The EMBO Journal vol.7 no.6 pp.1743–1748, 1988

Isolation of a human gene with protein sequence similarity to human and murine int-1 and the *Drosophila* segment polarity mutant *wingless*

Brandon J.Wainwright, Peter J.Scambler, Philip Stanier, Eila K.Watson, Gillian Bell, Carol Wicking, Xavier Estivill, Michael Courtney[1], Andre Boue[2], Peter S.Pedersen[3], Robert Williamson and Martin Farrall

Department of Biochemistry, St Mary's Hospital Medical School, Norfolk Place, London W2 1PG, UK, [1]Transgene S.A., 11, Rue de Molsheim, 67082 Strasbourg Cedex, [2]Unité de Recherches de Biologie Prenatale, U.73, Chateau de Longchamp, Bois de Boulogne, 75016 Paris, France and [3]Department of Clinical Genetics 4062, Rigshospitalet, 9 Blegdamsvej, DK-2100 Copenhagen, Denmark

Communicated by R.Williamson

An expressed gene sequence which was identified by the isolation of a methylation free CpG island from human chromosome 7 has been cloned from a human lung cDNA library. The deduced protein sequence contains 360 amino acids and has several features of a secreted protein; it is cysteine rich with a signal peptide sequence and two potential asn-linked glycosolation sites. The protein sequence shows marked similarity with human and murine int-1 and their *Drosophila* homolog *wingless* (Dint-1). This human int-1 related protein, int-1 and Dint-1 have diverse patterns of expression, but the inferred structural similarities suggest that some of the functional characteristics of these proteins may be shared.

Key words: int-1/HTF islands/human development/cystic fibrosis

Introduction

Regions of the vertebrate genome have been described which have a relatively high G/C content and are virtually free of methylation at the CpG dinucleotide. Such regions usually occur at discrete units 1–2 kb long which can be detected by analysis with methylation-sensitive restriction enzymes and are therefore known as HTF (*Hpa*II Tiny Fragments) islands (Bird *et al.*, 1985). HTF islands are associated with genes and in particular with the site of transcription initiation and the first exon(s). Both tissue-specific and housekeeping genes have been described in association with HTF islands (Bird, 1987). The characteristics of HTF islands make it possible to isolate selectively regions of the genome likely to contain structural genes.

We have isolated chromosome-mediated gene transfer cell lines using the activated *met* oncogene (Scambler *et al.*, 1986) which maps within 1 centimorgan of *CF* on human chromosome 7 (White *et al.*, 1985; Beaudet *et al.*, 1986) as part of a strategy to isolate candidate sequences for the cystic fibrosis (*CF*) gene. We have used DNA from these cell lines and a selective cloning method to isolate genomic regions relatively rich in the dinucleotide CpG (Estivill *et al.*, 1987). Two of these clones, NX4 and NX2, were shown to include the same HTF island and to be associated with a coding sequence. DNA polymorphisms defined by these sequences are in marked linkage disequilibrium with *CF* and suggest that these markers map within several tens of kilobases from *CF*.

Here we describe the isolation and characterization of a coding sequence associated with the NX2/NX4 HTF island which reveals a marked similarity to the murine proto-oncogene int-1 (Nusse, *et al.*, 1984a) and its *Drosophila* homolog *wingless* (Rijsewijk *et al.*, 1987; Baker, 1987; Cabrera *et al.*, 1987) (Dint-1), but is distinct from a human int-1 ortholog located on human chromosome 12 (van Ooyen *et al.*, 1985). The murine int-1 proto-oncogene was identified by cloning host DNA around the integration site of mouse mammary tumor virus (Nusse and Varmus, 1982; Nusse *et al.*, 1984) with misexpression of int-1 after retroviral mediated infection leading to transformation of mammary epithelial cells (Brown *et al.*, 1986). Int-1 has a highly specific (both temporal and spatial) pattern of expression in fetal brain and spinal cord from 9–10 day old mouse embryos (Wilkinson *et al.*, 1987; Shackleford and Varmus, 1987) but has only been demonstrated to be expressed in one adult tissue, postmeiotic spermatids (Jakobovits *et al.*, 1986; Shackleford and Varmus, 1987). *Drosophila* cDNA clones isolated by low-stringency probing with mouse int-1 cDNAs map by *in situ* hybridization to the *wingless* locus, a segment-polarity gene (Rijsewijk *et al.*, 1987).

Indirect evidence that int-1 is secreted (Papkoff *et al.*, 1987) and that the product of *wingless* is a diffusible gene product (Morata and Lawrence, 1977) suggest that these proteins are secreted 'growth factors'. The int-1 related protein (irp) we describe shows a comparable level of homology, and we propose that it is an additional member of the 'int-1 growth factor' gene family. Irp is expressed in a spectrum of fetal and adult human tissues that do not overlap with the express pattern for int-1. The function of int-1 related proteins in normal and pathological states in man remains an enigma.

Results

Isolation of lung cDNA clones

We have previously demonstrated a conserved sequence within a 50 kb 'contig' (XV contig) centered around cosmid NX2 (Estivill *et al.*, 1987). This cosmid contains part of a CpG enriched HTF island. Cosmid H147 which contains ~ 17 kb of DNA either side of the HTF island was selected to screen the lung cDNA library as it hybridizes strongly to human lung RNA dot blots after competitive hybridization with sheared total human DNA. Two independent cDNA clones were recovered after several rounds of screening; these clones were subsequently sequenced.

irp cDNA sequence

The complete sequence (2318 bp) of the irp cDNA is shown in Figure 1. This is a composite sequence and includes sequences from cDNA clone 1 (nucleotides 1−2019) and clone 2 (nucleotides 26−2318). Primer extension and S1 nuclease mapping experiments have demonstrated that the start of cDNA clone 1 is 50 nucleotides downstream from the site of transcription initiation (P.Stanier *et al.*, unpublished results). A long open reading frame (ORF) of 1101 nucleotides is identified in both cDNA clones (nucleotides 274− 1374). The initiator codon at position 295 is the third AUG triplet read from the 5′ end of the cDNA. Kozak (1987) has reported that ~10% of all eukaryotic mRNAs have AUG codons upstream of the known start of protein synthesis; however, two-thirds of proto-oncogene mRNAs have upstream AUG codons. The length of the 5′ non-coding sequence (345 bases) is considerably longer than expected for most eukaryotic mRNAs and is more typical of proto-oncogenes (Kozak, 1987). The ATG triplet at position 295 has a purine (A) three nucleotides upstream (position −3); this is a highly conserved feature of eukaryotic initiators (Kozak, 1987). The two potential initiators upstream from ATG(295)[ATG(245), ATG(273)] have pyrimidines at their −3 positions and are predicted to be poor candidates as initiators; there are also stop codons UGA(266) and UAA(291) in frame with these latter two potential initiators.

cDNA clones 1 and 2 vary in the length of their 3′ non-coding regions due to alternative utilization of polyadenylation signals. cDNA 1 most likely utilizes the sequence ATTAAA at position 1996, poly(A) is added to nucleotide 2019. Clone 2 has a canonical polyadenylation signal (AATAAA) at position 2283 with poly(A) added after nucleotide 2301. Heterogeneity of the 3′ non-coding sequence has been previously reported for several genes. There are two ATTTA motifs located close to the polyadenylation sites utilized in clones 1 and 2; these have been proposed to be the recognition signal for a processing pathway that specifically degrades the mRNAs for certain lymphokines, cytokines and proto-oncogenes (Shaw and Kamen, 1986).

A further cDNA was isolated from a placental library constructed from a 19-week gestation placenta from a fetus diagnosed by abnormal amniotic microvillar enzyme levels to be affected by CF. Sequence analysis of the coding region revealed no differences between this and the cDNAs isolated from the normal adult lung library.

The amino acid sequence

Features. The protein contains a classical hydrophobic leader sequence with α-helix forming potential (residues 8−19) and with two candidate sites for cleavage [between proline (21) and glutamate (22) or adjacent serines (25 and 26)]when a weight-matrix predictive method is used (von Heijne, 1986). The latter signalase site fits the '(−3,−1)' rule more closely but cleavage would ablate a potential asparagine-linked glycosylation site that is conserved in int-1. The protein is cysteine rich (24/360 residues; 6% by mol. wt) and includes two cysteine doublets (residues 308, 309 and 331, 332). Assuming the leader sequence is cleaved between amino acids 21 and 22, then there are two potential asn-linked glycosylation sites (shown in Figure 1); the mature protein will have an M_r 38 000 excluding any glycosolation or other post-translational modifications.

The hydrophobicity plot shown in Figure 2 (upper plot)

```
AGCAGAGCGGACGGGCGCGCGGGAGGCGCGCAGAGCTTTCGGGCTGCAGGCGCTCGCTGC
        10        20        30        40        50        60

CGCTGGGGAATTGGGCTGTGGGCGAGGCGGTCCGGGCTGGCCTTTATCGCTCGCTGGGCC
        70        80        90       100       110       120

CATCGTTTGAAACTTTATCAGCGAGTCGCCACTCGTCGCAGGACCGAGCGGGGGGCGGGG
       130       140       150       160       170       180

GCGCGGCGAGGCGGCGGCCGTGACGAGGCGCTCCCGGAGCTGAGCGCTTCTGCTCTGGGC
       190       200       210       220       230       240

                                                       1
    ___                         ___                    M  N
ACGCATGGCGCCCGCACACGGAGTCTGACCTGATGCAGACGCAAGGGGGTTAATATGAAC
       250       260       270       280       290       300

                     10                            20
 A  P  L  G  G  I  W  L  W  L  P  L  L  L  T  W  L  T  P  E
GCCCCTCTCGGTGGAATCTGGCTCTGGCTCCCTCTGCTCTTGACCTGGCTCACCCCCGAG
       310       320       330       340       350       360

                     30                            40
    _______
 V  N  S  S  W  W  Y  M  R  A  T  G  G  S  S  R  V  M  C  D
GTCAACTCTTCATGGTGGTACATGAGAGCTACAGGTGGCTCCTCCAGGGTGATGTGCGAT
       370       380       390       400       410       420

                     50                            60
 N  V  P  G  L  V  S  S  Q  R  Q  L  C  H  R  H  P  D  V  M
AATGTGCCAGGCCTGGTGAGCAGCCAGCGGCAGCTGTGTCACCGACATCCAGATGTGATG
       430       440       450       460       470       480

                     70                            80
 R  A  I  S  Q  G  V  A  E  W  T  A  E  C  Q  H  Q  F  R  Q
CGTGCCATTAGCCAGGGCGTGGCCGAGTGGACAGCAGAATGCCAGCACCAGTTCCGCCAG
       490       500       510       520       530       540

                     90                           100
 H  R  W  N  C  N  T  L  D  R  D  H  S  L  F  G  R  V  L  L
CACCGCTGGAATTGCAACACCCTGGACAGGGATCACAGCCTTTTTGGCAGGGTCCTACTC
       550       560       570       580       590       600

                    110                           120
 R  S  S  R  E  S  A  F  V  Y  A  I  S  S  A  G  V  V  F  A
CGAAGTAGTCGGGAATCTGCCTTTGTTTATGCCATCTCCTCAGCTGGAGTTGTATTTGCC
       610       620       630       640       650       660

                    130                           140
 I  T  R  A  C  S  Q  G  E  V  K  S  C  S  C  D  P  K  K  M
ATCACCAGGGCCTGTAGCCAAGGAGAAGTAAAATCCTGTTCCTGTGATCCAAAGAAGATG
       670       680       690       700       710       720

                    150                           160
 G  S  A  K  D  S  K  G  I  F  D  W  G  G  C  S  D  N  I  D
GGAAGCGCCAAGGACAGCAAAGGCATTTTTGATTGGGGTGGCTGCAGTGATAACATTGAC
       730       740       750       760       770       780

                    170                           180
 Y  G  I  K  F  A  R  A  F  V  D  A  K  E  R  K  G  K  D  A
TATGGGATCAAATTTGCCCGCGCATTTGTGGATGCAAAGGAAAGGAAAGGAAAGGATGCC
       790       800       810       820       830       840

                    190                           200
 R  A  L  M  N  L  H  N  N  R  A  G  R  K  A  V  K  R  F  L
AGAGCCCTGATGAATCTTCACAACAACAGAGCTGGCAGGAAGGCTGTAAAGCGGTTCTTG
       850       860       870       880       890       900

                    210                           220
 K  Q  E  C  K  C  H  G  V  S  G  S  C  T  L  R  T  C  W  L
AAACAAGAGTGCAAGTGCCACGGGGTGAGCGGCTCATGTACTCTCAGGACATGCTGGCTG
       910       920       930       940       950       960

                    230                           240
 A  M  A  D  F  R  K  T  G  D  Y  L  W  R  K  Y  N  G  A  I
GCCATGGCCGACTTCAGGAAAACGGGCGATTATCTCTGGAGGAAGTACAATGGGGCCATC
       970       980       990      1000      1010      1020

                    250                           260
 Q  V  V  M  N  Q  D  G  T  G  F  T  V  A  N  E  R  F  K  K
CAGGTGGTCATGAACCAGGATGGCACAGGTTTCACTGTGGCTAACGAGAGGTTTAAGAAG
      1030      1040      1050      1060      1070      1080

                    270                           280
 P  T  K  N  D  L  V  Y  F  E  N  S  P  D  Y  C  I  R  D  R
CCAACGAAAAATGACCTCGTGTATTTTGAGAATTCTCCAGACTACTGTATCAGGGACCGA
      1090      1100      1110      1120      1130      1140

                    290                           300
                                     _______
 E  A  G  S  L  G  T  A  G  R  V  C  N  L  T  S  R  G  M  D
GAGGCAGGCTCCCTGGGTACAGCAGGCCGTGTGTGCAACCTGACTTCCCGGGGCATGGAC
      1150      1160      1170      1180      1190      1200

                    310                           320
 S  C  E  V  M  C  C  G  R  G  Y  D  T  S  H  V  T  R  M  T
AGCTGTGAAGTCATGTGCTGTGGGAGAGGCTACGACACCTCCCATGTCACCCGGATGACC
      1210      1220      1230      1240      1250      1260

                    330                           340
 K  C  G  C  K  F  H  W  C  C  A  V  R  C  Q  D  C  L  E  A
AAGTGTGGGTGTAAGTTCCACTGGTGCTGCGCCGTGCGCTGTCAGGACTGCCTGGAAGCT
      1270      1280      1290      1300      1310      1320
```

```
                 350                           360
 L  D  V  H  T  C  K  A  P  K  N  A  D  W  T  T  A  T  *
CTGGATGTGCACACATGCAAGGCCCCCAAGAACGCTGACTGGACAACCGCTACATGACCC
      1330      1340      1350      1360      1370      1380

CAGCAGGCGTCACCATCCACCTTCCCTTCTACAAGGACTCCATTGGATCTGCAAGAACAC
      1390      1400      1410      1420      1430      1440

TGGACCTTTGGGTTCTTTCTGGGGGGATATTTCCTAAGGCATGTGGCCTTTATCTCAACG
      1450      1460      1470      1480      1490      1500

GAAGCCCCCTCTTCCTCCCTGGGGGCCCCAGGATGGGGGGCCACACGCTGCACCTAAAGC
      1510      1520      1530      1540      1550      1560

CTACCCTATTCTATCCATCTCCTGGTGTTCTGCAGTCATCTCCCCTCCTGGCGAGTTCTC
      1570      1580      1590      1600      1610      1620

TTTGGAAATAGCATGACAGGCTGTTCAGCCGGGAGGGTGGTGGGCCCAGACCACTGTCTC
      1630      1640      1650      1660      1670      1680

CACCCACCTTGACGTTTCTTCTTTCTAGAGCAGTTGGCCAAGCAGAAAAAAAAGTGTCTC
      1690      1700      1710      1720      1730      1740

AAAGGAGCTTTCTCAATGTCTTCCCACAAATGGTCCCAATTAAGAAATTCCATACTTCTC
      1750      1760      1770      1780      1790      1800

TCAGATGGAACAGTAAAGAAAGCAGAATCAACTGCCCCTGACTTAACTTTAACTTTTGAA
      1810      1820      1830      1840      1850      1860

AAGACCAAGACTTTTGTCTGTACAAGTGGTTTTACAGCTACCACCCTTAGGGTAATTGGT
      1870      1880      1890      1900      1910      1920

AATTACCTGGAGAAGAATGGCTTTCAATACCCTTTTAAGTTTAAAATGTGTATTTTTCAA
      1930      1940      1950      1960      1970      1980

GGCATTTATTGCCATATTAAAATCTGATGTAACAAGGTGGGGACGTGTGTCCTTTGGTAC
      1990      2000      2010      2020      2030      2040

TATGGTGTGTTGTATCTTTGTAAGAGCAAAAGCCTCAGAAAGGGATTGCTTTGCATTACT
      2050      2060      2070      2080      2090      2100

GTCCCCTTGATATAAAAAATCTTTAGGGAATGAGAGTTCCTTCTCACTTAGAATCTGAAG
      2110      2120      2130      2140      2150      2160

GGAATTAAAAAGAAGATGAATGGTCTGGCAATATTCTGTAACTATTGGGTGAATATGGTG
      2170      2180      2190      2200      2210      2220

GAAAATAATTTAGTGGATGGAATATCAGAAGTATATCTGTACAGATCAAGAAAAAAAGGA
      2230      2240      2250      2260      2270      2280

AGAATAAAATTCCTATATCATAAAAAAAAAAAAAAAAA
      2290      2300      2310
```

Fig. 1. Nucleotide sequence of irp cDNA and the deduced amino acid sequence of irp protein. The longest ORF spans nucleotides 274–1374, the deduced amino acid sequence (in one letter code) is shown and numbered above the DNA sequence, the stop codon is marked with a star. The two ATG triplets that precede the longest ORF, the two potential asn-linked glycosolation sites, polyadenylation signals and 'AU' messenger degradation motifs are all underlined.

reveals a predominantly hydrophobic domain (residues 108–129) which is the best candidate in the sequence for a transmembrane α-helix; the mean hydrophobicity (<H>) for this 21 residue peptide is +0.44. Eisenberg *et al.* (1984) have shown that for single transmembrane helices, the mean helical hydrophobicity is ≥ +0.68, thus this domain does not fulfill this criterion and probably represents a buried domain rather than a membrane spanning structure. The combination of a signal peptide and absence of a transmembrane domain or intracellular organelle targeting sequence motifs suggests that this protein is secreted.

The hydrophobic signature of the signal sequence is also prominent on the hydrophobicity plot. The hydrophobicity moment (Figure 2, lower plot), a measure to detect amphipathic sequence, shows a sharp peak for residues 192–206. This domain is strongly predicted to be α-helical using a computer predictive method (Eliopoulos *et al.*, 1982) and six basic residues (arginines and lysines) cluster on one face of the helix. This basic amphipathic feature is found in proteins like vasoactive intestinal peptide (VIP) that bind to the amphiphilic E helix of calmodulin (Cox *et al.*, 1985). Whilst there is no experimental evidence to support the binding of calmodulin by irp, we note that this amphipathic helix has the potential to be a binding site for a negatively charged ligand.

Homology with int-1 and wingless

Computer searches through the NEWAT 84 database provided by R.F.Doolittle (University of California, San Diego) with the protein searching PEPSCAN program [Martin Bishop, University of Cambridge; PEPSCAN is adapted from a fast DNA searching program—Bishop and Thompson (1984)] revealed a marked similarity to the murine oncogene int-1. There was a recent report of the isolation of a homologous sequence from *Drosophila* which when mutated is responsible for the segment polarity *wingless* phenotype; this sequence also shows significant similarity to irp.

An alignment of irp, int-1 and Dint is shown in Figure 3. All three share several structural domains; all have an amino-terminal hydrophobic signal sequence, a conserved hydrophobic domain (17 amino acids) which all fail to reach the Eisenberg criteria of single transmembrane spanning α-helices, and perhaps most strikingly, the 12 carboxy-terminal cysteines are all conserved with no introduction of gaps for alignment. For the alignment shown in Figure 3, 147/402 (36%) amino acids are identical comparing irp with int-1 and 143/487 (29%) comparing irp with Dint. 122/487 (25%) amino acids are identical for all three sequences, 139/487 (29%) if conservative substitutions are included. The carboxy-terminal 22 cysteines are all conserved between all three sequences. There are an additional 85 amino acids in the *wingless* protein that are not found in either int-1 or irp (see Figure 3, aa's 292–376). This is contributed by an additional exon (Rijsewijk *et al.*, 1987); the surrounding residues are poorly conserved between int-1 and irp (295–291, 377–381).

There is little sequence conservation in the signal sequences of irp, int-1 and Dint-1; each has a different predicted signalase cleavage site (von Heijne, 1986). If int-1 is cleaved between alanines 27 and 28 then it has four potential asn-linked glycosolation sites (with asparagines at positions 29, 317, 347 and 360); irp shares two of these sites, one site is conserved between all three proteins. Some secondary structural properties are predicted to be conserved between all three sequences; there are similar α-helical regions in int-1 and Dint corresponding to the irp helices between amino acids 70–84 and 166–205. However, the helix found for residues 192–206 does not have basic amphipathic properties for either int-1 or Dint.

More extensive and sensitive searches were carried out using the FASTP program (William Pearson, University of Virginia) with the SWISSPROT database (release 5.0; September 1987) but failed to detect any further significant similarities.

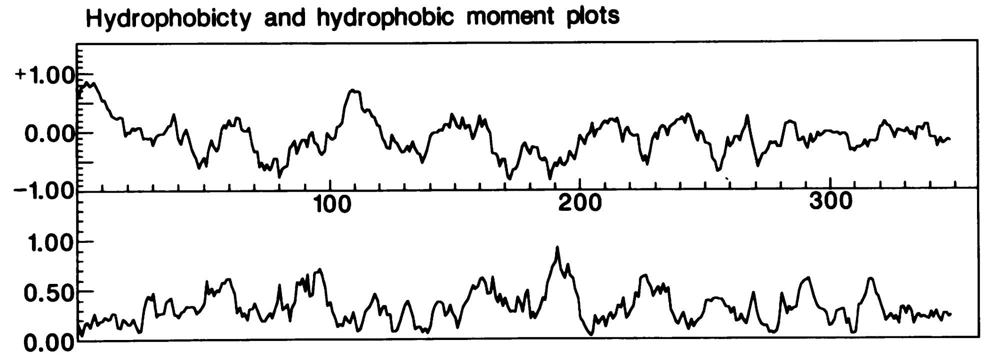

Fig. 2. Hydrophobicity and hydrophobic moment plot for irp. Mean hydrophobicities and moments are calculated for a 14 amino acid window using the scales of Eisenberg *et al.* (1984). Moments are calculated with an angle of 100° between residues corresponding to a repeat of 3.6 amino acids per helical turn.

```
  1                           30                            60
                                       **
        MNAPLGGIWLWLPLLLTWLTPEVNSS-----WWYMRATGGSSRVMCD-NVPGLV irp
   MGLWALLPGWVSATLLLALAALPAALAANSSGR---WWGIVNVASSTNLLTDSKSLQLV int-1
  MDISYIFVICLMALCSGGSSLSQVEGKQKSGRGRGSMWWGIAKVGEPNNI-T---PIMYM Dint-1

61                            90                           120
             *  *    *        *      **  ***  **** *         *
  ---S----SQRQ--LCHRHPDVMRAISQGVAEWTAECQHQFRQHRWNCNTLDRDHSL--F irp
  LEPSLQLLSRKQRRLIRQNPGILHSVSGGLQSAVRECKWQFRNRRWNCPTAPGPH-L--F int-1
  DPAIHSTLRRKQRRLVRDNPGVLGALVKGANLAISECQHQFRNRRWNCSTRNFSRGKNLF Dint-1

121                          150                           180
  *    *   **  *  ** ** *     * ** *   ** **
  GRVLLRSS-RESAFVYAISSAGVVFAITRACSQGEVKSCSCDPKKMGSAKDSK------- irp
  GKIVNRGGCRETAFIFAITSAGVTHSVARSCSEGSIESCTCDYRRRGP------------ int-1
  GKIVDR-GCRETSFIYAITSAAVTHSIARACSEGTIESCTCDYSHQSRSPQANHQAGSVA Dint-1

181                          210                           240
  *  **  *******  *  * * ***  *  *   *  ****** ***  *     ****
  GIFDW--GGCSDNIDYGIKFARAFVDAKERKGKDARALMNLHNNRAGRKAVKRFLKQECK irp
  GGPDWHWGGCSDNIDFGRLFGREFVDSGEK-GRDLRFLMNLHNNEAGRTTVFSEMRQECK int-1
  GVRDWEWGGCSDNIGFGFKFSREFVDTGER-GRNLREKMNLHNNEAGRAHVQAEMRQECK Dint-1

241                          270                           300
  *** *****  ***      *  ** *     **  *
  CHGVSGSCTLRTCWLAMADFRKTGDYLWRKYNGAIQVVMNQDGTGFTVANE--------- irp
  CHGMSGSCTVRTCWMRLPTLRAVGDVLRDRFDGASRVLYGNRGSNRASRAE--------- int-1
  CHGMSGSCTVKTCWMRLANFRVIGDNLKARFDGATRVQVTN--SLRATNALAPVSPNAAG Dint-1

301                          330                           360
  ------------------------------------------------------------ irp
  ------------------------------------------------------------ int-1
  SNSVGSNGLIIPQSGLVYGEEEERMLNDHMPDILLENSHPISKIHHPNMPSPNSLPQAGQ Dint-1

361                          390                           420
                       *           ****** **  *      *  ** **
  ----------------RFKK-PT--------KNDLVYFENSPDYCIRDREAGSLGTAGRV irp
  ----------------LLRLEPEDPAHKPPSPHDLVYFEKSPNFCTYSGRLGTAGTAGRA int-1
  RGGRNGRRQGRKHNRYHFQLNPHNPEHKPPGSKDLVYLEPSPSFCEKNLRQGILGTHGRQ Dint-1

421                          450                           480
  **  *   * *   *****           * * ***** * *  *        *
  CNLTSRGMDSCEVMCCGRGYDTSHVTRMTKCGCKFHWCCAVRCQDCLEALDVHTCKAPKN irp
  CNSSSPALDGCELLCCGRGHRTRTQRVTERCNCTFHWCCHVSCRNCTHTRVLHECL---- int-1
  CNETSLGVDGCGLMCCGRGYRRDEVVVVERCACTFHWCCEVKCKLCRTKKVIYTCL---- Dint-1


481
  ADWTTAT irp
  ------- int-1
  ------- Dint-1
```

Fig. 3. Sequences (one letter code) of irp, int-1 and Dint-1 aligned by hand. Gaps introduced to align the sequences are shown as horizontal lines. Identical residues are marked by *; 22 cysteines are conserved between all three sequences.

Expression studies

Figure 4 shows the result from a typical Northern blotting experiment showing expression of irp in human term placenta and fetal lung (18 week gestation). Equivalent levels of transcription are also observed in 10 week gestation chorionic villus, 18 week gestation placenta and adult lung. Multiple transcripts with sizes 2.9 kb, 2.4 kb and 2.1 kb are detected. The 2.4 kb and 2.1 kb transcripts correspond to the differential polyadenylation products evident from cDNAs 1 and 2. Since there is no evidence of heterogeneity at the 5′ end (P.Stanier *et al.*, unpublished results) and the irp locus contains a single gene which does not cross-hybridize to any other (Estivill *et al.*, 1987), it is possible that the 2.9 kb transcript represents an as yet uncloned product of further differential polyadenylation. We were unable to detect irp gene expression in the liver, kidney, pancreas and colon from an 18 week gestation fetus, adult peripheral lymphocytes and cerebral cortex, and cultured sweat duct epithelium. All Northern analyses were performed upon 5 μg of poly(A)$^+$ RNA. Comparison of expression in placental or lung RNA prepared from normal and CF affected tissues revealed no differences in either size of transcripts or gross level of expression.

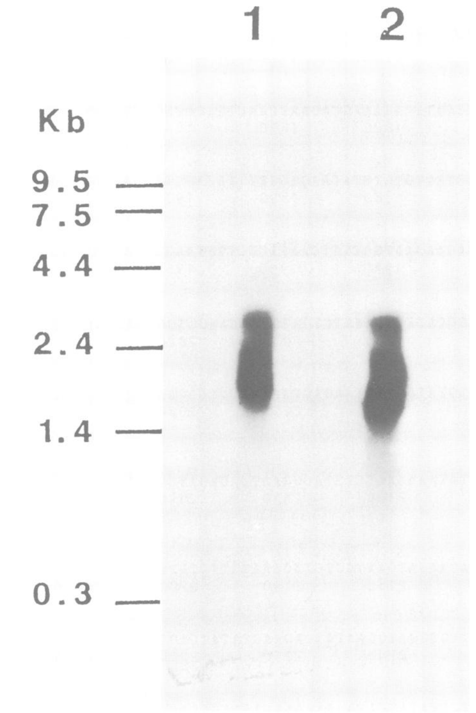

Fig. 4. Northern blot analysis showing irp gene expression in human term placenta (track 1) and 18 week gestation lung (track 2). Each sample consisted of 5 μg poly(A)$^+$ RNA extracted from fresh tissue. Exposure was for 6 h at −70°C.

Discussion

Relationship between irp, int-1 and Dint

Our human irp sequence shows a comparable and high level of similarity to both irp and Dint-1. All features [hydrophobic leader peptide, asn-linked glycosolation site(s) and cysteine rich domain] that are associated with secretion are conserved. The *wingless* protein has an additional 85 amino acid internal sequence contributed by an extra exon which is not found in int-1 or irp.

Irp is not the human ortholog of murine int-1; this has been previously cloned and maps to chromosome 12q14-pter (Nusse *et al.*, 1984b), while irp maps to chromosome 7q3.1. Human and murine int-1 share an almost identical protein sequence and also show extensive DNA homology for both exons and introns. In spite of the extensive sequence homology between human irp, murine int-1 and Dint, there is very little DNA sequence identity when the three gene sequences are compared. Using low stringency Southern blot analysis irp does not detect human int-1 sequences.

There is a remarkable degree of conservation of cysteines between all three proteins. The cysteine disulphide bridges are likely to constrain the tertiary structure so all three proteins present similar potential binding domains and we speculate that this int-1 family of 'growth factor' proteins may be recognized by a family of related receptors. The inferred common topology makes it likely that each protein will bind to alternative receptors leading to cross-reactivity. However related these proteins may be at the sequence and structural level, int-1 and Dint-1 have diverse functions and irp has as yet unknown functions.

Implications

Recent electrophysiological characterization has demonstrated that the basic defect in CF is probably in a membrane-associated component of chloride ion transport (Schoumacher *et al.*, 1987). The irp gene we have described here is most probably secreted from the cell, shows no differences between CF and normal individuals on sequence and Northern analysis, and is therefore unlikely to be the CF gene itself.

DNA hybridization studies have shown that there is an irp-related sequence in several mammalian species including mouse, chicken and *Xenopus* (Estivill *et al.*, 1987). We have attempted to detect by low stringency hybridization a conserved sequence in *Drosophila melanogaster, Anopheles aegyptensis* or *Caenorhabditis elegans* without success. This may indicate that although the various int-1 genes are present in both invertebrates and vertebrates, irp and its closely related DNA sequences are present only in vertebrates. Whilst the irp gene is associated with an HTF island there is as yet no defined role of such structures in the control of vertebrate gene expression.

The cellular role of irp is not yet understood. There is no apparent ontological expression pattern and of all of the tissues we have examined expression has been detected in only placental, fetal and adult lung. This is particularly interesting as adult lung is not usually regarded as an organ involved in the secretion of growth factors. Neither the human ortholog of int-1 nor irp have a characterized role in normal development in man nor in any pathological state.

A variety of methods now exist to test the transforming potential of irp (Brown *et al.*, 1986), as well as studying the effects of underexpressing irp in embryonic tissue, perhaps by the use of site-directed mutagenesis followed by homologous recombination (Thomas and Capecchi, 1987). The definition of the cellular function of int-1 and irp in man may provide a clue as to the common features of the signals required for the determination of growth and differentiation shared by such diverse organisms as *Drosophila*, mouse and man.

Materials and methods

Isolation of cDNA clones

A human adult lung cDNA library was constructed by annealing dG-tailed cDNA with a synthetic adaptor and cloning into the *Eco*RI site of gt10 (Le Bouc *et al.*, 1986). Two *Hin*dIII fragments (8.7 kb and 7 kb) from cosmid H147 were excised from a low-gelling temperature agarose gel and radiolabelled by random oligonucleotide priming ('oligolabelling') to a specific activity of $>10^8$ c.p.m./μg (Feinberg and Vogelstein, 1984). Probes were then competed with sheared total human DNA, lorist DNA and *Escherichia coli* DNA as described previously (Scambler *et al.*, 1987) before screening $\sim 10^6$ plaques immobilized on Hybond-N^{TM} nylon filters (Amersham International plc). Three positive plaques were picked and re-screened; recombinants were bulk prepared for subcloning and sequencing.

DNA sequencing strategy

Two independent recombinants, cDNA 1 and 2, were cleaved with *Eco*RI and subcloned into M13mp18 (Yanisch-Perron *et al.*, 1985). Both strands were sequenced by a modified dideoxy chain termination method (Sanger *et al.*, 1977; Tabor and Richardson, 1987) using several custom synthesized oligonucleotides as primers.

Construction and screening of CF placental DNA library

cDNA was synthesized from poly(A)$^+$ RNA extracted from a 19 week gestation placenta after termination of pregnancy for a '1 in 4 risk for CF' previously shown to have elevated microvillar enzymes in amniotic fluid consistent with a CF affected fetus (Carbans *et al.*, 1983; Boue *et al.*, 1986). On examination the fetal small bowel was found to be diagnostically obstructed by instipated meconium; albumin levels were grossly raised in the meconium (120 mg/ml) confirming the diagnosis. cDNA was ligated into gt10 and $\sim 10^6$ clones screened with a lung irp cDNA. Recombinants were subcloned and sequenced as above.

RNA isolation and hybridization

RNA was prepared from 'snap frozen' tissues by a modified LiCl/urea method, and poly(A)$^+$ RNA was size fractionated on a 1% agarose–formaldehyde gel and transferred to nylon membranes. Inserts from cDNA clones 1 or 2 were oligolabelled, hybridizations were carried out in 50% formamide at 42°C and washed at 60°C to a final stringency of at least 0.15 M NaCl.

Acknowledgements

We wish to thank our many clinical and scientific colleagues, the cystic fibrosis adults and children and their families, and our Departmental colleagues for help with this research. We are grateful to Professor Tom Blundell for helpful discussion, and thank Dr R.Staden, L.M.B., Cambridge, and Dr Geddes, University of Leeds, for providing DNA/protein sequence and secondary structure prediction programs. The Cystic Fibrosis Genetics Research Group is funded by generous grants from the Cystic Fibrosis Research Trust (UK) and the Medical Research Council. M.F. is a Foulkes Foundation Fellow. X.E. is supported by the British Council and the Spanish Ministry of Education and Science.

References

Baker,N.E. (1987) *EMBO J.*, **6**, 1765–1773.

Beaudet,A., Bowcock,A., Buchwald,M., Cavalli-Sforza,L., Farrall,M., King,M.-C., Klinger,K., Lalouel,J.-L., Lathrop,G., Naylor,S., Ott,J., Tsui,K.-C., Wainwright,B., Watkins,P., White,R. and Williamson,R. (1986) *Am. J. Hum. Genet.*, **39**, 681–693.

Bird,A.P. (1987) *Trends Genet.*, **3**, 342–347.

Bird,A., Taggart,M., Frommer,M., Miller,O.J. and Macleod,D. (1985) *Cell*, **40**, 91–99.

Bishop,M. and Thompson,E. (1984) *Nucleic Acids Res.*, **12**, 5471–5474.

Boue,A., Muller,F., Nezelof,C., Oury,J.F., Duchatel,F., Dumez,Y., Aubry,M.C. and Boue,J. (1986) *Hum. Genet.*, **74**, 288–297.

Brown,A.M.C., Wildin,R.S., Prendergast,T.J. and Varmus,H.E. (1986) *Cell*, **46**, 1001–1009.

Cabrera,C.V., Alonso,M.C., Johnston,P., Phillips,R.G. and Lawrence,P.A. (1987) *Cell*, **50**, 659–663.

Carbans,N.J.B., Gosden,C. and Brock,D.J.H. (1983) *Lancet*, **i**, 329–331.

Cox,J.A., Comte,M., Fitton,J.E. and DeGrado,W.F. (1985) *J. Biol. Chem.*, **260**, 2527–2534.

Eisenberg,D., Schwartz,E., Komaromy,M. and Wall,R. (1984) *J. Mol. Biol.*, **179**, 125–142.

Eliopoulos,E.E., Geddes,A.J., Brett,M., Pappin,D.J.C. and Findlay,J.B.C. (1982) *Int. J. Biol. Macromol.*, **4**, 263.

Estivill,X., Farrall,M., Scambler,P.J., Bell,G.M., Hawley,M.F., Lench, N.J., Bates,G.P., Kruyer,H.C., Frederick,P.A., Stanier,P., Watson,E.K., Williamson,R. and Wainwright,B.J. (1987) *Nature*, **326**, 840–845.

Feinberg,A.P. and Vogelstein,B. (1984) *Anal. Biochem.*, **137**, 266–267.

Jakobovits,A., Shackleford,G.M., Varmus,H.E. and Martin,G.R. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7806–7810.

Kozak,M. (1987) *Nucleic Acids Res.*, **15**, 8125–8148.

Le Bouc,Y., Dreyer,D., Jaeger,F., Binoux,M. and Sondermeyer,P. (1986) *FEBS Lett.*, **196**, 108–112.

Morata,G. and Lawrence,P.A. (1977) *Dev. Biol.*, **56**, 227–240.

Nusse,R. and Varmus,H.E. (1982) *Cell*, **31**, 99–109.

Nusse,R., van Ooyen,A., Cox,D., Fung,Y.K.T. and Varmus,H.E. (1984a) *Nature*, **307**, 131–136.

Nusse,R., Veer,L., Geurts van Kessel,A., van Agthoven,A., Bootsma,D. and Varmus,H. (1984b) *Cytogenet. Cell. Genet.*, **37**, 556–557.

Papkoff,J., Brown,A.M.C. and Varmus,H.E. (1987) *Mol. Cell. Biol.*, **7**, 3978–3984.

Rijsewijk,F., Schuermann,M., Wagenaar,E., Parren,P., Weigel,D. and Nusse,R. (1987) *Cell*, **50**, 649–657.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5466.

Scambler,P.J., Law,H.-Y., Williamson,R. and Cooper,C.S. (1986) *Nucleic Acids Res.*, **14**, 7159–7174.

Scambler,P.J., Estivill,X., Bell,G., Farrall,M., McLean,C., Newman,R., Little,P.F.R.,Frederick,P., Hawley,K., Wainwright,B.J., Williamson,R. and Lench,N. (1987) *Nucleic Acids Res.*, **15**, 3639–3652.

Shackleford,G.M. and Varmus,H.E. (1987) *Cell*, **50**, 89–95.

Shaw,G. and Kamen,R. (1986) *Cell*, **46**, 659–667.

Schoumacher,R.A., Shoemaker,R.L., Halm,D.R., Tallant,E.A., Wallace, R.W. and Frizzell,R.A. (1987) *Nature*, **330**, 752–754.

Tabor,S. and Richardson,C.C. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4767–4771.

Thomas,K.R. and Capecchi,M.R. (1987) *Cell*, **51**, 503–512.

van Ooyen,A., Kwee,V. and Nusse,R. (1985) *EMBO J.*, **4**, 2905–2909.

von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.

White,R., Woodward,S., Leppert,M., O'Connell,P., Hoff,M., Herbst,J., Lalouel,J.-M., Dean,M. and Vande Woude,G. (1985) *Nature*, **318**, 382–384.

Wilkinson,D.G., Bailes,J.A. and McMahon,A.P. (1987) *Cell*, **50**, 79–88.

Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, **33**, 103–119.

Received on February 8, 1988; revised on March 2, 1988