Vol. 139, No. 1, 1986 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
August 29, 1986 Pages 94-101

FREQUENT OCCURRENCE OF SHORT COMPLEMENTARY SEQUENCES IN NUCLEIC ACIDS

Ulf Segerstéen[1], Håkan Nordgren[2], and Jan-C. Biro[3]

[1]Reproductive Endocrinology Unit, [2]Department of Cellular Genetics, and [3]Pediatric Endocrinology Unit,
Karolinska Institute, 104 01 Stockholm, Sweden

Received June 11, 1986

The hypothesis, that nucleic acids which code specifically interacting receptor and ligand proteins contain complementary sequences was tested. Human insulin mRNA (HSINSU) contained 16 sequences which were 23.8 ± 1.4 nucleotides long and were complementary to the insulin receptor mRNA (HSIRPR, 74.8 ± 1.9% complementary matches, $p < 0.001$ compared to randomly occurring matches). However, when examining 10 different nucleic acids (coding proteins not interacting with the insulin receptor), 81 additional sequences were found which were also complementary to HSIRPR. Although the finding of short complementary sequences was statistically highly significant, we concluded that this is not specific for nucleic acids coding specifically interacting proteins.

Our previous hypothesis, that specifically interacting proteins may be coded by nucleic acids containing complementary sequences, was further examined (1).Several laboratories have seriously considered this possibility since our publication of the comparative computer analysis regarding the specificity of protein-protein interactions in 1981. Regions of complementarity between the mRNA for epidermal growth factor, transferrin, interleukin-2 and their respective receptors were found (2). In this study we investigated the occurrence of complementary nucleic acid sequences in mRNAs coding human insulin and its receptor.

METHODS

Human pre-proinsulin (HSINSU,3) and 10 other nucleic acid sequences (4-13) were studied in order to find messages which are complementary to the human mRNA for insulin receptor precursor (HSIRPR, 14). Nucleic acid sequences were taken from EMBL (Heidelberg,D-6900) and Dayhoff (NBRF, Washington)- Nucleotide Sequence Data Libraries. To find complementary nucleic acid sequences,we looked for homology between a nucleotide sequence and the reversed and converted sequence of HSIRPR. Homologue sequences (only those longer than 20 nucleotides) were investigated by BestFit, DotBlot and WordSearch programs in sequence analysis software package of the University of Wisconsin Genetics Computer Group (version 3, June 1985, Madison, Wisconsin), using a VAX computer.

0006-291X/86 $1.50

Statistical analysis of nucleotide homology

The significance of homology between two nucleotide sequences was estimated by calculating the probability of accidental occurrence (Pa), using the equation based on the Poisson distribution (2):

$$Pa = \sum_{i}^{N} \frac{e^{-Np}(Np)^{i}}{i!}$$

where N is the length of nucleotides in the homologue sequence, i is the number of matches over the sequence and p is the probability that any given nucleotide will match. The p value of the nucleic acids examined was very close to 0.25 (ideal randomness), except for some comparations concerning the approximately 700 nucleotides long end-part of the HSIRPR. We therefore used p=0.25 for the calculation of Pa values. When N = 26, in sequences with ideal randomness, the number of nucleotide matches (i) is 6.5 . The experimentally determined value varied around this mean within ±40% SD limits (3.9 - 9.1), resulting in Pa $\leq 8 \times 10^{-2}$ of homologies between random sequences.

For statistical significance "i" values had to be higher than 2SD from the mean (i.e. 11.7). Thus Pa values $\leq 2 \times 10^{-2}$ for N=26 were considered statistically significant.

For further statistical evaluation Student t-test and linear regression analysis were used

RESULTS

As many as 97 short complementary messages to HSIRPR were found in 11 nucleic acids (Table 1). All sequences examined - with the exception of HSALB - were codes of hormones or releasing factors. The average length of complementary messages was N = 26.1 ± 0.8 (S.E.M., n=97) and the average number of complementary matches was i= 73.4 ± 1.2%. The average N and i values of the 16 complementary sequences in HSINSU did not differ significantly from the respective values of complementary messages found in "control" nucleic acids which code other hormones or albumin. The mean Pa value of these short complementary messages was 3.4×10^{-5} (N=26, i=19), i.e. by 3rd order of magnitude (2000 times) less than the average ±2 S.D. limit of significance ($Pa=2 \times 10^{-2}$). This indicates that the probability of occasional occurrence of short complementary sequences was extremely small. Considering that we found 97 such sequences, the probability of detecting a randomly occurring phenomenon is even less.

The number of complementary sequences varied from one nucleic acid to the other, proportionally to its length. There was a significant linear correlation between the length of the nucleic acid examined and the total length of complementary sequences found in them (Fig.1.). The length of complementary messages

Table I. Characteristics of complementary sequences

Name	Reference	Number of nucleotides (N)	Number of complementary sequences found (n)	Length of complementary sequences (mean±SEM)	Total length of complementary sequences (nL)	Percentage of complementary sequences (nL/N100)	Average numbers of complementary maches (mean±SEM)
Insulin receptor precursor (HSIRPR)	(14)	5180	97	26.1±0.8	2536 (1546)* 48.9%(29.8%)	49.0 (29.8)*	75.5± 3.5
Corticotropin (HSACTH)	(4)	8658	24	24.9±2.0	598	6.9	75.5±3.5
Precursor for epidermal growth factor (HSEGFPRE)	(5)	5531	13	28.1±1.4	365	6.6	72.2±1.2
Preproinsulin (HSINSU)	(3)	4992	16	23.8±1.4	380	7.6	74.8±1.9
Preprocorticotropin releasing factor (HSPCRF	(7)	2685	10	26.5±1.6	265	9.8	76.3±1.6
Endorphin (SSENDO)	(6)	2333	7	26.0±2.7	182	7.8	73.4±2.0
Albumin (HSALBU)	(8)	2251	5	31.6±2.6	158	7.0	71.4±2.3
Presomatotropin (HSGROW2)	(9)	1964	7	24.0±3.2	168	8.5	65.4±9.6
Luteinizing hormone-α(RNLHØ1)	(10)	1662	7	28.8±2.6	201	12.1	70.5±1.2
Preprocalcitonin (HSCALC)	(11)	791	2	38.5±9.5	77	9.7	64.6±4.2
Thyrotropin (MMTHYR)	(12)	640	1	18.0	18	2.8 (+)	77.7
Somatomammotropin (HSSOMA)	(13)	551	5	24.8±5.7	124	22.5 (+)	75.9±5.3
Summa or average		32,058	97	26.1±0.8	2536(7.9%)	9.2±1.5 n = 11	73.4±1.2
Random		2,500	Ø (10)**	Ø (26)**	Ø (160)**	Ø (10)**	25.0±3.1

* length and % of complementary sequences considering overlays

** Randomly chosen sequences for calculation of frequency of occasionally occurring complementary nucleotides

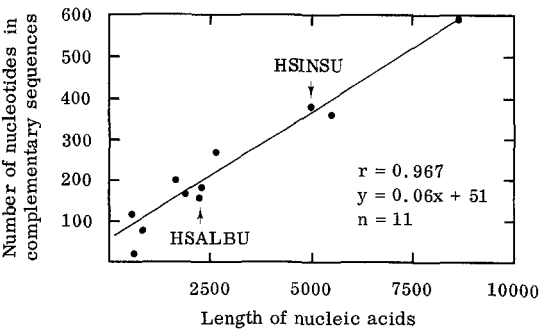

Fig.1.: Correlation between the length of nucleic acids examined and the total number of nucleotides found in HSIRPR-complementary sequences.

found in HSINSU and HSALBU did not differ from each other and those of other nucleic acids examined, when correlated with the total length of the sequence.

The 97 short complementary sequences had a total length of 2536 nucleotides. This corresponds to 48.9% of HSIRPR length and complements to 1546 long sequences of HSIRPR (29.8%) because of some overlap between the sequences.

With a frequency of 7-9% the short complementary sequences of the nucleic acid sequences examined, proved - without any exception- to be complementary to HSIRPR.

The short complementary sequences in HSIRPR and HSINSU did not show any preferred location at the protein coding parts of the nucleic acids in question. (Table II).

Nucleotides coding proinsulin peptide A and B did not contain any complementary sequences and putative α- and β-subunit coding sequences of HSIRPR contained only the average number of complementary nucleic acids.

However, it is notable that the signal peptide coding messages in HSIRPR and the proinsulin peptide-C-coding-region of HSINSU contained approximately 4 to 9 times more complementary nucleotide sequences than average.

Some of the 16 complementary sequences found in HSINSU were unique and we were not able to find homologue sequences in other nucleic acids. Others showed more or less homology to various nucleic acid messages (Fig.2.).

Table II. Distribution of complementary sequences

Description	Position	Length (L) (nucleotides)	Length of complementary sequences (C)	C/L (%)
I. HSIRPR				
- signal peptide	49-129	81	156	193
- put. α-subunit	130-2298	2169	628	28
- put. β-subunit	2299-4158	1860	785	42
- undefined regions	1-48 4159-5180	1070	967	90
|entire sequence	1-5180	5180	2536	48|
II. HSINSU				
- polymorphic region	1340-1823	484	50	10
- preproinsulin primary transcript	2186-3615	1430	101	7
· intron	2611-3396	785	52	6
· proinsulin peptide C (part. 2)	3397-3476	80	52	65
· proinsulin peptide A	3477-3539	63	Ø	Ø
" B	2496-2585	90	Ø	Ø
- undefined regions	1-1338 1824-2185 3616-4992	3078	229	7
|entire sequence	1-4992	4992	380	7|

DISCUSSION

The receptor and ligand, antigen and antibody protein molecules show exceptionally strong attraction to each other. The molecular background of these specific protein-protein interactions is not sufficiently understood. In 1981 we suggested the theory (1) that proteins which are able to interact specifically with each other, might be coded by nucleic acid sequences which are complementary to each other. Five years later the first opportunity for experimental tests came, when the sequence of some receptor-coding nucleic acids was published. When searching for complementary messages between mRNAs for epidermal growth factor, transferrin, interleukin-2 and their respective receptors, as many as 11 specific complements were determined which consisted

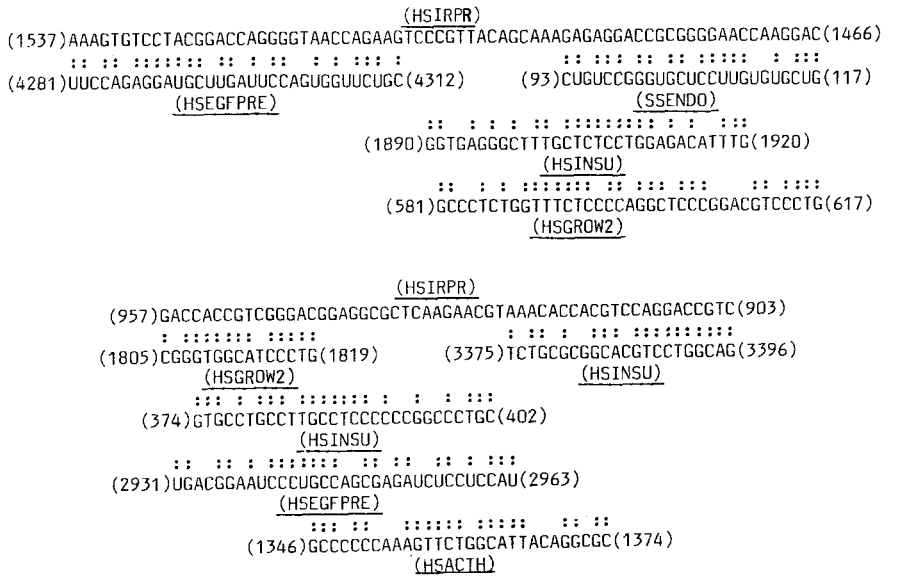

Fig.2.: Some examples of short complementary sequences found (9 out of 87).

of 15-18 nucleotides (2). Based on this and other previous experiments, Bost, Smith and Blalock (17) concluded that coding of proteins by complementary nucleic acids might result in proteins which specifically interact with each other.A hydropathic anti-complementarity of amino acids based on the genetic code may be the chemical background for this specificity (15).Codons for hydrophilic and hydrophobic amino acids on one strand of DNA were complemented by codons for hydrophobic and hydrophilic amino acids on the other DNA strand.

The co-existence of short complementary RNA messages was shown by Mizuno, Chou and Inouye (16). They described a sequence in E. coli which contained 84 nucleotides. This sequence was complementary to the ompF mRNA (67 complementary matches). This so called mRNA-interfering complementary RNA (micRNA) is believed to regulate the translation of ompF mRNA by specific interaction. It is unknown whether a micRNA can be translated into a protein but this possibility should not be disregarded.

A micRNA might possibly regulate translation, thus, a micRNA-derived protein might regulate the interactions, biological action or degradation of other proteins to which it is "complementary". This was proven by sophisticated experiments done by Bost, Smith and Blalock (17):

a) Corticotropin (ACTH) and β-endorphin bind specifically and with high affinity to synthetically derived counterparts, that were specified by RNA sequences which were complementary to the mRNA for ACTH and β-endorphin, respectively.
b) Antibody to the peptide (which was encoded by the complementary RNA for ACTH) recognized the adrenal cell ACTH-receptor. This observation strongly supported the supposition that two peptides coded by complementary RNA messages are "internal" images of each other.

When looking for nucleic acid sequences complementary to HSIRPR, we expected HSINSU to contain such messages. We did, indeed, find 16. The length and number of complementary matches were close to what had been described for micRNA (16) and other receptor-ligand coding RNA complements (2). The finding of each sequence was statistically highly significant, because
a) their Pa values were less than Pa values for ideal randomness by at least 3rd order of magnitude, and
b) no such sequence was found when computer generated random nucleic acid sequences were studied with the same method.
In addition to the 16 short complementary sequences in HSINSU we found 81 in 10 other nucleic acids which code other peptides. These peptides do not bind to the insulin receptor. Even though these sequences were in most cases completely different from those found in HSINSU, this finding made our initial hypothesis doubtful. The frequent occurrence of short complementary sequences in genes indicates that not only specifically interacting proteins are coded by nucleic acids which contain complementary messages. These short complementary messages might therefore have some completely different structural and functional importance (18). Their main role, for example, may be in the functional organization of different genes or simply in compact packing of nucleic acids in the nucleus.

REFERENCES

1. Biro, J.C. (1981) Medical Hypothesis 7, 969-1007.
2. Bost, K.L., Smith, E.M. and Blalock, J.E. (1985) Biochem. Biophys. Res. Comm. 128, 1373-1380.
3. Bell, G.I. et al (1980) Nature 284, 26-32.
4. Takahashi, H. et al (1983) Nucl. Acid. Res. 11, 6847-6858.
5. Ullrich, A. et al (1984) Nature 309, 418-425.
6. Kakidani, H. et al (1982) Nature 298, 245-249.
7. Shibahara, S. et al (1983) EMBO J. 2, 775-779.
8. Lawn, R.M. et al (1981) Nucl. Acid. Res. 9, 6103-6114.
9. Denoto, F.M. et al (1981) Nucl. Acid. Res.9,3719-3730
10. Gudine, J.E. et al (1982) J. Biol. Chem. 257,8368-8371

11. LeMoullec, J.M. et al (1984) FEBS Letters 167, 93-97
12. Chin, W.W. et al (1983) PNAS. USA 78, 5329-5333
13. Shine, J. et al (1977) Nature 270, 494-499
14. Ullrich, A. et al (1985) Nature 313, 756-761
15. Blalock, J.E. and Smith, E.M. (1984) Biochem. Biophys. Res. Comm. 121, 203-207
16. Mizuno, T., Chou M-Y. and Inouye,M. (1984) PNAS. USA 81, 1966-1970
17. Bost, K.L., Smith, E.M. and Blalock, J.E. (1984) PNAS. USA 82, 1372-1375
18. Biro, J.C. (1983) Med. Hypoth. 12, 203-226