## A distal dimerization domain is essential for DNA-binding by the atypical HNF1 homeodomain

Tanguy Chouard, Marta Blumenfeld, Ingolf Bach, Joël Vandekerckhove<sup>1</sup>, Silvia Cereghini and Moshe Yaniv<sup>\*</sup>

Unité des virus oncogènes, UA 1149 du CNRS, Département de Biologie Moléculaire, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris cedex 15, France and <sup>1</sup>Laboratorium voor Genetica, Rijksuniversiteit Gent, B 9000 Gent, Belgium

Received March 8, 1990; Revised and Accepted August 8, 1990

EMBL accession no. X54423

### ABSTRACT

Hepatic Nuclear Factor 1 (HNF1, also referred to as LFB1, HP1 or APF) is a liver-specific transcription factor required for the expression of many hepatocyte specific genes. We report here the purification of this rat liver nuclear protein and the cloning of its cDNA using a PCR-derived approach. Seven independent clones reveal 3 alternative polyadenylation sites and a unique open reading frame. Both a motif homologous to the homeodomain and a distal dimerization domain are required for specific DNA binding. Sequence comparisons reveal several atypical features at key positions in the segment corresponding to helices III and IV of the Antaennapedia homeodomain as well as a potential 24 amino acid loop in place of the universal turn between helices II and III. Together with its property to dimerize in the presence or absence of DNA, these features place HNF1 as the prototype of a novel subclass of transcription factors distantly related to homeoproteins.

### INTRODUCTION

Transcriptional analysis of the albumin promoter has led to the identification of the cis-elements (1, 2) and the corresponding transacting factors (3-5) that are required for efficient and hepatocyte-specific transcription. Mutations in the proximal element immediately upstream of the TATA motif cause the most drastic and most hepatocyte-specific effect on transcription (6, 7). The factor interacting with this site has been characterized in several laboratories (8-11) as binding to proximal control regions in the promoters of many liver-specific genes as well as to several of their distal enhancer elements (reviewed in 12, 13). The consensus sequence derived from these binding sites, g/aGTTAATNATTAACc/a, is palindromic, although it is never fully symmetrical in its various natural occurences. The HNF1 target site is able to drive tissue-specific transcription of heterologous, ubiquitous promoters when cloned in both orientations or even as single or multimerized copies upstream of a TATA motif (7, 14-16). In the Xenopus laevis albumin promoter, the HNF1 target site seems to be the only functional element upstream of the TATA motif (10).

In vitro studies confirm the crucial role of HNF1 in hepatocytespecific transcription. Titration of HNF1 protein in rat liver nuclear extracts using an excess of its specific DNA target abolishes the transcription driven by liver-specific promoters containing the HNF1 site (10, 11, 14). Furthermore, highly purified rat HNF1 protein is sufficient to complement a spleen nuclear extract for the *in vitro* transcription of the mouse albumin promoter (17). We report, here, the isolation of cDNA clones coding for HNF1 and the preliminary mapping of its DNAbinding domain which reveals several atypical features.

### MATERIALS AND METHODS

### Purification and microsequencing of HNF1

The double stranded oligonucleotide PE56a (generated by annealing the following oligonucleotides : 5' TCGAGTGTGG-TTAATGATCTACAGTTA-3' and 5'-TCGATAACTGTA-GATCATTAACCACAC-3'), encompassing the HNF1 binding site of the rat albumin promoter, was phosphorylated and stepwise ligated to a Sepharose-bound oligonucleotide bearing a XhoI cohesive end (18), to generate a specific DNA-affinity chromatography medium carrying  $50-75 \ \mu g$  DNA/ml gel. HNF1 purification was monitored by a band-shift assay using 5'-end-labeled PE56a. Liver nuclear extracts were prepared from 150 rats as described (11), except that the second sucrose cushion was omitted; proteins were resuspended in Elution Buffer (EB: 20% glycerol, 20 mM hepes pH 7.9, O.5 mM EDTA, O.5 mM DTT, O.5 mM PMSF, 12 mM MgCl<sub>2</sub>, 1µg/ml pepstatin, leupeptin and aprotinin) to a conductivity lower than that of 0.15 mM KCl in EB and loaded on five 200 ml columns of heparin Ultrogel (IBF). HNF1 containing fractions were eluted with 0.35 M KCl, diluted with EB to 0.3 M KCl supplemented with polydIdC (15 µg/mg protein) and O.1% NP40 and loaded on PE56a-DNA-affinity columns (20 ml total volume). Columns were washed with 0.3 M KCl and HNF1 activity was eluted at 0.6 M KCl, with a yield of 20% and a purification factor of 6600 fold, relative to nuclear extracts. 30 ml of pooled fractions

<sup>\*</sup> To whom correspondence should be addressed

containing  $25-30 \ \mu g$  of  $87-93 \ kDa$  HNF1 were concentrated by lyophilisation to 10 ml and dialysed against 10 mM ammonium acetate containing 0.01% SDS before further concentration to 300  $\mu$ l; the protein was electrophoresed on an SDS-10% polyacrylamide gel (19) and electroblotted onto a polyvinylidene

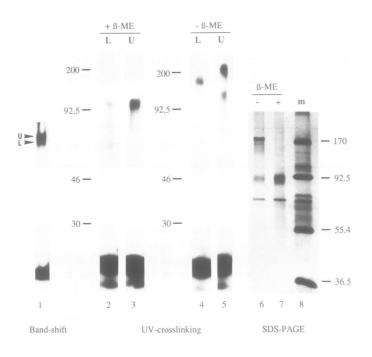


Figure 1: Analysis of affinity-purified HNF1. Dimers stabilized by S-S bonds. 10 ng of labelled PE56a oligonucleotide, in which several thymine residues had been replaced by bromodeoxyuridine residues, were incubated with 10  $\mu$ l of affinity-purified fraction and complexes were resolved by preparative 5% native PAGE (lane 1); proteins were UV-crosslinked to DNA and two bands (U and L for upper and lower bands respectively) separately cut and analyzed by 10% SDS-PAGE in reducing (lanes 2, 3; 0.1%  $\beta$ -mercaptoethanol) or nonreducing conditions (lanes 4, 5). The estimated molecular weights of UV-crosslinked peptides should be reduced by 10–15 kDa, due to the contribution of the PE-56a oligonucleotide (11). The fraction corresponding to the peak of DNA-binding activity was analyzed by silver-stained 8% SDS-PAGE in reducing (lane 7) or nonreducing conditions (lane 6). m : markers for molecular weights (given in kDa).

difluoride membrane (PVDF); the amidoblack stained 87-93 kd HNF1 band was cut out, digested *in situ* with trypsin and the eluted tryptic peptides were fractionated by Reversed Phase HPLC Chromatography followed by gas-phase amino acid sequence analysis as previously described (20 and references therein).

# Production of a nondegenerate DNA probe using polymerase chain reaction (PCR)

Enzymatic DNA amplifications with Taq DNA polymerase were performed in the buffers indicated by Saïki et al. (21) using 10 ng of oligo-dT primed cDNA from Fao cells (see below) and 1  $\mu$ g each of oligonucleotides C1 or C3 and D2 or D4 (see text and Figure 2). After 50 cycles of primary amplification (one cycle = 15" at 94°C, 1' at 60°C, 1' at 72°C) a 10% aliquot was futher amplified for 50 cycles and amplification products were analyzed on a 20% polyacrylamide gel. DNA of samples displaying only the expected amplified band (47 bp) in addition to the singlestranded primers was directly extracted with phenol-chloroform, phosphorylated, blunt-end ligated to a Bluescribe vector linearized with SmaI and sequenced. A 41-mer oligonucleotidic probe (CD41, see Figure 2) was derived from the sequence data.

### Preparation and screening of cDNA libraries

An amplified  $\lambda$ -gt10 cDNA library (oligo-dT primed) from rat hepatoma Fao cells treated with cycloheximide was prepared according to standard procedures (22) and was screened with 5'-end-labeled CD41 and A15 probes (Figure 2). Hybridizations were performed in 6×saline sodium citrate (SSC), 1×Denhardt's solution, 50 µg/ml tRNA, 0.05% sodium pyrophosphate at 42°C; washes were done at 55°C in 1×SSC. Inserts of positive clones were subcloned in a Bluescribe vector and sequenced on both strands by the dideoxynucleotide method modified for doublestranded DNA (23, 24) using successive primers.

### Construction of HNF1 expression vectors

Full length HNF1 cDNA was obtained as follows: the A4 clone was digested with SacI, treated with T4 DNA-polymerase and then digested with MluI to generate an MluI-SacI HNF1 fragment with a blunt SacI end (nt 200-2240, Figure 4); the 5' part of HNF1 cDNA was synthesized by reverse transcription of rat liver

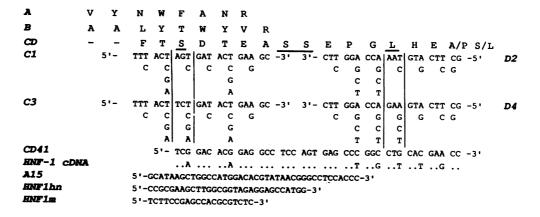


Figure 2: HNF1 tryptic peptides and PCR derived probe. Sequences of tryptic peptides A, B and CD are given in single letter code, the amino acids encoded by six codons in CD are underlined. PCR primers used to determine the CD41 sequence are aligned underneath the CD peptide sequence, two of the coding strand (C1 and C3) on the left and two of the noncoding strand on the right. The nucleotides in the coding strand of the HNF1 cDNA clones which matched CD41 are replaced by dots, those which did not are indicated by their letter code. Oligonucleotide A15 probe was derived from the 5' end of the CD13 clone coding strand; HNF1hn PCR primer, encompassing the initiation ATG, from the coding strand of the BP14 genomic clone with creation of an artificatial HindIII site; HNF1m, encompassing the unique MluI site, from the noncoding strand of the A4 clone.

total RNA (10  $\mu$ g) using a primer complementary to nt 230-250 (Figure 4); the purified cDNA then served as template to generate a PCR fragment using HNF1hn and HNF1m oligonucleotides (Figure 2) and the 241 bp fragment thus obtained was digested at the HindIII and MluI sites included in the PCR primers. Both 3' and 5' cDNA fragments were ligated to an HindIII/HpaI digested pRSV-CAT vector (25), thus generating the pRSV-HNF1 construct. pRSV-HNF1 was digested partially with NcoI and then with BamHI to obtain a 2369 bp fragment encompassing the segment 1-2240 of the HNF1 sequence (Figure 4) followed by the 129 bp HpaI-BamHI segment from pRSV-CAT. This fragment was cloned between the NcoI and BamHI sites of a (T7 promoter-ß globin leader)-vector, derived from pGEM1 (provided by Dr. R. Treisman), thus putting the whole HNF1 coding sequence in frame with the  $\beta$  globin initiator ATG (pT7 $\beta$ H plasmid).

HNF1 deletion vectors were constructed as follows: the 774 bp SmaI fragment from  $pT7\beta H$  was cloned in frame with the

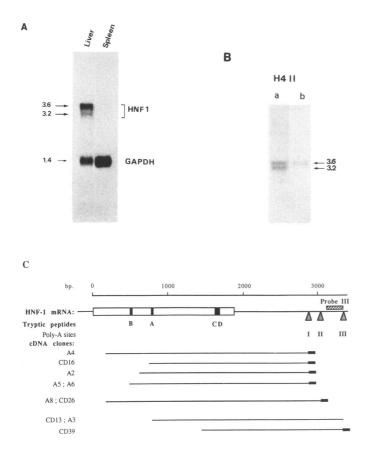


Figure 3A: Northern blot analysis of HNF1 transcripts. 3 µg of rat liver or spleen poly A+RNA were fractionated and hybridized with a mixture of full length HNF1 (clone CD26, Figure 3C) and GAPDH probes. Arrows to the left indicate the estimated size of the different bands. B: Northern blot analysis of alternative polyadenylation sites usage. In lane b, 12.5 µg of total RNA from H4II cells were hybridized with an HNF1 antisense RNA probe specific for transcripts of group III (see text; Probe III: nt 3158-3397, Figure 3C); after the final high stringency wash, the blot was rehybridized with a full length HNF1 cDNA probe (lane a). C: Schematic map of HNF1 cDNA clones presenting three polyadenylation sites. The extent of each of the seven independent HNF1 cDNA clones is indicated relative to the combined sequence of the HNF1 mRNA in front of the list of the corresponding redundant clones. The HNF1 open reading frame is indicated by a box; tryptic peptides are shown in black. PolyA tails are indicated by a black box at the end of each clone and the three polyadenylation sites included in the longest sequence are indicated by triangles. The group III specific probe (Probe III) is indicated by a hatched box.

first ATG in an NcoI digested pT7 $\beta$ H vector repaired with the Klenow enzyme, thus generating pT7 $\beta$ H-34/291; pT7 $\beta$ H- $\Delta$ 18/53 was obtained by XhoI digestion of pT7 $\beta$ H followed by repair with the Klenow enzyme and auto-ligation; pT7 $\beta$ H- $\Delta$ 34/208 was constructed by digestion of pT7 $\beta$ H with ApaI, further treatment with T4 DNA polymerase and auto-ligation.

### In vitro transcription and translation

Sense RNA was synthesized *in vitro* using 10 units of T7-RNA polymerase (Stratagene transcription kit) and 1  $\mu$ g of DNA templates linearized as follows: the T7 $\beta$ H-wt, T7 $\beta$ H-1/390, T7 $\beta$ H-1/281 and T7 $\beta$ H-1/264 transcripts were generated after digestion of pT7 $\beta$ H with BamHI, AatII, BalI and AccI respectively. The T7 $\beta$ H-34/291, T7 $\beta$ H- $\Delta$ 18/53 and T7 $\beta$ H- $\Delta$ 34/208 transcripts were obtained after digestion with BamHI of the corresponding deletion plasmids.

After 1/2 hour incubation at 37°C, phenol/chloroform extraction and ethanol precipitation, the synthesized RNA was resuspended in 20  $\mu$ l of water and 2  $\mu$ l was used for the *in vitro* translation with a rabbit reticulocyte lysate system (Amersham). The reaction mixture (15  $\mu$ l) included 12  $\mu$ l of the cell lysate and 1  $\mu$ l of diluted [<sup>35</sup>S]L-methionine (<1000 Ci/mmol, 1mmol/ml). The translation was carried out for 1 hour at 30°C, and quantity and quality of the products were checked by counting TCA-insoluble <sup>35</sup>S and by SDS-PAGE.

# Gel retardation assays, SDS-PAGE and UV-crosslinking experiments

Gel retardation assays and UV-crosslinking experiments with liver proteins were performed as previously described (11), without any competitor DNA in the case of affinity-purified fractions.

For each in vitro translated protein, the volume indicated in the legend for Figure 5A was preincubated for 10 min on ice with sonicated salmon sperm DNA at a ratio of 0.5  $\mu g/\mu l$ translation mixture and with competitor oligonucleotide where mentioned, in 20  $\mu$ l of binding buffer containing 10 mM Hepes, 4mM MgCl<sub>2</sub>, 0.1 mM EDTA, 4 mM spermidine, 15% glycerol. After addition of 1 ng of <sup>32</sup>P-labelled PE56 oligonucleotide probe, mixtures were further incubated for 10 min on ice and then analyzed by 5% PAGE in  $0.25 \times$  TBE. The gel was dried and exposed to two sheets of X-ray film with intensifying screen at  $-80^{\circ}$ C. The film closest to the gel shows both <sup>35</sup>S and <sup>32</sup>P signals and the second one only <sup>32</sup>P intensified signals. Figure 5A shows only the pattern of the second film. SDS-PAGE were done according to Laemmli (19) and, in the case of in vitro translated protein analysis, treated for fluorography (22) before exposure.

#### Northern blot analysis

Total RNA from rat liver, spleen or H4II cells was extracted using the guanidium thiocyanate-CsCl method (26). Poly A+ RNA was isolated by oligo dT cellulose chromatography (22), electrophoresed through 1.2% agarose-2.2 M formaldehyde gels and transferred to a Hybond N membrane (Amersham). RNA markers (0.24–9.5 kb; BRL) were run in parallel and visualized under UV. cDNA inserts were isolated from agarose gels, labelled by random priming to a specific activity of  $3-7\times10^8$ cpm/µg for HNF1 and  $1\times10^7$ cpm/µg for GAPDH and used as probes. Hybridization was performed at 42°C in 50% formamide,  $5\times$ SSPE,  $5\times$ Denhardt with the <sup>32</sup>P-labelled probe for 24 hours. The membranes were washed at 60°C with  $5\times$ SSC, 0.25% SDS

M V S K L S Q										w w c
ATGGTTTCTAAGTTGAGCCA	GCTGCAGACGGAGC	TCCTGGCTGC	TCTGCTCGAG	TCGCGCCTG	GCAAAGAGGC	TCTGATCCA	GCTCTGGGG	GAGCCCGGGC	CCTACCTGA	IGGTTGGA
10 20	30	40	50	60	70	80	90	100	110	120
D G P L D K G Gatggtcccctggacaaggg	ESCG	G T R G	D L T	E L P	N G L G	ETR	G S E	D D T		G E D
130 140		160	170	180	190	200	210	220	230	240
FAPPILK	ELEN	LSPE	EAA	нок		. S L L	OED	P W R	V A K I	4 V K
TTCGCGCCACCCATTCTGAA	AGAGCTGGAGAACC	TCAGCCCAGA	GGAGGCAGCC	CACCAGAAA	CCGTGGTGGA	GTCACTTCTI	CAGGAGGAC	CCATGGCGCG	TGGCAAAGA	FCGTCAAG
250 260	270	280	290	300	310	320	330	340	350	360
SYLQQHN										
TCGTACCTGCAGCAACACAA 370 380		400	410	420	430	440	450	460	470	480
Peptide B A L Y T W Y V			0 F T	H A G		TEE	р т с	<b>D F L</b>		
GCGCTGTACACCTGGTACGT	CCOCAAOCAGCGAG	AGGTGGCTCA	GCAATTCACC	CACGCGGGGG	AGGGCGGACT	GATTGAAGAG	CCCACAGGT	GATGAOCTCC	CAACCAAAA	AGCCCCC
490 500	510	520	530	540	550	560	570	580	590	600
RNRFKWG	PASQ	QILF	QAY	ERQ	K N P S	KEE	RET	LVE	ECNI	R A E
AGGAACCOGTTCAAGTOGGG 610 620		640	650	660	670	680	690	700	AGTGCAATAG 710	720
CIQRGVS	P S O A		<b>S N 1</b>	V T F	<b>v p v v</b>	Peptide				
TGCATCCAGAGAGGGGGTGTC	ACCATCGCAGGCCC	AGGGGGCTAGG	CTCCAACCTT	GTCACCGAG	TOCOTOTCTA	CAACTGGTTT	CCCAACCGG	CGCAAGGAAG		
730 740	750	760	770	780	790	800	810	820	830	840
LANDTYN										
CTOGCCATOGACACGTATAA 850 860		GGCCAGGCCC 880	CGGCCCTGCG 890	CTACCTGCCC 900	ACAGTTCCCC 910	OGGCCTGCCC 920	ACAACCACC	CTCTCTCCCA 940	GTAAGGTCC/ 950	ACCCTCTC 960
RYGQSAT										
CGGTATGGACAGTCTGCAAC	CAOCGAGGCAGCTG	AGGTGCCCTC	CAGCAGCGGA	GGTCCCTTA	TCACAGTGTC	TGCGGCCTTA	CACCAAGTG	TCCCCCACAG		
970 980	990	1000	1010	1020	1030	1040	1050	1060	1070	1080
LLSTEAK	LVSA	TGGP	LPP	V S T	LTAL	H S L	EQT	SPG	LNQO	Q P Q
CTGCTGAGCACCGAGGCCAA 1090 1100		CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	1130	GTCAGCACCO 1140	TGACAGCACT 1150	CCACAGCTTO 1160	GAGCAGACG	ILLED 1180	TCAACCAGCA 1190	AGCCGCAG 1200
NLIMASL	всчы	T T C P	6 F P				~	<b>T</b> I V		
AACCTTATCATGGCCTCOCT	OCCTOGGGTCATGA	CCATCGGCCC.	ACCCGACCCC	CCTCCCTG	GTCCCACGTT	CACTAACACO	GGTGCCTCT	ACCCTGGTTA	TTGGTCTGG	CTCCACA
1210 1220	1230	1240	1250	1260	1270	1280	1290	1300	1310	1320
Q A Q S V P V	INSM	GSSL	TTL	QPV	QFSQ	PLH	PSY	QQP	LMPI	PVQ
Q A Q S V P V CAGGCACAGAGCGTGCCAGT 1330 1340	CATCAACAGCATGG	G S S L GGAGCAGCCT 1360	T T L GACCACCCTG 1370	Q P V CAGCCGGTCC 1380	Q F S Q AGTTTTCCCA 1390	PLH GCCACTGCAC 1400	P S Y CCTTCCTATO 1410	Q Q P CAGCAGCCTC 1420	L M P I ICATGCCCCC 1430	PVQ CTGTACAG 1440
CAGGCACAGAGCGTGCCAGT 1330 1340	CATCAACAGCATGG 1350	GGAGCAGCCT 1360	DACCACCCTG 1370	CAGCCGGTCC 1380	AGTTTTCCCA 1390	GCCACTGCAC 1400	CCTTCCTATO 1410	1420	1430	TGTACAG 1440
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACGTGGCCCAGAGTCC	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT	Q S P GCAGAGCCCC	CAGCCGGTCC 1380 H A L CACGCCCTGT	AGTTTTCCCA 1390 Y S H K TACAGCCACAA	GCCACTGCAG 1400 P E V GCCTGAGGTG	A Q Y	T H T :	ICATGCCCCC 1430 L L E CCCTGCTTCC	CTGTACAG 1440 Q T CGCAGACC
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACCTGGCCCAGAGTCC 1450 1460	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480	Q S P GCAGAGCCCC 1490	H A L CACCCCCTCT H A L CACCCCCTCT 1500	AGTTTTCCCA 1390 Y S H K ACAGCCACAA 1510	GCCACTGCAG 1400 P E V GCCTGAGGTG 1520 Peptide C	A Q Y GCCCAGTAC. 1530	T H T : ACCCATACAAN 1540	ICATGCCCCC 1430 L L E CCCTGCTTCC 1550	Q T CCCAGACC 1560
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACGTGGCCCAGAGTCC 1450 1460 M L I T D T N	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T	GACCACCCTG 1370 Q S P GCAGAGCCCCC 1490 P T K	CAGCCGGTCC 1380 H A L CACGCCCTG1 1500	AGTTTTCCCA 1390 Y S H K TACAGCCACAA 1510 T S D T	GCCACTGCAG 1400 P E V GCCTGAGGTG 1520 Peptide C E A S	CCTTCCTAT 1410 A Q Y CCCCAGTAC 1530 D S E P	CAGCAGCCTC: 1420 T H T : ACGCATACAAN 1540 G L H 1	ICATGCCCCC 1430 L L F CCCTGCTTCC 1550 L P S S	P Q T CCCAGACC 1560 P A
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACCTGGCCCAGAGTCC 1450 1460	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGCACCCTTG	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T	GACCACCCTG 1370 Q S P GCAGAGCCCCC 1490 P T K	CAGCCGGTCC 1380 H A L CACGCCCTG1 1500	AGTTTTCCCA 1390 Y S H K TACAGCCACAA 1510 T S D T	GCCACTGCAG 1400 P E V GCCTGAGGTG 1520 Peptide C E A S	CCTTCCTAT 1410 A Q Y CCCCAGTAC 1530 D S E P	CAGCAGCCTC: 1420 T H T : ACGCATACAAN 1540 G L H 1	ICATGCCCCC 1430 L L F CCCTGCTTCC 1550 L P S S	P Q T CCCAGACC 1560 P A
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACGTGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGAGCACCAA 1570 1580	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGCACCCTTG 1590	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T CCAGCCTCAC 1600	Q S P GCAGAGCCCCC 1490 P T K GCCCACCAAG 1610	CAGCCGGTCC 1380 H A L CACGCCCTGT 1500 Q V F CAGGTCTTC/ 1620	AGTTTTCCCA 1390 Y S H K CACAGCCACAA 1510 T S D T ACCTCAGACAC 1630	GCCACTGCAG 1400 P E V GCCTGAGGTG 1520 Peptide (C E A S AGAGGCCTCC 1640	A Q Y GCCCAGTAC. 1530 D S E P AGTGAGCCTC 1650	CAGCAGCCTC: 1420 THT: ACCGCATACAAA 1540 CLH: DGGCTTCATG. 1660	ICATGCCCCC 1430 L L E CCCTGCTTCC 1550 E P S S CCCCGTCGTC 1670	CTGTACAG 1440 CGCAGACC 1560 CF A CTCCAGCC 1680
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGCAGCAGA 1570 1580 T T I H I P S ACCACCATTCACATCGCCG	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCAA	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T CCAGCCTCAC 1600 N I Q H ACATCCAGCA	Q S P GCAGAGCCCCC 1490 P T K GCCCACCAAG 1610 L Q P CCTGCACCCT	A CAGCCGGTCC 1380 H A L CACGCCCTGT 1500 Q V F CAGGTCTTC/ 1620 A H R GCTCACCGGC	AGTTTTCCCA 1390 Y S H K ACAGGCCACAA 1510 T S D T ACCTCAGACAC 1630 L S T S TTCAGCACCAG	GCCACTGCAG 1400 P E V GCCTGAGGTO 1520 Peptide (C E A S AGAGGCCTCC 1640 P T V TCCCACAGTO	CCTTCCTAT( 1410 A Q Y CCCCAGTAC. 1530 D S E P AGTGAGCCT( 1650 S S S TTCCTCCAGCA	CAGCAGCCTC: 1420 T H T : ACGCATACAAN 1540 G L H I DGGCTTCATC. 1660 S L V I AGCCTGGTGT:	ICATGCCCCC 1430 IL L F GCCTGCTTCC 1550 I P S S NGCCGTCGTC 1670 I Y Q S ICTACCAGAG	CTGTACAG 1440 CQ T CGCAGACC 1560 CTCCAGCC 1680 CS D CTTCTGAC
CAGGCACAGACCGTGCCAGT 1330 1340 S H V A Q S P AGCCACGTGGCCCCAGGTCC 1450 1460 H L I T D T N ATGCTGATCACAGACACGAA 1570 1580 T T I H I P S	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCAA	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T CCAGCCTCAC 1600 N I Q H	GACCACCCTG 1370 Q S P GCAGAGCCCC 1490 P T K GCCCCACCAAG 1610 L Q P	CAGCCGGTCC 1380 H A L CACCCCCTGT 1500 Q V F CAGGTCTTC/ 1620 A H R	AGTTTTCCCA 1390 Y S H K AGAGCGCACAA 1510 T S D T AGTCGAGACAG 1630 L S T S	GCCACTGCAG 1400 P E V GCCTGAGGTG 1520 Peptide C B A S AGAGGCCTCC 1640 P T V	CCTTCCTAT( 1410 A Q Y GCCCAGTAC. 1530 D S E P AGTGAGCCT( 1650 S S S	CAGCAGCCTC: 1420 T H T : ACGCATACAAA 1540 C L H : CGCCTTCATG. 1660 S L V :	ICATGCCCCC 1430 ILLE CCCTGCTTCC 1550 IPSS CCCGTCGTC 1670 ILYOS	CTGTACAG 1440 P Q T CGCAGACC 1560 S P A CTCCAGCC 1680 S S D
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCAGGTGGCCCAGAGTCC 1450 1460 H L I T D T N ATGCTGATCACAGGCAGCAA 1570 1580 T T I H I P S ACAACCATTCACATCCCCGAG 1690 1700 S N G H S H L	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGACCCTTG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N	GGAGCAGCCT 1360 M A Q L TCGCCCAGCT 1480 A S L T CCAGCCTCAC 1600 N I Q H ACATCCAGCA 1720 H G V I	GACCACCCTC 1370 Q S P GCAGAGCCCCC 1490 P T K GCCCACCAAG 1610 L Q P CCTGCACCCTT 1730 E T F	CAGCCGGTCC 1380 H A L CACCCCCTGT 1500 Q V F CAGGTCTTC/ 1620 A H R GCTCACCGCC 1740 I S T	AGTTITICCA         1390           Y         S         H         K           ACAGECCACAA         1510         T         S         D         T           IT         S         D         T         GCTCAGACAC         1630         1630           L         S         T         S         D         T         CCTCAGACACC         1630         1750         1750         Q         M         A         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S         S	GCCACTGCAG 1400 P E V GCCTGAGGTC 1520 Peptide C E A S AGAGGCCTCC 1640 P T V TCCCACAGTC 1760 S S Q	A Q Y A Q Y ACCCCAGTAC. 1530 D S E P AGTGAGCCTM 1650 S S S TCCTCCCAGCA 1770 *	CAGCAGCCTC: 1420 T H T : ACGCATACAAA 1540 C L H I CGCCTTCATCA 1660 S L V I ACCCTCGTGTT 1780	CATGCCCCC 1430 S L L E GCCTGCTTCC 1550 E P S S GCCCGTCGTC 1670 L Y Q S GTACCAGAG 1790	CTCTACAG 1440 Q T GCCAGACC 1560 S P A CTCCAGCC 1680 S D STTCTGAC 1800
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGGAGCAGCAG T I H I P S ACAACCATTCACATCCGCAG 1690 1700 S N G H S H L TCCAACGGGCACAGCCACCT 1810 1820	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGCACCCCTG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCCATCCAACCC 1830	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T CCAGCCTCAC 1600 N I Q H ACATCCAGCA 1720 H G V I ACGGTGTCAT 1840	GACCACCCTC 1370 Q S P GCAGAGCCCCC 1490 P T K GCCCACCAAG 1610 L Q P CCTGCAGCCTT 1730 E T F CGAGACTTTT 1850	CAGCCGGTCC 1380 H A L CACCCCCTGT 1500 Q V F CAGGTCTTCC 1620 A H R GCTCACCGGC 1740 I S T ATCTCCACCG 1860	AGTTITICCA 1390 Y S H K ACAGECCACAA 1510 T S D T ACCTCAGACACC 1630 L S T S TCAGECACCAG 1750 Q H A S AGATOCCTCC 1870	GCCACTGCAG 1400 P E V GCCTGAGGTC 1520 Peptide C E A 3 IAGAGGCCTCC 1640 P T V TCCCACAGTC 1760 S S Q CTCCTCCCAG 1880	CCTTCCTATO 1410 A Q Y CCCCAGTAC. 1530 D S P 1650 S S S TCCTCCAGCA 1770 * TAACCATGCC 1890	CAGCAGCCTC: 1420 T H T : ACGCATACAAN 1540 G L H I SGCCTTCATG 1660 S L V I AGCCTGGTGT: 1780 TGACTGCCTCC 1900	I A TOCCOCCI I A 30 I L L E SCCTGCTTCC I 550 I P S S NOCCGTCGTC I 670 I Y Q S IGTACCAGAG I 790 CCAGGAGCTC 1910	CTGTACÁG 1440 Q T GGCAGACC 1560 S P A TTCCAGCC 1680 S D TTCTGAC 1800 CGCCTCCC 1920
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACGTGGCCCCAGGTCC 1450 1460 H L I T D T N ATGCTGATCGACAGCAGCAGA 1570 1580 T T I H I P S ACAACCATTCACAGCACCCACC 1690 1700 S N G H S H L TCCAACGGGCACAGCCCACCT	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGACCCTTG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCCATCCAACC 1830 GGAGGGCCACACCC	GGAGCAGCCT 1360 M A Q L TGGCCCAGCT 1480 A S L T CCAGCCTCAC 1600 N I Q H ACATCCAGCA 1720 H G V I ACGGTGTCAT 1840	GACCACCCTC 1370 Q S P GCAGAGCCCC 1490 P T K GCCCACCAAG 1610 L Q P CCTGCACGCCT 1730 E T F CGAGACTTTT 1850 GAGCGTCTTG	CACCCCGCTCC 1380 H A L CACCCCCTCT 1500 Q V F CACGCTCTCA 1620 A H R GCTCACCGCC 1740 I S T ATCTCCACCGC 1860 GAGCCTGCCA	AGTITICCCA 1390 Y S H K ACAGCCACAA 1510 T S D T ICCTCAGACAC 1630 L S T S TCAGCACCAG 1750 Q H A S AGATOGCCTC 1870 1670	GCCACTCCAC 1400 P E V GCCTCACGTC 1520 Peptide C E A S AGAGGCCTCC 1640 S S Q CTCCCCCAC 1860 S S Q CTCCTCCCCAC	CCTTCCTATU 1410 A Q Y CCCCAGTAC: 1530 D S E P TAGTGACCCT 1550 S S S TCCTCCAGCA 1770 * TAACCATGG: 1890	CAGCAGCCTC: 1420 T H T : ACCCATACAA 1540 C L E 1 COCCTTCATG 1660 S L V 1 ACCCTCGTGT 1780 FGACTCCCTCC 1900 FTCTATGCCT	It and the second secon	CTGTACAG 1440 P Q T CGCAGACC 1560 S P A TTCCAGCC 1680 S S D FTTCTCAC 1800
CAGGCACAGAGCGTGCCACT 1330 1340 S H V A Q S P ACCCACGTGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGGACACCAA 1570 1580 T T I H I P S ACAACCATTCACATCCCCGA 1690 1700 S N C H S H L TCCAACGGGCACAGCCACCT 1810 1820 AGAGCCTGCACAGGGGGGAGA 1930 1940 CTCCATCATCAGAAAGGGAT	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 CCAGGACCCCTTCA 1710 L P S N GCTGCCATCCAACG 1830 GGAGGGCCACAGCC 1950 GGCTGTCAGGTGTC	GCAGCAGCCT 1360 M A Q L TGCCCAGCT 1480 A S L T CCAGCCTCAC 1640 N I Q H ACGCTCTCAT 1840 ATGCTGCCTG 1860 TGCCTGCTGCTG	ZACCACCCTO 1370 Q S P CACAGACCCCC 1490 P T K CCCCCACCAAG 1610 L Q P CCTCCAACCCT 1730 E T F CCACACCTTTC 1970 CACCCCTCTCA	CACCCOGCTCC 1380 H A L CACGCCTGT 1500 Q V P CACGTCTTCI 1620 A H R CGTCACCCC 1740 I S T ATCTCACCC 1860 GACCCTCACCC 1980 GACCCTCCACC	AGTITICCCA 1390 Y S H K ACAGCCCACA 1510 T S D T CCTCAGACAC 1630 L S T S CCTCAGACAC 1630 Q H A S ACAGTOCCTC 1870 Q H A S ACAGTOCCCTC 1870	GCCATCCAC 1400 P E V GCCTCACGTT 1520 Petile C E A 3 ACAGECCTCC 1640 S S Q CTCCTCCACT 1880 GCTCCTGCCC 2000	CCTTCCTAT 1410 A Q Y GCCCAGTAC 1530 D S E P AGTGAGCCT 1650 S S S TCCCCAGAC 1090 TTCCCCAGAM 2010 TTCCCCAGAM	CAGCAGCCTC: 1420 T H T :: ACGCATACAAN 1540 C L H :: 2020 CTCATCCTCATC 1660 S L V :: 460 CCCCGTCTCT 1780 CCCCCTCCTCC 1900 CTCATCCCTC 2020 CCTGCTGCTGCTC	CARGECCCC 1430 S L L F S L L F SCCCGTCCTC 1550 E P S S SCCCGTCCTC 1670 L Y Q S SCCCGTCCTCC 1790 CCAGGAGCTCC 1910 CCAGGAGCTCC 1910 CCAGGACTCCC	STGTACAG           1440           P         Q           T         CGCAGACC           1560         S           S         P           TTCCAGCC         1680           S         S           TTCTCACC         1800           COCCTCCC         1920           NGCTOCTCC         2040           AACTTAA         1800
CAGGCACAGACCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGAGTCC 1450 1460 H L I T D T N ATGCTGATCACAGACAGCAA 1570 1580 T T I H I P S ACAACCATTCACAGACACCAA 1690 1700 S N C H S H L TCCAACGGCACAGCGCACA 0 1930 1940 CTCCATCATCAGAAGGGGAA	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCCGATCCAACC 1830 GGCGCCCACACCC 1950 GGCTCTCAAGGTGTC 2070	GCACCARCCT 1360 M A Q L TGCCCACGCT 1680 N I Q H ACATCCACCA 1720 M G V I ACGCTGTCAT 1840 ATGCTGCCTG 1960 TCCTCAGTCC 2080	CACCACCCTC 1370 CACAGACCCCC 1490 P T K GCCCACCACC 1610 L Q P CCTCCACCCTC 1730 E T F CCACCCTCTC 1730 E T F CCACCCTCTC 1730 CACCCCCCC 1610 CACCCCCCCCC 1610 CACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGCTCC 1380 H A L CACCOCCCTTC 1500 Q Y F CACCTTCC 1520 A H R CCTCACCCCC 1540 I S T ATCTCCACC 1560 CACCCCCCCC 1580 CACCCCCCCCC 1580 CACCCCCCCCCCCCCC 1580 CACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTITICCCA 1390 Y S H K ACACCCCACA 1510 <b>I S D I</b> CCCCCCACACACAC 1530 L S T S TCCACCACACAC 1530 Q H A S ACATOCCCTC 1870 CCCCCCCACAG 1990 GCCCGCACACAC	CCCCACTCCAC 1400 P E V CCCTCACGTT 1520 Peptide C E A S ACAGECCTCC 1640 S S Q CTCCTCCCACACTC 1860 CCCCCCCCCC 2000 CCCCACTCCCCCCC 2000	A Q Y CCCTCCTAT 1410 A Q Y CCCCAGTAC. 1530 B S E P AGTGACCCT 1650 S S S TCCTCCAGCA 1770 * TAACCATGC 1890 TCCCCGCAA 2010 2010 2010 2010 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1770 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700 1700	CACCACCCTC:         1420           T         H         T:           CACCATACAN         1540           C         L         H           DSGCTTCATC         1660           S         L         V           NGCCTGCTGT:         1780           TCAATCGCTC         12000           TCTATGCCTC         2020           CCTGCTGGT:         2140	rcATCCCCCC 1430 S L L E SCCTCCTTC 1550 F P S S SCCCGTCCTC 1670 L Y Q S TGTACCACAGA 1790 CCAGGAGCTC 1910 CCAGGAGCTC 2030 CCAAGAACTC	TGTACAG           1440           ? Q T           GCGCAGACC           1550           ? F A           TCCAGC           1680           S S D           TTTCTCACC           1800           SGCCTCCC           1920           NGCTCCCC           2040           AACTTAA           2160
CAGGCACAGAGCGTGCCACT 1330 1340 S H V A Q S P AGCCACGTGCCCCCAGGTCC 1450 1460 H L I T D T N ATGCTGATCACAGACACAAA 1570 1580 T T I H I P S ACAACCATTCACAGACACCACA 1690 1700 S N C H S H L TCCAACGGGCACAGCCACCT 1810 1820 CTCCATCAGAAAGGGGACA 1930 1940 CTCCATCAGAAAGGGGACA 2050 2060 TGCTTGGAACAGAGGGGACA 2170 2180	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGACCCTTG 1590 Q D P S CCAGGACCCCTCAA 1710 L P S N GCTGCCATCCAACCC 1830 GGCTGCCACCACCC 2070 AGGCCTGCTGCTTGCTCACCC 2190	GCACCLACCT 1360 M A Q L TGCCCACCT 1680 A S L T CCACCTCAC 1600 N I Q H ACATCCACCA 1720 H C V I ACGGTGTCAT 1840 TCCTCAGT 1950 TCCTCAGTC 2080 TGCCACCCTC 2200	CACCACCCTC 1370 Q S P CCACACCCCC 1490 P T K GCCCCCCCACA L Q P CCCCCCCCACCAC 1610 L Q P CCCCCCCCACCAC 1730 CCCCCCCCACCC 1970 CCCCCCCCACCC 1970 CCCCCCCCCCCC 1970 CCCCCCCCCCCCCC 1970 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOCCTC 1380 H A L CACCCCCTTC 1500 Q V F CACCCCCTC 1520 A H R CCTCACCCC 1540 I S T ATCTCCACCC 1560 CCCCCAACC 1960 CCCCCAACC 2100 CCCCCAACC 1960 CCCCCCAACC 1200 CCCCCAACC 1200 CCCCCAACC 1200 CCCCCAACC 1200 CCCCCAACC 1200 CCCCCAACC 1200 CCCCCCACCCCCC 1360 CCCCCCCCCCC 1360 CCCCCCCCCCCC 1360 CCCCCCCCCCCCC 1360 CCCCCCCCCCCCCC 1360 CCCCCCCCCCCCCCC 1360 CCCCCCCCCCCCCCC 1360 CCCCCCCCCCCCCCCC 1360 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTITICCCA 1390 Y S H K KACAGCCACAA 1510 T S D T ICCCTCAGACAC 1630 L S T S TCCAGCACCA 1630 Q H A S XAGATOGCCT 1870 ICCCTCCCACAC 1990 ICCCCCGCAGACA 1990 ICCCCCGCACAC 1990 ICCCCCCCACACA 1990 ICCCCCCCACACAC 1990 ICCCCCCCACACAC 1990 ICCCCCCCCACACAC 1990 ICCCCCCCCACACAC 1990 ICCCCCCCCACACACAC 1990 ICCCCCCCCACACACACACACACACACACACACACACAC	CCCCATTCCAC 1400 P E V CCCTGAGGTC 1520 Peptide C E A S AGAGGCCTCC 1640 P T V TCCCACAGTC 1880 CTCCTCCCCAC 2000 CCCCATTCC 2120 CTCCTCGACCT 2240	A Q Y GCCCACGTAC 1530 B S F P AGTGACCCT 1550 S S S TCCCCCAAC 1770 * TAACCATGC 1890 TTCCCCAAC 100 GACCGCCATU 2130 CGGACGCCCT 2250	CACCACCTC:         1420           I         H         T           CACCCATACAN         1540           G         L         H           DGGCTTCATG         1660           S         L         V           IGGCTCCATC         1780           FGACTGCCTCT         2020           CCCTGCTGGTT         2140           FGACTGCACACA:         2260	TCATECECCC 1430 S L L F S L L F SCOTGCTTCC 1550 E P S S GCCGCTCCT 1670 L Y Q S TCATCCTCCL 1910 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCL 2030 TCATCCTCCCL 2030 TCATCCTCCCL 2030 TCATCCTCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCCL 2030 TCATCCTCCCCCL 2030 TCATCCTCCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCCTCCCCL 2030 TCATCTCTCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCL 2030 TCATCTCCCCL 2030 TCATCTCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCCL 2030 TCATCTCCCCL 2030 TCATCTCCCCL 2030 TCATCTCCCCCCL 2030 TCATCTCCCCCCL 2030 TCATCTCCCCCCL 2030 TCATCTCCCCCCCCCL 2030 TCATCTCCCCCCCCCCCL 2030 TCATCTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	TGTACAG           1440           ? Q T           GCGCAGACC           1560           S F A           TTCCAGCC           1680           S S D           TTTCTCAC           1800           SCCCTCCC           2040           AACTTAA           2280           2280
CAGGCACAGACCGTGCCACT 1330 1340 S H V A Q S P ACCCACGTGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGAGCACCAA 1570 1580 T T I H I P S ACAACCATTCACATCCCCAG 1690 1700 S N C H S H L TCCAACGGGCACAGCCACT 1810 1820 AGAGCCTGCACAGGGGGAGA 1930 1940 CTCCATCATCAGAAGGGGGA 2170 2180 CTGCTGCCCCCGCCCCCCC 2290 2300	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCCTG 1590 CCACGCCCGTCAA 1710 L P S N GCTGCCATCCAACCC 1830 GGAGGGCCCACAGCC 1950 GGCCTCTGAGGTGTC 2070 AGCCCTGCTATTCC 2190 CACCCTGCGCCAGA 2310	GCAGCAGCCT 1360 M A Q L TGCCCAGCT 1480 A S L T CCAGCCTCAC 1640 N I Q H ACGCTGTCAT 1840 ATGCTGCCTC 2080 TGCCACGTCC 2080 GCCCATGTG 2200 GACCATGTG 22320	ZACCACCCTO 1370 Q S P CACAGAGCCCC 1490 P T K GCCCACCAAG 1610 L Q P CCATCACCATT 1730 E T F CCACGCTOTTC 1970 CACGCCTOTTC 1970 CACGCCTOTTC 2010 CACTCACACC 2210 CCCTTCGTCG 2330	CACCCOGCTCC 1380 H A L CACGCCTGT 1500 Q V P CACGTCTCC 1620 A H R CACGTCTCCACCC 1740 I S T ATCTCCACCC 1860 CACCCTCACCC 2100 CTTCCACCCC 2220 CTTCCACCCC 22300	AGTITICCCA 1390 Y S H K ACAGCCCACA 1510 T S D T CCTCAGACAC 1630 L S T S TCACACCCCA 1630 Q H A S ACATOCCCTC 1870 Q H A S ACATOCCCCC 1870 CCTCCCACAG 2110 CTCAACACATC 2230	CCCCACTCCAC 1400 P E V CCCTCACGTC 1520 Petile C E A 3 ACAGECCTCC 1640 P T V CCCCCACGTC 1660 S S Q CTCCTCCCCAC 1880 CCCCCACGCC 2000 CCCCCACGCC 2120 CCCCCACGCC 2240 ATCTTCCACG	CCTTCCTAT 1410 A Q Y SCCCAGTAC. 1530 D S E P AGTGACCCT 1650 S S S TTCCCAGAC. 1770 + TTACCATGC: 1890 TTCCCAGAM 2010 CGGAGCCCC 2250 CGGAGCCCCTCTCT 2370	CAGGAGECTC:         1420           T H T :         ACGCATACAAN           ACGCATACAAN         1540           C L H :         BOGCTTCATG           1660         S L V :           SCCCTGTCGTT         1780           TGACTCCCTCC         1900           CTCTATGCCT:         2140           TCACATCAACA:         2260           SCCTGCTGGTCT         2380	FCATECCCC         1430           S         L         L           S         L         L           S         CCCTCCTTCC         1550           E         P         S           S         CCCGTCCTCC         1670           L         Y         Q           S         CCACGACCTCC         1910           S         CCACGACCTCC         1910           S         CCACGACCTCC         2030           CCACGACCTAC         1250         INGECTCCCTAC           S         CCACGCCTAC         2390	TGTACAG           1440           P Q T           CGCAGACC           1560           S P A           TTCCAGCC           1680           S S D           TTTCTCAC           1800           CGCTCCC           1920           NGCTCCCC           1920           CGCTCCC           1920           CGCTCCCC           1920           CGCTCCCC           1920           NGCTGCTCC           2400           XACTTAA           2160           CTGCCCGCG           2280           SGCAGCAG           2400
CAGGCACAGACCTGCCACT 1330 1340 S H V A Q S P ACCCACGTGCCCCAGGTCC 1450 1460 H L I T D T N ATGCTGATGACAGACAGA 1570 1580 T T I H I P S ACAACCATTCACAGACACCA 1690 1700 S N G H S H L TCCAACGGGCACAGCGCACT 1810 1820 AGGCCTGCACAGGGGGAA 2170 2180 CTGTGCCTCCCAGGCCACTC 2290 2300	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTO 1590 Q D P S CCAGGACCCCTCAA 1710 CGTGCCATCCAACC 1830 GGCTGCCATCCAACC 1950 GGCTCCTACACCCC 2190 CACCCTGCGCCCAA 2310 TACTAAGTCGCTCT	GCACCARCECT 1360 M A Q L TGCCCACGCT 1680 N I Q H ACATCCACCA 1720 N I Q H ACATCCACCA 1720 H G V I ACGCTGTCAT 1860 ATGCTGCCCTG 2080 GACCCATCTG 2200 GACCCATCTG 2200	CACCACCCTC 1370 Q S P CACAGACCCCC 1490 P T K GCCCACCACC 1610 L Q P CCTCCACCCCT 1730 E T F CCACCACCTT 1970 CACCACCCCTC 1970 CACCACCCCCC 2090 CACCACACCCCC 2330 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGCTCC 1380 H A L CACCOCCCTTC 1500 Q V F CACCCTCTCC 1520 A H R GCTCACCCCC 1540 I S T ATCTCCACC 1560 GACCTCCCCC 1580 GACCTCCCCC 2140 GCTTACCCCC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCTC 2140 GCTTACCCCCCC 2140 GCTTACCCCCCC 2140 GCTTACCCCCCCCC 1980 GCTTACCCCCCCCCCCC 1980 GCTTACCCCCCCCCCCCCCC 1980 GCTTACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTITICCCA 1390 Y S H K ACACCCACACA 1510 <b>I S D I</b> CCCCCAGACAG 1530 L S T S TCCACCACAG 1750 Q H A S XCATOCCCTC 1870 CCTCCCACAG 1990 CCCCCCCACAG 2230 CCTCCACAGG 2350 CCTCCACAGG 2350	CCCCATCCAC 1400 P E V CCCTCACGTT 1520 Peptide C E A S ACAGECCTCC 1540 P T V TCCCACACTC 1760 S S Q CTCCTCCCAC 2000 CCCCTCCCCCCA 2000 CCCCCCCCCC 2000 CCCCCCCCCC 2120 CTCCTCCCCCCC 2120 CCCCCCCCCC 2120 CCCCCCCCCC 2120 CCCCCCCCCCC 2120 CCCCCCCCCCC 2120 CCCCCCCCCCC 2120 CCCCCCCCCCC 2120 CCCCCCCCCCC 2120 CCCCCCCCCCCCCCC 2120 CCCCCCCCCCCCCCCCCCC 2120 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCTTCCTAT 1410 A Q Y GCCCACTACA 1530 S F P AGTGAGCCT 1650 S S S TCCTCCAGCA 1770 * TAACCATGC 1890 TTCCCAGCAT 2010 GGCCGCCTTC 2130 CGGAGCCCCC	CACCACCCTC         1420           T         H         T           CACCATACAN         1540           C         L         H           DOCCTTCATC         1660           S         L         V           NGCCTCGTGTT         1780           TCAATGCCTC         1900           TCTATGCCTC         2020           CCCTCGTGGT         2140           FACATCACAT         2260           SCCTAGCTATC         2380           ANGCTAGCTACATCA         2020	TCATECCCC         1430           S         L         L           S         C         L         F           S         C         L         F           S         C         L         F           S         C         CCTOCTTCC         1550           F         P         S         S           GCCCGTCGTC         1670         1670           C         AGGCGTCGTC         1910           FCATGCTGCA         2030         CAGGAGCCTG           1910         FCATGCTGCA         2030           FCGCGTCGCCAT         2030         CAGGAGCCTAG           2030         CAGCATAGAA         2470	TETACAG           1440           ? Q T           GCGCAGACC           1560           S P A           TCCCAGCC           1800           S S D           FTTCTGAC           1800           SGCCTCCC           1920           IGCTCCTC           2040           ALACTTAA           2160           CCTOTCTC           2280           GGAGCAGC           22800           GGCTCCT
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGGACACCAA 1570 1580 T T I H I P S ACAACCATTCACATCCCCGA 1690 1700 S N C H S H L TCCAACGGGCACAGCCACCT 1810 1820 AGAGCCTGCACAGGGGGAGA 1930 1940 CTCCATCATCAGAAGCGGGA 2170 2180 CTGCTGCCCCCGGCCATC 2290 2300 GGCTCTCC TAGCGTTTCCCC	CATCAACAGCATGG 1350 F H A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCCTG 1590 CCACGCCCGTCAACCA 1710 L P S N GCTGCCATCCAACCC 1830 GGAGGGCCCACAGCCC 1830 GGAGGGCCCACAGCCC 1950 GGCCTCTGAGGTGTC 2070 AGCCCTGCTATGCC 2190 CACCCTGCGCCAGA 2310 TACTAAGTCGCTCT 2430 CTTCTTCTCCTTAGGG	GCAGCACCTACT 1360 M A Q L TGCCCAGCT 1480 A S L T CCAGCCTCAC 1640 N I Q H ACGCTGTCAT 1840 ATGCTGCCTC 2080 TGCCCACTCC 2080 TGCCCACTCC 2080 TGCCCACTCC 2080 TGCCCACTCC 2080 TGCCCACTCC 2080 TGCCCACTCC 2080 TGCCCCCCCC 2080 TGCCCCCCCC 2080 TGCCCCCCCC 2080 TGCCCCCCCCC 2080 TGCCCCCCCCCC 2080 TGCCCCCCCCCC 2080 TGCCCCCCCCCCC 2080 TGCCCCCCCCCCCCC 2080 TGCCCCCCCCCCCCC 2080 TGCCCCCCCCCCCCC 2080 TGCCCCCCCCCCCCCCCC 2080 TGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	ZACCACCCTO 1370 Q S P CACAGAGCCCC 1490 P T K GCCCACCAAG 1610 L Q P CCACGCCACCAAG 1730 E T F CCACGCCTOTACCC 1970 CACGCCACACCT 2330 CCACGACCT 2450	CACCCOGGTCC 1380 H A L CACGCCTGT 1500 Q V P CACGTCTCC 1620 A H R CCTCACCCACC 1740 I S T ATCTCCACCC 1860 CACCTCACCACC 2100 CTTCCACCCC 2220 CCTCAACCC 2100 CCTCACCCCC 2230 CACTTCTTTC 2360 CACCTCCTAC	AGTITICCCA 1390 Y S H K ACAGCCCACA 1510 T S D T CCTCAGACAC 1630 L S T S TCACACCCCACAC 1750 Q H A S ACATOCCCT 1870 CCTCCCACACA 2110 CTCACACACCACACAC 2130 CCTCCACACAC 2350 CCTCACACACGTAC 2470 CCTCACACACGTAC	CCCCACTCCAC 1400 P E V CCCTCACGTC 1520 Petile C E A 3 ACAGGCCTCC 1640 P T V CCCCACATC 1760 S S Q CTCCTCCCAC 1880 CCCCCTCCCCCC 2120 CCCCCACCC 2240 ACTTCCACGC 2240 CTCTCCACCC 2350 TCTTCCACCC	CCTTCCTATC 1410 A Q Y SCCCAGTAC. 1530 D S E P AGTGAGCCT 1650 S S S TTCCTCCACC. 1770 * TTACCATGC: 1890 TTCCCAGAM 2010 CGGAGGCCC 2250 CCACCTTCCTC 2370 CCACCTTCCTC	CACCACCCTC:         1420           T H T :         ACCCATACAAN           ACGCATACAAN         1540           C L H :         1500           C L H :         1660           S L V :         1660           S CCCTGTCGTT         1780           FGACTGCCTCT         1780           CTCATGCCTTC         2140           FGACTGCGCTTC         2160           CACATCCAACAC         2280           CACTGCGGTT         2380           CACTTCCCTC         2300           CACTTCCCTC         2300	TCATECECCC         1430           1430         1430           S         L         L           S         L         L           S         CCCTCCTTCC         1550           E         P         S           S         CCCGTCCTCC         1670           L         Y         Q           S         CCACGACCTGTC         1910           S         CCACGACCTGC         2010           S         CCACGACCTGC         2390           S         CCACTAGAAA         2310           S         CCACTAGAAA         2510	TGTACAG           1440           P         Q           T         TGCAGACC           1560           S         P           TTCCCAGCC         1680           S         S           TTTCTAC         1800           GCCTCCC         1920           NGCTCCCC         1920           NGCTCCCC         2040           AACTTAA         2160           CTGTCTCC         2280           NGCAGCAGC         2400           AGCTGCTCC         2280           NGCAGCAGC         2400           AGCTGCTCC         2280           NGCAGCAGC         2400           AGCTGCTCC         2280           NGCAGCAGCAG         2400           AGCTGCTGCT         2280           NGCAGCAGCAG         2400
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCACGTGGCCCCAGGGTCC 1450 1460 H L I T D T N ATGCTGATCGACAGACAGCAG 1570 1580 T T I H I P S ACAACCATTCACAGGCACAGC 1690 1700 S N G H S H L TCCAACGGCGCACAGCCACCT 1810 1820 AGGCCTGCACGGCGACAGCCACCT 1810 1820 AGGCCTGCACGGCGACAGCGGGA 2170 2180 GCTGTGCTCGCAGGCGCACTC 2290 2300 GGCTTCCCCAGGCCTTCCCC 2410 2420 TCAGGGCTAGGATTGCACC 2530 2540	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTGG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCCATCCAACCC 1830 GGAGGGCCCACAACCC 1950 GGCGCCTTCAAGGGGTCT 2130 ACGCCCTGCGCCCAA 2310 TACTAGTCGCCTT 2430 CTTGTTCCTTAGGC 2550	GCACCACCCT 1360 M A Q L TGCCCACCT 1680 A S L T CCACCTCAC 1690 N I Q H ACATCCACCAC 1720 H G V I ACATCCACCA 1720 H G V I ACGCTGCCAT 2000 TGCACCCTC 2000 TGCACCCTCC 2000 TGCACCCTCC 2000 TGCCCCCCCCC 2000 CCCCAACTCC 2550	ZACCACCCTO 1370 Q S P CACAGACCCCC 1490 P I K GCCCACCAAG 1610 L Q P CCTOCACCCAT 1730 E T F CCACGACCTTTT 1730 E T F CCACGACCTTTT 1850 CACCCACCACT 2090 CACTCAGACC 2110 CCCTCTTGCTCTTC 2090 CACCCACGACC 2210 CCCTCTCCACGACC 2350	CACCCOGGTCC 1380 H A L CACCOCCTT 1500 Q Y F CACCTTCTC 1500 Q Y F CACCTCTCC 1620 A H R CCTCACCCCC 1740 I S T ACTTCACCCC 1740 I S T ACTTCACCCC 1740 I S T ACTTCACCCC 1740 I S T ACTTCACCCC 2200 CCTCACCCCCCC 2130 CCTCACCCCCCCCC 2130 CCTCACCCCCCCCCCCC 2130 CCTCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTTITECCA 1390 Y S H K KACACCCACA KACACCCACA 1510 I S D I CCCTCCACAGACAC 1530 L S T S TCAGCACCCC 1550 Q H A S XAGATGCCTC 1870 Q H A S XAGATGCCTCACAC 1990 CCCTCCCACAG 2130 TCAAGGACTCT 2230	CCCCTCCACC 1400 P E V CCCTCACGTT 1520 Peptide C E A S ACAGCCCTCC 1540 PT V TCCCACACTC 1540 S S Q CTCCTCCACC 2000 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCCCCCCCCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCTCACCC 2180 CCCCCCCCCCCC 2180 CCCCCCCCCCCC 2180 CCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCCCCCCCCCCC 2180 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCTTCCTAT 1410 A Q Y 530 5 E P 1650 5 S S TCCTCCAGCA 1650 5 S S TCCTCCAGCA 1770 * TAACCATGG: 1890 TTCCCAGCA 2010 CGAGGCCCT 2250 CGAGCTCTGG 2370 CCACCTTGG 2370 CCACCTTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 2450 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCTGG 250 CCACCGC CCACCTGG 250 CCACCGCCCTG 250 CCACCGCCCTG 250 CCACCGCCCTG 250 CCACCGCCCCTG 250 CCACCGCCCCTG 250 CCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CAGCAGCCTC: 1420 T H T : CGCCATACAM: 1540 C L E : DCCCTTCATC: 1540 C L E : DCCCTCCTCATC: 1780 C L V : AGCCTGCTGGT: 1780 CCTGCCTGGT: 2020 CCTGCTGGT: 2140 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 2240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGGT: 240 FCACTGCCTGCTGCTGCT: 240 FCACTGCCTGCTGCT: 240 FCACTGCCTGCTGCT: 240 FCACTGCCTGCTCCTGCT: 240 FCACTGCCTGCTGCT: 240 FCACTGCCTGCTGCT: 240 FCACTGCCTGCTCCTCCTCCTCCTCCTCCTCCTCCTGCT: 240 FCACTGCCTGCTCCTCCTCCTCCTCCTCCTCCTCCTCCTCCTC	TCATECCCC         1430           1430         1430           S         L         L           SCCTCCTTCC         1550           E         P         S           SCCTCCTTCC         1550           E         P         S           SCCTCCTCC         1570           C         1670           L         Y         Q           SCACGCATCACACAC         1910           TCCATECTCCCA         2030           TCCATECTCCCAT         2150           TCCCTCCCAT         2390           CCASCACTAGAAA         2350           CCASCACAAA         2370           SCGACCTAGAAA         2360           CCASCAGACTGC         2390           CCASCAGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	TGTACAG           1440           ? Q T           GCGCAGACC           1560           3 P A           TTCCAGACC           1680           5 S D           FTTCTCAC           1800           GGCTCCC           1920           GGCTCCC           1920           GGCTGCTC           2040           CGGCTCCC           2240           GGCAGCAG           2400           GGCAGCAG           2520           XAGAGACC           2540
CAGGCACAGACCGTGCCACT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGGGTCC 1450 1460 H L I T D T N ATGCTGATCGACAGACACGAA 1570 1580 T T I H I P S ACAACCATTCACAGCACCCACT 1690 1700 S N C H S H L TCCAACGGCGACAGCCACCT 1810 1820 AGGCCTGCACAGGGGGAA 1930 1940 CTCCATCGTACAGAGGGGAA 1930 1940 CTCTGCACCGCGACAGCGGGAA 2170 2180 CTGTGCCTCCCAGGCCACTC 2290 2300 CTGTGCCTCCCAGGCCACTC 2210 240 CTCGGCCTAGCCTTTCCCC 230 2540 TGTACACTCGCAAGACAGACA	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCCAA 1710 L P S N GCTGCCATCCAACCA 1830 GGCTGCAGTCCAACCA 1950 GGCTCGCACCACGCC 2190 CACCCTGCGCCAGA 2310 CACCCTGCGCCCTG 2430 CCTTGTCCTTAGGC 2550 GCCTCGGCAATGCC 2570	GCACCACCCT 1360 M A Q L TGCCCACGCT 1680 N I Q H ACATCCACCAC 1720 N I Q H ACATCCACCAC 1720 H G V I ACGGTGTCAT 1860 ATGCTGCCCG 2080 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2200 GACCCATCTCC 2560	CACCACCCTC 1370 Q S P CACAGACCCC 1490 P T K GCCACCACCA 1610 L Q P CCTCCACCCC CACCACT 1730 E T F CCACCACT 1730 E T F CCACCACT 2330 CCACCCCCTC 2330 CCACCACCTC 2330 CCACCACCTC 2330 CCACCACCTC 2370 CCACCACCTC 2570	CACCCOGGTCC 1380 H A L CACCOCCCTTC 1500 Q V F CACCOCCTTC 1500 Q V F CACCOCCTTC 1500 A H R CCTCACCCCC 1540 CACCCCCCC 1980 CACCCCCCCC 2140 CCCCCCCCCC 2140 CCCCCCCCCCCCCCCCC 1980 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTTITECCA 1390 Y S H K KACACCCACA KACACCCACA 1510 I S D I CCCCCCACACACAC 1530 L S T S TECACCACCAC 1530 CCCCCCACACACAC 1590 CCCCCCACACACACAC 1590 CCCCCCCACACACACACACACACACACACACACACACA	GCCACTCCAC           1400           P E V           CCCTGACGTC           SZ0           Peptide C           E A S           AGAGECCTCC           1540           P T V           TCCCACACTC           1760           S S Q           CTCCTCCCACACTC           1880           GCTGCTGCCC           2000           GCCCACATCTC           2120           CTGCTCCACACT           2360           CCTCTCCCACCC           2500           AAGGECCTCC           2500           AAGGECCTCC           2400           CCTCTCGCCCC           2500	CCTTCCTAT 1410 A Q Y CCCCACTAC 1530 S E P AGTGACCCT 1650 S S S TCCCTCCACCA 1770 * TTAACCATOC 1890 TTCCCACACATOC 2010 GACGCCCCTT 2130 CGACGTCCTCT 2330 CGACGTCCCTT 2330 CGACGCCCTA 2490 CGACCTCCCTACACATOC 2490 CGACCTCCCTACACATOC 2490 CGACCTCCCTACACATOC 2490 CGACCTCCCTACACATOC 2490 CGACCTCCCCTACACATOC 2490 CGACCTCCCCCTACACATOC 2400 CGACCTCCCCCCTACACATOC 2400 CGACCTCCCCCCTACACATOC 2400 CGACCTCCCCCCCTACACATOC 2400 CGACCTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCACCCTC:         1420           T         H         T:           CACCATACAN         1540           C         L         H           DGCCTTCATC         1660           S         L         V           MACCTCCTCATC         1780           TCAACCCCCCCCTGGT:         2020           CCCTCGTGGT:         2140           FACATCACACA:         2240           CACACCACACA:         2500           MACACTCTCTT         2620           CATACTCACAT         2260           CACACTCTCATC         2620           CATACTCACA:         2740	Conservation           1430           S         L           S         L           Controctruct           1550           E         P           S         Controctruct           1670           L         Y           S         Controctruct           1910           Prestorence           1910           Consegnation           1910           Consegnation           Consegnation           2030           Cantanana           Cantanana           Contractana           Contractana           Contractanana           Contractanana           Contanana           Contanananana           Contanananananananananananananananananana	TGTACAG           1440           ? Q T           GCGCAGACC           1550
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P AGCCAGCTGCCCCAGGGTCC 1450 1460 H L I T D T N ATGCTGATCACAGACAGCAGA 1570 1580 T T I H I P S ACAACCATTCACAGACAGCAGA 1690 1700 S N G H S H L TCCAACGGCACAGCGGAGA 1930 1940 CTCCATCATCAGAAAGGGGTA 2050 2060 TGCTTGCACGGGAGAAGGGGAGA 2170 2180 CTGTGCCTCCCAGGCCATTC 2290 2300 GGCTCTCCCAGGCGAGCAGCAGCA 2300 2400 CTGTGCCTCCCAGGCCATTC 2200 2300 GGCTCTCCCAGGCAGCAGCA 2300 2500 2560 CGTAGAACTGCGAAGCAGAA	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCGAACCA 1710 CACCCGCGCCACACCC 1950 GGCTCCTAGCGTGTC 2070 CACCCTGCGCCACACCC 2190 CACCCTGCGCCCACCC 2190 CACCCTGCGCCCACCC 2550 GCCTGAGCATGTC 2550 CCTCGCCCTCACCCTC 2550 CCTCGCCCTCACCCTCACCC 2550 CCTCGCCCCTCACCC 2570	GCACCLACCT 1360 M A Q L TGCCCACCT 1680 N I Q H A S L T CCACGCTCAC 1600 N I Q H ACATCCACCA 1720 M G V I ACGCTCTCAT 1840 ATGCTCCCTC 2080 TGCCACCTC 2080 TGCCACCTCC 2320 TGCCACGTCAC 2550 TGCTCACGAGAG 2660 TGCCAAGAGC	CACCACCCTC 1370 Q S P CACAGACCCCC 1490 P T K GCCCACCACC 1610 L Q P CCTCCACCACCC 1730 E T F CCCACCACCCT 1970 CCCCCACCACCC 2090 CCCTCTTOGTC 2210 CCCCCTCTCGTC 2210 CCCCCCCCCCCCC 2330 CCCCCCCCCCCCCCC 2690 CACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGGTCC 1380 H A L CACCOCCTTC 1500 Q V F CACGTCTTCJ 1520 A H R GCTCACCGC 1740 A H R GCTCACCGC 1740 GACCTCCCCC 2340 GGTTACCCTC 2440 GGTTACCCTC 2440 CCACTCCTACGC 2700 TCCACGCTACG 2700 TCCACCCTAC 2820	AGTITICCCA 1390 Y S H K ACACCCCACA ACACCCCACA 1510 T S D T CCCTCAGAACA 1510 L S T S TCCACCACCA 1630 L S T S TCCACCACACA 1750 Q H A S XAGATOCCTC 1870 CCTCCCACAGE 2350 CCTCCCACAGE 2350 CCTCCCACAGE 2350 CCTCCCACAGE 2350 CCTCCCCCCCACAGE 2350 CCTCCCCCCCACAGE 2350 CCTCCCCCCCCC 2710 GCCCGTCCCTCCCCCCC 2710 GCCCGTCCCTCCCCCCC 2710 GCCCGTCCCTCCCCCCC 2710 GCCCGTCCCTCCCCCCC 2710 GCCCGTCCCCCCCCCC 2710 GCCCGTCCCCCCCCC 2710 GCCCGTCCCCCCCCCCC 2710 GCCCGTCCCCCCCCCCC 2710 GCCCGTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCCCTCCCACA 1400 P E V CCCCTAGGTC 1520 Peptide C E A S ACAGECCTCC 1540 P T V TCCCCCCACA 1880 CCTCCTCCCAC 2000 CCCCCACACTC 2000 CCCCTCCCCCCAC 2000 CCCCCCCCCCCCCC 2000 ACCCCCCCCAC 2350 ACCCCCCCCCCCCCC 2350 CCCCCCCCCCCCCCC 2460 CCCCCCCCCCCCCCC 2720 CCCCCCCCCCCCCCCC 2720 CCCCCCCCCCCCCCCCC 2560 CCCCCCCCCCCCCCCC 2500 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCTTCCTAT 1410 A Q Y CCCCACTAC. 1530 S E P AGTGACCCT 1550 S S S TCCCTCCACC. 1770 * TTAACCATGCC 1890 TTCCCACAA. 2010 CGACCTCCTCC 2330 CCGCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTC 2490 CCACCTTCCTCCACC 2490 CCACCTTCCTCCACC 2490 CCACCTTCCTCCACC 2490 CCACCTTCCTCCACC 2490 CCACCTTCCTCCCACC 2490 CCACCTTCCTCCACCC 2490 CCACCTTCCTCCCACC 2490 CCACCTTCCTCCACC 2490 CCACCTTCCTCCACCCTTCCTC 2490 CCACCTTCCTCCACCCTTCCTC 2490 CCACCTTCCTCCCCTTCCTCCCCCCTTCCTCCCCCCCTTCCTCCCC	CACCACCCTC:         1420           T         H         T:           CACCCATACAN         1540           C         L         H           DGGCTTCATG.         1660           S         L         V           IAGCCTCGTGT:         1780           TCTATECCT:         2020           CCCTGCTGGT:         2140           GCACTACCAC:         2140           SGCTAGCCTMC         2500           CCAACTCCTCT         2620           CTATCTTCCT.         2740           TTAACTTTT         2740	Construction         Construction           1430         S         L         L         F         S         Construction         <	TGTACAG           1440           ? Q T           GCGCAGACC           1550
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGGGTCC 1450 1460 H L I T D T H ATGCTGATCGACAGACAGCAA 1570 1580 T T I H I P S ACAACCATTCACAGGCGCACAGC 1690 1700 S N G H S H L TCCAACGGCGCACAGCCACCT 1810 1820 AGGCCTGCACAGGGGGAA 2170 2180 GGCTTGC AAGGATGCCACC 2290 2300 GGCTTGCC TAGGCTTTCCCC 230 2540 1540 2420 TCAGGGCTAGGATGCCACGCAC 230 2540 1540 2420 TCAGGCTAGGCTAGCTTCCCC 2530 2540 1540 250 260 AGAGCCTGCCCCCGCCCCCAGTGGG 2770 2780 GGTGTCGTAGTAGCTTCCG	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTGG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCCATCCAACCA 1950 GGCGCCTCCAGCACCC 2190 CACCCTGCGCCCAGA 2310 TACTAGGCGCCTCT 2430 CTTGTTCCTTAGGC 2550 CCCCCGCCCAGCAATGTC 2790 TCCCCCCCCCCAGCACTGC	GCACCACCCC 1360 M A Q L TGCCCACGCT 1680 A S L T CCACCTCAC 1680 N I Q H ACATCCACCA 1720 N I Q H ACATCCACCA 1720 N I Q H ACATCCACCA 1720 N I Q H ACATCCACCAC 1840 ATCCTCCCTC 2080 TGCCACCCTC 2080 TGCCCCCCCCC 2200 TGCCCCCCCCC 2200 TGCCCCCCCCC 2200 TGCCCCCCCCCCC 2550 TGCACCACCACCCCC 2660 TGCACCACCCCCCC 2660 TGCACCACCCCCCC 2660 TGCACCACCCCCCC 2660 TGCACCACCCCCCC 2660 TGCACCACCCCCCCC 2660 TGCACCACCCCCCCCC 2660 TGCACCCCCCCCCCCCC 2660 TGCACCCCCCCCCCCCC 2660 TGCACCCCCCCCCCCCC 2660 TGCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCACCCTC 1370 Q S P CACACACCCC 1490 P I K GCCCACCACA 1610 L Q P CCTGCACCACC 1730 E T F CCACCACCT 1730 E T F CCACCACCT 1970 CACGCGCCCT 2330 CCACCACCT 2330 CCACCACCACCT 2330 CCACCACCACCT 2570 CACCCACCACCT 2570 CACCGCCCCCC 2650 CACCGCCCCCC 2650 CACCGCCCCCC 2650 CACCGCCCCCC 2650 CACCGCCCCCC 2650 CACCGCCCCCCCCC 2650 CACCGCCCCCCC 2650 CACCGCCCCCCCC 2650 CACCGCCCCCCCCC 2650 CACCGCCCCCCCCCCCCC 2650 CACCGCCCCCCCCCCCCC 2650 CACCGCCCCCCCCCCCCCCCCC 2650 CACCGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGCTCC 1380 H A L CACCOCCTCTC 1500 Q V F CACCCTTCC 1500 A H R CCTCACCCCC 1740 I S T ACTCCACCC 1740 I S T ACTCCACCCC 1740 I S T ACTCCACCCC 1740 I S T ACTCCACCCC 1740 I S T ACTCCACCCC 2100 TTCCACCCCCC 2100 TTCCACCCCCCC 2100 TTCCACCCCCCC 2200 TCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTTITECCA 1390 Y S H K ACACCCCACA ACACCCCACA ACACCCCCACA 1510 I S D I CCCTCCACAGACAC 1570 L S T S TCAGCACCCC 1570 Q H A S XAGATOCCTC 1870 Q H A S XAGATOCCTC 1870 Q H A S XAGATOCCTCACACA 1990 CCCTCCACACACACACA 2110 TCAAGGATCT 2210 CCTCCCACACCTAC 2550 CACCCCACCCCC 2550 CACCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	GCCACTCCAC           1400           P E V           GCCTCAAGTC           1520           Peptide C           E A S           AGAGCCTCC           1540           Peptide C           E A S           AGAGCCTCC           1540           P T V           TCCCCACAGTC           1880           GCTCATCGCC           2000           GCCCAATCTC           2120           CTCTCGAGCT           2480           CCTCTGCCC           2500           AAGCGCCTAGCC           2400           GCCCAAGCCT           2400           GCCCAAGCCT           2400           GCTCAAGCCT           2400           GCTCAAGCCT	CCTTCCTAT 1410 A Q Y CCCCACTAC 1530 S E P 1650 S S S TCCTCCAGCA 1770 * TAACCATCC 1890 TTCCCAGCA 2010 CGACCCCCTT 2130 CGACCCCCTT 2370 CCACCTTCCC 2250 CGACCTCCCCA 2370 CCACCTCCCCACA 2370 CCACCTTCCC 2550 CCACCTTCCCCACA 2550 CCACCTTCCCCACA 2550 CCACCTTCCCCACA 2550 CCACCTTCCCCACA 2550 CCACCTTCCCCACA 2550 CCACCTTCCCCACA 2550 CCACCTCCCCACA 2550 CCACCTTCCCCACA 2550 CCACCTCCCCACA 2550 CCACCTCCCCACA 2550 CCACCTCCCCACACA 2550 CCACCTCCCCACACA 2550 CCACCTCCCCACACACACACACACACACACACACACACA	CAGCAGCCTC:         1420           T         H         T:           CGCCATACAM:         1540           C         L         H           DGCCTTCATC:         1660           S         L         V           AGCCTGCTGT:         1780           TCTATGCCT:         2020           TCTATGCCTC:         1900           TCTATGCCT:         2020           CCTGCCTGGT:         2140           TCATCACACA:         2260           SCCTACCTTCT         2360           XCATCTTCCT         2520           TCTATCTTZ         240           TTACTTTCT         2380           ZT40         2740           TTAACTTTTZ         2860           CCTGTCGCCAC         2740	TCATECCCC         1430           1430         1430           S         L         L           S         CCTOCTTCC         1550           E         P         S           MCCCGTCGTC         1670           L         Y         Q           TGTACCACACA         1790           CCAGGAGCTC         1910           TCATECTCCAT         2030           TGGACCTACAA         2150           TGCAGCACTGTC         2390           CCAGGACTGTC         2630           CTCTGGCTGTC         2630           CTCTGGCGT         2750           GTAAACCCA         2870           GGTCAAGCCA         2870	TGTACAG           1440           ? Q T           GCGCAGACC           1560           : F A           :TCCAGACC           1560           : F A           :TCCAGACC           1680           : S D           :TTCCAC           1800           :GGCTCCC           1920           IGGTGCTC           2040           :GGCTCCC           :280           :GGCTCCC           :GGCTCCC           :GGCTCCC           :GGCTCCC      :GGCTCCC      :GGCTCCC
CAGGCACAGACCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGAGTCC 1450 1460 H L I T D T N ATGCTGATCACAGACAGCAA 1570 1580 T T I H I P S ACAACCATTCACAGCACCAC 1690 1700 S N C H S H L TCCAACGGCACAGCGCACAC 1930 1940 CTCCATCATCAGAAAGGGGAA 1930 1940 CTCGTGCACAGCAGCAGCAC 2170 2180 CTGTGCTAGCATTCCCC 2290 2300 TGTGTGCAACGTCAGCAC 2530 2540 TGTGTCAGCTCCCAGCCCAC 2530 2500 1640 1650 2420 1650 2650 250 260 2770 2780 CTGTGCTAGTAGCTTCCCGAT 2790 2900 1700 2180 CTGTGCTAGTAGCTTCCCGAT 2500 2500 1500 2500 1500 2500 2500 1700 2780 1700 2780 1000 2700 1000 2700	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCGAACCA 1710 L P S N GCTGCCATCCAACCC 1830 GGCGCCACACCCC 1950 GGCTCCATCCAACCC 2190 CACCCTGCGCCACACCC 2190 CACCCTGCGCCACACCC 2550 CCTCGCCCTCAGCCATGCTC 2670 TACCTAGGCATCTC 2910 CACCCTAGCATCTC 2910	GCACCAGCCT 1360 M A Q L TGCCCAGCT 1680 M A Q L TGCCCAGCT 1680 N I Q H ACATCCAGCA 1720 M G V I ACATCCAGCA 1720 H G V I ACGGTGTCAT 1860 N I Q H ACATCCAGCA 1720 GACCCAGCCT 2080 GACCCATCTG 2200 GACCCATCTG 2200 GACCCATCTG 2200 GACCCATCTG 2200 GACCCATCTG 2200 GACCCATCTG 2200 GACCCATCTG 2200 GACCCATCTG 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCT 2200 GACCCAGCCC 2200 GACCCAGCCCC 2200 GACCCAGCCCCCC 2400 GACCAGCGCCACCCC 2400 GACCAGCGCCACCCC 2560 GACCAGCCAGCCACCACCCCCC 2600 GACCAGCCACCACCACCCCC 2600 GACCAGCCACCACCACCACCACCCCCC 2600 GACCAGCCACCACCACCACCACCACCACCACCACCACCAC	CACCACCCTC CACCACCCTC 1370 Q S P CACACACCCC 1490 P T K GCCCACCACCA 1610 L Q P CCTCCACCCCC CACCACTT 1730 E T F CCACCACCTC 2090 CACCACCACCTC 2330 CCCCCCCCCC 2330 CCCCCCCCCCC 2330 CCCCCCCCCCC 2330 CCCCCCCCCCC 2330 CCCCCCCCCCC 2330 CCCCCCCCCCCCCC 2330 CCCCCCCCCCCCC 2330 CCCCCCCCCCCCCCCC 2330 CCCCCCCCCCCCCCCC 2330 CCCCCCCCCCCCCCCCC 2690 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGCTCC 1380 H A L CACCOCCTCTC 1500 Q V F CACCOCCTTC 1500 Q V F CACCOCCTCC 1500 A H R CCTCACCCCC 1540 CACCCCCCC 1540 CACCCCCCCC 1580 CACCCCCCCC 2340 CACTTCCTTC 2450 CCACCACCCTAC 2550 CCACCACCCTAC 2500 CCACCACCCTAC 2500 CCACCACCCTAC 2500 CCACCACCCTAC 2500 CCACCACCCTAC 2500 CCACCCCCCCACCA 2500 CCACCACCCTAC 2500 CCACCCCCCCACCAC 2500 CCACCCCCCCCACCAC 2500 CCACCCCCCCCCCCCCC 2500 CCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTTITECCA 1390 Y S H K ACACCCCACA ACACCCCACA ACACCCCACA 1510 I S D I CCCTCACAGACAC 1530 L S T S TCAACACCACA 1570 Q H A S XCATOCCCTC 1870 Q H A S XCATOCCCTC 1870 Q H A S XCATOCCCTC 1870 CCTCCCACAG 2350 CCTCCCACAG 2350 CCTCCACACG 2350 CCTCCACACG 2350 CCTCCACCCCC 2590 CAACCCACCCCCC 2590 CAACCACCCCCCCCCCCC 2590 CAACCACCACCACCAC 2590 CAACCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	GCCACTCCAC           1400           P E V           GCCTGAGTT           1520           Peptide C           E A S           AGAGCCTCC           1540           P T V           TCCCACAGTC           1760           S S Q           CTCCTCCACAGTC           1880           GCTGCTGCCC           2000           GCCACATCTC           2120           CTGCTCCAGCC           2500           AGGCGCTAGCC           2500           GGCCAAGCCT           2400           GGTCAAGCCTAGCC           2500           AAGGAAGCCA           2600           AAGGAAGCCAAGCCT           2840           GGTCAAGCCA           2840           GGTCAAGCCA           2840           GGTCAAGCCA           2840           CAGAGACCT           COCCAAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGCCAGC	CCTTCCTAT 1410 A Q Q Y CCCCACTAC 1530 S E P AGTGAGCCT 1650 S S S TCCCTCCAGCA 1770 * TTAACCATOC 1890 TTCCCAGCA 2010 GGACGCCCT 2130 CGAGCTCCTC 2490 CGACGTCCTCT 2490 CGACGTCCTCACA 2490 CGACCTCCCCACA 2490 CGACCTCCCCCACA 2490 CGCCCCTTCCCCACA 2490 CGCCCCCTTCCCACA 2490 CGCCCCCCTCCCACA 2730 CGCCCCCCCCCCCCAC 290 CGCCCCCCCCCCCCCCAC 290 CGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCACCCTC:         1420           T         H         T:           CACCATACAN         1540           C         L         H           DOCCTTCATC         1660           S         L         V           MACCTCCTCATC         1780           TCAACCCCTCC         1900           CTCATACCCTC         2020           CCCTCATGCTGGT:         2140           FACATCAACA:         2260           CCACTCCTCT         2500           CAACTTCTCCTT         2500           CTATCTTCCTT         2740           TTAACTTTTCT         2740           CTATCTCCCCCCC         2980           CCTOTCGCCAC         2980	TCCATEGECECC         1430           1430         1430           S         L         L           S         CCCTECTTCC         1550           E         P         S           MCCCGTCGTTC         1670           L         Y         Q           STGTACCAGAG         1790           CCAGGAGCTTC         1910           TCATACCAGAG         2030           CAGGAGCCTGC         2390           CAGCATAGAA         2510           CCCGGCGCT         2630           CCTCGGCCT         2750           GCTCAAGGC         2870           GCTCAAGGCC         3870           AAAACCAA         2870           GCTCAAGGC         2750           GCTCAAGGC         2870           GCTCAAGCAA         2870	TGTACAG           1440           ? Q T           GCGCAGACC           1550
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCCACGTGGCCCCAGAGTCC 1450 1460 M L I T D T N ATGCTGATCACAGCAGAGTCC 1570 1580 T T I H I P S ACAACCATTCACATCCCCAG 1690 1700 S N G H S H L TCCAACGGGCACAGCCACCT 1810 1820 CAGCCCTGCACAGGGGGAA 1930 1940 CTCCATCATCAGAGAGGGGA 2170 2180 CIGTIGG ACAGCAGGGGCA 2170 2180 CGTCTCC TAGGCTTTGCCCC 230 2500 TGTAGA ACTCC CAACACAAC 2530 2540 TGTAGA ACTCC CAACACAAC 2530 2540 TGTAGA ACTCCCACCACACAC 2770 2780 GGTGTGCTACTGTACCTCATCG	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGACCCTTG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCATCCAACG 1830 GGGGGCCACACCC 1950 GGCTCTGAGGTGTC 2070 AGGCCTTGTATTGC 2190 CACCTGCGCCCAAC 2310 TACTAAGGCT 2550 GCCTGAGCAATGTC 2670 CTCCCCCCCCACCC 2790 CACCTGCGCTCAGCC 2910 ACTCCAACTGCGTTG	GCAGCAGCCT 1360 M A Q L TGCCCAGCC 1480 A S L T CCAGCTCAC 1680 N I Q H ACATCCACCA 1720 H G V I ACGGTGTCAT 1840 NTCCTCATCC 2080 TCCTCCAGCA 2320 TCCTCCAGCA 2320 TCCTCCAGCA 2560 TACTCGAAGAGC 2580 TCCTCGAAGAGC 2580 TCCTCGAAGAGC 2590 TACTCGAAGAGC 2590 TACTCGAAGAGCAC 2590 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGAGCAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC 2500 TACTCGAAGACAC	CACCACCCTC 1370 Q S P CACACACCCAC 1490 P T K CCCACCCAC 1610 L Q P CCTCCACCAC 1610 E T F CCGACACCTTTT 1330 E T F CCGACACCTTTT 1300 E T F CCGACCCTCTTCT 1350 CACCCCCAC 2100 CCCCCACCT 2300 CCCCCACCT 2310 CCCCCACCT 2310 CCCCCACCT 2300 CCCCCACCT 2310 CCCCCACCT 2310 CCCCCACCT 2310 CCCCCACCT 2310 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCACCT 2570 CCCCACCT 2570 CCCCACCT 2570 CCCCACCT 2570 CCCCACCT 2570 CCCCACCT 2570 CCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCACCT 2570 CCCCCCCCCC 2570 CCCCCCCCC 2570 CCCCCCCCC 2570 CCCCCCCCC 2570 CCCCCCCCCCCC 2570 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGCTCC 1380 H A L CACCOCCTCT 1500 Q V F CACGTCTTC 1500 Q V F CACGTCTCC 1500 Q V F CACGTCTCC 1500 I S T ATCTCCACCC 1740 I S T ATCTCCACCC 1580 CACCTCCCC 2100 CCCCCACCACA 2100 CCCCCCACCACA 2580 CCACCACCACA 2580 CCACCACCACA 2580 CCACCACCACACA 2580 CCACCACCACACA 2580 CCACCACCACACA 2580 CCACCACCACACA 2580 CCACCACCACACA 2580 CCACCACCACACA 2580 CCACCACCCACACA 2580 CCACCACCACACA 2580 CCACCACCACACA 2580 CCACCACCACACACA 2580 CCACCACCACACACA 2580 CCACCACCACACACACA 2580 CCACCACCACACACACA 2580 CCACCACCACACACACA 2580 CCACCACCACACACACA 2580 CCACCACCACACACACACACACACA 2580 CCACCACCCACACACACACACACACACACACACACACA	AGTTITECCA 1390 Y S H K XCACCCCACA XCACCCCACA 1510 T S D T CCCTCAGACACA 1510 T S D T CCCTCAGACACA 1510 CCCTCACACACACA 150 Q H A S XAGATOCCCTC 1870 Q H A S XAGATOCCCCCACAG 1990 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCACAG 2150 CCCCCCCCCCCCCCCC 2110 CCCCCCCCCCCCCCCCCCC 210 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	GCCACTCCAC           1400           P E V           GCCTCAGGTT           1520           Peptide C           Fasting           1640           Paptide C           Intervention           1640           CCCCCACGTC           1640           CCCCCCCCCAC           1880           GCTCCTCCCCC           1880           GCTCCTCCCCC           1800           GCCCCATCGCC           2160           GCCCCTCGCCC           22600           ALGEGCCTAGC           22600           ALGEGCCTAGCC           22600           ALGEGCCTAGC           22600           ALGEGCCTAGC           22600           ALGEGCCTAGC           22600           ALGEGCCTAGC           2400           GCCCATCGCCT           2400           ALGEGCCTAGC           2560           ALGEGCATAGCC           2560           ALGEGCAACCCT           2560           ALGEGCAACCCT           3080	CCTCCTAT 1410 A Q Y CCCCACTAC 1530 CCCCACTAC 1530 CCCACCT 1530 CCCACCTA 1650 S S P 1650 CCCACCACA 1890 CCCACCACA 2130 CCCACCACAC 2370 CCCACCTCACA 2610 CCCACCCCACA 2730 CCCACCTCACACAC 2970 CCCACCTCACACACACACACACACACACACACACACACA	CAGCAGCCTC: 1420 T H T :: CGCCATACAM: 1540 C L H :: CGCCATCATCA 1540 C L H :: CGCCATCATCA 1540 C L H :: CGCCTCCTCATC 1660 C L H :: CGCCTCCTCATCATCA 1780 C L H :: CGCCTCCTCATCATCATCATCATCATCATCATCATCATCATC	TCATECCCCC 1430 S L L F S L L F SCCTCCTTCC 1550 E P S S NOCCGTCGTC 1570 L Y Q S TCATCGTGCC 1910 TCATCGTGCCA 2030 TCGTCCCAGA 2030 TCGCTGCCCC 2390 TCGCTGCCCC 2430 TCTTGGCCT 2750 GCTACGCCC 2870 GCTACGCCC 2870 GCTACCCCCC 2870 GCTCCCCCCC 2990 ATTCCCCCCC	TGTACAG           1440           P           Q           CCCAGACC           CSCAGACC           1560           S           FTCCAGCC           1680           S           FTTCTACC           1800           SGGCTCCC           1920           IGCTGCTC           2040           GGCTGCTC           2280           GGACGCG           2400           GGCTCCCC           2520           GGCTCCC           2400           GGCTCCC           2500           GGCTCCC
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCAGCTGCCCCAGAGTCC 1450 1460 H L I T D T N ATGCTGATCGACAGACAGCAG 1570 1580 T T I H I P S ACAACCATTCACAGCACAGCA 1690 1700 S N G H S H L TCCAACGGCGCACAGCCACCT 1810 1820 AGACCTGCACGGCACAGCCACCT 1810 1820 AGACCTGCACGGCACAGCCAGCG 2170 2180 GCTGTGCTGCAACGACGGGA 2170 2180 GGTGTGCTGCCAGCCCTTCCC 2330 2540 IGTAGAACTGCCAAGACAGGGCA 2170 2180 GGTGTCTACGCAAGCCCTCT 2290 2300 GGTGTCCTAGCCTTCCCCAGCCCACT 2330 2540 IGTAGAACTGCCAAGCAGCGCA 2770 2780 IGTAGAACTGCCAAGCAGCG 2770 2780 IGTAGAACTGCCAAGCAAGCG 2770 2780 ITTACACATCTGCAAGCAACC 10 3020 GCTGCGCTGCCTGCCTGCGC	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCGCCGTCCAGCGCCAA 2010 GGGCCGTCGAGCAGCCC 2070 AGGCCTTGAGCGCTCA 2130 CTTCCTCAGCCCCCAA 2310 TACTAGGCGCTCT 2430 CTTCCTCCCCCCCCAGCA 2790 TCCCCAGCATGGCG 2910 ACTCGAATGAGGT 3030 CCGGCGTCGGGCGG	GCACCACCCCC 1360 M A Q L TGCCCACCCC 1680 A S L T CCACCCTCAC 1690 N I Q H ACATCCACCACA 1720 H G V I ACATCCACCACA 1720 H G V I ACATCCACCACA 1840 TGCCACCCCCC 2000 TGCACCCCCCC 2000 TGCACCCCCCC 2200 TGCCACCCCCC 2200 TGCCACCCCCC 2200 TGCCACCCCCC 2200 TGCCACCCCCC 2550 TGCCCACGCCC 2500 TGCCACACCCCC 2500 TGCCCCCCCCCCC 2500 TGCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCCCC 2500 TGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCACCCTC CACCACCCTC 1370 Q S P CACACACCCC 1490 P I K CCCCACCAAC 1610 L Q P CCTCCACCACC 1730 E T F CCCACCACCT 1730 E T F CCACCACCTC 1730 CACCCACCACCT 1730 CACCACCACCT 1730 CACCACCACCT 2570 CACCACCACCC 2510 CACCACCACCC 2510 CACCACCACCC 2510 CACCACCACCCCCC 2510 CACCACCACCCCCC 2510 CACCACCACCCCCCC 2510 CACCACCACCCCCCCC 2510 CACCACCACCCCCCC 2510 CACCACCCCCCCCCC 2510 CACCACCCCCCCCCCC 2510 CACCACCCCCCCCCCCC 2510 CACCACCACCCCCCCCC 2510 CACCACCACCCCCCCCCCCCCCC 2610 CACCACCACCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CAGCCOGGTCC 1380 H A L CACGCCTTC 1500 Q Y F CAGGTCTTC 1500 Q Y F CAGGTCTTC 1500 A H R GCTCACCGC 1740 I S T ATCTCCAGC 1740 I S T ATCTCCAGCC 2100 TTCCAGGCCT 2100 TTCCAGGCCT 2340 CCATCATCTTAG 2580 CCACCAGGA 2540 CCACCAGGA 2540 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCACCAGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGCACCACCACCACCACCACACACACACACACAC	AGTTITECCA 1390 Y S H K KACACCCACA KACACCCACA 1510 I S D I CCCTCGAGACAC 1530 L S T S TCAGCACCAC 1750 Q H A S XAGATOCCTC 1870 Q H A S XAGATOCCTC 1870 Q H A S XAGATOCCTCACAC 1990 CCCTCCACACG 2130 TCAGCACCTACAC 2550 CCTCCACACGTA 2550 CCCCCACACCTACAC 2550 CCCCCCACACCTACAC 2550 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCACCCCCCC 250 CCCCCCACACCACCACCCAC 250 CCCCCCACACCACCACCACCAC 250 CCCCCCACACCACCACCACCAC 250 CCCCCCACACCACCACCAC 250 CCCCCCACACCACCACCACCACCACCACCACCAC 250 CCCCCCACACCACCACCACCACCACCACCACCACCACCA	CCCCTCCACC 1400 P E V CCCTCACGTT 1520 Peptide C E A S ACAGCCCTC 1540 C E T V TCCCACACTC 1540 S S Q CTCCTCCACC 1880 CCCCTCGCC 2000 CCCCTCGCC 2120 CTCCTCGACCT 2240 ACCTCCCCCA 2120 CCCCTGCCC 2350 CCCCTGCCC 2350 CCCCTGCCC 2360 CCCCTGCCC 240 ACCTCCCCCACC 250 ACCCCCCCACC 250 ACCCCCCCCACC 250 ACCCCCCCCCCC 230 ACCCCCCCCCCCC 230 ACCCCCCCCCCCCC 230 ACCCCCCCCCCCCCC 230 ACCCCCCCCCCCCCCCC 230 ACCCCCCCCCCCCCCCCC 230 ACCCCCCCCCCCCCCCCCCCCC 240 ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCTTCCTAT 1410 A Q Y CCCCACTAC 1530 B S E P 1650 S S S TCCTCCACAC 1770 * TACCACGCT 1890 TTCCCACACA 2010 CGACGCCCCACACA 2010 CGACGCCCCTC 2130 CCCACTCCCCACACA 2730 CCCACTCCCCACACA 2730 CCCACTCCCCACACA 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACACACACACACACACACACACACACA	CAGCAGCCTC:         1420           T         H         T:           CGCATACAM:         1540           G         L         H           DGCCTTCATG:         1660           S         L         V           AGCCTGCTGT:         1780           TGACTGCCTC:         1900           TGACTGCCTCG:         2020           CCTGCTGGT:         2140           TCATCTACACA:         2250           SCCTACCTGTCT         2380           CACTCACACA:         2620           CATCTTCCTCT         2480           CTTACCTTTCTA         2620           CTATTCTCAA         2980           CTATCTCCA:         23100	TCATECCCCC 1430 S L L F CCTCCTTCC 1550 E P S S CCCCCTCTC 1570 E P S S CCCCCCCC 1790 CCAGGACCTCC 1910 CCAGGACCTCC 2030 CCCCCCCCCC 2030 CCCCCCCCCCCCCCCCCC 2390 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	TGTACAG           1440           ? Q T           SCGCAGACC           1560           3 P A           TTCCAGACC           1560           5 D           TTCCAGC           1800           SGCCTCCC           1920           GGCTGCTC           2040           GGCTGCTC           2240           GGCGCTCC           2240           GGCGCTCC           22400           GGCAGCAG           2400           GGCTGCCA           2520           TAGAGACG           26400           GCCTCCAG           26400           GCTGCTA           2640           GCTGCTA           3240
CAGGCACAGAGCGTGCCAGT 1330 1340 S H V A Q S P ACCAGCTGCCCCAGAGTCC 1450 1460 H L I T D T N ATGCTGATCGACAGACAGCAG 1570 1580 T T I H I P S ACAACCATTCACAGCACAGCA 1690 1700 S N G H S H L TCCAACGGCGCAGAGCCACCT 1810 1820 AGACCTGCACGGCGAGAGCAGCG 1930 1940 CTCCATCGTACGGCGAGAGCGGGA 2170 2180 GGCTGCCTGCCAGGCGCAGC 2330 2540 IGTAGAACTGCGAGAGCAGCG 2770 2580 IGTAGAACTGCGAGAGCAGC 2770 2580 IGTAGAACTGCGAGAGCAGC 2770 2580 IGTAGAACTGCGAGAGCAGC 2770 2580 IGTAGAACTGCGAGAGCAGC 2770 2780 ITTACACTGTGCTTGTCA 3010 3020 GCTGGCGTGACAAGCGTTTGAC 3140 ITTACACGCCGGTTGCACGCTTCAC 1310 3140 ITTACACCCCGGTTCAC	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTCG 1590 Q D P S CCAGGACCCGTCAA 1710 L P S N GCTGCCATCCAACCC 1950 GGGCCTGCAGCACGCC 2070 AGGCCTTGAGGGCCAGACCC 2190 CACCCTGCGCCCAGA 2310 TACTAGGCGCTCT 2430 CTTGCTCACGCCCCCAGA 2310 CTCCCCCCCCCCAGCC 2790 TACCTAGGCGTCTGCC 2910 ACCCGGCACTCGACGCTCG 3030 CCGGGGCTGGGGTGG 3150 FCCCCGGCACTCAA 3720	GCACCACCCC 1360 M A Q L TGCCCAGCC 1480 A S L T CCAGCCTCAC 1600 N I Q H ACATCCACCACA 1720 H G V I ACATCCACCACA 1720 H G V I ACGCGTGTCAT 1840 TGCCACCCTC 2080 TGCACCCCTC 2080 TGCACCCCCCC 2000 TGCACCCCCCC 2000 TGCACCCCCCC 2000 TGCCACCCCCC 2000 TGCCACCCCCC 2000 TGCCCCCCCCC 2000 TGCCCCCCCCC 2000 TGCCCCCCCCC 2000 TGCCCCCCCCC 2000 TGCCCCCCCCCCC 2000 TGCCCCCCCCCCC 2000 TGCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCACCCTC CACCACCCTC 1370 Q S P CACACACCCC 1490 P I K GCCCACCAAC 1610 L Q P CCTCCACCACC 1730 E T F CCACCACCTT 1730 E T F CCACCACCTT 1730 CACCCACCACCT 2090 CACCACACCT 2090 CACCACACCT 2090 CACCACACCT 2090 CACCACACCT 2310 CCTCTTCGCAC 2310 CCTCTCACACC 2570 CACCACCACCT 2510 CACCCACCACCT 2510 CCTCCACCACCT 2510 CCTCCACCACCT 2510 CCTCCACCACCT 2510 CCTCCACCCCCCC 2510 CCTCCACCACCT 2510 CCTCCACCCCCCCCC 2510 CCTCCACCACCT 2510 CCTCCACCCCCCCCCCC 2510 CCTCCACCACCTCACCCCCCC 2510 CCTCCACCACCTCACCACCT 2510 CCTCCACCCCCCCCCCCCCC 2510 CCTCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CAGCCOGGTCC 1380 H A L CACGCCTTC 1500 Q Y F CAGGTCTTC 1500 Q Y F CAGGTCTTC 1500 A H R GCTCACCGC 1740 I S T ATCTCCAGC 1740 I S T ATCTCCAGCC 2100 TTCCAGGCCT 2100 TTCCAGGCCT 2340 CCATCATCTTAG 2580 CCACCAGGA 2540 CCACCAGGA 2540 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCACCAGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGGA 2500 CCACCAGCACCACCACCACCACCACACACACACACACAC	AGTTITECCA 1390 Y S H K KACACCCACA KACACCCACA 1510 I S D I CCCTCGAGACAC 1530 L S T S TCAGCACCAC 1750 Q H A S XAGATOCCTC 1870 Q H A S XAGATOCCTC 1870 Q H A S XAGATOCCTCACAC 1990 CCCTCCACACG 2130 TCAGCACCTACAC 2550 CCTCCACACGTA 2550 CCCCCACACCTACAC 2550 CCCCCCACAC 2550 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCTACAC 250 CCCCCCACACCACCAC 250 CCCCCCACACCACCACCAC 250 CCCCCCACACCACCACCAC 250 CCCCCCACACCACCACCACCACCAC 250 CCCCCCACACCACCACCACCACCACCACCACCACCACCA	CCCCTCCACC 1400 P E V CCCTCACGTT 1520 Peptide C E A S ACAGCCCTC 1540 C E T V TCCCACACTC 1540 S S Q CTCCTCCACC 1880 CCCCTCGCC 2000 CCCCTCGCC 2120 CTCCTCGACCT 2240 ACCTCCCCCA 2120 CCCCTGCCC 2350 CCCCTGCCC 2350 CCCCTGCCC 2360 CCCCTGCCC 240 ACCTCCCCCACC 250 ACCCCCCCACC 250 ACCCCCCCCACC 250 ACCCCCCCCCCC 230 ACCCCCCCCCCCC 230 ACCCCCCCCCCCCC 230 ACCCCCCCCCCCCCC 230 ACCCCCCCCCCCCCCCC 230 ACCCCCCCCCCCCCCCCC 230 ACCCCCCCCCCCCCCCCCCCCC 240 ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCTTCCTAT 1410 A Q Y CCCCACTAC 1530 B S E P 1650 S S S TCCTCCACAC 1770 * TACCACGCT 1890 TTCCCACACA 2010 CGACGCCCCACACA 2010 CGACGCCCCTC 2130 CCCACTCCCCACACA 2730 CCCACTCCCCACACA 2730 CCCACTCCCCACACA 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACAC 2730 CCCACTCCCCACACACACACACACACACACACACACACA	CAGCAGCCTC:         1420           T         H         T:           CGCATACAM:         1540           G         L         H           DGCCTTCATG:         1660           S         L         V           AGCCTGCTGT:         1780           TGACTGCCTC:         1900           TGACTGCCTCG:         2020           CCTGCTGGT:         2140           TCATCTACACA:         2250           SCCTACCTGTCT         2380           CACTCACACA:         2620           CATCTTCCTCT         2480           CTTACCTTTCTA         2620           CTATTCTCAA         2980           CTATCTCCA:         23100	TCATECCCCC 1430 S L L F CCTCCTTCC 1550 E P S S CCCCCTCTC 1570 E P S S CCCCCCCC 1790 CCAGGACCTCC 1910 CCAGGACCTCC 2030 CCCCCCCCCC 2030 CCCCCCCCCCCCCCCCCC 2390 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	TGTACAG           1440           ? Q T           SCGCAGACC           1560           3 P A           TTCCAGACC           1560           5 D           TTCCAGC           1800           SGCCTCCC           1920           GGCTGCTC           2040           GGCTGCTC           2240           GGCGCTCC           2240           GGCGCTCC           22400           GGCAGCAG           2400           GGCTGCCA           2520           TAGAGACG           26400           GCCTCCAG           26400           GCTGCTA           2640           GCTGCTA           3240
CAGCCACAGACCTGCCACA 1330 1340 S H V A Q S P AGCCACGTCGCCCCAGAGTCC 1450 1460 H L I T D T N ATGCTGATCACAGACACACA 1570 1580 T T I H I P S ACAACCATTCACAGACACCACA 1690 1700 S N G H S H L TCCAACGGGCACACCCACCT 1810 1820 CTCCATCATCAGAAAGGGGAT 2050 2060 TGCTACAGACAGAGGGAGA 2170 2180 CTGCTCCCCACGGCCATCC 22410 2420 GGCTCTCCTAGGCAGAGGGGACA 2530 2560 CGGTGCCCCCCAGTGGC 2770 2280 CGGTGTAGCAGACGCCTTCCCC 2410 2420 CGGCGCTCCCTAGCCCTCCGC 2530 2560 AGAGCCTGCCCCCGGCCATC 2590 2300 CGTGTAGCGCACGCGCTATT 2890 2900 TTACACATCTTTTGTTCACA 3010 3020 GCTGGGGGTGACAAAGCCGTAT 3130 3140	CATCAACAGCATGG 1350 F M A T CTTCATGGCAACCA 1470 L S T L CCTCAGGCACCCTTG 1590 Q D P S CCAGGACCCGTGCAACCA 1710 L P S N GCTGCCATCCAACCC 1830 GGCTGCGATCCAACCC 1950 GGCTCGCACCACCCC 2190 CACCCTGCGCCAACCC 2190 CACCCTGCGCCCAGAC 2310 CTTCCCAACTACCCT 2430 CCTGCCGCCCACCCC 2790 TACCTAGCCATGCC 2790 TACCTAGCCATGCC 2790 TACCTAGCCATGCCACCC 2310 CCCCGGCCTCGGCTG	GCACCACCCC 1360 M A Q L TGCCCACCCT 1680 A S L T CCCACCTCAC 1690 N I Q H ACATCCACCACA 1720 H G V I ACATCCACCACA 1720 H G V I ACACCACCCCC 2000 TGCACCCCTC 2000 TGCACCCCTC 2000 TGCACCCCCCC 2000 TGCACCCCCCC 2000 TGCCCCCCCCC 2000 TGCCCCCCCCC 2000 TGCCCCCCCCCCC 2000 TGCCCCCCCCCC 2000 TGCCCCCCCCCCC 2000 TGCCCCCCCCCCC 2000 TGCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCCC 2000 TGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCACCCTC CACCACCCTC 1370 Q S P CACACACCCC 1490 P I K GCCCACCAAC 1610 L Q P CCTCCACCACC 1730 E T F CCACCACCTT 1730 E T F CCACCACCTT 1730 CACCCACCACCT 2090 CACCACACCT 2090 CACCACACCT 2090 CACCACACCT 2090 CACCACACCT 2310 CCTCTTCGCAC 2310 CCTCTCACACC 2570 CACCACCACCT 2510 CACCCACCACCT 2510 CCTCCACCACCT 2510 CCTCCACCACCT 2510 CCTCCACCACCT 2510 CCTCCACCCCCCC 2510 CCTCCACCACCT 2510 CCTCCACCCCCCCCC 2510 CCTCCACCACCT 2510 CCTCCACCCCCCCCCCC 2510 CCTCCACCACCTCACCCCCCC 2510 CCTCCACCACCTCACCACCT 2510 CCTCCACCCCCCCCCCCCCC 2510 CCTCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CACCCOGCTCC 1380 H A L CACCCCCTTC 1500 Q V F CACCCCCTTC 1520 A H R CCTCACCCCC 1540 I S T ATCTCCACCC 1540 CCTCAACCC 1540 CCTCAACCC 2100 CTTCCACCC 2340 CACTTCCTTC 2460 CACCTCCCC 2340 CACTTCCTCC 2550 CCACCCCCAC 2500 CCACCCCCCAC 2500 CCACCCCCCAC 2500 CCACCCCCCC 2400 CACCTCCCCC 2500 CCACCCCCCC 2500 CCACCCCCCC 2500 CCACCCCCCC 2500 CCACCCCCCCC 2500 CCACCCCCCCC 2500 CCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	AGTITICCCA 1390 Y S H K ACAGCCLACAA 1510 Y S H K ACAGCCCACA 1510 Y S H K ACAGCCCACA 1630 L S T S TCCACCACAC 1630 L S T S TCCACCACAC 1630 L S T S TCCACCACAC 1630 CTCACACACCACAC 1950 GCCGGAGACACAC 2350 CCTCACCACACACACAC 2350 CCTCACCACACACACACACACACACACACACACACACAC	GCCACTCCAC           1400           P E V           GCCTCACGTC           SCTCACGTC           1520           Peptide C           Fast           AGAGECCTCC           1640           TCCCACACTC           1760           S S Q           CTCCTCCCCC           1880           GCTCCTGCCC           2000           GCCCCAATCTC           2160           CTCTCCACGC           2360           CCTCTGCCTC           2400           AAGEGCCTACC           22600           AAGEGCCTACC           2360           CCTCTGCACT           2400           CCTCTGCCTGC           2500           AAGEGCCTACCAC           2200           GGCAACCT           2400           CATCTCACAC           2500           AAGEGCCTACC           2960           AAGAACCCA           3080           TGACACACGACA           3200	CCTCCTATC 1410 A Q Y CCCCACTAC 1530 CCCCACTAC 1530 S E P 1650 S S F TACTGACCATGC 1770 * TAACCATGC 1890 TTCCCACACA 1890 TTCCCACACATGC 2130 GACCGCCCTTCG 2370 GACCGCCCTTCG 2370 GACCGCCCTTCG 2370 GACCGCCCTTCG 2370 GACCGCCCTCCA 2370 GACCGCCCTCCA 2370 GACCGCCCTCCA 2370 CCACTTCACCACA 2370 CCACTTCACCACAC 2370 CCACTTCACCACAC 2370 CCACTTCACCACAC 2370 CCACTTCACCACAC 2370 CCACTTCACCACACAC 2300 CCACTTCACCACACACACACACACACACACACACACACA	CACCACCTC: 1420 T H T :: CCCCATACAM: 1540 C L E :: CCCCATACAM: 1540 C L E :: CCCCATCATC 1540 C L E :: CCCCTCATC 1540 C :: CCCCTCATCATC 1780 C :: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	TCATECCCCC 1430 S L L F CCCCCTCC 1550 E P S S CCCCGTCGTC 1550 L Y Q S CCCGGTCGTC 1910 CCAGGAGCTC 1910 CCAGGAGCTC 2030 CCCGGTCGTC 2030 CCCGGTGTC 2390 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGGTGTC 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCGCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCGCCCCA 2500 CCCCGCCCCA 2500 CCCCGCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCCA 2500 CCCCCCCCA 2500 CCCCCCCCA 2500 CCCCCCCCA 2500 CCCCCCCCA 2500 CCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCA 2500 CCCCCCCCCCA 2500 CCCCCCCCCCA 2500 CCCCCCCCCCA 2500 CCCCCCCCCCCCCCCCCCCA 2500 CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	TGTACAG           1440           P           Q           CGCAGACC           CSCAGACC           1560           S           FTCCACC           1680           S           TTCTCACC           1800           SGGCTCCC           1920           MACTTAA           2040           GGCTCCTC           2280           GGACACC           2400           GGCTCCCT           2520           XAACTTAA           2500           GGTCGCAC           2760           GTACAA           3000           GTCTCAA           3000           GTCTCAA           3120           CTCGTAA           3120

Figure 4: HNF1 cDNA sequence. Amino acids are given in single letter code above the nucleotide sequence. The numbering adopted, for both the nucleotide and the amino acid sequences, is initiated at the first codon of the HNF1 open reading frame (see text). Underline bold regions indicate tryptic peptides that were sequenced. \* indicates the first nonsense codon.

and  $0.1 \times SSC$ , 0.25% SDS. Alternatively, an HNF1 antisense RNA probe (nt 3158-3397) was prepared by *in vitro* transcription with T7 RNA polymerase of a BamHI digested CD13 clone (Figure 3C). Membranes were hybridized in 50% formamide,  $5 \times SSC$ ,  $5 \times$  Denhardt at  $52^{\circ}C$  for 16 h. and washed at a final stringency of  $0.1 \times SSC$  at  $74^{\circ}C$  for 1 h.

### RESULTS

# Native HNF1 forms dimers in the presence or absence of its DNA-target

HNF1 protein was purified from rat liver in a three step procedure: preparation of nuclear extracts, chromatography on heparin

Ultrogel and specific DNA-affinity column as described in Materials and Methods. The peak of HNF1 DNA-binding activity produced two complexes in gel retardation assay (Figure 1, lane 1: arrows U and L) and displayed a major 87-93 kDa protein, a 72 kDa polypeptide and several minor polypeptides (Figure 1, lane 7). To determine which of these polypeptides directly interacted with DNA, we performed UV-crosslinking experiments, in which <sup>32</sup>P-labelled oligonucleotide containing BrdU residues was covalently linked to the proteins in the retardation gel, before the upper and lower band were excised and run on a SDS-polyacrylamide gel. The upper complex gave rise to a roughly100 kDa band corresponding to the 87-93 kDa protein linked to the 15 kDa oligonucleotide, which is in agreement with the molucular weight previously reported for HNF1 (27, 11, 12, 17). The heterogeneity of the 87-93 kDa band (lane 7) is probably caused by the glycosylation of HNF1 (see below). The lower complex gave rise to the same 100 kDa band in addition to faster migrating species, likely to be proteolyzed chains since the yield of this lower retarded band increased during the purification procedure. The fact that the HNF1 binding site was almost palindromic suggested a hypothesis in which the upper complex would contain two intact 87-93 kDa polypeptide chains and the lower band one intact and one partially degraded chains; the profile obtained in lanes 2 and 3 is compatible with this hypothesis. Analysis of the UV-crosslinked DNA-protein complexes in nonreducing conditions gave slower migrating bands, suggesting that HNF1 might either associate with another polypeptide chain or form dimers covalently linked by S-S bonds (Figure 1, lanes 4 and 5). This interaction appeared to be specific since the same patterns were observed whether crude liver nuclear extracts or various purified fractions were used (data not shown). Further support for potential dimerization came from the observation that a major fraction of the 87-93kDa HNF1 band ran as a 180 kDa species under nonreducing conditions (Figure 1, lane 6). In addition, this suggested that the protein might preexist as a dimer in the absence of DNA; neither addition of specific or unrelated DNA nor irradiation with UVlight increased the yield of dimerization observed in lane 6 (not shown). These observations were further strengthened using cloned HNF1 (see below). Gel retardation assays performed in the presence of different reducing agents, indicated that S-S bond formation was not required for DNA-binding in vitro (data not shown), nevertheless, the factor might still need to dimerize for this purpose.

In order to obtain protein sequence data for the 87-93 kDa chain, 150 pmoles were further purified by SDS-PAGE, since no other polypeptide was visible on two-dimensional gels at the same molecular weight level (not shown), then transferred to a PVDF membrane and digested *in situ* with trypsin. The tryptic peptides were fractionated on reversed phase HPLC and submitted to amino acid sequence analysis. The three sequences obtained are listed in Figure 2. Peptides A and B were used to derive redundant oligonucleotides for direct screening of cDNA libraries from rat liver or hepatoma cell lines. After several unsuccessful attempts, we turned to a PCR-derived approach to get better probes for HNF1.

### Use of PCR to generate a unique probe for HNF1

From the longest peptide (CD), four degenerate primers, two of the coding strand (C1 and C3) and two of the noncoding strand (D2 and D4), were designed to reduce the high degeneracy due to serine and leucine residues (see Figure 2). These oligonucleotides were used in an enzymatic amplification on Fao cells cDNA, followed by cloning and sequencing of the expected 47 bp PCR-amplified fragment. The nucleotides coding for the alanine and the two central serine residues of the CD peptide were thus unambiguously determined and a 41-mer nondegenerated probe was synthesized according to this sequence data (CD41 in Figure 2).

Since the C and D primers were highly degenerate (respectively 512 and 1024 mixed sequences) and PCR is a primer-limiting procedure, only about 1% of the 47 bp amplified fragments that were visible on a gel might display the exact HNF1 sequence. The remaining must contain some mismatches mainly in the ends of the fragments, since elongation by Taq-DNA-polymerase requires a better annealing at the 3' end of the primers. The nucleotides that were actually wrong in the single PCR fragment that we sequenced as compared to the final cDNA sequence are indicated in Figure 2.

## Seven independent HNF1 cDNA clones display three polyadenylation sites and a single open reading frame

With the CD41 probe, we screened  $10^6$  clones of a  $\lambda$ -gt10 cDNA library prepared from rat hepatoma Fao cells (28), treated with cycloheximide to increase the representation of potentially unstable transcripts (29). We first isolated 3 partial cDNA clones (CD13, CD16, CD39); one of them (CD13) included the two others. A 45 bp probe (A15, Figure 2) from the most 5' region of this clone was synthesized and used to screen the same library. Seven new clones (A-clones and CD26) were isolated that were also positive with CD41. The structure of the ten clones including three pairs of identical ones is described schematically in Figure 3C.

The total length of the combined nucleotide sequence (CD26 and the 3' end of CD39 without the oligoA tail) was 3205 bases. The cDNA clones could be ordered in three groups (I, II, III, see figure 3C) according to the position of their polyA tails relative to the combined sequence. When these clones were used to probe Northern blots with rat liver polyA RNA, a minor and a major species of 3.2 and 3.6 kb respectively were detected (figure 3A). Hence, our longest clone could not cover the entire length of the 3.6 kb mRNA and was probably missing a few hundred base pairs.

In addition, using a probe derived from the most 3' sequence of clones of the group III, only the 3.6 kb species was detected, thus demonstrating that more than one polyadenylation site were actually used in vivo (Figure 3B). Examination of the sequences upstream of the polyA tails found in the three groups of clones revealed three possible alternative polyadenylation signals. The four independent clones of group I utilize the AGTAAA sequence at position 2864 (numbering of the figure 4, see below) with a polyA chain 14 to 16 nt downstream; the single clone of group II might use the AATGAG sequence at position 3027 with a polyA chain 17 nt downstream; finally, the two clones of group III utilize the AATAAA sequence at position 3361 with a polyA chain 31 nt downstream (30). The Northern blot analysis suggested that, in rat liver, the most 3' site is used predominantly in rat liver. The situation was different in rat hepatoma cell lines (S.C., unpublished observations).

Conceptual translation of the cDNA sequence revealed a single open reading frame of 564 triplets, followed by a long nontranslated region. The open reading frame included peptides A, B and CD (underlined in figure 4) as well as two additional peptides that were not separated by the HPLC and gave short

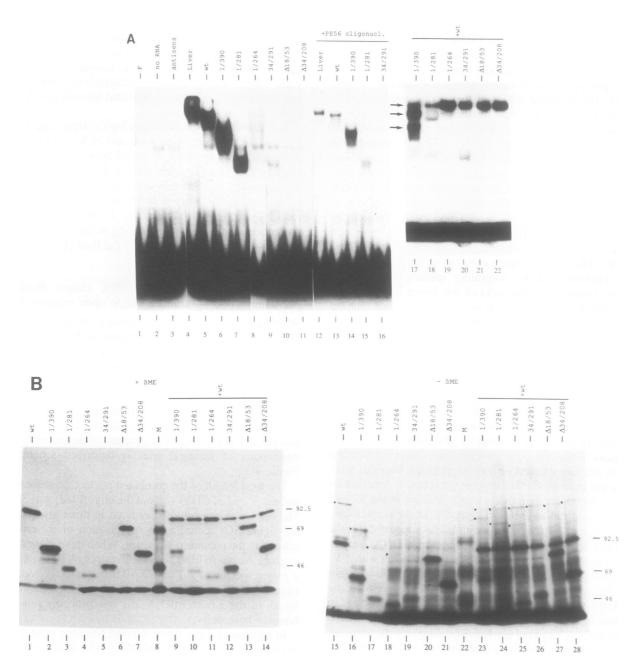


Figure 5: Functional mapping of the HNF1 protein. A: In vitro translated HNF1 analyzed by Band-shift assay. 1 ng of <sup>32</sup>P-labelled PE56a probe was incubated as described in Materials and Methods with 1 µg of rat liver nuclear proteins (lanes 4, 12) or with in vitro products of following constructions, in absence (lanes 1-11 and 17-22) or presence of 9 ng of unlabelled PE56a as competitor (lanes 12-16); the respective volumes of translation mixture used in binding assay and lanes in the Figure are indicated in paranthesis: no proteins (lane 1); reticulocyte lysate incubated without exogenous RNA (1 µl, lane 2); antisens HNF1 RNA (1 μl, lane 3); T7βH-wt (1 μl, lanes 5, 13); T7βH-1/390 (1 μl, lanes 6, 14); T7βH-1/281 (1 μl, lanes 7, 15); T7βH-1/264 (1 μl, lane 8); T7βH-34/291 (3 μl, lanes 9, 16);  $T7\beta H-\Delta 18/53$  (3  $\mu$ l, lane 10);  $T7\beta H-\Delta 34/208$  (3  $\mu$ l, lane 11); co-translation of approximately 1  $\mu$ g of both mRNA from  $T7\beta H$ -wt and  $T7\beta H$ -1/390 (1  $\mu$ l, lane 17), T7βH-1/281 (1 μl, lane 18), T7βH-1/264 (1 μl, lane 19), T7βH-34/291 (1 μl, lane 20), T7βH-Δ18/53 (1 μl, lane 21) or T7βH-Δ34/208 (1 μl, lane 22). Band-shift assay were performed as described in Materials and Methods. Note that the band visible in lane 2 is common to all lanes where in vitro translated proteins were used and is due to a reticulocyte DNA-binding activity. Arrows to the left of lane 17 indicate the positions, from top to bottom, of wilt type homodimers, heterodimers and T7βH1/390 homodimers. B: In vitro translated HNF1 analyzed by SDS-PAGE. 2 µl of translation mixtures were diluted with 8 µl of water, left 1 hour at room temperature, mixed with 10  $\mu$ l of Laemmli's sample buffer (19) in presence (lanes 1–14) or absence (lanes 15–28) of 0.1%  $\beta$ -mercaptoethanol and separated on two identical SDS-8%-polyacrylamide gels that were subsequently autoradiographed for <sup>35</sup>S-met-labelled protein detection. In the translation or co-translation reactions, approximately 1  $\mu$ g of mRNA from each of the following constructions were used and the respective lanes on gels are indicated in paranthesis: T7βH-wt (lanes 1, 15); T7βH-1/390 (lanes 2, 16); T7βH-1/281 (lanes 3, 17); T7βH-1/264 (lanes 4, 18); T7βH-34/291 (lanes 5, 19); T7βH-Δ18/53 (lanes 6, 20); T7βH-Δ34/208 (lanes 7, 21); T7βH-wt and T7βH-1/390 (lanes 9, 23), T7βH-1/281 (lanes 10, 24), T7βH-1/264 (lanes 11, 25), T7βH-34/291 (lanes 12, 26), T7βH-Δ18/53 (lanes 13, 27) or T7βH-Δ34/208 (lanes 14, 28). <sup>14</sup>C -labelled molecular weight markers (92.5, 69, 46 and 27 kDa; Amersham) were loaded in lanes 8 and 22. Dots on the left of bands on the autoradiogram indicate the homo- or hetero-dimers. Since an excess of intact protein was used in lanes 9-11 and 23-25, we do not detect homodimers of truncated proteins in these experiments.

mixed sequences; all were preceded by arginine or lysine residues as expected for tryptic peptides. The first methionine in this open reading frame was in the 54th position (amino acid no. 118 in Figure 4), thus generating a polypeptide of 511 amino acids with a molecular weight of 54.7 kDa, probably too short when compared with the 87-93 kDa protein that we purified. The absence of the initiation ATG in the cDNA clones was confirmed by analysis of a genomic clone (BP14) obtained by screening a rat library with a cDNA probe. The sequence of this clone partially overlapped the 5' sequence of the cDNA clones. The first ATG codon preceded by a nonsense codon in phase was located 192 nucleotides upstream of our cDNA 5' end (I.B., unpublished results). Using a PCR primer overlapping this putative initiation codon and a second primer encompassing a unique MluI restriction site in our cDNA sequence (respectively HNF1hn, and HNF1m in Figure 2), we amplified a 241 bp fragment from a rat liver cDNA preparation, thus making sure that there was no splicing event within the first 192 coding base pairs. This was confirmed by cloning and sequencing of the PCR fragment. This sequence, combined with that of the cDNA clones, gives rise to an open reading frame encoding a polypeptide of 628 amino acids with a calculated molecular weight of 69 kDa (see Figure 4). This sequence is virtually identical to recently published cDNA sequence for the LFB1 rat liver factor that was purified as a 45 kDa protein : the exceptions are 1 nt in the codon no. 434, coding for a valine in both sequences, and 7 nt in the noncoding region (31). A partial cDNA sequence of rat HNF1 was recently published by another group (32).

### In vitro translated HNF1 requires a motif homologous to the putative recognition helix of the homeodomain for specific DNA-binding

To confirm that the cDNA we obtained coded for a protein that could indeed bind specifically to the albumin proximal element. we inserted its complete coding sequence in a pGEM1 derived vector and synthesized HNF1 by in vitro transcription and translation. The <sup>35</sup>S-met-labelled protein was analyzed by both a gel retardation assay and SDS polyacrylamide electrophoresis. When incubated with its specific DNA-binding site as probe (PE56a), in vitro translated HNF1 gave rise to a complex that was specifically displaced by the homologous oligonucleotide (see lanes 5 and 13 in Figure 5A) but not by a mutated HNF1 DNA target (DS34 (11); not shown), thus demonstrating that our cDNA clone actually encoded the activity that had been purified from rat liver. In vitro translated HNF1 had an apparent molecular weight of 80 kDa, lower than that observed for the purified rat liver protein (87-93 kDa; Figure 1, lane 7, Figure 5A, lanes 4 and 5 and 5B) and these variations are likely to be due to the glycosylation of HNF1 (17 and our unpublished data). Examination of the complete amino acid sequence of HNF1 (Figure 4) reveals essentially two domains separated by a flexible junction rich in prolines and glycines (residues 288 to 297). The amino acid composition of the two domains is clearly different suggesting that they could have evolved separately. General features of the HNF1 protein primary sequence are schematically represented in figure 6A.

The sequence VYNWFANR (residues V264 to R271) of peptide A matches the consensus sequence I/V--WF--NRR highly conserved in the putative DNA-recognition helix of homeodomains (36). Comparison of the HNF1 sequence and homeoboxes identified to date is discussed below. To check whether this homology was functionally relevant, we analyzed the effect of progressive C-terminal deletions on HNF1 DNAbinding. Figure 5A shows that HNF1 can still bind DNA specifically when the whole C-terminal domain has been deleted (deletions T7 $\beta$ H-1/390, T7 $\beta$ H-1/281; Figure 5A, lanes 6 and 7 resp.) but not when further deletion removes a short segment containing the putative recognition helix (deletion T7 $\beta$ H-1/264; Figure 5A, lane 8). Thus, the C-terminal end of the HNF1 DNAbinding domain is localized within the 17 amino acids 265 to 281.

# The HNF1 homeodomain is unable to bind specifically to its DNA target in the absence of a distal dimerization domain

In order to map the N-terminal end of the HNF1 DNA-binding domain, we constructed deletions using different restriction sites. The N-terminal domain encoded by the longest SmaI fragment of the HNF1 cDNA, in which the first 33 amino acids are removed and 59 extra amino acids encoded in the second phase are added in C-terminal, retained only 1-2% of the wild type activity (deletion T7 $\beta$ H-34/291; Figure 5A, lane 9). Deletion of 35 amino acids from E18 to T53 encoded by the single XhoI fragment of the HNF1 cDNA led to a complete loss of specific DNA-binding (deletion T7 $\beta$ H- $\Delta$ 18/53; Figure 5A, lane 10). This deletion mutant could bind the PE56 probe with very low efficiency but only in the absence of any competitor DNA (not shown). Finally, the deletion of 175 amino acids from G34 to P208, encoded by the unique ApaI fragment of the HNF1 sequence, abolished specific as well as nonspecific DNA-binding activity (deletion T7 $\beta$ H- $\Delta$ 34/208; Figure 5A, lane 11 and data not shown).

Thus, the HNF1 homeodomain appeared to be unable to stably bind DNA by itself and the question arose as to whether it would need to dimerize to do so, since dimers of native HNF1 had been observed as discussed above. This prompted us to analyse, by band-shift assay, the ability of the different deletion mutants to form DNA-bound heterodimers with the co-translated wild type protein. In addition, in order to analyse dimerization independantly of DNA binding, we ran the same mutants on SDSpolyacrylamide gels in nonreducing conditions, thus taking advantage of the potential spontaneous S-S bond formation between two HNF1 monomers.

The first two C-terminal deletion mutants presented, on nonreducing SDS gels, the variations in mobility that were expected for truncated HNF1 homodimers (Figure 5B, lanes 15-17), thus confirming the observations made with the native protein. Co-translation of wild-type HNF1 with those mutants revealed heterodimers by both techniques (Figure 5A, lanes 17 and 18 and 5B, lanes 23, 24). As mentioned above, deletion of the putative recognition helix abolished DNA binding, thus making it impossible to check by band-shift analysis whether or not the resulting mutant could dimerize. In fact, it could still form homodimers and more clearly heterodimers which were visualized on a SDS-polyacrylamide gel (Figure 5B: lanes 18 and 25). Thus, it appeared that amino acids C-terminal to the V264 were not required for HNF1 dimerization.

In contrast, the three deletions in the N-terminal domain strongly reduced or abolished both DNA-binding and dimerization, although in different manners. First, as mentioned above, the  $T7\beta H-\Delta 18/53$  deletion mutant lost the ability to bind specifically to the HNF1 DNA target site; it also failed to form heterodimers with the intact protein in both the DNA binding assay and SDS PAGE in nonreducing conditions (figure 5A, lanes 10 and 21; figure 5B, lanes 20 and 27). It should be noted however, that the deletion of C50 makes the method irrelevant Α

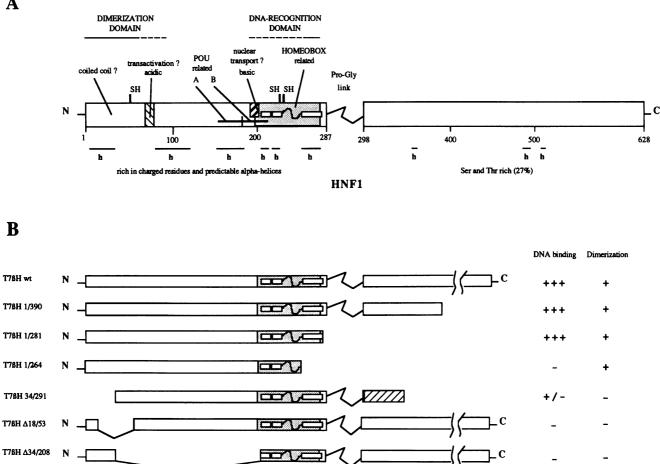


Figure 6A: Global architecture of the HNF1 protein. The domains discussed in the text are represented by boxes or bars linearly arranged along the sequence, given by the amino acid numbers used in the Figure 4. Motifs that might be involved in transactivation and nuclear transport (according to 33 and 34 resp) correspond to amino acids E71-D80 and K197-K207 respectively. SH indicate cysteine residues. N and C indicate the N- and C-terminal parts of the protein. Segments more than 3 amino acid long that were predicted as  $\alpha$ -helical using the algorithm of Garnier et al. (35) are indicated by horizontal bars with an **h** letter beneath. The three helices and the 24 amino acid loop present in the HNF1 with a loop model of homeobox are also represented (see text and Figure 7). B: Deletion analysis of the HNF1 DNA binding domain. The 7 constructions described in text are schematically represented with same scale as in Figure 6A. Parts of the HNF1 with a loop homeodomain that are present in each of them are also represented. A hatched box indicates the 59 extra amino acids, encoded in the second phase, that are C-terminal in the T7βH-34/291 protein. N and C indicate the terminal parts of the wild type HNF1 protein when present in the deletion mutants. Results of Figure 5 are summarized in the right part of the Figure.

for this mutant without precise identification of the cysteine residues involved in S-S bond formation. In any case, these results demonstrated that the homeodomain is not sufficient for specific DNA binding and indicated that formation of the HNF1 dimers requires sequences far from the homeodomain. Second, the residual DNA binding activity observed with  $T7\beta H-34/291$ suggested that the N-terminal part of the HNF1 dimerization domain (upstream of G34) is still required but perhaps less crucial than the rest of it (which includes some amino acids from G34 to T53). Indeed, the fact that DNA binding is completely lost by  $T7\beta H-\Delta 18/53$  and that the  $T7\beta H-34/291$ - and T7βH-1/281-DNA complexes have similar electrophoretic mobilities makes unlikely, though does not exclude, the hypothesis of a monomeric interaction of the HNF1 - 34/291molecule with DNA and rather suggests a weak residual ability to dimerize for this mutant. Nevertheless, this mutant did not form heterodimers when co-translated with wild type HNF1 and dimers including it have not yet been detected in nonreducing protein gels (Figure 5A, lane 20 and 5B lanes 19 and 26). It is possible that it dimerizes poorly in the absence of DNA and

cannot compete for the dimerization process with the wild type protein, when co-translated. Finally, the third N-terminal deletion T7 $\beta$ H- $\Delta$ 34/208, which was also unable to dimerize, confirmed that there are some amino acids crucial for dimerization downstream of P33 (Figure 5B, lanes 21, 28). In addition, it had lost any affinity for any type of DNA which, we assume, was retained by T7 $\beta$ H- $\Delta$ 18/53 thanks to an intact homeodomain moiety. This suggested that the N-terminus of the HNF1 homeodomain could lie upstream of A209 (see also discussion below).

### DISCUSSION

### The HNF1 DNA-recognition domain is probably not a classical helix-turn-helix motif

To document the homology between the DNA-recognition domain defined above and the homeodomain, we aligned the predicted amino acid sequence of HNF1 in the region of peptide A with 87 of the 60 amino acid-long homeobox motifs compiled to date (37, 38). The mean overall homology was approximately

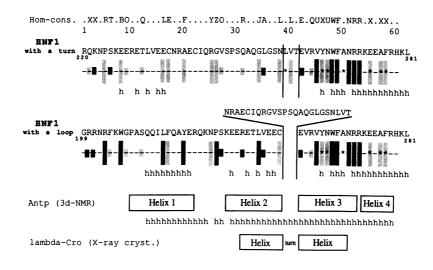


Figure 7: HNF1 homeodomain compared to the homeodomains sequenced to date. The HNF1 with a 3 amino acid turn (residues R220 to H279) and the HNF1 with a 24 amino acid loop (residues G199 to H279) sequences were aligned separately with the 31 positions where a consensus could be found among more than 60% of the homeoprotein (hom-cons, 37). Amino acids are single letter coded and 5 additional codes are used for the hom-cons sequence as follows: **B**= Big hydrophobic i.e. F, Y or I; **J**= I or L; **O** = T or S; **U** = V or I; **X** = K or R; **Z** = P or L. A diagram below each sequence illustrates the agreement with the consensus as follows: each of the 31 positions conserved in the homeobox consensus sequence is represented by a small or a big bar when between 60 to 80% or more than 80% of the homeoproteins agree with the consensus respectively. The bar is black when the corresponding HNF1 amino acid is in structural agreement with the consensus and pale when it differs. A star (\*) indicates an atypical feature specific to HNF1 (see text). Boxes at the bottom indicate the location of helices identified by three-dimensional NMR in the *Antp* homeodomain (40) and by X-ray crystallography in the  $\lambda$ -Cro helix-turn-helix motif (41). h letters indicate the  $\alpha$ -helical segments predicted from the HNF1 sequences and from the *Antp* homeobox sequence (18). Numbers below the HNF1 sequences indicate positions of their ends in the complete HNF1 amino acid sequence (Figure 4).

17%, clearly concentrated in the third helix. As recently proposed by Finney (39), allowing the looping out of 24 amino acids ('HNF1 with a loop') in the place of the canonical 3 amino acid turn ('HNF1 with a turn') between helices II and III in the HNF1 sequence led to significant improvement in homology (mean: 23%).

Figure 7 illustrates the degree of structural homology between the HNF1 sequence, in both configurations discussed here, with the consensus for the homeodomain (see criteria adopted in legend). The homology is clearly restricted to the third helix in the case of HNF1 with a turn and extends largely to the whole N-terminal region in the case of HNF1 with a loop. Moreover, predicted  $\alpha$ -helices, using the algorithm that gave the best result with the Antp sequence, as compared to available structural data (40), perfectly match those of the Antp homeodomain in the case of HNF1 with a loop while they are hardly compatible in the case of HNF1 with a turn.

If further structural analysis were to confirm the validity of the HNF1 with a loop model, one might wonder about the exact nature of the selective pressure that led to an almost universal conservation of a strict 3 amino acid turn among the helix-turnhelix proteins and about the properties of the loop that allowed HNF1 to escape this selection.

Proteins of a new family, including two that are mammalian transcription factors restricted to a specific cell type, share a highly homologous homeobox and two sequences upstream of it, defined as 'POU-A' and 'POU-B' (see references 38, 42 and 43 for review). The two latter show some homology with the HNF1 sequence, however it is very weak (highest scores : 6/26 and 4/34 respectively) as compared to homologies between the already described POU proteins themselves (42, 38). Moreover, these sequences overlap the HNF1 homeodomain whereas they are separated by roughly 20-30 amino acids in POU proteins (42, 38).

Further careful examination shows that several amino acids of the segment corresponding to the putative DNA-recognition helix in the HNF1 sequence differ radically in their nature from those found in most if not all other homeobox sequences (37). These positions are outlined as stars (\*) in Figure 7. Let us retain two features that are exclusive to HNF1: An alanine at position 50: the residue in this position is believed to determine the specificity of DNA-recognition of homeoboxes of the Antp subclass (44, 45). At this key position most known homeoboxes contain a glutamine (the other residues that have been found until now are K, C, S, H, I). An aliphatic residue at this location may seriously impair binding to DNA through hydrogen bonding or polar interactions which might explain why HNF1 should bind to DNA as a dimer. A glutamic acid and an alanine at positions 55 and 57 respectively: all known homeoproteins have an arginine or a lysine at these extremely conserved functional positions and almost all at the adjacent position 58 also. The replacement, exclusive to HNF1, of these three basic residues by one acidic and two hydrophobic ones is striking, since these amino acids have been recently demonstrated to be exposed on the external face of the fourth helical segment of the Antp homeodomain and proposed to be involved in general electrostatic contacts with DNA (40); they are usually taken as a distinctive character of all homeoproteins (43). More precise mutagenesis is needed to define which amino acids in the long third helix of HNF1 are involved in contacting DNA, since they are likely to differ from those of both the canonical homeodomain and prokaryotic helixturn-helix motif, possibly in relation to the presence of the 24 amino acid loop.

## The HNF1 DNA-binding domain includes a distal dimerization domain of new type

The similarities and differences between HNF1 and other homeoproteins prompted us to compare the HNF1 target site to other homeoprotein binding sequences determined to date. Homeoprotein DNA binding sites are AT-rich and similar to the average consensus sequence for one half of the HNF1 palindromic binding site (not shown). However, two points should be mentioned: first, except in yeast, homeoprotein target sites identified so far lack dyad symmetry; second, the HNF1 monomer has never been shown to form a stable complex with DNA. Rather, taken together, the different experiments reported here, using both native and cloned HNF1, strongly suggest that HNF1 forms homodimers, in the absence of DNA as well as when bound to its DNA target and that dimerization is essential for specific and high affinity DNA-binding. A domain required for dimerization lies outside of the homeodomain in a distal Nterminal region of the HNF1 protein. This region is rich in predicted  $\alpha$ -helical structures and shows remote homologies with sequences involved in coiled coil formation but is not included in any existing dimerization motif family (not shown). During revision of our manuscript, anoter group outlined the crucial role of the N-terminal segment of HNF1 for dimerization and DNAbinding (48) An N-terminal dimerization domain was also observed in the Mat- $\alpha$ 2 homeoprotein (46) to which it confers a very interesting flexibility in its interactions both with DNA and with other proteins (46, 47). Mat- $\alpha$ 2 dimers are like here stabilized in vitro in nonreducing conditions, the physiological relevance of which is unknown. The need to dimerize for DNAbinding has so far been observed with no other homeoprotein. By contrast, this property is shared by most of the prokaryotic factors of the  $\lambda$ -Cro-type (41).

### HNF1 properties extend the field of the homeoprotein superfamily

The example of HNF1, in addition to that of the POU proteins, further documents the homeodomain as the DNA recognition moiety of transcription factors implicated in the control of cell specific genes (43). As we discussed before, several predicted structural features extend the properties encountered among homeoproteins; this is also true at the functional level. HNF1 was initially identified only in hepatic nuclear extracts (8-11), however, northern blot and in situ hybridization showed that HNF1 transcripts are present at high levels in nonhepatic tissues like intestine and kidney (32 and our unpublished observations). On the other hand, expression of exogenous HNF1 is not sufficient to drive a high rate of transcription of a co-transfected albumin promoter in several cell types (F. Tronche, unpublished results). Thus, achievement of tissue specific patterns of transcription does not appear to be based on the action of strictly tissue specific trans-acting factors, but rather on various combinations of a small number of proteins. Homeoproteins, like HNF1, might fulfill complex regulation networks by alternative homo- or hetero-dimerizations and the low specificity observed in DNA recognition by the homeoproteins might be overcome by specific protein-protein interactions.

### ACKNOWLEDGEMENTS

We are grateful to S. Hirai and B. Arcangioli for advice in protein purification, to S. Pochet for providing primed affinity resin, to J. Van Damme and M. Puype for their help in protein blotting and amino acid sequence analysis, to F. Tronche for help in construction of pRSV-HNF1, to R. Treisman for the gift of  $T7-\beta$ globin vector, to J. Ars for typing, to N. Dostatni, R. Sousa, J. Ham and M. Weiss for valuable comments on the manuscript.

This work was supported by grants from the EEC BAP Program, from the ARC, the LNFCC, and the FMRF to M.Y. and from the Belgian National Fund for Scientific Research (N.F.W.O.) to J.V. I.B. was supported by a Boehringer Ingelheim Fonds fellowship.

#### REFERENCES

- 1. Gorski, K., Carneiro, M. and Schibler, U. (1986) Cell, 47, 767-776.
- 2. Heard, J.M., Herbomel, P., Ott M.O., Mottura-Rollier, A., Weiss, M. and
- Yaniv, M. (1987) Mol. Cell. Biol., 7, 2425-2434. Cereghini, S., Raymondjean, M., Garcia Carranca, A., Herbomel, P. and Yaniv, M. (1987) Cell, 50, 627-638.
- Lichsteiner, S., Wuarin, J. and Schibler, U. (1987) Cell, 51, 963-973.
- Babiss, L.E., Herbst, R.S., Bennet, A.L. and Darnell, Jr., J.E. (1987) Genes and Develpment, 1, 256-267.
- Herbomel, P., Rollier, A., Tronche, F., Ott, M.O., Yaniv, M. and Weiss, M. (1989). Mol. Cell. Biol., 9, 4750-4758.
- 7. Tronche, F., Rollier, A., Bach, I., Weiss, M. and Yaniv, M. (1989) Mol. Cell. Biol., 9, 4759-4766.
- 8. Courtois, G., Morgan, J.G., Campbell, L.A., Fourel, G. and Crabtree, G.R. (1987). Science, 238, 688-692.
- Hardon, E.M., Frain, M., Paonessa, G. and Cortese, R. (1988). EMBO J., 7, 1711-1719.
- 10. Schorpp, M., Kugler, W., Wagner, U. and Ryffel, G.U. (1988). J. Mol. Biol., 202, 307-320.
- 11. Cereghini, S., Blumenfeld, M. and Yaniv, M. (1988). Genes and Development, 2, 957-974.
- 12. Courtois, G., Baumhueter, S. and Crabtree, G.R. (1988). Proc. Natl. Acad. Sci. USA, 85, 7937-7941.
- 13. Blumenfeld, M., Cereghini, S., Raymondjean, M., Chouard, T. and Yaniv, M. (1988). UCLA symposia on Molecular and Cellular Biology, New Series, Vol.95, Alan R. Liss Inc., New York, NY, 91-105.
- 14. Monaci, P., Nicosia, A. and Cortese, R. (1988) EMBO J., 7, 2075-2087.
- 15. Ryffel, G.U., Kugler, W., Wagner, U. and Kaling, M. (1989) Nucl. Acids
- Res., 17, 939-953. 16. Maire, P., Wuarin, J. and Schibler, U. (1989) Science, 244, 343-346.
- 17 Lichtsteiner, S. and Schibler, U. (1989) Cell, 57, 1179-1187.
- 18. Arcangioli, B., Pochet, S., Sousa, R.M. and Huynh-Dinh, T. (1989) Europ. J. Biochem., 179, 359-369.
- 19. Laemmli, U.K. (1970) Nature, 227, 680-685.
- 20. Bauw, G., de Loose, M., Inge, D., Van Montagu, M. and Vandekerckhove, J. (1987). Proc. Natl. Acad. Sci. USA, 84, 4806-4810.
- 21. Saïki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Ehrlich, H.A. (1988) Science, 239, 487-491.
- 22. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular cloning: a laboratory manual (Cold Spring Harbor, New york).
- 23. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA, 74, 5463-5467.
- 24. Chen, C. and Seeburg, P. (1985). DNA, 4, 165-170.
- 25. Gorman, C.M., Merlino, G.T., Willingham, M.C., Pastan, I. and Howard, B.H. (1982). Proc. Natl. Acad. Sci. USA, 79, 6777-6781.
- 26. Chirngwin, J.M., Przybyla, A.E., Mac Donald, R.J. and Rutter, W.J. (1979) Biochemistry, 18, 5294-5299.
- 27. Baumhueter, S., Courtois, G. and Crabtree, G.R. (1988) EMBO J., 7, 2485-2493.
- 28. Deschatrette, J. and Weiss, M.C. (1974) Biochimie, 56, 1603-1611.
- 29. Ryseck, R.P., Hirai, S.I., Yaniv, M. and Bravo, R. (1988) Nature, 334, 535 - 537
- 30. Birnsteil, M., Busslinger, M. and Strub, K. (1985) Cell, 41, 349-359.
- 31. Frain, M., Swart, G., Monaci, P., Nicosia, A., Stämpfli, S., Frank, R. and Cortese, R. (1989) Cell, 59, 145-157.
- 32. Baumhueter, S., Mendel, D.B., Conley, P.B., Kuo, C.J., Turk, C., Graves, M.K., Edwards, C.A., Courtois, G. and Crabtree, G.R. (1990). Genes and Development, 4, 372-379
- 33. Ptashne, M. (1988) Nature, 335, 683-689.
- 34. Dingwall, C. and Laskey, R.A. (1986) Ann. Rev. Cell. Biol., 2, 367-390. 35. Garnier, J., Osguthrope, D. and Robson, B. (1978) J. Mol. Biol., 120, 97 - 120
- 36. Gehring, W.J. (1987) Science, 236, 1245-1251.
- 37. Scott, M.P., Tamkun, J.W. and Hartzell, GW.III (1989) BBA Rev., 989, 25 - 48
- 38. He, X., Treacy, M.N., Simmons, D.M., Ingraham, H.A., Swanson, L.W. and Rosenfeld, M.G. (1989) Nature, 340, 35-42.

- 39. Finney, M. (1990). Cell, 60, 5-6.
- 40. Qian, Y.Q., Billeter, M., Otting, G., Müller, M., Gehring, W.J. and Wütrich, K. (1989) Cell, 59, 573-580.
- 41. Pabo, C.O. and Sauer, R.T. (1984) Ann. Rev. Biochem., 53, 293-321.
- 42. Herr, W., Sturm, R., Clerc, R.G., Corcoran, L.M., Baltimore, D., Sharp, P.A., Ingraham, H.A., Rosenfeld, M.G., Finney, M., Ruvkun, G. and Horvitz, H.R. (1988) Genes and Development, 2, 1513-1516. 43. Levine, M. and Hoey, T. (1988) Cell, 55, 537-540.
- 44. Hanes, S.D. and Brent, R. (1989) Cell, 57, 1275, 1283.
- 45. Treisman, J., Gönczy, P., Vashishta, M., Harris, E. and Desplan, C. (1989) Cell, 59, 553-562.
- 46. Sauer, R.T., Smith, D.L. and Johnson, A.D. (1988) Genes and Development, 2, 807-816.
- 47. Keleher, C.A., Goutte, C. and Johnson, A.D. (1988) Cell, 53, 927-936.
- 48. Nicosia, A., Monaci, P., Tomei, L., De Francesco, R., Nuzzo, M., Stunnenberg, H. and Cortese, R. (1990) Cell, 61, 1225-1236.