J Mol Evol (1995) 40:570–577

JOURNAL OF MOLECULAR EVOLUTION

Investigating Hypothetical Products from Noncoding Frames (HyPNoFs)

Antonio Facchiano

Raggio Italgene S.p.A., via delle Antille 29, 00040 Pomezia, Roma, Italy

Received: 6 May 1994 / Accepted: 7 November 1994

Abstract. Hypothetical Products from Noncoding Frames (i.e., HyPNoFs) are hypothetical, not-coded proteins, translated from alternate reading frames (i.e., coding+1 and coding+2) of cDNAs. HyPNoFs of CD4, PKC, oncostatin, bcl-2 proto-oncogene, tumor suppressor p53, cystic fibrosis transmembrane regulator (CFTR), and tumor necrosis factors α and β were searched as query sequences vs the SWISS-PROT data bank. Homology searches carried out revealed that hypothetical products (i.e., HyPNoFs) may share high similarity with real protein products actually coded. Sequence similarity of hypothetical products to real proteins is sometimes very high, suggesting common conformational features, according to the Sander and Schneider cutoff value. This finding supports the hypothesis that eukaryotic DNA, currently considered to be monocistronic, might occasionally have polycistronic regions, carrying different protein messages on overlapping frames. As yet, polycistronic genes have been observed in viral genomes only. The presence of polycistronic regions in eukaryotic genes is likely reminiscent of an ancient strategy, rather than a present feature of the genome in eukaryotes.

These data suggest that thorough investigation of HyPNoFs is likely to improve our ability to trace genes' evolution and to investigate structure-function relationships of protein and DNA sequences.

Key words: Alternate reading frames — Evolution — Overlapping frames — Homology search — Primary sequence analysis — Polycistronic genes

Introduction

cDNA of eukaryotes is commonly reported to be monocistronic—i.e., the only protein message associated to a given cDNA is the one carried on the coding-reading frame. Polycistronic genes have been reported in viral genomes, where different protein messages may lie in-phase side by side, or alternatively, on overlapping frames (Dinesh-Kumar et al. 1992; Jacks et al. 1988; Kirchner et al. 1992; Lamb and Choppin 1979; Lamb and Horvath 1991; Montell et al. 1982; Ray et al. 1992; Suzuki et al. 1992). The benefit of carrying different genes on overlapping frames is twofold: (1) it is a useful strategy to save space in microorganisms with a short genome, like viruses and (2) it is a way to keep the expression of the overlapping proteins under the same regulatory control. The latter might be useful in viral as well as nonviral genomes.

We recently reported that the protein product of a noncoding (overlapping) frame of human CD4 gene shares similarity with HIV protein gp41. This was, to our knowledge, the first report indicating significant similarity between a real viral protein and a not-coded (HyPNoFs) protein from eukariotes. On this basis, an evolutionary divergent link between immunoglobulins and HIV-env genes was suggested (Facchiano et al. 1993).

In the present study the hypothesis is proposed that polycistronic genes might be present in the evolutionary history of eukaryotic genes—likely as an ancient strategy currently not active. A novel approach was followed to perform comparison analyses and homology searches of protein sequences. The classic approach consists of aligning query, "real" gene products to target sequences, the products of noncoding frames not usually being analyzed or only investigated as target sequences (Needleman and Wunsch 1970; Gotoh 1982; Wilbur and Lipman 1983; Lipman and Pearson 1985; Myers and Miller 1988; Green et al. 1993; Karlin and Brendel 1992; Gonnet et al. 1992; Sali et al. 1990; Brendel et al. 1992). In the current study, products of noncoding frames, i.e.,

Table 1. Significant homologies of CD4 HyPNoFs to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
CD4 RF2 shows similarity to:				
HLA class I histocompat. antigen B-62	55	20	103	P30513
Let-23 receptor tyrosine kinase	29	59	102	P24348
Plectin	27	80	102	P30427
CD4 RF3 shows similarity to:				
Major centromere antigene	24	88	132	P27790
Circumsporozoite protein	26	125	101	P08672
Nonstructural protein, Ross river virus	28	74	93	P13888

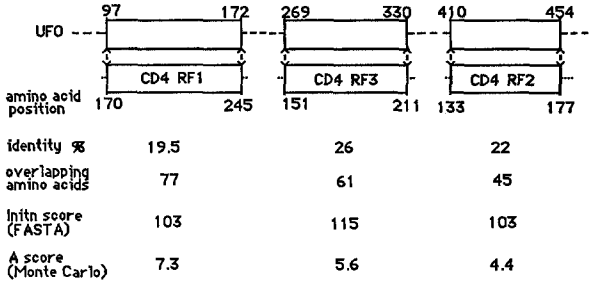

Fig. 1. Schematic representation of protein alignment between different portions of receptor tyrosine-protein kinase UFO precursor (E.C. 2.7.1.112., SWISS-PROT accession number P30530) and the products of reading frames 1, 2, and 3 (i.e., coding, coding+1, and coding+2, respectively) of the 5′ end portion of the human CD4 gene. Alignments were achieved without any gap insertion. *Numbers* above and below *boxes* indicate position on the corresponding protein sequences. FASTA and Monte Carlo analyses were carried out as reported in the Materials and Methods section.

hypothetical, not-coded sequences translated from coding+1 and coding+2 frames, were taken as query sequences and searched for any similarity to coded products listed in SWISS-PROT data bank. The presence of significant homology between products of coding and noncoding frames indicates that eukaryotic DNA may occasionally carry protein signals in overlapping frames.

Materials and Methods

A number of widely investigated genes were randomly selected from GenBank: human CD4, human protein kinase C (PKC), human oncostatin, human bcl-2 proto-oncogene, human p53 gene, human cystic fibrosis transmembrane conductance regulator (CFTR), human tumor necrosis factor alpha (TNFα), and mouse tumor necrosis factor beta (TNFβ). DNA sequences used for the current study code the real mature proteins on reading frame 1 (RF1) from the first to the last codon. Hypothetical protein sequences were obtained by translating DNA sequences in RF2 and RF3 (i.e., coding+1 and coding+2, respectively) by means of GENEPRO software (by Riverside Scientific Enterprises, Bainbridge Island, Washington) and the TRANSLATE routine from the GCG package; these hypothetical sequences are here referred to as HyPNoFs (hypothetical products of noncoding frames). Homology searches of HyPNoFs were carried out by means of the FASTA routine, from the GCG package. Stop-codon signals were eliminated by default; ktup was set at 1, and scores were computed according to the Dayhoff PAM 250 matrix (Dayhoff 1978). Searches were carried out vs the SWISS-PROT data bank, which contains 31,808 sequences. From the top hundred alignments (sorted by Initn score) of each search, the ones falling above the significance threshold computed according to Sander and Schneider (1991) were selected. They are here reported in Tables 1–8. Sander and Schneider recently reported a threshold of identity percentage as a function of the number of residues aligned. According to their analyses, identity percentage above this threshold infers a similar conformation of the aligned sequences. Homologies to "probable" proteins and to repetitive strands (from collagen, sulphated surface glycoproteins, and proline-rich proteins) are not reported here.

A-scores reported in Figs. 1–3 refer to the statistical significance evaluated by Monte Carlo analysis, implemented on PCGENE software (by Intelligenetics, Mountain View, California). Briefly, the homology score was calculated of a pair of real sequences identified by means of the FASTA routine. Then, 500 pairs of corresponding scrambled sequences were aligned and the mean homology score and standard deviation were calculated. Nonrandom relatedness and high probability of common ancestry are assumed when the homology score of the real sequences exceeds by a factor of at least 5 standard deviation units the mean homology of the scrambled sequences. Alignments with A-score between 4 and 5 were also reported, as this range was considered possibly significant (Doolittle 1981; Feng et al. 1984; Lipman and Pearson 1985).

Results

Sequence homologies listed in Tables 1–8 fall above the significance threshold computed by Sander and Schneider (Sander and Schneider 1991) and were selected from the top 100 alignments found by each FASTA search. Few additional alignments selected from the top 100 scores of each search are reported in the text. As opposed to those reported in Tables 1–8, they fall below the Sander-Schneider cutoff point.

Comparison of CD4 HyPNoFs to SWISS-PROT Data Bank

The product of the CD4 gene is an immunoglobulin-family member expressed on the membrane of a subset of T-lymphocytes. It has been identified as the receptor of the env protein of human immunodeficiency retrovirus HIV (Dalgleish et al. 1984). HyPNoFs of CD4 share a high identity percentage with at least six proteins present in SWISS-PROT (Table 1). Some are functionally related to the real product of CD4, i.e., HLA class I histocompatibility antigen (which belongs to the immu-

Table 2. Significant homologies of PKC HyPNoFs to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
PKC RF2 shows similarity to:				
None				
PKC RF3 shows similarity to:				
Negative regulator of mitosis	24	90	120	P24686
Cardiac-muscle calcium channel	30	84	110	P15381

noglobulin superfamily, like CD4) and LET-23 receptor tyrosine kinase. Notice that CD4 is known to associate to an intracellular p56-lck tyrosine kinase. In addition, CD4 RF2 aligns with the highest Initn score and the longest amino acid stretch (i.e., 22% identity over 158 residues, with an Initn score of 147) to env protein of murine leukemia virus (data not shown). Regions of CD4 RF1 (i.e., the real product) as well as CD4 RF2 and CD4 RF3 products share high similarity with the receptor of tyrosine-protein kinase UFO (19.5% identity over 77 residues, 22% identity over 45 residues, and 26% identity over 61 residues, respectively), as summarized in Fig. 1. Although alignments do not exceed the Sander and Schneider cutoff point, A-scores computed by means of Monte Carlo analysis indicate high statistical significance in all cases (Fig. 1), suggesting a possible common ancestry. Significance is reinforced by two additional observations: (1) alignments reported in Fig. 1 were achieved without any gap insertion (data not shown), and (2) scrambled versions of CD4 RF1, CD4 RF2, and CD4 RF3 do not show any significant similarity to tyrosine-protein kinase UFO.

Comparison of PKC HyPNoFs to SWISS-PROT Data Bank

Table 2 indicates that HyPNoFs of PKC show similarities above the Sander and Schneider (1991) cutoff value in only two cases, whereas many of the top hundred alignments fall slightly below the threshold. They are membrane-bound proteins or receptors, i.e., brain calcium channel, inositol-triphosphate binding type 2 receptor, hemagglutinin of influenza A virus (24% over 54 residues, 23% over 60 residues, and 25% over 72 residues, respectively, identical to PKC RF2), calcitonin receptor and α 2B adrenergic receptors (22% over 88 residues and 22% over 58 residues, respectively identical to PKC RF3).

FASTA analysis identified high local homologies of cardiac-muscle calcium channel with both PKC RF2 and PKC RF3 products. This similarity is summarized in Fig. 2. In two cases Monte Carlo A-scores are significantly (i.e., 8.5 and 11.5) higher than the accepted cutoff value. In addition, scrambled PKC RF2 and scrambled PKC

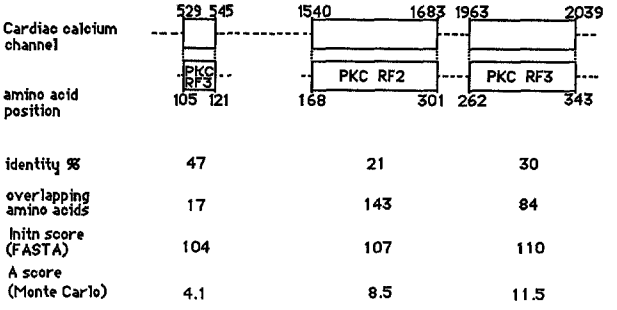

Fig. 2. Schematic representation of protein alignment between different portions of dihydropyridine-sensitive L-type cardiac muscle calcium channel (SWISS-PROT accession number P15381) and products of reading frames 2 and 3 (i.e., coding+1 and coding+2) of a central portion of the human PKC gene. *Numbers* above and below *boxes* indicate position on the corresponding protein sequences.

RF3 do not show any similarity to cardiac-muscle calcium channel. These data suggest very unlikely random similarity and possibly common ancestry for PKC and cardiac-muscle calcium channel.

Comparison of Oncostatin's HyPNoFs to SWISS-PROT Data Bank

Oncostatin is a glycoprotein expressed by stimulated macrophages and T-cells. It is known to inhibit in vitro proliferation of tumor cell lines. It is a cytokine with structural similarity and partial overlapping activity with cytokines such as leukemia inhibitor factor (LIF), interleukin-6, and granulocyte colony-stimulating factor (Rose and Bruce 1991). The similarity reported in Table 3 to interleukin-2-receptor and plasminogen, which are functionally related to oncostatin (Wegenka et al. 1993; Taga et al. 1992), is remarkable. Pecanex is a *Drosophila* neurogenic gene (Gilbert et al. 1992) sharing high similarity with oncostatin RF2 product (see Table 3). Functional relation might underlie the structural similarity observed, according to the neurotrophic role suggested for oncostatin (Taga et al.). Transcription/replication regulators such as DNA binding protein and transcription regulator IE63 are also similar to HyPNoFs of oncostatin (24% over 63 residues and 22% over 151 residues, respectively, identical to oncostatin RF3, i.e., slightly below the Sander and Schneider [1991] threshold).

Table 3. Significant homologies of oncostatin HyPNoFs to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
Oncostatin RF2 shows similarity to:				
CAD protein	39	43	116	P05990
Plasminogen	24	122	115	P12545
Pecanex	31	58	106	P18490
Protein glut.-gamma-glutamyltransf. K	24	79	93	P22735
Oncostatin RF3 shows similarity to:				
HSV protein UL8	24	98	103	P10192
Tyrosine-kinase-related protein	37	49	88	P14083
NIF-specific regulatory protein	28	110	85	P30667
Interleukin-2 receptor	26	89	82	P26898

Comparison of bcl-2 HyPNoFs to SWISS-PROT Data Bank

bcl-2 is a membrane-bound protein whose role in regulating apoptosis has been reported (Hockenbery et al. 1990). Its biochemical function has not been clarified yet, although tyrosine-kinase activity has been associated to its action. HyPNoFs of bcl-2 show a large number of significant similarities to SWISS-PROT (Table 4). Many sequences similar to bcl-2 HyPNoFs are nucleotide-interacting factors, i.e., transacting transcription protein ICP0, DNA cytosine methyltransferase (23% over 137 residues and 24% over 92 residues, respectively, identical to bcl-2 RF2), and alkaline exonuclease (25% over 131 residues identical to bcl-2 RF3).

Comparison of p53 HyPNoFs to SWISS-PROT Data Bank

The product of the p53 gene is a tumor suppressor, DNA binding protein with kinase activity. RF2 and RF3 products of the p53 gene show homology with a variety of sequences (Table 5), including kinase-related proteins such as FPS tyrosine kinase transforming protein, metabotropic glutamate receptor 2 (24% over 50 residues and 26% over 68 residues, respectively, identical to p53 RF2; i.e., slightly below the Sander and Schneider threshold), kinase-related transforming protein DRAF-1, transcription regulatory protein XYLR (20% over 105 residues and 27% over 59 residues, respectively, identical to p53 RF3), or electron transport proteins such as monophenol monooxygenase and benzenediol oxygen oxidoreductase (21% over 82 residues identical to p53 RF2) and L-amino acid oxidase (24% over 90 residues identical to p53 RF3).

Comparison of HyPNoFs of CFTR Gene (Cystic Fibrosis Transmembrane Conductance Regulator) to SWISS-PROT Data Bank

CFTR is a chloride channel. It belongs to an ATP binding protein superfamily with transport/channel function (Riordan et al. 1989; Rich et al. 1990). Two sequences share similarity above the Sander and Schneider threshold (Table 6). However, HyPNoFs of CFTR share similarity to a larger variety of sequences functionally related to the real product of the CFTR gene, including NF-1 protein, which is related to the neurofibromatosis (Table 6); transport/channel proteins such as SEC-2, chromate transport protein (21% over 85 residues and 19% over 185 residues, respectively, identical to CFTR RF2); transport protein YGLO22; brain calcium channel B II (25% over 52 residues and 27% over 51 residues, respectively, identical to CFTR RF3); nucleotide-interacting proteins such as DNA-directed RNA polymerases (18% over 444 residues and 19% over 131 residues identical to CFTR RF2 and RF3 products); and RNA polymerase and adenylate cyclase type IV (24% over 63 residues and 21% over 72 residues, respectively, identical to CFTR RF2 and RF3 products).

Comparison of HyPNoFs of Tumor Necrosis Factor α and β Genes (TNF-α and TNF-β) to SWISS-PROT Data Bank

HyPNoFs of human TNFα and mouse TNFβ show similarity to a number of sequences in SWISS-PROT (Tables 7 and 8). Both HyPNoFs of TNFβ show high similarity to different portions of the laminin B2 chain, as summarized in Fig. 3. Monte Carlo A-scores show high statistical significance, and scrambled TNF-β RF2 and scrambled TNF-β RF3 do not show any similarity to the laminin B2 chain. This suggests a very unlikely random similarity and a possible common ancestry between TNF-β and the laminin B2 chain. Alignment showing high A-score (i.e., 11) is reported in Fig. 4. Besides the high identity percentage (29%), significance of the alignment is reinforced by the very few gaps inserted and by the matching of four cystein residues and one triptophan residue.

Discussion

Data reported here indicate that a number of hypothetical sequences translated from coding+1 and coding+2 frames of eukaryotic genes share significant similarity with real sequences present in SWISS-PROT data bank. This analysis was carried out on hypothetical sequences

Table 4. Significant homologies of bcl-2 HyPNoFs to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
BCL-2 RF2 shows similarity to:				
DNA cytosine methyltransferase	24	92	120	P13864
Structural polyprotein	26	113	113	P21480
Dynein-associated polypeptide	26	76	105	P13496
Plectin	26	85	105	P30427
Folate binding protein	28	71	98	P15328
Lysosomal α-glucosidase precursor	24	89	96	P10253
BCL-2 RF3 shows similarity to:				
α 2A adrenergic receptor	29	131	104	P08913
Wilm's tumor protein	28	58	100	P22561
Retinoic acid receptor RXR	25	106	98	P28703
Glycoprotein GI precursor	28	85	92	P08354
Capsid assembly and DNA maturation prot.	32	47	91	P10222
Neurovirulence factor	29	83	90	P28283
Acetylcholinesterase	35	65	87	P23795
Alkaline exonuclease	25	131	87	P06489
Adrenal-specific 30 KD protein	25	126	83	P15803
T-cell-surface CD5 glycoprotein	25	71	83	P13379

Table 5. Significant homologies of p53 HyPNoFs to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
p53 RF2 shows similarity to:				
General secretion pathway prot. L	26	95	100	P31707
Monophenol monooxygenase (tyrosinase)	30	69	99	P06845
Phosphoenol pyruvate carboxylase	25	87	99	P30694
HSV large-tegument protein	29	59	98	P10220
p53 RF3 shows similarity to:				
HCV genome polyprotein	27	77	117	P27958
HSV gene 68 protein	31	98	116	P28964
L-amino acid oxydase	24	90	111	P23623
Clathrin-coated vesicle prot. pump	29	63	109	P25286

Table 6. Significant homologies of HyPNoFs cystic fibrosis transmembrane regulator (CFTR) to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
CFTR RF2 shows similarity to:				
Photosystem I P700 chlorophyll A prot.	26	98	138	P29254
Neurofibromatosis-related NF-1 protein	24	82	124	P21359
CFTR RF3 shows similarity to:				
None				

translated from noncoding frames of CD4, PKC, oncostatin, bcl-2, p53, CFTR a, TNFα, and TNFβ genes. A number of alignments fall on—or well above—the Sander and Schneider threshold. This cutoff value is usually considered an indication of overall structural similarity and might indicate common ancestry. In the current study, the finding that HyPNoFs significantly share conformational features with real proteins was interpreted to be biologically meaningful and as suggesting the presence of polycistronic regions on eukaryotic genes.

Polycistronic genes, coding for different proteins on overlapping frames, have been already found in viruses and retroviruses. Data reported here suggest that polycistronic regions might be present in eukaryotes genes, too. Three examples are reported in detail (Figs. 1–3). Tyrosine-protein kinase UFO appears to be closely re-

Table 7. Significant homologies of HyPNoFs tumor necrosis factor α (TNF-α) to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
TNF-α RF2 shows similarity to:				
ABL proto-oncogene tyrosine kinase	25	101	99	P00520
Homoserine acetyltransferase	41	56	98	P12917
Endoglucanase A precursor	25	103	92	P22533
Lymphocyte activation gene-3 protein	25	76	89	P18627
TNF-α RF3 shows similarity to:				
Structural polyprotein	33	51	107	P075566
Lα-aminoadipyl-*l*-cysteinil-*d*-valine synthetase	28	81	106	P25464
HUPK protein	35	86	93	P30797

Table 8. Significant homologies of HyPNoFs tumor necrosis factor β (TNF-β) to SWISS-PROT (similarity above the Sander and Schneider threshold)

	Identity (%)	No. of amino acids aligned	Initn score	Accession number
TNF-β RF2 shows similarity to:				
Inhibin α-chain precursor	38	42	87	P17490
Homeobox protein HOX-7.1	28	72	75	P13297
Phosphorous acquisition-controlling protein	31	71	71	P20824
70-kD protein	25	91	71	P20129
TNF-β RF3 shows similarity to:				
Laminin B2-chain precursor	29	79	94	P15215
DNA-directed RNA polym. I large subunit	24	82	87	P16355
NAD(P) transhydrogenase, mithoc. prec.	26	87	81	P11024
Epidermal growth factor receptor	43	30	81	P04412

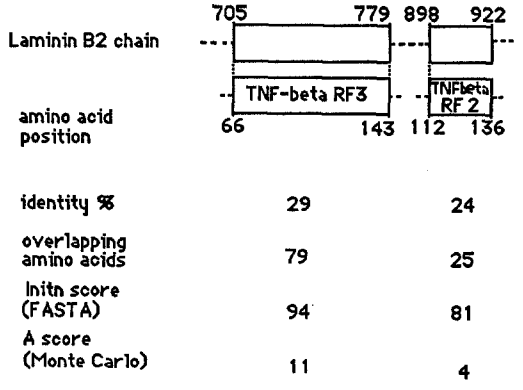

Fig. 3. Schematic representation of protein alignment between different portions of drosophila and mouse laminin B2 chain (SWISS-PROT accession number P15215) to HyPNoFs of the mouse tumor necrosis factor β gene. *Numbers* above and below *boxes* indicate position on the corresponding protein sequences.

lated to the human CD4 gene. Three regions of UFO protein show significant similarity to products of overlapping frames of the same stretch of CD4 gene (Fig. 1). This may indicate duplication of portions of a CD4 gene ancestor and following assembly in different frames, giving origin to different regions of UFO gene ancestor. Similarity reported in Fig. 2 suggests duplication and translation in different frames of stretches of PKC gene ancestor, originating different portions of the cardiac-muscle calcium channel. Finally, duplication of stretches of tumor necrosis factor β ancestor and translation in different reading frames might have originated different portions of laminin B2 chain (Fig. 3). The high statistical significance, computed according to Monte Carlo analysis, indicates that structural relations reported in Figs. 1–3 are very unlikely due to a random phenomenon (P estimated <0.00001).

The observed sequence similarities are supported by functional relations in many cases. In fact, cDNA coding for mature CD4 on RF1 carries protein messages on RF2 and RF3 which are similar to proteins functionally related to CD4, i.e., retrovirus env protein, HLA class I histocompatibility antigen, and tyrosine kinases. Similarity of CD4 RF2 to the env protein of murine leukemia retrovirus falls below the Sander and Schneider threshold (see Results section), but it is consistent with our previous report on a possible evolutionary-divergent link between env retrovirus gene and immunoglobulins (Facchiano et al. 1993). Further, cDNA coding for PKC on RF1 codes on RF2 and RF3 for sequences similar to

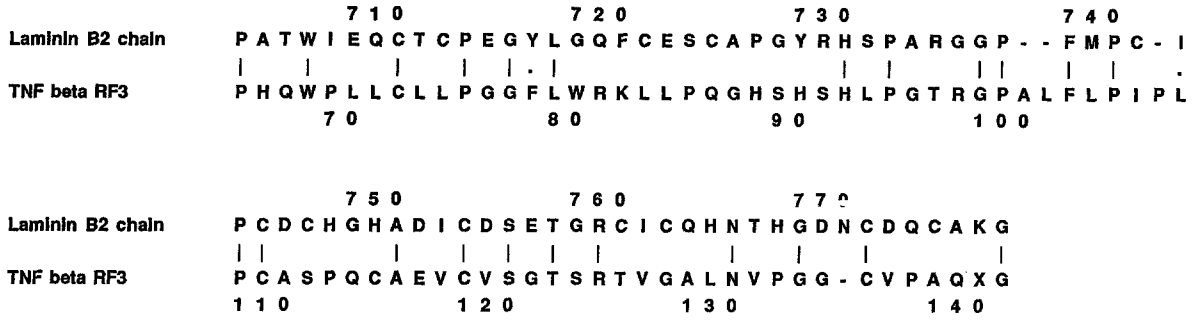

Fig. 4. Alignment of fragment 66–143 of TNFβ RF3 product to *Drosophila melanogaster* laminin B2 chain, performed by means of FASTA routine. Corresponding identity percentage and Monte Carlo A-score are shown in Fig. 3. TNFβ RF3 region reported in the alignment does not contain any stop codon, except in position 142 (i.e., the penultimate position), indicated by an *X* symbol.

kinases, calcium channels, and receptors. The structural similarity of PKC HyPNoFs to cardiac calcium channel (summarized in Fig. 2) may underlie recent data indicating PKC to be an inhibitor of cardiac calcium-channel activity (Merelli et al. 1992). RF2 and RF3 products of CFTR cDNA code for sequences similar to a number of proteins with transport/channel activity (SEC2, chromate transport, brain calcium channel, transport protein YGLO22), and CFTR RF2 is significantly similar to NF-1 protein, which is related to neurofibromatosis. These sequences appear to be functionally related to the real product of the CFTR gene, which is an ATP binding membrane protein with chloride transport activity. On this basis, analysis of bcl-2 HyPNoFs was particularly interesting. The biochemical role of bcl-2 protein has not been clarified yet; HyPNoFs of bcl-2 show similarity to transcriptional regulators, nucleotide-interacting proteins, and tyrosine kinases, suggesting their possible functional relation to bcl-2 protein.

Data reported here indicate that eukaryote genes may carry functional protein messages on reading frame 1 as well as on the overlapping reading frames 2 and 3. This finding supports the hypothesis that polycistronic genes may had been present in eukaryotes, possibly as an ancient strategy, transferred to and kept on viruses genomes, and perhaps gradually lost in eukaryotes. Exhaustive analysis of HyPNoFs might help to trace the structural and functional evolution of DNA and proteins through the identification of putative ancestor-polycistronic genes coding for different (and possibly functionally related) proteins on overlapping frames.

Acknowledgments. I would like to thank Dr. A. Tramontano for helpful comments and critical reading of the manuscript.

References

Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 89:2002–2006

Dalgleish AG, Beverley PCL, Clapham PR, Crawford DH, Greaves MF, Weiss RA (1984) The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature 312:763–767

Dayhoff MO (1978) Atlas of protein sequences and structure. National Biomedical Research Foundation, Washington, DC

Dinesh-Kumar SP, Brault V, Miller WA (1992) Precise mapping and in vitro translation of a trifunctional subgenomic RNA of barley yellow dwarf virus. Virology 187(2):711–722

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? Science 214:149–159

Facchiano A, Facchiano F, van Renswoude J (1993) Divergent evolution may link human immunodeficiency virus gp411 to human CD4. J Mol Evol 36:448–457

Feng DF, Johnson MS, Doolittle RF (1984) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21: 112–125

Gilbert TL, Haldeman BA, Mulvihill E, O'Hara PJ (1992) A mammalian homologue of a transcript from the *Drosophila pecanex* locus. J Neurogenet 8(3):181–187

Gonnet GH, Cohen MA, Benner SA (1992) Exaustive matching of the entire protein sequence database. Science 256:1443–1445

Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162:705–70

Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM (1993) Ancient conserved regions in new gene sequences and the protein databases. Science 259:1711–1716

Hockenbery D, Nunez G, Milliman C, Schreiber RD, Korsmeyer SJ (1990) BCL-2 is an inner mitochondrial membrane protein that blocks programmed cell death. Nature 348:334–336

Jacks T, Madhani HD, Masiarz FR, Varmus HE (1988) Signals for ribosomal frameshifting in the Rous Sarcoma Virus gag-pol region. Cell 55:447–458

Lamb RA, Choppin PW (1979) Segment 8 of the influenza virus genome is unique in coding for two polipeptides. Proc Natl Acad Sci USA 76:4908–4912

Lamb RA, Horvath CM (1991) Diversity of coding strategies in influenza viruses. Trends Genet 7(8):261–267

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Karlin S, Brendel V (1992) Chance and statistical significance in protein and DNA sequence analysis. Science 257:39–49

Kirchner J, Sandmeyer SB, Forrest DB (1992) Transposition of a TY3 GAG3-POL3 fusion mutant is limited by availability of capsid protein. J Virol 66(10):6081–6092

Merelli F, Stojilkovic SS, Iida T, Krsmanovic LZ, Zheng L, Mellon PL, Catt KJ (1992) Gonadotropin-releasing hormone-induced calcium signaling in clonal pituitary gonadotrophs. Endocrinology 131;2: 925–932

Montell C, Fisher EF, Caruthers MH, Berk AJ (1982) Resolving the function of overlapping viral genes by site-specific mutagenesis at a mRNA splice site. Nature 295:380–384

Myers EW, Miller W (1988) Optimal alignments in linear space. Comput Appl Biosci 4:11–17

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Ray R, Jameel S, Manivel V, Ray R (1992) Indian hepatits E virus shows a major deletion in the small open reading frame. Virology 189(1):359–362

Rich DP, Anderson MP, Gregory RJ, Cheng SH, Paul S, Jefferson DM, et al. (1990) Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. Nature 347:358–363

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066–1073

Rose TM, Bruce AG (1991) Oncostatin M is a member of cytokine family that includes leukemia-inhibitory factor, granulocyte colony stimulating factor, and interleukin 6. Proc Natl Acad Sci USA 88(19):8641–8645

Sali A, Overington JP, Johnson MS, Blundell TL (1990) From comparison of protein sequences and structures to protein modelling and design. In: Bradshow RA, Purton M (eds) Proteins: form and function. Elsevier Trends Journals, Cambridge, pp 163–171

Sander C, Scheneider R (1991) Database of homology derived protein structure and the structural meaning of sequence alignment. Proteins Struct Funct Genet 9:56–68

Suzuki N, Sugawara M, Kusano T (1992) Rice dwarf phytoreovirus segment S12 transcript is tricistronic in vitro. Virology 191(2):992–995

Taga T, Narazaki M, Yasukawa K, Saito T, Miki D, Hamaguchi M, Davis S, Shoyab M, Yancopoulos GD, Kishimoto T (1992) Functional inhibition of hematopoietic and neurotrophic cytokines by blocking the interleukin-6 signal transducer gp130. Proc Natl Acad Sci USA 89(22):10998–11001

Wegenka UM, Buschmann J, Luttichen C, Heinrich PC, Horn F (1993) Acute-phase response factor, a nuclear factor binding to acute-phase response elements, is rapidly activated by interleukin-6 at the posttranslational level. Mol Cell Biol 13(1):276–288

Wilbur WJ, Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci USA 80:726–730