



DERWENT  
GENESEQ FASTA*Alert*

**User guide**

© 2000 Thomson. All rights reserved

Edition 1  
ISBN: 0 901157 67 8

© 2000 Thomson  
Published by Thomson Scientific  
14 Great Queen Street, London, WC2B 5DF,  
United Kingdom

Visit the Thomson Scientific web site at [www.thomson.com/scientific](http://www.thomson.com/scientific)

Edition 1 published March 2000

ISBN: 0 901157 67 8 (Edition 1)

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

Derwent GENESEQ FASTAlert File Content Description .....	1
Data Specification .....	2
IUPAC Definitions .....	4
Sample Records .....	6
Contact Details .....	8
Appendix .....	9
Index .....	15
Notes .....	17



# Derwent GENESEQ FASTAlert File Content Description

Derwent GENESEQ FASTAlert is a set of two sequence files in FASTA format. **FastAlert\_N.txt** contains nucleic acid sequences and **FastAlert\_P.txt** contains protein or peptide sequences. The sequences are taken from patents including claims, examples, disclosures and Japanese ID listings and they represent the preliminary entries for subsequent issue in fully annotated form in Derwent GENESEQ.

Both files are sequential listings, by patent number, made up of a series of discrete sequences delineated by an annotation line between each sequence.

# Data Specification

## Annotation Lines

Each annotation line has the format:

```
>CCNNNNNNN.SID DD-MMM-YYYYY
```

where:

CC	Country Code of the patent from which the sequence has been taken
NNNNNNN	Serial of the patent (patent number). The number of digits varies from country to country
SID	represents the SEQ ID NO associated with the sequence as shown in the patent, or an arbitrary SEQ ID NO if none is given in the patent. This number can be further qualified by the suffix 'a', '_dupN' or '_dupNa', where: <ul style="list-style-type: none"> <li>SIDa = the sequence ID number of a protein sequence which has been separated out from within a nucleotide sequence SID</li> <li>SID_dupN = represents the N<sup>th</sup> duplicate of SID when this SEQ ID number has been erroneously used more than once to refer to different sequences in the same patent</li> <li>SID_dupNa = the SEQ ID number of a protein sequence which has been separated out from within the nucleotide sequence SID_dupN</li> </ul>

DD-MMM-YYYYY represents a system date to identify the update uniquely.

## Nucleic acid sequences

These are represented as lower case letters, 60 per line, no spaces. The following IUPAC sequence letters are valid:

**a, b, c, d, g, h, k, m, n, r, s, t, u, v, w, y**

## Protein/Peptide sequences

These are represented as capital letters, 60 per line, no spaces. The following IUPAC sequence letters are valid:

**A, B, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, X, Y, Z**

# IUPAC Definitions

The IUPAC system letters are defined as follows:

## 1 Nucleic acid bases

<b>IUPAC</b>	<b>Definition</b>	<b>IUPAC</b>	<b>Definition</b>
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	a, c, g or t/u; Unknown or
Other			

## 2 Amino acids

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

# Sample Records

## Nucleic Acid Sequences

>EP931833.1 03-SEP-1999

```
ggatccggacacacgtgacaaaattgtagaaaattggatgattttgtcacgcctgtctgg
tttagctctgggttcgggacgggctggaatggaggtagcgcaccgagaccttgaccgcg
gcccagacaagccaaaagtccccaaaaaaaaccacctcgccggagacgtgaataaaattc
gcagctcattccatcagcgtaaacgcagctttttgcatggtgagacacctttgggggtaa
atctcacagcatgaatctctggggttagatgactttctgggtgggggagggttagaatgt
ttctagtcgcacgccaaaaccggcgtggacacgtctgcagccgacgcggctcgtgcctgt
tgtaggcggacattcctagttttccaggagtaacttgtgagccagaatggcgcctcgggt
agtccctcatcgccgataagcttgcgcagctccactgttgacgcgccttggagatgcagtaga
agtcggttgggttgacggacctaaccgcccagaactgcttgatgcagttaaggaagcggga
cgcactgctcgtgcggttctgctaccactgtcgcgatgctgaagtcatcgccgctgccctaa
cttgaagatcgtcggctcgtgccggcgtgggcttggacaacgttgacatccctgctgccac
tgaagctggcgtcatgggtgctaacgcaccgaccttaacattcactctgcttgtgagca
cgcaatcttcttggctgctgctactgctcgccagatccctgctgctgatgcgacgctgcg
tgaggggcagtggaagcggctcttcttcaacgggttggaatcttccggaaaaactgctcgg
tatcgtcgggtttggccacattgggtcagttggttctcagcgtcttgctgcgcttgagac
caccattggtgcttacgatccttacgccaacctgctcgtgcagctcagctgaacgttga
gttggttgagttggatgagctgatgagccggttctgactttgtcaccattcaccttcttaa
gaccaaggaaactgctgcatggttgatgcgacgctccttgctaagtccaagaaggcca
gatcatcatcaacgctgctcgtgggtggccttgttgatgagcaggcttggctgatgcgat
tgagtcgggtcacattcgtggcgctgggttccgatgtgtactccaccgagccttgcactga
ttctccttggttcaagttgcctcagggttgggtgactcctcacttgggtgcttctactga
agaggctcaggatcgtgcccgtactgacgttgcctgattctgtgctcaaggcgtggctgg
cgagttcgtggcggatgctgtgaacgtttccggtggctcgcgtgggcgaagagggttgcgtg
gtggatggatctggctcgcaagcttggctcttcttgcctggcaagcttgtcgcac
```

>EP931833.3 03-SEP-1999

```
ggacacacgtgacaaaattgtag
```

>EP931833.4 03-SEP-1999

```
gccagcaagaagaccaagcttgc
```

>EP931833.5 03-SEP-1999

```
gtacatattgtcgttagaacgcgtaatacgactca
```

## Protein or Peptide Sequences

>EP931833.2 03-SEP-1999

VSQNGRPVVLIIADKLAQSTVDALGDAVEVRWVDGPNRPELLDVAKEADALLVRSATTVDA  
EVIAAAPNLKIVGRAGVGLDNVDI PAATEAGVMVANAPTSNIHSACEHAISLLLSTARQI  
PAADATLREGGEWKRSSFNNGVEIFGKTVGIVGFGHIGQLFAQRLAAFETTIVAYDPYANPA  
RAAQLNVELVELDELMSRSDFTVIHLPKTKETAGMFDAQLLAKSKKQIIINAARGGLVD  
EQALADAIESGHIRGAGFDVYSTEPCTDSPLFKLPQVVVTPHLGASTEEAQDRAGTDVAD  
SVLKALAGEFVADAVNVSGGRVGEVAVWMDLARKLGLLAGKLV

>EP931833.12 03-SEP-1999

VSQNGRPVVLIIADKLAQSTVDALGDAVEVRWVDGPNRPELLDVTKEADALLVRSATTVDA  
EVIAAAPNLKIVGRAGVGLDNVDI PAATEAGVMVANAPTSNIHSACEHAISLLLSTARQI  
PAADATLREGGEWKRSSFNNGVEIFGKTVGIVGFGHIGQLFAQRLAAFETTIVAYDPYANPA  
RAAQLNVELVELDELMSRSDFTVIHLPKTKETAGMFDAQLLAKSKKQIIINAARGGLVD  
EQALADAIESGHIRGAGFDVYSTEPCTDSPLFKLPQVVVTPHLGASTEEAQDRAGTDVAD  
SVLKALAGEFVADAVNVSGGRVGEVAVWMDLARKLGLLAGKLVDAAPVSVIEVEARGELS  
SEQVDALGLSAVRGLFSGII EESVTFVNAPRIAEERGLDISVKTNSESVTHRSVLQVKVI  
TGSGASATVVGALTGLERVEKISTRINGRGLDLRAEGLNFLQYTDAPGALGTVGTKLGAA  
GINIEAAALTQAEKGDGAVLILRVESAVSVEELEAEINAELGATSFQVDLD

>EP931833.14 03-SEP-1999

VSQNGRPVVLIIADKLAQSTVDALGDAVEVRWVDGPNRPELLDVTKEADALLVRSATTVDA  
EVIAAAPNLKIVGRAGVGLDNVDI PAATEAGVMVANAPTSNIHSACEHAISLLLSTARQI  
PAADATLREGGEWKRSSFNNGVEIFGKTVGIVGFGHIGQLFAQRLAAFETTIVAYDPYANPA  
RAAQLNVELVELDELMSRSDFTVIHLPKTKETAGMFDAQLLAKSKKQIIINAARGGLVD  
EQALADAIESGHIRGAGFDVYSTEPCTDSPLFKLPQVVVTPHLGASTEEAQDRAGTDVAD  
SVLKALAGEFVADAVNVSGGRVGEKVAWMDLARKLGLLAGKLVDAAPVSVIEVEARGELS  
SEQVDALGLSAVRGLFSGII EESVTFVNAPRIAEERGLDISVKTNSESVTHRSVLQVKVI  
TGSGASATVVGALTGLERVEKISTRINGRGLDLRAEGLNFLQYTDAPGALGTVGTKLGAA  
GINIEAAALTQAEKGDGAVLILRVESAVSVEELEAEINAELGATSFQVDLD

>W09931269.2 03-SEP-1999

MRQFQIILISLVVSI IRCVVADV DITSPKSGETFSGSSGSASIKITWDDSDSDSPKSLD  
NAKGYSISLCTGPTSDGDIQCLDPLVKNEAIAGKSKTVSIPQNSVPNGYFFQIYVFTFN  
GGTTIHYSPRFKLTGMSGPTATLDVTETGSPADQASGFDATTADSKSFTVPYTLQTGK  
TRYAPMQMPGKVTATTWSMKFPTS AVTYYSTKAGTPNVASTITPGWSYTAESAVNYAS  
VAPYPTYWYPASERVSKATISAATKRRRWLD

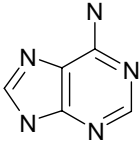
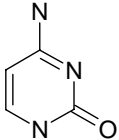
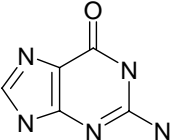
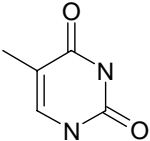
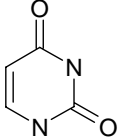
## Contact Details

	<b>North &amp; South America</b>	<b>Japan</b>
Email:	custserv@derwentus.com	ts.support.jp@thomson.com
Postal Address:	Thomson Scientific 1725 Duke Street Suite 250 Alexandria VA 22314 USA	Derwent Information Japan Thomson Corp. Japan Ltd 5F East, Palaceside Building 1-1 Hitotsubashi 1-Chome Chiyoda-Ku, Tokyo 100-0003 Japan
Telephone:	<b>+1 800 451 3551</b>	<b>+81 (0)3 5218 6500</b>
Fax:	+1 703 838 0450	+81 (0)3 5218 7840

	<b>Europe &amp; Rest of World</b>
Email:	ts.support.emea@thomson.com
Postal Address:	Thomson Scientific 14 Great Queen Street London WC2B 5DF UK
Telephone:	+44 (0)20 7344 2999
Fax:	+44 (0)20 7344 2900

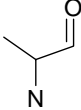
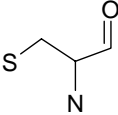
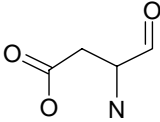
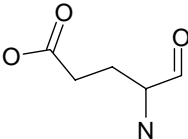
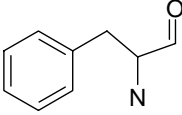

# Appendix

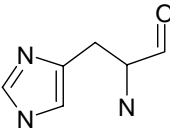
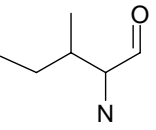
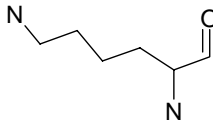
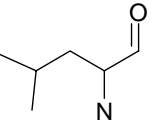
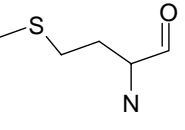
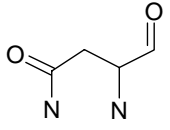
## Nucleic Acid Structures

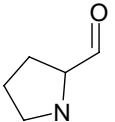
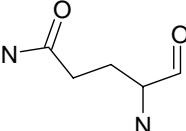
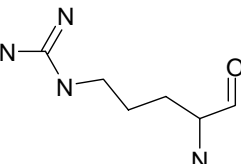
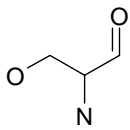
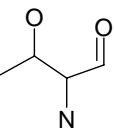
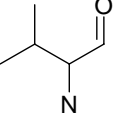
a	Adenine	
c	Cytosine	
g	Guanine	
t	Thymine	
u	Uracil	

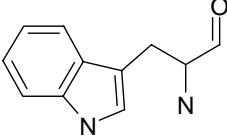
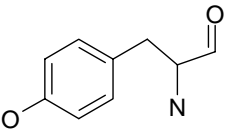
m	a or c
r	a or g
w	a or t/u
s	c or g
y	c or t/u
k	g or t/u
v	a, c or g; not t/u
h	a, c or t/u; not g
d	a, g or t/u; not c
b	c, g or t/u; not a
n	a, c, g or t/u; Unknown or Other

## Amino Acid Structures

A	Alanine	
C	Cysteine	
D	Aspartic acid	
E	Glutamic acid	
F	Phenylalanine	
G	Glycine	

H	Histidine	
I	Isoleucine	
K	Lysine	
L	Leucine	
M	Methionine	
N	Asparagine	

P	Proline	
Q	Glutamine	
R	Arginine	
S	Serine	
T	Threonine	
V	Valine	

W	Tryptophan	
Y	Tyrosine	
B	Aspartic acid or Asparagine	
Z	Glutamine or Glutamic acid	
X	Undetermined or atypical	

# Index

## A

Adenine 4  
Alanine 5  
Amino Acid Structures 11  
Annotation Lines 2  
Arginine 5  
Asparagine 5  
Aspartic acid 5

## C

Contact Details 8  
    Europe & Rest of World 8  
    Japan 8  
    North & South America 8  
Country Code 2  
Cysteine 5  
Cytosine 4

## D

Data Specification 2  
Definitions 4

## F

FastAlert\_P.txt 1  
FastAlert\_N.txt 1  
File Content Description 1

## G

Glutamic acid 5  
Glutamine 5  
Glycine 5  
Guanine 4

## H

Histidine 5

## I

Isoleucine 5  
IUPAC Definitions 4

## L

Leucine 5  
Lysine 5

## M

Methionine 5

## N

Nucleic acid bases 4  
Nucleic acid sequences 3, 6  
Nucleic acid structures 9

## P

Patent number 2  
Phenylalanine 5  
Proline 5  
Protein or Peptide Sequences 3, 7

## S

Sample Records 6  
Sequence ID number 2  
Sequences  
    Nucleic acids 3, 6  
    Protein/peptide 3, 7  
Serial of the patent 2  
Serine 5

**Structures**

Amino acid 11

Nucleic acid 9

**T**

Threonine 5

Thymine 4

Tryptophan 5

Tyrosine 5

**U**

Universal Freephone 8

Uracil 4

**V**

Valine 5

## Notes

## Notes