Chemistry of Natural Compounds, Vol. 34, No. 1, 1998

PEPTIDE DICTIONARIES FOR GROWTH FACTORS

D. Kh. Khamidov, R. S. Salikhov, and M. G. Khafizova

UDC 576.12:575.321

Peptide fragments in the amino acid sequences in various families of growth factors that are characteristic and unique for each group of proteins have been determined with the aid of computer methods. Peptide maps of growth factors have been compared. A method is proposed for identifying proteins on the basis of peptide maps.

Features of the primary structure of molecules of a definite family of proteins determine in considerable measure the uniqueness of their spatial configuration and the specificity of their physiological action. Within a family, a functional similarity of molecules is determined by the presence of common conservative fragments in their amino acid sequences. Some of these fragments may be characteristic, i.e., be present exclusively in proteins of a given family. For such fragments we have used the term "unique." It is just with these, in all probability, that the specific functions of a given family of proteins are connected. The aim of the present work was to find and make a comparative study of sets of peptide fragments unique for each family in the superfamily of growth proteins. We have previously determined peptide fragments unique for the family of nerve growth factors [1].

The result of the analysis is given in detail for, as an example, the family of growth factors from thrombocytes (PDGF, platelet-derived growth factor). PDGF is a powerful mitogen for cells of mesenchymal origin and is a dimer of disulfide-bound A- and B-chains.

Computer analysis of the available amino acid sequences of the A- and B-chains of PDGFs from various species of animals and from man have shown the presence in them of 10 conservative fragments, 8 of which form part of a linear consensus (Fig. 1).

All the proteins present in the Atlas of proteins were scanned for the presence in their amino acid sequences of peptides corresponding to the conservative fragments of the PDGFs. It must be mentioned that PIR1 – PIR3 may simultaneously contain information on the amino acid sequences of both a fragment and a whole protein molecule. Consequently, information on PDGFs is presented in 14 entries even though only 8 variants of PDGF from various sources are known.

We have found that the peptide WPPCVEV is unique for this group of proteins, i.e., it is absent from absolutely any other proteins represented in the Atlas. Other peptides are found in the sequences both of PDGF and of some other proteins. Thus, the GCCN fragment has been found in 44 entries, 14 of which relate to PDGF and the other 30 to different proteins (a precursor of alpha-fetoprotein, pyruvate decarboxylase, collagen-2, endothelial growth factor, and others).

The fragment VRK is found in 2903 entries. The full results of the analysis of the specificity of other peptides and their combinations are shown in Table 1.

Thus, we have established a definite peptide dictionary for PDGF: the peptide WPPCVEV and also the combinations CKTRT + ANFL, CKTRT + GCCN, and GCCN + ANFL are unique.

The principle of the analysis and comparison of peptide dictionaries given above have been used in work with other families of growth proteins, attention being concentrated on those proteins, information on which is given in fuller volume. The analysis of such proteins as, for example, ciliary neurotropic factor, with a single entry is difficult. The results of the investigations performed on the main groups of growth proteins are given in Table 2.

As a rule, each of the growth factors from different families contains a characteristic unique peptide or even several such peptides. An exception is the family of epidermal growth factors.

Institute of Biochemistry, Academy of Sciences of the Republic of Uzbekistan, 56 Abdullaeva Prosp., Tashkent.
Translated from Khimiya Prirodnykh Soedinenii, No. 1, pp. 89-93, January-February, 1998. Original article submitted December 9, 1997.

0009-3130/98/3401-0068$20.00

```
 45                  65                  85
1.-EKRPLPIRRKRSIEEAV**PAVCKTRT**VIYEIPRSQVDPTS**ANFLIWPPCVEV**K-
 47                  67                  87
2.-PPAPVPVRRKRTIEEA**IPAICKTRT**VIYEIPRSQVDPTS**ANFLIWPPCVEV**K-
 38                  58                  78
3.-EKRPVPIRRKRSIEEAI**PAVCKTRT**VIYEIPRSQVDPTS**ANFLIWPPCVEV**K-
 50                  70                  90
4.-EKRSVPSRRKRSVEEAV**PAICKTRT**VIYEIPRSQIDPTS**ANFLIWPPCVEV**K-
 46                  66                  86
5.-ARGRRSLGSLTIAE**PA**MIAE**CKTRT**EVFEISRRLIDRTN**ANFLVWPPCVEV**Q-
 50                  70                  90
6.-RSLGEAAGSPTVAE**PA**MIAE**CKTRT**EVFEVSRRLIDRTN**ANFLVWPPCVEV**Q-
 38                  58                  78
7.-SRGRRSLGSLAAAE**PA**VIAE**CKTRT**EVFQISRNLIDRTN**ANFLVWPPCVEV**Q-
 46                  66                  86
8.-SRGRRSLGSLAAAE**PA**VIAE**CKTRT**EVFQISRNLIDRTN**ANFLVWPPCVEV**Q-

          105                 125                 145
1.-**RCTGCCN**TSSVKCQPSRVHHRSVKVAKVEY**VRK**KPKLKEVQVR**LEEHL**ECAC-
          107                 127                 147
2.-**RCTGCCN**TSSVKCQPSRVHHRSVKVAKVEY**VRK**KPKLKEVQVR**LEEHL**ECAC-
          98                  118                 138
3.-**RCTGCCN**TSSVKCQPSRVHHRSVKVAKVEY**VRK**KPKLKEVQVR**LEEHL**ECAC-
          110                 130                 150
4.-**RCTGCCN**TSSVKCQPSRIHHRSVKVAKVEY**VRK**KPKLKEVLVR**LEEHL**ECTC-
          106                 126                 146
5.-**RCSGCCN**NRNVQCRPTQVQLRPVQVRKIEI**VRK**KPIFKKATVT**LEDHL**ACKC-
          110                 130                 150
6.-**RCSGCCN**NRNVQCRPTQVQLRLVQVRKIEI**VRK**RPVFKKATVT**LEDHL**ACKC-
          98                  118                 138
7.-**RCSGCCN**NRNVQCRASQVQMRPVQVRKIEI**VRK**KPVFKKATVT**LEDHL**ACKC-
          106                 126                 146
8.-**RCSGCCN**NRNVQCRASQVQMRPVQVRKIEI**VRK**KPIFKKATVT**LEDHL**ACKC-

                  165
1.-ATTSLNPDYREEDTGRPRES**GK**KRKRKRLKPT
                  167
2.-AASSAGPEHREEEADVR
                  158
3.-ATSNLNPDREEETGRRRES**GK**KRK
                  170
4.-TANSNSDYREEETGRTRET**GK**KRKRKKLKPT
                  166
5.-ETVAAARPVTRSPGGSQEQRAKTPQTRVTIRTVRVRRPPK**GK**RKFKHTHDKT-
                  170
6.-ETVVAARPVTRSPGSSQEQRARTPQTRVTIRTVRVRRPPK**GK**HQKFKHTHDK-
                  158
7.-ETVTPRPTRSPGTSREHRAKTPQTRVTVRTVRIRRPPK**GK**HRKFKHTHDKK
                  166
8.-ETITPRPVTRSPGTSREQRAKTPQARVTIRTVRIRRPPK**GK**HRKFKHTHDKA-
```

Fig. 1. Conservative fragments in the primary structures of PDGFs: 1) Human A-chain (*Homo sapiens*); 2) rabbit A-chain (*Oryctolagus cuniculus*); 3) rat A-chain (*Rattus norvegicus*); 4) xenopus A-chain (*Xenopus laevis*); 5) human B-chain; 6) cat B-chain (*Felis silvestris catus*; 7) rat B-chain; 8) mouse B-chain (*Mus musculus*). The conservative fragments are given in bold print.

The unique peptides may consist of different numbers of amino acids. It is obvious that the greater the number of amino acids in the protein the lower is the probability of finding this combination of amino acids in other proteins. We have established that the smallest number of amino acids in unique peptides is five. In particular, the unique peptide for the growth factor from erythrocytes consists of seven amino acids (WPPCVEV) as revealed in computer analysis. However, a further check showed that its fragments PPCVEV and WPPCVE, and even WPPCV, are also unique for this family of growth factors. Nevertheless, unique fragments of colony-stimulating growth factors and transforming growth factors beta, consisting of seven and six amino acids, lost their "uniqueness" when the fragment was diminished by even a single amino acid in any variant.

In the long conservative fragment CFHGTCRFLVQE of transforming growth factors alpha the following combinations are unique: CFHGT, FHGTC, HGTCR, GTCRF, and TCRFLV, i.e., in an analysis of seven possible combinations of five amino acids, five proved to be unique and concentrated in one part of the fragment. No regularities at all were found in the qualitative compositions of the amino acids in the unique peptides, although amino acid (C) — cysteine — was in fact found

TABLE 1. Number (N) of Proteins Revealed in the Atlas that Possess Fragments from the PDGF Peptide Dictionary in Their Primary Structures

Fragment	N	Fragment	N	Fragment	N
CKTRT	24	ANFL	165	GCCN	44
CKTRT+ANFL	14	ANFL+GCCN	14	GCCN+CKTRT	14
CKTRT+VRK	21	ANFL+VRK	41	GCCN+VRK	22
				VRK	2903

TABLE 2. Peptide Dictionaries of Growth Factors

Name of the factor	Set of consensus peptides	Unique peptide	Classification set of peptides
Growth factor from thrombocytes A- and B-forms	ANFL CKTRT WPPCVEV GCCN VRK	WPPCVEV	KCTRT + ANFL CKTRT + GCCN GCCN + ANFL.
Colony-stimulating factor from granulocytes	CPPTPET IPF QTRL SFK	CPPTPET	QTRL QTRL+SFK +IPF
Transforming growth factor, alpha-form	VVSHFN CPDSHTQ CFHGTCRFLV QE KPACVCHSGY . VG RCEHADLL	VSHFN PDSHTQ CPDCHT CFHGT FHGTC HGTCR GTCRF TCRFLV KPACVC PACVC HSGYV EHADLL RCEHA	
Transforming growth factor, beta-form	YCF NCC LYI WKW HEPKGY PCCV CPY IYN EQLSNM	HEPKGY	EQLSNM + PCCV EQLSNM+ CPY EQLSNM+ LYI EQLSNM+ LYN EQLSNM+ NCC EQLSNM+ WKW EQLSNM+ YCF PCCV + CPY PCCV + LYN PCCV + NCC PCCV + WKW PCCV + YCF PCCV + LYI.
Epidermal growth factor	DGYCL GVCM CNCV RCQ RDL		DGYCL+GVCM DGYCL+CNCV DGYCL+RCQ+RDL GVCM+CNCV GVCM+RCQ CNCV+RCQ CNCV-RDL
Insulin-like growth factor	ETLCG ELVD LQFVC RGFYF GIV ECCFRSCDL	LQFVC ECCFRS CCFRSC CFRSCD FRSCDL	RGFYF + GIV ETLCG + GIV

in each of them. It is possible that the specificity of the spatial configuration of the molecules is also connected with the unique peptides.

All the families of growth factors that have been analyzed may contain a set of two-three short conservative peptide fragments that are also characteristic. The presence of one or a combination of several peptides in an amino acid sequence in the given variants ("peptide dictionaries") may serve as an identification feature of a protein; for example, the presence of the peptide HEPKGY or the combination PCCV + CPY in an amino acid sequence is a necessary and sufficient indication that the protein belongs to the group of transforming growth factors beta.

EXPERIMENTAL

We used active data bases from the compact disc "Atlas of Protein and Genomic Sequences," version 03.93: PIR-1, including 11,252 systematized and annotated entries on amino acid sequences of proteins; PIR-2 with 27,383 annotated entries; and PIR-3 with 13,622 unchecked entries. Details about growth factors were extracted from the selected mass of information. A specialized information base on growth factors was created with the aid of the "Protein Adviser" packet of applied computer programs developed in the laboratory, and peptide fragments conservative for each group in their primary structures were determined. Among these fragments we selected only those consisting of three or more amino acids, which corresponds to the words of the Atlas computer program. The "uniqueness" of each conservative peptide or combination of peptides (peptide dictionary) was evaluated after the scanning of the whole mass of information in the Atlas.

The present work was performed under grant No. 112-96 of the Fund for Fundamental Investigations of the Academy of Sciences of the Republic of Uzbekistan.

REFERENCES

1. R. S. Salikhov, I. I. Parilis, M. G. Khafizova, E. Yu. Kazanov, and D. Kh. Khamidov, Mol. Biol. (Moscow), **28**, 201 (1994).
2. I. I. Parilis, E. Yu. Kazanov, R. S. Salikhov, and D. Kh. Khamidov, Biopolim. Kletka, **7**, 109 (1991).