Nucleic Acids Research, 2006, Vol. 34, Database issue ***D261–D266***
doi:10.1093/nar/gkj008

The EROP-Moscow oligopeptide database

Alexander A. Zamyatnin*, Alexander S. Borchikov, Michail G. Vladimirov and Olga L. Voronina

A.N. Bach Institute of Biochemistry, Russian Academy of Sciences, 33 Leninsky prosp., Moscow 119071, Russian Federation

Received July 29, 2005; Revised August 30, 2005; Accepted September 12, 2005

ABSTRACT

Natural oligopeptides may regulate nearly all vital processes. To date, the chemical structures of nearly 6000 oligopeptides have been identified from >1000 organisms representing all the biological kingdoms. We have compiled the known physical, chemical and biological properties of these oligopeptides—whether synthesized on ribosomes or by non-ribosomal enzymes—and have constructed an internet-accessible database, EROP-Moscow (Endogenous Regulatory OligoPeptides), which resides at http://erop.inbi.ras.ru. This database enables users to perform rapid searches via many key features of the oligopeptides, and to carry out statistical analysis of all the available information. The database lists only those oligopeptides whose chemical structures have been completely determined (directly or by translation from nucleotide sequences). It provides extensive links with the Swiss-Prot-TrEMBL peptide-protein database, as well as with the PubMed biomedical bibliographic database. EROP-Moscow also contains data on many oligopeptides that are absent from other convenient databases, and is designed for extended use in classifying new natural oligopeptides and for production of novel peptide pharmaceuticals.

INTRODUCTION

For more than a century, natural oligopeptides have attracted scientific attention (1) as biochemical regulators. The very first such oligopeptide, carnosine (β-ala-his), was discovered by Gulevitch and Amiradzhibi in 1900 (2), but its chemical structure was not determined until 1918 (3). Since that time, thousands oligopeptide regulators have been described, and now ~500 new natural oligopeptides emerge annually, out of a literature of >20 000 publications each year on oligopeptide chemistry and biology.

Regulatory oligopeptides generally do not exceed ~50 amino acid residues (4), and they differ substantially from larger polypeptides (proteins) in their physicochemical and biological properties. Specifically, the smaller peptides rarely possess strong enough intramolecular attractions to form stable globules (5), so they are able to shift configurations (6) and to fit themselves into specific receptor molecules, a process which is further aided by high diffusional mobility (7).

The terms 'natural' oligopeptide and 'regulatory' oligopeptide can be considered synonymous. Few integrated biological processes are known which are not regulated, or at least modulated, by small peptides. Such roles are especially well known in the regulatory organ systems, viz. nervous, endocrine and immune systems (1), but their functions extend well beyond the bounds of single organ systems or even of single biological species. Antimicrobial oligopeptides produced by prokaryotes themselves, for example, regulate competition for ecological niches and simultaneously function as signaling molecules for species-specific intercellular communication (8). And even eukaryotic oligopeptide toxins seem to play important roles in regulating interspecies reactions (9).

It has been clear for >15 years that detailed understanding of the complex regulatory processes involving oligopeptides requires a system for classifying these molecules and for cataloguing their major properties. Our first attempt at such a system, in 1991 (4,10), yielded the MS-DOS version of EROP-Moscow, which contained structures, functions and sources of the oligopeptides then known. That database, however, was not widely accessible to the research community. In the meanwhile, a number of extensive and highly utilized peptide–protein databases have been created (e.g. Swiss-Prot-TrEMBL), but their data on small natural peptides is far from comprehensive and constitutes only a small fraction of their total information, so that retrieving relevant oligopeptide data from them can be laborious and excessively time-consuming period.

Here we present an internet version of the compact specialized database EROP-Moscow, recreated to provide a comprehensive description of all presently known natural oligopeptides. Neuropeptides, peptide hormones, antimicrobial agents and toxins represent the largest functional classes.

*To whom correspondence should be addressed. Tel: +7 095 9543066; Fax: +7 095 9542732; Email: aaz@inbi.ras.ru

This database should now enable investigators to search easily for oligopeptides via a wide variety of different features, to compare their properties quickly, and to retrieve statistics about all relevant information in the database.

INFORMATION SOURCES

Since the objectives of an internet version of EROP-Moscow are to collect and disseminate all essential information about currently known oligopeptides, authenticity is of paramount importance. Therefore, all information in this database has been extracted directly from the primary sources, the great majority of which are publications in scientific journals. More than 100 journals in biochemistry, biophysics, physiology, genetics and general biology are being continuously screened and descriptions of the structures of newly found natural oligopeptides are being retrieved. In many of these articles, the authors compare the novel structures with known ones and provide references to publications not included in our systematic screening. Such publications then become an additional source of primary information for the EROP-Moscow. Finally, initial reports on novel oligopeptides are sometimes found in book chapters, patent descriptions and other protein-peptide databases, and these sources are used, as well, and are appropriately documented. The total number of useful sources for basic information on natural oligopeptides is now >250.

In addition to the client–server features of EROP-Moscow, a library of publications has been created, containing the actual primary descriptions of oligopeptides, along with their pdf files.

SELECTION OF OLIGOPEPTIDES

Only those oligopeptides are entered into EROP-Moscow, whose chemical structures have been completely determined (either directly or by translation from nucleotide sequences), and can be described by the standard single-letter amino acid code. Although most peptides included in this version of EROP-Moscow are formed by ribosomal synthesis, a small number formed by non-ribosomal enzymes (11)—mostly from bacteria and fungi—are also included, provided they comprise residues fitting the standard one-letter code. Oligopeptides with still ambiguous structures, such as asparagine/aspartic acid or glutamine/glutamic acid at single residues, have been deliberately excluded from this database, as have artificially synthesized molecules that are not found in nature.

ORGANIZATION OF EROP-MOSCOW

The EROP-Moscow database presents multilevel bioinformation via an HTML-based interface. This interface includes: Home page, Query page, Peptide page, Results page, Family page and Statistics pages (Figure 1). All of these pages contain internal EROP-Moscow links (including Home page, Site map, Contact us and Help) as well as links to external databases such as Swiss-Prot, Protein Identification Resource (PIR), PDB and PubMed.

The following programming elements, freely available on the basis of the GNU License, have been used as server software, and these are updated as new versions appear.

(i) MySQL database server, version 4.0.14-max;

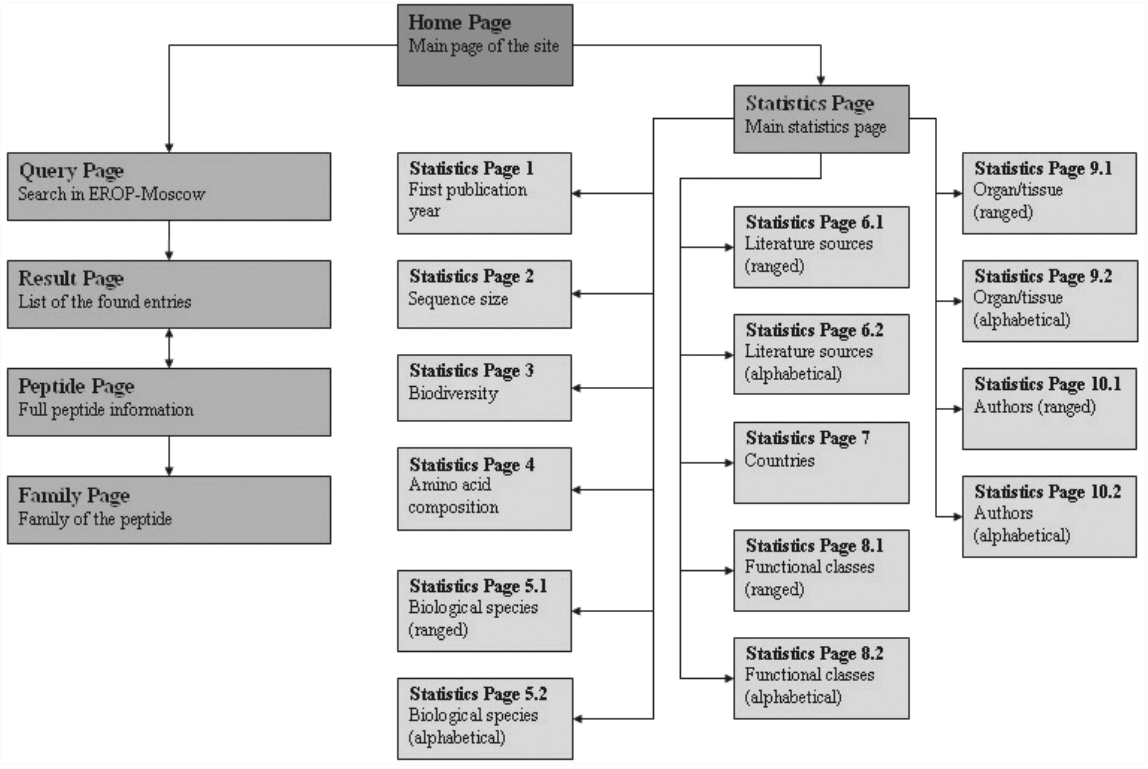

Figure 1. EROP-Moscow architecture illustrated on the Site Map page.

(ii) Apache web server, version 2.0.47, compiled with PHP support;
(iii) PHP language, version 4.0;
(iv) Remote management server with the HTML interface WebMin, version 1.0.70.

The basic operational unit of EROP-Moscow is an entry (= record). Each individual entry describes the physical, chemical and biological characteristics of one unique natural oligopeptide. Each entry is tagged by a unique accession number, beginning with character 'E' (from EROP) followed by five numerical digits.

Individual sequences which are found in multiple organisms, even taxonomically remote ones, are presented only once in EROP-Moscow, with the names of all known organisms possessing that oligopeptide being suspended.

On the other hand, oligopeptides existing in two or more distinct chemical modifications (usually accompanied by clear functional differences) are registered as separate entries and are assigned distinct names and accession numbers. Good examples of this are two natural chemical forms of gastrin: one having a simple tyrosine residue and the other, a sulfated tyrosine residue (12).

Home page

Users would normally enter EROP-Moscow via the site address http://erop.inbi.ras.ru. The Home page lists various information about the database itself, including the date of the most recent version, the current number of entries, on-going changes in the content (EROP-news), the list of database authors and some descriptive information. Home page is linked to the Query page and to Statistics pages for individual peptides. A Contact-us button facilitates ordinary E-mail messages and inqueries to the database manager.

Query page

The Query page, entered from Home page via the 'Query page' button, provides a rapid search for the peptide records signaled by specific characteristics. These characteristics are subdivided into the following groups: general information (such as oligopeptide name or accession number), organismic classification (including multiple trivial species names), physicochemical properties (such as partial amino acid sequences), biochemical or biologic functions and literature references. Query examples (single words, phrases or numbers) are provided adjacent to each Query window, and pull-down menus are provided with most query options. An 'Abbreviations' button, located near the Query window for amino acid sequence, elicits display of the standard one-letter code for amino acid residues, along with optional abbreviations.

Because some oligopeptides, particularly the smallest ones, are chemically modified at the N- and C-termini, six more symbols augment the standard one-letter code. These are:

(i) '+' to denote $^{+}H_2$, which is the open N-terminus,
(ii) 'b' for an acetyl residue or other chemical group at the N-terminus,
(iii) '−' to denote O^{-}, which is the open C-terminus,
(iv) 'z' for an amide bond at the C-terminus,
(v) 'J' to denote the pyroglutaminyl linkage, formed by an N-terminal glutamine, owing to side-chain reaction with the terminal amine residue (13,14), and finally,
(vi) 'U' for the (occasional) aminoisobutyric acid residue.

Results page

After entering a search word or phrase on the Query page, the user should click on the 'Submit query' button, which initiates the search and returns with the Results page, containing a list of oligopeptides that meet the specified characteristics. Each item in this list will contain the preferred name of oligopeptide, the trivial and taxonomic names of organisms where the peptide has been identified, and the accession number. The accession number, in turn, links to the appropriate Peptide page (record). When the query returns only a single oligopeptide, its record opens immediately.

Peptide page

This page, reached via accession number, presents the collected data on each oligopeptide, including the number of amino acid residues, primary structure, precursors, known posttranslational modification(s), affinity to any definite structure–function family, taxon(s) of biological sources, tissue/cell localization in each organism, major known biological functions, molecular mass (Da), isoelectric point, pI (calculated and experimentally observed), literature sources and linking accession numbers (if any) in other peptide-protein databases (see above) or PubMed.

Family page

Tentative homologous family assignments, for each oligopeptide in EROP-Moscow, have been developed by sequence alignment, and the entire family can be reached from the Peptide page via a 'View family' button. Equally located amino acid residues are highlighted in red and the attached oligopeptide name for each sequence links back to the appropriate Peptide page.

Statistics pages

A special set of pages is devoted to the overall characteristics of data on oligopeptides listed in EROP-Moscow. The starting Statistics page is reached from Home page via the 'EROP Statistics' button, and it contains the list of statistical parameters compiled, each named parameter being a link to one of 15 additional Statistics pages (pp. 1–10.2). These in turn present graphic and tabular information on oligopeptides currently available in EROP-Moscow, information including:

(i) a chronological diagram, by years, for decoding chemical structures of new oligopeptides (Figure 2),
(ii) size distribution of oligopeptides (number of amino acid residues; Figure 3),
(iii) current numerical yield of oligopeptides per taxonomic group,
(iv) total amino acid residue content of all listed oligopeptides,
(v) relative contributions of international scientists, by home country, to the discovery of new oligopeptide structures,

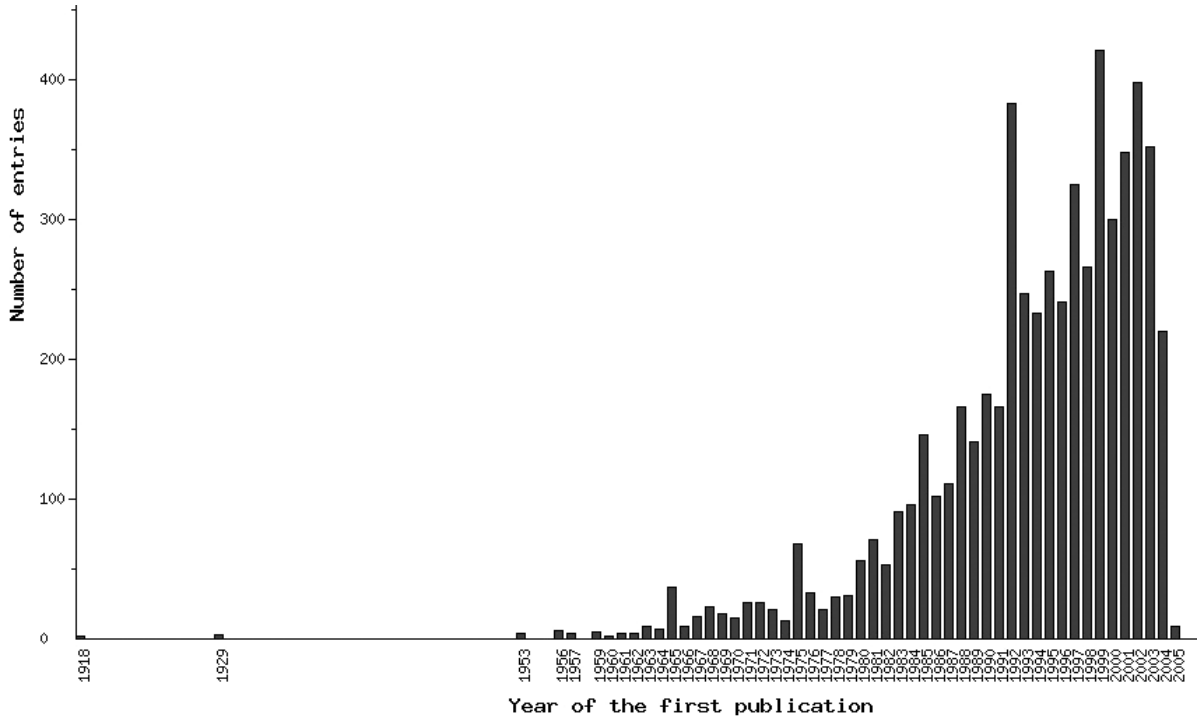

Figure 2. Increase in the number of decoded amino acid sequences of the natural oligopeptides (Statistics page 1).

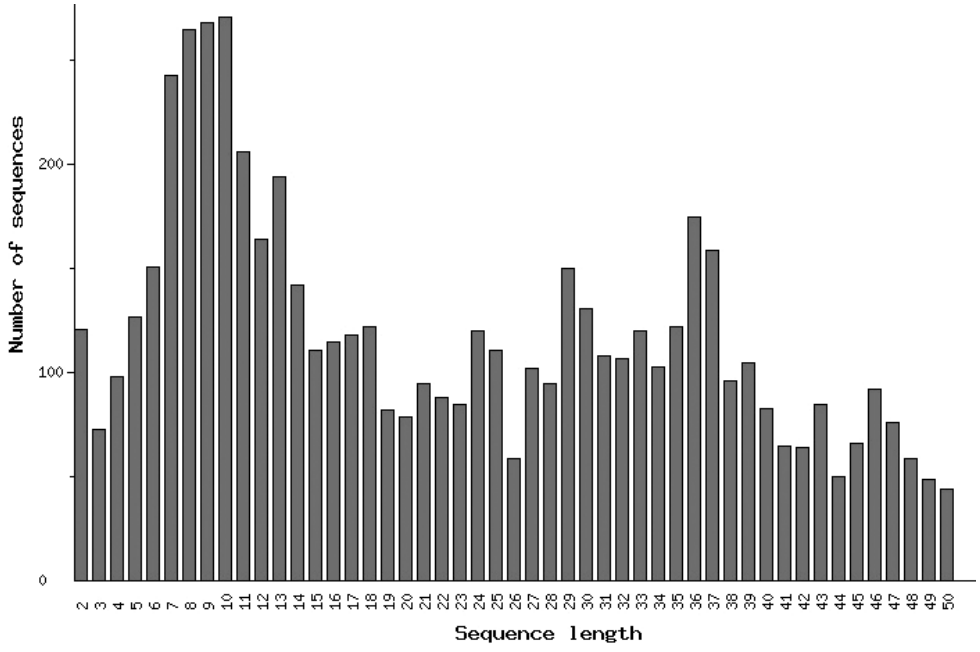

Figure 3. Distribution of known oligopeptides by amino acid residue number (Statistics page 2).

(vi) organisms covered (>1000), and tissues and organs (>500),
(vii) functional classes of oligopeptides (~100),
(viii) primary literature sources (>300), and
(ix) authors of the original publications devoted to decoding of oligopeptide chemical structures (>8000).

This statistical information demonstrates, for example, that the greatest number of natural oligopeptides have been identified in humans, especially in the human brain; to date neuropeptides represent the single largest functional class. The data also show that the most prolific journal for publication of new oligopeptides is the *Journal of Biological*

Chemistry and that an American laboratory, that of J. M. Conlon, leads the discovery of new oligopeptides.

SERVICE MODULES

The EROP-Moscow database accesses a set of special software tools for alignment, and for calculating molecular masses and isoelectric points. These operations are executed outside the EROP-Moscow site, and results are returned to EROP-Moscow (as a part of its update capability) and displayed on the Peptide page and the Family page.

This system creates dynamic web pages using *cgi*-scripts written in the PHP language. In response to each appropriate user query, the required HTML-pages are generated interactively. Once the user's web browser has sent the HTTP query to the web server, the required script containing the database query is executed. After the survey of needed records, the PHP script dynamically generates the results, in the form of an HTML-page sent to the user's computer.

Graphic information is also generated and plotted dynamically, e.g. in the statistical processing of data, thus permitting online display of statistical summaries for all natural oligopeptides currently available in the EROP-Moscow database.

COMPARISON OF EROP-MOSCOW WITH OTHER PEPTIDE–PROTEIN DATABASES

The internet now provides free access to a large number of both generalized and specialized peptide-protein databases. Best known among the generalized databases are PIR (15) and Swiss-Prot (Swiss Protein), which is linked with the database of amino acid sequences translated from nucleotides TrEMBL (Translated, European Molecular Biology Laboratory) (16). Smaller databases, containing information about selected classes of oligopeptides, include Peptaibol (Peptide aminoisobuturic, for data on peptides possessing at least one residue of aminoisobutyric acid) (17), ANTIMIC (ANTIMIcrobial, concerned with antimicrobial peptides) (18) and SCORPION (19), especially created for the peptide-protein toxins from a single order of arachnoids, the scorpions.

About half the natural oligopeptide structures now available in EROP-Moscow, however, are absent from the above-named databases, for several reasons, especially (i) that little attention is paid to oligopeptides formed by means of non-ribosomal or pure enzymatic synthesis, and (ii) that precursor-product series are not handled systematically. In particular, very many oligopeptides generated as natural fragments of large precursors are not specifically indexed in the above databases and can be found there only by their amino acid sequences. Swiss-Prot contains one record (P01019), e.g. on human angiotensinogen, which includes the amino acid sequences for angiotensins I and II (20), but omits angiotensins V and VI (21). EROP-Moscow contains these (records E00165 and E00166), as well as the more familiar oligopeptides.

In addition, EROP-Moscow lists a considerable number of oligopeptides with unique amino acid sequences that are simply not included in the other databases—owing either to the source journals (e.g. *Biological Bulletin*) being out of view of most database managers, or to failure in tracing primary sources keyed from secondary publications.

Table 1. Comparison of numbers of oligopeptides (from 2 to 7 amino acids) contained in Swiss-Prot/TrEMBL (Release of July 19, 2005) and EROP-Moscow (Release of May 17, 2005)

Amino acid residue number	Oligopeptide number Swiss-Prot-TrEMBL	EROP-Moscow
2	1	120
3	6	71
4	25	90
5	28	120
6	19	139
7	106	232

Information on very short oligopeptides is particularly deficient in the other major databases. Table 1 displays a useful comparison of these (di- to hepta-) peptides in EROP-Moscow, versus Swiss-Prot.

CONCLUDING REMARKS

The database EROP-Moscow has been developed for simple and rapid retrieval of information on natural regulatory oligopeptides and their structurally homologous families. In addition to solving true informational problems, EROP-Moscow can serve as a basis for new research and for elucidating general principles of structural and functional organization for these substances. For example, the current size distribution of oligopeptides, by number of amino acid residues (Figure 3), shows a numerical peak ~8–10 residues, but this peak has no proper rationale at present. Study of structurally homologous families should facilitate prediction of the functional properties of newly found oligopeptide molecules, should provide bases for classifying newly discovered molecules and in turn should promote creation of novel, highly efficient pharmaceutics derived from the natural regulatory oligopeptides.

Because the discovery of new regulatory oligopeptides is a vigorous and continuing process (Figure 2), continuous revision and upgrading of EROP-Moscow will be essential. In this cause, we would ask users of EROP-Moscow to alert us to newly discovered oligopeptides which may not yet have been entered into the EROP-Moscow database. For this purpose, the Contact-us button on Home page should be convenient.

CITING EROP-MOSCOW

Users of the EROP-Moscow database are asked to cite this paper, in their relevant published research.

ACKNOWLEDGEMENTS

We thank Prof. Clifford Slayman, of Yale University, for fruitful discussions and comments on the manuscript, two anonymous reviewers for their valuable suggestions, and Margarita Il'ina for technical assistance. This work has been supported by Grant 02-07-90175 from the Russian Foundation for Basic Research (RFBR) and by the Program 'Molecular and Cellular Biology', RAS-10P (Russian Academy of Sciences). Funding to pay the Open Access publication charges for this article was provided by A.N. Bach Institute of Biochemistry, Russian Academy of Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Sewald,N. and Jakubke,H.-D. (2002) *Pepides: Chemistry and Biology*. WILEY-VCH Verlag, GmbH, Weinheim.
2. Gulevitch,V.S. and Amiradzhibi,S. (1900) Ueber das Carnosin, eine neue organishe Base des Fleischextrakt. *Deutsch. Chem. Ges*., **33**, 1902–1903.
3. Baumann,L. and Ingwaldsen,T. (1918) Conserning histidine and carnosine. *J. Biol. Chem*., **35**, 263–276.
4. Zamyatnin,A.A. (1991) EROP-Moscow specialized data bank for endogenous regulatory oligopeptides. *Protein Seq. Data Anal*., **4**, 49–52.
5. Privalov,P.L. (1985) Energy characteristics of the structure of protein molecules. *Biophysics (Moscow)*, **30**, 722–733.
6. Karle,I.L. (1981) X-ray analysis conformation of peptides in the crystalline state. In Gross,E. and Meienhofer,J. (eds), *The Peptides: Analysis, Synthesis.* Academic Press, NY, pp. 1–54.
7. Zamyatnin,A.A. (2003) Biophysical problems of oligopeptide regulation. *Biophysics (Moscow)*, **48**, 950–958.
8. Woo,P.C., To,A.P., Lau,S.K. and Yuen,K.Y. (2003) Facilitation of horizontal transfer of antimicrobial resistance by transformation of antibiotic-induced cell-wall-deficient bacteria. *Med. Hypotheses*, **61**, 503–508.
9. Zamyatnin,A.A. (1996) Physicochemical and biological features of endogenous oligopeptide toxins. *Neirokhimia*, **13**, 243–259, (in Russian).
10. Zamyatnin,A.A. (1991) Structural classification of endogenous regulatory oligopeptides. *Protein Seq. Data Anal*., **4**, 53–56.
11. Egorov,N.S., Silaev,A.B. and Katrukha,G.S. (1987) *Antibiotics-polypeptides (structure, function, biogenesis).* Moscow University Publishers, Moscow, p. 263.
12. Bentley,P.H., Kenner,G.W. and Sheppard,R.C. (1966) Structures of human gastrins I and II. *Nature*, **209**, 583–585.
13. Kizer,J.S., Busby,W.H., Cottle,C. and Youngblood,W.W. (1984) Glycine-directed peptide amidation: presence in rat brain of two enzymes that convert p-Glu-His-Pro-Gly-OH into p-Glu-His-Pro-NH2 (thyrotropin-releasing hormone). *Proc. Natl Acad. Sci*., **81**, 3228–3232.
14. Bateman,R.C., Yougblood,W.W., Busby,W.H. and Kizer,J.S. (1985) Nonenzymatic peptide alpha-amidation. Implications for a novel enzyme mechanism. *J. Biol. Chem*., **260**, 9088–9091.
15. Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.W., Orcutt,B.C., Pfeiffer,F., Tsugita,A. *et al*. (2001) Protein information resource: a community resource for expert annotation of protein data. *Nucleic Acid Res*., **29**, 29–32.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acid Res*., **31**, 365–370.
17. Whitmore,L. and Wallace,B.A. (2004) The peptaibol database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acid Res*., **32**, D593–D594.
18. Brahmachary,M., Krishnan,S.P.T., Koh,J.L.Y., Khan,A.M., Seah,S.H., Tan,T.W., Brusic,V. and Bajic,A.V. (2004) ANTIMIC: a database of antimicrobial sequences. *Nucleic Acid Res*., **32**, D586–D589.
19. Srinivasan,K.N., Gopalakrishnakone,P., Tan,P.T., Chew,K.C., Cheng,B., Kini,R.M., Koh,J.L., Seah,S.H. and Brusic,V. (2002) SCORPION, a molecular database of scorpion toxins. *Toxicon*, **40**, 23–31.
20. Skeggs,L.T., Lentz,K.L., Kahn,J.R., Shumway,N.P. and Woods,K.R. (1956) *J. Exp. Med*., **104**, 193–197.
21. Semple,P.F., Boyd,A.S., Daves,P.M. and Morton,J.J. (1976) Angiotensin II and its heptapeptide (2–8), hexapeptide (3–8), and pentapeptide (4–8) metabolites in arterial and venous blood of man. *Circ. Res*., **39**, 671–678.