Chicken NFI/TGGCA proteins are encoded by at least three independent genes: NFI-A, NFI-B and NFI-C with homologues in mammalian genomes

Ralph A.W.Rupp[+], Ulrich Kruse[§], Gerd Multhaup, Ulrike Göbel, Konrad Beyreuther and Albrecht E.Sippel*[§]
Zentrum für Molekulare Biologie der Universität Heidelberg (ZMBH), INF 282,
D-6900 Heidelberg, FRG

Received February 12, 1990; Revised and Accepted March 27, 1990 EMBL accession nos. X51483–X51486 (incl.)

ABSTRACT

Chicken TGGCA proteins belong to the ubiquitous, eukaryotic family of NFI-like nuclear proteins, which share an identical DNA binding specificity. They are involved in viral and cellular aspects of transcriptional regulation and they are capable of stimulating Adenovirus initiation of replication. Using microsequencing data from peptides of isolated proteins and PCR supported cloning, we have derived four cDNAs for NFI/TGGCA proteins, which are encoded by three separate chicken genes. Sequence alignments of NFI proteins from chicken and various mammalian species provide evidence for a common genetic equipment among higher eukaryotes, in which several related genes, employing each differential RNA splicing generate an unexpectedly large family of diverse NFI proteins. The extensive similarity of the amino acid sequence throughout the complete coding regions between products of the same gene type in different species indicates a uniform selection pressure on all protein parts, also on those outside the DNA-binding domain.

INTRODUCTION

During recent years a rapidly increasing number of cis-active DNA elements has been identified by transfection experiments, which control mRNA-synthesis in eukaryotes. They can be divided into promoters, guiding the correct initiation by RNA polymerase II, and proximal and distal elements, e.g. enhancers, which modulate the extent and specificity of transcription (1, 2). These regulatory sites are composed of multiple motifs for sequence-specific DNA-binding proteins acting synergistically. For example, this was shown in great detail for the SV40 enhancer (3, 4).

Investigating cis-elements, which control the tissue- and stage-specific expression of the chicken lysozyme gene in oviduct and myeloid cells of the hematopoietic system, we detected a cell-specific enhancer element located 6.1kb upstream of the transcriptional start-site (5). One component of the protein-enhancer DNA-complex was identified as the TGGCA protein by in vitro (6) and in vivo footprinting (Borgmeyer et al., manuscript in preparation). This DNA-binding protein was characterized by us in nuclear protein extracts of various chicken cells (7, 6, 8, 9). The prototype recognition sequence for NFI/TGGCA protein is the palindrome 5′-PyTGGCANNNTG-CCAPu-3′ (8), which is bound by protein-dimers (10, 11). It soon became obvious, however, that there is a family of proteins, sharing this DNA-binding specificity, which was found to be ubiquitously present in all cells of vertebrate species tested so far (7, 12, 13, 14, 15, 9, 16). Independently purified protein populations like NFI (13) and CTF (17), which were shown to be involved in Adenovirus replication and transcription of the TK promoter respectively, turned out to be identical by various functional and biochemical criteria (14). We have purified the TGGCA protein from chicken liver by preparative mobility shift electrophoresis (15) and we could show that it is able to substitute for HeLa cell NFI in the reconstituted Adenovirus replication system (18).

The recent cloning of NFI/CTF cDNAs from human (19) and other mammalian species (20, 16, 11), revealed indeed highly conserved domains among these proteins. It was shown that individual cDNAs can stimulate both transcription and viral replication (19). Structural characteristics of CTF cDNAs suggested that at least part of the protein diversity is generated by differential splicing (19).

In this paper we report the cloning of four cDNAs for NFI/TGGCA proteins from chicken. Despite extensive amino acid sequence identity within aminoterminal domains, the cDNAs do not crossreact under stringent hybridization conditions. Their isolation was successful by using PCR-amplified DNA-fragments derived from specific TGGCA protein peptides for the screening of cDNA-libraries. NFI/TGGCA proteins are encoded by at least

* To whom correspondence should be addressed

[+] Present addresses: Department of Genetics, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA and [§]Institut für Biologie III, Schänzlestrasse 1, D-7800 Freiburg/Br., FRG three different genes. Most strikingly, sequence alignments with other NFI-cDNAs reveal the existence of subgroups, which are derived from a small number of related genes in each vertebrate species.

MATERIALS AND METHODS

cDNA-Libraries (λgt11) made from poly A^+-RNA

Library 1: liver tissue of adult Leg-Horn rooster (Clontech Laboratories Inc.; CL1002, 2.1×10^6 independent clones; oligo(dT)primed); library 2: BM2 cell line (AMV-transformed myeloblasts (32); 1.4×10^6 independent clones; oligo(dT)plus random primed); library 3: 10 days old decapitated chicken embryos (1×10^6 independent clones; oligo(dT)primed) (33).

Oligonucleotides

ON1: 5′-GCCTG(C/G)AGGT(A/G)AACCA(A/G/T)GTGTA-(A/G)GCAAA(A/G)GC-3′;
ON3: 5′-TT(A/G)AACCAIGT(A/G)TAIGC(A/G)AAIGC-3′;
ON5: 5′-*GGAATTC*CAT(A/C/T)CI(A/T)CGIGA(C/T)CA-(A/G)GA-3′;
ON6: 5′-*GGAATTC*(G/T)TG(A/G)AA(A/G)TCIGTA(C/G)-(A/T)-3′;
ON13: 5′-*GGAATTC*(C/T)GA(A/G)TT(C/T)CA(C/T)CCITT-(C/T)ATIGA(A/G)GC-3′;
ON14: 5′-*GGAATTC*AA(C/T)CCIGA(C/T)CA(A/G)AA(A/G)-GGIAA(A/G)ATG-3′;
ON15: 5′-*GGAATTC*(A/G)AA(A/G)TTIGGICCIGTICCIG-CIGC(A/G/T)ATIG-3′;
ON16: 5′-*GGAATTC*AGCGTCTTGGGCTTGGT-3′;
ON21:
5′-*GGAATTC*(G/T)AT(A/G)TGGTG(G/T)GGCTG(G/T)-A(C/T)GCA-3′.

Oligonucleotides were synthesized by the solid phase phosphoramidite method. 'I' represents deoxyinosine. Underlined nucleotides indicate additional EcoRI-linkers to facilitate cloning of PCR-products. Alternative nucleotides in brackets are present simultaneously.

Sequence analysis of TGGCA protein species

TGGCA Protein species were purified from chicken liver as described (15). Total protein of final DNA-cellulose fraction was subjected to preparative SDS-PAGE (34; gel buffer contained 0.1mM sodium thioglycolate). TGGCA protein species (36.8-32.5kd) were electroeluted into elution buffer (50mM NH_4CO_3, 0.05% SDS, 0.05mM sodium thioglycolate, 0.001% dithioerythritol) in a biotrap (Schleicher and Schüll) and aceton-precipitated (35). The proteins were redissolved in digestion buffer (50 mM NH_4CO_3, 0.1% Nonidet P40) and quantitated by acid hydrolysis (36) followed by analysis of phenylthiocarbamyl derivatives (ABI 420A amino acid analyser system with on-line PTC-analysis; data analysis module 920A; Applied Biosystems). Appr. 40μg of protein were digested with 5μg Trypsin (Boehringer Mannheim, sequence grade) by incubation at 37°C for 3h. Peptide fragments were directly separated on a wide pore C8-column (J.T. Baker) by a linear gradient of 0–100% acetonitrile in 0.1% trifluoracetic acid (70min, 1ml/min) and lyophilized. Selected peptides were redissolved in 30μl formic acid and aliquots (100–300pmol) were dried on TFA-activated glass filter discs, preconditioned with polybrene. They were sequenced using the ABI 477A pulse liquid gas phase sequencer with on-line PTH-analysis (120A PTH analyser, Applied Biosystems). The separation of the PTH-amino acids was performed according to the manufacturer's protocol.

PCR-Amplification

Amplification of DNA fragments with Taq-DNA polymerase was performed as published (37, 38) in 100μl reaction volumes at final concentrations of 20μM nucleotide triphosphates and 1 μM oligonucleotide primers. 30 cycles were employed under the following conditions: 0.5min, 93.5°C/0.5min, 55°C/1min, 70°C with all primers listed above, except for the cloning of p125.2 by ON5 and ON6. In this case temperatures were lowered to 45°C (annealing) and 60°C (synthesis). 10–100pg cDNA plasmids or a 10μl aliquot of the BM2 cDNA library were used as templates. Products were purified by polyacrylamide gelelectrophoresis, subcloned via EcoRI linkers attached to PCR primers into pbluescript KSM13+ and verified by DNA sequencing.

cDNA Cloning

NC-filterbound λ-phages of library 1 were hybridized with kinased oligonucleotides ON1 and ON3 under reduced stringency (5×SET, 10×Denhardt's, 0.1% SDS, 0.1% $Na_2P_4O_7$, 100μg/ml salmon sperm DNA; 20×SET = 0.6M TrisHCl, pH8.0, 3M NaCl, 20mM EDTA) at 42°C (ON1) or 37°C (ON3). Wash steps were performed with 2×SET/0.1% SDS for 20min each at 45°C and 48°C (ON1) or 42°C (ON3), respectively. Screening libraries 2 and 3 with nick-translated probes was performed under stringent conditions. Hybridizations were carried out in 50% formamide, 20mM HEPES pH7.0, 5×SSC, 5×Denhardt's, 0.1% SDS at 42°C, final washes with 0.1×SSC/0.1% SDS at 68°C. After sub-cloning of cDNA inserts in pbluescript KSM13+ DNA sequences were determined by dideoxy-sequencing of both strands (39).

Southern Blot Analysis

DNAs (genomic DNA: chicken hybrid line Niechos) were electrophoretically separated on a 0.8% agarose/0.5×TBE gel and processed for southern blotting according to a standard protocol (40). Nick-translated DNA-fragments (30–45ng, specific activities: $4-7\cdot10^8$ cpm/μg) were hybridized to Hybond N-bound DNA in 50% formamide, 20mM HEPES pH7.0, 5×SSC, 5×Denhardt's, 0.5% SDS, 50μg/ml salmon sperm DNA at 42°C (20h). Filters were washed in a final step with 0.1×SSC/0.1% SDS at 68°C (2×10min).

RESULTS

Amino acid sequencing of TGGCA protein peptides

TGGCA Protein has been purified from chicken liver as a microheterogenic family of polypeptides. The comparison of V8-proteolytic peptide maps of individual species indicated a high relationship in their primary structure (15). For amino acid sequencing we purified the bulk of TGGCA proteins (i.e. the 36.8-32.5kd species) via SDS-PAGE from DNA cellulose fractions (15). The smallest variant (29.8kd) was omitted because of its possible contamination with unrelated protein. After tryptic digest of gel-eluted proteins, peptides were separated by HPLC. Fig. 1 shows the column's elution profile with the corresponding amino acid sequences of selected peptides. From this analysis we gained a data base of 127 amino acids from 11 independent peptides, which were necessary for cloning and identification of the various NFI/TGGCA protein cDNA species.

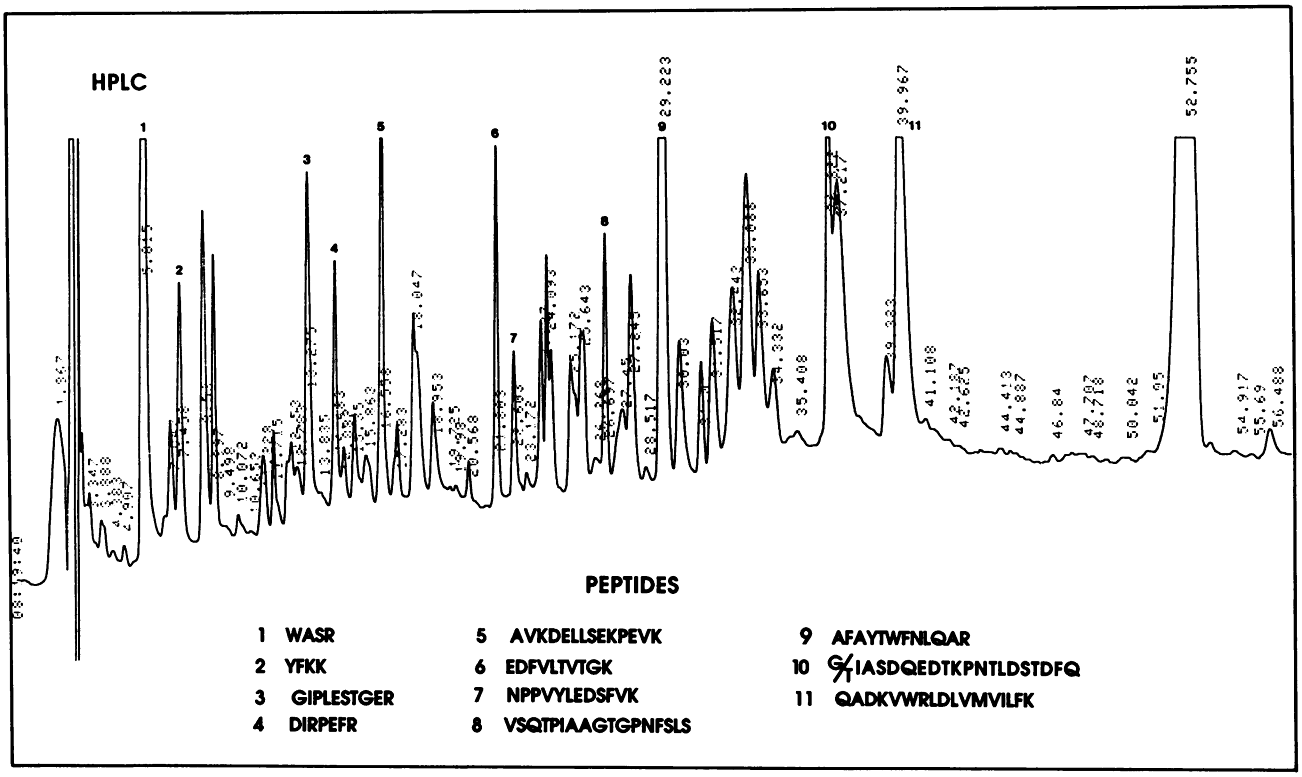


Figure 1. HPLC-separation and sequences of tryptic peptides of chicken liver TGGCA proteins. 40μg of trypsin digested TGGCA protein were separated by hydrophobic chromatography. Peptides from numbered peaks were microsequenced and their amino acid sequence is given below in one letter code.

Cloning of NFI-B cDNAs

Two degenerate oligonucleotides were derived from the amino acid sequence of peptide 9 (Fig.1) to screen a chicken liver cDNA library. The first oligonucleotide mixture (ON1) made use of preferential codon usage predictions (21) and CpG-depletion (22), and the second oligonucleotide mixture (ON2) included inosin (23) at highly ambigous positions to keep the oligonucleotide complexity low. Parallel screening with both probes of 6×10^6 λ-phages resulted in the isolation of several identical cDNA clones.

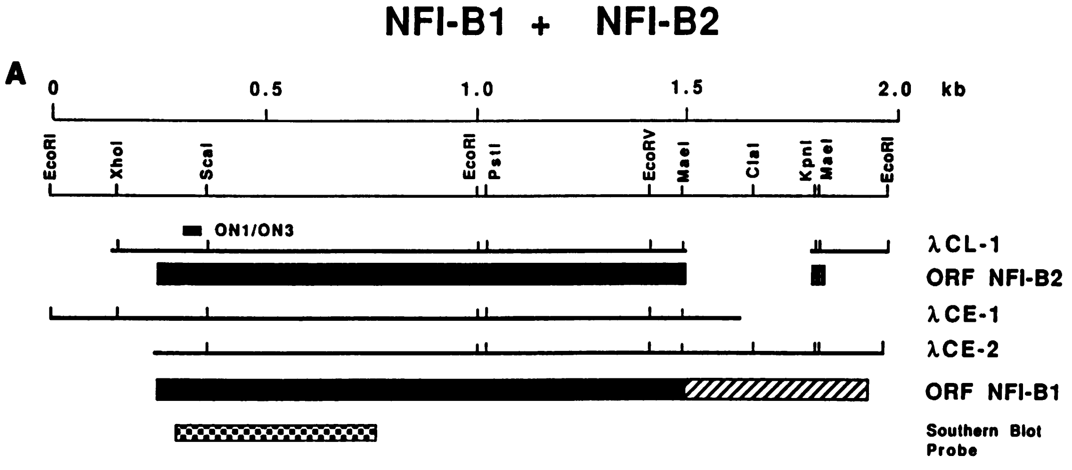
The clone λCL-1 (Fig.2B) contains an open reading frame (ORF) of 1257 bp coding for a protein of 419 amino acids. This ORF is complete since all three reading frames contain stop-codons upstream of the double-methionine at positions 262/265. Furthermore, 5′-end analysis of RNA determines the transcriptional start site to a position approximately 50 bp upstream of the 5′-end of λCE-1 (Rupp and Qian, unpublished results). 9 out of 11 TGGCA Protein peptide sequences were found in the sequence of the ORF (marked in Fig.2B). The peptides 4, 5 and 9 contain each a single amino acid substitution. Surprisingly, the sequence of amino acids 226–242 (Fig.2B) resembles only partly peptide 8, but differs in 8 out of 17 positions. The peptide 10 sequence is missing completely from λCL-1.

This finding prompted us to look for cDNA variants. We screened a second cDNA library derived from chicken embryonic mRNA. Besides cDNAs resembling the λCL-1 type, we isolated two overlapping cDNA clones called λCE-1 and λCE-2 (Fig.2). Their open reading frame is identical with the one of λ-CL-1 besides a C-terminal insertion of 104 amino acids (boxed in Fig.2B). Due to a frameshift at the 3′-border of the insert, the λCE-2 ORF terminates 37 codons further downstream than λCL-1, resulting in a total coding capacity of 560 amino acids.

Genomic southern blot analysis gives no indication for more than one gene hybridizing to these cDNAs (see Fig.5). Therefore, the two types of coding regions, which we name NFI-B1 and NFI-B2, are most probably differentially spliced products of a common RNA precursor. The observed differences between peptide sequences 4, 5 and 9 and the amino acid sequence of NFI-B2 were also found in NFI-B1, which has been isolated from a different cDNA library. The low homology to peptide 8 and the absence of the peptide 10 sequence from all our initially isolated clones made us expect further cDNAs belonging to the NFI/TGGCA protein family. Instead of screening cDNA libraries under hybridization conditions of low stringency, we empioyed a PCR-supported method to isolate clones carrying the missing peptide sequences.

Cloning of a cDNA containing the peptide 10 sequence

In a first step we cloned a DNA-fragment coding for the central 7 amino acids of peptide 10 (i.e. TKPNTLD—see Fig.1). This was accomplished by PCR-amplification of the respective nucleotide sequence from total λ-DNA of a cDNA library made from mRNA of the chicken myeloid BM2 cell line. The degenerate primers ON5 and ON6 for the PCR reaction were derived from amino acid sequences located at the borders of peptide 10 (Fig.3A). DNA-sequencing of the expected PCR-product (p125.2, Fig.3A) proved the authenticity of the amplified DNA-fragment, since the nucleotides between the primers coded indeed for the central amino acids of peptide 10. In a second round of PCR-amplification we cloned a longer DNA-fragment to increase the reliability of cDNA-screening. This time we used as primers the oligonucleotides ON16, which contains the unique sequence coding for the central part of peptide 10, and ON14,



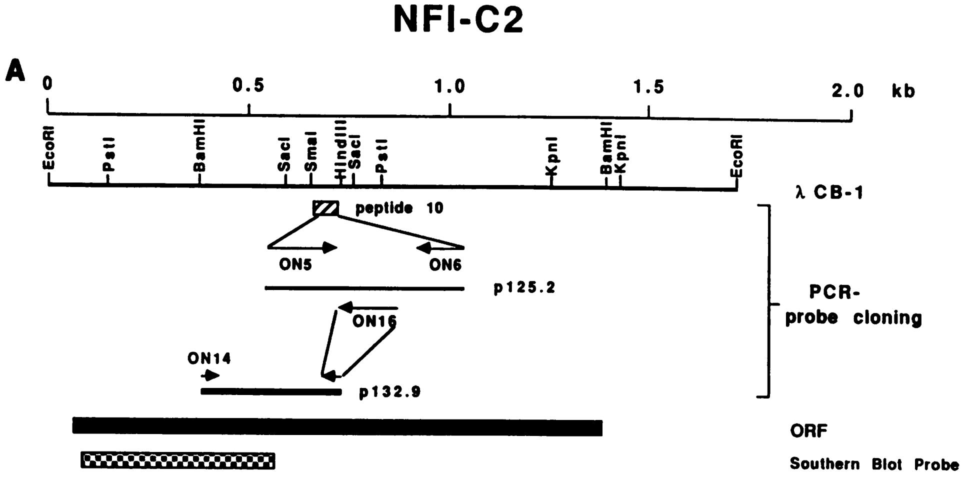
B

```
   1 GAATTCCACTACAAGAGGGTTAAAAGTGTAGATTGGATTTCACCCCGGGAAATCGAGCGCACGGAGTGAACTTGAATCTTTGGCTATTTAAGGAGGACTGGGGTTTGTTGTGGAGTTGTG

 121 GTGATTCAGGGCAGAGCCCCGTCCTGATTGATTTCATCTTCCTCGAGTCTCAGATGACTGTAAAATGAACAGATGAAATTATTCCTTTTTTGAGGCTTTTTCTTAGGGGCTCTCCGGGCA
                                                                       ON13
 241 GCGCGTTTGCAAAAGCGAAGTCATGATGTATTCTCCAATCTGTCTCACTCAGGATGAATTTCACCCGTTCATTGAGGCACTTCTTCCGCATGTCCGTGCAATTGCCTATACTTGGTTCAA
   1                       M  M  Y  S  P  I  C  L  T  Q  D  E  F  H  P  F  I  E  A  L  L  P  H  V  R  A  I  A  Y  T  W  F  N
                                                                                                               F        9
 361 CCTCCAGGCTCGAAAACGCAAGTACTTCAAAAAACATGAGAAACGAATGTCGAAGGATGAAGAAAGAGCAGTCAAGGATGAGCTACTAAGCGAAAAGCCCGAAATTAAACAGAAGTGGGC
  34  L  Q  A  R  K  R  K  Y  F  K  K  H  E  K  R  M  S  K  D  E  E  R  A  V  K  D  E  L  L  S  E  K  P  E  I  K  Q  K  W  A
                             2                                                    5                 V
 481 ATCCAGGCTCCTCGCCAAGCTGCGCAAAGATATTCGCCAAGAGTTCCGGGAGGATTTTGTGCTCACGGTGACTGGGAAGAAGCACCCGTGCTGTGTATTGTCCAATCCTGACCAGAAGGG
  74  S  R  L  L  A  K  L  R  K  D  I  R  Q  E  F  R  E  D  F  V  L  T  V  T  G  K  K  H  P  C  C  V  L  S  N  P  D  Q  K  G
      1                                4 P              6
 601 TAAGATTAGGAGGATCGACTGCCTGAGACAGGCTGACAAAGTCTGGCGTCTGGATCTAGTCATGGTGATCCTGTTCAAAGGCATCCCCTTGGAAAGTACGGATGGAGAGCGGCTCATGAA
 114  K  I  R  R  I  D  C  L  R  Q  A  D  K  V  W  R  L  D  L  V  M  V  I  L  F  K  G  I  P  L  E  S  T  D  G  E  R  L  M  K
                                       ON21           11                                   3
 721 ATCCCCACATTGCACAAATCCAGCACTTTGCGTCCAGCCACACCACATAACAGTATCAGTCAAGGAGCTTGATTTGTTTTTGGCATACTACGTGCAGGAGCAAGATTCTGGACAACCAGG
 154  S  P  H  C  T  N  P  A  L  C  V  Q  P  H  H  I  T  V  S  V  K  E  L  D  L  F  L  A  Y  Y  V  Q  E  Q  D  S  G  Q  P  G

 841 AAGTCCAAGCCACAATGATCCTGCCAAGAATCCACCAGTGTATCTTGAGGACAGTTTTGTAAAATCTGGAGTGTTCAACGTGTCAGAACTTGTGAGAGTGTCCAGAACACCTATAACCCA
 194  S  P  S  H  N  D  P  A  K  N  P  P  V  Y  L  E  D  S  F  V  K  S  G  V  F  N  V  S  E  L  V  R  V  S  R  T  P  I  T  Q
                                               7
 961 AGGAACTGGAGTGAACTTCCCAATTGGAGAACTTCCGAGCCAGCCGTACTACCATGATATGAATTCTGGTATAAATCTGCAGAGGTCTTTATCCTCTCCACCAAGCAGCAAAAGACCCAA
 234  G  T  G  V  N  F  P  I  G  E  L  P  S  Q  P  Y  Y  H  D  M  N  S  G  I  N  L  Q  R  S  L  S  S  P  P  S  S  K  R  P  K

1081 AACTATCTCCATAGATGAGAATATGGAGCCAAGCCCGACAGGAGACTTTTATCCTTCTCCAAATTCACCAGCTGCGGGAAGTCGGACATGGCATGAAAGAGATCAAGATATGTCTTCTCC
 274  T  I  S  I  D  E  N  M  E  P  S  P  T  G  D  F  Y  P  S  P  N  S  P  A  A  G  S  R  T  W  H  E  R  D  Q  D  M  S  S  P

1201 TACAATGAAGAAACCTGAAAAGCCTTTGTTTAGTCCTACTTCTCCCCAGGATTCTTCACCAAGACTGAGCACTTTCCCCCAGCATCACCACCCAGGAATCCCAGGAGTTGCACAGAGTGT
 314  T  M  K  K  P  E  K  P  L  F  S  P  T  S  P  Q  D  S  S  P  R  L  S  T  F  P  Q  H  H  H  P  G  I  P  G  V  A  Q  S  V

1321 CATCTCAACTAGAACTCCACCTTCACCTTCACCATTGCCATTTCCAGCACAAGCTATTCTCCCTCCAGCCCCGTCTAGTTACTTCTCTCATCCAACGATCAGATATCCTCCCCACCTGAA
 354  I  S  T  R  T  P  P  S  P  S  P  L  P  F  P  A  Q  A  I  L  P  P  A  P  S  S  Y  F  S  H  P  T  I  R  Y  P  P  H  L  N

1441 TCCTCAGGATACTCTGAAGAATTATGTACCTTCTTACGACCCGTCCAGTCCACAGACTAGCCAGCCTAACAGCAGTGGTCAAGTAGTAGGGAAAGTGCCTGGCCATTTTACTCCAGTTTT
 394  P  Q  D  T  L  K  N  Y  V  P  S  Y  D  P  S  S  P  Q  T  S  Q  P  N  S  S  G  Q  V  V  G  K  V  P  G  H  F  T  P  V  L

1561 GGCACCCTCTCCCCATCACAGTGCGGTGCGACCTGTGACCCTGACCATGACAGATACTAAGCCCATCACTACATCCACTGAAGCCTACACAGCTTCAGGCACATCCCAAGCCAATCGATA
 434  A  P  S  P  H  H  S  A  V  R  P  V  T  L  T  M  T  D  T  K  P  I  T  T  S  T  E  A  Y  T  A  S  G  T  S  Q  A  N  R  Y

1681 TGTGGGACTAAGCCCAAGAGACCCATCCTTCCTACACCAGCAACAGCTGAGGATTTGTGACTGGACCATGAATCAAAACGGCAGGCATTTATACCCCAGTGCCAGTGAGGATACATTGGG
 474  V  G  L  S  P  R  D  P  S  F  L  H  Q  Q  Q  L  R  I  C  D  W  T  M  N  Q  N  G  R  H  L  Y  P  S  A  S  E  D  T  L  G

1801 AATTACTTGGCAAAGTCCTGGTACCTGGGCTAGCTTGGTTCCTTTTCAAGTATCAAACAGGACACCGATCCTGCCGACAAATGTCCCAAATTATGGTTTGAACATAATTGGAGAGCCTTT
 514  I  T  W  Q  S  P  G  T  W  A  S  L  V  P  F  Q  V  S  N  R  T  P  I  L  P  T  N  V  P  N  Y  G  L  N  I  I  G  E  P  F
                   S  W  Y  L  G  *

1921 CCTTCAAGCAGAAAGCAGCAACTGAGGGAAAAAATAAAAAGAAAAATTAGGAAAAAAAAAGGAATTC
 554  L  Q  A  E  S  S  N  *
```

Figure 2. Structure and sequence of NFI-B cDNAs. **Part A** depicts a schematical view of individual and shared sequences among NFI-B cDNA clones. The position of ON1/ON3 in respect to the ORF is indicated by a black rectangle. Differently striped boxes represent the NFI-B1 specific insert and the NFI-B2 specific C-terminus. **Part B** shows the complete nucleotide and amino acid sequence of the three clones. The NFI-B1 specific insert and peptide 9 are boxed, the other peptide sequences present on these clones are underlined and numbered according to Fig.1. At positions, where peptide and cDNA sequences differ, amino acids present on the peptides are written below the cDNA-translation. The predicted C-terminal sequence of NFI-B2 is written below the translation of the NFI-B1 reading frame. Additionally, arrows indicate positions of degenerate PCR-primers ON13/ON21 for cloning of the NFI-B specific southern blot probe. Extension of individual λ-clones: λCL-1: pos 151–1504/1816–1987; λCE-1: pos. 1–1643; λCE-2: pos. 244–1976.

which is again a degenerate primer. ON14 (Fig.3A) is derived from part of a large region, which is highly conserved among all known NFI/TGGCA protein species (Fig.6). The cloned PCR-fragment (p132.9, Fig.3A) of approximately 300 bp was then used for screening the BM2-cDNA library under stringent hybridization conditions.

From 1.2×10^6 recombinant λ-phages 25 clones were initially isolated. In Fig.3B we show the sequence of λCB-1, which carries an open reading frame of 439 codons calculated from the first methionine at position 65. Although this frame contains no stop-codon upstream of this ATG, it represents most likely the start of the mRNA's coding region as is suggested by cDNA-



B

```
                                                                                                                  ON13
1    GAATTCCACGACCACGCCTCGAGTCTCAGCTCTGCAATTAGCCACTTTCCCAGCGCCGCTCGGGATGTATTCGTCTCCTCTCTGCCTGACCCAGGATGAGTTCCACCCCTTCATCGAGGC
1                                                                     M  Y  S  S  P  L  C  L  T  Q  D  E  F  H  P  F  I  E  A

121  GCTGCTCCCCCACGTGCGGGCCTTCGCCTACACCTGGTTCAACCTGCAGGCCCGCAAGCGCAAGTACTTCAAGAAGCACGAGAAGAGGATGACGAAGGACGAGGAGCGGGCGGTGAAGGA
20    L  L  P  H  V  R  A  F  A  Y  T  W  F  N  L  Q  A  R  K  R  K  Y  F  K  K  H  E  K  R  M  T  K  D  E  E  R  A  V  K  D
                                          9                                2
241  CGAGCTGCTGAGCGAGAAGCCCGAGGTGAAGCAGAAATGGGCCTCGCGCCTCCTGGCCAAGCTGCGCAAGGACATCCGGCCCGAGTGCCGCGAGGACTTCGTGCTCTCCATCACCGGCAA
60    E  L  L  S  E  K  P  E  V  K  Q  K  W  A  S  R  L  L  A  K  L  R  K  D  I  R  P  E  C  R  E  D  F  V  L  S  I  T  G  K
               5                                1                                   4     F                   6T  V
361  GAAGCCGTCGTGCTGCGTCCTCTCCAACCCCGACCAGAAGGGCAAAATGCGCCGCATCGACTGCCTGCGCCAGGCCGACAAGGTGTGGAGGCTGGACCTGGTGATGGTCATCCTCTTCAA
100  _K  P  S  C  C  V  L  S  N  P  D  Q  K  G  K  M  R  R  I  D  C  L  R  Q  A  D  K  V  W  R  L  D  L  V  M  V  I  L  F  K
                                                                                    ON21            11
481  AGGGATCCCGCTGGAGAGCACCGACGGCGAGCGGCTGGTCAAGGCGGGCCAATGCACCAACCCCATCCTCTGCATCCAGCCCCACCACATCAGCGTCTCCGTCAAAGAGCTCGACCTCTA
140  _G  I  P  L  E  S  T  D  G  E  R  L  V  K  A  G  Q  C  T  N  P  I  L  C  I  Q  P  H  H  I  S  V  S  V  K  E  L  D  L  Y
                      3
601  CTTGGCCTACTTCGTCCGCGAGAGAGATTCCGAGCAGAGCAGCAGTCCCCGGACGGGCATCGCCTCGGACCAGGAGGACACCAAGCCCAACACGCTGGACTCCACAGACTTCCAGGAAAG
180   L  A  Y  F  V  R  E  R  D  S  E  Q  S  S  S  P  R  T [G  I  A  S  D  Q  E  D  T  K  P  N  T  L  D  S  T  D  F  Q] E  S
                                                                                          10
721  CTTCGTCACCTCGGGGGTGTTCAGTGTCACCGAGCTCATCCAGGTGTCCCGAACGCCGGTGGTGACGGGCACGGGCCCAAATTTCTCCCTGGGGGAGCTGCAGGGCCACCTGGCCTACGA
220   F  V  T  S  G  V  F  S  V  T  E  L  I  Q  V  S  R  T  P  V  V  T  G  T  G  P  N  F  S  L  G  E  L  Q  G  H  L  A  Y  D
                                                      Q        I  A  A                          S
841  CCTGAACCCCTCCAGCACGGGCATGAGGAGGACGCTGCCCAGCACCTCCTCCAGCGGGAGCAAACGGCACAAGTCTGGCTCCATGGAGGATGACATCGACACGAGCCCCGGCGGCGAGTA
260   L  N  P  S  S  T  G  M  R  R  T  L  P  S  T  S  S  S  G  S  K  R  H  K  S  G  S  M  E  D  D  I  D  T  S  P  G  G  E  Y

961  CTACACCTCCTCCAACTCACCCACGAGTAGCAGCCGCAACTGGACAGAGGACATGGAAGGGGGCATCTCCCCCAACGTGAAGACAGAGATGGACAAATCGCCCTTCAACAGCCCCTCGCC
300   Y  T  S  S  N  S  P  T  S  S  S  R  N  W  T  E  D  M  E  G  G  I  S  P  N  V  K  T  E  M  D  K  S  P  F  N  S  P  S  P

1081 GCAGGACTCCTCGCCCCGCCTGAGCAGCTTCACGCAGCACCACCGTCCCGTCATCGCGGTGCACAGCGGTATCGCTCGCAGCCCGCACCCTTCGTCCACGCTGCATTTCCCCACCACCTC
340   Q  D  S  S  P  R  L  S  S  F  T  Q  H  H  R  P  V  I  A  V  H  S  G  I  A  R  S  P  H  P  S  S  T  L  H  F  P  T  T  S
                        .
1201 CATTCTCCCCCAGACGGCCTCCACCTATTTCCCCCACACGGCCATCCGGTACCCGCCTCATCTCAACCCGCAGGACCCACTGAAAGATCTCGTCTCGCTGGCCTGCGACCCGTCCAACCA
380   I  L  P  Q  T  A  S  T  Y  F  P  H  T  A  I  R  Y  P  P  H  L  N  P  Q  D  P  L  K  D  L  V  S  L  A  C  D  P  S  N  Q

1321 GCAGCCCGGACCGCCTACTCTGCACCAGGCACGCCCCCTGCGAACCGTTCCTTCGTGGGATTAGGACCAAGGGATCCTGGGAGTATCTATCAGGCACAGTCCTGGTACCTGGGATAGCGC
420   Q  P  G  P  P  T  L  H  Q  A  R  P  L  R  T  V  P  S  W  D  *

1441 GGCCGTGCCACCCTTCGCTCCCTTCTCCTCTGCTTTAGGCACCCTGAGAGGACCCCCCGGCGCGGAGCCCCCGGCGGACAGCGTGCATTTACCACCACCTCCACCACCTCCATCGTGAC

1561 GACGACGACGACGACGACGACGACGACAAAGCAAAACCAAACCCCCCCCTCCCCCCTCCCCTCGGCCCCCCCATCCCATCCCCGGACGGAAGGAGAAGAGGGAAAAGGAAGGAAGGT

1681 TGGAACTTTCTTTTATGAAAAAAAAAAAAAACCCGGAATTC
```

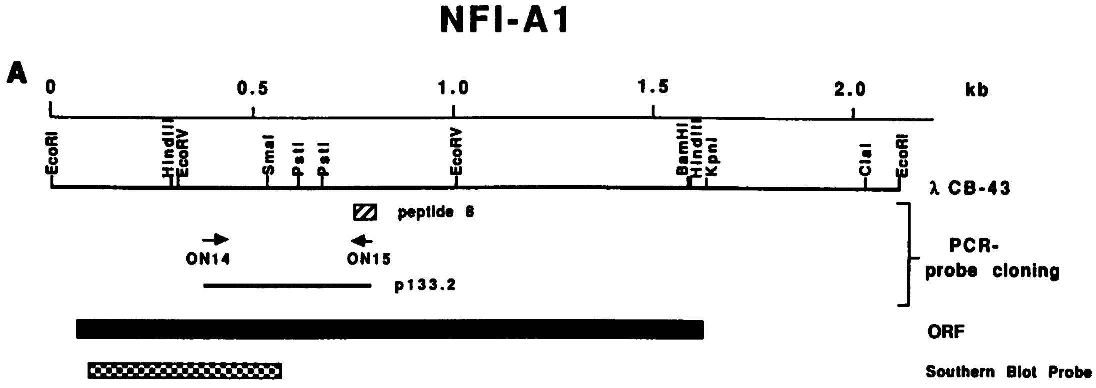
Figure 3. Structure and sequence of NFI-C2. **Part A** schematizes features of the cDNA clone NFI-C2. The positions of peptide 10 and of the probe used for southern blot analysis are shown in respect to the ORF. In addition, the two-step PCR-cloning of the probe P132.9, used for screening the BM2-cDNA library, is outlined. **Part B** displays both nucleotide and corresponding amino acid sequences of NFI-C2. The diagnostic peptide 10 is boxed, while all other peptide sequences present on this clone are underlined (compare with Fig.1). At positions, where peptide and cDNA sequence differ, amino acids present on the peptides are written below the amino acid sequence. Arrows indicate the positions of the PCR-primers ON13/ON21 used for cloning of the NFI-C specific southern blot probe.

alignments (Fig.6). Including peptide 10, this ORF contains again 9 out of 11 peptide sequences of our data base derived from TGGCA protein species isolated from liver. In contrast to NFI-B1/B2, peptide 7 is missing from this cDNA. Furthermore, peptides 9 and 5, which diverged each in one position from the NFI-B coding region, are perfectly fitting to the new cDNA, while new amino acid substitutions are now found within sequences of peptides 4 and 6 (see Fig.3B).

The sequence alignment (Fig.6) of this clone with NFI cDNAs from other species (19, 16, 20, 11) demonstrates its similarity with CTF-2 from HeLa cells (19). We therefore name this ORF-type NFI-C2.

Cloning of a cDNA containing peptide 8

As in the case of NFI-C2, we chose to clone first a specific hybridization probe by PCR for the isolation of peptide 8 containing cDNAs (Fig.4A). We derived a degenerate oligonucleotide ON15 from part of the sequence of peptide 8 (pos. 3–12, i.e. QTP...PNF; Fig.1). To make the primer discriminate against the so far isolated cDNA-types, its 3′-end was complementary not to codons for arginine, but for glutamine. Using ON15 and ON14, a 420bp long DNA-fragment was generated by PCR from total λ-DNA of the BM2-cDNA library (p133.2, Fig.4A).



B

```
                                                                                                   ON13
   1 GAATTCCGGCCTCTTTTCTCTCTCCCGCACACGCACACCTCCGCCCGCAGCCCCCCGCAGCAGTCATGTATTCTCCGCTCTGTCTCACGCAGGATGAATTTCATCCCTTCATCGAAGCAC
   1                                                                   M  Y  S  P  L  C  L  T  Q  D  E  F  H  P  F  I  E  A

 121 TTCTGCCTCACGTCCGAGCGTTTGCGTACACGTGGTTCAACCTGCAAGCCCGAAAGAGAAAATACTTCAAAAAACATGAGAAACGCATGTCAAAAGAGGAAGAGAGGGCCGTGAAGGACG
  19 L  L  P  H  V  R  A  F  A  Y  T  W  F  N  L  Q  A  R  K  R  K  Y  F  K  K  H  E  K  R  M  S  K  E  E  E  R  A  V  K  D
                                       9                                2
 241 AACTCTTAAGTGAAAAACCCGAGGTAAAGCAGAAATGGGCGTCCAGACTTCTGGCAAAGCTTCGGAAAGATATCAGGCCTGAATTTCGGGAGGATTTTGTTCTCACAGTCACAGGAAAAA
  59 E  L  L  S  E  K  P  E  V  K  Q  K  W  A  S  R  L  L  A  K  L  R  K  D  I  R  P  E  F  R  E  D  F  V  L  T  V  T  G  K
                  5                          1                                     4                         6
 361 AGCCTCCGTGTTGCGTCCTTTCCAATCCAGACCAGAAGGGCAAGATGAGAAGAATTGACTGCCTTCGCCAGGCAGATAAGGTCTGGAGGTTGGACCTCGTTATGGTGATTTTATTTAAAG
  99 K  P  P  C  C  V  L  S  N  P  D  Q  K  G  K  M  R  R  I  D  C  L  R  Q  A  D  K  V  W  R  L  D  L  V  M  V  I  L  F  K
                                                                                   ON21           11
 481 GTATTCCGCTCGAAAGTACTGATGGCGAGCGCCTTGTAAAGTCCCCACAGTGCTCTAACCCCGGGCTGTGCGTACAGCCCCACCATATAGGAGTTTCTGTTAAGGAACTCGATTTATATT
 139 G  I  P  L  E  S  T  D  G  E  R  L  V  K  S  P  Q  C  S  N  P  G  L  C  V  Q  P  H  H  I  G  V  S  V  K  E  L  D  L  Y
                    3
 601 TGGCATACTTTGTGCACGCTGCAGATTCAAGTCAATCTGAAAGTCCCAGCCAGCCAAGTGAAGCTGATATTAAGGACCAGCCAGAAAATGGACATTTGGGCTTCCAGGACAGTTTTGTCA
 179 L  A  Y  F  V  H  A  A  D  S  S  Q  S  E  S  P  S  Q  P  S  E  A  D  I  K  D  Q  P  E  N  G  H  L  G  F  Q  D  S  F  V

 721 CATCAGGTGTTTTCAGTGTGACCGAGCTAGTAAGAGTCTCACAAACACCAATAGCTGCAGGAACAGGCCCTAATTTCTCCCTCTCAGATTTGGAAAGTTCTTCGTACTACAGCATGAGCC
 219 T  S  G  V  F  S  V  T  E  L  V  R  V  S  Q  T  P  I  A  A  G  T  G  P  N  F  S  L  S  D  L  E  S  S  S  Y  Y  S  M  S
                                                                    8
 841 CAGGAGCAATGAGGAGGTCTTTACCTAGCACATCCTCTACCAGCTCTACAAAACGTATCAAATCTGTGGAGGATGAGATGGACAGTCCTGGTGAGGAGCCGTTCTATACAAGCCAAGGGC
 259 P  G  A  M  R  R  S  L  P  S  T  S  S  T  S  S  T  K  R  I  K  S  V  E  D  E  M  D  S  P  G  E  E  P  F  Y  T  S  Q  G

 961 GCTCTCCAGGAAGTGGCAGCCAGTCAAGTGGATGGCATGAAGTGGAACCAGGATATCTAAGAAACCCCGAACACCGCGGAGCACTACATGGAATGCCATCGCCAACTGCATTAAAGAAGT
 299 R  S  P  G  S  G  S  Q  S  S  G  W  H  E  V  E  P  G  Y  L  R  N  P  E  H  R  G  A  L  H  G  M  P  S  P  T  A  L  K  K

1081 CAGAGAAGTCTGGTTTCAGCAGTCCCTCGCCTTCGCAGACCTCCTCCCTTGGAACGGCTTTCACACAGCACCATCGACCTGTCATTACAGGACCCAGAGCAAGTCCACATGCAACACCAT
 339 S  E  K  S  G  F  S  S  P  S  P  S  Q  T  S  S  L  G  T  A  F  T  Q  H  H  R  P  V  I  T  G  P  R  A  S  P  H  A  T  P

1201 CGACTCTTCATTTTCCAACATCACCCATTATCCAACAGCCTGGGCCATACTTCTCACACCCAGCAATCCGCTATCATCCTCAGGAGACTCTGAAAGAGTTTGTCCAACTTGTCTGCCCCG
 379 S  T  L  H  F  P  T  S  P  I  I  Q  Q  P  G  P  Y  F  S  H  P  A  I  R  Y  H  P  Q  E  T  L  K  E  F  V  Q  L  V  C  P

1321 ACGCTGGTCAGCAGGCTGGACAGGTGGGGTTCCTAAATCCCAATGGTAGCAGCCAAGGCAAGGTGCACAATCCATTCCTTCCTACCCCAATGTTGCCACCACCACCTCCGCCACCAATGG
 419 D  A  G  Q  Q  A  G  Q  V  G  F  L  N  P  N  G  S  S  Q  G  K  V  H  N  P  F  L  P  T  P  M  L  P  P  P  P  P  P  P  M

1441 CTAGGCCTGTGCCTCTGCCAGTGCCAGACACAAAACCTCCAACTACATCAACAGAAGGAGGGGCCACCTCTCCCACTTCTCCAACCTACTCGACACCCAGCACCTCCCCCGCAAACCGAT
 459 A  R  P  V  P  L  P  V  P  D  T  K  P  P  T  T  S  T  E  G  G  A  T  S  P  T  S  P  T  Y  S  T  P  S  T  S  P  A  N  R

1561 TCGTCAGTGTTGGACCACGGGATCCAAGCTTTGTAAATATCCCTCAACAGACTCAGTCCTGGTACCTGGGATAAAAGTCTCCGCTTCCCACCACGCACCGGACAGACCACCTGACCCCTT
 499 F  V  S  V  G  P  R  D  P  S  F  V  N  I  P  Q  Q  T  Q  S  W  Y  L  G  *

1681 CCCACCTCTGTAACACGGAAGCAGCAGCATCGCAGCGCAGCGACTTCACATCGGCGGAAGGGGAGGCGAGCAGACACCGACAGCAAACTCGTACATGGAAATGGCAAACGTTATGGTTGA

1801 ATGGCAAAGGCCGTAACTTTTTCGAGATTTTTTTTTTTTTTAAACAACTTAGGACTGTTGTAATTTCTCCTATGGTGCTGGAAATGGTTGGGCTTCATAACTTTTGAAGTGTTTCATTGG

1921 TAGTGTAAGCGTTAGGTGACACTGAGTAGAACTAGCCTCTGCTGCTGCTTACCAACTCGCTGTTGTAACACACTTTCCATATGGAGCCTAACTGTTTGACAGTATCCTCATCGATTTTTC

2041 CCATACAGGTGTGAGACTTCTTGAGAAATCATGAAACACACTTAAACTATAACTTTTGTAGTAGCTGAATTC
```

Figure 4. Structure and sequence of NFI-A1. **Part A** presents schematically the structure of the cDNA-clone NFI-A1. The positions of peptide 8 and of the southern blot probe are indicated by marked rectangles. PCR-amplification with primers ON14/ON15 produced the DNA fragment p133.2, which was used for isolation of NFI-A1. **Part B** shows the nucleotide and deduced amino acid sequence of NFI-A1. The sequence of the diagnostic peptide 8 is boxed, all other peptides present on this clone are underlined and numbered according to Fig.1. Additionally, arrows indicate the positions of ON14 and ON21 used to clone the NFI-A1 specific southern blot probe.

Screening 1.2×10^6 λ-phages with this probe under stringent hybridization conditions resulted in the isolation of 22 clones, from which one was chosen for sequence analysis (λCB-43, Fig.4). The cDNA displays an open reading frame for a protein of 522 amino acids, which we call NFI-A1. Although there are no stop-codons upstream of the first ATG present on the cDNA (Fig.4B), amino acid sequence comparisons (Fig.6) again suggest this to be the site of translational initiation. Similar to the previously described clones, the majority of peptides from our data base (9 out of 11) are found also within this cDNA. Among these is peptide 8, which was used for PCR-probe cloning. But in contrast to NFI-B and -C types, where we found some differences between several peptides and cDNA sequences, NFI-A1 shows no amino acid substitution. It lacks peptides 7 and 10, however, which can be looked at momentarily as being specific for NFI-B and -C coding regions respectively.

So far we have cloned four different cDNA types, whose ORFs contain all peptide sequences we have derived from chicken liver TGGCA protein species. These clones were isolated from three different cDNA libraries with different probes. In order to assess

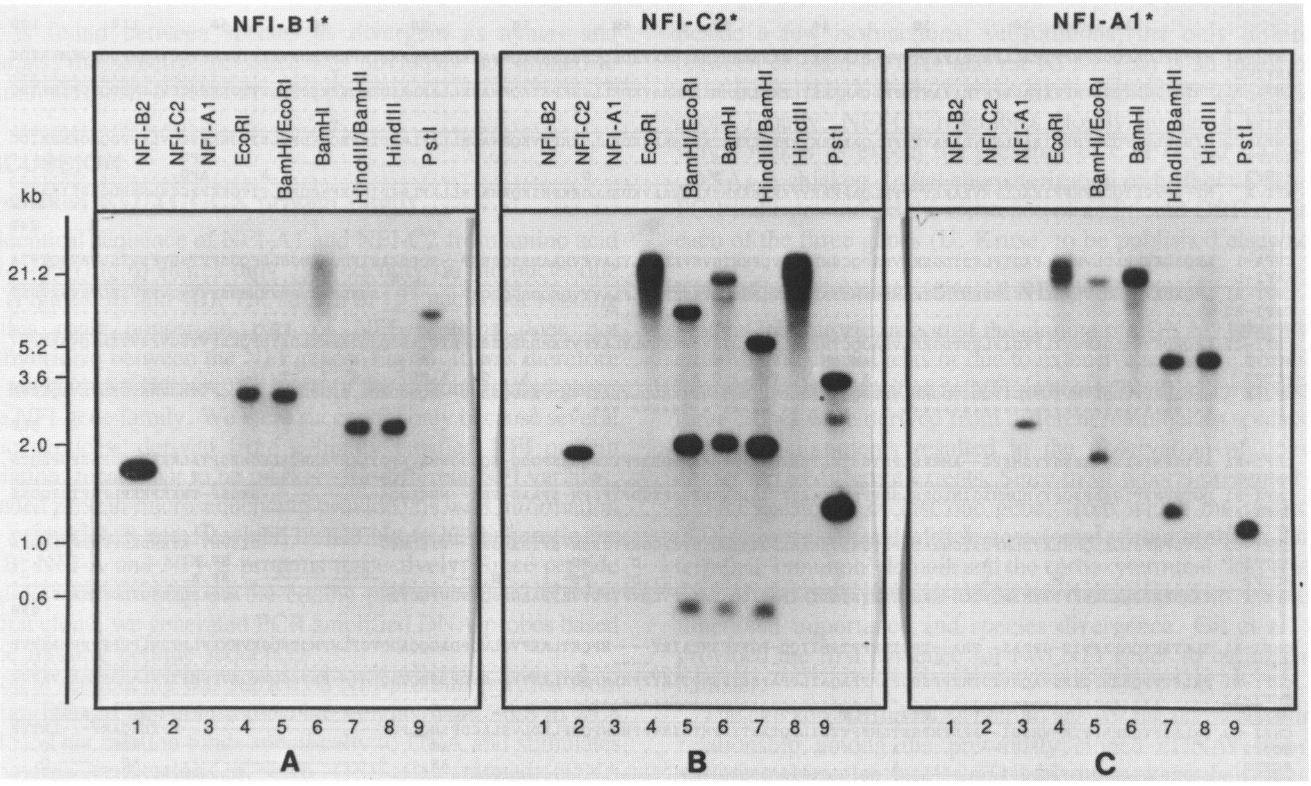


Figure 5. Genomic southern blot analysis. Duplicate filters carried 60pg of cDNA plasmids (lanes 1–3) or 10 μg genomic chicken DNA (lanes 4–9) per lane. Respective cDNAs and restriction endonucleases are named on top. Hybridizations were carried out under stringent conditions. Radioactive probes derived from the highly conserved N-terminal domains as indicated in Figures 2 to 4 are given above each panel.

questions concerning both their crossreactivity and the origin of the corresponding cDNAs, we performed genomic southern blot analyses.

The four TGGCA protein cDNAs are derived from three different genes

Sequence comparison of the different coding regions resulted in the identification of an aminoterminally located domain, which is extremely conserved among the cDNAs (see below). For southern blot analysis we used 477bp-long DNA-fragments, which were derived from corresponding positions within this domain of NFI-B1, -C2 and -A1, respectively. These probes were constructed by three separate PCR-amplifications with degenerate primers ON13 and ON21 (positions displayed in part B of Figures 2 to 4) and each of the three cDNAs as template. The three duplicate filters shown in Figure 5 reveal no crossreactivity between the three cDNA-types under stringent conditions. This is clearly seen in each set of control lanes 1–3 by the exclusive hybridization of each probe with the plasmid insert from which it has been derived. In respect to genomic DNA, the hybridization signals of each probe comprise a unique pattern of DNA-fragments. The patterns are completely consistent with the idea, that each probe recognizes a separate genomic locus. In addition, the NFI-C2 probe hybridized weakly to further DNA-fragments visible in BamHI- and PstI-digests (lanes 5–7 and 9 in Fig.5B). Since the size of these fragments does not correlate with any signal observed with the other two probes, it probably indicates the presence of a fourth independent locus, with which only the NFI-C2 probe crossreacts. In all other cases, where hybridization to two fragments per lane was observed (lanes 5, 7 and 9 in Fig. 5B; lanes 7 and 9 in Fig.5C), this is explained by restriction cuts within the probes (compare with restriction site maps in part A of Figures 2 to 4). We expected such a result also for hybridization of the NFI-A1 probe to HindIII-digested genomic DNA (Fig.5C, lanes 7 and 8) because of the presence of a HindIII-site at position 297 of λCB-43 (Fig.4). But neither in this blot, nor by additional analysis did we detect a second signal. It might therefore indicate a restriction fragment length polymorphism.

We conclude from these data that the family of chicken TGGCA proteins is encoded by at least three different genes. The weak crosshybridization of the NFI-C2 probe with an up to now unidentified fourth gene might further increase the genomic complexity. Despite almost identical amino acid sequences in the region covered by the probes, the NFI genes do not crosshybridize under stringent hybridization conditions. This clearly hampered the search for NFI gene variants.

Multiple Sequence Alignment of NFI-ORF types

The sequence alignment of a variety of NFI-cDNAs led to informations concerning common features of NFI proteins (19, 20, 16). In general, one observed a high similarity among aminoterminal sequences paired with low conservation in the carboxyterminal parts of the proteins. Since most of the cDNAs were isolated from different species, the evaluation of observed differences, however, was difficult. They could either indicate species-specific divergence of one common gene during evolution or represent differences between open reading frames of several related genes in each of the species.

When we aligned the coding regions of our three cDNA clones NFI-A1, -B1 and -C2 with each other, we detected primarily the same result. Their overall identity of 51–60% is mainly based

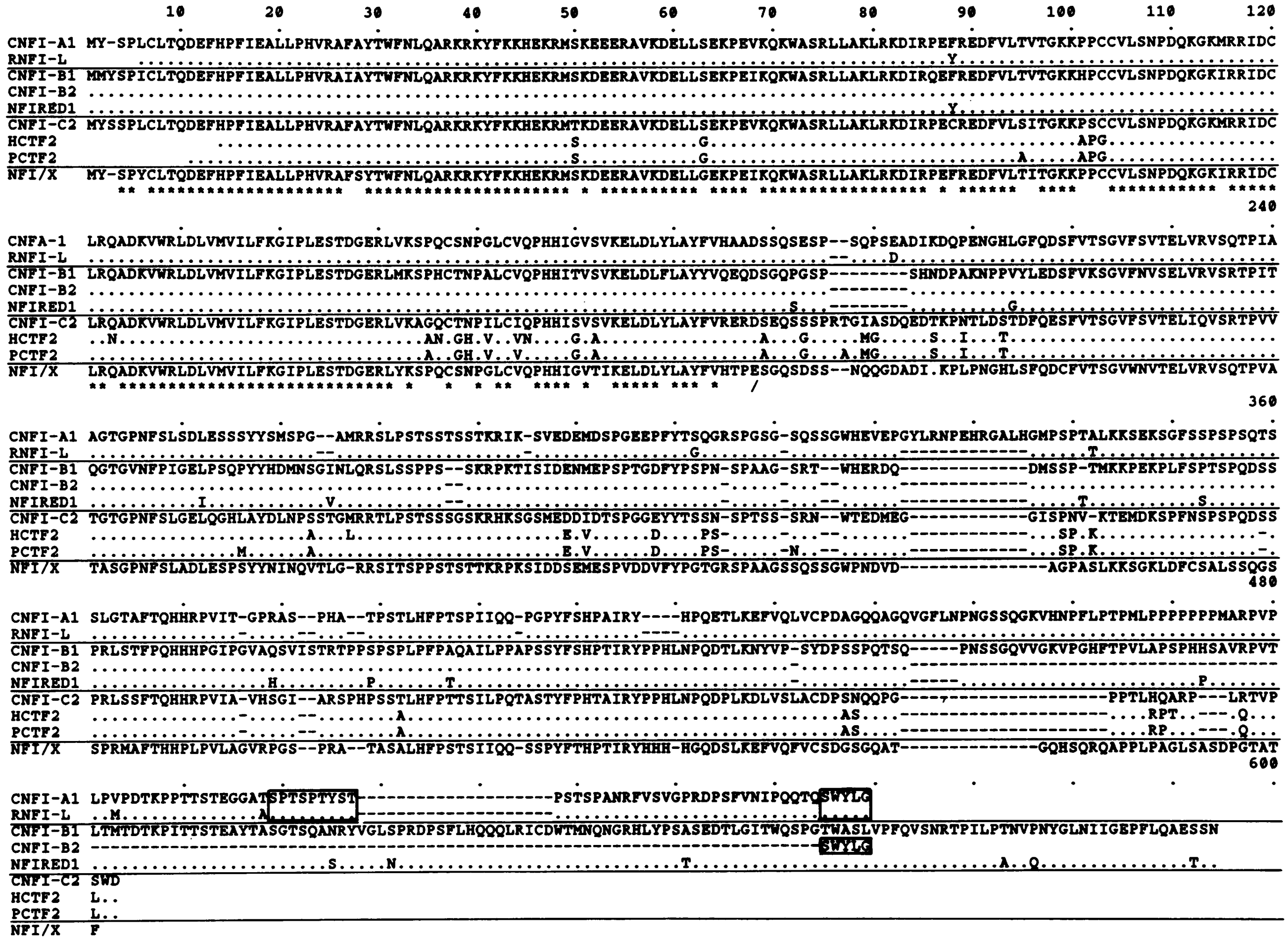


Figure 6. Sequence alignment of NFI/TGGCA cDNAs. Sequences were aligned according to Feng and Doolittle (41). Residue numbers refer to NFI-A1. Gaps introduced to maximize similarity are indicated by dashes. Within subgroups, which are separated by horizontal lines, dots represent identical amino acids in comparison to chicken NFI-A1, NFI-B1 and NFI-C2, respectively. Only amino acids differing from the respective chicken sequence are printed. NFI-prefices: C-chicken, R-rat; CTF-prefices: H-human, P-porcine. Asterisks indicate positions, which are identical among all cDNAs up to position 187 (marked by a slash). Boxes display short sequence motifs outlined in the text. The sequences were either deduced from Fig.2–4, or from NFI-L (20); NFI/RED1 and NFI/X (16); HCTF2 (19); PCTF2 (11).

upon the presence of a highly similar aminoterminal region showing 87–91% identity (per centages deduced from Fig.6). The C-terminal border of this region was set to position 187 (Fig.6) because it coincides with an exon-intron boundary found in the CTF-gene from porcine (11) and in the chicken NFI-B gene (F. Qian and A.E. Sippel, unpublished results).

A simultaneous sequence alignment (41) of all known NFI sequences, however, revealed that subgroups of cDNAs can be formed, in which the similarity between the members of each group is high throughout the whole coding region. To visualize this fact, we ordered the cDNAs in Fig.6 according to such a grouping. While displaying the complete sequences of NFI-A1, -B1, -C2 as references, only those amino acids are indicated, which are different from the corresponding chicken cDNA. CTF1 and CTF3 (19) were not included in this alignment, since none of our cDNA clones corresponds to these types. NFI-A1 displays 98.6% identity to rat NFI-L, NFI-B1 shares 97% identity with hamster NFI/RED1, and finally NFI-C2 is in 90%, respectively 91% of the positions identical with human or porcine NFI/CTF2. NFI/X most likely represents currently the only member of a fourth subgroup, since it fits into none of the previous groups (52–59% identity with NFI-A1, -B1 and -C2). This shows that within the subgroups there exists nearly the same high homology throughout the entire sequence than it was previously seen only in the first 187 amino acids (20, 16).

Some cDNAs (including human and porcine NFI/CTF1) share the short motif 'SWYLG' (boxed in Fig.6) at their carboxytermini. With the exception of NFI/X, the respective nucleotide sequence is found also within all other cDNAs near the ends of their open reading frames (NFI-B1, Fig.2B), or in the 3′-untranslated region (NFI-C2-like proteins) as the result of differential splicing (19). Interestingly, the chicken NFI-A1 clone contains an insert of 13 amino acids in comparison to rat NFI-L (pos.323–335, Fig.7), which might represent an additional exon. Finally, Meisterernst et al. (11) reported the presence of a single motif within CTF1-coding regions resembling a permutation of the (CT7)-repeat typical for the largest subunit of yeast and mouse RNA-polymeraseII (26). We note that also NFI-A1 and rat NFI-L cDNAs contain a similar repeat with one conservative substitution when compared to the motif found in porcine NFI/CTF1 (pos. 504, thr to ser, Fig.6).

Due to the outstanding conservation found among the members of each sub-group, we look at these cDNAs as products of homologous genes. Since the conservation throughout the entire

ORF is found between species as divergent as avians and mammals, the selective pressure was obviously not limited to the aminoterminal portion of the proteins.

DISCUSSION

The chicken NFI/TGGCA protein family

The identical sequence of NFI-A1 and NFI-C2 from amino acid 103 to 153 (Fig.6) shares only 79% identity on the nucleotide level (Figures 3B and 4B). The stretch of 153 nucleotides coding for this most conserved part of NFI proteins does not crosshybridize between the NFI genes (Fig.5). It was therefore not a trivial task to clone cDNAs from the various related genes of the NFI-gene family. We were successful only because several tryptic peptides, derived from a highly purified NFI protein population, turned out to be markers for different NFI variants. Extended protein microsequencing provided us with information from peptide 7, 8 and 10 which turned out to be diagnostic for NFI-B, NFI-A and NFI-C proteins respectively. Since peptide 8 and 10 were missing on the coding part of NFI-B2 cDNA, our first clone, we generated PCR amplified DNA-probes based on the missing peptide sequences.

Protein sequencing was started on NFI proteins purified from chicken liver of apparent molecular weights from 36.8 to 29.8 kd (15). This fraction binds specifically to DNA and stimulates Adenovirus replication in vitro (18). Our cloned cDNA sequences, however, carry open reading frames from 419 to 560 amino acids, corresponding to calculated protein molecular weights of 46 to 62 kd. Most likely the isolated proteins represent proteolytic products comprising the N-terminal part of various NFI protein species. This assumption is consistent with the facts that all 11 isolated tryptic peptides are contained within the N-terminal 247 amino acids and that the protein domain conferring specific DNA-binding activity and stimulation of Adenovirus replication were mapped by others within the N-terminal 220 residues of NFI/CTF proteins (10, 11). Genomic Southern blots were probed with fragments from the most highly conserved amino terminal region of the cDNAs (Figures 2, 3, 4 and 5). Since each of the three probes detects a unique set of genomic fragments (the weak hybridization of NFI-C2 with a second locus will be discussed later), we conclude that mRNAs must be derived from three independent transcriptional units. An alternative to this interpretation would be a situation in which multiple, non-crosshybridizing exons of a single gene region are spliced to yield completely exclusive cDNA-types. This is rather unlikely, especially since southern blot analysis of genomic clones covering most of the NFI-B gene revealed no hybridization of NFI-A1 or -C2 cDNAs to this locus (F. Qian and A.E. Sippel, unpublished results).

All three genes can be active in the same cell type. This is concluded from the presence of peptides encoded by different genes in liver, from the cloning of cNFI-A1 and cNFI-C2 from the same cDNA library prepared from a transformed myeloid cell line and from transcript analysis of a number of chicken cells and tissues (U. Kruse, F. Qian, R. Rupp, and A.E. Sippel, in preparation). There is good evidence that differential splicing produces the isoformes observed with NFI/CTF cDNAs (19, 11). We conclude from several observations that alternative splicing is not restricted to NFI/CTF, but is rather a common means among NFI genes to generate protein diversity . With NFI-B1 and -B2 we have cloned two different ORF-types from a gene, which is not the chicken homologue to the human NFI/CTF-gene. Beside a few isofunctional substitutions, the only difference between chicken NFI-A1 and rat NFI-L (20) is an insertion of 13 residues (Fig.6), which most likely represents an additional exon. Finally, NFI-C2 resembles closely human CTF2 (19), which makes us expect the presence of CTF1- and CTF3-like cDNAs in chicken. Initial characterization of further cDNAs via PCR indicates indeed the existence of additional ORF-types for each of the three genes (U. Kruse, to be published elsewhere).

Evolutionary conservation of NFI-genes

Several laboratories reported the cloning of cDNAs, which were either by functional tests or due to extensive sequence homology unambiguously identified as NFI proteins (19, 20, 11). However, these clones were derived from different mammalian species and sequence alignments resulted in the observation of domains conserved to different extents. Since most reports presented data for the existence of just one gene, from which the respective cDNAs were derived, differences found within both the aminoterminal 'common' domain and the carboxyterminal 'less related' part of the NFI proteins could not be assessed in respect to functional importance and species divergence. Gil et al. (16) provided the first evidence for two NFI genes to be present in hamster.

The chicken NFI/cDNAs help to answer the question of interrelationship among the previously cloned cDNAs from the various mammalian species and help to reevaluate their sequence divergence. Multiple sequence alignment including NFI types from mammalian species led to the surprising result that all known cDNAs fall into four distinct subgroups. As it is outlined in Fig.6, the members of each subgroup, irrespective of their source, share a significantly higher degree of amino acid identity than do cDNAs from one species belonging to different subgroups. While we have cloned representatives of three subgroups from chicken, for the moment the fourth NFI subgroup consists solely of hamster NFI/X (16). However, the weak crosshybridization of the NFI-C-specific Southern Blot probe to a second locus (Fig.5) might indicate the presence of a NFI/X related gene in chicken. This could be consistant with the independent observation from Meisterernst et al. (11), who found weak crossreaction of the porcine NFI/CTF probe with a second procine gene, whose partially determined exon sequences were suggested to belong to the NFI/X type.

All these data are consistent with the assumption that the NFI protein family in vertebrates is encoded by a set of at least 3, most likely of 4 different genes, which have been highly conserved. On the one hand, the considerable divergence on the nucleotide level among even the most conserved parts of the coding regions of NFI-A, -B or -C genes from chicken (i.e. 75% identity, pos.1−187, Fig.6) suggests an early separation of these genes in evolution. On the other hand, the extensive amino acid similarity found throughout the complete coding regions between the same gene type in different species indicates a uniform selection pressure on all protein parts, also the the corresponding parts outside the N-termi-nal DNA-binding domain.

NFI protein diversity

The characterization of eukaryotic transcription factors has led to the identification of several protein families, whose members share an identical or a very similar DNA-binding activity, for example POU-proteins (27) or AP-1 like proteins (28). In case of the NFI-family, protein diversity is set up by many different means. Three, most likely more genes, give rise each to differentially spliced mRNAs, which in turn can give rise to protein variants with posttranslational secondary modifications, e.g. glycosylation (29) or phosphorylation. Since we have indications for the presence of products of different NFI genes in the same cell, heterodimerization could further increase NFI diversity to unexpectedly high levels of variants.

The tight conservation of homologues in the various species argues for a scenario in which cells establish by differential synthesis and by balanced concentrations of each member a distinct NFI-milieu through which these proteins take part in the regulation of gene expression and perhaps cellular replication.

Proposal for a NFI nomenclature

The cloning of cDNAs finally proved the identity of proteins, whose names 'NFI' (12), 'TGGCA Protein' (6) and 'CTF' (17) can now be looked at as synonyms. Since there is evidence for a NFI gene family conserved among higher eukaryotes, whose transcripts are additionally subjected to differential splicing, we are facing a puzzling complexity of NFI proteins. It seems therefore helpful to simplify the NFI nomenclature. We propose the following system, according to which we have already named our cDNAs.

We have chosen NFI to be the 'family name'. It is widely accepted, has the right of priority and is connected with Adenovirus replication. The name 'CCAAT-binding transcription factor' CTF is complicated by the fact that there is clear evidence for a variety of transcription factors binding to CCAAT-motifs (30), some of which definitely are genetically unlinked to the NFI protein family, e.g. C/EBP (31). Finally, our own functional designation 'TGGCA protein' was not commonly accepted and was more or less used for chicken NFI proteins only. The species origin could be indicated by NFI-prefices, c for chicken, r for rat, h for human (Fig.6).

We have named the genes A to C, trying either to conserve a connection with present names or to follow the sequence of publication. The NFI/CTF gene, from which the first cDNAs were cloned (19), is therefore named NFI-C. The gene, which gives rise to rat NFI-L (20) is named NFI-A, while the gene encoding hamster NFI/Red1 (16) is called NFI-B. Any gene detected in future can now simply be added to the end of this list. We further suggest NFI-X to be the name for the gene, to which hamster NFI/X (16) belongs. The various splice products from individual genes could be numbered according to their size, if known, or according to their appearance (Fig.6).

ACKNOWLEDGEMENTS

We are grateful to B. Vennström for the gift of the chicken embryo cDNA library, to M. Nassal for the synthesis of oligonucleotides, to G. Pecht for technical assistance and to E. Koepff for typing the manuscript. U.K. is a fellow of the Studienstiftung des deutschen Volkes. Sections of the work were done in partial fulfilment of the thesis requirements for U.K. The research was supported by grants from the Bundesministerium für Forschung und Technologie (BCT 0381/5) and the Fonds der Chemischen Industrie to A.E.S. and K.B.

REFERENCES

1. Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) *Science* **236**, 1237–1245.
2. Mitchell, J.P. and Tjian, R. (1989) *Science* **245**, 371–378.
3. Ondek, B., Gloss, L. and Herr, W. (1988) *Nature* **333**, 40–45.
4. Fromental, C., Kanno, M., Nomiyama, H. and Chambon, P. (1988) *Cell* **54**, 943–953.
5. Theisen, M., Stief, A. and Sippel, A.E. 81986) *EMBO J.* **5**, 719–724.
6. Borgmeyer, U., Nowock, J. and Sippel, A.E. (1984) *Nucleic Acids Res.* **12**, 4295–4311.
7. Nowock, J. and Sippel, A.E. (1982) *Cell* **30**, 607–615.
8. Nowock, J., Borgmeyer, U., Püschel, A.W., Rupp, R.A.W. and Sippel, A.E. (1985) *Nucleic Acids Res.* **13**, 2045–2061.
9. Rupp, R.A.W., Nicholas, R.H., Borgmeyer, U., Lobanenkow, V.V., Plumb, M.A. Sippel, A.E. and Goodwin, G.H. (1988) *Eur. J. Biochem.* **177**, 505–511.
10. Mermod, N., O'Neill, E.A., Kelly, T.J. and Tjian, R. (1989) *Cell* **58**, 741–753.
11. Meisterernst, M., Rogge, L., Foeckler, R., Karaghiosoff, M. and Winnacker, E.-L. (1989) *Biochem.* **28**, 8191–8200.
12. Nagata, K., Guggenheimer, R.A. and Hurwitz, J. (1983) *Proc. Nat. Acad. Sci. USA* **80**, 6177–6181.
13. Rosenfeld, P.J. and Kelly, T.J. (1986) *J. Biol. Chem.* **261**, 1398–1408.
14. Jones, K.A., Kadonaga, J.T., Rosenfeld, P.J., Kelly, T.J. and Tjian, R. (1987) *Cell* **48**, 79–89.
15. Rupp, R.A.W. and Sippel, A.E. (1987) *Nucleic Acids Res.* **15**, 9707–9726.
16. Gil, G., Smith, J.R., Goldstein, J.L., Slaughter, C.A., Orth, K., Brown, M.S. and Osborne, T.F. (1988) *Proc. Nat. Acad. Sci. USA* **85**, 8963–8967.
17. Jones, K.A., Yamamoto, K.R. and Tjian, R. (1985) *Cell* **42**, 559–572.
18. Leegwater, P.A.J., van der Vliet, P.C., Rupp, R.A.W., Nowock, J. and Sippel, A.E. (1986) *EMBO J.* **5**, 381–386.
19. Santoro, C., Mermod, N., Andrews, P.C. and Tjian, R. (1988) *Nature* **334**, 218–224.
20. Paonessa, G., Gounari, F., Frank, R. and Cortese, R. (1988) *EMBO J.* **7**, 3115–3123.
21. Maruyama, T., Gojobori, T., Aota, S. and Ikemura, T. (1986) *Nucleic Acids Res.* **14**, r151–r190.
22. Lathe, R. (1985) *J. Mol. Biol.* **183**, 1–12.
23. Ohtsuka, E., Matshuki, S., Ikehara, M., Takahashi, Y. and Matsubara, K. (1985) *J. Biol. Chem.* **250**, 2605–2608.
24. Cleveland, D.W., Lopata, M.A., MacDonald, R.J., Cowan, N.J., Rutter, W.J. and Kirschner, M.W. (1980) *Cell* **20**, 95–105.
25. Fort, Ph., Marty, L., Piechaczyk, M., El Sabrouty, S., Dani, Ch., Chanteur, P.H. and Blanchard, J.M. (1985) *Nucleic Acids Res.* **13**, 1431–1442.
26. Sigler, P.B. (1988) *Nature* **333**, 210–212.
27. Robertson, M. (1988) *Nature* **336**, 522–524.
28. Kouzarides, T. and Ziff, E. (1989) *Cancer Cells* **1**, 7–76.
29. Jackson, S.P. and Tjian, R. (1988) *Cell* **55**, 125–133.
30. Dorn, A., Bollekens, J., Staub, A., Benoist, C. and Mathis, D. (1987) *Cell* **50**, 863–872.
31. Landschulz, W.H., Johnson, P.F., Adashi, E.Y., Graves, B.J. and McKnight, S.L. (1988) *Genes Dev.* **2**, 786–800.
32. Moscovici, C., Zeller, N. and Moscovici, M.G. (1982) Expression of differentiated function in cancer cells. Raven Press New York, pp 435–449.
33. Sap, J., Munoz, A., Damm, K., Goldberg, Y., Ghysdael, J., Leutz, A., Beug, H. and Vennström, B. (1986) *Nature* **324**, 635–640.
34. Laemmli, U.K. (1970) *Nature* **227**, 680–685.
35. Konigsberg, W.N. and Henderson, L. (1983) *Meth. Enzymol.* **91**, 224–236.
36. Hirs, C.H.W. (1967) *Meth. Enzymol.* **11**, 197–199.
37. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, J.T., Mullis, B. and Ehrlich, H.A. (1988) *Science* **239**, 487–491.
38. Friedmann, K.D., Rosen, N.L., Newmann, P.J. and Montgomery, R.R. (1988) *Nucleic Acids Res.* **16**, 8718.
39. Sanger, F. and Coulson, A. (1975) *J. Mol. Biol.* **94**, 441–448.
40. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York
41. Feng, D. and Doolittle, R.F. (1987) *J. Mol. Evol.* **25**, 351–360.