(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 September 2002 (12.09.2002)

PCT

(10) International Publication Number
WO 02/070558 A2

(51) International Patent Classification⁷: C07K 14/705

(21) International Application Number: PCT/GB02/00941

(22) International Filing Date: 5 March 2002 (05.03.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0105401.4 5 March 2001 (05.03.2001) GB

(71) **Applicant** *(for all designated States except US)*: **INPHARMATICA LIMITED** [GB/GB]; 60 Charlotte Street, London W1T 2NU (GB).

(72) **Inventors; and**
(75) **Inventors/Applicants** *(for US only)*: **FAGAN, Richard, Joseph** [US/GB]; Inpharmatica Limited, 60 Charlotte Street, London W1T 2NU (GB). **PHELPS, Christopher, Benjamin** [GB/GB]; Inpharmatica Limited, 60 Charlotte Street, London W1T 2NU (GB).

(74) **Agents: MERCER, Christopher, Paul** et al.; Carpmaels & Ransford, 43 Bloomsbury Square, London WC1 2RA (GB).

(81) **Designated States** *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

WO 02/070558 A2

(54) Title: NUCLEAR HORMONE RECEPTOR LIGAND BINDING DOMAIN

(57) Abstract: The invention relates to a novel protein, termed BAA07892.2, herein identified as a Nuclear Hormone Receptor Ligand Binding Domain and to the use of this protein and nucleic acid sequence from the encoding gene in the diagnosis, prevention and treatment of disease.

NUCLEAR HORMONE RECEPTOR LIGAND BINDING DOMAIN

This invention relates to a novel protein, termed BAA07892.2 herein identified as a Nuclear Hormone Receptor Ligand Binding Domain and to the use of this protein and nucleic acid sequence from the encoding gene in the diagnosis, prevention and treatment of disease.

All publications, patents and patent applications cited herein are incorporated in full by reference.

BACKGROUND

The process of drug discovery is presently undergoing a fundamental revolution as the era of functional genomics comes of age. The term "functional genomics" applies to an approach utilising bioinformatics tools to ascribe function to protein sequences of interest. Such tools are becoming increasingly necessary as the speed of generation of sequence data is rapidly outpacing the ability of research laboratories to assign functions to these protein sequences.

As bioinformatics tools increase in potency and in accuracy, these tools are rapidly replacing the conventional techniques of biochemical characterisation. Indeed, the advanced bioinformatics tools used in identifying the present invention are now capable of outputting results in which a high degree of confidence can be placed.

Various institutions and commercial organisations are examining sequence data as they become available and significant discoveries are being made on an on-going basis. However, there remains a continuing need to identify and characterise further genes and the polypeptides that they encode, as targets for research and for drug discovery.

Recently, a remarkable tool for the evaluation of sequences of unknown function has been developed by the Applicant for the present invention. This tool is a database system, termed the Biopendium search database, that is the subject of co-pending International Patent Application No. PCT/GB01/01105. This database system consists of an integrated data resource created using proprietary technology and containing information generated from an all-by-all comparison of all available protein or nucleic acid sequences.

The aim behind the integration of these sequence data from separate data resources is to combine as much data as possible, relating both to the sequences themselves and to information relevant to each sequence, into one integrated resource. All the available data relating to each sequence, including data on the three-dimensional structure of the encoded protein, if this is available, are integrated together to make best use of the information that is known about each sequence and thus to allow the most educated predictions to be made from comparisons of these sequences. The annotation that is generated in the database and which accompanies each sequence entry imparts a biologically relevant context to the sequence information.

This data resource has made possible the accurate prediction of protein function from sequence alone. Using conventional technology, this is only possible for proteins that exhibit a high degree of sequence identity (above about 20%-30% identity) to other proteins in the same functional family. Accurate predictions are not possible for proteins that exhibit a very low degree of sequence homology to other related proteins of known function.

In the present case, a protein whose sequence is recorded in a publicly available database as KIAA0097 protein (NCBI Genebank nucleotide accession number D43948 and a Genebank protein accession number BAA07892.2), is implicated as a novel member of the Nuclear Hormone Receptor Ligand Binding Domain family.

Introduction to Nuclear Hormone Receptor Ligand Binding Domains

The Nuclear Hormone Receptor gene superfamily (see Table 1) encodes structurally related proteins that regulate the transcription of target genes. These proteins include receptors for steroid and thyroid hormones, vitamins, and other proteins for which no ligands have been found. Nuclear Receptors are composed of two key domains, a DNA-Binding Domain (DBD) and a Ligand Binding Domain (LBD). The DBD directs the receptors to bind specific DNA sequences as monomers, homodimers, or heterodimers. The DBD is a particular type of zinc-finger, found only in Nuclear Receptors. Nuclear Receptors with DBDs can be readily identified at the sequence level by searching for matches to the PROSITE consensus sequence (PS00031).[1]

The Ligand Binding Domain (LBD) binds and responds to the cognate hormone. Ligand binding to the LBD triggers a conformational change which expels a bound "Nuclear Receptor Co-Repressor". The site previously occupied by the Co-Repressor is then free to recruit a "Nuclear Receptor Co-Activator". This Ligand-triggered swap of a Co-Repressor for a Co-Activator is the mechanism by which Ligand binding leads to the transcriptional activation of target genes. All ligand binding domains contain a consensus sequence, the "LBD motif" (see Table 2) which mediates Co-Repressor and Co-Activator binding. The LBD is the binding site for all Nuclear Hormone Receptor targeted drugs to date and it is thus desirable to identify novel Ligand Binding Domains since these will be attractive drug targets. Ligand Binding Domains share low sequence identity (~15%) but have very similar structures and so present ideal targets for a structure-based relationship tool such as Genome Threader.

Many protein sequences have already been annotated in the public domain as Nuclear Hormone Receptors by their possession of DBDs using basic search tools like PROSITE, and their LBDs inferred on the basis of this. Because of this it is anticipated that any novel LBDs identified by Genome Threader *which are not annotated as nuclear receptors* will lack the DBD entirely. A precedent for a protein which has an LBD but lacks a DBD is provided by DAX1. Thus we annotate these DBD-less hits not as "Nuclear Hormone Receptors" but rather as containing a "Nuclear Hormone Receptor Ligand Binding Domain".

Family: Steroid Hormone Receptors	
Subfamilies	Glucocorticoid Receptors
	Progesterone Receptors
	Androgen Receptors
	Estrogen Receptors
Family: Thyroid Hormone Receptor-like Factors	
Subfamilies	Retinoic Acid Receptors (RARs)

	Retinoid X Receptors (RXRs)
	Thyroid Hormone Receptors
	Vitamin D Receptor
	NGFI-B
	FTZ-F1
	Peroxisome Proliferator Activated Receptors (PPARs)
	Ecdysone Receptors
	Retinoid Orphan Receptors (RORs)
	Tailess/COUP
	HNF-4
	CF1
	Knirps
Family: DAX1	
Subfamilies	DAX1

Table 1: Nuclear hormone Receptor Superfamily

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
L	Any 2 residues		L	Any 3 residues (or 2 residues or 4 residues)			D	Q	Any 2 residues (or 1 or 3 residues)		L	L
I			I				E	N			I	I
A			A					R			A	A
V			V					H			V	V
M			M					K			M	M
F			F					S			F	F
Y			Y					T			Y	Y
W			W								W	W

Table 2: The “LBD motif”. Numbers along the top row refer to residue position within the motif. Letters refer to amino acids by the 1-letter code. Letters within one column are all acceptable for that position within the motif. For example L, I, A, V, M, F, Y or W can occupy the first position of the “LBD motif”. Note that there is observed variation in the number of residues found between position 4 and 8, and position 9 and 12. The “LBD motif” was constructed by aligning 681 sequences of Nuclear Hormone Receptor Ligand Binding Domains, and identifying conserved patterns of residues.

II. Nuclear Hormone Receptors and Disease

Nuclear Hormone Receptors have been shown to play a role in diverse physiological functions, many of which can play a role in disease processes (see Table 3).

Table 3. Nuclear Hormone Receptors and disease.

Nuclear Hormone Receptor	Disease
Androgen Receptor	Androgen Insensitivity Syndrome (Lubahn *et al.* 1989 Proc. Natl. Acad. Sci. USA 86, 9534-9538).
	Reifenstein syndrome (Wooster *et al.* 1992 Nat. Genet. 2, 132-134).
	X-linked recessive spinal and bulbar muscular atrophy (MacLean *et al.* 1995 Mol. Cell. Endocrinol. 112,133-141).
	Male breast cancer ((Wooster *et al.* 1992 Nat. Genet. 2, 132-134).
Glucocorticoid Receptor	Nelson's syndrome (Karl *et al.* 1996 J. Clin. Endocrinol. Metab. 81, 124-129).
	Glucocorticoid resistant acute T-cell leukemia (Hala *et al.* 1996 Int. J. Cancer 68, 663-668).
Mineralocorticoid Receptor	Pseudohypoaldosteronism (Chung *et al.* 1995 J. Clin. Endocrinol. Metab. 80, 3341-3345).
Estrogen Receptor alpha	ER alpha expression is elevated in a subset of human breast cancers. The application of Tamoxifen is the major therapy to prevent breast tumour progression. Unfortunately 35% of ER alpha positive breast cancers are Tamoxifen resistant (Petrangeli *et al.* 1994 J. Steroid Biochem. Mol. Biol. 49, 327-331).
Vitamin D3 Receptor	Mutations in the Vitamin D3 receptor produce a hereditary disorder similar in phenotype to Vitamin D3 deficiency (Rickets) (Hughes *et al.* 1988 Science 242, 1702-1725).
Retinoic Acid Receptor alpha	Acute Myeloid Leukemia (Lavau and Dejean 1994 Leukemia 8, 9-15).
Thyroid Hormone Receptor beta	"Generalised Resistance to Thyroid Hormones" (GRTH) (Refetoff 1994 Thyroid 4, 345-349).
DAX1	X-linked Adrenal Hypoplasia Congenita (AHC) and Hypogonadism (Ito *et al.* 1997 Mol. Cell. Biol. 17, 1476-1483).

Alteration of Nuclear Hormone Receptors by ligands which bind to their LBD thus provides a means to alter the disease phenotype. There is thus a great need for the identification of novel Nuclear Hormone Receptor Ligand Binding Domains, as these proteins may play a role in the diseases identified above, as well as in other disease states. The identification of novel Nuclear Hormone Receptor Ligand Binding Domains is thus highly relevant for the treatment and diagnosis of disease, particularly those identified in Table 3.

THE INVENTION

The invention is based on the discovery that the BAA07892.2 protein functions as a Nuclear Hormone Receptor Ligand Binding Domain.

For the BAA07892.2 protein, it has been found that a region including residues 1507-1742 of this protein sequence adopts an equivalent fold to residues 4 (GLN185) to 234 (MET415) of the Human Retinoic Acid Receptor gamma (PDB code 3LBD). Human Retinoic Acid Receptor gamma is known to function as a Nuclear Hormone Receptor Ligand Binding Domain. Furthermore, the "LBD motif" residues PHE251, LEU254, ASP258, GLN259, LEU262 and LEU263 of the Human Retinoic Acid Receptor gamma are conserved as VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583 in BAA07892.2, respectively. This relationship is not just to Human Retinoic Acid Receptor gamma, but rather to the Nuclear Hormone Receptor Ligand Binding Domain family as a whole. Thus, by reference to the Genome Threader™ alignment of BAA07892.2 with the Human Retinoic Acid Receptor gamma (3LBD) VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583 of BAA07892.2 are predicted to form the "LBD motif" residues.

The combination of equivalent fold and conservation of "LBD motif" residues allows the functional annotation of this region of BAA07892.2, and therefore proteins that include this region, as possessing Nuclear Hormone Receptor Ligand Binding Domain activity.

In a first aspect, the invention provides a polypeptide, which polypeptide:

(i) comprises the amino acid sequence as recited in SEQ ID NO:2;

(ii) is a fragment thereof having Nuclear Hormone Receptor Ligand Binding Domain activity or having an antigenic determinant in common with the polypeptides of (i); or

(iii) is a functional equivalent of (i) or (ii).

Preferably, the polypeptide:

(i) consists of the amino acid sequence as recited in SEQ ID NO:2;

(ii) is a fragment thereof having Nuclear Hormone Receptor Ligand Binding Domain activity or having an antigenic determinant in common with the polypeptides of (i); or

(iii) is a functional equivalent of (i) or (ii).

The polypeptide having the sequence recited in SEQ ID NO:2 is referred to hereafter as "the LBDS1 polypeptide".

According to this aspect of the invention, a preferred polypeptide fragment according to part ii) above includes the region of the LBDS1 polypeptide that is predicted as that responsible for Nuclear Hormone Receptor Ligand Binding Domain activity (hereafter, the "LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region"), or is a variant thereof that possesses the "LBD motif" (VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues). As defined herein, the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region is considered to extend between residue 1507 and residue 1742 of the LBDS1 polypeptide sequence.

This aspect of the invention also includes fusion proteins that incorporate polypeptide fragments and variants of these polypeptide fragments as defined above, provided that said fusion proteins possess activity as a Nuclear Hormone Receptor Ligand Binding Domain.

In a second aspect, the invention provides a purified nucleic acid molecule that encodes a polypeptide of the first aspect of the invention. Preferably, the purified nucleic acid molecule has the nucleic acid sequence as recited in SEQ ID NO:1 (encoding the LBDS1 polypeptide), or is a redundant equivalent or fragment of this sequence. A preferred nucleic acid fragment is one that encodes a polypeptide fragment according to part ii) above, preferably a polypeptide fragment that includes the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region, or that encodes a variant of these fragments as this term is defined above.

In a third aspect, the invention provides a purified nucleic acid molecule which hybridizes under high stringency conditions with a nucleic acid molecule of the second aspect of the invention.

In a fourth aspect, the invention provides a vector, such as an expression vector, that contains a nucleic acid molecule of the second or third aspect of the invention.

In a fifth aspect, the invention provides a host cell transformed with a vector of the fourth aspect of the invention.

In a sixth aspect, the invention provides a ligand which binds specifically to, and which preferably inhibits the Nuclear Hormone Receptor Ligand Binding Domain activity of, a polypeptide of the first aspect of the invention.

In a seventh aspect, the invention provides a compound that is effective to alter the expression of a natural gene which encodes a polypeptide of the first aspect of the invention or to regulate the activity of a polypeptide of the first aspect of the invention.

A compound of the seventh aspect of the invention may either increase (agonise) or decrease (antagonise) the level of expression of the gene or the activity of the polypeptide. Importantly, the identification of the function of the region defined herein as the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide, respectively, allows for the design of screening methods capable of identifying compounds that are effective in the treatment and/or diagnosis of diseases in which Nuclear Hormone Receptor Ligand Binding Domains are implicated.

In an eighth aspect, the invention provides a polypeptide of the first aspect of the invention, or a nucleic acid molecule of the second or third aspect of the invention, or a vector of the fourth aspect of the invention, or a ligand of the fifth aspect of the invention, or a compound of the sixth aspect of the invention, for use in therapy or diagnosis. These molecules may also be used in the manufacture of a medicament for the treatment of cell proliferative disorders, including neoplasm, cancer and myeloproliferative disorders and angiogenesis, metabolic disorders, including obesity, osteoporosis, non-insulin dependent diabetes, lipid metabolism disorder, hyperthyroidism, hyperparathyroidism, hypercalcemia, hypercholestrolemia, hyperlipidemia, cardiovascular disorders including, hypertension, athersclerosis, hyperinsulinemia, heart arrhythmia, renal disorders, including glomerulonephritis, renovascular hypertension, skin disorders including dermatological disease, psoriasis, acne, aging, eczema, wound healing, inflammation, including inflammatory bowel disease, respiratory tract inflammation, emphysema, asthma, immune disorder, autoimmune disease, multiple sclerosis, nervous system disorders, including anxiety, depression, neurodegenerative disease, Alzheimer's disease Parkinson's disease, brain injury, stroke and pain, infection, including virus infection, and other conditions in which nuclear hormones are implicated.

In a ninth aspect, the invention provides a method of diagnosing a disease in a patient, comprising assessing the level of expression of a natural gene encoding a polypeptide of the first aspect of the invention or the activity of a polypeptide of the first aspect of the invention in tissue from said patient and comparing said level of expression or activity to a control level, wherein a level that is different to said control level is indicative of disease. Such a method will preferably be carried out *in vitro*. Similar methods may be used for monitoring the therapeutic treatment of disease in a patient, wherein altering the level of expression or activity of a polypeptide or nucleic acid molecule over the period of time towards a control level is indicative of regression of disease.

A preferred method for detecting polypeptides of the first aspect of the invention comprises the steps of: (a) contacting a ligand, such as an antibody, of the sixth aspect of the invention with a biological sample under conditions suitable for the formation of a ligand-polypeptide complex; and (b) detecting said complex.

A number of different such methods according to the ninth aspect of the invention exist, as the skilled reader will be aware, such as methods of nucleic acid hybridization with short probes, point mutation analysis, polymerase chain reaction (PCR) amplification and methods using antibodies to detect aberrant protein levels. Similar methods may be used on a short or long term basis to allow therapeutic treatment of a disease to be monitored in a patient. The invention also provides kits that are useful in these methods for diagnosing disease.

In a tenth aspect, the invention provides for the use of a polypeptide of the first aspect of the invention as a Nuclear Hormone Receptor Ligand Binding Domain. The invention also provides for the use of a nucleic acid molecule according to the second or third aspects of the invention to express a protein that possesses Nuclear Hormone Receptor Ligand Binding Domain activity. The invention also provides a method for effecting Nuclear Hormone Receptor Ligand Binding Domain activity, said method utilising a polypeptide of the first aspect of the invention.

In an eleventh aspect, the invention provides a pharmaceutical composition comprising a polypeptide of the first aspect of the invention, or a nucleic acid molecule of the second or third aspect of the invention, or a vector of the fourth aspect of the invention, or a ligand of the sixth aspect of the invention, or a compound of the seventh aspect of the invention, in conjunction with a pharmaceutically-acceptable carrier.

In a twelfth aspect, the present invention provides a polypeptide of the first aspect of the invention, or a nucleic acid molecule of the second or third aspect of the invention, or a vector of the fourth aspect of the invention, or a ligand of the sixth aspect of the invention, or a compound of the seventh aspect of the invention, for use in the manufacture of a medicament for the diagnosis or treatment of a disease, such as cell proliferative disorders, including neoplasm, cancer and myeloproliferative disorders and angiogenesis, metabolic disorders, including obesity, osteoporosis, non-insulin dependent diabetes, lipid metabolism disorder, hyperthyroidism, hyperparathyroidism, hypercalcemia, hypercholestrolemia, hyperlipidemia, cardiovascular disorders including, hypertension, athersclerosis, hyperinsulinemia, heart arrhythmia, renal disorders, including glomerulonephritis, renovascular hypertension, skin disorders including dermatological disease, psoriasis, acne, aging, eczema, wound healing, inflammation, including inflammatory bowel disease, respiratory tract inflammation, emphysema, asthma, immune disorder, autoimmune disease, multiple sclerosis, nervous system disorders, including anxiety, depression, neurodegenerative disease, Alzheimer's disease

Parkinson's disease, brain injury, stroke and pain, infection, including virus infection, and other conditions in which nuclear hormones are implicated.

In a thirteenth aspect, the invention provides a method of treating a disease in a patient comprising administering to the patient a polypeptide of the first aspect of the invention, or a nucleic acid molecule of the second or third aspect of the invention, or a vector of the fourth aspect of the invention, or a ligand of the sixth aspect of the invention, or a compound of the seventh aspect of the invention.

For diseases in which the expression of a natural gene encoding a polypeptide of the first aspect of the invention, or in which the activity of a polypeptide of the first aspect of the invention, is lower in a diseased patient when compared to the level of expression or activity in a healthy patient, the polypeptide, nucleic acid molecule, ligand or compound administered to the patient should be an agonist. Conversely, for diseases in which the expression of the natural gene or activity of the polypeptide is higher in a diseased patient when compared to the level of expression or activity in a healthy patient, the polypeptide, nucleic acid molecule, ligand or compound administered to the patient should be an antagonist. Examples of such antagonists include antisense nucleic acid molecules, ribozymes and ligands, such as antibodies.

In a fourteenth aspect, the invention provides transgenic or knockout non-human animals that have been transformed to express higher, lower or absent levels of a polypeptide of the first aspect of the invention. Such transgenic animals are very useful models for the study of disease and may also be using in screening regimes for the identification of compounds that are effective in the treatment or diagnosis of such a disease.

A summary of standard techniques and procedures which may be employed in order to utilise the invention is given below. It will be understood that this invention is not limited to the particular methodology, protocols, cell lines, vectors and reagents described. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only and it is not intended that this terminology should limit the scope of the present invention. The extent of the invention is limited only by the terms of the appended claims.

Standard abbreviations for nucleotides and amino acids are used in this specification.

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of molecular biology, microbiology, recombinant DNA technology and immunology, which are within the skill of those working in the art.

Such techniques are explained fully in the literature. Examples of particularly suitable texts for consultation include the following: Sambrook Molecular Cloning; A Laboratory Manual, Second Edition (1989); DNA Cloning, Volumes I and II (D.N. Glover ed. 1985); Oligonucleotide Synthesis (M.J. Gait ed. 1984); Nucleic Acid Hybridization (B.D. Hames & S.J. Higgins eds. 1984); Transcription and Translation (B.D. Hames & S.J. Higgins eds. 1984); Animal Cell Culture (R.I. Freshney ed. 1986); Immobilized Cells and Enzymes (IRL Press, 1986); B. Perbal, A Practical Guide to Molecular Cloning (1984); the Methods in Enzymology series (Academic Press, Inc.), especially volumes 154 & 155; Gene Transfer Vectors for Mammalian Cells (J.H. Miller and M.P. Calos eds. 1987, Cold Spring Harbor Laboratory); Immunochemical Methods in Cell and Molecular Biology (Mayer and Walker, eds. 1987, Academic Press, London); Scopes, (1987) Protein Purification: Principles and Practice, Second Edition (Springer Verlag, N.Y.); and Handbook of Experimental Immunology, Volumes I-IV (D.M. Weir and C. C. Blackwell eds. 1986).

As used herein, the term "polypeptide" includes any peptide or protein comprising two or more amino acids joined to each other by peptide bonds or modified peptide bonds, i.e. peptide isosteres. This term refers both to short chains (peptides and oligopeptides) and to longer chains (proteins).

The polypeptide of the present invention may be in the form of a mature protein or may be a pre-, pro- or prepro- protein that can be activated by cleavage of the pre-, pro- or prepro- portion to produce an active mature polypeptide. In such polypeptides, the pre-, pro- or prepro- sequence may be a leader or secretory sequence or may be a sequence that is employed for purification of the mature polypeptide sequence.

The polypeptide of the first aspect of the invention may form part of a fusion protein. For example, it is often advantageous to include one or more additional amino acid sequences which may contain secretory or leader sequences, pro-sequences, sequences which aid in purification, or sequences that confer higher protein stability, for example during recombinant production. Alternatively or additionally, the mature polypeptide may be fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol).

Polypeptides may contain amino acids other than the 20 gene-encoded amino acids, modified either by natural processes, such as by post-translational processing or by chemical modification techniques which are well known in the art. Among the known modifications which may commonly be present in polypeptides of the present invention are glycosylation, lipid attachment, sulphation, gamma-carboxylation, for instance of glutamic acid residues, hydroxylation and ADP-ribosylation. Other potential modifications include acetylation, acylation, amidation, covalent attachment of flavin, covalent attachment of a haeme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid derivative, covalent attachment of phosphatidylinositol, cross-linking, cyclization, disulphide bond formation, demethylation, formation of covalent cross-links, formation of cysteine, formation of pyroglutamate, formylation, GPI anchor formation, iodination, methylation, myristoylation, oxidation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination.

Modifications can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. In fact, blockage of the amino or carboxyl terminus in a polypeptide, or both, by a covalent modification is common in naturally-occurring and synthetic polypeptides and such modifications may be present in polypeptides of the present invention.

The modifications that occur in a polypeptide often will be a function of how the polypeptide is made. For polypeptides that are made recombinantly, the nature and extent of the modifications in large part will be determined by the post-translational modification capacity of the particular host cell and the modification signals that are present in the amino acid sequence of the polypeptide in question. For instance, glycosylation patterns vary between different types of host cell.

The polypeptides of the present invention can be prepared in any suitable manner. Such polypeptides include isolated naturally-occurring polypeptides (for example purified from cell culture), recombinantly-produced polypeptides (including fusion proteins), synthetically-produced polypeptides or polypeptides that are produced by a combination of these methods.

The functionally-equivalent polypeptides of the first aspect of the invention may be polypeptides that are homologous to the LBDS1 polypeptide. Two polypeptides are said to be "homologous", as the term is used herein, if the sequence of one of the polypeptides has a high enough degree of identity or similarity to the sequence of the other polypeptide. "Identity" indicates that at any particular position in the aligned sequences, the amino acid residue is identical between the sequences. "Similarity" indicates that, at any particular position in the aligned sequences, the amino acid residue is of a similar type between the sequences. Degrees of identity and similarity can be readily calculated (Computational Molecular Biology, Lesk, A.M., ed., Oxford University Press, New York, 1988; Biocomputing. Informatics and Genome Projects, Smith, D.W., ed., Academic Press, New York, 1993; Computer Analysis of Sequence Data, Part 1, Griffin, A.M., and Griffin, H.G., eds., Humana Press, New Jersey, 1994; Sequence Analysis in Molecular Biology, von Heinje, G., Academic Press, 1987; and Sequence Analysis Primer, Gribskov, M. and Devereux, J., eds., M Stockton Press, New York, 1991).

Homologous polypeptides therefore include natural biological variants (for example, allelic variants or geographical variations within the species from which the polypeptides are derived) and mutants (such as mutants containing amino acid substitutions, insertions or deletions) of the LBDS1 polypeptide. Such mutants may include polypeptides in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code. Typical such substitutions are among Ala, Val, Leu and Ile; among Ser and Thr; among the acidic residues Asp and Glu; among Asn and Gln; among the basic residues Lys and Arg; or among the aromatic residues Phe and Tyr. Particularly preferred are variants in which several, i.e. between 5 and 10, 1 and 5, 1 and 3, 1 and 2 or just 1 amino acids are substituted, deleted or added in any combination. Especially preferred are silent substitutions, additions and deletions, which do not alter the properties and activities of the protein. Also especially preferred in this regard are conservative substitutions.

Such mutants also include polypeptides in which one or more of the amino acid residues includes a substituent group.

Typically, greater than 80% identity between two polypeptides (preferably, over a specified region) is considered to be an indication of functional equivalence. Preferably, functionally equivalent polypeptides of the first aspect of the invention have a degree of sequence identity with the LBDS1 polypeptide, or with active fragments thereof, of greater than 80%. More preferred polypeptides have degrees of identity of greater than 85%, 90%, 95%, 98% or 99%, respectively with the LBDS1 polypeptide, or with active fragments thereof.

Percentage identity, as referred to herein, is as determined using BLAST version 2.1.3 using the default parameters specified by the NCBI (the National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/) [Blosum 62 matrix; gap open penalty=11 and gap extension penalty=1].

In the present case, preferred active fragments of the LBDS1 polypeptide are those that include the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region and which possess the “LBD motif” of residues VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues. By “equivalent residues” is meant residues that are equivalent to the “LBD motif” residues, provided that the Nuclear Hormone Receptor Ligand Binding Domain region retains activity as a Nuclear Hormone Receptor Ligand Binding Domain. For example VAL1573 may be replaced by LEU, ILE, ALA, MET, PHE, TYR or TRP. For example LEU1574 may be replaced by ILE, ALA, VAL, MET, PHE, TYR or TRP. For example ASP1578 may be replaced by GLU. For example LYS1579 may be replaced by GLN, ASN, ARG, HIS, SER or THR. For example ALA1582 may be replaced by LEU, ILE, VAL, MET, PHE, TYR or TRP. For example MET1583 may be replaced by LEU, ILE, ALA, VAL, PHE, TYR or TRP. Accordingly, this aspect of the invention includes polypeptides that have degrees of identity of greater than 80%, preferably, greater than 85%, 90%, 95%, 98% or 99%, respectively, with the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide and which possess the “LBD motif” of VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues. As discussed above, the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region is considered to extend between residue 1507 and residue 1742 of the LBDS1 polypeptide sequence.

The functionally-equivalent polypeptides of the first aspect of the invention may also be polypeptides which have been identified using one or more techniques of structural alignment. For example, the Inpharmatica Genome Threader^{TM} technology that forms one aspect of the search tools used to generate the Biopendium search database may be used (see co-pending International patent application PCT/GB01/01105) to identify polypeptides of presently-unknown function which, while having low sequence identity as compared to the LBDS1 polypeptide, are predicted to have Nuclear Hormone Receptor Ligand Binding Domain activity, by virtue of sharing significant structural homology with the LBDS1 polypeptide sequence.

By "significant structural homology" is meant that the Inpharmatica Genome Threader^{TM} predicts two proteins, or protein regions, to share structural homology with a certainty of at least 10% more preferably, at least 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and above. The certainty value of the Inpharmatica Genome Threader^{TM} is calculated as follows. A set of comparisons was initially performed using the Inpharmatica Genome Threader^{TM} exclusively using sequences of known structure. Some of the comparisons were between proteins that were known to be related (on the basis of structure). A neural network was then trained on the basis that it needed to best distinguish between the known relationships and known not-relationships taken from the CATH structure classification (www.biochem.ucl.ac.uk/bsm/cath). This resulted in a neural network score between 0 and 1. However, again as the number of proteins that are related and the number that are unrelated were known, it was possible to partition the neural network results into packets and calculate empirically the percentage of the results that were correct. In this manner, any genuine prediction in the Biopendium search database has an attached neural network score and the percentage confidence is a reflection of how successful the Inpharmatica Genome Threader^{TM} was in the training/testing set.

Structural homologues of LBDS1 should share structural homology with the LBDS1

Nuclear Hormone Receptor Ligand Binding Domain region and possess the "LBD motif" residues VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues. Such structural homologues are predicted to have Nuclear Hormone Receptor Ligand Binding Domain activity by virtue of sharing significant structural homology with this polypeptide sequence and possessing the "LBD motif" residues.

The polypeptides of the first aspect of the invention also include fragments of the LBDS1 polypeptide, functional equivalents of the fragments of the LBDS1 polypeptide, and fragments of the functional equivalents of the LBDS1 polypeptides, provided that those functional equivalents and fragments retain Nuclear Hormone Receptor Ligand Binding Domain activity or have an antigenic determinant in common with the LBDS1 polypeptide.

As used herein, the term "fragment" refers to a polypeptide having an amino acid sequence that is the same as part, but not all, of the amino acid sequence of the LBDS1 polypeptides or one of its functional equivalents. The fragments should comprise at least n consecutive amino acids from the sequence and, depending on the particular sequence, n preferably is 7 or more (for example, 8, 10, 12, 14, 16, 18, 20 or more). Small fragments may form an antigenic determinant.

Preferred polypeptide fragments according to this aspect of the invention are fragments that include a region defined herein as the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptides, respectively. These regions are the regions that have been annotated as Nuclear Hormone Receptor Ligand Binding Domain.

For the LBDS1 polypeptide, this region is considered to extend between residue 1507 and residue 1742.

Variants of this fragment are included as embodiments of this aspect of the invention, provided that these variants possess activity as a Nuclear Hormone Receptor Ligand Binding Domain.

In one respect, the term "variant" is meant to include extended or truncated versions of this polypeptide fragment.

For extended variants, it is considered highly likely that the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide will fold correctly and show Nuclear Hormone Receptor Ligand Binding Domain activity if additional residues C terminal and/or N terminal of these boundaries in the LBDS1 polypeptide sequence are included in the polypeptide fragment. For example, an additional 5, 10, 20, 30, 40 or even 50 or more amino acid residues from the LBDS1 polypeptide sequence, or from a homologous sequence, may be included at either or both the C terminal and/or N terminal of the boundaries of the Nuclear Hormone Receptor Ligand Binding Domain regions of the LBDS1 polypeptide, without prejudicing the ability of the polypeptide fragment to fold correctly and exhibit Nuclear Hormone Receptor Ligand Binding Domain activity.

For truncated variants of the LBDS1 polypeptide, one or more amino acid residues may be deleted at either or both the C terminus or the N terminus of the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide, although the "LBD motif" residues (VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583), or equivalent residues should be maintained intact; deletions should not extend so far into the polypeptide sequence that any of these residues are deleted.

In a second respect, the term "variant" includes homologues of the polypeptide fragments described above, that possess significant sequence homology with the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide and which possess the "LBD motif" residues (VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583), or equivalent residues, provided that said variants retain activity as an Nuclear Hormone Receptor Ligand Binding Domain.

Homologues include those polypeptide molecules that possess greater than 30% identity with the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain regions, of the LBDS1 polypeptides, respectively. Percentage identity is as determined using BLAST version 2.1.3 using the default parameters specified by the NCBI (the National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/) [Blosum 62 matrix; gap open penalty=11 and gap extension penalty=1]. Preferably, variant homologues of polypeptide fragments of this aspect of the invention have a degree of sequence identity with the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain regions, of the LBDS1 polypeptides, respectively, of greater than 80%. More preferred variant polypeptides have degrees of identity of greater than 85%, 90%, 95%, 98% or 99%, respectively with the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain regions of the LBDS1, polypeptides, provided that said variants retain activity as a Nuclear Hormone Receptor Ligand Binding Domain. Variant polypeptides also include homologues of the truncated forms of the polypeptide fragments discussed above, provided that said variants retain activity as a Nuclear Hormone Receptor Ligand Binding Domain.

The polypeptide fragments of the first aspect of the invention may be polypeptide fragments that exhibit significant structural homology with the structure of the polypeptide fragment defined by the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain regions, of the LBDS1 polypeptide sequence, for example, as identified by the Inpharmatica Genome Threader™. Accordingly, polypeptide fragments that are structural homologues of the polypeptide fragments defined by the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain regions of the LBDS1 polypeptide sequence should adopt the same fold as that adopted by this polypeptide fragment, as this fold is defined above.

Structural homologues of the polypeptide fragment defined by the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region should also retain the “LBD motif” residues VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues.

Such fragments may be "free-standing", i.e. not part of or fused to other amino acids or polypeptides, or they may be comprised within a larger polypeptide of which they form a part or region. When comprised within a larger polypeptide, the fragment of the invention most preferably forms a single continuous region. For instance, certain preferred embodiments relate to a fragment having a pre - and/or pro- polypeptide region fused to the amino terminus of the fragment and/or an additional region fused to the carboxyl terminus of the fragment. However, several fragments may be comprised within a single larger polypeptide.

The polypeptides of the present invention or their immunogenic fragments (comprising at least one antigenic determinant) can be used to generate ligands, such as polyclonal or monoclonal antibodies, that are immunospecific for the polypeptides. Such antibodies may be employed to isolate or to identify clones expressing the polypeptides of the invention or to purify the polypeptides by affinity chromatography. The antibodies may also be employed as diagnostic or therapeutic aids, amongst other applications, as will be apparent to the skilled reader.

The term "immunospecific" means that the antibodies have substantially greater affinity for the polypeptides of the invention than their affinity for other related polypeptides in the prior art. As used herein, the term "antibody" refers to intact molecules as well as to fragments thereof, such as Fab, F(ab')2 and Fv, which are capable of binding to the antigenic determinant in question. Such antibodies thus bind to the polypeptides of the first aspect of the invention.

If polyclonal antibodies are desired, a selected mammal, such as a mouse, rabbit, goat or horse, may be immunised with a polypeptide of the first aspect of the invention. The polypeptide used to immunise the animal can be derived by recombinant DNA technology or can be synthesized chemically. If desired, the polypeptide can be conjugated to a carrier protein. Commonly used carriers to which the polypeptides may be chemically coupled include bovine serum albumin, thyroglobulin and keyhole limpet haemocyanin. The coupled polypeptide is then used to immunise the animal. Serum from the immunised animal is collected and treated according to known procedures, for example by immunoaffinity chromatography.

Monoclonal antibodies to the polypeptides of the first aspect of the invention can also be readily produced by one skilled in the art. The general methodology for making monoclonal antibodies using hybridoma technology is well known (see, for example, Kohler, G. and Milstein, C., Nature 256: 495-497 (1975); Kozbor *et al.*, Immunology Today 4: 72 (1983); Cole *et al.*, 77-96 in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc. (1985).

Panels of monoclonal antibodies produced against the polypeptides of the first aspect of the invention can be screened for various properties, i.e., for isotype, epitope, affinity, etc. Monoclonal antibodies are particularly useful in purification of the individual polypeptides against which they are directed. Alternatively, genes encoding the monoclonal antibodies of interest may be isolated from hybridomas, for instance by PCR techniques known in the art, and cloned and expressed in appropriate vectors.

Chimeric antibodies, in which non-human variable regions are joined or fused to human constant regions (see, for example, Liu *et al.*, Proc. Natl. Acad. Sci. USA, 84, 3439 (1987)), may also be of use.

The antibody may be modified to make it less immunogenic in an individual, for example by humanisation (see Jones *et al.*, Nature, 321, 522 (1986); Verhoeyen *et al.*, Science, 239: 1534 (1988); Kabat *et al.*, J. Immunol., 147: 1709 (1991); Queen *et al.*, Proc. Natl Acad. Sci. USA, 86, 10029 (1989); Gorman *et al.*, Proc. Natl Acad. Sci. USA, 88: 34181 (1991); and Hodgson *et al.*, Bio/Technology 9: 421 (1991)). The term "humanised antibody", as used herein, refers to antibody molecules in which the CDR amino acids and selected other amino acids in the variable domains of the heavy and/or light chains of a non-human donor antibody have been substituted in place of the equivalent amino acids in a human antibody. The humanised antibody thus closely resembles a human antibody but has the binding ability of the donor antibody.

In a further alternative, the antibody may be a "bispecific" antibody, that is an antibody having two different antigen binding domains, each domain being directed against a different epitope.

Phage display technology may be utilised to select genes which encode antibodies with binding activities towards the polypeptides of the invention either from repertoires of PCR amplified V-genes of lymphocytes from humans screened for possessing the relevant antibodies, or from naive libraries (McCafferty, J. *et al.*, (1990), Nature 348, 552-554; Marks, J. *et al.*, (1992) Biotechnology 10, 779-783). The affinity of these antibodies can also be improved by chain shuffling (Clackson, T. *et al.*, (1991) Nature 352, 624-628).

Antibodies generated by the above techniques, whether polyclonal or monoclonal, have additional utility in that they may be employed as reagents in immunoassays, radioimmunoassays (RIA) or enzyme-linked immunosorbent assays (ELISA). In these applications, the antibodies can be labelled with an analytically-detectable reagent such as a radioisotope, a fluorescent molecule or an enzyme.

Preferred nucleic acid molecules of the second and third aspects of the invention are those which encode the polypeptide sequences recited in SEQ ID NO:2, and functionally equivalent polypeptides, including active fragments of the LBDS1 polypeptide, such as a fragment including the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide sequence, or a homologue thereof.

Nucleic acid molecules encompassing these stretches of sequence form a preferred embodiment of this aspect of the invention.

These nucleic acid molecules may be used in the methods and applications described herein. The nucleic acid molecules of the invention preferably comprise at least n consecutive nucleotides from the sequences disclosed herein where, depending on the particular sequence, n is 10 or more (for example, 12, 14, 15, 18, 20, 25, 30, 35, 40 or more).

The nucleic acid molecules of the invention also include sequences that are complementary to nucleic acid molecules described above (for example, for antisense or probing purposes).

Nucleic acid molecules of the present invention may be in the form of RNA, such as mRNA, or in the form of DNA, including, for instance cDNA, synthetic DNA or genomic DNA. Such nucleic acid molecules may be obtained by cloning, by chemical synthetic techniques or by a combination thereof. The nucleic acid molecules can be prepared, for example, by chemical synthesis using techniques such as solid phase phosphoramidite chemical synthesis, from genomic or cDNA libraries or by separation from an organism. RNA molecules may generally be generated by the *in vitro* or *in vivo* transcription of DNA sequences.

The nucleic acid molecules may be double-stranded or single-stranded. Single-stranded DNA may be the coding strand, also known as the sense strand, or it may be the non-coding strand, also referred to as the anti-sense strand.

The term "nucleic acid molecule" also includes analogues of DNA and RNA, such as those containing modified backbones, and peptide nucleic acids (PNA). The term "PNA", as used herein, refers to an antisense molecule or an anti-gene agent which comprises an oligonucleotide of at least five nucleotides in length linked to a peptide backbone of amino acid residues, which preferably ends in lysine. The terminal lysine confers solubility to the composition. PNAs may be pegylated to extend their lifespan in a cell, where they preferentially bind complementary single stranded DNA and RNA and stop transcript elongation (Nielsen, P.E. *et al.* (1993) Anticancer Drug Des. 8:53-63).

A nucleic acid molecule which encodes the polypeptide of SEQ ID NO:2, or an active fragment thereof, may be identical to the coding sequence of the nucleic acid molecule shown in SEQ ID NO:1. These molecules also may have a different sequence which, as a result of the degeneracy of the genetic code, encodes the polypeptide SEQ ID NO:2, or an active fragment of the LBDS1 polypeptide, such as a fragment including the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region, or a homologue thereof. The LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region is considered to extend between residue 1507 and residue 1742 of the LBDS1 polypeptide sequence. In SEQ ID NO:1 the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region is thus encoded by a nucleic acid molecule including nucleotide 4548 to 5255. Nucleic acid molecules encompassing this stretch of sequence, and homologues of this sequence, form a preferred embodiment of this aspect of the invention.

Such nucleic acid molecules that encode the polypeptide of SEQ ID NO:2 may include, but are not limited to, the coding sequence for the mature polypeptide by itself; the coding sequence for the mature polypeptide and additional coding sequences, such as those encoding a leader or secretory sequence, such as a pro-, pre- or prepro- polypeptide sequence; the coding sequence of the mature polypeptide, with or without the aforementioned additional coding sequences, together with further additional, non-coding sequences, including non-coding 5' and 3' sequences, such as the transcribed, non-translated sequences that play a role in transcription (including termination signals), ribosome binding and mRNA stability. The nucleic acid molecules may also include additional sequences which encode additional amino acids, such as those which provide additional functionalities.

The nucleic acid molecules of the second and third aspects of the invention may also encode the fragments or the functional equivalents of the polypeptides and fragments of the first aspect of the invention.

As discussed above, a preferred fragment of the LBDS1 polypeptide is a fragment including the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region, or a homologue thereof. The Nuclear Hormone Receptor Ligand Binding Domain region is encoded by, a nucleic acid molecule including nucleotide 4548 to 5255 of SEQ ID NO:1.

Functionally equivalent nucleic acid molecules according to the invention may be naturally-occurring variants such as a naturally-occurring allelic variant, or the molecules may be a variant that is not known to occur naturally. Such non-naturally occurring variants of the nucleic acid molecule may be made by mutagenesis techniques, including those applied to nucleic acid molecules, cells or organisms.

Among variants in this regard are variants that differ from the aforementioned nucleic acid molecules by nucleotide substitutions, deletions or insertions. The substitutions, deletions or insertions may involve one or more nucleotides. The variants may be altered in coding or non-coding regions or both. Alterations in the coding regions may produce conservative or non-conservative amino acid substitutions, deletions or insertions.

The nucleic acid molecules of the invention can also be engineered, using methods generally known in the art, for a variety of reasons, including modifying the cloning, processing, and/or expression of the gene product (the polypeptide). DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides are included as techniques which may be used to engineer the nucleotide sequences. Site-directed mutagenesis may be used to insert new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, introduce mutations and so forth.

Nucleic acid molecules which encode a polypeptide of the first aspect of the invention may be ligated to a heterologous sequence so that the combined nucleic acid molecule encodes a fusion protein. Such combined nucleic acid molecules are included within the second or third aspects of the invention. For example, to screen peptide libraries for inhibitors of the activity of the polypeptide, it may be useful to express, using such a combined nucleic acid molecule, a fusion protein that can be recognised by a commercially-available antibody. A fusion protein may also be engineered to contain a cleavage site located between the sequence of the polypeptide of the invention and the sequence of a heterologous protein so that the polypeptide may be cleaved and purified away from the heterologous protein.

The nucleic acid molecules of the invention also include antisense molecules that are partially complementary to nucleic acid molecules encoding polypeptides of the present invention and that therefore hybridize to the encoding nucleic acid molecules (hybridization). Such antisense molecules, such as oligonucleotides, can be designed to recognise, specifically bind to and prevent transcription of a target nucleic acid encoding a polypeptide of the invention, as will be known by those of ordinary skill in the art (see, for example, Cohen, J.S., Trends in Pharm. Sci., 10, 435 (1989), Okano, J. Neurochem. 56, 560 (1991); O'Connor, J. Neurochem 56, 560 (1991); Lee *et al.*, Nucleic Acids Res 6, 3073 (1979); Cooney *et al.*, Science 241, 456 (1988); Dervan *et al.*, Science 251, 1360 (1991).

The term "hybridization" as used here refers to the association of two nucleic acid molecules with one another by hydrogen bonding. Typically, one molecule will be fixed to a solid support and the other will be free in solution. Then, the two molecules may be placed in contact with one another under conditions that favour hydrogen bonding. Factors that affect this bonding include: the type and volume of solvent; reaction temperature; time of hybridization; agitation; agents to block the non-specific attachment of the liquid phase molecule to the solid support (Denhardt's reagent or BLOTTO); the concentration of the molecules; use of compounds to increase the rate of association of molecules (dextran sulphate or polyethylene glycol); and the stringency of the washing conditions following hybridization (see Sambrook *et al.* [*supra*]).

The inhibition of hybridization of a completely complementary molecule to a target molecule may be examined using a hybridization assay, as known in the art (see, for example, Sambrook *et al.* [*supra*]). A substantially homologous molecule will then compete for and inhibit the binding of a completely homologous molecule to the target molecule under various conditions of stringency, as taught in Wahl, G.M. and S.L. Berger (1987; Methods Enzymol. 152:399-407) and Kimmel, A.R. (1987; Methods Enzymol. 152:507-511).

"Stringency" refers to conditions in a hybridization reaction that favour the association of very similar molecules over association of molecules that differ. High stringency hybridisation conditions are defined as overnight incubation at 42°C in a solution comprising 50% formamide, 5XSSC (150mM NaCl, 15mM trisodium citrate), 50mM sodium phosphate (pH7.6), 5x Denhardts solution, 10% dextran sulphate, and 20 microgram/ml denatured, sheared salmon sperm DNA, followed by washing the filters in 0.1X SSC at approximately 65°C. Low stringency conditions involve the hybridisation reaction being carried out at 35°C (see Sambrook *et al.* [*supra*]). Preferably, the conditions used for hybridization are those of high stringency.

Preferred embodiments of this aspect of the invention are nucleic acid molecules that are at least 80% identical over their entire length to a nucleic acid molecule encoding the LBDS1 polypeptide (SEQ ID NO:2), and nucleic acid molecules that are substantially complementary to such nucleic acid molecules. A preferred active fragment is a fragment that includes an LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide sequences, resepctively. Accordingly, preferred nucleic acid molecules include those that are at least 80% identical over their entire length to a nucleic acid molecule encoding the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide sequence.

Percentage identity, as referred to herein, is as determined using BLAST version 2.1.3 using the default parameters specified by the NCBI (the National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/).

Preferably, a nucleic acid molecule according to this aspect of the invention comprises a region that is at least 80% identical over its entire length to the nucleic acid molecule having the sequence given in SEQ ID NO:1 to a region including nucleotides 4548-5255 of this sequence, or a nucleic acid molecule that is complementary to any one of these regions of nucleic acid. In this regard, nucleic acid molecules at least 90%, preferably at least 95%, more preferably at least 98% or 99% identical over their entire length to the same are particularly preferred. Preferred embodiments in this respect are nucleic acid molecules that encode polypeptides which retain substantially the same biological function or activity as the LBDS1 polypeptide.

The invention also provides a process for detecting a nucleic acid molecule of the invention, comprising the steps of: (a) contacting a nucleic probe according to the invention with a biological sample under hybridizing conditions to form duplexes; and (b) detecting any such duplexes that are formed.

As discussed additionally below in connection with assays that may be utilised according to the invention, a nucleic acid molecule as described above may be used as a hybridization probe for RNA, cDNA or genomic DNA, in order to isolate full-length cDNAs and genomic clones encoding the LBDS1 polypeptide and to isolate cDNA and genomic clones of homologous or orthologous genes that have a high sequence similarity to the gene encoding this polypeptide.

In this regard, the following techniques, among others known in the art, may be utilised and are discussed below for purposes of illustration. Methods for DNA sequencing and analysis are well known and are generally available in the art and may, indeed, be used to practice many of the embodiments of the invention discussed herein. Such methods may employ such enzymes as the Klenow fragment of DNA polymerase I, Sequenase (US Biochemical Corp, Cleveland, OH), Taq polymerase (Perkin Elmer), thermostable T7 polymerase (Amersham, Chicago, IL), or combinations of polymerases and proof-reading exonucleases such as those found in the ELONGASE Amplification System marketed by Gibco/BRL (Gaithersburg, MD). Preferably, the sequencing process may be automated using machines such as the Hamilton Micro Lab 2200 (Hamilton, Reno, NV), the Peltier Thermal Cycler (PTC200; MJ Research, Watertown, MA) and the ABI Catalyst and 373 and 377 DNA Sequencers (Perkin Elmer).

One method for isolating a nucleic acid molecule encoding a polypeptide with an equivalent function to that of the LBDS1 polypeptide, particularly with an equivalent function to the LBDS1 Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide, is to probe a genomic or cDNA library with a natural or artificially-designed probe using standard procedures that are recognised in the art (see, for example, "Current Protocols in Molecular Biology", Ausubel *et al.* (eds). Greene Publishing Association and John Wiley Interscience, New York, 1989,1992). Probes comprising at least 15, preferably at least 30, and more preferably at least 50, contiguous bases that correspond to, or are complementary to, nucleic acid sequences from the appropriate encoding gene (SEQ ID NO:1), particularly a region from nucleotides 4548-5255 of SEQ ID NO:1, are particularly useful probes.

Such probes may be labelled with an analytically-detectable reagent to facilitate their identification. Useful reagents include, but are not limited to, radioisotopes, fluorescent dyes and enzymes that are capable of catalysing the formation of a detectable product. Using these probes, the ordinarily skilled artisan will be capable of isolating complementary copies of genomic DNA, cDNA or RNA polynucleotides encoding proteins of interest from human, mammalian or other animal sources and screening such sources for related sequences, for example, for additional members of the family, type and/or subtype.

In many cases, isolated cDNA sequences will be incomplete, in that the region encoding the polypeptide will be cut short, normally at the 5' end. Several methods are available to obtain full length cDNAs, or to extend short cDNAs. Such sequences may be extended utilising a partial nucleotide sequence and employing various methods known in the art to detect upstream sequences such as promoters and regulatory elements. For example, one method which may be employed is based on the method of Rapid Amplification of cDNA Ends (RACE; see, for example, Frohman *et al.*, Proc. Natl. Acad. Sci. USA (1988) 85: 8998-9002). Recent modifications of this technique, exemplified by the Marathon^{TM} technology (Clontech Laboratories Inc.), for example, have significantly simplified the search for longer cDNAs. A slightly different technique, termed "restriction-site" PCR, uses universal primers to retrieve unknown nucleic acid sequence adjacent a known locus (Sarkar, G. (1993) PCR Methods Applic. 2:318-322). Inverse PCR may also be used to amplify or to extend sequences using divergent primers based on a known region (Triglia, T., *et al.* (1988) Nucleic Acids Res. 16:8186). Another method which may be used is capture PCR which involves PCR amplification of DNA fragments adjacent a known sequence in human and yeast artificial chromosome DNA (Lagerstrom, M. *et al.* (1991) PCR Methods Applic. 1: 111-119). Another method which may be used to retrieve unknown sequences is that of Parker, J.D. *et al.* (1991); Nucleic Acids Res. 19:3055-3060). Additionally, one may use PCR, nested primers, and PromoterFinder^{TM} libraries to walk genomic DNA (Clontech, Palo Alto, CA). This process avoids the need to screen libraries and is useful in finding intron/exon junctions.

When screening for full-length cDNAs, it is preferable to use libraries that have been size-selected to include larger cDNAs. Also, random-primed libraries are preferable, in that they will contain more sequences that contain the 5' regions of genes. Use of a randomly primed library may be especially preferable for situations in which an oligo d(T) library does not yield a full-length cDNA. Genomic libraries may be useful for extension of sequence into 5' non-transcribed regulatory regions.

In one embodiment of the invention, the nucleic acid molecules of the present invention may be used for chromosome localisation. In this technique, a nucleic acid molecule is specifically targeted to, and can hybridize with, a particular location on an individual human chromosome. The mapping of relevant sequences to chromosomes according to the present invention is an important step in the confirmatory correlation of those sequences with the gene-associated disease. Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. Such data are found in, for example, V. McKusick, Mendelian Inheritance in Man (available on-line through Johns Hopkins University Welch Medical Library). The relationships between genes and diseases that have been mapped to the same chromosomal region are then identified through linkage analysis (coinheritance of physically adjacent genes). This provides valuable information to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once the disease or syndrome has been crudely localised by genetic linkage to a particular genomic region, any sequences mapping to that area may represent associated or regulatory genes for further investigation. The nucleic acid molecule may also be used to detect differences in the chromosomal location due to translocation, inversion, etc. among normal, carrier, or affected individuals.

The nucleic acid molecules of the present invention are also valuable for tissue localisation. Such techniques allow the determination of expression patterns of the polypeptide in tissues by detection of the mRNAs that encode them. These techniques include in situ hybridization techniques and nucleotide amplification techniques, such as PCR. Results from these studies provide an indication of the normal functions of the polypeptide in the organism. In addition, comparative studies of the normal expression pattern of mRNAs with that of mRNAs encoded by a mutant gene provide valuable insights into the role of mutant polypeptides in disease. Such inappropriate expression may be of a temporal, spatial or quantitative nature.

The vectors of the present invention comprise nucleic acid molecules of the invention and may be cloning or expression vectors. The host cells of the invention, which may be transformed, transfested or transduced with the vectors of the invention may be prokaryotic or eukaryotic.

The polypeptides of the invention may be prepared in recombinant form by expression of their encoding nucleic acid molecules in vectors contained within a host cell. Such expression methods are well known to those of skill in the art and many are described in detail by Sambrook *et al.* (*supra*) and Fernandez & Hoeffler (1998, eds. "Gene expression systems. Using nature for the art of expression". Academic Press, San Diego, London, Boston, New York, Sydney, Tokyo, Toronto).

Generally, any system or vector that is suitable to maintain, propagate or express nucleic acid molecules to produce a polypeptide in the required host may be used. The appropriate nucleotide sequence may be inserted into an expression system by any of a variety of well-known and routine techniques, such as, for example, those described in Sambrook *et al.*, (*supra*). Generally, the encoding gene can be placed under the control of a control element such as a promoter, ribosome binding site (for bacterial expression) and, optionally, an operator, so that the DNA sequence encoding the desired polypeptide is transcribed into RNA in the transformed host cell.

Examples of suitable expression systems include, for example, chromosomal, episomal and virus-derived systems, including, for example, vectors derived from: bacterial plasmids, bacteriophage, transposons, yeast episomes, insertion elements, yeast chromosomal elements, viruses such as baculoviruses, papova viruses such as SV40, vaccinia viruses, adenoviruses, fowl pox viruses, pseudorabies viruses and retroviruses, or combinations thereof, such as those derived from plasmid and bacteriophage genetic elements, including cosmids and phagemids. Human artificial chromosomes (HACs) may also be employed to deliver larger fragments of DNA than can be contained and expressed in a plasmid.

Particularly suitable expression systems include microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with virus expression vectors (for example, baculovirus); plant cell systems transformed with virus expression vectors (for example, cauliflower mosaic virus, CaMV; tobacco mosaic virus, TMV) or with bacterial expression vectors (for example, Ti or pBR322 plasmids); or animal cell systems. Cell-free translation systems can also be employed to produce the polypeptides of the invention.

Introduction of nucleic acid molecules encoding a polypeptide of the present invention into host cells can be effected by methods described in many standard laboratory manuals, such as Davis *et al.*, Basic Methods in Molecular Biology (1986) and Sambrook *et al.*, (*supra*). Particularly suitable methods include calcium phosphate transfection, DEAE-dextran mediated transfection, transvection, microinjection, cationic lipid-mediated transfection, electroporation, transduction, scrape loading, ballistic introduction or infection (see Sambrook *et al.*, 1989 [*supra*]; Ausubel *et al.*, 1991 [*supra*]; Spector, Goldman & Leinwald, 1998). In eukaryotic cells, expression systems may either be transient (for example, episomal) or permanent (chromosomal integration) according to the needs of the system.

The encoding nucleic acid molecule may or may not include a sequence encoding a control sequence, such as a signal peptide or leader sequence, as desired, for example, for secretion of the translated polypeptide into the lumen of the endoplasmic reticulum, into the periplasmic space or into the extracellular environment. These signals may be endogenous to the polypeptide or they may be heterologous signals. Leader sequences can be removed by the bacterial host in post-translational processing.

In addition to control sequences, it may be desirable to add regulatory sequences that allow for regulation of the expression of the polypeptide relative to the growth of the host cell. Examples of regulatory sequences are those which cause the expression of a gene to be increased or decreased in response to a chemical or physical stimulus, including the presence of a regulatory compound or to various temperature or metabolic conditions.

Regulatory sequences are those non-translated regions of the vector, such as enhancers, promoters and 5' and 3' untranslated regions. These interact with host cellular proteins to carry out transcription and translation. Such regulatory sequences may vary in their strength and specificity. Depending on the vector system and host utilised, any number of suitable transcription and translation elements, including constitutive and inducible promoters, may be used. For example, when cloning in bacterial systems, inducible promoters such as the hybrid lacZ promoter of the Bluescript phagemid (Stratagene, LaJolla, CA) or pSportl^{TM} plasmid (Gibco BRL) and the like may be used. The baculovirus polyhedrin promoter may be used in insect cells. Promoters or enhancers derived from the genomes of plant cells (for example, heat shock, RUBISCO and storage protein genes) or from plant viruses (for example, viral promoters or leader sequences) may be cloned into the vector. In mammalian cell systems, promoters from mammalian genes or from mammalian viruses are preferable. If it is necessary to generate a cell line that contains multiple copies of the sequence, vectors based on SV40 or EBV may be used with an appropriate selectable marker.

An expression vector is constructed so that the particular nucleic acid coding sequence is located in the vector with the appropriate regulatory sequences, the positioning and orientation of the coding sequence with respect to the regulatory sequences being such that the coding sequence is transcribed under the "control" of the regulatory sequences, i.e., RNA polymerase which binds to the DNA molecule at the control sequences transcribes the coding sequence. In some cases it may be necessary to modify the sequence so that it may be attached to the control sequences with the appropriate orientation; i.e., to maintain the reading frame.

The control sequences and other regulatory sequences may be ligated to the nucleic acid coding sequence prior to insertion into a vector. Alternatively, the coding sequence can be cloned directly into an expression vector that already contains the control sequences and an appropriate restriction site.

For long-term, high-yield production of a recombinant polypeptide, stable expression is preferred. For example, cell lines which stably express the polypeptide of interest may be transformed using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Following the introduction of the vector, cells may be allowed to grow for 1-2 days in an enriched media before they are switched to selective media. The purpose of the selectable marker is to confer resistance to selection, and its presence allows growth and recovery of cells that successfully express the introduced sequences. Resistant clones of stably transformed cells may be proliferated using tissue culture techniques appropriate to the cell type.

Mammalian cell lines available as hosts for expression are known in the art and include many immortalised cell lines available from the American Type Culture Collection (ATCC) including, but not limited to, Chinese hamster ovary (CHO), HeLa, baby hamster kidney (BHK), monkey kidney (COS), C127, 3T3, BHK, HEK 293, Bowes melanoma and human hepatocellular carcinoma (for example Hep G2) cells and a number of other cell lines.

In the baculovirus system, the materials for baculovirus/insect cell expression systems are commercially available in kit form from, inter alia, Invitrogen, San Diego CA (the "MaxBac" kit). These techniques are generally known to those skilled in the art and are described fully in Summers and Smith, Texas Agricultural Experiment Station Bulletin No. 1555 (1987). Particularly suitable host cells for use in this system include insect cells such as Drosophila S2 and Spodoptera Sf9 cells.

There are many plant cell culture and whole plant genetic expression systems known in the art. Examples of suitable plant cellular genetic expression systems include those described in US 5,693,506; US 5,659,122; and US 5,608,143. Additional examples of genetic expression in plant cell culture has been described by Zenk, (1991) Phytochemistry 30, 3861-3863.

In particular, all plants from which protoplasts can be isolated and cultured to give whole regenerated plants can be utilised, so that whole plants are recovered which contain the transferred gene. Practically all plants can be regenerated from cultured cells or tissues, including but not limited to all major species of sugar cane, sugar beet, cotton, fruit and other trees, legumes and vegetables.

Examples of particularly preferred bacterial host cells include streptococci, staphylococci, *E. coli*, Streptomyces and *Bacillus subtilis* cells.

Examples of particularly suitable host cells for fungal expression include yeast cells (for example, *S. cerevisiae*) and Aspergillus cells.

Any number of selection systems are known in the art that may be used to recover transformed cell lines. Examples include the herpes simplex virus thymidine kinase (Wigler, M. *et al.* (1977) Cell 11:223-32) and adenine phosphoribosyltransferase (Lowy, I. *et al.* (1980) Cell 22:817-23) genes that can be employed in tk- or aprt± cells, respectively.

Also, antimetabolite, antibiotic or herbicide resistance can be used as the basis for selection; for example, dihydrofolate reductase (DHFR) that confers resistance to methotrexate (Wigler, M. *et al.* (1980) Proc. Natl. Acad. Sci. 77:3567-70); npt, which confers resistance to the aminoglycosides neomycin and G-418 (Colbere-Garapin, F. *et al.* (1981) J. Mol. Biol. 150:1-14) and als or pat, which confer resistance to chlorsulfuron and phosphinotricin acetyltransferase, respectively. Additional selectable genes have been described, examples of which will be clear to those of skill in the art.

Although the presence or absence of marker gene expression suggests that the gene of interest is also present, its presence and expression may need to be confirmed. For example, if the relevant sequence is inserted within a marker gene sequence, transformed cells containing the appropriate sequences can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with a sequence encoding a polypeptide of the invention under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the tandem gene as well.

Alternatively, host cells that contain a nucleic acid sequence encoding a polypeptide of the invention and which express said polypeptide may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations and protein bioassays, for example, fluorescence activated cell sorting (FACS) or immunoassay techniques (such as the enzyme-linked immunosorbent assay [ELISA] and radioimmunoassay [RIA]), that include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein (see Hampton, R. *et al.* (1990) Serological Methods, a Laboratory Manual, APS Press, St Paul, MN) and Maddox, D.E. *et al.* (1983) J. Exp. Med, 158, 1211-1216).

A wide variety of labels and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid and amino acid assays. Means for producing labelled hybridization or PCR probes for detecting sequences related to nucleic acid molecules encoding polypeptides of the present invention include oligolabelling, nick translation, end-labelling or PCR amplification using a labelled polynucleotide. Alternatively, the sequences encoding the polypeptide of the invention may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesise RNA probes *in vitro* by addition of an appropriate RNA polymerase such as T7, T3 or SP6 and labelled nucleotides. These procedures may be conducted using a variety of commercially available kits (Pharmacia & Upjohn, (Kalamazoo, MI); Promega (Madison WI); and U.S. Biochemical Corp., Cleveland, OH)).

Suitable reporter molecules or labels, which may be used for ease of detection, include radionuclides, enzymes and fluorescent, chemiluminescent or chromogenic agents as well as substrates, cofactors, inhibitors, magnetic particles, and the like.

Nucleic acid molecules according to the present invention may also be used to create transgenic animals, particularly rodent animals. Such transgenic animals form a further aspect of the present invention. This may be done locally by modification of somatic cells, or by germ line therapy to incorporate heritable modifications. Such transgenic animals may be particularly useful in the generation of animal models for drug molecules effective as modulators of the polypeptides of the present invention.

The polypeptide can be recovered and purified from recombinant cell cultures by well-known methods including ammonium sulphate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. High performance liquid chromatography is particularly useful for purification. Well known techniques for refolding proteins may be employed to regenerate an active conformation when the polypeptide is denatured during isolation and or purification.

Specialised vector constructions may also be used to facilitate purification of proteins, as desired, by joining sequences encoding the polypeptides of the invention to a nucleotide sequence encoding a polypeptide domain that will facilitate purification of soluble proteins. Examples of such purification-facilitating domains include metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilised metals, protein A domains that allow purification on immobilised immunoglobulin, and the domain utilised in the FLAGS extension/affinity purification system (Immunex Corp., Seattle, WA). The inclusion of cleavable linker sequences such as those specific for Factor XA or enterokinase (Invitrogen, San Diego, CA) between the purification domain and the polypeptide of the invention may be used to facilitate purification. One such expression vector provides for expression of a fusion protein containing the polypeptide of the invention fused to several histidine residues preceding a thioredoxin or an enterokinase cleavage site. The histidine residues facilitate purification by IMAC (immobilised metal ion affinity chromatography as described in Porath, J. *et al.* (1992) Prot. Exp. Purif. 3: 263-281) while the thioredoxin or enterokinase cleavage site provides a means for purifying the polypeptide from the fusion protein. A discussion of vectors which contain fusion proteins is provided in Kroll, D.J. *et al.* (DNA Cell Biol. 199312:441-453).

If the polypeptide is to be expressed for use in screening assays, generally it is preferred that it be produced at the surface of the host cell in which it is expressed. In this event, the host cells may be harvested prior to use in the screening assay, for example using techniques such as fluorescence activated cell sorting (FACS) or immunoaffinity techniques. If the polypeptide is secreted into the medium, the medium can be recovered in order to recover and purify the expressed polypeptide. If polypeptide is produced intracellularly, the cells must first be lysed before the polypeptide is recovered.

The polypeptide of the invention can be used to screen libraries of compounds in any of a variety of drug screening techniques. Such compounds may activate (agonise) or inhibit (antagonise) the level of expression of the gene or the activity of the polypeptide of the invention and form a further aspect of the present invention. Preferred compounds are effective to alter the expression of a natural gene which encodes a polypeptide of the first aspect of the invention or to regulate the activity of a polypeptide of the first aspect of the invention.

Agonist or antagonist compounds may be isolated from, for example, cells, cell-free preparations, chemical libraries or natural product mixtures. These agonists or antagonists may be natural or modified substrates, ligands, enzymes, receptors or structural or functional mimetics. For a suitable review of such screening techniques, see Coligan *et al.*, Current Protocols in Immunology 1(2):Chapter 5 (1991).

Compounds that are most likely to be good antagonists are molecules that bind to the polypeptide of the invention without inducing the biological effects of the polypeptide upon binding to it. Potential antagonists include small organic molecules, peptides, polypeptides and antibodies that bind to the polypeptide of the invention and thereby inhibit or extinguish its activity. In this fashion, binding of the polypeptide to normal cellular binding molecules may be inhibited, such that the normal biological activity of the polypeptide is prevented.

The polypeptide of the invention that is employed in such a screening technique may be free in solution, affixed to a solid support, borne on a cell surface or located intracellularly. In general, such screening procedures may involve using appropriate cells or cell membranes that express the polypeptide that are contacted with a test compound to observe binding, or stimulation or inhibition of a functional response. The functional response of the cells contacted with the test compound is then compared with control cells that were not contacted with the test compound. Such an assay may assess whether the test compound results in a signal generated by activation of the polypeptide, using an appropriate detection system. Inhibitors of activation are generally assayed in the presence of a known agonist and the effect on activation by the agonist in the presence of the test compound is observed.

Alternatively, simple binding assays may be used, in which the adherence of a test compound to a surface bearing the polypeptide is detected by means of a label directly or indirectly associated with the test compound or in an assay involving competition with a labelled competitor. In another embodiment, competitive drug screening assays may be used, in which neutralising antibodies that are capable of binding the polypeptide specifically compete with a test compound for binding. In this manner, the antibodies can be used to detect the presence of any test compound that possesses specific binding affinity for the polypeptide.

Assays may also be designed to detect the effect of added test compounds on the production of mRNA encoding the polypeptide in cells. For example, an ELISA may be constructed that measures secreted or cell-associated levels of polypeptide using monoclonal or polyclonal antibodies by standard methods known in the art, and this can be used to search for compounds that may inhibit or enhance the production of the polypeptide from suitably manipulated cells or tissues. The formation of binding complexes between the polypeptide and the compound being tested may then be measured.

Another technique for drug screening which may be used provides for high throughput screening of compounds having suitable binding affinity to the polypeptide of interest (see International patent application WO84/03564). In this method, large numbers of different small test compounds are synthesised on a solid substrate, which may then be reacted with the polypeptide of the invention and washed. One way of immobilising the polypeptide is to use non-neutralising antibodies. Bound polypeptide may then be detected using methods that are well known in the art. Purified polypeptide can also be coated directly onto plates for use in the aforementioned drug screening techniques.

The polypeptide of the invention may be used to identify membrane-bound or soluble receptors, through standard receptor binding techniques that are known in the art, such as ligand binding and crosslinking assays in which the polypeptide is labelled with a radioactive isotope, is chemically modified, or is fused to a peptide sequence that facilitates its detection or purification, and incubated with a source of the putative receptor (for example, a composition of cells, cell membranes, cell supernatants, tissue extracts, or bodily fluids). The efficacy of binding may be measured using biophysical techniques such as surface plasmon resonance and spectroscopy. Binding assays may be used for the purification and cloning of the receptor, but may also identify agonists and antagonists of the polypeptide, that compete with the binding of the polypeptide to its receptor. Standard methods for conducting screening assays are well understood in the art.

The invention also includes a screening kit useful in the methods for identifying agonists, antagonists, ligands, receptors, substrates, and/or enzymes that are described above.

The invention includes the agonists, antagonists, ligands, receptors, substrates and enzymes, and other compounds which modulate the activity or antigenicity of the polypeptide of the invention discovered by the methods that are described above.

The invention also provides pharmaceutical compositions comprising a polypeptide, nucleic acid, ligand or compound of the invention in combination with a suitable pharmaceutical carrier. These compositions may be suitable as therapeutic or diagnostic reagents, as vaccines, or as other immunogenic compositions, as outlined in detail below.

According to the terminology used herein, a composition containing a polypeptide, nucleic acid, ligand or compound [X] is "substantially free of" impurities [herein, Y] when at least 85% by weight of the total X+Y in the composition is X. Preferably, X comprises at least about 90% by weight of the total of X+Y in the composition, more preferably at least about 95%, 98% or even 99% by weight.

The pharmaceutical compositions should preferably comprise a therapeutically effective amount of the polypeptide, nucleic acid molecule, ligand, or compound of the invention. The term "therapeutically effective amount" as used herein refers to an amount of a therapeutic agent needed to treat, ameliorate, or prevent a targetted disease or condition, or to exhibit a detectable therapeutic or preventative effect. For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays, for example, of neoplastic cells, or in animal models, usually mice, rabbits, dogs, or pigs. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

The precise effective amount for a human subject will depend upon the severity of the disease state, general health of the subject, age, weight, and gender of the subject, diet, time and frequency of administration, drug combination(s), reaction sensitivities, and tolerance/response to therapy. This amount can be determined by routine experimentation and is within the judgement of the clinician. Generally, an effective dose will be from 0.01 mg/kg to 50 mg/kg, preferably 0.05 mg/kg to 10 mg/kg. Compositions may be administered individually to a patient or may be administered in combination with other agents, drugs or hormones.

A pharmaceutical composition may also contain a pharmaceutically acceptable carrier, for administration of a therapeutic agent. Such carriers include antibodies and other polypeptides, genes and other therapeutic agents such as liposomes, provided that the carrier does not itself induce the production of antibodies harmful to the individual receiving the composition, and which may be administered without undue toxicity. Suitable carriers may be large, slowly metabolised macromolecules such as proteins, polysaccharides, polylactic acids, polyglycolic acids, polymeric amino acids, amino acid copolymers and inactive virus particles.

Pharmaceutically acceptable salts can be used therein, for example, mineral acid salts such as hydrochlorides, hydrobromides, phosphates, sulphates, and the like; and the salts of organic acids such as acetates, propionates, malonates, benzoates, and the like. A thorough discussion of pharmaceutically acceptable carriers is available in Remington's Pharmaceutical Sciences (Mack Pub. Co., N.J. 1991).

Pharmaceutically acceptable carriers in therapeutic compositions may additionally contain liquids such as water, saline, glycerol and ethanol. Additionally, auxiliary substances, such as wetting or emulsifying agents, pH buffering substances, and the like, may be present in such compositions. Such carriers enable the pharmaceutical compositions to be formulated as tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, suspensions, and the like, for ingestion by the patient.

Once formulated, the compositions of the invention can be administered directly to the subject. The subjects to be treated can be animals; in particular, human subjects can be treated.

The pharmaceutical compositions utilised in this invention may be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intraarterial, intramedullary, intrathecal, intraventricular, transdermal or transcutaneous applications (for example, see WO98/20734), subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, intravaginal or rectal means. Gene guns or hyposprays may also be used to administer the pharmaceutical compositions of the invention. Typically, the therapeutic compositions may be prepared as injectables, either as liquid solutions or suspensions; solid forms suitable for solution in, or suspension in, liquid vehicles prior to injection may also be prepared.

Direct delivery of the compositions will generally be accomplished by injection, subcutaneously, intraperitoneally, intravenously or intramuscularly, or delivered to the interstitial space of a tissue. The compositions can also be administered into a lesion. Dosage treatment may be a single dose schedule or a multiple dose schedule.

If the activity of the polypeptide of the invention is in excess in a particular disease state, several approaches are available. One approach comprises administering to a subject an inhibitor compound (antagonist) as described above, along with a pharmaceutically acceptable carrier in an amount effective to inhibit the function of the polypeptide, such as by blocking the binding of ligands, substrates, enzymes, receptors, or by inhibiting a second signal, and thereby alleviating the abnormal condition. Preferably, such antagonists are antibodies. Most preferably, such antibodies are chimeric and/or humanised to minimise their immunogenicity, as described previously.

In another approach, soluble forms of the polypeptide that retain binding affinity for the ligand, substrate, enzyme, receptor, in question, may be administered. Typically, the polypeptide may be administered in the form of fragments that retain the relevant portions.

In an alternative approach, expression of the gene encoding the polypeptide can be inhibited using expression blocking techniques, such as the use of antisense nucleic acid molecules (as described above), either internally generated or separately administered. Modifications of gene expression can be obtained by designing complementary sequences or antisense molecules (DNA, RNA, or PNA) to the control, 5' or regulatory regions (signal sequence, promoters, enhancers and introns) of the gene encoding the polypeptide. Similarly, inhibition can be achieved using "triple helix" base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature (Gee, J.E. *et al.* (1994) In: Huber, B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, NY). The complementary sequence or antisense molecule may also be designed to block translation of mRNA by preventing the transcript from binding to ribosomes. Such oligonucleotides may be administered or may be generated in situ from expression *in vivo*.

In addition, expression of the polypeptide of the invention may be prevented by using ribozymes specific to its encoding mRNA sequence. Ribozymes are catalytically active RNAs that can be natural or synthetic (see for example Usman, N, *et al.*, Curr. Opin. Struct. Biol (1996) 6(4), 527-33). Synthetic ribozymes can be designed to specifically cleave mRNAs at selected positions thereby preventing translation of the mRNAs into functional polypeptide. Ribozymes may be synthesised with a natural ribose phosphate backbone and natural bases, as normally found in RNA molecules. Alternatively the ribozymes may be synthesised with non-natural backbones, for example, 2'-O-methyl RNA, to provide protection from ribonuclease degradation and may contain modified bases.

RNA molecules may be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at the 5' and/or 3' ends of the molecule or the use of phosphorothioate or 2' O-methyl rather than phosphodiesterase linkages within the backbone of the molecule. This concept is inherent in the production of PNAs and can be extended in all of these molecules by the inclusion of non-traditional bases such as inosine, queosine and butosine, as well as acetyl-, methyl-, thio- and similarly modified forms of adenine, cytidine, guanine, thymine and uridine which are not as easily recognised by endogenous endonucleases.

For treating abnormal conditions related to an under-expression of the polypeptide of the invention and its activity, several approaches are also available. One approach comprises administering to a subject a therapeutically effective amount of a compound that activates the polypeptide, i.e., an agonist as described above, to alleviate the abnormal condition.

Alternatively, a therapeutic amount of the polypeptide in combination with a suitable pharmaceutical carrier may be administered to restore the relevant physiological balance of polypeptide.

Gene therapy may be employed to effect the endogenous production of the polypeptide by the relevant cells in the subject. Gene therapy is used to treat permanently the inappropriate production of the polypeptide by replacing a defective gene with a corrected therapeutic gene.

Gene therapy of the present invention can occur *in vivo* or *ex vivo*. *Ex vivo* gene therapy requires the isolation and purification of patient cells, the introduction of a therapeutic gene and introduction of the genetically altered cells back into the patient. In contrast, *in vivo* gene therapy does not require isolation and purification of a patient's cells.

The therapeutic gene is typically "packaged" for administration to a patient. Gene delivery vehicles may be non-viral, such as liposomes, or replication-deficient viruses, such as adenovirus as described by Berkner, K.L., in Curr. Top. Microbiol. Immunol., 158, 39-66 (1992) or adeno-associated virus (AAV) vectors as described by Muzyczka, N., in Curr. Top. Microbiol. Immunol., 158, 97-129 (1992) and U.S. Patent No. 5,252,479. For example, a nucleic acid molecule encoding a polypeptide of the invention may be engineered for expression in a replication-defective retroviral vector. This expression construct may then be isolated and introduced into a packaging cell transduced with a retroviral plasmid vector containing RNA encoding the polypeptide, such that the packaging cell now produces infectious viral particles containing the gene of interest. These producer cells may be administered to a subject for engineering *cells in vivo* and expression of the polypeptide *in vivo* (see Chapter 20, Gene Therapy and other Molecular Genetic-based Therapeutic Approaches, (and references cited therein) in Human Molecular Genetics (1996), T Strachan and A P Read, BIOS Scientific Publishers Ltd).

Another approach is the administration of "naked DNA" in which the therapeutic gene is directly injected into the bloodstream or muscle tissue.

In situations in which the polypeptides or nucleic acid molecules of the invention are disease-causing agents, the invention provides that they can be used in vaccines to raise antibodies against the disease causing agent.

Vaccines according to the invention may either be prophylactic (ie. to prevent infection) or therapeutic (ie. to treat disease after infection). Such vaccines comprise immunising antigen(s), immunogen(s), polypeptide(s), protein(s) or nucleic acid, usually in combination with pharmaceutically-acceptable carriers as described above, which include any carrier that does not itself induce the production of antibodies harmful to the individual receiving the composition. Additionally, these carriers may function as immunostimulating agents ("adjuvants"). Furthermore, the antigen or immunogen may be conjugated to a bacterial toxoid, such as a toxoid from diphtheria, tetanus, cholera, *H. pylori*, and other pathogens.

Since polypeptides may be broken down in the stomach, vaccines comprising polypeptides are preferably administered parenterally (for instance, subcutaneous, intramuscular, intravenous, or intradermal injection). Formulations suitable for parenteral administration include aqueous and non-aqueous sterile injection solutions which may contain anti-oxidants, buffers, bacteriostats and solutes which render the formulation isotonic with the blood of the recipient, and aqueous and non-aqueous sterile suspensions which may include suspending agents or thickening agents.

The vaccine formulations of the invention may be presented in unit-dose or multi-dose containers. For example, sealed ampoules and vials and may be stored in a freeze-dried condition requiring only the addition of the sterile liquid carrier immediately prior to use. The dosage will depend on the specific activity of the vaccine and can be readily determined by routine experimentation.

This invention also relates to the use of nucleic acid molecules according to the present invention as diagnostic reagents. Detection of a mutated form of the gene characterised by the nucleic acid molecules of the invention which is associated with a dysfunction will provide a diagnostic tool that can add to, or define, a diagnosis of a disease, or susceptibility to a disease, which results from under-expression, over-expression or altered spatial or temporal expression of the gene. Individuals carrying mutations in the gene may be detected at the DNA level by a variety of techniques.

Nucleic acid molecules for diagnosis may be obtained from a subject's cells, such as from blood, urine, saliva, tissue biopsy or autopsy material. The genomic DNA may be used directly for detection or may be amplified enzymatically by using PCR, ligase chain reaction (LCR), strand displacement amplification (SDA), or other amplification techniques (see Saiki *et al.*, Nature, 324, 163-166 (1986); Bej, *et al.*, Crit. Rev. Biochem. Molec. Biol., 26, 301-334 (1991); Birkenmeyer *et al.*, J. Virol. Meth., 35, 117-126 (1991); Van Brunt, J., Bio/Technology, 8, 291-294 (1990)) prior to analysis.

In one embodiment, this aspect of the invention provides a method of diagnosing a disease in a patient, comprising assessing the level of expression of a natural gene encoding a polypeptide according to the invention and comparing said level of expression to a control level, wherein a level that is different to said control level is indicative of disease. The method may comprise the steps of:

a) contacting a sample of tissue from the patient with a nucleic acid probe under stringent conditions that allow the formation of a hybrid complex between a nucleic acid molecule of the invention and the probe;

b) contacting a control sample with said probe under the same conditions used in step a);

c) and detecting the presence of hybrid complexes in said samples;

wherein detection of levels of the hybrid complex in the patient sample that differ from levels of the hybrid complex in the control sample is indicative of disease.

A further aspect of the invention comprises a diagnostic method comprising the steps of:

a) obtaining a tissue sample from a patient being tested for disease;

b) isolating a nucleic acid molecule according to the invention from said tissue sample; and,

c) diagnosing the patient for disease by detecting the presence of a mutation in the nucleic acid molecule which is associated with disease.

To aid the detection of nucleic acid molecules in the above-described methods, an amplification step, for example using PCR, may be included.

Deletions and insertions can be detected by a change in the size of the amplified product in comparison to the normal genotype. Point mutations can be identified by hybridizing amplified DNA to labelled RNA of the invention or alternatively, labelled antisense DNA sequences of the invention. Perfectly-matched sequences can be distinguished from mismatched duplexes by RNase digestion or by assessing differences in melting temperatures. The presence or absence of the mutation in the patient may be detected by contacting DNA with a nucleic acid probe that hybridises to the DNA under stringent conditions to form a hybrid double-stranded molecule, the hybrid double-stranded molecule having an unhybridised portion of the nucleic acid probe strand at any portion corresponding to a mutation associated with disease; and detecting the presence or absence of an unhybridised portion of the probe strand as an indication of the presence or absence of a disease-associated mutation in the corresponding portion of the DNA strand.

Such diagnostics are particularly useful for prenatal and even neonatal testing.

Point mutations and other sequence differences between the reference gene and "mutant" genes can be identified by other well-known techniques, such as direct DNA sequencing or single-strand conformational polymorphism, (see Orita *et al.*, Genomics, 5, 874-879 (1989)). For example, a sequencing primer may be used with double-stranded PCR product or a single-stranded template molecule generated by a modified PCR. The sequence determination is performed by conventional procedures with radiolabelled nucleotides or by automatic sequencing procedures with fluorescent-tags. Cloned DNA segments may also be used as probes to detect specific DNA segments. The sensitivity of this method is greatly enhanced when combined with PCR. Further, point mutations and other sequence variations, such as polymorphisms, can be detected as described above, for example, through the use of allele-specific oligonucleotides for PCR amplification of sequences that differ by single nucleotides.

DNA sequence differences may also be detected by alterations in the electrophoretic mobility of DNA fragments in gels, with or without denaturing agents, or by direct DNA sequencing (for example, Myers *et al.*, Science (1985) 230:1242). Sequence changes at specific locations may also be revealed by nuclease protection assays, such as RNase and S1 protection or the chemical cleavage method (see Cotton *et al.*, Proc. Natl. Acad. Sci. USA (1985) 85: 4397-4401).

In addition to conventional gel electrophoresis and DNA sequencing, mutations such as microdeletions, aneuploidies, translocations, inversions, can also be detected by in situ analysis (see, for example, Keller *et al.*, DNA Probes, 2nd Ed., Stockton Press, New York, N.Y., USA (1993)), that is, DNA or RNA sequences in cells can be analysed for mutations without need for their isolation and/or immobilisation onto a membrane. Fluorescence in situ hybridization (FISH) is presently the most commonly applied method and numerous reviews of FISH have appeared (see, for example, Trachuck *et al.*, Science, 250: 559-562 (1990), and Trask *et al.*, Trends, Genet. 7:149-154 (1991)).

In another embodiment of the invention, an array of oligonucleotide probes comprising a nucleic acid molecule according to the invention can be constructed to conduct efficient screening of genetic variants, mutations and polymorphisms. Array technology methods are well known and have general applicability and can be used to address a variety of questions in molecular genetics including gene expression, genetic linkage, and genetic variability (see for example: M.Chee *et al.*, Science (1996) 274: 610-613).

In one embodiment, the array is prepared and used according to the methods described in PCT application WO95/11995 (Chee *et al*); Lockhart, D. J. *et al.* (1996) Nat. Biotech. 14: 1675-1680); and Schena, M. *et al.* (1996) Proc. Natl. Acad. Sci. 93: 10614-10619). Oligonucleotide pairs may range from two to over one million. The oligomers are synthesized at designated areas on a substrate using a light-directed chemical process. The substrate may be paper, nylon or other type of membrane, filter, chip, glass slide or any other suitable solid support. In another aspect, an oligonucleotide may be synthesized on the surface of the substrate by using a chemical coupling procedure and an ink jet application apparatus, as described in PCT application WO95/251116 (Baldeschweiler *et al.*). In another aspect, a "gridded" array analogous to a dot (or slot) blot may be used to arrange and link cDNA fragments or oligonucleotides to the surface of a substrate using a vacuum system, thermal, UV, mechanical or chemical bonding procedures. An array, such as those described above, may be produced by hand or by using available devices (slot blot or dot blot apparatus), materials (any suitable solid support), and machines (including robotic instruments), and may contain 8, 24, 96, 384, 1536 or 6144 oligonucleotides, or any other number between two and over one million which lends itself to the efficient use of commercially-available instrumentation.

In addition to the methods discussed above, diseases may be diagnosed by methods comprising determining, from a sample derived from a subject, an abnormally decreased or increased level of polypeptide or mRNA. Decreased or increased expression can be measured at the RNA level using any of the methods well known in the art for the quantitation of polynucleotides, such as, for example, nucleic acid amplification, for instance PCR, RT-PCR, RNase protection, Northern blotting and other hybridization methods.

Assay techniques that can be used to determine levels of a polypeptide of the present invention in a sample derived from a host are well-known to those of skill in the art and are discussed in some detail above (including radioimmunoassays, competitive-binding assays, Western Blot analysis and ELISA assays). This aspect of the invention provides a diagnostic method which comprises the steps of: (a) contacting a ligand as described above with a biological sample under conditions suitable for the formation of a ligand-polypeptide complex; and (b) detecting said complex.

Protocols such as ELISA, RIA, and FACS for measuring polypeptide levels may additionally provide a basis for diagnosing altered or abnormal levels of polypeptide expression. Normal or standard values for polypeptide expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, preferably humans, with antibody to the polypeptide under conditions suitable for complex formation The amount of standard complex formation may be quantified by various methods, such as by photometric means.

Antibodies which specifically bind to a polypeptide of the invention may be used for the diagnosis of conditions or diseases characterised by expression of the polypeptide, or in assays to monitor patients being treated with the polypeptides, nucleic acid molecules, ligands and other compounds of the invention. Antibodies useful for diagnostic purposes may be prepared in the same manner as those described above for therapeutics. Diagnostic assays for the polypeptide include methods that utilise the antibody and a label to detect the polypeptide in human body fluids or extracts of cells or tissues. The antibodies may be used with or without modification, and may be labelled by joining them, either covalently or non-covalently, with a reporter molecule. A wide variety of reporter molecules known in the art may be used, several of which are described above.

Quantities of polypeptide expressed in subject, control and disease samples from biopsied tissues are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease. Diagnostic assays may be used to distinguish between absence, presence, and excess expression of polypeptide and to monitor regulation of polypeptide levels during therapeutic intervention. Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials or in monitoring the treatment of an individual patient.

A diagnostic kit of the present invention may comprise:

(a) a nucleic acid molecule of the present invention;

(b) a polypeptide of the present invention; or

(c) a ligand of the present invention.

In one aspect of the invention, a diagnostic kit may comprise a first container containing a nucleic acid probe that hybridises under stringent conditions with a nucleic acid molecule according to the invention; a second container containing primers useful for amplifying the nucleic acid molecule; and instructions for using the probe and primers for facilitating the diagnosis of disease. The kit may further comprise a third container holding an agent for digesting unhybridised RNA.

In an alternative aspect of the invention, a diagnostic kit may comprise an array of nucleic acid molecules, at least one of which may be a nucleic acid molecule according to the invention.

To detect polypeptide according to the invention, a diagnostic kit may comprise one or more antibodies that bind to a polypeptide according to the invention; and a reagent useful for the detection of a binding reaction between the antibody and the polypeptide.

Such kits will be of use in diagnosing a disease or susceptibility to disease, particularly cell proliferative disorders, including neoplasm, cancer and myeloproliferative disorders and angiogenesis, metabolic disorders, including obesity, osteoporosis, non-insulin dependent diabetes, lipid metabolism disorder, hyperthyroidism, hyperparathyroidism, hypercalcemia, hypercholestrolemia, hyperlipidemia, cardiovascular disorders including, hypertension, athersclerosis, hyperinsulinemia, heart arrhythmia, renal disorders, including glomerulonephritis, renovascular hypertension, skin disorders including dermatological disease, psoriasis, acne, aging, eczema, wound healing, inflammation, including inflammatory bowel disease, respiratory tract inflammation, emphysema, asthma, immune disorder, autoimmune disease, multiple sclerosis, nervous system disorders, including anxiety, depression, neurodegenerative disease, Alzheimer's disease Parkinson's disease, brain injury, stroke and pain, infection, including virus infection, and other conditions in which nuclear hormones are implicated.

Various aspects and embodiments of the present invention will now be described in more detail by way of example, with particular reference to the LBDS1 polypeptide.

It will be appreciated that modification of detail may be made without departing from the scope of the invention.

Brief description of the Figures

Figure 1: This is the front end of the Biopendium Target Mining Interface. A search of the database is initiated using the PDB code "3LBD".

Figure 2A: A selection is shown of the Inpharmatica Genome Threader results for the search using 3LBD. The arrow indicates *Homo Sapiens* Retinoic Acid Receptor Gamma-1, which has a typical Nuclear Hormone Receptor Ligand Binding Domain.

Figure 2B: A selection is shown of the Inpharmatica Genome Threader results for the search using 3LBD. The arrow indicates BAA07892.2 (LBDS1).

Figure 2C: Full list of forward PSI-BLAST results for the search using 3LBD. BAA07892.2 (LBDS1) is not identified.

Figure 3: The Redundant Sequence Display results page for BAA07892.2 (LBDS1). There is an equivalent sequence called Q14008 (see arrow).

Figure 4: InterPro PFAM search results for BAA07892.2 (LBDS1), see arrow ①.

Figure 5A: NCBI protein report for BAA07892.2 (LBDS1).

Figure 5B: SWISS-PROT entry for a sequence equivalent to BAA07892.2 (LBDS1), called Q14008, CH-TOG protein.

Figure 6A: This is the front end of the Biopendium database. A search of the database is initiated using BAA07892.2 (LBDS1), as the query sequence.

Figure 6B: A selection of the Inpharmatica Genome Threader results of search using BAA07892.2 (LBDS1), as the query sequence. The arrow points to 3LBD.

Figure 6C: A selection of the reverse-maximised PSI-BLAST results obtained using BAA07892.2 (LBDS1), as the query sequence. The arrows numbered ① and ② point to homologues of BAA07892.2 (LBDS1) in *X.laevis* and *D.melanogaster* respectively. Arrow ③ points to a *Homo sapiens* variant of BAA07892.2 (LBDS1).

Figure 7: AlEye sequence alignment of BAA07892.2 (LBDS1) and 3LBD.

Figure 8A: LigEye for 3LBD that illustrates the sites of interaction of 9-cis Retinoic Acid (REA420) with the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma, 3LBD.

Figure 8B: iRasMol view of 3LBD, the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma.

Figure 9: AlEye sequence alignment of BAA07892.2 (LBDS1) with *X.laevis* (CAB61894.1) and *D.melanogaster* (CAB55772.1) homologues of BAA07892.2 (LBDS1).

Figure 10: AlEye sequence alignment of BAA07892.2 (LBDS1) with CAA63212.1, a *Homo sapiens* variant of BAA07892.2 (LBDS1).

Figure 11: NCBI protein report for CAA63212.1, a *Homo sapiens* variant of BAA07892.2 (LBDS1). CAA63212.1 was derived from a human breast cancer cell library (arrow).

Figure 12: Abstract of the article by by Charrasse *et al.* in the Eur: J. Biochem. 1995 Dec 1;234(2)406-13.

Example 1: BAA07892.2 (LBDS1)

In order to initiate a search for novel, distantly related Nuclear Hormone Receptor Ligand Binding Domains, an archetypal family member is chosen, the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma. More specifically, the search is initiated using a structure from the Protein Data Bank (PDB) which is operated by the Research Collaboratory for Structural Bioinformatics.

The structure chosen is the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma (PDB code 3LBD; see Figure 1).

A search of the Biopendium (using the Target Mining Interface) for relatives of 3LBD takes place and returns 4475 Genome Threader results. The 4475 Genome Threader results include examples of typical Nuclear Hormone Receptor Ligand Binding Domains, such as that found between residues 182-418 of the *Homo sapiens* Retinoic Acid Receptor Gamma-1 (see arrow in Figure 2A).

Among the proteins known to contain a Nuclear Hormone Receptor Ligand Binding Domain appears a protein which is not annotated as containing a Nuclear Hormone Receptor Ligand Binding Domain, BAA07892.2 (LBDS1; see arrow in Figure 2B). The Inpharmatica Genome Threader has identified a region of the sequence BAA07892.2 (LBDS1), between residues 1507-1742, as having a structure similar to the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma. The possession of a structure similar to the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma suggests that residues 1507-1742 of BAA07892.2 (LBDS1) function as a a Nuclear Hormone Receptor Ligand Binding Domain. The Genome Threader identifies this with 71% confidence.

The search of the Biopendium (using the Target Mining Interface) for relatives of 3LBD also returns 850 Forward PSI-Blast results. Forward PSI-Blast (see Figure 2C) is unable to identify this relationship; only the Inpharmatica Genome Threader is able to identify BAA07892.2 (LBDS1) as containing a Nuclear Hormone Receptor Ligand Binding Domain.

In order to assess what is known in the public domain databases about BAA07892.2 (LBDS1) the Redundant Sequence Display Page (Figure 3) is viewed. The Redundant Sequence Display Page reveals a near-identical sequence to BAA07892.2 (LBDS1) called Q14008. Q14008 is a sub-sequence of BAA07892.2 (LBDS1), lacking only the six N-terminal residues of BAA07892.2.

There are no PROSITE or PRINTS hits which identify BAA07892.2 (LBDS1) as containing a Nuclear Hormone Receptor Ligand Binding Domain. PROSITE and PRINTS are databases that help to describe proteins of similar families. Returning no Nuclear Hormone Receptor Ligand Binding Domain hits from both databases means that BAA07892.2 (LBDS1) is unidentifiable as containing a Nuclear Hormone Receptor Ligand Binding Domain using PROSITE or PRINTS.

In order to identify if any other public domain annotation vehicle is able to annotate BAA07892.2 (LBDS1) as containing a Nuclear Hormone Receptor Ligand Binding Domain, the BAA07892.2 (LBDS1) protein sequence is searched against the PFAM database (Protein Family Database of Alignment and hidden Markov models) at the InterPro website (see Figure 4 arrow ①). No PFAM-A matches are returned. Thus PFAM does not identify BAA07892.2 (LBDS1) as containing a Nuclear Hormone Receptor Ligand Binding Domain.

Interestingly, PROSITE PFscan (see Figure 4 arrow ②) identifies a bipartite nuclear localisation signal in BAA07892.2 (LBDS1) at residues 550-567. A typical (although non-diagnostic) feature of Nuclear Hormone Receptors is the possession of a bipartite nuclear localisation signal.

The National Center for Biotechnology Information (NCBI) Genebank protein database is then viewed to examine if there is any further information that is known in the public domain relating to BAA07892.2 (LBDS1). This is the US public domain database for protein and gene sequence deposition (Figure 5A). BAA07892.2 (LBDS1) is a *Homo sapiens* sequence, its Genebank protein ID is BAA07892.2 and it is 2038 amino acids in length. BAA07892.2 (LBDS1) was cloned by a group of scientists at the Kazusa DNA Research Institute, Kisarazu, Japan. The public domain information for this gene does not annotate it as containing a Nuclear Hormone Receptor Ligand Binding Domain.

The Swiss Institute of Bioinformatics (SIB) SWISS-PROT protein database is then viewed to examine if there is any further information that is known in the public domain relating to the BAA07892.2 (LBDS1) subsequence Q14008. This is the Swiss public domain database for protein and gene sequence deposition (Figure 5B). The BAA07892.2 (LBDS1) subsequence Q14008 is a *Homo Sapiens* sequence, its SWISS-PROT protein ID is Q14008 and it is 2032 amino acids in length. Q14008 is called CH-TOG protein (Colonic and Hepatic Tumour Over-Expressed Gene) protein. Q14008, CH-TOG was cloned by Charrasse *et al.* at INSERM, Montpellier, France. The public domain information for this gene does not annotate it as containing a Nuclear Hormone Receptor Ligand Binding Domain.

Therefore, it can be concluded that using all public domain annotation tools, BAA07892.2 (LBDS1) may not be annotated as containing a Nuclear Hormone Receptor Ligand Binding Domain. Only the Inpharmatica Genome Threader is able to annotate this protein as containing a Nuclear Hormone Receptor Ligand Binding Domain.

The reverse search is now carried out. BAA07892.2 (LBDS1) is now used as the query sequence in the Biopendium (see Figure 6A). The Inpharmatica Genome Threader identifies residues 1507-1742 of BAA07892.2 (LBDS1) as having a structure that is the same as the Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma with 71% confidence (see arrow in Figure 6B). The Ligand Binding Domain of *Homo sapiens* Retinoic Acid Receptor Gamma (3LBD) was the original query sequence. Positive iterations of PSI-Blast do not return this result (Figure 6C). It is only the Inpharmatica Genome Threader that is able to identify this relationship.

The sequence of the *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain is chosen against which to view the sequence alignment of BAA07892.2 (LBDS1). Viewing the AlEye alignment (Figure 7) of the query protein against the protein identified as being of a similar structure helps to visualize the areas of homology.

The *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain contains an "LBD motif" which has been found in all annotated Nuclear Hormone Receptor Ligand Binding Domains to date. The "LBD motif" is involved in recruiting Nuclear Hormone Receptor Co-Activators and Co-Repressors. The residues PHE251, LEU254, ASP258, GLN259, LEU262 and LEU263 constitute this motif in the *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain (see square boxes Figure 7). The residues VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583 in BAA07892.2 (LBDS1) align with and satisfy the LBD motif criteria (full and broken square boxes in Figure 7). This indicates that BAA07892.2 (LBDS1) contains a Nuclear Hormone

Receptor Ligand Binding Domain similar to The *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain.

In order to ensure that the protein identified is in fact a relative of the query sequence, the visualization programs "LigEye" (Figure 8A) and "iRasmol" (Figure 8B) are used. These visualization tools identify the active site of known protein structures by indicating the amino acids with which known small molecule inhibitors interact at the active site. These interactions are either through a direct hydrogen bond or through hydrophobic interactions. In this manner, one can see if the active site fold/structure is conserved between the identified homologue and the chosen protein of known structure. The LigEye view of the *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain reveals 14 residues which bind 9-cis Retinoic Acid (circled in Figure 7). Only 3 (ALA1556, LEU1591, LEU1725, circled in Figure 7) of these 14 residues are precisely conserved in BAA07892.2 (LBDS1), suggesting that BAA07892.2 (LBDS1) may not bind 9-cis Retinoic Acid in particular. However, there are 11 hydrophobic residues which line the pocket of the *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain; PHE230, LEU233, ALA234, LEU268, LEU271, MET272, ILE275, PHE288, PHE304, LEU400 and ILE412. In BAA07892.2 (LBDS1), 3 (ALA1556, LEU1591, LEU1725, circled in Figure 7) of these 11 residues are perfectly conserved. A further 5 residues (ILE1552, VAL1555, ILE1592, PHE1595, LEU1739, broken circles in Figure 7) are conserved in hydrophobic nature. This conservation of hydrophobicity in 8 out of the 11 hydrophobic residues which line the binding pocket indicates that BAA07892.2 (LBDS1) will bind a hydrophobic steroid-like hydrophobic ligand. Typically, the binding pocket of Nuclear Hormone Receptor Ligand Binding Domains will contain a crucial arginine residue. The guanidinium group of this arginine functions to bind and stabilise a hydroxyl or carboxyl group present on the steroid/steroid-like ligand. For example, ARG394 of the *Homo sapiens* Estrogen Receptor Ligand Binding Domain binds and stabilises the phenolic hydroxyl group of the A ring of Estradiol (Brzozowski Nature 1997 vol.389:p753-758). LigEye shows that ARG278 of the *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain stabilises the carboxylate group of Retinoic Acid. In BAA07892.2 (LBDS1) this ARG is not conserved, but there is ARG1599 (asterisk) which is displaced one residue towards the C-terminus. ARG1599 in

BAA07892.2 (LBDS1) could function to bind and stabilise a hydroxyl or carboxyl group present on a steroid-like ligand. This indicates that indeed as predicted by the Inpharmatica Genome Threader, BAA07892.2 (LBDS1) folds in a similar manner to of the *Homo sapiens* Retinoic Acid Receptor Gamma Ligand Binding Domain and as such is identified as containing a Nuclear Hormone Receptor Ligand Binding Domain.

Reverse-maximised PSI-BLAST of BAA07892.2 (LBDS1) identifies a *X.laevis* homologue of BAA07892.2 (LBDS1) called CAB61894.1, XMAP215, which has 76.0% sequence identity to BAA07892.2 (LBDS1), see Figure 6C arrow ①. Reverse-maximised PSI-BLAST of BAA07892.2 (LBDS1) also identifies a *D.melanogaster* homologue of BAA07892.2 (LBDS1) called CAB55772.1, Microtubule Associated Protein, which has 45.0% sequence identity to BAA07892.2 (LBDS1), see Figure 6C arrow ②.

BAA07892.2 (LBDS1), *X.laevis* CAB61894.1 (XMAP215) and *D.melanogaster* CAB55772.1 (Microtubule Associated Protein) are aligned and viewed in AlEye (Figure 9). AlEye reveals that the 8 hydrophobic residues of BAA07892.2 (LBDS1) predicted to line the ligand binding pocket (ILE1552, VAL1555, ALA1556, LEU1591, ILE1592, PHE1595, LEU1725 and LEU1739) are all precisely conserved in *X.laevis* CAB61894.1, XMAP215 (see Figure 9). Furthermore, the six residues matching the "LBD motif" in BAA07892.2 (LBDS1), (VAL1573, LEU1574, ASP1578, LYS1579, ALA1582, MET1583) are all precisely conserved in *X.laevis* CAB61894.1, XMAP215 (see Figure 9). In addition the predicted ligand binding ARG1599 is precisely conserved in *X.laevis* CAB61894.1, XMAP215 (marked with an asterisk, see Figure 9). Residues which are essential for the function of a protein will be conserved in homologues of that protein. Thus the absolute conservation of all the residues which would be essential for the function of the predicted BAA07892.2 (LBDS1) Nuclear Hormone Receptor Ligand Binding Domain in the *X.laevis* homologue CAB61894.1 (XMAP215) strongly supports the annotation of BAA07892.2 (LBDS1) as containing a Nuclear Hormone Receptor Ligand Binding Domain.

The AlEye alignment also reveals conserved relationships between BAA07892.2 (LBDS1) and *D.melanogaster* CAB55772.1 (Microtubule Associated Protein). AlEye indicates that 2 (ILE1552 and VAL1555) of the 8 hydrophobic residues of BAA07892.2

(LBDS1) are precisely conserved in *D.melanogaster* CAB55772.1 (Microtubule Associated Protein). Another 4 of the 8 BAA07892.2 (LBDS1) hydrophobic residues (LEU1591, PHE1595, LEU1725, LEU1739) are conservatively substituted for other hydrophobic residues in *D.melanogaster* CAB55772.1 (Microtubule Associated Protein). Furthermore, the Lys1579 of the BAA07892.2 (LBDS1) "LBD motif" is precisely conserved in *D.melanogaster* CAB55772.1 (Microtubule Associated Protein). The other five residues of the BAA07892.2 (LBDS1) "LBD motif" (VAL1573, LEU1574, ASP1578, ALA1582, MET1583) are conservatively substituted in *D.melanogaster* CAB55772.1 (Microtubule Associated Protein). In addition the predicted ligand binding ARG1599 is precisely conserved in *D.melanogaster* CAB55772.1 (Microtubule Associated Protein), marked with an asterisk, see Figure 9. Residues which are essential for the function of a protein will be conserved in homologues of that protein. Thus the precise conservation of 4 (and conservative substitution of a further 9) of the 15 residues which would be essential for the function of the predicted BAA07892.2 (LBDS1) Nuclear Hormone Receptor Ligand Binding Domain in the *D.melanogaster* homologue CAB55772.1 (Microtubule Associated Protein) supports the annotation of BAA07892.2 (LBDS1) as containing a Nuclear Hormone Receptor Ligand Binding Domain.

Reverse-maximised PSI-BLAST of BAA07892.2 (LBDS1) also identifies *a Homo sapiens* sequence CAA63212.1 which has 95% sequence identity to BAA07892.2 (LBDS1), see Figure 6C arrow ③. An AlEye alignment of CAA63212.1 with BAA07892.2 (LBDS1) reveals that there are only two regions of difference (see Figure 10). The first difference is that CAA63212.1 lacks the 6 N-terminal amino acids of BAA07892.2 (LBDS1). The second difference is that CAA63212.1 has an internal deletion compared to BAA07892.2 (LBDS1) which removes 60 amino acids. This internal deletion may represent a splice variant or disease-linked deletion since the cDNA was obtained from breast cancer tissue (see Figure 11 arrow). Interestingly this 60 amino acid deletion means that CAA63212.1 lacks the "LBD motif", ARG1599 and three hydrophobic predicted ligand binding residues (LEU1591, ILE1592 and PHE1595). As such CAA63212.1 would be predicted to be non-competent at binding a steroid-like ligand and recruiting Nuclear Hormone Co-Activators and Co-Repressors.

There is an associated article by Charrasse *et al.* in the Eur. J. Biochem. 1995 Dec 1;234(2)406-13 (Figure 12). This article demonstrates that mRNA encoding the BAA07892.2 (LBDS1) subsequence Q14008, CH-TOG is over-expressed in cancerous colon tissue compared to healthy colon. In addition, mRNA encoding the BAA07892.2 (LBDS1) subsequence Q14008, CH-TOG is over-expressed in cancerous liver tissue compared to healthy colon. This establishes a direct correlation between levels of the BAA07892.2 (LBDS1) subsequence Q14008, CH-TOG and the cancerous state. Connecting BAA07892.2 (LBDS1) to the cancerous state paves the way for the development of agents that target the BAA07892.2 (LBDS1) gene or its encoded protein, to diagnose and/or treat these conditions. In particular, the identification of this gene as containing a Nuclear Hormone Receptor Ligand Binding Domain facilitates the development of agents that are specific to the BAA07892.2 (LBDS1) protein, for example, through the use of Nuclear Hormone Receptor Ligand Binding Domain agonists or antagonists.

Sequence Listing

SEQ ID NO: 1 Nucleotide coding sequence for BAA07892.2 (LBDS1) protein (nucleotides 30-6146 translate to give SEQ ID NO:2)

```
1    cccagctgag gaaatactct taattctaag gaaaacctgg aagcacaatg ggagatgaca
61   gtgagtggtt gaaactgcca gttgatcaga aatgtgaaca caagctgtgg aaagcaaggt
121  taagtgggta tgaagaggcc ctgaagatct tccagaaaat aaaggatgaa aagagcccag
181  agtggtccaa atttttagga ttgatcaaaa aatttgtcac tgattccaat gcagtggttc
241  aattgaaagg attagaagct gcacttgttt atgttgaaaa tgcccatgta gcaggaaaaa
301  ccacaggaga agttgtgtca ggtgttgtaa gtaaggtgtt caatcaacct aaagctaaag
361  ccaaggagct gggcatagag atctgtctta tgtacataga gattgagaaa ggagaggctg
421  ttcaagaaga gctcctgaaa ggcttggaca ataagaatcc caagatcata gtggcctgta
481  tagagacact gaggaaagcc ttaagtgaat ttggttccaa aatcatcttg cttaagccaa
541  ttatcaaagt gttgccaaaa ctctttgagt ctcgagagaa ggctgttcga gatgaagcca
601  aactaattgc tgtggagatt tacagatgga ttcgggatgc tctgagaccc ccattacaaa
661  atataaactc tgttcagttg aaagaactag aagaagaatg ggtcaaactg ccaacaagtg
721  ctcctagacc tactcgattt cttcgttccc aacaagaact agaagctaaa ttggaacaac
781  aacagtctgc tggtggagat gctgaaggag gtggtgatga tggtgatgag gtgccacaaa
841  tagatgctta tgagctttta gaagctgtag aaatcctttc caaacttccc aaagactttt
901  atgacaaaat tgaggcaaaa aaatggcaag agagaaaaga ggccctggag tctgtagaag
961  tactaataaa aaaccccaaa ctggaagctg gcgattatgc agatttagta aaagcattaa
1021 agaaggttgt tggaaaggac accaatgtca tgttggtggc tttggcagca aaatgtctta
1081 ctggcctggc tgttgggcta aggaagaaat ttggacaata tgcaggacat gttgtgccaa
1141 ccatcttgga gaaattcaaa gagaagaaac ctcaagtggt acaagccctg caggaggcaa
1201 ttgatgcaat cttccttact accacactac agaacatcag tgaggatgtt ttagcagtaa
1261 tggataataa aaatccaacc atcaagcagc agacatctct ttttattgca agaagtttcc
1321 gccactgcac tgcttctacc ctgccaaaga gcttgctaaa gccctttgt gctgcactac
1381 ttaagcacat caatgattct gctcctgaag tcagagatgc cgcatttgaa gcattgggta
1441 ctgctttgaa ggtggttggc gagaaagcag taaaaccatt cctagctgat gtggacaaac
1501 tcaagcttga taagatcaaa gaatgttcag aaaaggtaga actgatacat ggtaagaaag
1561 ctggactagc tgctgataag aaggaattca aacctctgcc tggaaggact gctgcttcag
1621 gggctgcagg agataaggac acaaaggaca tttctgcacc caaaccagga cctctaaaaa
1681 aggcacctgc tgctaaggct ggtgggccac caaaaaaggg gaaaccagct gcaccaggag
1741 gcgcagggaa tactggaacc aagaacaaga aaggactgga gactaaagaa atagtggagc
1801 ctgagctctc gatagaagta tgtgaagaaa aagcttcagc tgttcttccc cctacctgta
1861 tacagcttct tgacagcagt aactggaaag aaaggctggc ttgtatggaa gagttccaga
1921 aggctgttga gctaatggac cgaactgaaa tgccatgcca ggcattagtg aggatgctag
1981 ccaagaaacc tggatggaaa gaaactaatt ttcaggtgat gcaaatgaag cttcatatag
2041 ttgctttgat tgcccagaag ggaaattttt ccaaaacgtc agctcaggtt gtattagatg
2101 gccttgtgga caagattgga gatgtgaaat gtgggaacaa tgcaaaagaa gctatgacag
2161 caatagccga agcctgtatg ttaccatgga ctgctgaaca ggttgtgtca atggctttct
```

```
2221 cacaaaagaa tcccaaaaat cagtcagaaa ctctgaattg gctatcaaat gccataaaag
2281 aatttggttt ttctgggttg aatgtcaaag ctttcattag caatgtgaag acagctcttg
2341 ctgcaacaaa cccagctgtg aggactgctg ccataaccct gcttggcgtg atgtatctgt
2401 atgttggtcc ctctttgcga atgttctttg aggatgagaa gcctgccctc ctatcccaga
2461 tagatgcaga atttgagaag atgcagggac aaagtccacc tgctccaacc agaggaattt
2521 ccaagcatag cacaagtggt acagatgaag gagaagatgg agatgaacca gatgacggga
2581 gcaatgatgt cgttgatctt ttgccgagga cggagatcag tgataaaatc acttcagagt
2641 tggtatctaa gattggtgat aagaattgga agattaggaa agaaggccta gatgaagtgg
2701 caggtattat taatgacgca aaatttatcc aaccgaatat aggtgaactt ccaactgcct
2761 tgaagggtcg actcaatgat tcaaataaaa tcttggtaca gcaaacgctg aatatcctgc
2821 aacaactggc agtagccatg ggcccaaata ttaagcaaca tgtaaaaaat ttaggcatcc
2881 ctatcatcac agtccttgga gacagcaaga acaatgttcg agctgctgcc ctagcgactg
2941 tgaatgcttg ggcagaacag actggcatga aggaatggct ggaaggagaa gatctttctg
3001 aagagctcaa aaaggaaaat cctttcttga ggcaagagct tctgggctgg ctggctgaga
3061 aactacctac tcttcgttcc acccctacag accttatcct ttgtgttcct catctctact
3121 cctgcctaga agatcgaaat ggagatgtgc gaaagaaggc ccaagatgcc ttgccattct
3181 tcatgatgca tttaggatat gaaaaaatgg ccaaggctac tgggaaacta aagccaactt
3241 ctaaagatca ggtattggcc atgctagaga aagccaaagt taacatgcca gccaagcctg
3301 ctccacccac taaagcaact tctaaaccaa tgggagggtc cgctccagcc aaattccagc
3361 ctgcatcagc acctgctgaa gattgtattt ccagcagtac agaacccaaa cctgatccaa
3421 aaaaggccaa agctccagga ttatcctcta aagcaaagag tgcacaaggg aagaagatgc
3481 caagcaaaac cagcttaaag gaggatgaag acaaatccgg gcctattttt attgttgttc
3541 caaatggaaa agagcaaagg atgaaagatg aaaaaggatt gaaggtgcta aagtggaatt
3601 ttactacccc acgggatgaa tacattgagc aactaaagac tcaaatgtct agctgtgtgg
3661 ctaaatggtt acaagatgag atgtttcact cagactttca gcatcataac aaagcccttg
3721 ctgttatggt tgatcacttg gagagtgaaa aagaaggagt tattggttgc ctggatctta
3781 tcttaaagtg gcttaccctg aggtttttg acaccaatac aagcgtcctg atgaaagcac
3841 tagaatattt aaaattgctc ttcaccttgc taagtgaaga agaatatcat cttactgaga
3901 atgaagcatc ttccttcatc ccctatcttg tcgtcaaggt tggagaacca aaggatgtca
3961 ttcgtaaaga tgttcgtgcc atcctgaacc ggatgtgcct tgtctaccca gctagcaaga
4021 tgtttccctt tatcatggaa ggaaccaaat ccaaaaactc taagcagaga gcagagtgcc
4081 tggaagagct gggatgtctg gttgagtcct atggcatgaa tgtttgccaa ccaaccccag
4141 gaaaagcctt aaaggaaata gctgttcaca taggagaccg tgacaatgct gtacgcaatg
4201 ctgcactcaa caccattgta acggtgtaca atgtacatgg ggatcaggtg ttcaaactga
4261 ttggaaatct ttctgaaaag gatatgagca tgctcgagga gaggattaag cggtcagcaa
4321 agagaccctc tgctgcacca ataaaacagg tggaagagaa acctcagcgt gcacagaaca
4381 taagctccaa tgccaacatg ttacgcaagg gaccagctga ggacatgtct tccaaactca
4441 accaagcccg aagcatgagt gggcatcctg aggcagccca gatggtccgc cgagaattcc
4501 agctggatct agatgagatt gagaatgaca atggtacagt ccgatgtgaa atgccagaac
4561 ttgttcagca caaactggat gacatttttg agccagtcct tattcctgaa cccaagatcc
4621 gggctgtttc tccacacttc gatgacatgc acagtaatac agcatccaca atcaatttca
4681 ttatctccca agtagccagt ggtgacatca acacaagtat ccaagctctg acacagatcg
4741 atgaggtcct gagacaggaa gacaaagctg aagccatgtc cggccatatt gatcagtttc
```

```
4801 tgatagccac ttttatgcag ctaagactca tctacaacac acacatggca gatgagaaat
4861 tggagaagga cgagatcatc aagttgtata gctgtatcat tggcaacatg atttcgctgt
4921 ttcagataga gagccttgcc cgggaggcct ccactggagt actaaaagac ctaatgcatg
4981 gcctcatcac cttaatgctg gattctcgga ttgaagatct tgaggaagga caacaggtca
5041 tccgctctgt gaacctcttg gtggtgaagg ttctggagaa gtcagaccag accaacatcc
5101 tgagtgccct acttgttttg ctccaagaca gcctgctagc aacagccagt tctcccaaat
5161 tctcagagct tgttatgaag tgtctctgga gaatggttcg actgttgcct gataccatca
5221 atagcattaa cctagacaga attcttctgg atatccacat tttcatgaag gtcttcccca
5281 aagagaaact gaagcaatgc aaaagtgaat ttcccataag gaccctaaag accctgctac
5341 acaccttatg caaattaaaa gggcccaaga tcctggacca cctaacgatg atcgacaaca
5401 aaaacgagtc tgagctggag gcccatctct gccggatgat gaagcacagt atggaccaga
5461 ctgggagcaa gtctgataag gaaacagaaa agggagcatc tcgaatagat gaaaaatcat
5521 caaaggccaa agtgaatgat ttcttagctg agatttttaa gaagattggc tctaaagaaa
5581 acactaaaga gggactagca gagttatatg aatataagaa gaaatactca gatgctgaca
5641 ttgaaccatt tctgaaaaat tcctcacagt tcttccagag ctatgtcgaa agaggccttc
5701 gggtgattga gatggagagg gagggcaaag gtcgtatttc cacttcaaca ggcatctccc
5761 ctcagatgga agtcacatgt gtgcccacgc ccacaagcac agtgtcctcc ataggtaaca
5821 caaatgggga agaagtgggg ccatctgtct acttggaaag gctaaagatc ctccgacagc
5881 gatgtggtct ggacaacaca aagcaagatg accgacctcc tttgacctct ttgctctcca
5941 aaccagcagt tcctactgtc gcctcttcca cagacatgct ccacagcaaa ctctctcagc
6001 tccgggagtc acgggagcag caccagcatt cagacctgga ttctaaccag actcactctt
6061 caggaactgt gacctcctcc tcctccacag ctaacataga cgacttgaaa aaaagactgg
6121 agagaataaa gagcagtcgc aaatgaagct gccccactcc cccggcaccc tgcagcttta
6181 gtttactaaa ctagaagtcc tcatagttta aaatggcctc agcaggccta gtgtatacaa
6241 actggttgta tgtatcatgc cgtggagcta gggggaggag tcattgtggc acaagtattt
6301 gtacatactc tgcttctctc tgtcagcgtc ctgctgctct agaagactgt ccgtggatga
6361 gtttagtgta cagacttgta aacagctgcc ccctctctgc tcagtctagt tcccagatcc
6421 ttttcttttt aattgctcat ttgtaaaatt gtcctaatct ttcctagctt tttaatagtt
6481 aatattagaa actctttaat agttttcctt tcagtttgtg agctcttctc tgtcgccctg
6541 aagggtcact gtattctgta tgaatgcatg gcatgataca actaatttaa gagtctttta
6601 taaataaagt ttgcattaac tatacctg
```

SEQ ID NO: 2 Protein sequence for BAA07892.2 (LBDS1) MET7=start methionine

```
  1 gkpgstmgdd sewlklpvdq kcehklwkar lsgyeealki fqkikdeksp ewskflglik
61   kfvtdsnavv qlkgleaalv yvenahvagk ttgevvsgvv skvfnqpkak akelgieicl
121   myieiekgea vqeellkgld nknpkiivac ietlrkalse fgskiillkp iikvlpklfe
181   srekavrdea kliaveiyrw irdalrpplq ninsvqlkel eeewvklpts aprptrflrs
241   qqeleakleq qqsaggdaeg ggddgdevpq idayelleav eilsklpkdf ydkieakkwq
301   erkealesve vliknpklea gdyadlvkal kkvvgkdtnv mlvalaakcl tglavglrkk
361   fgqyaghvvp tilekfkekk pqvvqalqea idaiflttl qnisedvlav mdnknptikq
421   qtslfiarsf rhctastlpk sllkpfcaal lkhindsape vrdaafealg talkvvgeka
481   vkpfladvdk lkldkikecs ekvelihgkk aglaadkkef kplpgrtaas gaagdkdtkd
```

```
541  isapkpgplk kapaakaggp pkkgkpaapg gagntgtknk kgletkeive pelsievcee
601  kasavlpptc iqlldssnwk erlacmeefq kavelmdrte mpcqalvrml akkpgwketn
661  fqvmqmklhi valiaqkgnf sktsaqvvld glvdkigdvk cgnnakeamt aiaeacmlpw
721  taeqvvsmaf sqknpknqse tlnwlsnaik efgfsglnvk afisnvktal aatnpavrta
781  aitllgvmyl yvgpslrmff edekpallsq idaefekmqg qsppaptrgi skhstsgtde
841  gedgdepddg sndvvdllpr teisdkitse lvskigdknw kirkegldev agiindakfi
901  qpnigelpta lkgrlndsnk ilvqqtlnil qqlavamgpn ikqhvknlgi piitvlgdsk
961  nnvraaalat vnawaeqtgm kewlegedls eelkkenpfl rqellgwlae klptlrstpt
1021 dlilcvphly scledrngdv rkkaqdalpf fmmhlgyekm akatgklkpt skdqvlamle
1081 kakvnmpakp apptkatskp mggsapakfq pasapaedci ssstepkpdp kkakapglss
1141 kaksaqgkkm psktslkede dksgpifivv pngkeqrmkd ekglkvlkwn fttprdeyie
1201 qlktqmsscv akwlqdemfh sdfqhhnkal avmvdhlese kegvigcldl ilkwltlrff
1261 dtntsvlmka leylkllftl lseeeyhlte neassfipyl vvkvgepkdv irkdvrailn
1321 rmclvypask mfpfimegtk sknskqraec leelgclves ygmnvcqptp gkalkeiavh
1381 igdrdnavrn aalntivtvy nvhgdqvfkl ignlsekdms mleerikrsa krpsaapikq
1441 veekpqraqn issnanmlrk gpaedmsskl nqarsmsghp eaaqmvrref qldldeiend
1501 ngtvrcempe lvqhklddif epvlipepki ravsphfddm hsntastinf iisqvasgdi
1561 ntsiqaltqi devlrqedka eamsghidqf liatfmqlrl iynthmadek lekdeiikly
1621 sciignmisl fqieslarea stgvlkdlmh glitlmldsr iedleegqqv irsvnllvvk
1681 vleksdqtni lsallvllqd sllatasspk fselvmkclw rmvrllpdti nsinldrill
1741 dihifmkvfp keklkqckse fpirtlktll htlcklkgpk ildhltmidn kneseleahl
1801 crmmkhsmdq tgsksdkete kgasrideks skakvndfla eifkkigske ntkeglaely
1861 eykkkysdad iepflknssq ffqsyvergl rviemeregk gristsgis pqmevtcvpt
1921 ptstvssign tngeevgpsv ylerlkilrq rcgldntkqd drppltslls kpavptvass
1981 tdmlhsklsq lresreqhqh sdldsnqths sgtvtsssst aniddlkkrl erikssrk
```

CLAIMS

1. A polypeptide, which polypeptide:
   (i) comprises or consists of the amino acid sequence as recited in SEQ ID NO:2;
   (ii) is a fragment thereof having Nuclear Hormone Receptor Ligand Binding Domain activity or having an antigenic determinant in common with the polypeptide of (i); or
   (iii) is a functional equivalent of (i) or (ii).
2. A polypeptide which is a fragment according to claim 1(ii), which includes the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide, said Nuclear Hormone Receptor Ligand Binding Domain region being defined as including residues 1507 and 1742 inclusive, of the amino acid sequence recited in SEQ ID NO:2, wherein said fragment possesses the "LBD motif" residues VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues, and possesses Nuclear Hormone Receptor Ligand Binding Domain activity.
3. A polypeptide which is a functional equivalent according to claim 1(iii), is homologous to the amino acid sequence as recited in SEQ ID NO:2, possesses the catalytic residues VAL1573, LEU1574, ASP1578, LYS1579, ALA1582 and MET1583, or equivalent residues, and has Nuclear Hormone Receptor Ligand Binding Domain activity.
4. A polypeptide according to claim 3, wherein said functional equivalent is homologous to the Nuclear Hormone Receptor Ligand Binding Domain region of the LBDS1 polypeptide.
5. A fragment or functional equivalent according to any one of claims 1-4, which has greater than 80% sequence identity with an amino acid sequence as recited in SEQ ID NO:2, or with a fragment thereof that possesses Nuclear Hormone Receptor Ligand Binding Domain activity, preferably greater than 85%, 90%, 95%, 98% or 99% sequence identity, as determined using BLAST version 2.1.3 using the default

parameters specified by the NCBI (the National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/) [Blosum 62 matrix; gap open penalty=11 and gap extension penalty=1].

6. A functional equivalent according to any one of claims 1-5, which exhibits significant structural homology with a polypeptide having the amino acid sequence given in any one of SEQ ID NO:2, or with a fragment thereof that possesses Nuclear Hormone Receptor Ligand Binding Domain activity.

7. A fragment as recited in claim 1, 2, or 5, having an antigenic determinant in common with the polypeptide of claim 1(i), which consists of 7 or more (for example, 8, 10, 12, 14, 16, 18, 20 or more) amino acid residues from the sequence of SEQ ID NO:2.

8. A purified nucleic acid molecule which encodes a polypeptide according to any one of the preceding claims.

9. A purified nucleic acid molecule according to claim 8, which has the nucleic acid sequence as recited in SEQ ID NO:1, or is a redundant equivalent or fragment thereof.

10. A fragment of a purified nucleic acid molecule according to claim 8 or claim 9, which comprises, nucleotides 4548 to 5255 inclusive of SEQ ID NO:1, or is a redundant equivalent thereof.

11. A purified nucleic acid molecule which hybridizes under high stringency conditions with a nucleic acid molecule according to any one of claims 8-10.

12. A vector comprising a nucleic acid molecule as recited in any one of claims 8-11.

13. A host cell transformed with a vector according to claim 12.

14. A ligand which binds specifically to, and which preferably inhibits the Nuclear Hormone Receptor Ligand Binding Domain activity of, a polypeptide according to any one of claims 1-7.

15. A ligand according to claim 14, which is an antibody.

16. A compound that either increases or decreases the level of expression or activity of a polypeptide according to any one of claims 1-7.

17. A compound according to claim 16 that binds to a polypeptide according to any one of claims 1-7 without inducing any of the biological effects of the polypeptide.

18. A compound according to claim 16 or claim 17, which is a natural or modified substrate, ligand, enzyme, receptor or structural or functional mimetic.

19. A polypeptide according to any one of claims 1-7, a nucleic acid molecule according to any one of claims 8-11, a vector according to claim 12, a ligand according to claim 14 or 15, or a compound according to any one of claims 16-18, for use in therapy or diagnosis of disease.

20. A method of diagnosing a disease in a patient, comprising assessing the level of expression of a natural gene encoding a polypeptide according to any one of claims 1-7, or assessing the activity of a polypeptide according to any one of claims 1-7, in tissue from said patient and comparing said level of expression or activity to a control level, wherein a level that is different to said control level is indicative of disease.

21. A method according to claim 20 that is carried out *in vitro*.

22. A method according to claim 20 or claim 21, which comprises the steps of: (a) contacting a ligand according to claim 14 or claim 15 with a biological sample under conditions suitable for the formation of a ligand-polypeptide complex; and (b) detecting said complex.

23. A method according to claim 20 or claim 21, comprising the steps of:

a) contacting a sample of tissue from the patient with a nucleic acid probe under stringent conditions that allow the formation of a hybrid complex between a nucleic acid molecule according to any one of claims 8-11 and the probe;

b) contacting a control sample with said probe under the same conditions used in step a); and

c) detecting the presence of hybrid complexes in said samples;

wherein detection of levels of the hybrid complex in the patient sample that differ from levels of the hybrid complex in the control sample is indicative of disease.

24. A method according to claim 20 or claim 21, comprising:

a) contacting a sample of nucleic acid from tissue of the patient with a nucleic acid primer under stringent conditions that allow the formation of a hybrid complex between a nucleic acid molecule according to any one of claims 8-11 and the primer;

b) contacting a control sample with said primer under the same conditions used in step a); and

c) amplifying the sampled nucleic acid; and

d) detecting the level of amplified nucleic acid from both patient and control samples;

wherein detection of levels of the amplified nucleic acid in the patient sample that differ significantly from levels of the amplified nucleic acid in the control sample is indicative of disease.

25. A method according to claim 20 or claim 21 comprising:

a) obtaining a tissue sample from a patient being tested for disease;

b) isolating a nucleic acid molecule according to any one of claims 8-11 from said tissue sample; and

c) diagnosing the patient for disease by detecting the presence of a mutation which is associated with disease in the nucleic acid molecule as an indication of the disease.

26. The method of claim 25, further comprising amplifying the nucleic acid molecule to form an amplified product and detecting the presence or absence of a mutation in the amplified product.

27. The method of either claim 25 or 26, wherein the presence or absence of the mutation in the patient is detected by contacting said nucleic acid molecule with a nucleic acid probe that hybridises to said nucleic acid molecule under stringent conditions to form a hybrid double-stranded molecule, the hybrid double-stranded molecule having an unhybridised portion of the nucleic acid probe strand at any portion corresponding to a mutation associated with disease; and

detecting the presence or absence of an unhybridised portion of the probe strand as an indication of the presence or absence of a disease-associated mutation.

28. A method according to any one of claims 20-27, wherein said disease is a cell proliferative disorders, including neoplasm, cancer and myeloproliferative disorders and angiogenesis, metabolic disorders, including obesity, osteoporosis, non-insulin dependent diabetes, lipid metabolism disorder, hyperthyroidism, hyperparathyroidism, hypercalcemia, hypercholestrolemia, hyperlipidemia, cardiovascular disorders including, hypertension, athersclerosis, hyperinsulinemia, heart arrhythmia, renal disorders, including glomerulonephritis, renovascular hypertension, skin disorders including dermatological disease, psoriasis, acne, aging, eczema, wound healing, inflammation, including inflammatory bowel disease, respiratory tract inflammation, emphysema, asthma, immune disorder, autoimmune disease, multiple sclerosis, nervous system disorders, including anxiety, depression, neurodegenerative disease, Alzheimer's disease Parkinson's disease, brain injury, stroke and pain, infection, including virus infection, and other conditions in which nuclear hormones are implicated.

29. Use of a polypeptide according to any one of claims 1-7 as a Nuclear Hormone Receptor Ligand Binding Domain.

30. Use of a nucleic acid molecule according to any one of claims 8-11 to express a protein that possesses Nuclear Hormone Receptor Ligand Binding Domain activity.

31. A method for effecting cell-cell adhesion, utilising a polypeptide according to any one of claims 1-7.

32. A pharmaceutical composition comprising a polypeptide according to any one of claims 1-7, a nucleic acid molecule according to any one of claims 8-11, a vector according to claim 12, a ligand according to claim 14 or 15, or a compound according to any one of claims 16-18.

33. A vaccine composition comprising a polypeptide according to any one of claims 1-7 or a nucleic acid molecule according to any one of claims 8-11.

34. A polypeptide according to any one of claims 1-7, a nucleic acid molecule according to any one of claims 8-11, a vector according to claim 12, a ligand according to claim 14 or 15, a compound according to any one of claims 16-18, or a pharmaceutical

composition according to claim 32 for use in the manufacture of a medicament for the treatment of cell proliferative disorders, including neoplasm, cancer and myeloproliferative disorders and angiogenesis, metabolic disorders, including obesity, osteoporosis, non-insulin dependent diabetes, lipid metabolism disorder, hyperthyroidism, hyperparathyroidism, hypercalcemia, hypercholestrolemia, hyperlipidemia, cardiovascular disorders including, hypertension, athersclerosis, hyperinsulinemia, heart arrhythmia, renal disorders, including glomerulonephritis, renovascular hypertension, skin disorders including dermatological disease, psoriasis, acne, aging, eczema, wound healing, inflammation, including inflammatory bowel disease, respiratory tract inflammation, emphysema, asthma, immune disorder, autoimmune disease, multiple sclerosis, nervous system disorders, including anxiety, depression, neurodegenerative disease, Alzheimer's disease Parkinson's disease, brain injury, stroke and pain, infection, including virus infection, and other conditions in which nuclear hormones are implicated.

35. A method of treating a disease in a patient, comprising administering to the patient a polypeptide according to any one of claims 1-7, a nucleic acid molecule according to any one of claims 8-11, a vector according to claim 12, a ligand according to claim 14 or 15, a compound according to any one of claims 16-18, or a pharmaceutical composition according to claim 32.

36. A method according to claim 35, wherein, for diseases in which the expression of the natural gene or the activity of the polypeptide is lower in a diseased patient when compared to the level of expression or activity in a healthy patient, the polypeptide, nucleic acid molecule, vector, ligand, compound or composition administered to the patient is an agonist.

37. A method according to claim 35, wherein, for diseases in which the expression of the natural gene or activity of the polypeptide is higher in a diseased patient when compared to the level of expression or activity in a healthy patient, the polypeptide, nucleic acid molecule, vector, ligand, compound or composition administered to the patient is an antagonist.

38. A method of monitoring the therapeutic treatment of disease in a patient, comprising monitoring over a period of time the level of expression or activity of a polypeptide according to any one of claims 1-7, or the level of expression of a nucleic acid molecule according to any one of claims 8-11 in tissue from said patient, wherein altering said level of expression or activity over the period of time towards a control level is indicative of regression of said disease.

39. A method for the identification of a compound that is effective in the treatment and/or diagnosis of disease, comprising contacting a polypeptide according to any one of claims 1-7, a nucleic acid molecule according to any one of claims 8-11, or a host cell according to claim 13 with one or more compounds suspected of possessing binding affinity for said polypeptide or nucleic acid molecule, and selecting a compound that binds specifically to said nucleic acid molecule or polypeptide.

40. A kit useful for diagnosing disease comprising a first container containing a nucleic acid probe that hybridises under stringent conditions with a nucleic acid molecule according to any one of claims 8-11; a second container containing primers useful for amplifying said nucleic acid molecule; and instructions for using the probe and primers for facilitating the diagnosis of disease.

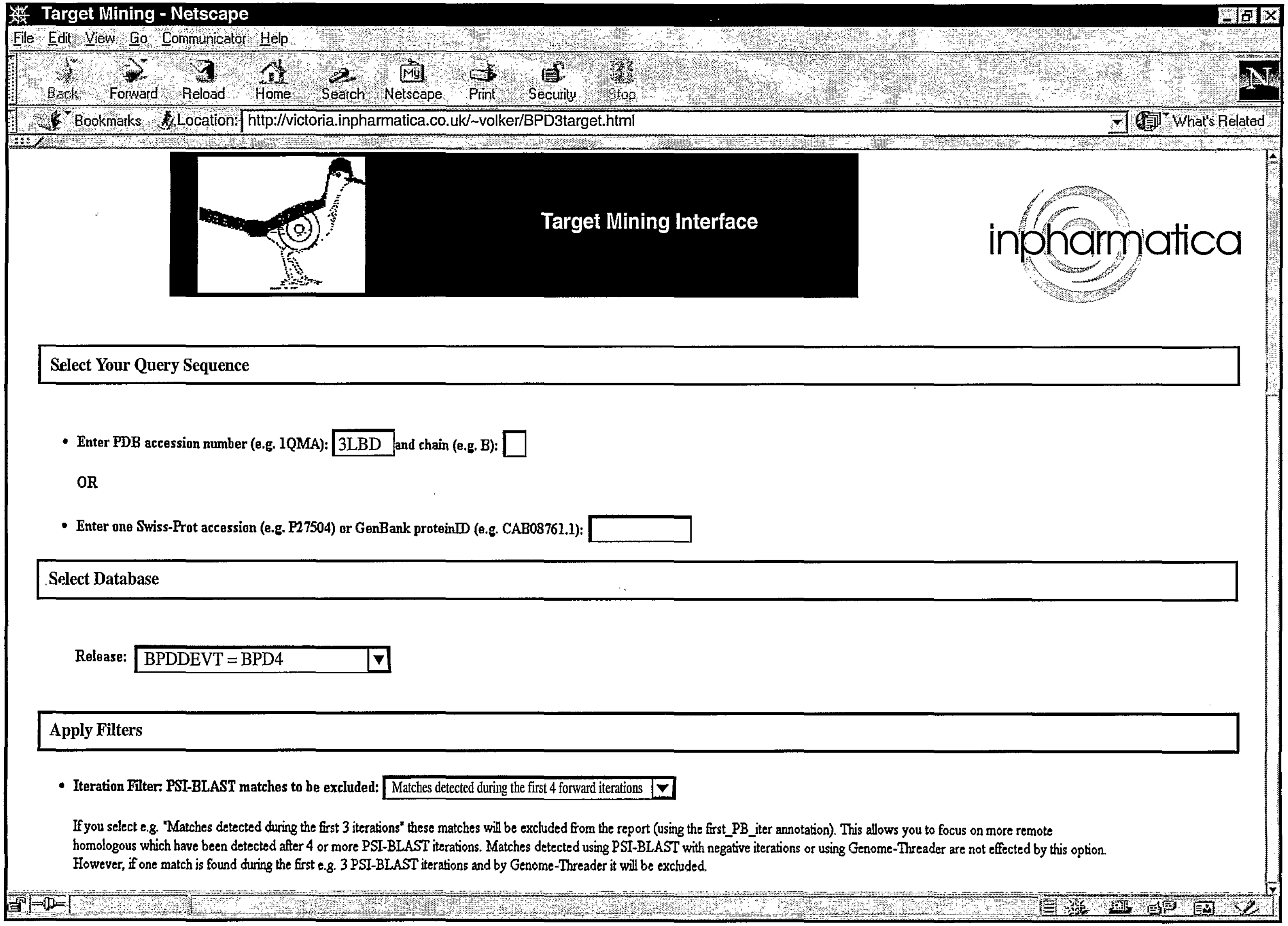
41. The kit of claim 40, further comprising a third container holding an agent for digesting unhybridised RNA.

42. A kit comprising an array of nucleic acid molecules, at least one of which is a nucleic acid molecule according to any one of claims 8-11.

43. A kit comprising one or more antibodies that bind to a polypeptide as recited in any one of claims 1-7; and a reagent useful for the detection of a binding reaction between said antibody and said polypeptide.

44. A transgenic or knockout non-human animal that has been transformed to express higher, lower or absent levels of a polypeptide according to any one of claims 1-7.

45. A method for screening for a compound effective to treat disease, by contacting a non-human transgenic animal according to claim 44 with a candidate compound and determining the effect of the compound on the disease of the animal.



Target Mining Interface

Select Your Query Sequence

- Enter PDB accession number (e.g. 1QMA): 3LBD and chain (e.g. B): ☐

OR

- Enter one Swiss-Prot accession (e.g. P27504) or GenBank proteinID (e.g. CAB08761.1): ☐

Select Database

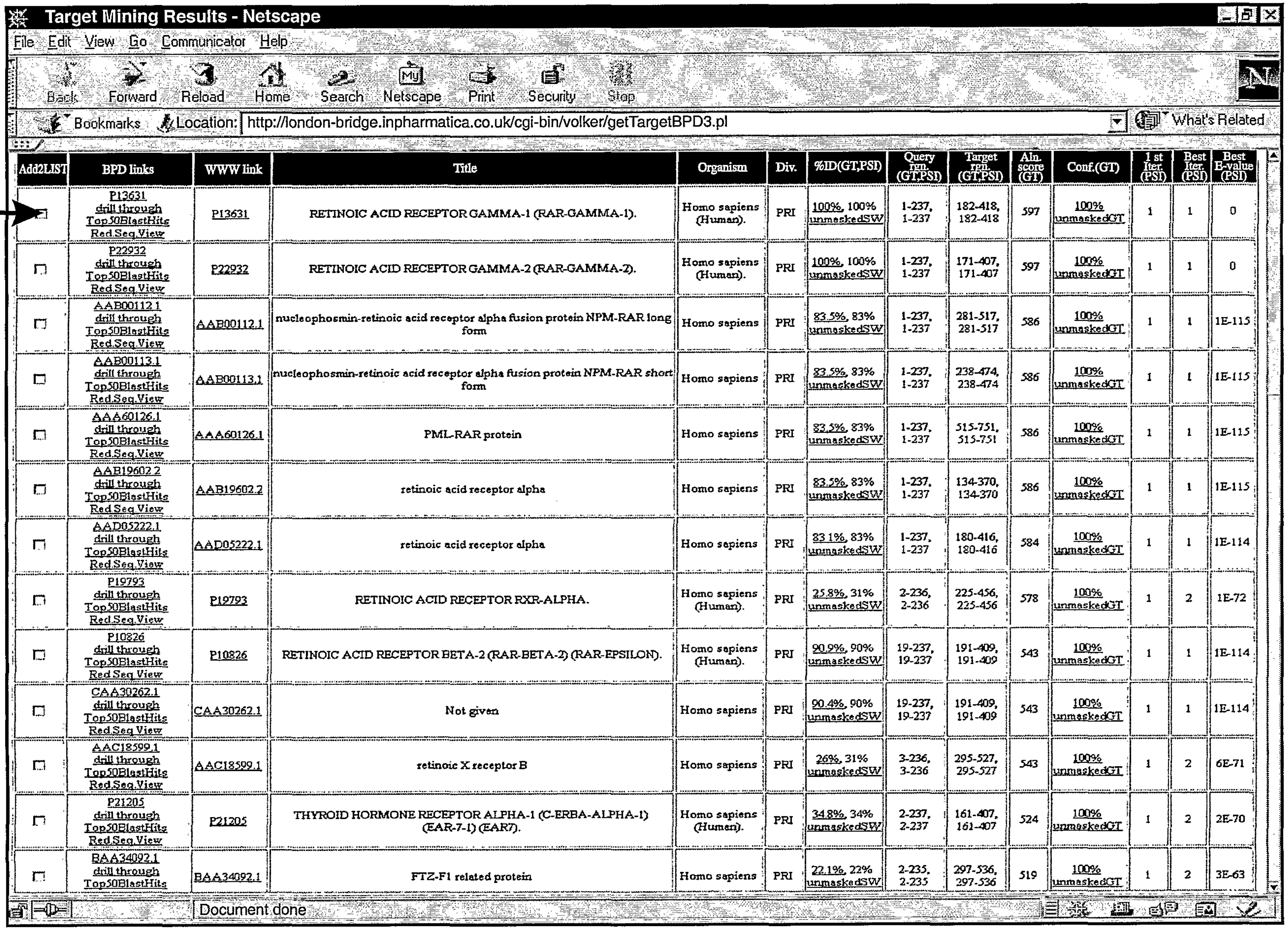
Release: BPDDEVT = BPD4 ▼

Apply Filters

- Iteration Filter: PSI-BLAST matches to be excluded: Matches detected during the first 4 forward iterations ▼

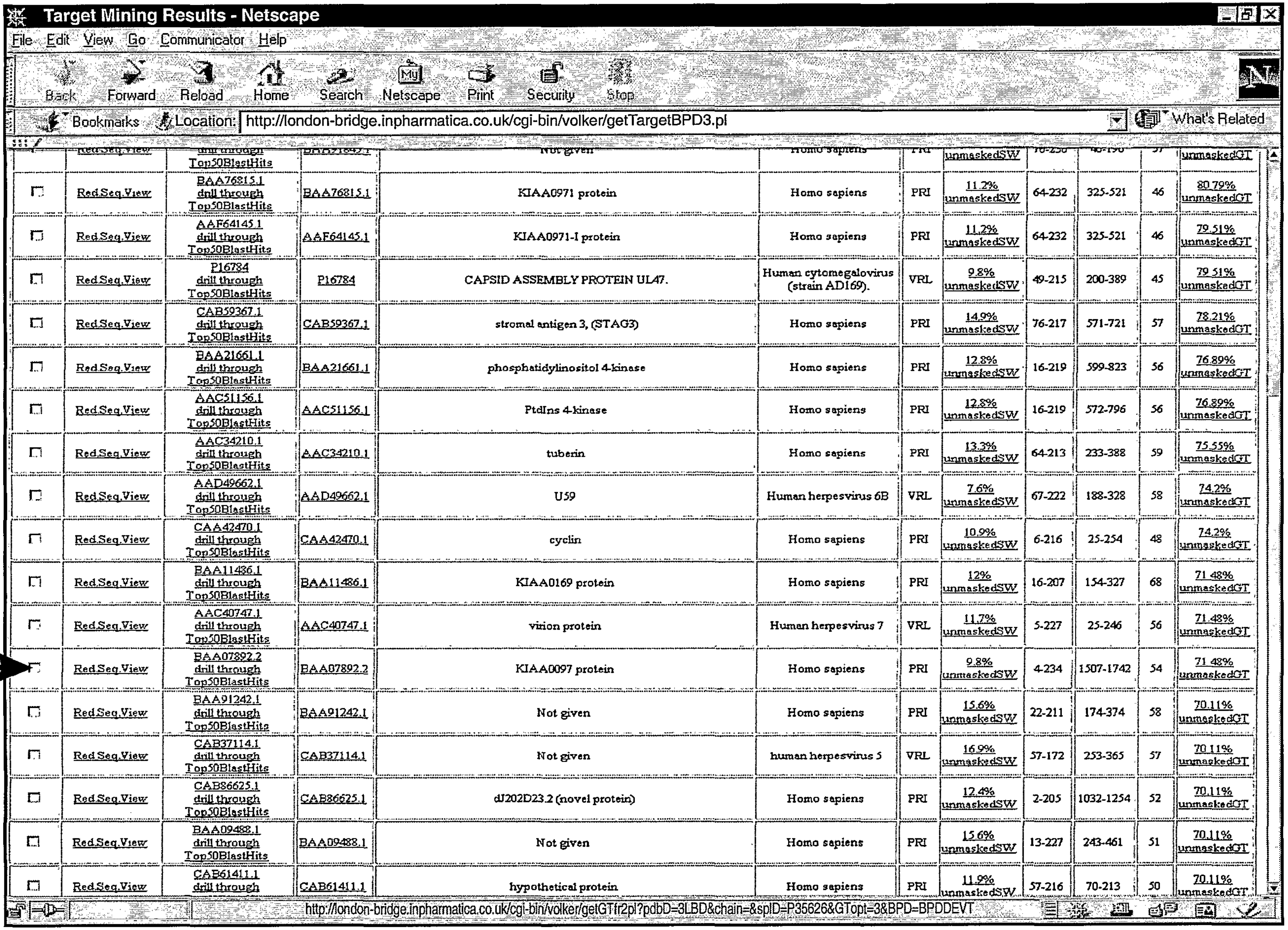
If you select e.g. "Matches detected during the first 3 iterations" these matches will be excluded from the report (using the first_PB_iter annotation). This allows you to focus on more remote homologous which have been detected after 4 or more PSI-BLAST iterations. Matches detected using PSI-BLAST with negative iterations or using Genome-Threader are not effected by this option. However, if one match is found during the first e.g. 3 PSI-BLAST iterations and by Genome-Threader it will be excluded.

FIG. 1



Add2LIST	BPD links	WWW link	Title	Organism	Div.	%ID(GT,PSI)	Query rgn. (GT,PSI)	Target rgn. (GT,PSI)	Aln. score (GT)	Conf.(GT)	1 st Iter. (PSI)	Best Iter. (PSI)	Best E-value (PSI)
☐	P13631 drill through Top50BlastHits Red.Seq.View	P13631	RETINOIC ACID RECEPTOR GAMMA-1 (RAR-GAMMA-1).	Homo sapiens (Human).	PRI	100%, 100% unmaskedSW	1-237, 1-237	182-418, 182-418	597	100% unmaskedGT	1	1	0
☐	P22932 drill through Top50BlastHits Red.Seq.View	P22932	RETINOIC ACID RECEPTOR GAMMA-2 (RAR-GAMMA-2).	Homo sapiens (Human).	PRI	100%, 100% unmaskedSW	1-237, 1-237	171-407, 171-407	597	100% unmaskedGT	1	1	0
☐	AAB00112.1 drill through Top50BlastHits Red.Seq.View	AAB00112.1	nucleophosmin-retinoic acid receptor alpha fusion protein NPM-RAR long form	Homo sapiens	PRI	83.5%, 83% unmaskedSW	1-237, 1-237	281-517, 281-517	586	100% unmaskedGT	1	1	1E-115
☐	AAB00113.1 drill through Top50BlastHits Red.Seq.View	AAB00113.1	nucleophosmin-retinoic acid receptor alpha fusion protein NPM-RAR short form	Homo sapiens	PRI	83.5%, 83% unmaskedSW	1-237, 1-237	238-474, 238-474	586	100% unmaskedGT	1	1	1E-115
☐	AAA60126.1 drill through Top50BlastHits Red.Seq.View	AAA60126.1	PML-RAR protein	Homo sapiens	PRI	83.5%, 83% unmaskedSW	1-237, 1-237	515-751, 515-751	586	100% unmaskedGT	1	1	1E-115
☐	AAB19602.2 drill through Top50BlastHits Red.Seq.View	AAB19602.2	retinoic acid receptor alpha	Homo sapiens	PRI	83.5%, 83% unmaskedSW	1-237, 1-237	134-370, 134-370	586	100% unmaskedGT	1	1	1E-115
☐	AAD05222.1 drill through Top50BlastHits Red.Seq.View	AAD05222.1	retinoic acid receptor alpha	Homo sapiens	PRI	83.1%, 83% unmaskedSW	1-237, 1-237	180-416, 180-416	584	100% unmaskedGT	1	1	1E-114
☐	P19793 drill through Top50BlastHits Red.Seq.View	P19793	RETINOIC ACID RECEPTOR RXR-ALPHA.	Homo sapiens (Human).	PRI	25.8%, 31% unmaskedSW	2-236, 2-236	225-456, 225-456	578	100% unmaskedGT	1	2	1E-72
☐	P10826 drill through Top50BlastHits Red.Seq.View	P10826	RETINOIC ACID RECEPTOR BETA-2 (RAR-BETA-2) (RAR-EPSILON).	Homo sapiens (Human).	PRI	90.9%, 90% unmaskedSW	19-237, 19-237	191-409, 191-409	543	100% unmaskedGT	1	1	1E-114
☐	CAA30262.1 drill through Top50BlastHits Red.Seq.View	CAA30262.1	Not given	Homo sapiens	PRI	90.4%, 90% unmaskedSW	19-237, 19-237	191-409, 191-409	543	100% unmaskedGT	1	1	1E-114
☐	AAC18599.1 drill through Top50BlastHits Red.Seq.View	AAC18599.1	retinoic X receptor B	Homo sapiens	PRI	26%, 31% unmaskedSW	3-236, 3-236	295-527, 295-527	543	100% unmaskedGT	1	2	6E-71
☐	P21205 drill through Top50BlastHits Red.Seq.View	P21205	THYROID HORMONE RECEPTOR ALPHA-1 (C-ERBA-ALPHA-1) (EAR-7-1) (EAR7).	Homo sapiens (Human).	PRI	34.8%, 34% unmaskedSW	2-237, 2-237	161-407, 161-407	524	100% unmaskedGT	1	2	2E-70
☐	BAA34092.1 drill through Top50BlastHits	BAA34092.1	FTZ-F1 related protein	Homo sapiens	PRI	22.1%, 22% unmaskedSW	2-235, 2-235	297-536, 297-536	519	100% unmaskedGT	1	2	3E-63

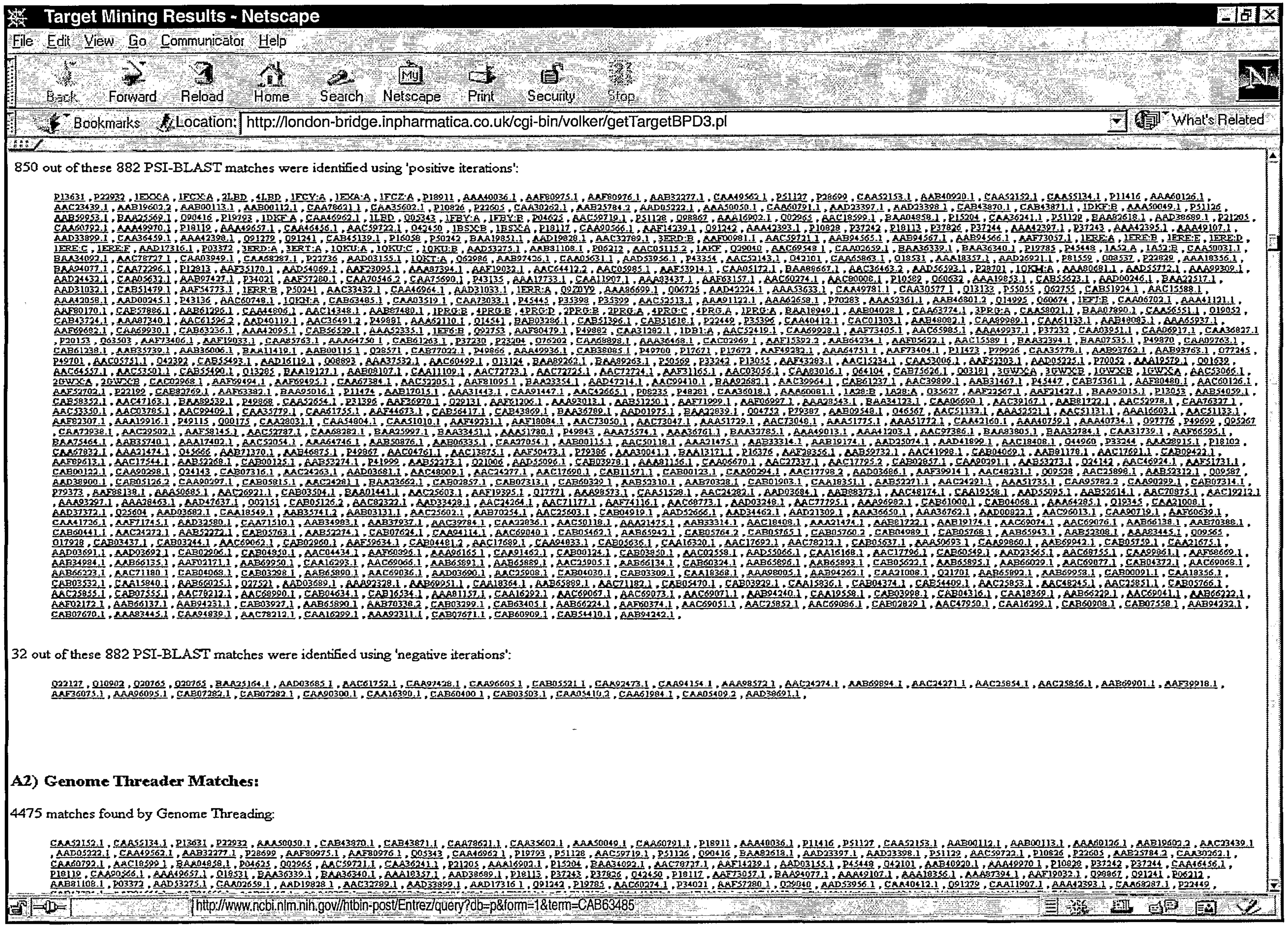
FIG. 2A



	Red.Seq.View	Top50BlastHits		Not given	Homo sapiens		unmaskedSW				unmaskedGT
☐	Red.Seq.View	BAA76815.1 drill through Top50BlastHits	BAA76815.1	KIAA0971 protein	Homo sapiens	PRI	11.2% unmaskedSW	64-232	325-521	46	80.79% unmaskedGT
☐	Red.Seq.View	AAF64145.1 drill through Top50BlastHits	AAF64145.1	KIAA0971-I protein	Homo sapiens	PRI	11.2% unmaskedSW	64-232	325-521	46	79.51% unmaskedGT
☐	Red.Seq.View	P16784 drill through Top50BlastHits	P16784	CAPSID ASSEMBLY PROTEIN UL47.	Human cytomegalovirus (strain AD169).	VRL	9.8% unmaskedSW	49-215	200-389	45	79.51% unmaskedGT
☐	Red.Seq.View	CAB59367.1 drill through Top50BlastHits	CAB59367.1	stromal antigen 3, (STAG3)	Homo sapiens	PRI	14.9% unmaskedSW	76-217	571-721	57	78.21% unmaskedGT
☐	Red.Seq.View	BAA21661.1 drill through Top50BlastHits	BAA21661.1	phosphatidylinositol 4-kinase	Homo sapiens	PRI	12.8% unmaskedSW	16-219	599-823	56	76.89% unmaskedGT
☐	Red.Seq.View	AAC51156.1 drill through Top50BlastHits	AAC51156.1	PtdIns 4-kinase	Homo sapiens	PRI	12.8% unmaskedSW	16-219	572-796	56	76.89% unmaskedGT
☐	Red.Seq.View	AAC34210.1 drill through Top50BlastHits	AAC34210.1	tuberin	Homo sapiens	PRI	13.3% unmaskedSW	64-213	233-388	59	75.55% unmaskedGT
☐	Red.Seq.View	AAD49662.1 drill through Top50BlastHits	AAD49662.1	U59	Human herpesvirus 6B	VRL	7.6% unmaskedSW	67-222	188-328	58	74.2% unmaskedGT
☐	Red.Seq.View	CAA42470.1 drill through Top50BlastHits	CAA42470.1	cyclin	Homo sapiens	PRI	10.9% unmaskedSW	6-216	25-254	48	74.2% unmaskedGT
☐	Red.Seq.View	BAA11486.1 drill through Top50BlastHits	BAA11486.1	KIAA0169 protein	Homo sapiens	PRI	12% unmaskedSW	16-207	154-327	68	71.48% unmaskedGT
☐	Red.Seq.View	AAC40747.1 drill through Top50BlastHits	AAC40747.1	virion protein	Human herpesvirus 7	VRL	11.7% unmaskedSW	5-227	25-246	56	71.48% unmaskedGT
☐	Red.Seq.View	BAA07892.2 drill through Top50BlastHits	BAA07892.2	KIAA0097 protein	Homo sapiens	PRI	9.8% unmaskedSW	4-234	1507-1742	54	71.48% unmaskedGT
☐	Red.Seq.View	BAA91242.1 drill through Top50BlastHits	BAA91242.1	Not given	Homo sapiens	PRI	15.6% unmaskedSW	22-211	174-374	58	70.11% unmaskedGT
☐	Red.Seq.View	CAB37114.1 drill through Top50BlastHits	CAB37114.1	Not given	human herpesvirus 5	VRL	16.9% unmaskedSW	57-172	253-365	57	70.11% unmaskedGT
☐	Red.Seq.View	CAB86625.1 drill through Top50BlastHits	CAB86625.1	dJ202D23.2 (novel protein)	Homo sapiens	PRI	12.4% unmaskedSW	2-205	1032-1254	52	70.11% unmaskedGT
☐	Red.Seq.View	BAA09488.1 drill through Top50BlastHits	BAA09488.1	Not given	Homo sapiens	PRI	15.6% unmaskedSW	13-227	243-461	51	70.11% unmaskedGT
☐	Red.Seq.View	CAB61411.1 drill through	CAB61411.1	hypothetical protein	Homo sapiens	PRI	11.9% unmaskedSW	57-216	70-213	50	70.11% unmaskedGT

http://london-bridge.inpharmatica.co.uk/cgi-bin/volker/getGTfr2pl?pdbD=3LBD&chain=&spID=P35626&GTopt=3&BPD=BPDDEVT

FIG. 2B



Location: http://london-bridge.inpharmatica.co.uk/cgi-bin/volker/getTargetBPD3.pl

850 out of these 882 PSI-BLAST matches were identified using 'positive iterations':

P13631 , P22932 , 1EXX:A , 1FCX:A , 2LBD , 4LBD , 1FCY:A , 1EXA:A , 1FCZ:A , P18911 , AAA40036.1 , AAF80975.1 , AAF80976.1 , AAB32277.1 , CAA49562.1 , P51127 , P28699 , CAA52153.1 , AAB40920.1 , CAA52152.1 , CAA55134.1 , P11416 , AAA60126.1 ,
AAC23439.1 , AAB19602.2 , AAB00113.1 , AAB00112.1 , CAA78621.1 , CAA35602.1 , P10826 , P22605 , CAA30262.1 , AAB25784.2 , AAD05222.1 , AAA50050.1 , CAA60791.1 , AAD23397.1 , AAD23398.1 , CAB43870.1 , CAB43871.1 , 1DKF:B , AAA50049.1 , P51126 ,
AAB59953.1 , BAA25569.1 , Q90416 , P19793 , 1DKF:A , CAA46962.1 , 1LBD , Q05343 , 1FBY:A , 1FBY:B , P04625 , AAC59719.1 , P51128 , Q98867 , AAA16902.1 , Q02965 , AAC18599.1 , BAA04858.1 , P15204 , CAA36241.1 , P51129 , BAA82618.1 , AAD38689.1 , P21205 ,
CAA60792.1 , AAA49970.1 , P18119 , AAA49657.1 , CAA46456.1 , AAC59722.1 , O42450 , 1BSX:B , 1BSX:A , P18117 , CAA90566.1 , AAF14239.1 , Q91242 , AAA42393.1 , P10828 , P37242 , P18113 , P37826 , P37244 , AAA42397.1 , P37243 , AAA42395.1 , AAA49107.1 ,
AAD33899.1 , CAA36459.1 , AAA42398.1 , Q91279 , Q91241 , CAB45139.1 , P16058 , P50242 , BAA19851.1 , AAD19828.1 , AAC32789.1 , 3ERD:B , AAF00981.1 , AAC59721.1 , AAB94565.1 , AAB94567.1 , AAB94566.1 , AAF73057.1 , 1ERE:A , 1ERE:B , 1ERE:E , 1ERE:D ,
1ERE:C , 1ERE:F , AAD17316.1 , P03372 , 3ERD:A , 3ERT:A , 1QKU:A , 1QKU:C , 1QKU:B , AAD53275.1 , AAB81108.1 , P06212 , AAC05115.2 , 1AKF , Q39040 , AAC69548.1 , CAA02659.1 , BAA36339.1 , BAA36340.1 , P19785 , P45448 , 1A52:A , 1A52:B , CAA50031.1 ,
BAA34092.1 , AAC78727.1 , CAA03949.1 , CAA68287.1 , P22736 , AAD03155.1 , 1QKT:A , Q62986 , AAB97426.1 , CAA05631.1 , AAD53956.1 , P43354 , AAC52143.1 , O42101 , CAA65863.1 , Q18531 , AAA18357.1 , AAD26921.1 , P81559 , O08537 , P22829 , AAA18356.1 ,
BAA94077.1 , CAA72296.1 , P12813 , AAF35170.1 , AAD54069.1 , AAF23095.1 , AAA87394.1 , AAF19032.1 , AAC64412.2 , AAC05985.1 , AAF53914.1 , CAA05172.1 , BAA88667.1 , AAC36463.2 , AAD56593.1 , P28701 , 1QKM:A , AAA80681.1 , AAD55772.1 , AAA99309.1 ,
AAD24432.1 , CAA05632.1 , AAB97427.1 , P34021 , AAF57280.1 , CAA70546.2 , CAA75690.1 , P43135 , AAA17733.1 , CAA11907.1 , AAA83437.1 , AAF63157.1 , AAC60274.1 , AAC80008.1 , P10589 , O60632 , AAA19853.1 , CAB55623.1 , AAD00246.1 , BAA22517.1 ,
AAD31032.1 , CAB51479.1 , AAF54773.1 , 1ERR:B , P50241 , AAC33432.1 , CAA46964.1 , AAD31033.1 , 1ERR:A , Q9Z0Y9 , AAA86699.1 , Q06725 , AAD42224.1 , AAA53633.1 , CAA49781.1 , CAA30577.1 , Q13133 , P55055 , Q62755 , CAB51924.1 , AAC15588.1 ,
AAA42058.1 , AAD00245.1 , P43136 , AAC60748.1 , 1QKN:A , CAB63485.1 , CAA03519.1 , CAA73033.1 , P45445 , P35398 , P35399 , AAC52513.1 , AAA91122.1 , AAA62658.1 , P70283 , AAA52361.1 , AAB46801.2 , Q14995 , O60674 , 1EFJ:B , CAA06702.1 , AAA41121.1 ,
AAF80170.1 , CAB57886.1 , AAB61296.1 , CAA44806.1 , AAC14348.1 , AAB87480.1 , 1PRG:B , 4PRG:B , 4PRG:D , 2PRG:B , 2PRG:A , 4PRG:C , 4PRG:A , 1PRG:A , BAA18949.1 , AAB04028.1 , CAA63774.1 , 3PRG:A , CAA58021.1 , BAA07890.1 , CAA56551.1 , O19052 ,
CAB43724.1 , AAA87340.1 , AAC61596.2 , AAD40119.1 , AAC36491.2 , P49881 , AAA62110.1 , Q14541 , BAB03286.1 , CAB51396.1 , CAB51618.1 , P22449 , P35396 , CAA40412.1 , CAC01303.1 , AAB48082.1 , CAA89989.1 , CAA61133.1 , AAB48083.1 , AAA65937.1 ,
AAF69682.1 , CAA69930.1 , CAB63236.1 , AAA42095.1 , CAB56539.1 , AAA52335.1 , 1EF6:B , Q92753 , AAF80479.1 , P49882 , CAA31282.1 , 1DB1:A , AAC52419.1 , CAA69928.1 , AAF73405.1 , AAC65985.1 , AAA49937.1 , P37232 , CAA03951.1 , CAA06917.1 , CAA36827.1
, P20153 , Q63503 , AAF73406.1 , AAF19033.1 , CAA85763.1 , AAA64750.1 , CAB61263.1 , P37230 , P23204 , O76202 , CAA68898.1 , AAA36468.1 , CAC02969.1 , AAF15392.2 , AAB64234.1 , AAF05622.1 , AAC15589.1 , BAA32394.1 , BAA07535.1 , P49870 , CAA09763.1 ,
CAB61238.1 , AAB35739.1 , AAB36006.1 , BAA11419.1 , AAB00115.1 , Q28571 , CAB77022.1 , P49866 , AAA49936.1 , CAB38085.1 , P49700 , P17671 , P17672 , AAF49282.1 , AAA64751.1 , AAF73404.1 , P11473 , P79926 , CAA35778.1 , AAB93762.1 , AAB93763.1 , O77245 ,
P49701 , AAC05751.1 , O42392 , CAB55493.1 , AAD16119.1 , Q08893 , AAA37532.1 , AAC60499.1 , O13124 , BAA89262.1 , BAA89263.1 , P50569 , P33242 , P13055 , AAF43283.1 , AAC15234.1 , CAA53006.1 , AAF52303.1 , AAD05225.1 , P70052 , AAA19579.1 , Q01639 ,
AAC64557.1 , AAC53501.1 , CAB55490.1 , Q13285 , BAA19127.1 , AAB08107.1 , CAA11109.1 , AAC72723.1 , AAC72725.1 , AAC72724.1 , AAF31165.1 , AAC03056.1 , CAA83016.1 , Q64104 , CAB75626.1 , O03181 , 3GWX:A , 3GWX:B , 1GWX:B , 1GWX:A , AAC53066.1 ,
2GWX:A , 2GWX:B , CAC02968.1 , AAF69494.1 , AAF69495.1 , CAA67384.1 , AAC52205.1 , AAF81095.1 , BAA23354.1 , AAD47214.1 , AAC99410.1 , BAA92682.1 , AAC39964.1 , CAB61237.1 , AAC39899.1 , AAB31467.1 , P45447 , CAB75361.1 , AAF80480.1 , AAC60126.1 ,
AAF52702.1 , P22199 , CAB82769.1 , AAF63382.1 , BAA95016.1 , P11474 , AAB17015.1 , AAA31443.1 , CAA91447.1 , AAC42665.1 , P08235 , P48281 , CAA36018.1 , AAA60081.1 , 1A28:B , 1A28:A , O35627 , AAF22567.1 , AAF21427.1 , BAA95015.1 , P13053 , AAB54059.1 ,
CAB58352.1 , AAC47163.1 , BAA89539.1 , P49868 , CAA52654.1 , P31396 , AAF36970.1 , Q29131 , AAF61206.1 , AAA93013.1 , AAB51250.1 , AAF71999.1 , AAF06997.1 , AAA28543.1 , BAA34123.1 , CAA06690.1 , AAC39167.1 , AAB81722.1 , AAC52978.1 , CAA76327.1 ,
AAC53350.1 , AAC03785.1 , AAC99409.1 , CAA35779.1 , CAA61755.1 , AAF44673.1 , CAB56417.1 , CAB43869.1 , BAA36789.1 , AAD01975.1 , BAA22839.1 , Q04752 , P79387 , AAB09548.1 , O46567 , AAC51132.1 , AAA52521.1 , AAC51131.1 , AAA16603.1 , AAC51133.1 ,
AAF82307.1 , AAA19916.1 , P49115 , Q00175 , CAA28031.1 , CAA54804.1 , CAA51010.1 , AAF49231.1 , AAF18084.1 , AAC73050.1 , AAC73047.1 , AAA51729.1 , AAC73048.1 , AAA51775.1 , AAA51772.1 , CAA42160.1 , AAA40759.1 , AAA40734.1 , O97776 , P49699 , Q95267
, CAA72938.1 , AAC29502.1 , AAF58145.1 , AAC52787.1 , CAA68282.1 , BAA25997.1 , BAA33451.1 , AAA51780.1 , P49843 , AAA75574.1 , AAA36761.1 , BAA32785.1 , AAA49013.1 , AAA41203.1 , AAC97386.1 , BAA83805.1 , BAA32784.1 , CAA31739.1 , AAF66595.1 ,
BAA75464.1 , AAB35740.1 , AAA17402.1 , AAC52054.1 , AAA64746.1 , AAB50876.1 , AAB06335.1 , CAA27054.1 , AAB00115.1 , AAC50118.1 , AAA21475.1 , AAB33314.1 , AAB19174.1 , AAD25074.1 , AAD41899.1 , AAC18408.1 , Q44960 , P33244 , AAA28915.1 , P18102 ,
CAA67832.1 , AAA21474.1 , O45666 , AAB71370.1 , AAB46875.1 , P49867 , AAC04761.1 , AAC13875.1 , AAF50473.1 , P79386 , AAA30041.1 , BAA13171.1 , P16376 , AAF28356.1 , AAB59732.1 , AAC41998.1 , CAB04069.1 , AAB81178.1 , AAC17691.1 , CAB09422.1 ,
AAF89613.1 , AAC17544.1 , AAB52268.1 , CAB00125.1 , AAB53274.1 , P41999 , AAB52273.1 , Q21006 , AAD55096.1 , CAB03978.1 , AAA81156.1 , CAA06670.1 , AAC27337.1 , AAC17795.2 , CAB02857.1 , CAA90391.1 , AAB53273.1 , Q24142 , AAC46924.1 , AAF51731.1 ,
CAB00122.1 , CAA90298.1 , Q24143 , CAB07316.1 , AAC24263.1 , AAD03681.1 , AAC48009.1 , AAC24277.1 , AAC17690.1 , CAB11571.1 , CAB00123.1 , CAA90294.1 , AAC17798.2 , AAD03686.1 , AAF39914.1 , AAC48231.1 , Q09528 , AAC25898.1 , AAB52312.1 , Q09587 ,
AAD38900.1 , CAB05126.2 , CAA90297.1 , CAB05815.1 , AAC24281.1 , BAA23662.1 , CAB02857.1 , CAB07313.1 , CAB60329.1 , AAB52310.1 , AAB70328.1 , CAB01903.1 , CAA18351.1 , AAB52271.1 , AAC24291.1 , AAA51735.1 , CAA95782.2 , CAA90299.1 , CAB07314.1 ,
P79373 , AAF88138.1 , AAA50685.1 , AAC26921.1 , CAB03504.1 , BAA01441.1 , AAC25603.1 , AAF19395.1 , Q17771 , AAA98573.1 , CAA51528.1 , AAC24282.1 , AAD03684.1 , AAB88373.1 , AAC48174.1 , CAA19558.1 , AAD55095.1 , AAB52614.1 , AAC70875.1 , AAC19212.1
, AAA93297.1 , AAA28463.1 , AAD47637.1 , O02151 , CAB05126.2 , AAC82322.1 , AAD33428.1 , AAC24264.1 , AAC71177.1 , AAF74116.1 , AAC68773.1 , AAD03248.1 , AAC77795.1 , AAA96982.1 , CAB61000.1 , CAB04068.1 , AAA64285.1 , Q19345 , CAA21008.1 ,
AAD37372.1 , Q25604 , AAD03682.1 , CAA18549.1 , AAB35741.2 , AAB03131.1 , AAC25602.1 , AAB70254.1 , AAC25603.1 , CAB04919.1 , AAD52666.1 , AAD34462.1 , AAD21309.1 , AAA36650.1 , AAA36762.1 , AAD00822.1 , AAC96013.1 , CAA90719.1 , AAF60639.1 ,
CAA41726.1 , AAF71745.1 , AAD32580.1 , CAA71510.1 , AAB34983.1 , AAB37937.1 , AAC39784.1 , CAA22836.1 , AAC50118.1 , AAA21475.1 , AAB33314.1 , AAC18408.1 , AAA21474.1 , AAB81722.1 , AAB19174.1 , AAC69074.1 , AAC69076.1 , AAB66138.1 , AAB70388.1 ,
CAB60441.1 , AAC24273.1 , AAB52272.1 , CAB05763.1 , AAB52274.1 , CAB07624.1 , CAA94114.1 , AAC69040.1 , CAB05462.1 , AAB65943.1 , CAB05764.2 , CAB05765.1 , CAB05760.2 , CAB04989.1 , CAB05768.1 , AAB65943.1 , AAB52308.1 , AAA83445.1 , Q09565 ,
O17928 , CAB03437.1 , CAB03244.1 , AAC69062.1 , CAB02960.1 , AAF59634.1 , CAB04481.2 , AAC17689.1 , CAA94833.1 , CAB05636.1 , CAA16320.1 , AAC17692.1 , AAC78212.1 , CAB05637.1 , AAA50693.1 , CAA99860.1 , AAB69942.1 , CAB05759.1 , CAA21675.1 ,
AAD03691.1 , AAD03692.1 , CAB02906.1 , CAB04850.1 , AAC04434.1 , AAF60396.1 , AAA96165.1 , CAA91462.1 , CAB00124.1 , CAB03850.1 , AAC02558.1 , AAD55066.1 , CAA16168.1 , AAC17796.1 , CAB60549.1 , AAD23565.1 , AAC68755.1 , CAA99861.1 , AAF68669.1 ,
AAB34984.1 , AAB66135.1 , AAF02171.1 , AAB69950.1 , CAA16293.1 , AAC69066.1 , AAB65891.1 , AAB65889.1 , AAC25905.1 , AAB66134.1 , CAB60324.1 , AAB65896.1 , AAB65893.1 , CAB05622.1 , AAB65895.1 , AAB66029.1 , AAC69077.1 , CAB04372.1 , AAC69068.1 ,
AAB66223.1 , AAC71180.1 , CAB04068.1 , CAB03298.1 , AAB65890.1 , AAC69036.1 , AAD03690.1 , AAC25908.1 , CAB04030.1 , CAB03309.1 , CAA18368.1 , AAA98005.1 , AAB94262.1 , CAA21008.1 , Q21701 , AAB65892.1 , AAB69958.1 , CAB00091.1 , CAA18356.1 ,
CAB03532.1 , CAA15840.1 , AAB66025.1 , Q27521 , AAD03689.1 , AAA92328.1 , AAB69951.1 , CAA18364.1 , AAB65889.1 , AAC71182.1 , CAB05470.1 , CAB03929.1 , CAA15836.1 , CAB04374.1 , CAB54409.1 , AAC25853.1 , AAC48245.1 , AAC25851.1 , CAB05766.1 ,
AAC25855.1 , CAB07555.1 , AAC78212.1 , AAC68990.1 , CAB04634.1 , CAB16534.1 , AAA81157.1 , CAA16292.1 , AAC69067.1 , AAC69073.1 , AAC69071.1 , AAB94240.1 , CAA19558.1 , CAB03998.1 , CAB04316.1 , CAA18369.1 , AAB66229.1 , AAC69041.1 , AAB66222.1 ,
AAF02172.1 , AAB66137.1 , AAB94231.1 , CAB03927.1 , AAB65890.1 , AAB70338.2 , CAB03299.1 , CAB63405.1 , AAB66224.1 , AAF60374.1 , AAC69051.1 , AAC25852.1 , AAC69086.1 , CAB02829.1 , AAC47950.1 , CAA16299.1 , CAB60908.1 , CAB07558.1 , AAB94232.1 ,
CAB07670.1 , AAA83445.1 , CAA94839.1 , AAC78212.1 , CAA16299.1 , AAA92311.1 , CAB07671.1 , CAB60909.1 , CAB54410.1 , AAB94242.1 ,

32 out of these 882 PSI-BLAST matches were identified using 'negative iterations':

Q22127 , Q10902 , Q20765 , Q20765 , BAA25164.1 , AAD03685.1 , AAC61752.1 , CAA97428.1 , CAA96605.1 , CAB05521.1 , CAA92473.1 , CAA94154.1 , AAA98572.1 , AAC24274.1 , AAB69894.1 , AAC24271.1 , AAC25854.1 , AAC25856.1 , AAB69901.1 , AAF39918.1 ,
AAF36075.1 , AAA96095.1 , CAB07282.1 , CAB07282.1 , CAA90300.1 , CAA16390.1 , CAB60400.1 , CAB03503.1 , CAA05410.2 , CAA61984.1 , CAA05409.2 , AAD38691.1 ,

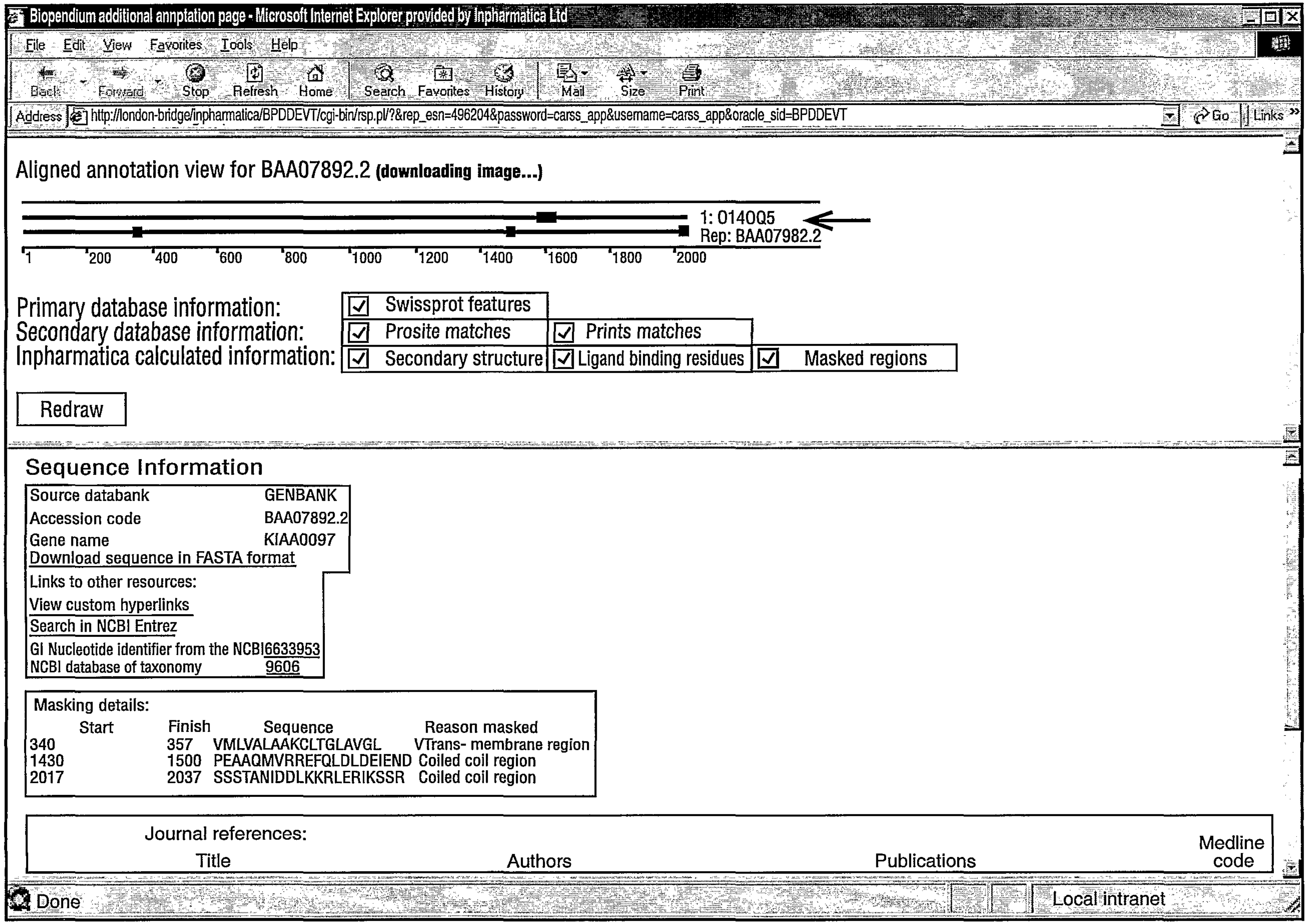
A2) Genome Threader Matches:

4475 matches found by Genome Threading:

CAA52152.1 , CAA55134.1 , P13631 , P22932 , AAA50050.1 , CAB43870.1 , CAB43871.1 , CAA78621.1 , CAA35602.1 , AAA50049.1 , CAA60791.1 , P18911 , AAA40036.1 , P11416 , P51127 , CAA52153.1 , AAB00112.1 , AAB00113.1 , AAA60126.1 , AAB19602.2 , AAC23439.1
, AAD05222.1 , CAA49562.1 , AAB32277.1 , P28699 , AAF80975.1 , AAF80976.1 , Q05343 , CAA46962.1 , P19793 , P51128 , AAC59719.1 , P51126 , Q90416 , BAA82618.1 , AAD23397.1 , AAD23398.1 , P51129 , AAC59722.1 , P10826 , P22605 , AAB25784.2 , CAA30262.1 ,
CAA60792.1 , AAC18599.1 , BAA04858.1 , P04625 , Q02965 , AAC59721.1 , CAA36241.1 , P21205 , AAA16902.1 , P15204 , BAA34092.1 , AAC78727.1 , AAF14239.1 , AAD03155.1 , P45448 , O42101 , AAB40920.1 , AAA49970.1 , P10828 , P37242 , P37244 , CAA46456.1 ,
P18119 , CAA90566.1 , AAA49657.1 , Q18531 , BAA36339.1 , BAA36340.1 , AAA18357.1 , AAD38689.1 , P18113 , P37243 , P37826 , O42450 , P18117 , AAF73057.1 , BAA94077.1 , AAA49107.1 , AAA18356.1 , AAA87394.1 , AAF19032.1 , Q98867 , Q91241 , P06212 ,
AAB81108.1 , P03372 , AAD53275.1 , CAA02659.1 , AAD19828.1 , AAC32789.1 , AAD33899.1 , AAD17316.1 , Q91242 , P19785 , AAC60274.1 , P34021 , AAF57280.1 , Q39040 , AAD53956.1 , CAA40412.1 , Q91279 , CAA11907.1 , AAA42393.1 , CAA68287.1 , P22449 ,

http://www.ncbi.nlm.nih.gov//htbin-post/Entrez/query?db=p&form=1&term=CAB63485

FIG. 2C



http://london-bridge/inpharmatica/BPDDEVT/cgi-bin/rsp.pl/?&rep_esn=496204&password=carss_app&username=carss_app&oracle_sid=BPDDEVT

Aligned annotation view for BAA07892.2 **(downloading image...)**

1: 014OQ5
Rep: BAA07982.2

Primary database information: ☑ Swissprot features
Secondary database information: ☑ Prosite matches ☑ Prints matches
Inpharmatica calculated information: ☑ Secondary structure ☑ Ligand binding residues ☑ Masked regions

Redraw

Sequence Information

Source databank	GENBANK
Accession code	BAA07892.2
Gene name	KIAA0097
Download sequence in FASTA format	
Links to other resources:	
View custom hyperlinks	
Search in NCBI Entrez	
GI Nucleotide identifier from the NCBI	6633953
NCBI database of taxonomy	9606

Masking details:

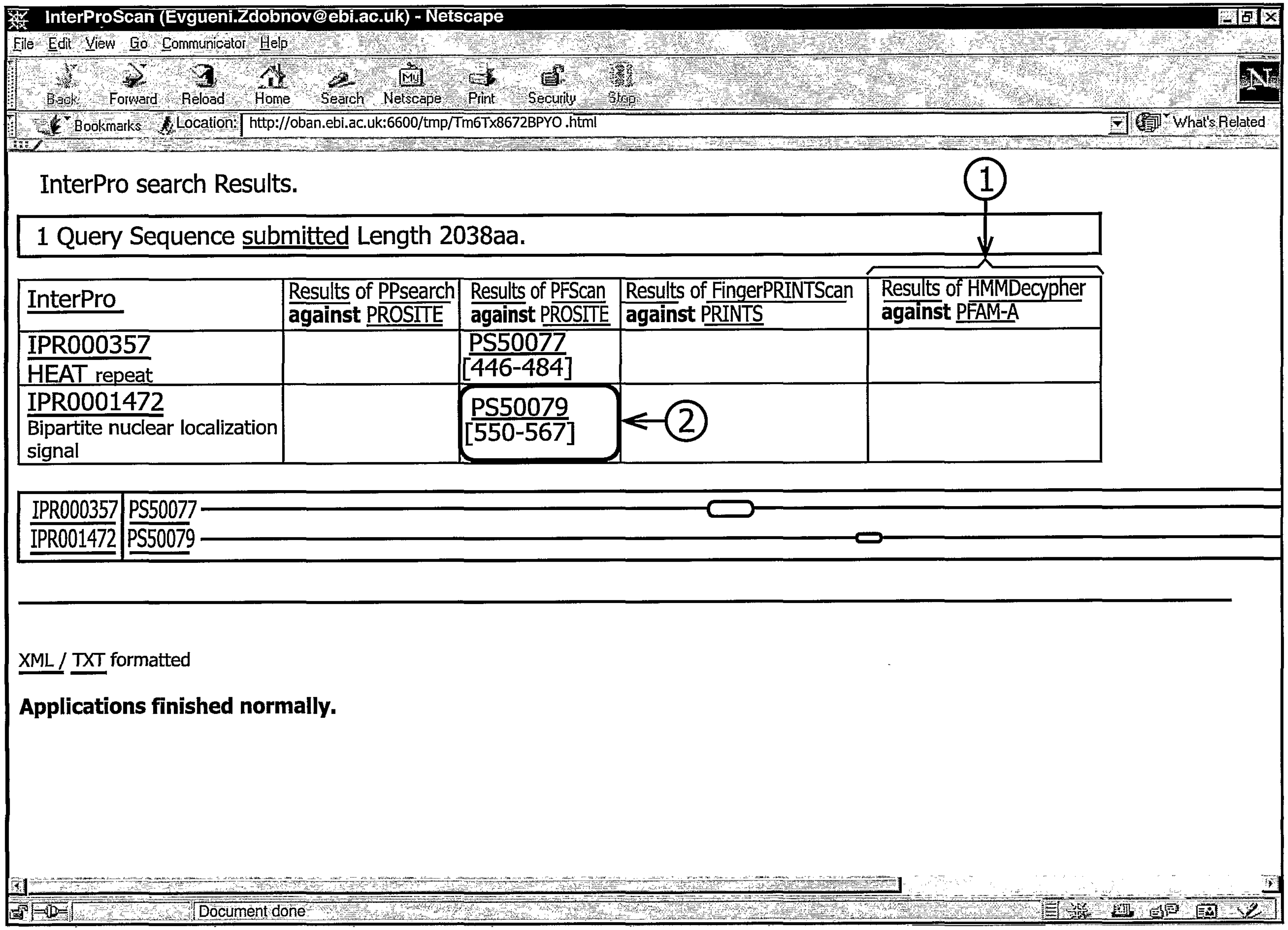
Start	Finish	Sequence	Reason masked
340	357	VMLVALAAKCLTGLAVGL	VTrans- membrane region
1430	1500	PEAAQMVRREFQLDLDEIEND	Coiled coil region
2017	2037	SSSTANIDDLKKRLERIKSSR	Coiled coil region

Journal references:

Title	Authors	Publications	Medline code

FIG. 3

FIG. 4



InterPro search Results.

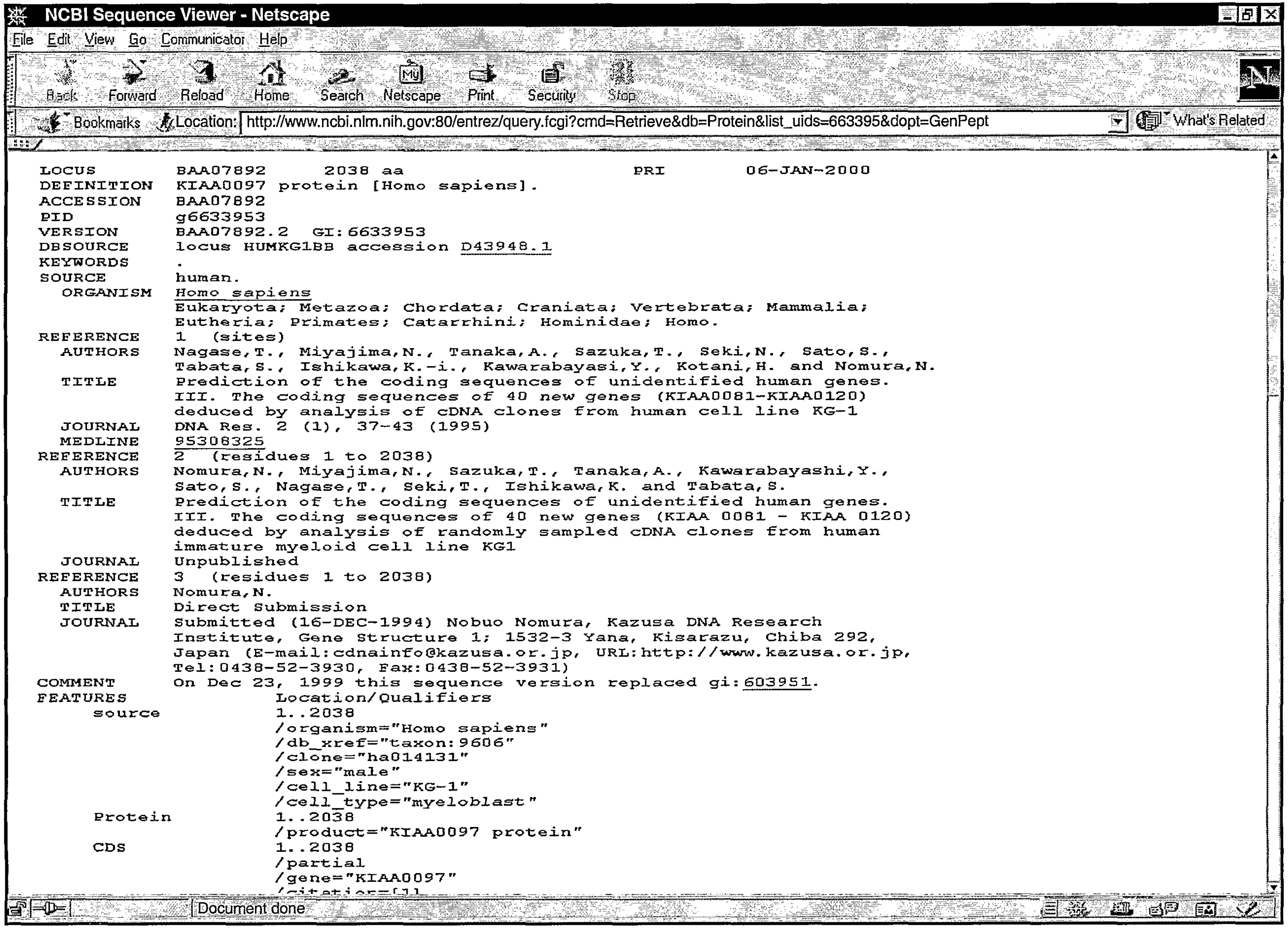
1 Query Sequence submitted Length 2038aa.

InterPro	Results of PPsearch **against** PROSITE	Results of PFScan **against** PROSITE	Results of FingerPRINTScan **against** PRINTS	Results of HMMDecypher **against** PFAM-A
IPR000357 HEAT repeat		PS50077 [446-484]		
IPR0001472 Bipartite nuclear localization signal		PS50079 [550-567]		

IPR000357	PS50077
IPR001472	PS50079

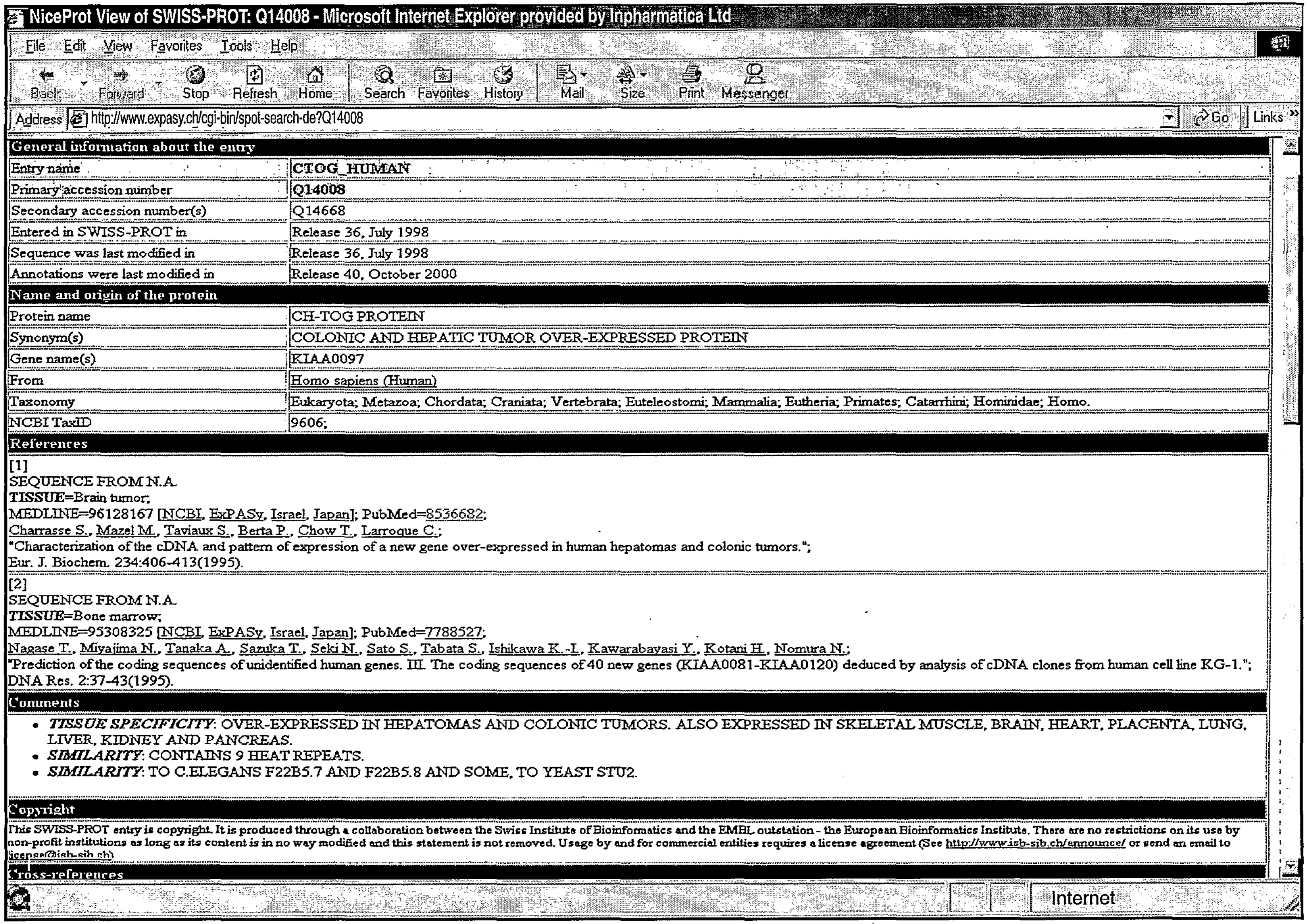
XML / TXT formatted

Applications finished normally.



```
LOCUS       BAA07892      2038 aa                      PRI       06-JAN-2000
DEFINITION  KIAA0097 protein [Homo sapiens].
ACCESSION   BAA07892
PID         g6633953
VERSION     BAA07892.2  GI:6633953
DBSOURCE    locus HUMKG1BB accession D43948.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
            Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (sites)
  AUTHORS   Nagase,T., Miyajima,N., Tanaka,A., Sazuka,T., Seki,N., Sato,S.,
            Tabata,S., Ishikawa,K.-i., Kawarabayasi,Y., Kotani,H. and Nomura,N.
  TITLE     Prediction of the coding sequences of unidentified human genes.
            III. The coding sequences of 40 new genes (KIAA0081-KIAA0120)
            deduced by analysis of cDNA clones from human cell line KG-1
  JOURNAL   DNA Res. 2 (1), 37-43 (1995)
  MEDLINE   95308325
REFERENCE   2  (residues 1 to 2038)
  AUTHORS   Nomura,N., Miyajima,N., Sazuka,T., Tanaka,A., Kawarabayashi,Y.,
            Sato,S., Nagase,T., Seki,T., Ishikawa,K. and Tabata,S.
  TITLE     Prediction of the coding sequences of unidentified human genes.
            III. The coding sequences of 40 new genes (KIAA 0081 - KIAA 0120)
            deduced by analysis of randomly sampled cDNA clones from human
            immature myeloid cell line KG1
  JOURNAL   Unpublished
REFERENCE   3  (residues 1 to 2038)
  AUTHORS   Nomura,N.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-DEC-1994) Nobuo Nomura, Kazusa DNA Research
            Institute, Gene Structure 1; 1532-3 Yana, Kisarazu, Chiba 292,
            Japan (E-mail:cdnainfo@kazusa.or.jp, URL:http://www.kazusa.or.jp,
            Tel:0438-52-3930, Fax:0438-52-3931)
COMMENT     On Dec 23, 1999 this sequence version replaced gi:603951.
FEATURES             Location/Qualifiers
     source          1..2038
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /clone="ha014131"
                     /sex="male"
                     /cell_line="KG-1"
                     /cell_type="myeloblast"
     Protein         1..2038
                     /product="KIAA0097 protein"
     CDS             1..2038
                     /partial
                     /gene="KIAA0097"
                     /citation=[1]
```

FIG. 5A



General information about the entry

Entry name	**CTOG_HUMAN**
Primary accession number	**Q14008**
Secondary accession number(s)	Q14668
Entered in SWISS-PROT in	Release 36, July 1998
Sequence was last modified in	Release 36, July 1998
Annotations were last modified in	Release 40, October 2000

Name and origin of the protein

Protein name	CH-TOG PROTEIN
Synonym(s)	COLONIC AND HEPATIC TUMOR OVER-EXPRESSED PROTEIN
Gene name(s)	KIAA0097
From	Homo sapiens (Human)
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
NCBI TaxID	9606;

References

[1]
SEQUENCE FROM N.A.
TISSUE=Brain tumor;
MEDLINE=96128167 [NCBI, ExPASy, Israel, Japan]; PubMed=8536682;
Charrasse S., Mazel M., Taviaux S., Berta P., Chow T., Larroque C.;
"Characterization of the cDNA and pattern of expression of a new gene over-expressed in human hepatomas and colonic tumors.";
Eur. J. Biochem. 234:406-413(1995).

[2]
SEQUENCE FROM N.A.
TISSUE=Bone marrow;
MEDLINE=95308325 [NCBI, ExPASy, Israel, Japan]; PubMed=7788527;
Nagase T., Miyajima N., Tanaka A., Sazuka T., Seki N., Sato S., Tabata S., Ishikawa K.-I., Kawarabayasi Y., Kotani H., Nomura N.;
"Prediction of the coding sequences of unidentified human genes. III. The coding sequences of 40 new genes (KIAA0081-KIAA0120) deduced by analysis of cDNA clones from human cell line KG-1.";
DNA Res. 2:37-43(1995).

Comments

- ***TISSUE SPECIFICITY***: OVER-EXPRESSED IN HEPATOMAS AND COLONIC TUMORS. ALSO EXPRESSED IN SKELETAL MUSCLE, BRAIN, HEART, PLACENTA, LUNG, LIVER, KIDNEY AND PANCREAS.
- ***SIMILARITY***: CONTAINS 9 HEAT REPEATS.
- ***SIMILARITY***: TO C.ELEGANS F22B5.7 AND F22B5.8 AND SOME, TO YEAST STU2.

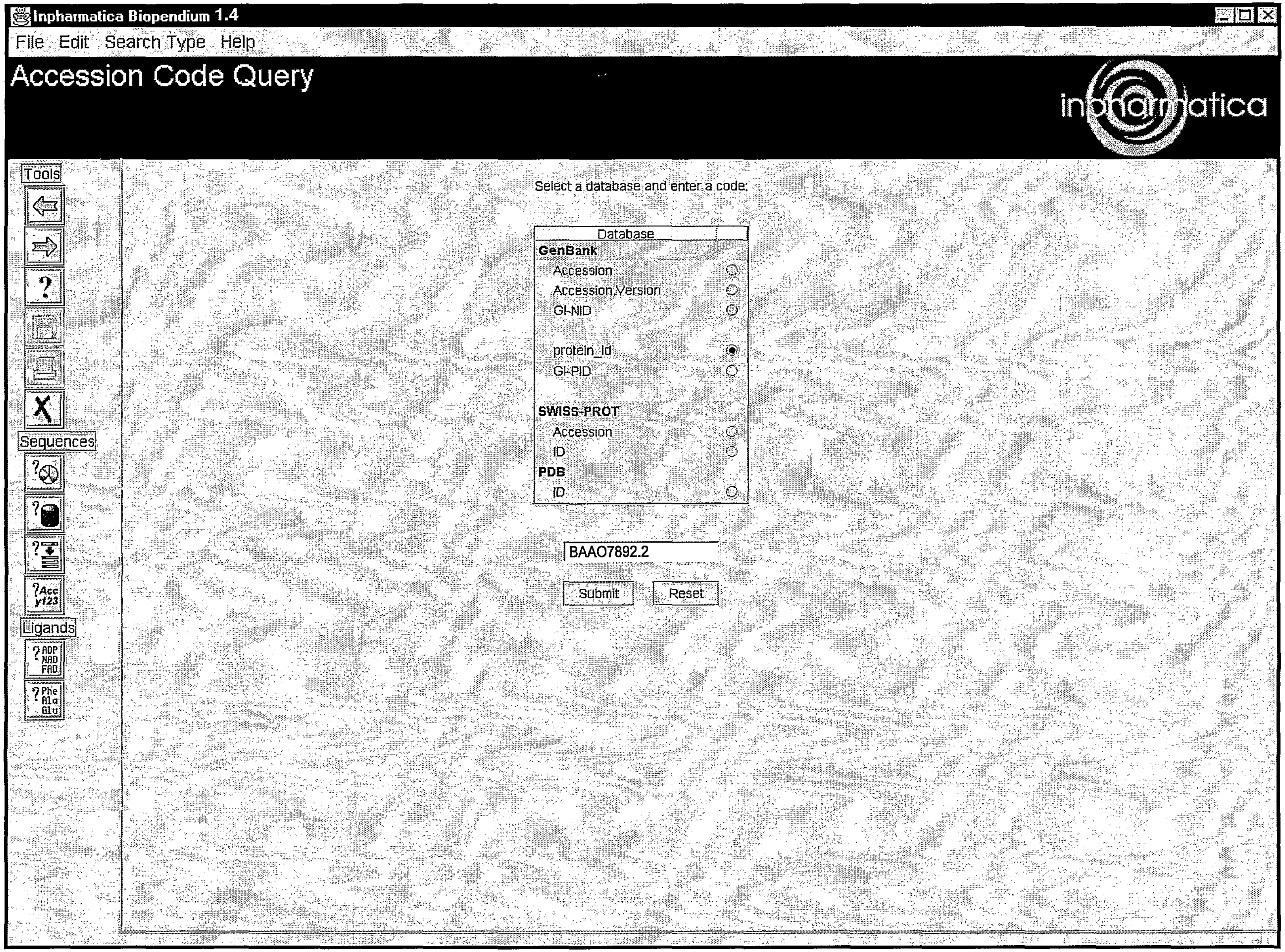
Copyright

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch)

Cross-references

FIG. 5B

FIG. 6A



Accession Code Query

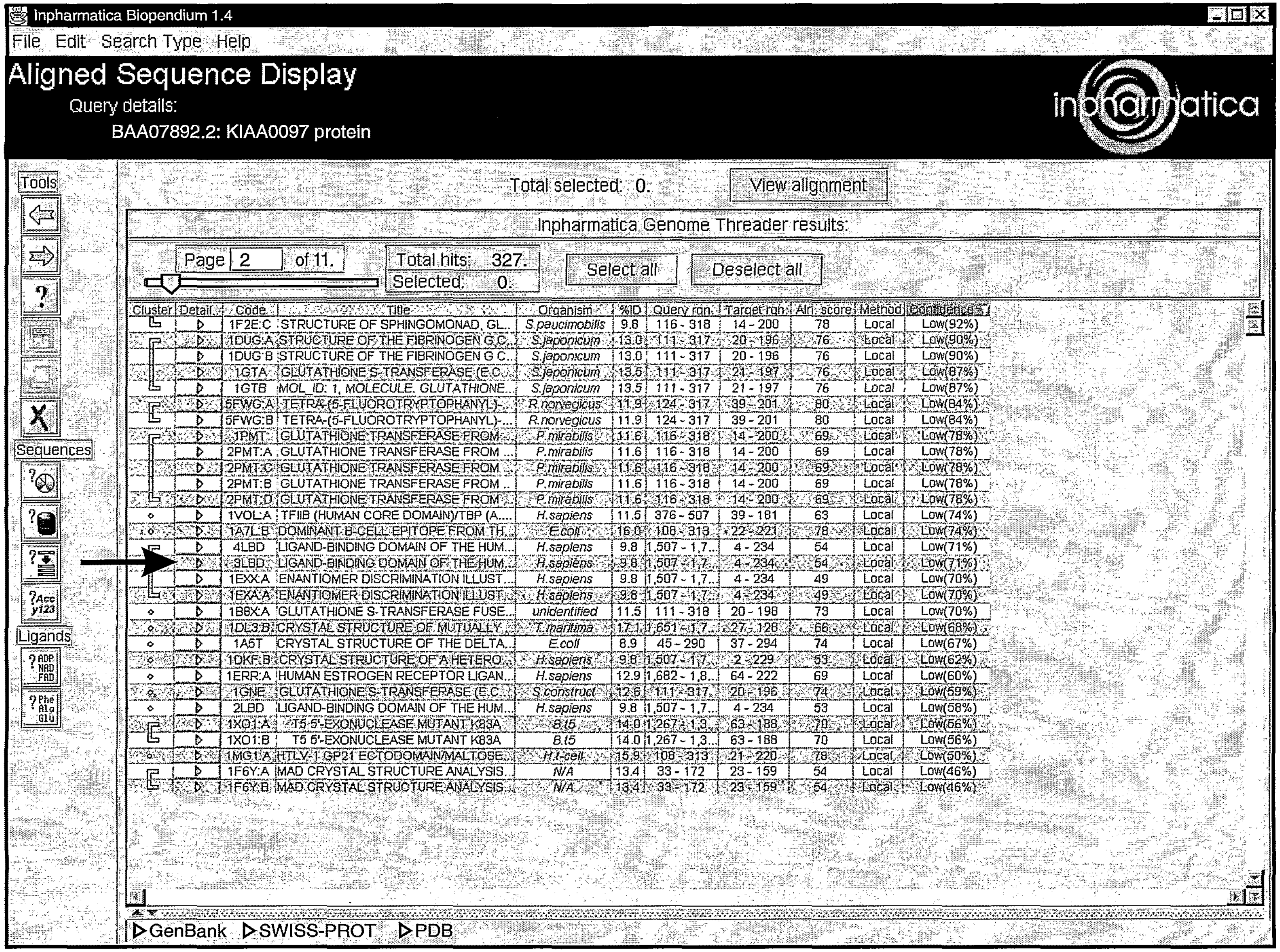
Select a database and enter a code:

Database	
GenBank	
Accession	○
Accession.Version	○
GI-NID	○
protein_id	●
GI-PID	○
SWISS-PROT	
Accession	○
ID	○
PDB	
ID	○

BAAO7892.2

Submit Reset

FIG. 6B



Aligned Sequence Display

Query details:

BAA07892.2: KIAA0097 protein

Total selected: 0.

Inpharmatica Genome Threader results:

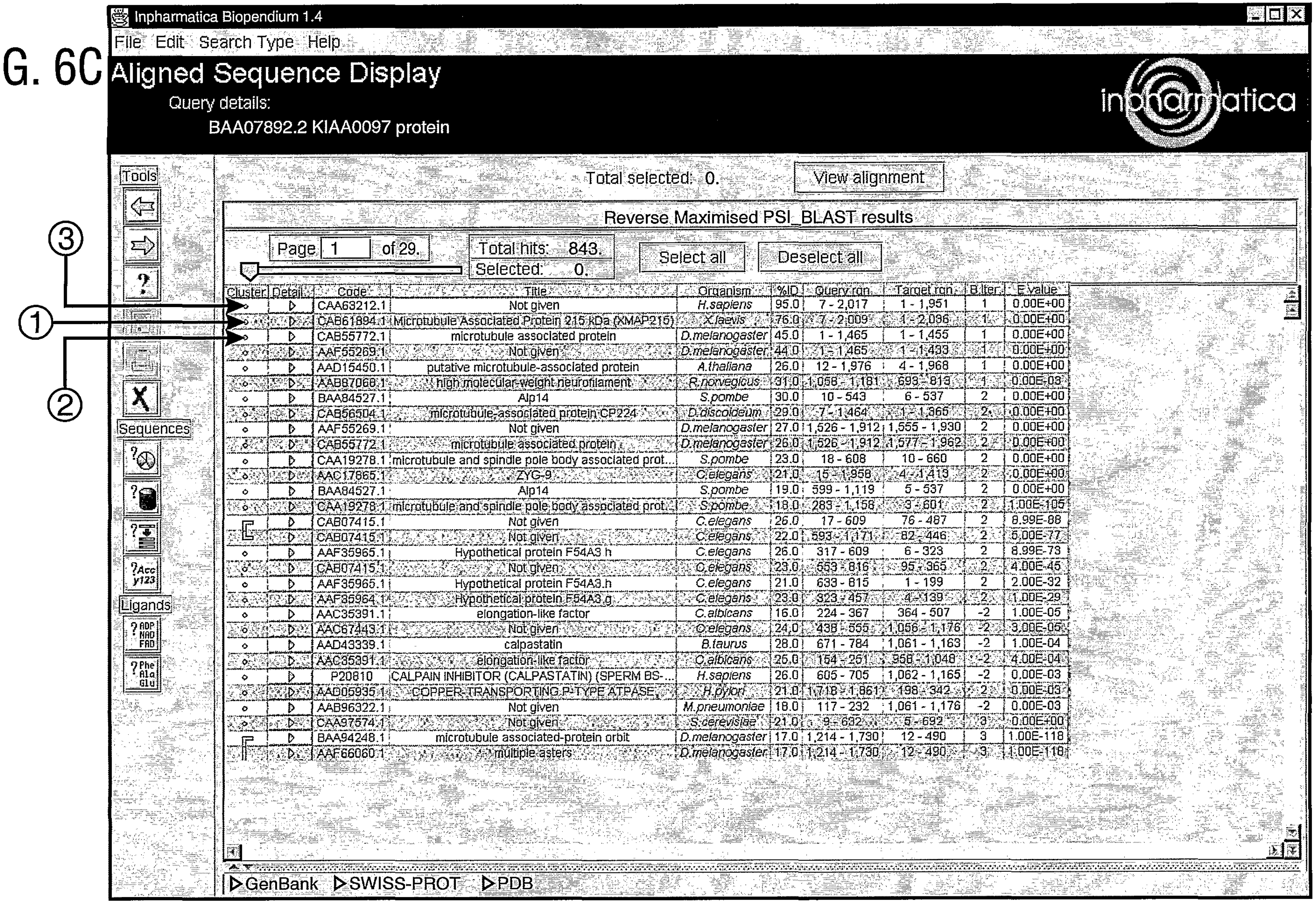
Page 2 of 11.

Total hits: 327.

Selected: 0.

Cluster	Detail	Code	Title	Organism	%ID	Query rgn.	Target rgn.	Aln. score	Method	Confidence
	▷	1F2E:C	STRUCTURE OF SPHINGOMONAD, GL...	S.paucimobilis	9.8	116 - 318	14 - 200	78	Local	Low(92%)
	▷	1DUG:A	STRUCTURE OF THE FIBRINOGEN G C...	S.japonicum	13.0	111 - 317	20 - 196	76	Local	Low(90%)
	▷	1DUG:B	STRUCTURE OF THE FIBRINOGEN G C...	S.japonicum	13.0	111 - 317	20 - 196	76	Local	Low(90%)
	▷	1GTA	GLUTATHIONE S-TRANSFERASE (E.C...	S.japonicum	13.5	111 - 317	21 - 197	76	Local	Low(87%)
	▷	1GTB	MOL_ID: 1, MOLECULE: GLUTATHIONE...	S.japonicum	13.5	111 - 317	21 - 197	76	Local	Low(87%)
	▷	5FWG:A	TETRA-(5-FLUOROTRYPTOPHANYL)-...	R.norvegicus	11.9	124 - 317	39 - 201	80	Local	Low(84%)
	▷	5FWG:B	TETRA-(5-FLUOROTRYPTOPHANYL)-...	R.norvegicus	11.9	124 - 317	39 - 201	80	Local	Low(84%)
	▷	1PMT	GLUTATHIONE TRANSFERASE FROM ...	P.mirabilis	11.6	116 - 318	14 - 200	69	Local	Low(78%)
	▷	2PMT:A	GLUTATHIONE TRANSFERASE FROM ...	P.mirabilis	11.6	116 - 318	14 - 200	69	Local	Low(78%)
	▷	2PMT:C	GLUTATHIONE TRANSFERASE FROM ...	P.mirabilis	11.6	116 - 318	14 - 200	69	Local	Low(78%)
	▷	2PMT:B	GLUTATHIONE TRANSFERASE FROM ..	P.mirabilis	11.6	116 - 318	14 - 200	69	Local	Low(78%)
	▷	2PMT:D	GLUTATHIONE TRANSFERASE FROM ...	P.mirabilis	11.6	116 - 318	14 - 200	69	Local	Low(78%)
◇	▷	1VOL:A	TFIIB (HUMAN CORE DOMAIN)/TBP (A....	H.sapiens	11.5	376 - 507	39 - 181	63	Local	Low(74%)
◇	▷	1A7L:B	DOMINANT B-CELL EPITOPE FROM TH...	E.coli	16.0	108 - 313	22 - 221	78	Local	Low(74%)
	▷	4LBD	LIGAND-BINDING DOMAIN OF THE HUM...	H.sapiens	9.8	1,507 - 1,7...	4 - 234	54	Local	Low(71%)
	▷	3LBD	LIGAND-BINDING DOMAIN OF THE HUM...	H.sapiens	9.8	1,507 - 1,7...	4 - 234	54	Local	Low(71%)
	▷	1EXX:A	ENANTIOMER DISCRIMINATION ILLUST...	H.sapiens	9.8	1,507 - 1,7...	4 - 234	49	Local	Low(70%)
	▷	1EXA:A	ENANTIOMER DISCRIMINATION ILLUST...	H.sapiens	9.8	1,507 - 1,7...	4 - 234	49	Local	Low(70%)
◇	▷	1B8X:A	GLUTATHIONE S-TRANSFERASE FUSE...	unidentified	11.5	111 - 318	20 - 198	73	Local	Low(70%)
◇	▷	1DL3:B	CRYSTAL STRUCTURE OF MUTUALLY ...	T.maritima	17.1	1,651 - 1,7...	27 - 128	66	Local	Low(68%)
◇	▷	1A5T	CRYSTAL STRUCTURE OF THE DELTA...	E.coli	8.9	45 - 290	37 - 294	74	Local	Low(67%)
◇	▷	1DKF:B	CRYSTAL STRUCTURE OF A HETERO...	H.sapiens	9.8	1,507 - 1,7...	2 - 229	53	Local	Low(62%)
◇	▷	1ERR:A	HUMAN ESTROGEN RECEPTOR LIGAN...	H.sapiens	12.9	1,682 - 1,8...	64 - 222	69	Local	Low(60%)
◇	▷	1GNE	GLUTATHIONE S-TRANSFERASE (E.C...	S.construct	12.6	111 - 317	20 - 196	74	Local	Low(59%)
◇	▷	2LBD	LIGAND-BINDING DOMAIN OF THE HUM...	H.sapiens	9.8	1,507 - 1,7...	4 - 234	53	Local	Low(58%)
	▷	1XO1:A	T5 5'-EXONUCLEASE MUTANT K83A	B.t5	14.0	1,267 - 1,3...	63 - 188	70	Local	Low(56%)
	▷	1XO1:B	T5 5'-EXONUCLEASE MUTANT K83A	B.t5	14.0	1,267 - 1,3...	63 - 188	70	Local	Low(56%)
◇	▷	1MG1:A	HTLV-1 GP21 ECTODOMAIN/MALTOSE...	H.t-cell	15.9	108 - 313	21 - 220	78	Local	Low(50%)
	▷	1F6Y:A	MAD CRYSTAL STRUCTURE ANALYSIS...	N/A	13.4	33 - 172	23 - 159	54	Local	Low(46%)
	▷	1F6Y:B	MAD CRYSTAL STRUCTURE ANALYSIS...	N/A	13.4	33 - 172	23 - 159	54	Local	Low(46%)

FIG. 6C



Aligned Sequence Display

Query details:

BAA07892.2 KIAA0097 protein

Total selected: 0.

Reverse Maximised PSI_BLAST results

Page 1 of 29.

Total hits: 843.

Selected: 0.

Cluster	Detail	Code	Title	Organism	%ID	Query rgn.	Target rgn.	B.Iter.	E value
○	▷	CAA63212.1	Not given	*H.sapiens*	95.0	7 - 2,017	1 - 1,951	1	0.00E+00
○	▷	CAB61894.1	Microtubule Associated Protein 215 kDa (XMAP215)	*X.laevis*	76.0	7 - 2,009	1 - 2,096	1	0.00E+00
○	▷	CAB55772.1	microtubule associated protein	*D.melanogaster*	45.0	1 - 1,465	1 - 1,455	1	0.00E+00
○	▷	AAF55269.1	Not given	*D.melanogaster*	44.0	1 - 1,465	1 - 1,433	1	0.00E+00
○	▷	AAD15450.1	putative microtubule-associated protein	*A.thaliana*	26.0	12 - 1,976	4 - 1,968	1	0.00E+00
○	▷	AAB87068.1	high molecular-weight neurofilament	*R.norvegicus*	31.0	1,058 - 1,181	693 - 813	1	0.00E-03
○	▷	BAA84527.1	Alp14	*S.pombe*	30.0	10 - 543	6 - 537	2	0.00E+00
○	▷	CAB56504.1	microtubule-associated protein CP224	*D.discoideum*	29.0	7 - 1,464	1 - 1,366	2	0.00E+00
○	▷	AAF55269.1	Not given	*D.melanogaster*	27.0	1,526 - 1,912	1,555 - 1,930	2	0.00E+00
○	▷	CAB55772.1	microtubule associated protein	*D.melanogaster*	26.0	1,526 - 1,912	1,577 - 1,962	2	0.00E+00
○	▷	CAA19278.1	microtubule and spindle pole body associated prot...	*S.pombe*	23.0	18 - 608	10 - 660	2	0.00E+00
○	▷	AAC17865.1	ZYG-9	*C.elegans*	21.0	15 - 1,958	4 - 1,413	2	0.00E+00
○	▷	BAA84527.1	Alp14	*S.pombe*	19.0	599 - 1,119	5 - 537	2	0.00E+00
○	▷	CAA19278.1	microtubule and spindle pole body associated prot...	*S.pombe*	18.0	283 - 1,158	3 - 601	2	1.00E-105
	▷	CAB07415.1	Not given	*C.elegans*	26.0	17 - 609	76 - 487	2	8.99E-88
	▷	CAB07415.1	Not given	*C.elegans*	22.0	593 - 1,171	82 - 446	2	5.00E-77
○	▷	AAF35965.1	Hypothetical protein F54A3.h	*C.elegans*	26.0	317 - 609	6 - 323	2	8.99E-73
○	▷	CAB07415.1	Not given	*C.elegans*	23.0	553 - 816	95 - 365	2	4.00E-45
○	▷	AAF35965.1	Hypothetical protein F54A3.h	*C.elegans*	21.0	633 - 815	1 - 199	2	2.00E-32
○	▷	AAF35964.1	Hypothetical protein F54A3.g	*C.elegans*	23.0	323 - 457	4 - 139	2	1.00E-29
○	▷	AAC35391.1	elongation-like factor	*C.albicans*	16.0	224 - 367	364 - 507	-2	1.00E-05
○	▷	AAC67443.1	Not given	*C.elegans*	24.0	438 - 555	1,056 - 1,176	-2	3.00E-05
○	▷	AAD43339.1	calpastatin	*B.taurus*	28.0	671 - 784	1,061 - 1,163	-2	1.00E-04
○	▷	AAC35391.1	elongation-like factor	*C.albicans*	25.0	154 - 261	958 - 1,048	-2	4.00E-04
○	▷	P20810	CALPAIN INHIBITOR (CALPASTATIN) (SPERM BS-...	*H.sapiens*	26.0	605 - 705	1,062 - 1,165	-2	0.00E-03
○	▷	AAD05935.1	COPPER-TRANSPORTING P-TYPE ATPASE,	*H.pylori*	21.0	1,718 - 1,861	198 - 342	2	0.00E-03
○	▷	AAB96322.1	Not given	*M.pneumoniae*	18.0	117 - 232	1,061 - 1,176	-2	0.00E-03
○	▷	CAA97574.1	Not given	*S.cerevisiae*	21.0	9 - 632	5 - 692	3	0.00E+00
	▷	BAA94248.1	microtubule associated-protein orbit	*D.melanogaster*	17.0	1,214 - 1,730	12 - 490	3	1.00E-118
	▷	AAF66060.1	multiple asters	*D.melanogaster*	17.0	1,214 - 1,730	12 - 490	3	1.00E-118

FIG. 7-1

AlEye output (February 20, 2001 1:05 PM)

```
                              10                  20                  30                  40                  50                  60                  70
                              |                   |                   |                   |                   |                   |                   |
BAA07892.2  G K P G S T M G D D S E W L K L P V D Q K C E H K L W K A R L S G Y E E A L K I F Q K I K D E K S P E W S K F L G L I K K F V T D S N A V V Q L K G L E A A L
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

            80                  90                  100                 110                 120                 130                 140                 150
            |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V Y V E N A H V A G K T T G E V V S G V V S K V F N Q P K A K A K E L G I E I C L M Y I E I E K G E A V Q E E L L K G L D N K N P K I I V A C I E T L R K A L
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

              160                 170                 180                 190                 200                 210                 220                 230
              |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  S E F G S K I I L L K P I I K V L P K L F E S R E K A V R D E A K L I A V E I Y R W I R D A L R P P L Q N I N S V Q L K E L E E E W V K L P T S A P R P T R F
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                240                 250                 260                 270                 280                 290                 300                 310
                |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  L R S Q Q E L E A K L E Q Q Q S A G G D A E G G G D D G D E V P Q I D A Y E L L E A V E I L S K L P K D F Y D K I E A K K W Q E R K E A L E S V E V L I K N P
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                  320                 330                 340                 350                 360                 370                 380                 390
                  |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  K L E A G D Y A D L V K A L K K V V G K D T N V M L V A L A A K C L T G L A V G L R K K F G Q Y A G H V V P T I L E K F K E K K P Q V V Q A L Q E A I D A I F
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                    400                 410                 420                 430                 440                 450                 460                 470
                    |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  L T T T L Q N I S E D V L A V M D N K N P T I K Q Q T S L F I A R S F R H C T A S T L P K S L L K P F C A A L L K H I N D S A P E V R D A A F E A L G T A L K
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                      480                 490                 500                 510                 520                 530                 540                 550
                      |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V V G E K A V K P F L A D V D K L K L D K I K E C S E K V E L I H G K K A G L A A D K K E F K P L P G R T A A S G A A G D K D T K D I S A P K P G P L K K A P
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                        560                 570                 580                 590                 600                 610                 620                 630
                        |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  A A K A G G P P K K G K P A A P G G A G N T G T K N K K G L E T K E I V E P E L S I E V C E E K A S A V L P P T C I Q L L D S S N W K E R L A C M E E F Q K A
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                          640                 650                 660                 670                 680                 690                 700                 710
                          |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V E L M D R T E M P C Q A L V R M L A K K P G W K E T N F Q V M Q M K L H I V A L I A Q K G N F S K T S A Q V V L D G L V D K I G D V K C G N N A K E A M T A
3LBD        - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

FIG. 7-2

AlEye output (February 20, 2001 1:05 PM)

```
                  720       730       740       750       760       770       780       790
                    |         |         |         |         |         |         |         |
BAA07892.2  IAEACMLPWTAEQVVSMAFSQKNPKNQSETLNWLSNAIKEFGFSGLNVKAFISNVKTALAATNPAVRTAAITLLGVMYL
3LBD        -------------------------------------------------------------------------------

                   800       810       820       830       840       850       860
                     |         |         |         |         |         |         |
BAA07892.2  YVGPSLRMFFEDEKPALLSQIDAEFEKMQGQSPPAPTRGISKHSTSGTDEGEDGDEPDDGSNDVVDLLPRTEISDKITS
3LBD        -------------------------------------------------------------------------------

          870       880       890       900       910       920       930       940
            |         |         |         |         |         |         |         |
BAA07892.2  ELVSKIGDKNWKIRKEGLDEVAGIINDAKFIQPNIGELPTALKGRLNDSNKILVQQTLNILQQLAVAMGPNIKQHVKNL
3LBD        -------------------------------------------------------------------------------

           950       960       970       980       990      1000      1010      1020
             |         |         |         |         |         |         |         |
BAA07892.2  GIPIITVLGDSKNNVRAAALATVNAWAEQTGMKEWLEGEDLSEELKKENPFLRQELLGWLAEKLPTLRSTPTDLILCVP
3LBD        -------------------------------------------------------------------------------

           1030      1040      1050      1060      1070      1080      1090      1100
              |         |         |         |         |         |         |         |
BAA07892.2  HLYSCLEDRNGDVRKKAQDALPFFMMHLGYEKMAKATGKLKPTSKDQVLAMLEKAKVNMPAKPAPPTKATSKPMGGSAP
3LBD        -------------------------------------------------------------------------------

            1110      1120      1130      1140      1150      1160      1170      1180
               |         |         |         |         |         |         |         |
BAA07892.2  AKFQPASAPAEDCISSSTEPKPDPKKAKAPGLSSKAKSAQGKKMPSKTSLKEDEDKSGPIFIVVPNGKEQRMKDEKGLK
3LBD        -------------------------------------------------------------------------------

             1190      1200      1210      1220      1230      1240      1250      1260
                |         |         |         |         |         |         |         |
BAA07892.2  VLKWNFTTPRDEYIEQLKTQMSSCVAKWLQDEMFHSDFQHHNKALAVMVDHLESEKEGVIGCLDLILKWLTLRFFDTNT
3LBD        -------------------------------------------------------------------------------

              1270      1280      1290      1300      1310      1320      1330      1340
                 |         |         |         |         |         |         |         |
BAA07892.2  SVLMKALEYLKLLFTLLSEEEYHLTENEASSFIPYLVVKVGEPKDVIRKDVRAILNRMCLVYPASKMFPFIMEGTKSKN
3LBD        -------------------------------------------------------------------------------

               1350      1360      1370      1380      1390      1400      1410      1420
                  |         |         |         |         |         |         |         |
BAA07892.2  SKQRAECLEELGCLVESYGMNVCQPTPGKALKEIAVHIGDRDNAVRNAALNTIVTVYNVHGDQVFKLIGNLSEKDMSML
3LBD        -------------------------------------------------------------------------------
```

FIG. 7-3

AlEye output (February 20, 2001 1:05 PM)

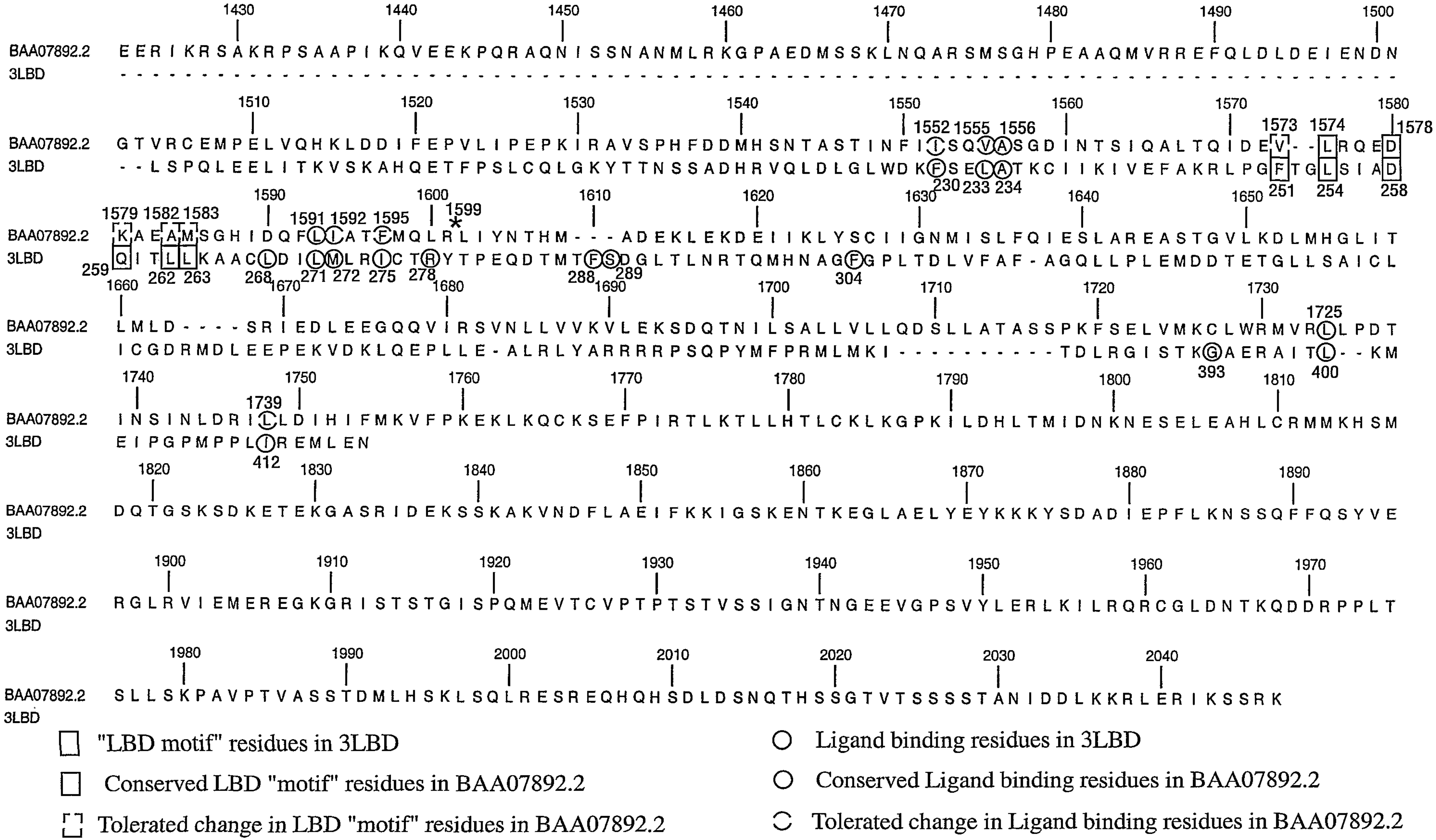
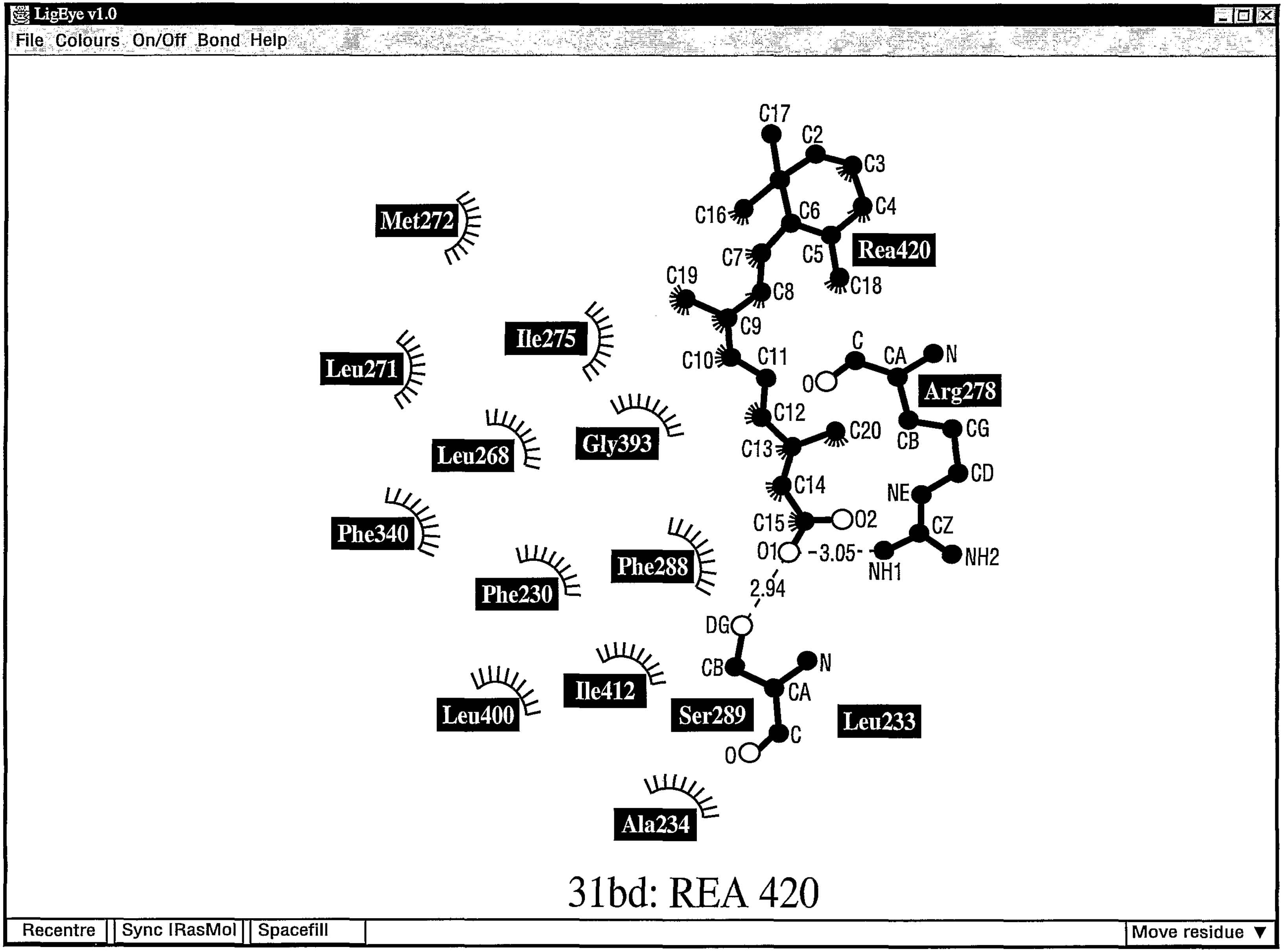


FIG. 8A


LigEye v1.0
File Colours On/Off Bond Help
C17
C2
C3
C4
C16
C6
C5
Rea420
C7
C19
C8
C18
C9
C10
C11
C
CA
N
O
Arg278
C12
C20
CB
CG
C13
C14
CD
NE
C15
O2
CZ
O1
-3.05-
NH1
NH2
2.94
DG
CB
N
CA
C
O
Ser289
Leu233
Met272
Ile275
Leu271
Leu268
Gly393
Phe340
Phe230
Phe288
Leu400
Ile412
Ala234
31bd: REA 420
Recentre
Sync IRasMol
Spacefill
Move residue ▼

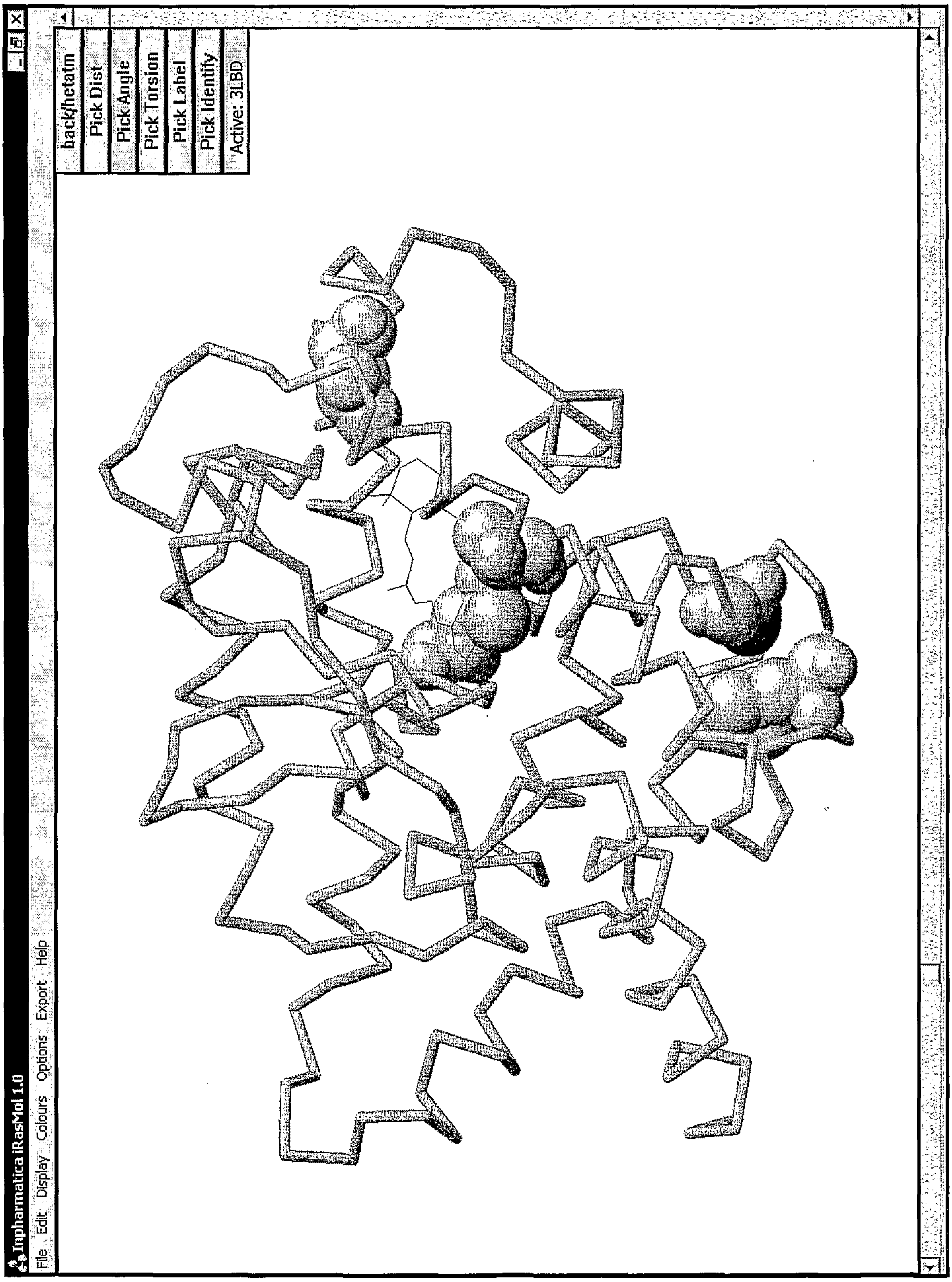

Inpharmatica iRasMol 1.0
File
Edit
Display
Colours
Options
Export
Help
back/hetatm
Pick Dist
Pick Angle
Pick Torsion
Pick Label
Pick Identify
Active: 3LBD

FIG. 8B

FIG. 9-1

AlEye output (February 21, 2001 10:52 AM)

```
                    10        20        30        40        50        60        70
                    |         |         |         |         |         |         |
BAA07892.2 GKPGSTMGDDSEWLKLPVDQKCEHKLWKARLSGYEEALKIFQKIKDEKSPEWSKFLGLIKKFVTDSNAVVQLKGLEAAL
CAB61894.1 ------MGDDSEWMKLPIDQKCEHKVWKARLNGYEEAVKLFQKIVDEKSPEWSKYLGLIKRFVTESNAVAQLKGLEAAL
CAB55772.1 ------MAEDTEYKKLPVEERCVHKLWKARVDGYEEAAKIFRELDDEKSPEWSKFAGLIKKMVVDSNALAQEKGLEAAL

           80        90        100       110       120       130       140       150
           |         |         |         |         |         |         |         |
BAA07892.2 VYVENAHVAGKTTGEVVSGVVSKVFNQPKAKAKELGIEICLMYIEIEKGEAVQEELLKGLDNKNPKIIVACIETLRKAL
CAB61894.1 VYVENAHVAGKTTGEVVNGVVNKVFNQPKARAKELGADICLMYVEIEKAEVVQEELLKGLDNKNPKIVVACVETVRKAL
CAB55772.1 IFVENSGLAGRTVGDVMTGIVQKCIAAPKTKTKELSVQVALMYVEIEKQEAVVEELVKGMEAKNPKIVSACVAATTLAL

            160       170       180       190       200       210       220       230
            |         |         |         |         |         |         |         |
BAA07892.2 SEFGSKIILLKPIIKVLPKLFESREKAVRDEAKLIAVEIYRWIRDALRPPLQNINSVQLKELEEEWVKLPTSAPRPTRF
CAB61894.1 SEFGSKIMTLKPIIKVLPKLFESREKAIRDEAKLLAVEIYRWIRDALRPPLQNINPVQLKELEEEWVKLPQSAPKQTRF
CAB55772.1 REFGHKVIGVKPLIKKLAPLMSDRDKTVRDEGKQLAVEIYRWIGAAMKAQISTLPQVTLKELEDEFDKLKGERVEPSRY

             240       250       260       270       280       290       300       310
             |         |         |         |         |         |         |         |
BAA07892.2 LRSQQELEAKLEQQQSAGGDAEGGGDDGDEVPQIDAYELLEAVEILSKLPKDFYDKIEAKKWQERKEALESVEVLI-KN
CAB61894.1 LRSQQDLKAKFEQQQAAGDDG-GDDGEEEIVPQVDAYELLEAVEILSKLPKDFYDKIEAKKWQERKEALEAVEALV-KN
CAB55772.1 LKSQQEKQAKI-ADAAATEDAYNEDDGEAGVEEIDPMDLLDPVDILSKMPKDFYDKLEEKKWTLRKESLEVLEKLLTDH

              320       330       340       350       360       370       380       390
              |         |         |         |         |         |         |         |
BAA07892.2 PKLEAGDYADLVKALKKVVGKDTNVMLVALAAKCLTGLAVGLRKKFGQYAGHVVPTILEKFKEKKPQVVQALQEAIDAI
CAB61894.1 PKIEAGDFADLVKALKTVVGKDTNVMLVALAAKCIAGLAAGLRKKFGSYAGHIVPTILEKFKEKKPQVVQALQEAIDAV
CAB55772.1 PKLENGEYGALVSALKKVITKDSNVVLVAMAGKCLALLAKGLAKRFSNYASACVPSLLEKFKEKKPNVVTALREAIDAI

               400       410       420       430       440       450       460       470
               |         |         |         |         |         |         |         |
BAA07892.2 FLTTTLQNISEDVLAVMDNKNPTIKQQTSLFIARSFRHCTASTLPKSLLKPFCAALLKHINDSAPEVRDAAFEALGTAL
CAB61894.1 FLTTTLQNISEDVLAVMDNKNPAIKQQTSLFLARSFRHCTPSTLPKSLLKPFCVALLKQINDSAPEVRDAAFEALGTAQ
CAB55772.1 YASTSLEAQQESIVESLANKNPSVKSETALFIARALTRTQPTALNKKLLKLLTTSLVKTLNEPDPTVRDSSAEALGTLI

                480       490       500       510       520       530       540       550
                |         |         |         |         |         |         |         |
BAA07892.2 KVVGEKAVKPFLADVDKLKLDKIKECSEKVELIHGKKAGLAADKKEFKPLPGRTAAS-----GAAGDKDTKDISAPKPG
CAB61894.1 KVVGEKAVNPFLAEVDKLKLDRIKECADKAELANGKKGGAAAGEKKETKAPAAAPGKPVPNQGAAAEKDAG----KAAA
CAB55772.1 KLMGDKAVTPLLADVDPLKMAKIKECQEKAE----IKIKVAGPKKETRPASAPTAKAAAPAKTVAGSVDPKPVTRPATT
```

FIG. 9-2

AlEye output (February 21, 2001 10:52 AM)

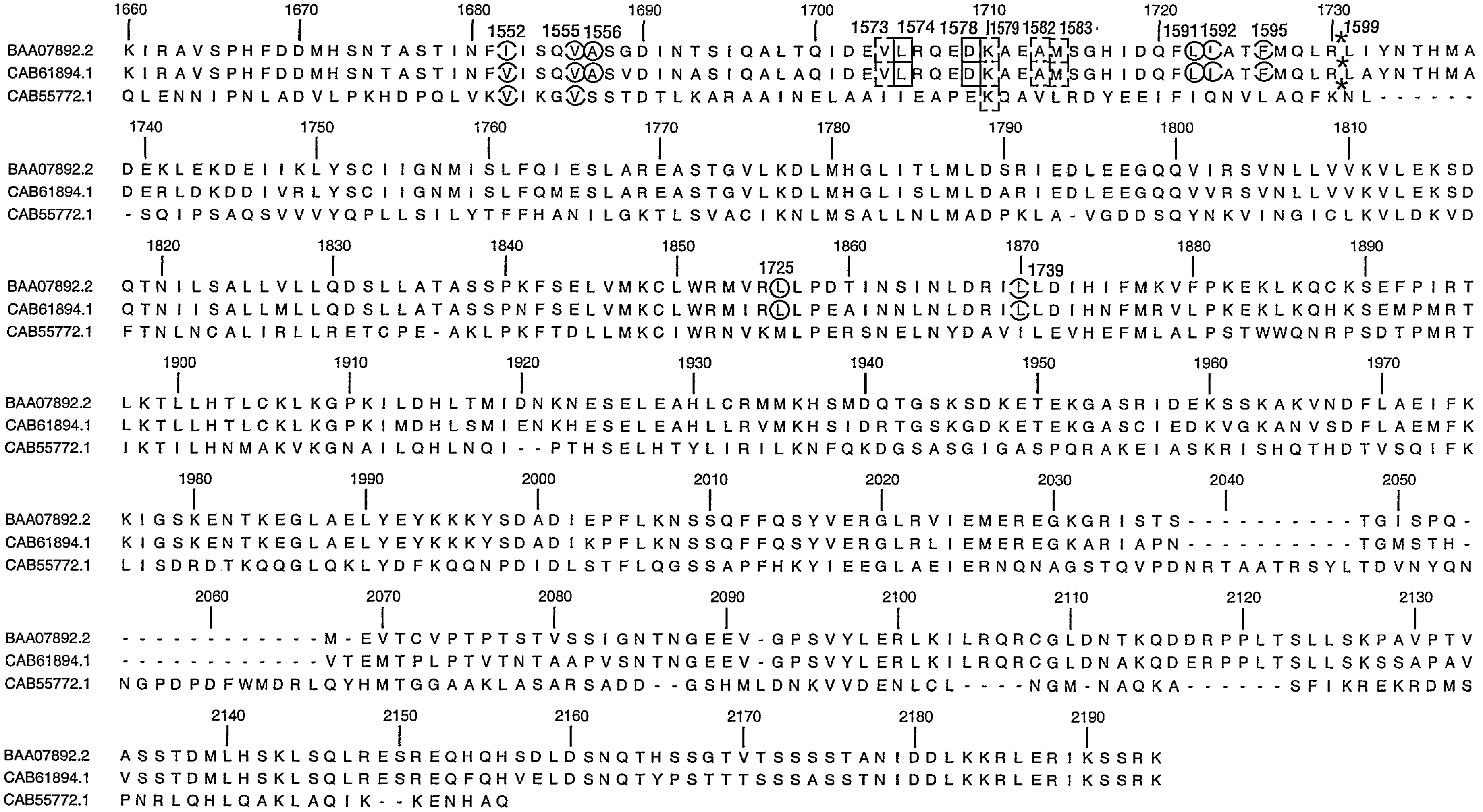
```
                       560                 570                 580                 590                 600                 610                 620                 630
                        |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  P L K K A P A A K A G G P P K K G K P A A P G G A G N T G T K N K K G L E T K E I V E P E L S I E V C E E K A S A V L P P T C I Q L L D S S N W K E R L A C M
CAB61894.1  A P K K A P A A K P G G P V K K A K A P A S S G A T - - - A K G K K A V E N K E I I E Q E L S P E A C E E R A A A V L P A S C M Q Q L D S S N W K E R L A S M
CAB55772.1  G A R K V - - - - - - - - - L K K P A T V S G G A T S A P T A A L K A G G K P L A T E R E I T P E E L Q E K S E E I L P A E I L N G L V D S N W K N R L A A V

                         640                 650                 660                 670                 680                 690                 700                 710
                          |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  E E F Q K A V E L M D R T E - M P C Q A L V R M L A - K K P G W K E T N F Q V M Q M K L H I V A L I A Q K G N F S K T S A Q V V L D G L V D K I G D V K C G N
CAB61894.1  E E F Q K T V E S M E R N D - I P C Q A L V K M L A - K K P G F K E T N F Q V M Q M K L H I V A L I A Q K G N F S K T S A C A V L D G L V D K V G D V K C G G
CAB55772.1  E Q L L G E I S G F D A K Q A G I S Q I L I R T I S G R K P G L K E M N F Q V L K F K L D I I R S V A E N Y P L T T T T V D L V I N E I I E K L A D A K N G A

                           720                 730                 740                 750                 760                 770                 780                 790
                            |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  N A K E A M T A I A E A C M L P W T A E Q V V S M A F S Q K N P K N Q S E T L N W L S N A I K E F G F S G L N V K A F I S N V K T A L A A T N P A V R T A A I
CAB61894.1  N A K E A L S G I A E A C T L P W T A E Q V V S L A F A Q K N P K N Q S E T L N W L S N A I K E F G F T G I N V K A F I S N V K T A L A A T N P A I R T S A I
CAB55772.1  A A A D V L S A F A E A T K L E Y V V G K V L S F A F E Q K S P K V Q S E A F N W V N R S I I E F G F - Q L Q P K T L I E D V R K G V Q S T N P T V R A S A I

                             800                 810                 820                 830                 840                 850                 860
                              |                   |                   |                   |                   |                   |                   |
BAA07892.2  T L L G V M Y L Y V G P S L R M F F E D E K P A L L S Q I D A E F E K M Q G Q S P P A P T R G I S K H S T S G T D E G E D G D E P D - D G S N - - - - - - D V
CAB61894.1  T L L G V M Y L Y M G A P L R M F F E E E K P A L L S Q I D A E F E K M K G Q T P P V S I R G - - S K H G S G R D E G E E G E E Q D E D A P A - - - - - - D V
CAB55772.1  Q M V G T M S M Y M G K A L M M F F D S E K P A L K S Q I Q V E F D K N V G E K P P K P V R G V Q - R S S G G T A G N S P D N E D D D G G A A G E E E P I N M

           870                 880                 890                 900                 910                 920                 930                 940
            |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V D L L P R T E I S D K I T S E L V S K I G D K N W K I R K E G L D E V A G I I N D A K F I Q P N I G E L P T A L K G R L N D S N K I L V Q Q T L N I L Q Q L
CAB61894.1  T D L L P R T D I S D K I S S D L V S K I E D K N W K I R K E G L D E V T A I I N E A K F I Q P S I G E L P S A L K G R L N D S N K I L V Q Q T L T I L Q Q L
CAB55772.1  A D L L P R V D I A P Q I T E A L L K E M S D K D W K T R N E G L T K L Q A I I S E A R L I K P S I G D L A P A L A H R L V D S N A K I A Q T T L A I C E Q L

             950                 960                 970                 980                 990                 1000                1010                1020
              |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  A V A M G P N I K Q H V K N L G I P I I T V L G D S K N N V R A A A L A T V N A W A E Q T G M K E W L E G E D L S E E L K K E N P F L R Q E L L G W L A E K L
CAB61894.1  S T A M G H N I K Q H V K N L G M P I I T V L G D S K A N V R A A A L G T L K S W V D Q T G M K D W L E G E D L S E E L K K E N P F L R Q E L L G W L A E K L
CAB55772.1  A T A M G A G C R N H V R N L F P G F L H A L G D N K S F V R A A A L N C I N S F G E K G G Y K E F F E S E M I A D A L K G G S P A L K T E L W A W L A D K L

               1030                1040                1050                1060                1070                1080                1090                1100
                |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  P T L - - - R S T P T D L I L C V P H L Y S C L E D R N G D V R K K A Q D A L P F F M M H L G Y E K M A K A T G K L K P T S K D Q V L A M L E K A K V N M P A
CAB61894.1  P S M - - - R T V P S D L Q L C V P Y L Y N C L E D R N G D V R K K A Q E A L P I F M M H I G F E K M S K A T S K L K P A S K D Q V V A L L E K A K A S M P A
CAB55772.1  P G L P P K S V S K E D I H S M V P H L Y A H I C D R N A D V R K N A N E A V L G I M I H L G F D A M N R A L D K Q K P A S K K D I L A A L E K A R P N L P V
```

FIG. 9-3

AlEye output (February 21, 2001 10:52 AM)

```
              1110      1120      1130      1140      1150      1160      1170      1180
               |         |         |         |         |         |         |         |
BAA07892.2  KPA-PPTKATSKPMGGSAPAKFQPASAPAEDCISSSTEPKPDPKKAKAPGLSSKAK-----------------------
CAB61894.1  KPAGPPGKASSKQPPAVAQASASPPPAASSDSGSSTSDYKPDPKKTKPGTQASKAKTQSVSSEGNTSLNPSNTSLTPSK
CAB55772.1  KP--------LPKGKHQAPIPEEPKLKTVRGGGAGG-----------APGIQKSATARVAGG-----------------

               1190      1200      1210      1220      1230      1240      1250      1260
                |         |         |         |         |         |         |         |
BAA07892.2  -------------SAQGKKMPSKTSLKEDEDKSGPIFIVVPNGKEQRMKDEKGLKVLKWNFTTPRDEYIEQLKTQMSS-
CAB61894.1  ANTSLSKAKPAKQTLPGKKAPSKPNAKDEEDKSGPIYIIVPNGKEQRVKDEKALKVLKWNFTTPRDEYIEQLKTQMSP-
CAB55772.1  ---------------QDKQVPAR---KKDEDIDTSPLLCANSAKNQRLLDEQKMKVLKWTFVTPREEFTELLRDQMMTA

                1270      1280      1290      1300      1310      1320      1330      1340
                 |         |         |         |         |         |         |         |
BAA07892.2  CVAKWLQDEMFHSDFQHHNKALAVMVDHLESEKEGVIGCLDLILKWLTLRFFDTNTSVLMKALEYLKLLFTLLSEEEYH
CAB61894.1  CIARWLQDELFHADFQRQIKGLAVMTEHLESEKEGVISCLDLVLKWFTLRFFDTNTSVLMKCLEYLKLLFIMLSQEEYH
CAB55772.1  NVNKALIANMFHDDFRYHLKVIEQLSEDLAGNSKALVCNLDLILKWLTLRFYDTNPSVLIKGLEYLVQVFQVLIDEEYI

                 1350      1360      1370      1380      1390      1400      1410      1420
                  |         |         |         |         |         |         |         |
BAA07892.2  LTENEASSFIPYLVVKVGEPKDVIRKDVRAILNRMCLVYPASKMFPFIMEGTKSKNSKQRAECLEELGCLVESYGMNVC
CAB61894.1  LTEMEGTSFLPYLMLKVGEPKDIVRKDVRAILTKMCQVYPASKMFNFVMEGTKSKNSKQRAECLEELGCLVESYGMNVC
CAB55772.1  LAENEGSSFVPHLLLK-ANPKDAVRNGVRRVLRQVILVFPFVKVFGYVMEGLKSKNARQRTECLDELTFLIESYGMNIC

                  1430      1440      1450      1460      1470      1480      1490      1500
                   |         |         |         |         |         |         |         |
BAA07892.2  QPTPGKALKEIAVHIGDRDNAVRNAALNTIVTVYNVHGDQVFKLIGNLSEKDMSMLEERIKRSAKR-PSAAPIKQVEEK
CAB61894.1  QPTPAKALKEIAIHIGDRDTTVRNAALNTIVTVYNVHGEQVFKLIGNLSEKDMSMLEERIKRAGKKQAAAAPAKQVEEK
CAB55772.1  ---PQSAMREIARQISDRDNSVRNAALNCIVQVFFLSGEKTYKMIGHLNEKDLSMLDERIKRAKKTKKPTPPPSVDVPA

                   1510      1520      1530      1540      1550      1560      1570      1580
                    |         |         |         |         |         |         |         |
BAA07892.2  PQRAQNISSNANML-----RKGPAE--------DMS------SKLNQARSMSGHPEAAQMVRREFQLDLDEIENDNGTV
CAB61894.1  PQRVQ--SANASIL-----RKAPPE--------DMS------SKLNQARNMGGHTEPSHSVPREFQLDLDEIENDNGTV
CAB55772.1  PQRHDSIEIEDAEVGNGCDELPPPDEDGTFDQAPSSQLLLQQQLQQLQQQAQQQKPSGPFGLDSQVISEIEKDWVRVD

                    1590      1600      1610      1620      1630      1640      1650
                     |         |         |         |         |         |         |
BAA07892.2  RCEMPELVQHKLDDIFEPV-------------------------------------------------------LIPEP
CAB61894.1  RCEMPALVQHKLDEIFEPV-------------------------------------------------------LIPEP
CAB55772.1  QMEQKPLLNVDISSLDEPIKVRPTRAGIHYPQEKFDRLISRQHYMQQTLTTSPSSTAGMTSGVSPYRSPMRLQHQQPQQ
```

AIEye output (February 21, 2001 10:52 AM)



```
             1660                1670                1680                1690                1700                1710                1720                1730
BAA07892.2   K I R A V S P H F D D M H S N T A S T I N F I I S Q V A S G D I N T S I Q A L T Q I D E V L R Q E D K A E A M S G H I D Q F L I A T E M Q L R L I Y N T H M A
CAB61894.1   K I R A V S P H F D D M H S N T A S T I N F V I S Q V A S V D I N A S I Q A L A Q I D E V L R Q E D K A E A M S G H I D Q F L L A T E M Q L R L A Y N T H M A
CAB55772.1   Q L E N N I P N L A D V L P K H D P Q L V K V I K G V S S T D T L K A R A A I N E L A A I I E A P E K Q A V L R D Y E E I F I Q N V L A Q F K N L - - - - - -

               1740                1750                1760                1770                1780                1790                1800                1810
BAA07892.2   D E K L E K D E I I K L Y S C I I G N M I S L F Q I E S L A R E A S T G V L K D L M H G L I T L M L D S R I E D L E E G Q Q V I R S V N L L V V K V L E K S D
CAB61894.1   D E R L D K D D I V R L Y S C I I G N M I S L F Q M E S L A R E A S T G V L K D L M H G L I S L M L D A R I E D L E E G Q Q V V R S V N L L V V K V L E K S D
CAB55772.1   - S Q I P S A Q S V V V Y Q P L L S I L Y T F F H A N I L G K T L S V A C I K N L M S A L L N L M A D P K L A - V G D D S Q Y N K V I N G I C L K V L D K V D

                 1820                1830                1840                1850                1860                1870                1880                1890
BAA07892.2   Q T N I L S A L L V L L Q D S L L A T A S S P K F S E L V M K C L W R M V R L L P D T I N S I N L D R I L L D I H I F M K V F P K E K L K Q C K S E F P I R T
CAB61894.1   Q T N I I S A L L M L L Q D S L L A T A S S P N F S E L V M K C L W R M I R L L P E A I N N L N L D R I L L D I H N F M R V L P K E K L K Q H K S E M P M R T
CAB55772.1   F T N L N C A L I R L L R E T C P E - A K L P K F T D L L M K C I W R N V K M L P E R S N E L N Y D A V I L E V H E F M L A L P S T W W Q N R P S D T P M R T

                   1900                1910                1920                1930                1940                1950                1960                1970
BAA07892.2   L K T L L H T L C K L K G P K I L D H L T M I D N K N E S E L E A H L C R M M K H S M D Q T G S K S D K E T E K G A S R I D E K S S K A K V N D F L A E I F K
CAB61894.1   L K T L L H T L C K L K G P K I M D H L S M I E N K H E S E L E A H L L R V M K H S I D R T G S K G D K E T E K G A S C I E D K V G K A N V S D F L A E M F K
CAB55772.1   I K T I L H N M A K V K G N A I L Q H L N Q I - - P T H S E L H T Y L I R I L K N F Q K D G S A S G I G A S P Q R A K E I A S K R I S H Q T H D T V S Q I F K

                     1980                1990                2000                2010                2020                2030                2040                2050
BAA07892.2   K I G S K E N T K E G L A E L Y E Y K K K Y S D A D I E P F L K N S S Q F F Q S Y V E R G L R V I E M E R E G K G R I S T S - - - - - - - - - - T G I S P Q -
CAB61894.1   K I G S K E N T K E G L A E L Y E Y K K K Y S D A D I K P F L K N S S Q F F Q S Y V E R G L R L I E M E R E G K A R I A P N - - - - - - - - - - T G M S T H -
CAB55772.1   L I S D R D T K Q Q G L Q K L Y D F K Q Q N P D I D L S T F L Q G S S A P F H K Y I E E G L A E I E R N Q N A G S T Q V P D N R T A A T R S Y L T D V N Y Q N

                       2060                2070                2080                2090                2100                2110                2120                2130
BAA07892.2   - - - - - - - - - - - - M - E V T C V P T P T S T V S S I G N T N G E E V - G P S V Y L E R L K I L R Q R C G L D N T K Q D D R P P L T S L L S K P A V P T V
CAB61894.1   - - - - - - - - - - - - V T E M T P L P T V T N T A A P V S N T N G E E V - G P S V Y L E R L K I L R Q R C G L D N A K Q D E R P P L T S L L S K S S A P A V
CAB55772.1   N G P D P D F W M D R L Q Y H M T G G A A K L A S A R S A D D - - G S H M L D N K V V D E N L C L - - - - N G M - N A Q K A - - - - - - S F I K R E K R D M S

                         2140                2150                2160                2170                2180                2190
BAA07892.2   A S S T D M L H S K L S Q L R E S R E Q H Q H S D L D S N Q T H S S G T V T S S S S T A N I D D L K K R L E R I K S S R K
CAB61894.1   V S S T D M L H S K L S Q L R E S R E Q F Q H V E L D S N Q T Y P S T T T S S S A S S T N I D D L K K R L E R I K S S R K
CAB55772.1   P N R L Q H L Q A K L A Q I K - - K E N H A Q
```

☐ Precisely conserved "LBD motif " residue in BAA07892.2
☐ Same residue conserved in homolog

⸢⸣ Tolerated change in "LBD motif" residue in BAA07892.2
⸢⸣ Same residue conserved in homolog

FIG. 9-4

○ Precisely conserved ligand binding residue in BAA07892.2
○ Same residue conserved in homolog

◌ Tolerated change in ligand binding residue in BAA07892.2
◌ Same residue conserved in homolog

FIG. 10-1

AlEye output (February 21, 2001 4:29 PM)

```
                              10                  20                  30                  40                  50                  60                  70
                              |                   |                   |                   |                   |                   |                   |
BAA07892.2  G K P G S T M G D D S E W L K L P V D Q K C E H K L W K A R L S G Y E E A L K I F Q K I K D E K S P E W S K F L G L I K K F V T D S N A V V Q L K G L E A A L
CAA63212.1  - - - - - - M G D D S E W L K L P V D Q K C E H K L W K A R L S G Y E E A L K I F Q K I K D E K S P E W S K F L G L I K K F V T D S N A V V Q L K G L E A A L
            80                  90                  100                 110                 120                 130                 140                 150
            |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V Y V E N A H V A G K T T G E V V S G V V S K V F N Q P K A K A K E L G I E I C L M Y I E I E K G E A V Q E E L L K G L D N K N P K I I V A C I E T L R K A L
CAA63212.1  V Y V E N A H V A G K T T G E V V S G V V S K V F N Q P K A K A K E L G I E I C L M Y I E I E K G E A V Q E E L L K G L D N K N P K I I V A C I E T L R K A L
              160                 170                 180                 190                 200                 210                 220                 230
              |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  S E F G S K I I L L K P I I K V L P K L F E S R E K A V R D E A K L I A V E I Y R W I R D A L R P P L Q N I N S V Q L K E L E E E W V K L P T S A P R P T R F
CAA63212.1  S E F G S K I I L L K P I I K V L P K L F E S R E K A V R D E A K L I A V E I Y R W I R D A L R P P L Q N I N S V Q L K E L E E E W V K L P T S A P R P T R F
              240                 250                 260                 270                 280                 290                 300                 310
              |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  L R S Q Q E L E A K L E Q Q Q S A G G D A E G G G D D G D E V P Q I D A Y E L L E A V E I L S K L P K D F Y D K I E A K K W Q E R K E A L E S V E V L I K N P
CAA63212.1  L R S Q Q E L E A K L E Q Q Q S A G G D A E G G G D D G D E V P Q I D A Y E L L E A V E I L S K L P K D F Y D K I E A K K W Q E R K E A L E S V E V L I K N P
                320                 330                 340                 350                 360                 370                 380                 390
                |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  K L E A G D Y A D L V K A L K K V V G K D T N V M L V A L A A K C L T G L A V G L R K K F G Q Y A G H V V P T I L E K F K E K K P Q V V Q A L Q E A I D A I F
CAA63212.1  K L E A G D Y A D L V K A L K K V V G K D T N V M L V A L A A K C L T G L A V G L R K K F G Q Y A G H V V P T I L E K F K E K K P Q V V Q A L Q E A I D A I F
                  400                 410                 420                 430                 440                 450                 460                 470
                  |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  L T T T L Q N I S E D V L A V M D N K N P T I K Q Q T S L F I A R S F R H C T A S T L P K S L L K P F C A A L L K H I N D S A P E V R D A A F E A L G T A L K
CAA63212.1  L T T T L Q N I S E D V L A V M D N K N P T I K Q Q T S L F I A R S F R H C T A S T L P K S L L K P F C A A L L K H I N D S A P E V R D A A F E A L G T A L K
                    480                 490                 500                 510                 520                 530                 540                 550
                    |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V V G E K A V K P F L A D V D K L K L D K I K E C S E K V E L I H G K K A G L A A D K K E F K P L P G R T A A S G A A G D K D T K D I S A P K P G P L K K A P
CAA63212.1  V V G E K A V K P F L A D V D K L K L D K I K E C S E K V E L I H G K K A G L A A D K K E F K P L P G R T A A S G A A G D K D T K D I S A P K P G P L K K A P
                      560                 570                 580                 590                 600                 610                 620                 630
                      |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  A A K A G G P P K K G K P A A P G G A G N T G T K N K K G L E T K E I V E P E L S I E V C E E K A S A V L P P T C I Q L L D S S N W K E R L A C M E E F Q K A
CAA63212.1  A A K A G G P P K K G K P A A P G G A G N T G T K N K K G L E T K E I V E P E L S I E V C E E K A S A V L P P T C I Q L L D S S N W K E R L A C M E E F Q K A
                        640                 650                 660                 670                 680                 690                 700                 710
                        |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V E L M D R T E M P C Q A L V R M L A K K P G W K E T N F Q V M Q M K L H I V A L I A Q K G N F S K T S A Q V V L D G L V D K I G D V K C G N N A K E A M T A
CAA63212.1  V E L M D R T E M P C Q A L V R M L A K K P G W K E T N F Q V M Q M K L H I V A L I A Q K G N F S K T S A Q V V L D G L V D K I G D V K C G N N A K E A M T A
```

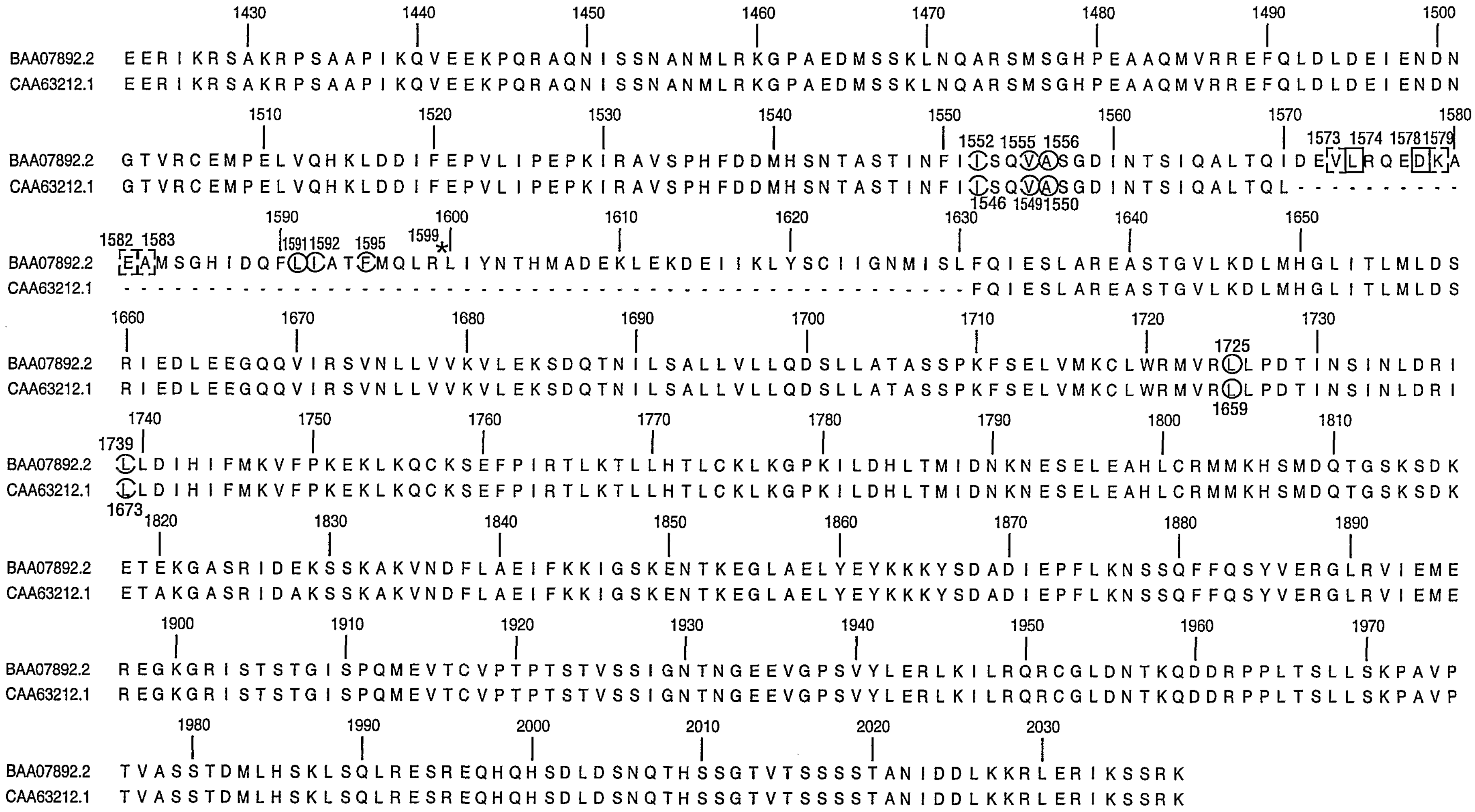
FIG. 10-2

AlEye output (February 21, 2001 4:29 PM)

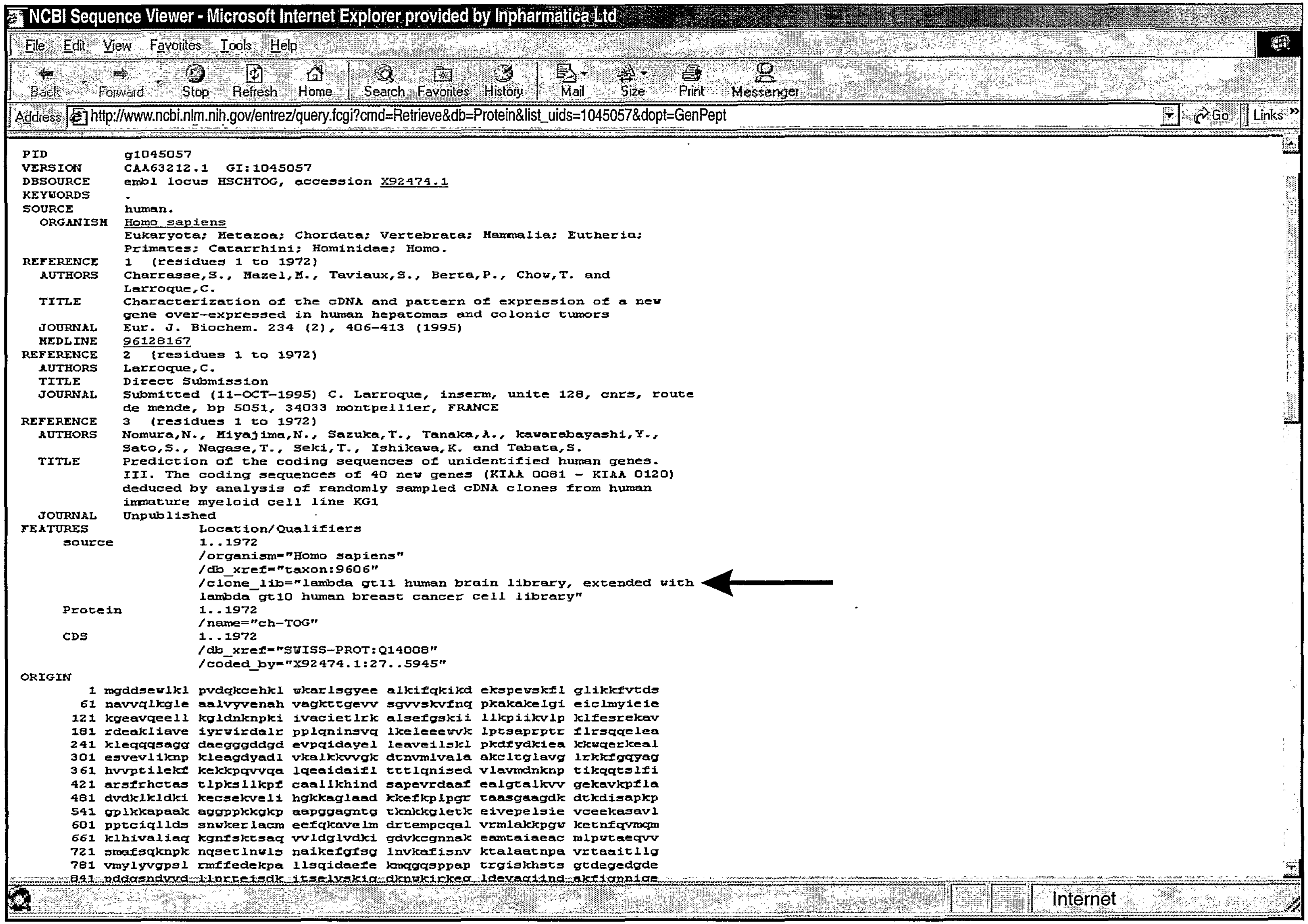
```
                           720                 730                 740                 750                 760                 770                 780                 790
                            |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  I A E A C M L P W T A E Q V V S M A F S Q K N P K N Q S E T L N W L S N A I K E F G F S G L N V K A F I S N V K T A L A A T N P A V R T A A I T L L G V M Y L
CAA63212.1  I A E A C M L P W T A E Q V V S M A F S Q K N P K N Q S E T L N W L S N A I K E F G F S G L N V K A F I S N V K T A L A A T N P A V R T A A I T L L G V M Y L
                             800                 810                 820                 830                 840                 850                 860
                              |                   |                   |                   |                   |                   |                   |
BAA07892.2  Y V G P S L R M F F E D E K P A L L S Q I D A E F E K M Q G Q S P P A P T R G I S K H S T S G T D E G E D G D E P D D G S N D V V D L L P R T E I S D K I T S
CAA63212.1  Y V G P S L R M F F E D E K P A L L S Q I D A E F E K M Q G Q S P P A P T R G I S K H S T S G T D E G E D G D E P D D G S N D V V D L L P R T E I S D K I T S
           870                 880                 890                 900                 910                 920                 930                 940
            |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  E L V S K I G D K N W K I R K E G L D E V A G I I N D A K F I Q P N I G E L P T A L K G R L N D S N K I L V Q Q T L N I L Q Q L A V A M G P N I K Q H V K N L
CAA63212.1  E L V S K I G D K N W K I R K E G L D E V A G I I N D A K F I Q P N I G E L P T A L K G R L N D S N K I L V Q Q T L N I L Q Q L A V A M G P N I K Q H V K N L
             950                 960                 970                 980                 990                1000                1010                1020
              |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  G I P I I T V L G D S K N N V R A A A L A T V N A W A E Q T G M K E W L E G E D L S E E L K K E N P F L R Q E L L G W L A E K L P T L R S T P T D L I L C V P
CAA63212.1  G I P I I T V L G D S K N N V R A A A L A T V N A W A E Q T G M K E W L E G E D L S E E L K K E N P F L R Q E L L G W L A E K L P T L R S T P T D L I L C V P
              1030                1040                1050                1060                1070                1080                1090                1100
                |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  H L Y S C L E D R N G D V R K K A Q D A L P F F M M H L G Y E K M A K A T G K L K P T S K D Q V L A M L E K A K V N M P A K P A P P T K A T S K P M G G S A P
CAA63212.1  H L Y S C L E D R N G D V R K K A Q D A L P F F M M H L G Y E K M A K A T G K L K P T S K D Q V L A M L E K A K V N M P A K P A P P T K A T S K P M G G S A P
                1110                1120                1130                1140                1150                1160                1170                1180
                  |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  A K F Q P A S A P A E D C I S S S T E P K P D P K K A K A P G L S S K A K S A Q G K K M P S K T S L K E D E D K S G P I F I V V P N G K E Q R M K D E K G L K
CAA63212.1  A K F Q P A S A P A E D C I S S S T E P K P D P K K A K A P G L S S K A K S A Q G K K M P S K T S L K E D E D K S G P I F I V V P N G K E Q R M K D E K G L K
                  1190                1200                1210                1220                1230                1240                1250                1260
                    |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  V L K W N F T T P R D E Y I E Q L K T Q M S S C V A K W L Q D E M F H S D F Q H H N K A L A V M V D H L E S E K E G V I G C L D L I L K W L T L R F F D T N T
CAA63212.1  V L K W N F T T P R D E Y I E Q L K T Q M S S C V A K W L Q D E M F H S D F Q H H N K A L A V M V D H L E S E K E G V I G C L D L I L K W L T L R F F D T N T
                    1270                1280                1290                1300                1310                1320                1330                1340
                      |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  S V L M K A L E Y L K L L F T L L S E E E Y H L T E N E A S S F I P Y L V V K V G E P K D V I R K D V R A I L N R M C L V Y P A S K M F P F I M E G T K S K N
CAA63212.1  S V L M K A L E Y L K L L F T L L S E E E Y H L T E N E A S S F I P Y L V V K V G E P K D V I R K D V R A I L N R M C L V Y P A S K M F P F I M E G T K S K N
                      1350                1360                1370                1380                1390                1400                1410                1420
                        |                   |                   |                   |                   |                   |                   |                   |
BAA07892.2  S K Q R A E C L E E L G C L V E S Y G M N V C Q P T P G K A L K E I A V H I G D R D N A V R N A A L N T I V T V Y N V H G D Q V F K L I G N L S E K D M S M L
CAA63212.1  S K Q R A E C L E E L G C L V E S Y G M N V C Q P T P G K A L K E I A V H I G D R D N A V R N A A L N T I V T V Y N V H G D Q V F K L I G N L S E K D M S M L
```

FIG. 10-3

AlEye output (February 21, 2001 4:29 PM)

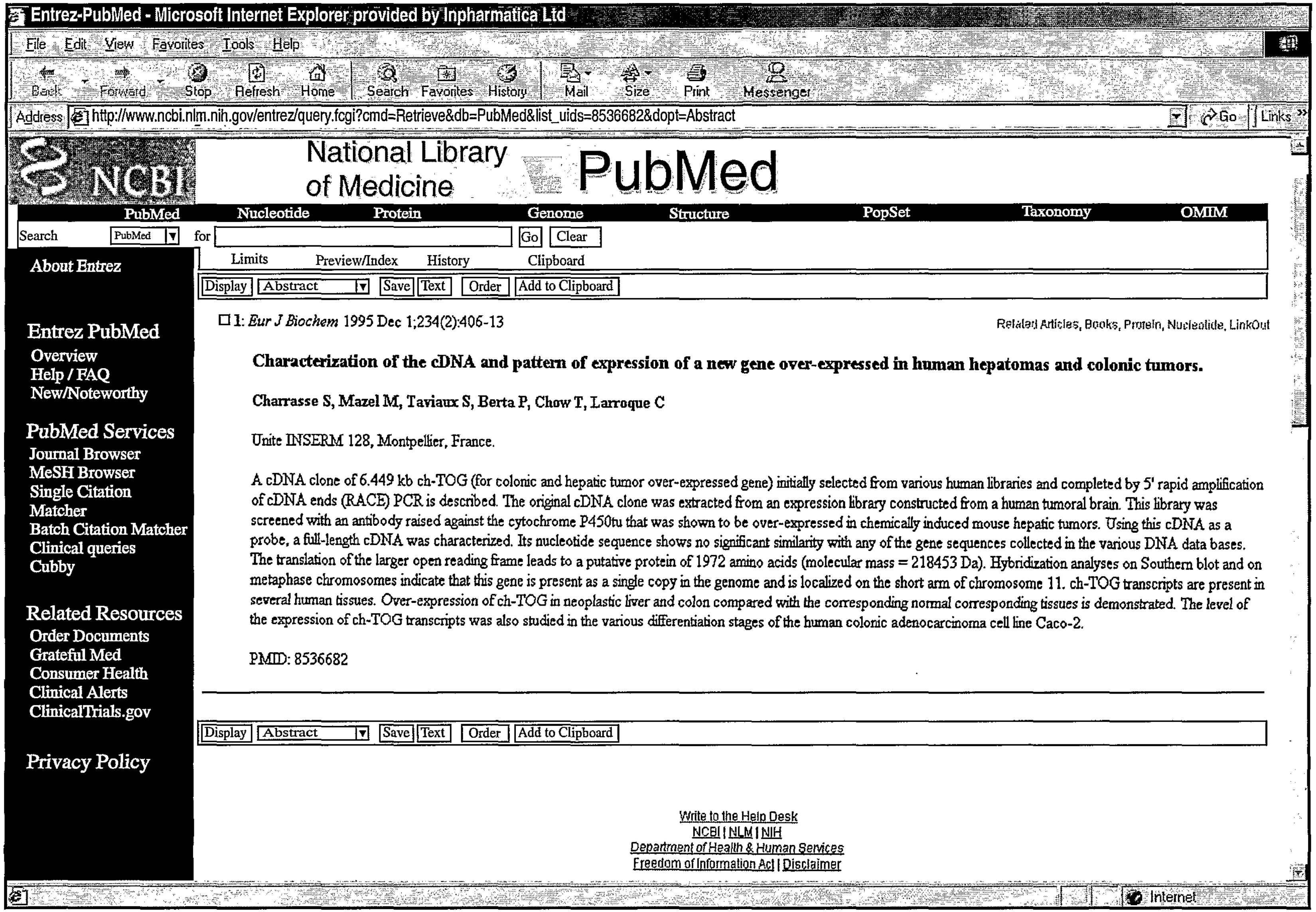


```
                1430      1440      1450      1460      1470      1480      1490      1500
BAA07892.2 E E R I K R S A K R P S A A P I K Q V E E K P Q R A Q N I S S N A N M L R K G P A E D M S S K L N Q A R S M S G H P E A A Q M V R R E F Q L D L D E I E N D N
CAA63212.1 E E R I K R S A K R P S A A P I K Q V E E K P Q R A Q N I S S N A N M L R K G P A E D M S S K L N Q A R S M S G H P E A A Q M V R R E F Q L D L D E I E N D N

                1510      1520      1530      1540      1550 1552 1555 1556 1560      1570 1573 1574 1578 1579 1580
BAA07892.2 G T V R C E M P E L V Q H K L D D I F E P V L I P E P K I R A V S P H F D D M H S N T A S T I N F I L S Q V A S G D I N T S I Q A L T Q I D E V L R Q E D K A
CAA63212.1 G T V R C E M P E L V Q H K L D D I F E P V L I P E P K I R A V S P H F D D M H S N T A S T I N F I L S Q V A S G D I N T S I Q A L T Q L - - - - - - - - - -
                                                                     1546 1549 1550

           1582 1583  1590 1591 1592 1595 1599 * 1600   1610      1620      1630      1640      1650
BAA07892.2 E A M S G H I D Q F L I A T E M Q L R L I Y N T H M A D E K L E K D E I I K L Y S C I I G N M I S L F Q I E S L A R E A S T G V L K D L M H G L I T L M L D S
CAA63212.1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - F Q I E S L A R E A S T G V L K D L M H G L I T L M L D S

           1660      1670      1680      1690      1700      1710      1720  1725    1730
BAA07892.2 R I E D L E E G Q Q V I R S V N L L V V K V L E K S D Q T N I L S A L L V L L Q D S L L A T A S S P K F S E L V M K C L W R M V R L L P D T I N S I N L D R I
CAA63212.1 R I E D L E E G Q Q V I R S V N L L V V K V L E K S D Q T N I L S A L L V L L Q D S L L A T A S S P K F S E L V M K C L W R M V R L L P D T I N S I N L D R I
                                                                                          1659

           1739 1740      1750      1760      1770      1780      1790      1800      1810
BAA07892.2 L L D I H I F M K V F P K E K L K Q C K S E F P I R T L K T L L H T L C K L K G P K I L D H L T M I D N K N E S E L E A H L C R M M K H S M D Q T G S K S D K
CAA63212.1 L L D I H I F M K V F P K E K L K Q C K S E F P I R T L K T L L H T L C K L K G P K I L D H L T M I D N K N E S E L E A H L C R M M K H S M D Q T G S K S D K
           1673

             1820      1830      1840      1850      1860      1870      1880      1890
BAA07892.2 E T E K G A S R I D E K S S K A K V N D F L A E I F K K I G S K E N T K E G L A E L Y E Y K K K Y S D A D I E P F L K N S S Q F F Q S Y V E R G L R V I E M E
CAA63212.1 E T A K G A S R I D A K S S K A K V N D F L A E I F K K I G S K E N T K E G L A E L Y E Y K K K Y S D A D I E P F L K N S S Q F F Q S Y V E R G L R V I E M E

             1900      1910      1920      1930      1940      1950      1960      1970
BAA07892.2 R E G K G R I S T S T G I S P Q M E V T C V P T P T S T V S S I G N T N G E E V G P S V Y L E R L K I L R Q R C G L D N T K Q D D R P P L T S L L S K P A V P
CAA63212.1 R E G K G R I S T S T G I S P Q M E V T C V P T P T S T V S S I G N T N G E E V G P S V Y L E R L K I L R Q R C G L D N T K Q D D R P P L T S L L S K P A V P

              1980      1990      2000      2010      2020      2030
BAA07892.2 T V A S S T D M L H S K L S Q L R E S R E Q H Q H S D L D S N Q T H S S G T V T S S S S T A N I D D L K K R L E R I K S S R K
CAA63212.1 T V A S S T D M L H S K L S Q L R E S R E Q H Q H S D L D S N Q T H S S G T V T S S S S T A N I D D L K K R L E R I K S S R K
```



```
PID         g1045057
VERSION     CAA63212.1  GI:1045057
DBSOURCE    embl locus HSCHTOG, accession X92474.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria;
            Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 1972)
  AUTHORS   Charrasse,S., Mazel,M., Taviaux,S., Berta,P., Chow,T. and
            Larroque,C.
  TITLE     Characterization of the cDNA and pattern of expression of a new
            gene over-expressed in human hepatomas and colonic tumors
  JOURNAL   Eur. J. Biochem. 234 (2), 406-413 (1995)
  MEDLINE   96128167
REFERENCE   2  (residues 1 to 1972)
  AUTHORS   Larroque,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-OCT-1995) C. Larroque, inserm, unite 128, cnrs, route
            de mende, bp 5051, 34033 montpellier, FRANCE
REFERENCE   3  (residues 1 to 1972)
  AUTHORS   Nomura,N., Miyajima,N., Sazuka,T., Tanaka,A., kawarabayashi,Y.,
            Sato,S., Nagase,T., Seki,T., Ishikawa,K. and Tabata,S.
  TITLE     Prediction of the coding sequences of unidentified human genes.
            III. The coding sequences of 40 new genes (KIAA 0081 - KIAA 0120)
            deduced by analysis of randomly sampled cDNA clones from human
            immature myeloid cell line KG1
  JOURNAL   Unpublished
FEATURES             Location/Qualifiers
     source          1..1972
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /clone_lib="lambda gt11 human brain library, extended with
                     lambda gt10 human breast cancer cell library"
     Protein         1..1972
                     /name="ch-TOG"
     CDS             1..1972
                     /db_xref="SWISS-PROT:Q14008"
                     /coded_by="X92474.1:27..5945"
ORIGIN
        1 mgddsewlkl pvdqkcehkl wkarlsgyee alkifqkikd ekspewskfl glikkfvtds
       61 navvqlkgle aalvyvenah vagkttgevv sgvvskvfnq pkakakelgi eiclmyieie
      121 kgeavqeell kgldnknpki ivacietlrk alsefgskii llkpiikvlp klfesrekav
      181 rdeakliave iyrwirdalr pplqninsvq lkeleeewvk lptsaprptr flrsqqelea
      241 kleqqqsagg daegggddgd evpqidayel leaveilskl pkdfydkiea kkwqerkeal
      301 esvevlikmp kleagdyadl vkalkkvvgk dtnvmlvala akcltglavg lrkkfgqyag
      361 hvvptilekf kekkpqvvqa lqeaidaifl tttlqnised vlavmdnknp tikqqtslfi
      421 arsfrhctas tlpksllkpf caallkhind sapevrdaaf ealgtalkvv gekavkpfla
      481 dvdklkldki kecsekveli hgkkaglaad kkefkplpgr taasgaagdk dtkdisapkp
      541 gplkkapaak aggppkkgkp aapggagntg tknkkgletk eivepelsie vceekasavl
      601 pptciqllds snwkerlacm eefqkavelm drtempcqal vrmlakkpgw ketnfqvmqm
      661 klhivaliaq kgnfsktsaq vvldglvdki gdvkcgnnak eamtaiaeac mlpwtaeqvv
      721 smafsqknpk nqsetlnwls naikefgfsg lnvkafisnv ktalaatnpa vrtaaitllg
      781 vmylyvgpsl rmffedekpa llsqidaefe kmqgqsppap trgiskhsts gtdegedgde
      841 pddgsndvvd llprteisdk itselvskig dknwkirkeg ldevagiind akfiqnnige
```

FIG. 11



☐ 1: *Eur J Biochem* 1995 Dec 1;234(2):406-13

Characterization of the cDNA and pattern of expression of a new gene over-expressed in human hepatomas and colonic tumors.

Charrasse S, Mazel M, Taviaux S, Berta P, Chow T, Larroque C

Unite INSERM 128, Montpellier, France.

A cDNA clone of 6.449 kb ch-TOG (for colonic and hepatic tumor over-expressed gene) initially selected from various human libraries and completed by 5' rapid amplification of cDNA ends (RACE) PCR is described. The original cDNA clone was extracted from an expression library constructed from a human tumoral brain. This library was screened with an antibody raised against the cytochrome P450tu that was shown to be over-expressed in chemically induced mouse hepatic tumors. Using this cDNA as a probe, a full-length cDNA was characterized. Its nucleotide sequence shows no significant similarity with any of the gene sequences collected in the various DNA data bases. The translation of the larger open reading frame leads to a putative protein of 1972 amino acids (molecular mass = 218453 Da). Hybridization analyses on Southern blot and on metaphase chromosomes indicate that this gene is present as a single copy in the genome and is localized on the short arm of chromosome 11. ch-TOG transcripts are present in several human tissues. Over-expression of ch-TOG in neoplastic liver and colon compared with the corresponding normal corresponding tissues is demonstrated. The level of the expression of ch-TOG transcripts was also studied in the various differentiation stages of the human colonic adenocarcinoma cell line Caco-2.

PMID: 8536682

FIG. 12