Vol. 186, No. 1, 1992
July 15, 1992

BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
Pages 211-218

# INTERMOLECULAR HOMOLOGIES OF HUMAN INTERFERON-ALPHA

Fedor E. Romantsev*, Nikolai N. Khodarev, and Igor I. Votrin

Center of Medical Biotechnology, Schukinskaia Street, 6,
Moscow 123436, Russia

Received May 28, 1992

SUMMARY: Human interferon-alpha 2 (IFN) was analyzed by homology search computer program with the use of protein primary structures data bases. Results indicate that four domains with heightened ability to form homology pairs with different proteins exist in the IFN molecule. These domains occupy regions 35-56, 72-85, 97-110 and 124-136, mainly between the alpha-helical cylinders on the tertiary structure models. Additionally, results show in IFN structure the presence of amino-acid motifs that create the opportunity for this cytokine to influence directly the processes of DNA functioning in cell nuclei. 

The interferons (IFNs), are the members of the complex network of cytokines. They demonstrate three different types of activity: antiviral, immunomodulating and antiproliferative and act through the induction of specific genes (1-11).

The IFNs primary structures were well characterized and functional roles of some amino-acid residues were described (1, 13-16). Recently, the three-dimensional structure of murine interferon-beta was reported (17). There is some evidence for IFN molecule to have definite specific domains which determine distinct activities (14, 16, 17).

Investigations of structure-function relationships in IFNs molecules are necessary to understand the molecular basis of their actions on the cell. One can suggest that IFNs structure comparison with different proteins which have well defined biological activities and mechanisms of interaction with cells may be useful approach in these investigations.

In this paper we present results of homology search using computer data bases between IFN and different proteins.

Results show that in IFN molecule exist four regions containing amino-acid sequences which are found by homology search in the used data bases with relatively high frequency. These regions well coincide with parts of the molecule located between alpha-helical regions in IFN three-dimensional model (12, 15-17).

Results also reveal definite IFN sequences which are similar to some DNA-binding and nuclear proteins.

* To whom correspondence should be addressed.

0006-291X/92 $4.00

## Materials and Methods

The work was performed on the PC Olivetti M290 with the CDROM laser driver and DNASIS/PROSIS data base (Pharmacia-LKB).

The initial structure, i.e., key sequence for homology search was the primary structure of human interferon-alpha 2 precursor, but signal sequence in the precursor from Met1 till Gly23 was deleted from the analysis. So, the primary structure of mature IFN from Cys1 till Gly165 was used in every run of homology search in order to find and extract sequences homologous to IFN.

Sequence comparison was performed according to Pearson and Lipman rules (18, 19).

Scheme of search. First step: homologous sequences were determined and collected from laser disc General Data Base (GDB) with the use of Amino Acid Homology Search (AAHS) subprogram. Second step: the Restricted Data Base (RDB) was constructed by Data Base Access and Retrieval subprogram with the use of Short Directory and Keyword program options. Homologous sequences were determined and collected from RDB by AAHS subprogram and then they were united with sequences from the first step.

Our RDB for IFN homology search was based on key words and short directories, corresponding to those protein superfamilies, which revealed the maximum extents of homology (evaluated on the basis of matching score) during the initial GDB screening. 213 homologous sequences used for the further analysis were found and selected from RDB also on the basis of matching score calculated by PROSIS. Matching score values were from 28 to 87. Homologous sequences with matching scores less than 28 were not taken into further analysis. This limitation determined the total quantity of selected homologous sequences.

## Results

The first step of investigation was the random search through all available sequences from GDB in order to find most similar protein structures with the highest values of initial homology matching scores calculated by PROSIS AAHS program. Homology search in GDB permitted to find out and to look at only 50 most homologous sequences, because the first 50 positions were occupied by IFNs superfamily representatives and the total capacity of accumulating program was 100 sequences. At the same time, while watching the current status of the program one could notice that when only about 4 % of GDB were screened, the number of discovered homologous sequences (HS) became more than 100.

This situation led to the necessity to create RDB and to divide it into several parts in order to have an opportunity to take into account all HS from chosen data base. The RDB was constructed by means of Short Directories (SD) and Keywords (KW) on the base of protein families found in the first step. Some sequences were found both in SD and KW searches and doubling was excluded. In general, RDB contained 1228 proteins. Additional 213 structures homologous to IFN were extracted from RDB. So, the total amount of HS found in GDB and RDB consisted from 263 sequences.

It was assumed that identical amino-acids are distributed randomly within the sequence of definite length. Thus, it was supposed that the dependence of identical amino-acids' content in HS upon the length of these sequences must be described by simple multiplicative function $y=ax^{-b}$. In fact, the regression analysis of identical amino-acids content in found sequences versus their length is most correctly described by corresponding regression curve with correlation coefficient -0.8521 (fig. 1, B). In whole, this indicates the really random character of our RDB. It was also supported by curves designed for cytochrome P-450 and albumen which were taken as controls.

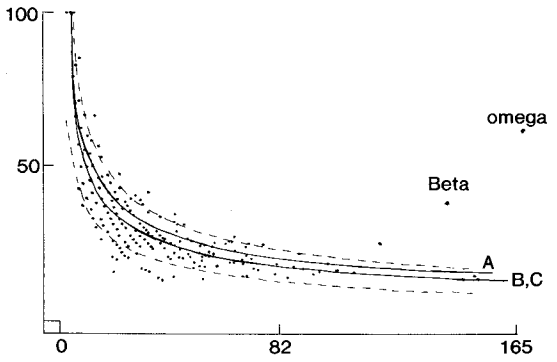


*Figure 1.* Content of identical amino-acids in homology overlap regions of HS.
X axis - length of homology overlap region; Y axis - % of identical amino-acids; A - cytochrome P-450; B - IFN; C - albumen; omega - interferon-omega sequence; gamma - interferon-gamma sequence. Points specific only for IFN HS are represented. Dotted line indicates the 85% predictive interval.

Comparison of discovered 263 sequences with IFN primary structure revealed that different amino-acids in IFN have different frequencies of homology. Fig. 2 shows that N- and C-terminal regions of IFN molecule are characterized by low frequencies of homologies, but in the central part of the molecule four regions with high frequencies exist. We called these regions NT (N-terminal), NC (N-central), CC (C-central) and CT (C-terminal). They are occupying amino-acid positions 35-56, 72-85, 97-110 and 124-136 respectively, if the peak width is determined at the level of the half of their amplitude towards the common linear base line, going through the left and right bottoms of NC peak.

One of the possible explanations might be the following: the RDB was subjectively formed by protein sequences with maximum homologies to four above mentioned regions. Possibly, some changes in the homologies frequencies per residue along the IFN molecule might appear if the comparison with

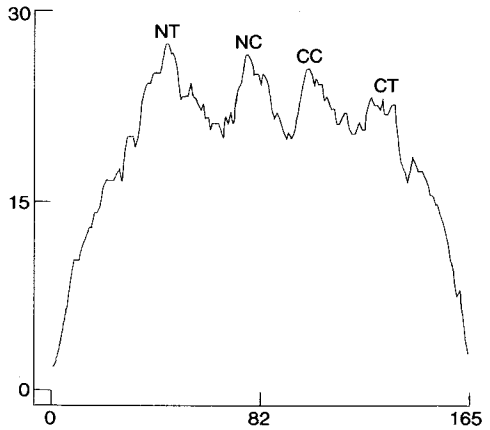


*Figure 2.* Frequency of inclusions of definite amino-acid residue in IFN sequence into the formation of homology pairs.
X axis - number of residue, starting from N-terminal Cys1; Y axis - relative quantity of homologous sequences in which definite IFN residue was found, %.

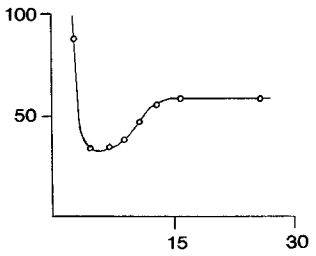


*Figure 3.* Dependence of quantity of homologous sequences upon the length of IFN fragment.
X axis - length of key sequence fragment; Y axis - quantity of homologous sequences.

GDB was performed. But as it was said, owing to the restrictions of the used program it was impossible to carry out the HS search in GDB with the use of the whole IFN molecule. We carried out the analysis of dependence of HS quantity found in GDB from the length of key sequence fragment (fig. 3). As it is seen from the fig. 3, the plot has the "U" - form manner. While increasing the length of key sequence fragment over three amino-acids the amount of found HS sharply reduces. Further increase in length of key sequence fragment up to 7-11 residues leads to the growth of HS quantity gradually reaching the plateau.

On the basis of this data we divided IFN molecule into fragments having 7, 9 and 11 residues in length and for every fragment we carried out homology search in GDB. HS quantities found by such mode for every part of the molecule are represented on fig. 4 A, B and C. There is a good agreement between the regions found by this method and the regions found on the basis of RDB (fig. 2).

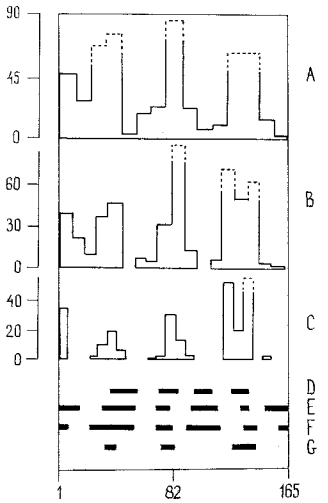


*Figure 4.* IFN domains.
A, B and C drawings show the effect of length of IFN fragments on the quantity of homologous sequences.
X axis - IFN primary structure; Y axis - quantity of homologous sequences; A - 11 amino-acids fragments, B - 9 amino-acid fragments, C - 7 amino-acid fragments, D - this article, fig. 2; E - domains positions calculated from Zavialov's model of IFN tertiary structure (12); F - non helical regions positions determined from Senda's three dimensional model (17); G - positions of three crucial segments in interferon-alpha and -beta families (17).

TABLE 1. Proteins with maximal lengths of homologous to IFN sequences

| No. | Target protein - source | Overlap region | | | |
|---|---|---|---|---|---|
| | | Length, amino acids | Position | | Identical amino acids, % |
| | | | IFN | Target | |
| 1. | Retrovirus - related pol polyprotein (transposon gypsy) - Fruit fly | 146 | 10 - 149 | 787 - 930 | 14.4 |
| 2. | rpoC2 protein; superfamily: DNA - directed RNA polymerase - Liverwort | 145 | 17 - 157 | 1047 - 1185 | 15.2 |
| 3. | D4 protein; related to DNA replication - Vaccinia virus | 112 | 59 - 160 | 22 - 130 | 25.9 |
| 4. | rec N protein; may be involved in DNA recombination and DNA repair - E. coli | | | | |

This fact confirms the representative character of RDB. We tried to classify the proteins in RDB by their activities and possible functions. It appeared that the most representative group is formed by different enzymes and proteins that are participants of DNA and RNA metabolism (34% of total amount of HS).

More of all, 4 proteins (from 263 homologous to IFN proteins found in GDB and RDB) which formed the most long homology overlap regions with IFN belonged to the same group. They are summarized in the Table 1.

These facts give rise to the question about the existence in the IFN molecule definite sequences specific for some of DNA-enzymes and proteins. From this point of view, one of the interesting regions in IFN is the sequence from 79 till 137 residue, which is analogous to Leu-zipper motif of well-known mammalian DNA-binding proteins (fig. 5). It is interesting, that this region is involved in IFN homology pairs with pol-polyprotein, rpoC2 protein and D4 protein.

| Row N° | | | | LEUCINE ZIPPER | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 | | 2 | | 3 | | 4 | | 5 | |
| 1 | TREB5 | ( 96-130) | SE | L | EQQVVD | L | EEENQK | L | LLENQL | L | REKTHG | L | VVEN |
| 2 | TREB7 | (378-412) | QS | L | EKKAED | L | SSLNGQ | L | QSEVTL | L | RNEVAQ | L | KQLL |
| 3 | TREB36 | (239-267) | KC | L | ENRVAV | L | ENQNKT | L | IEELKT | L | KDLYSNKSV | | |
| 4 | CREB1 | (309-337) | KC | L | ENRVAV | L | ENQNKT | L | IEELKA | L | KDLYCHKSD | | |
| 5 | cJUN | (278-312) | AR | L | EEKVKT | L | KAQNSE | L | ASTANM | L | REQVAQ | L | KQKV |
| 6 | cFOS | (163-297) | DT | L | QAETDQ | L | EDEKSA | L | QTEIAN | L | LKEKEK | L | EFIL |
| 7 | IFN | ( 79-137) | TL | L | DKFYTE | L | YQQLND | L | --EALY | L | KEKKYSP | | |

*Figure 5.* Sequence homology between mammalian proteins containing Leu zippers.
The name of the protein and the numbers of the first and last amino acids are shown on the left of the sequences (26). Underlined regions indicate positions where definite IFN residue is homologous or identical to at least one residue in rows 1-6.

## Discussion

As a result of PROSIS data base screening we found high amount of proteins which contained homologous with IFN sequences. It was reported previously that usage of computer data base and homology search programs make this high level of homologies a general rule for any protein (18, 19). May be, this situation correlates with the idea, that great variety of functions is not accompanied by the great variety of structures (20). This creates the question about the presence of definite structural-functional prototypes in the amino-acid sequences of proteins, and the computer homology search can be the useful tool in solving this question.

The regression analysis of HS found in RDB and GDB (see fig. 1) show their random character. Non random homologies of omega- and beta-interferons to alpha-interferon are clearly seen on that graph, obviously representing the protein membership in definite family or superfamily.

Data represented on fig. 2 and 4 indicate that IFN molecule contains sequences with non equal frequencies of homology with other proteins. The sequences with maximal frequencies of homology are localized in the regions occupying 35-56, 72-85, 97-110 and 124-136 amino-acid residues. These regions were called NT, NC, CC and CT respectively.

These regions can contain high level of most frequently met in the proteins amino-acids. We calculated the average frequencies of met of amino-acids in that regions, between them and in positions 1-34 and 137-165, using the data of (21). The calculated values varied from 44.54% to 65.03% and did not coincide with patterns of homology frequencies.

Thus, the existence of these 4 regions is the result of non random distribution in IFN molecule of amino-acid sequences which differs in frequency of homology with other proteins.

It was interesting to compare IFN regions found by us with structural-functional domains which were proposed by other authors (fig. 4 E, F and G). These regions are in an agreement with reported data (12, 17), in which some possible variants of IFN tertiary structures are supposed. As it is seen from the fig. 4 NT, NC, CC and CT regions coincide with non helical regions and loops described by Zavialov (12) and Senda (17). At the same time, NC, CC and CT domains occupy such regions in IFN molecule, that correspond to three crucial polypeptide segments which are most directly involved in the expression of antiviral and antiproliferative activity of IFN-alpha and -beta family (17). Such closely coincidence of regions obtained by alternative methods is unlikely to have the random nature. This mean that non helical and loop regions of IFN molecule contain amino-acid sequences which are most commonly distributed among the other protein families then the others IFN sequences. The same sequences are most important for biological activity of IFN. One can suggest that it really indicates the existence of evolutionary conservative structure-functional prototypes of amino-acid sequences, but this must be the object of special investigation.

As it was mentioned above, the most highly represented protein group among homologous to IFN proteins was formed by proteins and/or enzymes that are involved in DNA metabolism. Thus, it arose the question about the existence of specific sequences in the interferon molecule that are functionally significant for interactions between IFN and cell nuclei's structures. Such question is also based on experimental data indicating that interferon-beta may translocate through nuclear membrane and besides internalization may influence on nuclear structures (22), and that interferon IFN can inhibit the mammalian DNA polymerase (5).

From that point of view, the main interest is formed by the presence of Leu-zipper sequence in the IFN molecule (fig. 5). Need to be mentioned, that this sequence overlaps with NC and CC regions and is involved in such parts of IFN structure that are forming the maximally long homology pairs with DNA-proteins (Table 1).

The presence of Leu-zipper sequence providing the formation of dimeric molecules can explain the brightly expressed IFN tendency to form dimers in aqueous solutions. The second consequence is the possible direct (without receptor mediation) IFN action on genome functioning. Such opportunity is supported by data about the formation of heterodimeric molecules of DNA-binding proteins (27-32). IFN sequence does not contain positively charged block necessary for DNA binding near the Leu-zipper analog, but the IFN formation of heterodimeric molecules with specific DNA-binding proteins containing Leu-zipper motifs may modulate their binding with DNA and therefore may influence the regulation of genome functions.

It is interesting to mark that IFN contain reversed Nuclear Location Signal (NLS) (23-25) in position 131-136 (Ser-Tyr-Lys-Lys-Glu-Lys) flanked by Cys-138. According to Wagner, one of the main rules for NLS is the high proportion of positively charged amino-acids associated with Pro (25). Reversed NLS is connected with Pro 137. In three-dimensional structure Cys-138 is connected with Cys-29 which too has positive neighbours; may be it can lead to formation of three-dimensional NLS-like structure, though special investigations need to study this possibility.

## Acknowledgments

The authors would like to thank Dr. Igor Nikulin for help in computer design of this work and Mrs. Olga Sokirko for help in manuscript preparation.

## References

1. Beilharz, M.W., Nisbet, I.T., Tymms, M.J., Hertzog, P.J., and Linnane, A.W. (1986) J.Interferon Res. 6, 677-685.
2. Faltynek, C.R., and Baglioni, C. (1984) Microbiol. Sci. 1, 81- 85. 3. Faltynek, C.R., and Princler, G.L. (1986) J.Interferon Res. 6, 639-653.
4. Stiem, E.R., Kronenberg, L.H., Rosenblatt, H.M., Bryson, Y., and Merigan, T.C. (1982) Ann. Intern. Med. 96, 80-93.
5. Tanaka, M., Kimura, K., and Yoshida, S. (1987) Cancer Res. 47, 80-93.
6. Tokuda, Y., Ebina, N., and Golub, S.H. (1989) Cancer Immunol. Immunother. 30, 205-212.
7. Zasukhina, G.D., Makedonov, G.P., Shvetsova, T.P., Chekova, V.V., Andronova, A.V., and Alekhina, N.I. (1986) Journal of General Biology XLVII, 42-50 (In Russian).
8. Heremans, H., and Billiau, A. (1989) Drugs 38, 957-972.
9. Staeheli, P. (1990) Adv. Virus Res. 38, 147-200.
10. Arnheiter, H., and Haller, O. (1988) EMBO J. 7, 1315-1320.
11. Shan, B., Vazquez, E., and Lewis, J.A. (1990) EMBO J. 9, 4307- 4314.
12. Zavyalov, V.P., and Denesiuk, A.I. (1984) Proc. Natl. Acad. Sci. USSR 275, 242-246 (In Russian).
13. Sternberg, M., and Cohen, F. (1982) Int. J. Biol. Macromol. 4, 137-144.
14. Fish, E.N., Banerjee, E.N., and Stebbing, N. (1989) J. Interferon Res. 9, 97-114.
15. Borukhov, S.I., and Strongin, A.Ya. (1990) Biochem. Biophys. Res. Comm. 169, 282-288.
16. Raj, N.B.K., Israeli, R., Kelley, K.A., Leach, S.J., Minasian, E., Sikaris, K., Parry, D.A.D., and Pitha, P.M. (1987) J. Biol.Chem. 263, 8943-8952.
17. Senda, T., Matsuda, S., Kurihara H., Nakamura, K.T., Kawano, G., Shimizu, H., Mizuno, H., and Mitsui, Y. (1990) Proc. Japan. Acad. 66, Ser. B, 77-80.

18. Lipman, D.J., and Pearson, W.R. (1985) Science 227, 435-1441.
19. Pearson, W.R., and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444-2448.
20. Zuckerkandl, E. (1975) J. Mol. Evol. 7, No 1, 1-57.
21. Shulz, G.E., and Schirmer, R.H. In: Principles of protein structure (1982) 10, Moscow, Mir, (in Russian).
22. Kushnaryov, V.M., MacDonald, H.S., Lemense, G.P., DeBruin, J., Sedmak, J.J., and Grossberg, S.E. (1988) Cytobios 53, 185-197.
23. Roberts, B. (1989) Biochim. et Biophys. Acta. 1008, 263- 280.
24. Kubota, S., Siomi, H., Satoh, T., Endo, S., Maki, M., and Hatanaka, M. (1989) Biochem. Biophys. Res. Comm. 162, 963- 970.
25. Wagner, P., Kunz, J., Koller, A., and Hall, M.N. (1990) FEBS Lett. 275, 1-5.
26. Yoshimura, T., Fujisawa, J., and Yoshida, M. (1990) EMBO J. 9, 2537-2542.
27. Landschulz, W.H., Johnson, P.F., and McKnight, S.L. (1988) Science 240, 1759-1764.
28. Johnson, P.F., and McKnight, S.L. (1989) Annual Review of Biochemistry 58, 799-839.
29. O'Shea, E.K., Rutkowski, R., and Kim, P.S. (1989) Science 243, 538-542.
30. Landschulz, W.H., Johnson, P.F., and McKnight, S.L. (1989) Science 243, 1681-1688.
31. Abel, T., and Maniatis, T. (1989) Nature 341, No 6237, 24- 25.
32. Vinson, C.R., Sigler, P.B., and McKnight, S.L. (1989) Science 246, 911-916.