Coding capacity of complementary DNA strands

A.Casino, M.Cipollaro⁺, A.M.Guerrini, G.Mastrocinque⁺⁺, A.Spena and V.Scarlato

International Institute of Genetics and Biophysics, Naples, ⁺III^o Servizio di Analisi, 2nd Medical School, University of Naples, and ⁺⁺Istituto Elettrotecnico, Facoltà di Ingegneria, University of Naples, Italy

Received 25 November 1980

ABSTRACT

A Fortran computer algorithm has been used to analyze the nucleotide sequence of several structural genes. The analysis performed on both coding and complementary DNA strands shows that whereas open reading frames shorter than 100 codons are randomly distributed on both DNA strands, open reading frames longer than 100 codons ("virtual genes") are significantly more frequent on the complementary DNA strand than on the coding one.

These "virtual genes" were further investigated by looking at intron sequences, splicing points, signal sequences and by analyzing gene mutations. On the basis of this analysis coding and complementary DNA strands of several eukaryotic structural genes cannot be distinguished. In particular we suggest that the complementary DNA strand of the human ϵ -globin gene might indeed code for a protein.

INTRODUCTION

Since the nucleotide sequence of a number of structural genes have recently become available, we performed a theoretical analysis to investigate the possibility that the DNA strand complementary to the coding one, if transcribed, might indeed code for a protein. Symmetric transcription is in fact clearly demonstrated in many different biological systems where complementary RNA species were proved to be synthesized on opposite DNA strands (1-8).

We asked if it was possible to distinguish the coding (DNA sequence equal to the mRNA sequence) from the complementary DNA strand (DNA sequence complementary to the mRNA sequence) on the basis of the present knowledge of the nucleotide sequence of genes. For this purpose, we tentatively applied to the complementary DNA sequence the same criteria used for the identification of exons, introns, splicing points, leader and poly A tailing site sequences on the coding sequence. The analysis was performed on a sample of gene sequences

which are part of very large genomes and which are not believed to contain overlapping genes. In most of the cases analyzed, the amino acid sequence of the corresponding product was independently known (9-29). The results of this analysis indicate that in many instances it is not possible to distinguish on the basis of the definition of gene sequence, the coding from the complementary DNA strand. We call the open reading frames longer than 100 codons "virtual genes" and their putative products "complementary inverted proteins" or c.i.p.

STRATEGY FOR IDENTIFYING OPEN READING FRAMES

In order to analyze DNA sequences, a Fortran computer algorithm has been used to display the open reading frames (o.r.f.) present in a sequence in its three different frames. An o.r.f. is defined as a sequence, different from the real gene, identified by either of the following two statements:

1. The sequence starts with an initiation I (AUG) codon and terminates with the first subsequent termination T1, T2 or T3 (UAA, UAG, UGA) codon.
2. The sequence follows an initiation codon and no termination codon is found within the known sequence. This definition is applied under the assumption that a termination codon will be found in the unknown part of the sequence. This last assumption underestimates the length of the o.r.f.

The computer code analyzes the sequence and recognizes initiation and termination codons. When an o.r.f. is found its length is computed by the Fortran code starting from the first sense codon which follows AUG and terminating either to the first termination codon or to the end of the sequence. The number of o.r.f. is then plotted as a function of length for each frame and DNA strand. The results of the computer analysis for all the DNA sequences were combined to produce the histograms which will be discussed in the next section. Spliced genes are examined in detail separately, since in most cases intron sequences are not completely known.

The computer program correctly identified all the real genes present in the coding DNA strand, thus providing a positive check of the numerical code operations. Certain o.r.f. have been further analyzed in greater detail following other criteria proposed to identify gene sequences.

We have analyzed 3 S. cerevisiae mitochondrial DNA sequences, 23 nuclear

sequences, 1 phage sequence and 4 bacterial sequences (Table I). The former have been considered separately since in yeast mitochondria: i) UGA, which is a termination codon according to the standard codon assignment, codes for tryptophan (30, 21); ii) the codon reading pattern is unusual and iii) the AT/GC ratio outside the coding sequences is very high (32-34).

COMPUTER RESULTS

A) Nuclear, phage and bacterial sequences

The analyzed sequences are listed in Table I. Fig. 1 shows the o.r.f. coding length distribution of nuclear, phage and bacterial genes in the case of the coding (panel A) and of the complementary (panel B) DNA strands.

Two regions can be tentatively distinguished in panel A. In region I the o.r.f. (white bars) shorter than 100 sense codons have a random length distribution (one sided normal distribution). In this region no real genes (black bars) are found. Region II of panel A (codon length > 100) contains all the real genes and only one additional o.r.f.. To check the randomness of the distribution of o.r.f. a "a priori" probability function can be applied; i.e. a particular case of the binomial distribution. Once an initiation codon is found in a sequence, if the appearance of one of the three termination codons is a random event, the relative frequency (q) should be

$$q = q(U) q(A) \{2 q(G) + q(A)\} \quad (1)$$

with q(N) being the probability of the corresponding N base. For a sequence in which the four bases occur with equal probability:

$$q = 3/64 = 0.0469.$$

Then the "a priori" probability (P) of finding a termination codon at a distance equal to 0, 1, 2 K condons from the initial AUG is:

$$P(K) = q(1-q)^K \quad (2)$$

This probability distribution satisfies the normalization condition:

$$\sum_{k=0}^{\infty} q(1-q)^K = 1$$

Thus the probability P(K) of finding an o.r.f. of K codons length is simply given by equation (2). The expected number of o.r.f. calculated using equation (2) (i.e. by chance alone) has been plotted on the histogram (open circles). The mean coding length L as calculated using equation (2) is:

GENE PORTFOLIO	References	Nucleotide number	CODING DNA STRAND			COMPLEMENTARY DNA STRAND		
			Number of sense codons	Number of introns	Longest c.r.f.	Longest c.r.f.	Number of putative introns	2nd longest c.r.f.
<u>S.cerevisiae</u> oxi II	(31)	906	250	0	47 ³	24 ¹	0	20 ²
<u>S.cerevisiae</u> oli 1	(11)	1770	75	0	29 ²	9 ³	0	8 ¹
<u>S.cerevisiae</u> oli 2	(35)	1419	258	0	16 ²	22 ³	0	11 ²
Human α -globin	(12)	573	141		23 ²	33 ³		4 ³
Human β -globin	(13)	627	146	2	38 ²	>132 ¹	2	14 ²
Human γ -globin*	(12)	588	146		37 ²	>132 ¹		82 ²
Human ϵ -globin	(14)	1242	146	2	50 ²	243 ¹	2	46 ³
Human alpha chorionic gonadotropin	(15)	618	115	0	57 ²	37 ¹	0	32 ³
Human growth hormone	(16)	801	216	0	90 ²	47 ³	0	40 ¹
Human insulin	(17)	825	108	2	73 ³	48 ³	1	20 ¹
Rabbit α -globin	(18)	546	141		41 ²	>132 ¹		37 ²
Rabbit β -globin	(36)	585	146	2	35 ²	198 ¹	2	24 ²
Rabbit β_1 -globin	(19)	585	146	2	35 ²	>182 ¹	2	24 ²
Mouse β -globin	(20)	804	146	2	38 ²	52 ³	0	48 ¹
Mouse immunoglobulin λ_1 -type light chain	(21)	606	173	2	32 ³	38 ²	1	33 ³
Chicken β -globin	(22)	600	146		75 ²	>156 ¹		76 ³
Chicken ovalbumin	(23)	1857	385	6	76 ²	145 ³	2	45 ¹
<u>P. miliaris</u> H4 Histone*	(24)	408	102	0	0	16 ¹	0	12 ²
<u>P. miliaris</u> H2B histone*	(24)	570	122	0	23 ²	59 ¹	0	42 ¹
<u>P. miliaris</u> H3 histone	(24)	417	137	0	0	39 ¹	0	38 ²
<u>P. miliaris</u> H2A histone	(24)	735	123	0	34 ¹	47 ¹	0	28 ³
<u>P. miliaris</u> H1 histone*	(24)	690	189	0	22 ³	61 ¹	0	29 ³
<u>S.purpuratus</u> H4 histone	(25)	550	102	0	>162 ²	70 ²	0	32 ³
<u>S.purpuratus</u> H2B histone	(26)	939	123	0	38 ¹	135 ¹	0	44 ³

<u>S. purpuratus</u> H3 histone (26)	978	135	0	21 ²	44 ¹	0	37 ¹
<u>S. purpuratus</u> H2A histone(26)	684	123	0	26 ²	28 ³	0	17 ²
Phage MS2 replicase	(27) 1638	544	0	56 ²	83 ¹	0	49 ³
<u>E. coli</u> lactose permease (37)	1503	416	0	48 ²	76 ³	0	50 ³
<u>E. coli</u> lac I	(28) 1113	359	0	80 ³	56 ²	0	34 ²
<u>E. coli</u> Ampicillin resistance	(38) 1104	285	0	32 ²	87 ²	0	55 ³
<u>E. coli</u> ribosomal 16S	(39) 1541		0	83 ²	49 ¹	0	44 ²

Table I: Features of the 31 DNA sequences analyzed by the computer.

Sequences not completely determined are indicated by an asterix. The codon number of the o.r.f.'s which do not contain termination codons is preceded by > . The number in superscript indicates the reading frame with respect to the real gene. When a long o.r.f. is found on a complementary DNA strand of a cDNA derived sequence, if it is known that the real gene is spliced but no intron sequence is available yet, the result is temporary because the presence of extra sequences might interrupt the o.r.f., unless appropriate consensus signals are found.

$$L = \sum_{k=0}^{\infty} qK (1-q)^K = 1/q-1 = 20 \quad (3)$$

which is consistent with the computer results ($L = 18.4$).

Since from equation (2) the number of o.r.f. expected in the region between 100 and 200 codons is 1.58 the finding of one o.r.f. in the 160 codons decade, can be considered in agreement with the theoretical prediction. Obviously real genes (black bars) must not be taken into account in this statistical analysis since the investigated sequences have been selected for the very reason that they contain these genes.

The o.r.f. distribution on the complementary DNA strand (Fig. 1, panel B) is consistent with the theoretical prediction only in region I. In region II instead, the 9 o.r.f. found by the computer are significantly in excess of the figure of 2.08 predicted by equation (2). It should also be noted that the computed length of these o.r.f. is certainly an underestimate since 5 of them do not contain terminators.

We decide to call any open reading frame longer than 100 codons a "virtual gene". In fact, only the knowledge of the amino acid sequence of human β -, γ -, and α de-, rabbit α -, β -, and β_1 -, chicken β -globins, chicken ovalbumin and

Figure 1: o.r.f. length distributions found by the computer in the coding (panel A) and complementary (panel B) DNA strands of bacterial, phage and nuclear genes sequences as a function of their coding length.

Coding and complementary DNA strands are defined in the text. Complementary DNA strand is either the sequence obtained by means of the complementation rules (A:T; G:C) when only the mRNA sequence is known, or when available, the DNA sequence directly obtained from published data. Sequences that are not completely known have been treated as if they would not introduce termination codons in the unknown part of the sequence: the relative results are clearly temporary.

The total number of nucleotides per DNA strand is 27.822 whereas the codogenic part is 16.932 nucleotides. It has been assumed that all codons are used in the classical way.

Symbols: (▨) real genes, (□) o.r.f., (○) expected number of o.r.f. in each decade as calculated from equation (2). Regions I and II are discussed in the text. Left and right ordinates refer to region I and II, respectively.

sea urchin S. purpuratus H2B histone genes, allows one to distinguish between coding and complementary DNA strands.

All these "virtual genes" except one (chicken ovalbumin) are found in the same reading frame of the real genes (Table I): i.e. the initial AUG codon is complementary to a CAU codon in phase with the AUG of the real gene. In some cases (rabbit β^- and β_1^- , chicken β^- , human ϵ -globins and sea urchin S. purpuratus H2B histone genes) they are even longer than the corresponding real genes. For what concerns the codon usage it has to be noticed that the number of Leu and Ser residues present on the polypeptides coded by the sense DNA strand of these nine "virtual genes" is 161 and 106 respectively in agreement with the average percent composition of proteins.

These findings support the hypothesis that "virtual genes" are not the result of a random event.

B) S. cerevisiae mitochondrial sequences

Due to the particular base composition of the mitochondrial DNA sequences (32, 34) we have performed a separate analysis on coding DNA sequences and (A, T) reach sequences flanking the genes. Since UGA codes in the mitochondrial system for tryptophan (22, 23) and the remaining 62 codons have been proposed to be recognized as sense codons (41), only UAG and UAA have been regarded by the computer as termination codons. Furthermore since each base in both coding and (A, T) reach regions occurs with a peculiar frequency, we have used to following equation:

$$q = q(U) q(A) \{ q(G) + q(A) \} \quad (4)$$

where $q(N)$ has been taken equal to the average relative frequency of the N base calculated for the coding regions (A = 31%, T = 41%, G = 15%, C = 12%) and (A, T) reach regions (A = 46%, T = 43%, G = 5%, C = 5%) of the sense DNA strands of oli 1, oli 2 and oxi II genes. The figures for the complementary DNA strand have been calculated according to the complementation rules.

Fig. 2 shows the o.r.f. coding length distribution of S. cerevisiae mitochondrial sequences. The histogram is divided into two regions: region I which contains only o.r.f. shorter than 70 codons on both coding (panel A) and complementary (panel B) DNA strands and region II, which contains all the real genes (panel A) and no o.r.f.. The expected number of o.r.f. calculated

Figure 2: o.r.f. length distributions found by the computer in the coding regions of the sense (panel A) and complementary (panel B) DNA strands of oli 1, oli 2 and oxi II *S. cerevisiae* mitochondrial genes.

The insert shows the o.r.f. length distributions found by the computer in the (A, T) rich DNA sequences, flanking the genes. The number of nucleotides per DNA strand in the (A, T) rich sequences is 2445. The total number of nucleotides is 4419, whereas the number of coding nucleotides is 1974. Symbols are the same as in Fig. 1.

using eq. (2) has been plotted on the histogram (open circles). The distribution of o.r.f. is at random on both mitochondrial DNA strands. In fact, the mean coding length L as calculated from eq. (3) is 16 and 14 for the coding and the complementary DNA strand respectively, which is consistent with the computer result ($L = 14$ for the coding strand; $L = 10$ for the complementary DNA strand). The insert to Fig.2 shows the o.r.f. coding length distribution in the (A,T) rich sequences flanking the genes. The mean coding length ($L=10$) resulting from the computer for either coding and complementary DNA strand is consistent with the theoretical value $L=9$ for the coding strand and $L=9.5$ for the complementary DNA strand.

With respect to the codon usage in the mitochondrial system, it has been found that Leu and Ser are coded almost exclusively by UUA and CUA (9, 11) to which correspond on the complementary DNA strand the UAA and UGA termination codons respectively. This fact is in contrast with the situation reported previously for the nuclear sequences.

From this analysis we conclude that in S. cerevisiae mitochondria if the complementary DNA strand would be transcribed these RNA molecules could not be translated into protein.

C) Spliced genes

The results shown in Fig. 1 included a number of structural genes (see Table I) which are known to contain introns. We have done a separate analysis of the five cases (human ϵ -, rabbit β - and β_1 -, mouse β -globin and human insulin) for which the complete chromosomal sequence is presently available.

Since the frequency of termination codons inside the intron sequence is very high all the real genes and the o.r.f. longer than 100 codons (see Fig. 3) disappear. The number of o.r.f. shorter than 100 codons almost doubles (Fig. 4). Among them we found o.r.f. which encompass splicing points on the coding DNA strand (see Fig. 5 insert) and putative splicing points on the complementary DNA strand. These o.r.f. extend themselves considerably into the adjacent intron sequence. A recent report suggests the existence of the translation product of a sequence which is included in the first intron of the *oxi 2* gene of S. cerevisiae mitochondrial DNA (42).

Three new "virtual genes" are found on the coding DNA strand: one on the human ϵ -globin and two on the human insulin genes. The first one includes part of the ϵ -globin first intron, the second exon and part of the second intron. The human insulin "virtual gene" (150 codons) is completely included in the second intron, while the other "virtual gene" (190 codons) begins in the first exon and terminates in the second intron.

D) Splicing rules

The analysis of the exon-intron and intron-exon boundaries of several spliced genes has revealed that the excision-ligation events occur in all cases at unique positions with respect to the two dinucleotides GT and AG (43). In particular, GT specifies the exon-intron boundary whereas AG speci-

Figure 3: o.r.f. length distributions found by the computer in the coding and complementary DNA strands excluding from the analysis intron sequences.

The genes analyzed are human ϵ , rabbit β - and β_1 -, mouse β -globin and human insulin genes. Symbols are the same as in figures 1 and 2.

fies the intron-exon boundary. This feature obviously is not sufficient for the identification of splicing points on the precursors RNA molecules. Splicing points can tentatively be assigned through the analysis of quite a large number of nucleotides flanking the intron boundaries with the use of a computer program which shows in any given sequence the most probable splicing points (Staden and Brownlee, personal communication). In their program the overall average ratio of correct to false prediction is 10:8 for the exon-intron junctions and 11:8 for the intron-exon junctions.

This analysis, applied to the human ϵ -globin complementary DNA strand which contains the longest "virtual gene" found so far by the computer (see Table I, line 7), shows the existence of several putative splicing points. Four of them, almost aligned with the intron-exon boundaries of the coding

Figure 4: Analysis of the same gene sequences of Fig.3 including intron sequences.

DNA strand, show the correct polarity (see Fig. 5). This result indicates that the coding and the complementary DNA strands of the human ϵ -globin gene cannot be distinguished on the basis of the known splicing rules.

ARE ALSO "VIRTUAL GENES" IDENTIFIED BY SIGNAL SEQUENCES?

Signal sequences involved in the initiation of transcription are presumed to be located very close to or within the 5' end of an eukaryotic gene sequence which is transcribed into mature messenger RNA molecule. In several cases, a sequence equal or similar to 5'TATAAA was found to precede the transcribed region (44). Another sequence AAUAAA is found near the 3' end of the mRNA and is thought to signal the addition of the poly A tail (45). Moreover, sequences complementary to the 3'GCGGAAGGA sequence at the 3' end of the 18S ribosomal RNA, ribosomal binding site sequence (leader), have been observed in eukaryotic mRNA (46, 47).

Most of the analyzed sequences in our sample were obtained from cDNA clones. As a consequence, the proposed promoter sequence is seen only in the case of the mouse β and human ϵ -globin genes, whose sequences derive from chromosomal DNA clones. We looked for the presence of these signals in the appropriate nucleotide regions flanking the "virtual genes" (Table II). Seven out of ten of the "virtual genes" considered in our analysis contain no TATAAA sequence in the 5' flanking region. In these cases the absence of a promoter signal could be explained assuming either that the AUG, considered by the computer as the initiation codon, codes instead for an internal Met or that the known sequence upstream of the AUG is too short to include also the promoter site. The "virtual genes" found on the complementary DNA strand of the chicken ovalbumin and human ϵ -globin genes, have a 5'TATAAA sequence 409 and 217 nucleotides before the initiation codon, respectively.

Putative leader sequences are found upstream of the initial AUG of the "virtual genes" present on the complementary DNA strand of human β -, γ - and ϵ -, rabbit β - and β_1 -, chicken β -globin, chicken ovalbumin and sea urchin S.purpuratus H2B histone genes. The two cases where putative leader sequences are not found are the "virtual genes" present on the complementary DNA strand of the rabbit α -globin gene and that found on the coding DNA strand of the sea urchin H4 histone gene. A negative result has been obtained also taking into account the specific leader sequence proposed for the histone genes (47). However, these results are not conclusive since no termination codons are found among the codons preceding in phase the AUG found by the computer inside the known sequence. Similar considerations apply to the search for the poly A tailing site sequence (Table II).

HUMAN β -GLOBIN VARIANTS ANALYSIS

The analysis of one particular system, i.e. the human β -globin, has suggested us that the human β -globin "complementary inverted protein" exists and is essential. Indeed, since many viable mutations of this gene exist and have no obvious phenotypic effect, one would have to assume that c.i.p. is not seriously affected.

The availability of the amino acid sequence of 166 different point mutations (48) makes it possible to determine the correspondent nucleotide se-

Gene	Real gene		Virtual gene	
	Leader sequence	3' poly A tailing site	Leader sequence	3' poly A tailing site
Human ϵ -globin	CUUCCG ⁻⁴¹	AAUAAA	CUUCCU ⁻¹⁰²	AAGAAA U
Human β -globin	CUUCUG ⁻⁴¹	AAUAAA	CCUUCA ⁻⁴⁴	-
Human γ -globin	CUUCUG ⁻⁴¹	AAUAAA	CUUCUG ⁻³⁹	-
Rabbit α -globin	CUUCUG ⁻³⁷	AAUAAA	-	-
Rabbit β -globin	CUUUG ⁻⁴¹	AAUAAA	CUUCUG ⁻⁴¹	-
Rabbit β_1 -globin	CUUUG ⁻⁴¹	AAUAAA	CCUUCU ⁻⁴¹	-
Chicken β -globin	CCUCCG ⁻³⁶	AAUAAA	CUCCGU ⁻⁴³	-
Chicken ovalbumin	CCUUUAGCA ⁻³²	AAUAAA	GCGUGC ⁻⁶⁵	0
<i>S. purpuratus</i> H2B histone*	CCATTCA ⁻⁷³	AAUAAA	AUCAUUC	0
<i>S. purpuratus</i> H4 histone*	TCATTCG ⁻⁶³	-	-	-

Table II: Signal sequences on real genes and on "virtual" genes found on complementary DNA strands.

The first 9 "virtual" genes listed are found on the complementary DNA strand whereas the last one is located on the coding DNA strand (see Table I).

The sequence reported as "leader" might be involved in mRNA - 18S rRNA base pairing. Dots indicate G:U base pairs (46). The leader sequence in the case of the two histone genes, flanked by an asterix, are different (47).

Poly A tailing site sequence: 5'AAUAAA.

In the case of "virtual" genes only the putative leader sequence found upstream the initiation codon and the putative poly A tailing site sequence found downstream the termination codon are shown. (0) indicates the absence of the corresponding signal sequence, (-) indicates that the corresponding signal sequence was not found, (*) indicates the leader sequence proposed for the histone genes.

quence of the mutated gene (49). These facts have allowed us to verify whether the 132 codon long "virtual gene" is also present on the complementary DNA strand of these variants.

The initiation codon of this "virtual gene" faces the His 116 residue of the β -globin chain and it is coded by CAU. None of the variants so far analyzed affect this codon, yet at least one variant has been found for each of the other 8 His residues present in the wild type β -globin chain. Therefore, the initiation codon of the "virtual gene" is not removed by mutation affecting the β -globin polypeptide chain.

We have also analyzed the codon usage of all the known β -globin variants (48) to verify if any of them would introduce a termination codon on the complementary DNA strand. Since it has been demonstrated that termination codons introduced by mutations inside sequences translated in erythroid cell are effective (50), we infer that the presence of termination codons on the complementary DNA strand would seriously affect the c.i.p. Table III shows that none of the existing variants necessarily introduces a termination codon in-

I			II		III	
C U G	C U C	U C C	A C A	G C A	U U C	U U U
13	3	2	1	1	2	4
Leu ▼ Leu	Leu ▼ Leu	Ser ▼ Ser	Thr ▼ Ser	Ala ▼ Ser	Phe ▼ Leu	Phe ▼ Leu
-	-	-	0	0	1	1

Table III: Human β -globin variants.

The analysis is limited to the part of the β -globin coding sequence which is complementary to the "virtual gene". The first line shows the codons which by a single nucleotide substitution might become termination codons on the complementary DNA strand. The second line shows the number of times that these codons are used. The third line shows the correspondent amino acid substitutions on the β -globin polypeptide chain. The last line shows their occurrence: (-) indicates that the amino acid substitution is not observable since involves synonymous codons. Columns I, II, and III are discussed in the text.

side the "virtual gene". Class I consists of variants which cannot be detected since synonymous codons are involved. Class II shows that none of the variants that would have necessarily introduced a termination codon in the "virtual gene" has yet been found. Class III shows that two variants Phe $\rightarrow$ Leu have been found which could introduce a termination codon on the "virtual gene". However, UUC (Phe) and UUU (Phe) by single nucleotide substitution can become UUA or CUC, and UUA or UUG or CUU, respectively. Since all these five codons specify Leu, but only UUA corresponds a termination codon on the "virtual gene", it would be necessary to know the nucleotide sequence of the two variants to answer the question. Therefore, the existence of the 132 codon "virtual gene" is compatible with all the known β -globin variants. On the other hand, we can exclude the possibility that the three codons to which corresponds a termination codon on the complementary DNA strand (CUA and UUA that code for Leu and UCA for Ser) are not translatable in reticulocyte cell. In fact, 59 sense codons are used to translate human α and β -globin genes.

It would be interesting also to analyze the effect of deletion on c.i.p. existence. Indeed, there exist homozygotes with fused δ - β hemoglobin (Lepore) genes generated by deletions that extend from inside the δ -globin to inside the β -globin gene. Also this genetic context should not necessarily abolish the "virtual gene". In fact, the δ -globin gene is almost identical to the β -globin gene sequence. Therefore, in the δ - β fused gene it might occur that no termination codons are introduced in the complementary DNA strand by the new 5' part of the sequence corresponding to the moiety of the δ -globin gene.

It has also been proposed (49) that in the human α and β -globin genes, the codon usage and the pattern of single nucleotide substitutions are biased in ways that reduce the probability of occurrence of sense ($\rightarrow$) non-sense mutations and non-polar hydrophilic ($\rightleftharpoons$) hydrophobic amino acid substitutions. In fact, in the β -globin gene, the codons which by a single nucleotide substitution can become termination codons are used only when synonymous codons are not available to specify the same amino acid residue (49). We have verified that in the "virtual gene" the same rule is valid. For what

concerns the use of codons which introduce non-polar hydrophilic ($\rightleftharpoons$) hydrophobic amino substitutions, their frequency on both β -globin gene and "virtual gene", is low and comparable (data not shown).

HUMAN ϵ -GLOBIN

Fig. 5 shows the human ϵ -globin gene, as an example of our analysis. Both DNA strands show all the known features of the nucleotide sequence of a spliced eukaryotic gene. It has to be noted that the putative splicing points found on the B strand do not coincide with the splicing points of the A strand. The putative product (c.i.p.) of the 243 codon long "virtual gene", after splicing of the two putative introns, shows the typical features as far as protein secondary structure is concerned.

The 170 codon long "virtual gene" (Fig. 4, panel A) covers most of the small intron sequence, the second exon, and extends for 70 codons in the second intron, and is found in the same reading frame of real ϵ -globin. The termination codon at 1071 is then followed by a poly A tailing site at position 1277. If the AUG codon in position 801 of the second exon is used as initiation codon also a putative leader sequence is found (5'CUUCCU) upstream the AUG at position 775 (see Fig. 5 insert).

POSSIBLE ROLE OF THE C.I.P. GENES

Assuming that c.i.p. genes are functional we can make the following considerations:

- i) since "virtual genes" have been found "vis a vis" on DNA sequences which code for the main products of highly specialized cell lines, some hypothesis can be made on their role. We can imagine for instance, that the human c.i.p. β -globin gene is involved in the mechanism(s) that switch -on -off the different globins that are synthesized during embryonic, fetal and adult life. Alternatively it might be involved in determining the red cell line specificity or the level of β -globin expression. As a consequence mutations drastically affecting either only the β -globin or the c.i.p. β -globin products or both of them, might show the same phenotypic defect. This hypothesis could account for the extreme heterogeneity of the β -thalassemia phenotype.
- ii) Pseudogenes occurrence has been reported in mammalian globin gene clusters. Sequence analysis has revealed the presence of deletions, insertions and

Figure 5: Human ϵ -globin chromosomal DNA

All signal sequences present on both DNA strands of the human ϵ -globin gene are shown. The processing of precursor mRNA for the ϵ -globin and the proposed splicing process of the "virtual gene" precursor mRNA molecule are shown. The position of signal sequences and splicing points are indicated with the relative nucleotide number. The insert is discussed in the text.

A) coding DNA strand, symbols are: () : intron sequences; P: proposed promoter sequence; L: leader sequence; I: initial AUG codon; P1,2: splicing point of the proximal site of the introns; D1,2: splicing point of the distal site of the introns; T3: termination codon UGA; S: proposed poly A tailing site sequence.

B) complementary DNA strand, symbols are: (----): putative intron sequences of the "virtual gene"; VP: putative introns; VD1,2: splicing points for the distal site of the putative introns; VTi: termination codon UAA; VS: putative poly A tailing site.

The translated part of the human ϵ -globin gene (438 nucleotides) and the translatable sequence of the "virtual gene" (738 nucleotides) are indicated by a continuous line (—) flanked by dotted lines (---).

base changes that abolish the possibility to code for a functional globin molecule. We have found (data not shown) that an o.r.f. 141 codon long is present on the complementary DNA strand of the $\Psi\beta_2$ rabbit globin pseudogene. We think that our findings should be taken into account in the evaluation of the evolutionary processes of globin gene clusters.

CONCLUSIONS

The results of this investigation allow us to conclude that some of the longest "virtual genes" might indeed be transcribed and/or translated. Alternatively (52) something is either preventing their expression or missing in the definition of a gene as it can be deduced today on the basis of DNA sequences. In both cases, the study of the "virtual genes" will contribute to improve our knowledge on gene evolution and structure.

ACKNOWLEDGMENT

We thank Dr. G. Battistuzzi and Dr. G. Modiano for providing us their data on codon usage in the human globin genes and for helpful discussions; Dr. J. Quartieri and M. Candurro for the computer program; Dr. T.H. Jukes, Dr. E. P. Geiduschek and Dr. V. Pirrotta for reviewing the manuscript; Mrs. R. Behrend for editing the manuscript.

Address correspondence to: Dr. A.Cascino, International Institute of Genetics and Biophysics, Via Marconi 10, 80125 Naples, Italy

REFERENCES

1. Geiduschek, E.P. and Grau, O. (1969) Proc. 1st Le Petit Colloquium, L. Silvestri ed., pp. 190-201, North Holland Publishing Co.
2. Spiegelman, W.G., Reichardt, L.F., Yaniv, M., Heinemann, F., Kaiser, A.D. and Eisen, H. (1972) Proc. Nat. Acad. Sci. USA 69, 3156-3160
3. Jacquemont, B. and Roizman, B. (1975) J. Virol. 15, 707-713
4. Colby, C. and Duesberg, P.H. (1969) Nature 222, 940-944
5. Aloni, Y. (1973) Nature New Biol. 243, 2-6
6. Zimmer, S.G. and Raskas, H.J. (1976) Virology 70, 118-126.
7. Aloni, Y. and Locker, H. (1973) Virology 54, 495-505
8. Aloni, Y. and Attardi, G. (1971) Proc. Nat. Acad. Sci. USA 68, 1757-1761
9. Coruzzi, G. and Tzagoloff, A. (1979) J. Biol. Chem. 254, 9324-9330
10. Macino, G. and Tzagoloff, A. (1979) Proc. Nat. Acad. Sci. USA 76, 131-135.
11. Hensgens, L.A., Grivell, L.A., Borst, P. and Bos, J.L. (1979) Proc. Nat. Acad. Sci. USA 76, 1663-1667.

12. Forget, B., Cavallero, J., deRiel, J., Spritz, R., Choudary, P., Wilson, J., Wilson, L., Reddy, V. and Weissman, S. (1979) R. Axel, T. Maniatis and C.F. Fox, eds. (New York, Academic Press) in press
13. Kafatos, F.C., Efstratiadis, A., Forget, B.G. and Weissman, S.M. (1977) *Proc. Nat. Acad. Sci. USA* 74, 5618-5622
14. Baralle, F.E., Shoulders, C.C. and Proudfoot N.J. (1980) *Cell* 21, 621-626
15. Fiddes, J.C. and Goodman, H.M. (1979) *Nature* 281, 351-356
16. Roskam, W.G. and Rougeon, F. (1979) *Nucl. Acid Res.* 7, 305-320
17. Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischler, E. and Goodman, H.M. (1980) *Nature* 284, 26-32
18. Heindeli, H.C., Liu, A., Paddock, G.V., Studnicka, G.M. and Salser, W.A. (1978) *Cell* 15, 43-54
19. Hardison, R.C., Butler, E.T., Lacy, E., Maniatis, T., Rosenthal, N. and Efstratiadis, A. (1979) *Cell* 18, 1283-1297
20. Konkel D.A., Tilghman, S.M. and Leder, P. (1978) *Cell* 15, 1125-1132
21. Bernard, O., Hozumi, N. and Tonegawa, S. (1978) *Cell* 15, 1133-1144
22. Richards, R.I., Shine, J., Ullrich, A., Wells, J.R.E. and Goodman, H.M. (1979) *Nucl. Acid Res.* 7, 1137-1146
23. McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M. and Brownlee, G.G. (1978) *Nature* 273, 723-728
24. Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. and Birnstiel, M.L. (1978) *Cell* 14, 655-671
25. Grunstein, M. and Grunstein, J.E. (1977) *Cold Spring Harbor Symp. Quant. Biol.* XLIII, 1083-1092
26. Sures, I., Lowry, J. and Kedes, L.H. (1978) *Cell* 15, 1033-1044
27. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volchaert, G. and Ysebaert, M. (1976) *Nature* 260, 500-507
28. Farabaugh, P.J. (1978) *Nature* 274, 765-769
29. Ambler, R.P. and Scott, G.K. (1978) *Proc. Nat. Acad. Sci. USA* 75, 3732-3736
30. Macino, G., Coruzzi, G., Nobrega, F.G., May Li, and Tzagoloff, A. (1979) *Proc. Nat. Acad. Sci. USA* 76, 3784-3785
31. Fox, T.D. (1979) *Proc. Nat. Acad. Sci. USA* 76, 6534-6538
32. Bernardi, G., Prunell, A., Fonty, G., Kopecka, H. and Strauss, F. (1976) *In the Genetic Function of Mitochondrial DNA*. C. Saccone and A.M. Kroon, eds. (Amsterdam, North Holland) 185-198
33. Bernardi, G., Piperno, G. and Fonty, G. (1972) *J. Mol. Biol.* 65, 173-189
34. Prunell, A. and Bernardi, G. (1977) *J. Mol. Biol.* 110, 53-74
35. Macino, G. and Tzagoloff, A. (1980) *Cell* 20, 507-517
36. Van Ooyen, A., Van den Berg, J., Mantei, N. and Weissmann, C. (1979) *Science* 206, 337-344.
37. Buchel, D., Gronenborn, B. and Muller-Hill, B. (1980) *Nature* 283, 541-545
38. Calos, M.P. (1978) *Nature* 274, 762-775
39. Brosius, J., Palmer, M.L., Kennedy, P.J. and Noller, H.F. (1978) *Proc. Nat. Acad. Sci. USA* 75, 4801-4805
40. Bonitz, S.G., Berliani, R., Coruzzi, G., Li M., Macino, G., Nobrega, F.G., Nobrega, M.P., Thalenfeld, B.E. and Tzagoloff, A. (1980) *Proc. Nat. Acad. Sci. USA* 77, 3167-3170

41. Barrell, B.G., Anderson, S., Bankier, A.T., de Bruijn, M.H.L., Chen, E., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger F., Schreier, P.H., Smith, A.J.H., Staden, R. and Young, I.G. (1980) *Proc. Nat. Acad. Sci. US* 77, 3164-3166
42. Lazowska, J., Jacq, C. and Slonimski, P.G. (1980) *Cell* 22, 333-348
43. Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. and Steitz, J.A. (1980) *Nature* 283, 220-224
44. Baker, C.C., Herisse, J., Courtois, G., Galibert, F. and Ziff, E. (1979) *Cell* 18, 369-380
45. Proudfoot, N.J. and Brownlee, G.G. (1976) *Nature* 263, 211-214
46. Hagenbuckle, O., Santer, M., Steitz, J.A. and Mans, R.J. (1978) *Cell* 13, 551-563
47. Sures, I., Levy, S. and Kedes, L.H. (1980) *Proc. Nat. Acad. Sci. USA* 77, 1265-1269
48. International Haemoglobin Information Center (IHIC) Comprehensive Sickle Cell Center, August 1979
49. Modiano, G., Battistuzzi, G. and Motulsky, A.G. (1980) (submitted for publication)
50. Chang, J.C. and Kan, Y.W. (1979) *Proc. Nat. Acad. Sci. USA* 76, 2886-2889
51. Mears, J.G., Ramirez, F., Leibowitz, D. and Bank, A. (1978) *Cell* 15, 15-23
52. Wilson, C.N., Steggles, A.W. and Neinhuis, A.W. (1975) *Proc. Nat. Acad. Sci. USA* 72, 4835-4839