWENT INFORMATION LTD
PI      WO8808457  A  881103 (8845)* EN  42 pp
        RW:AT BE CH DE FR GB IT LU NL SE
        W: AU DK JP US
        AU8817280  A  881202 (8908)
ADT     WO8808457  A  88WO-US01402 880428
PRAI    87US-0043437 870430; 87US-0093347 870904

```

**Search for references on the detection of adeno virus using probes. Carry out a sequence search\* using part of the probe sequence as the query search.**

\* Use the following types of polypeptide sequence searches: SQSP (exact answers plus embedded sequences); SQEP (exact answers only - not embedded sequences); SQEFP (exact answers plus sequences with family equivalent substitutions); SQSFP (exact answers plus embedded sequences plus family equivalent substitutions).

```
=> s probe?
L6          36013 PROBE?

=> d L6 all 1

L6          ANSWER 1 OF 36013  DGENE COPYRIGHT 1996 DERWENT INFORMATION LTD
AN          96P-R97603  Polypeptide          DGENE
TI          Detection and identification of adenovirus using serotype and sub-type
            specific oligo:nucleotide(s) - and ***probes*** and primers used in
            the method
PA          (MITP)          MITSUBISHI YUKA BCL KK
PI          JP07327700 A  951219          42 pp
AI          94JP-0126163  940608
PRAI       94JP-0126163  940608
PSL        Claim 43; Page 37-39
DED        25 JUL 1996  (first entry)
DT         Patent
LA         Japanese
OS         96-072347 [08]
CR         N-PSDB: 96N-T29174
DESC       Adenovirus 37 subtype D (Ad-37) polypeptide fragment
KW         Polymerase chain reaction; PCR; amplify; primer; ***probe***; detection;
            identification; adenovirus; exon region; serotype; subtype
ORGN       Mastadenovirus
AB         The sequences given in R97594-603 represent polypeptide fragments
            derived
            from different subtypes of adenovirus. These sequences are encoded by
            DNA fragments which were isolated using the primer and ***probe***
            sequences given in T10167-80. These primers and ***probes*** are used in
            the detection and identification of adenovirus. An adenoviral exon
            region is amplified which has a serotype and subtype specific sequence
            and then the amplified fragment is detected. The primers used for the
            amplification are complementary to sequence which are serotype specific.
            The method allows reliable, rapid and easy detection and identification
            of the subtype and serotype of adenovirus
AA         25 A; 20 R; 28 N; 20 D; 0 B; 2 C; 12 Q; 12 E; 0 Z; 22 G; 11 H; 13 I;
            31 L; 14 K; 16 M; 30 F; 28 P; 29 S; 22 T; 6 W; 22 Y; 25 V; 33 Others;
```

```

SQL      421
SEQ
  1  npfnhhrnaa lryrsmllgn gryvpfhiqv pqkffaiknl lllpgsytye
 51  wnfatxvnmj lqssxgndlr vdgasvrfds vnlyxtfffx ahxtsxxex
101  mlrndtxdxs xhxxlsaam lypipakatn vpsipsrnw aafrrwsftr
151  lkktetpslg sgfdpyfvys gsipyldgtf ylnhtfkkvs imfdsxvswp
201  gndxlltpne xeikrxvdge gxnvaxcmt kdwflxqmxs xynxgyxgfh
251  vpeaykdrmy sffrnfpms  rqvvdainyk dykavtlpfq hnnsgftgyl
301  aptmrggppy panfpyplig stavpsvtgk kflcdrvmwr ipfssnfmxm
351  galtdlgqnm lyansahald mtfevdpnde ptxyyllfev fdvvrvhqph
401  pgxieavylr tpfagnaxx x

FEATURE TABLE:
Key          |Location|Qualifier|
=====+=====+=====+=====
Misc_difference |56      |label    |Tyr, His, Asn, Asp
| |note    |"encoded by codon NAT"
Misc_difference |65      |label    |Ile, Phe, Val, Leu
| |note    |"encoded by codon NTC"
Misc_difference |85      |label    |Ser, Thr, Ala, Pro
| |note    |"encoded by codon NCC"
Misc_difference |89      |label    |Leu, pro, His, Gln, Arg
| |note    |"encoded by codon CNN"
Misc_difference |90      |label    |Leu, Met, Val
| |note    |"encoded by codon NTG"
Misc_difference |93      |label    |Ile, Thr, Asn, Ser
| |note    |"encoded by codon ANC"
Misc_difference |95      |label    |Val, Ala, Asp, Glu, Gly
| |note    |"encoded by codon GNN"
Misc_difference |97      |label    |Ser, Thr, Ala, Pro
| |note    |"encoded by codon NCC"
Misc_difference |98      |label    |Met, Leu, Val
| |note    |"encoded by codon NTG"
Misc_difference |100     |label    |Ser, pro, Thr, Ala
| |note    |"encoded by codon NCN"
Misc_difference |107     |label    |Asp, Tyr, Asn, His
| |note    |"encoded by codon NAC"
Misc_difference |109     |label    |Gln, Lys, Glu, STOP
| |note    |"encoded by codon NAG"
Misc_difference |111     |label    |Phe, Leu
| |note    |"encoded by codon TTN"
Misc_difference |113     |label    |Asp, Val, Gly, Ala
| |note    |"encoded by codon GNC"
Misc_difference |114     |label    |Phe, leu, Ser, Tyr, Cys, Trp, STOP
| |note    |"encoded by codon TNN"
Misc_difference |196     |label    |Leu, Ser, Trp, STOP
| |note    |"encoded by codon TNG"
Misc_difference |204     |label    |Trp Gly, Arg
| |note    |"encoded by codon NGG"

```

Misc_difference	211	label	Phe, Leu
	note	"encoded by codon	TTN"
Misc_difference	216	label	Ile, Met, Thr, Asn, Lys, Ser, Arg
	note	"encoded by codon	ANN"
Misc_difference	222	label	Asp, Tyr, Asn, His
	note	"encoded by codon	NAC"
Misc_difference	226	label	Gln, Lys, Glu, STOP
	note	"encoded by codon	NAA"
Misc_difference	236	label	Phe, Leu, Ile, Met, Val
	note	"encoded by codon	NTN"
Misc_difference	239	label	Ile, Met, Val, Phe, leu
	note	"encoded by codon	NTN"
Misc_difference	241	label	Ile, Thr, Asn, Ser
	note	"encoded by codon	ANC"
Misc_difference	244	label	Met, Ile
	note	"encoded by codon	ATN"
Misc_difference	247	label	Lys, Glu, Gln, STOP
	note	"any amino acid, encoded by codon	NNN"
Misc_difference	349	label	Gly, Pro, Arg, Ala
	note	"encoded by codon	SCC"
Misc_difference	383	label	Phe, Leu, Ile, Met, Val
	note	"encoded by codon	NTN"
Misc_difference	384	label	Phe, Leu, Ile, Met, Val
	note	"encoded by codon	NTN"
Misc_difference	403	label	Asp, Tyr, Asn, His
	note	"encoded by codon	NAC"
Misc_difference	419	label	Ile, Met, Thr, Asn, Lys, Ser, Arg
	note	"encoded by codon	ANN"
Misc_difference	420	note	"unspecified amino acid"
Misc_difference	421	note	"unspecified amino acid"

You can then carry out a sequence search using part of the probe sequence as the query search.

```
=> query npfnhhrnag
L7  QUE  NPFNHHRNAG
```

Search for exact answers plus embedded sequences.

```
=> run getseq
PLEASE ENTER SEQUENCE QUERY :17
TYPE OF SEARCH ? (SQSP):.
L8  RUN STATEMENT CREATED
L8  3      NPFNHHRNAG  /SQSP
```

Search for exact answers only/

```
=> run getseq l7
TYPE OF SEARCH ? (SQSP):sqep
L9  RUN STATEMENT CREATED
L9  0      NPFNHHRNAG /SQEP
```

Search for exact answers plus sequences with family equivalent substitutions.

```
=> run getseq l7
TYPE OF SEARCH ? (SQSP):sqefp
L10 RUN STATEMENT CREATED
L10 0      NPFNHHRNAG /SQEFP
```

Search for exact answers plus embedded sequences plus family equivalent substitutions.

```
=> run getseq npfnhhrnag/sqsfp
L11 RUN STATEMENT CREATED
L11 6      NPFNHHRNAG/SQSFP
```

You can then download one of the sequences found.

```
=> d L11 seq 1

L11ANSWER 1 OF 6  DGENE COPYRIGHT 1996 DERWENT INFORMATION LTD
SEQ
    1 npfnhhrnaa lryrsmllgn gryvpfhiqv pqkffaiknl lllpgsytye
      =====
    51 wnfatxvnmi lqssxgndlr vdgasvrfds vnlyxtffxx ahxtsxxex
   101 mlrndtxdxs xhxxlsaam lypipakatn vpsipsrnw aafrrwsftr
   151 lktketpslg sgfdpyfvys gsipyldgtf ylnhtfkkvs imfdsxvswp
   201 gndxlltpne xeikrxvdge gxnvaxcmt kdwflxqmxs xynxgyxgfh
   251 vpeaykdrmy sffrnfqpm s rqqvdeinyk dykavtlpfq hnnsgftgyl
   301 aptmrqqgpy panfpyplig stavpsvtqk kflcdrvmwr ipfssnfmxm
   351 galtdlgqnm lyansahald mtfevdpmde ptxyllfev fdvrvrhqph
   401 pgxieavylr tpfsagnaxx x
HITS AT:  1-10
```



```
L2  RUN STATEMENT CREATED
L2  600 HSDGIFTDSYSRYRQMAVKKYLAAVLGKRYKQRVK/SQP
Answer set arranged by accession number; to sort by descending similarity score,
enter at an arrow prompt (=>) "sor score d".
=> sort score d
ENTER (L2), L#, OR L# RANGE:l2
SORT ENTIRE ANSWER SET? (Y)/N:y
PROCESSING COMPLETED FOR L2
L3  600 SORT L2 SCORE D

=> d bib seq align 1 271

L3  ANSWER 1 OF 600  DGENE COPYRIGHT 1997 DERWENT INFORMATION LTD
AN  1996P-R97718  peptide          DGENE
TI  Stimulants of central nervous system contg. PACAP - useful e.g.
    as appetite control agent, for treating amnesia or dementia
    associated with Pick's disease or Huntington's chorea, etc
PA  (TAKE)  TAKEDA CHEM IND LTD
PI  JP 08092120 A  960409          23 pp
AI  1994JP-0229137  19940926
PRAI 1994JP-0229137  19940926
DT  Patent
LA  Japanese
OS  1996-235987 [24]
SEQ
    1 hsdgiftdsy sryrkqmvk kylaavlgkr ykqrvk
ALIGN Smith-Waterman score: 232
    36 aa overlap starting at 1
    hsdgiftdsysryrkqmvkkylaavlgkrykqrvk
    ::::::::::::::::::::::::::::::::::::::::::::
    hsdgiftdsysryrkqmvkkylaavlgkrykqrvk
```

```

L3 ANSWER 271 OF 600 DGENE COPYRIGHT 1997 DERWENT INFORMATION LTD
AN 1996P-R91063 Polypeptide DGENE
TI DNA encoding turkey (prepro) vasoactive intestinal peptide -
   useful to study the role of turkey VIP in prolactin regulation
IN El Halawani M E
PA (MINU) UNIV MINNESOTA
PI WO 9605310 A1 19960222 51 pp
AI 1995WO-US10075 19950809
PRAI 1995US-0437612 19950509
     1994US-0287668 19940809
DT Patent
LA English
OS 1996-139705 [14]
SEQ
 1 mehrgtsp11 lalallsalc wraralpprg aafpavprlg nrlpfdaase
51 sdrahgslks esdilqntlp enekfyfdls riidssqdsp vkrhsdavft
101 dnysrfrkqm avkkylnsvl tgkrsqeeln paklrdeaei lepsfsenyd
151 dvsvdellsh lpldl
ALIGN Smith-Weatherman score: 157
     37 aa overlap starting at 94
     hsdgiftsysryrkqmvkkyllaavl_gkrykqrvk
     :.....:
     hsdavftdnysrfrkqmvkkylnsvltgkrsqeeln

```

## Appendix V: Countries Covered by Derwent GENESEQ and GENESEQ FASTA/ert with start date for GENESEQ\*

Country	Start Date (Derwent Week)	Country	Start Date (Derwent Week)
Argentina	1975 only	Luxembourg	1984 (198443)
Australia	1963 - 1969, 1983 (198301)	Mexico	1998 (pending)
Austria	1975 (197515)	Netherlands	1963
Belgium	1963	Norway	1974 (197448)
Brazil	1976 (197601)	New Zealand	1993 (199301)
Canada	1963	PCT (World)	
China	1987 (198701)	Patents	1978 (197849)
Czechoslovakia	1975 (197520) - 1994	Philippines	1995 (199511)
Czech Republic	1994 (199417)	Portugal	1974 (197452)
Denmark	1974 (197445)	Romania	1975 (197532)
European Patents	1978 (197849)	Russian Federation	1994 (199406)
Finland	1974 (197445)	Singapore	1995 (199513)
France	1963	South Africa	1963
Germany (Democartic Republic)	1963	Soviet Union	- 1994
Germany (Federal Republic)	1963	Slovakia	1994 (199417)
Germany - Utility Models	1996 (199626)	Spain	1983 (198334)
Hungary	1975 (197526)	Sweden	1963
Ireland	1963-1969, 1995 (199521)	Switzerland	1963
Israel	1975 (197515)	Taiwain	1993 (199324)
Italy	1966 - 1969 Sect. A, 1978 (197801)	United Kingdom	1963
Japan	1963	United States	1963
Republic of Korea (South Korea)	1986 (198640)	PLUS:	
		Research Disclosure	1978 (197809)
		International Technology	
		Disclosure	1984 (198408) - 1993 (199351)

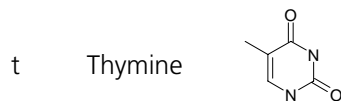
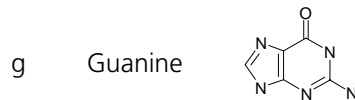
\*Sequences in GENESEQ FASTA/ert remain in the database until they appear in GENESEQ, when they are removed from the file so have no start date.

# Appendix VI: Genetic Code Table

FIRST POSITION (5' END)	SECOND POSITION				THIRD POSITION (3' END)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gin	Arg	A
	Leu	Pro	Gin	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Vale	Ala	Glu	Gly	G

# Appendix VII: Nucleic Acid and Amino Acid Structures

## Nucleic Acid Structures



m a or c

r a or g

w a or t/u

s c or g

y c or t/u

k g or t/u

v a, c or g; not t/u

h a, c or t/u; not g

d a, g or t/u; not c

b c, g or t/u; not a

n a, c, g or t/u;  
Unknown or Other

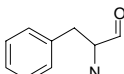
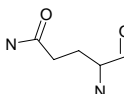

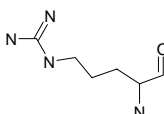
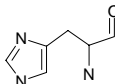

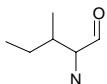

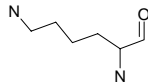

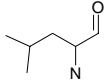
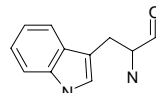
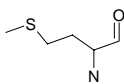
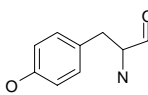
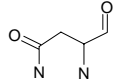

## Amino Acid Structures

A Alanine 

C Cysteine 

D Aspartic acid 

E Glutamic acid 

F	Phenylalanine		Q	Glutamine	
G	Glycine		R	Arginine	
H	Histidine		S	Serine	
I	Isoleucine		T	Threonine	
K	Lysine		V	Valine	
L	Leucine		W	Tryptophan	
M	Methionine		Y	Tyrosine	
N	Asparagine		B	Aspartic acid or Asparagine	
P	Proline		Z	Glutamine or Glutamic acid	
			X	Undetermined or atypical	

# Frequently Asked Questions

## What does Derwent GENESEQ contain and when did coverage start?

Derwent GENESEQ contains sequence information taken from patent applications and granted basic patents (all nucleic acids which are 10 or more bases in length, amino acids which are 4 or more residues in length plus all PCR primers and probes, of any length, are included in the database).

Sequences from patents from 1981 onwards are included in Derwent GENESEQ.

## How does Derwent GENESEQ on STN (DGENE) differ from Derwent GENESEQ flat file or in GCG format?

The content of the Derwent GENESEQ database is **identical** on all platforms and with all the various software packages. The only differing characteristics are the software packages which are used to search the Derwent GENESEQ database. However, some of the formats are different. In particular, sequence Accession numbers in DGENE are preceded with YYYYN – for nucleotides, and YYYYYP – for polypeptides/

peptides, where YYYY is the year of the WPIL accession number for the patent.

## What are the main differences between the Derwent GENESEQ software packages?

a Derwent GENESEQ on STN uses the standard STN command language to search the entire record apart from the SEQ field i.e. the sequence. To search the sequence part of the record, the run GETSEQ and run GETSIM packages are used (see page 18). GETSEQ and GETSIM do not have to be purchased as separate software packages.

GETSEQ will allow the following types of **string** searches:

Exact polypeptide searches, exact polypeptide searches with family-equivalent substitutions, subsequence polypeptide searches, subsequence polypeptide searches with family-equivalent substitutions, exact nucleic acid searches and subsequence nucleic acid searches.

Subsequence searches will allow the use of variability symbols (e.g. repeat preceding sequence m times etc.) and specified gaps.

GETSIM performs similarity (homologous) sequence searches, and retrieves polypeptide

sequences which include the exact or similar sequence query and assigns a similarity score.

b Derwent GENESEQ in GCG format is available for use with the Wisconsin Package produced by GCG (Genetics Computer Group). Please contact Derwent for more details.

### What are the differences between Derwent GENESEQ and other sequence databases?

The major difference is that Derwent GENESEQ only covers sequences which appear in patents. Every patent is intellectually scanned and processed, which means that there is no unnecessary duplication in the database. All records have the same sort of content and format and only the most relevant information is provided. The patent issuing authorities coverage is greater on Derwent GENESEQ than on other databases which cover patent information: Derwent GENESEQ covers 40 patent issuing authorities.

Derwent GENESEQ concentrates on both the bioinformatic and the IP aspect of the sequence whereas the other databases tend

to concentrate just on the bioinformatic aspects.

### How can I search for D-amino acid residues?

D-amino acid residues are represented as the parent amino acid in the sequence and the Feature Table will show the position of any D-modified amino acids (annotated by "Misc-difference").

### What are the selection rules for Derwent GENESEQ and GENESEQ FASTA/ert?

All sequences appearing in the patent claims section are included in Derwent GENESEQ and GENESEQ FASTA/ert.

All other sequences not specifically stated to be known or to have been published are included.

Specific examples of generically claimed peptides are included.

**Note:** Derwent GENESEQ and GENESEQ FASTA/ert only cover basic patents. A basic patent is one in which Derwent has seen the details of this invention published for the first time.

## **How can I search cyclic peptides in Derwent GENESEQ?**

All cyclic peptides have the word "cyclic" appearing in the Keyword field.

Cyclic peptides can only be searched by carrying out a series of searches with each amino acid residue acting as the first residue in the query sequence.

## **Does Derwent GENESEQ have single or double stranded DNA?**

Derwent GENESEQ covers both single-stranded and double-stranded DNA. The two types are distinguished by the terms "ss" and "ds" in the keyword field. However, in the case of double-stranded DNA, only the 5' to 3' strand is inputted. This follows the USPTO requirements for all nucleotide sequences in the sequence listing section, which have to be presented on diskette, as single stranded, 5' to 3', left to right (for more information see "An overview of the PTO sequence rules". The Law Works, February 1996, page 4).

# Customer Support

## Derwent Help Desk Assistance

Expert advice and support is available via our Derwent Help Desk staff, to provide a fast and efficient response to all your enquiries. The experienced Help Desk staff have an in-depth knowledge of all Derwent's products and services and are familiar with the command languages of the various online hosts

In addition to your online and other product queries, the staff will assist with delivery and invoice enquiries, take details of amendments to contact address details, and action correction of any reported errors. In fact, from general customer queries through to technical questions on chemical or polymer indexing or how to set up your internet browser, the Help Desk is there to help you.

Contact your local Help Desk by phone, fax or e-mail or visit the Customer Services area on the Derwent World Wide Web Site.

## Universal Freephone

The European Help Desk has recently launched a Universal Freephone Number

**+800 33 44 2999**

At Present this is available in the following countries

---

Belgium	- 00800 33 44 2999
Denmark	- 00800 33 44 2999
France	- 00800 33 44 2999
Germany	- 00800 33 44 2999
Norway	- 001800 33 44 2999
Sweden	- 009800 33 44 2999
Switzerland	- 00800 33 44 2999
The Netherlands	- 00800 33 44 2999
UK*	- 00800 33 44 2999

---

Additional countries will be made available as they join the Universal Freephone system

(\* UK users should note that it is necessary to dial the international prefix '00')

### **Europe/International**

Tel: +44 171 344 2999

Fax: +44 171 344 2900

Email: [custserv@derwent.co.uk](mailto:custserv@derwent.co.uk)

### **Universal Freephone**

00800 33 44 2999 UK

+800 33 44 2999 International

### **North America**

Tel: +1 800 336 5010

Fax: +1 800 457 0850

Email: [custserv@derwentus.com](mailto:custserv@derwentus.com)  
[search@Derwentus.com](mailto:search@Derwentus.com)  
[docdel@derwentus.com](mailto:docdel@derwentus.com)

### **Japan**

Tel: +81 3 5218 6500

Fax: +81 3 5222 1280

### **Derwent Word Wide Web Site**

<http://www.derwent.com/>

<http://www.derwent.co.jp/>

### **Online Training**

Specialist staff run regular training programmes to help you search our online databases effectively and comprehensively.

Training classes are held in major towns and cities in Europe, North America and Japan.

Training sessions can be found at <http://www.derwent.com/training>

A wide range of classes are available introducing Derwent files to both beginners and experienced searchers. Subject specialist classes are also available providing specialist training on Derwent's indexing. We can also develop training classes or presentations tailored to your company's specific needs.

### **Derwent International Patent Copy Service**

Having completed your online search you can order quality copies of patent documents issued around the globe. As holders of the world's largest private collection of international patents, Derwent provides a fast and efficient service. In addition, through a global network of contacts, Derwent regularly locates and supplies old and unusual patents.

The complete file history of a patent can also be supplied. This detailed document enables you to track the entire life of a patent from application through amendments to grant (if this occurred).