2. M. D. Mashkovskii, Drugs [in Russian], Meditsina, Moscow, Part 2 (1984), p. 134.
3. A. M. Sobolev, Usp. Biol. Khim., 4, 248 (1983).
4. L. F. Johnson and M. E. Tate, Can. J. Chem., 47, No. 1, 63 (1969).
5. R. E. Grosselin and E. R. Coghean, Arch. Biochem. Biophys., 45, No. 2, 301, 318 (1953).
6. P. Karrer, Organic Chemistry, Elsevier, Amsterdam (4th English edn.) (1950) [Russian translated from the 13th German edition, Goskhimizdat, Leningrad (1963), p. 821].
7. The Merck Index: An Encyclopedia of Chemicals and Drugs, 8th edn., Merck, Rahway, N.J. (1968), p. 829.
8. K. Nakanashi, Infrared Absorption Spectroscopy. Practical, Holden-Day, San Francisco (1962) [Russian translation, Mir, Moscow (1965), p. 209].
9. L. Bellamy, Infrared Spectra of Complex Molecules, 2nd edn., Methuen, London/Wiley, New York (1958) [Russian translation, IL, Moscow (1963), p. 590].
10. J. Lewis and R. Wilkins, Modern Coordination Chemistry, Interscience, New York (1960) [Russian translation, IL, Moscow (1963), p. 270].
11. Yu. N. Kukushkin and R. I. Bobokhodzhaev, Chernyaev's Law of Trans Influence [in Russian, Nauka, Moscow (1977), p. 64].
12. F. Basalo and R. C. Johnson, Coordination Chemistry: The Chemistry of Metal Complexes, W. A. Benjamin, New York (1964) [Russian translation, Mir, Moscow (1966), p. 168].
13. A. D. O. Troitskaya, The Trans Influence of Organophosphorus Ligands in Platinum-Metal Complexes. The Reactivity of Coordination Compounds. Problems of Coordination Chemistry Series [in Russian], Nauka (1976), p. 7.

ANALYSIS AND PREDICTION OF PRIMARY STRUCTURES OF PROTEINS USING A COMPUTER

L. Ya. Yukel'son, I. I. Parilis,
G. L. Bussel', and D. Kh. Khamidov

UDC 577.007

On the basis of a computer analysis of four classes of toxic polypeptides, generalized primary structures of each class have been obtained which can be used for planning the synthesis of biologically active compounds.

The increase in the number of deciphered amino acids and nucleotide sequences has made it necessary to create special automated systems for their storage, search, and processing. Banks of nucleotide sequences have been created in the USA (Los Alamos Bank, the Dayhoff Bank in Washington), the FRG (Heidelberg Bank of EMBL, GenBank, the PIR Protein Data Base (NBRSF)), and other countries. Here in Moscow we have created the All-Union Bank of Nucleotide Sequences. Data bases exist in IMG AN SSSR, [Institute of Molecular Genetics of the USSR Academy of Sciences], IMB AN SSSR [Institute of Molecular Biology of the USSR Academy of Sciences], NIVTs AN SSSR [Scientific-Research Computer Center of the USSR Academy of Sciences], and VINITI [All-Union Institute of Scientific and Technical Information] and the formation is taking place of a data bank on molecular biology of ITsiG SO AN SSSR [Institute of Cytology and Genetics, Siberian Division of the USSR Academy of Sciences]; other banks are also being created. A basis for the spatial structures of biological macromolecules already exists — the National Protein Data Bank, containing the atomic coordinates of about 220 proteins. Here the rate of accumulation of information is 10-20 proteins per year, which is considerably lower than the rate of acquisition of information on primary structures [1-5].

Together with these "accumulators" of data, bibliographic information exists on publications in the field of molecular biology and genetics which contains thousands of names of sequences. They all solve problems of the collection and processing of published sequences, the input, editing, and storage of the corresponding information, and the systematization and search for required sequences. However, in some cases possibilities are being considered of the prediction of the chemical structures of proteins from the aspect of a definite function. In connection with this, the question of the generation of structures with predetermined properties requires special consideration.

Institute of Biochemistry, Uzbek SSR Academy of Sciences, Tashkent. Translated from Khimiya Prirodnykh Soedinenii, Vol. 6, pp. 822-830, November-December, 1989. Original article submitted September 13, 1988; revision submitted April 3, 1989.

0009-3130/89/2506-0698$12.50

Existing information-searching systems (ISSs) ensure a fairly effective search for concrete primary structures or groups of them united by some characteristic or other. The most complete information on primary structures of proteins is contained in M. O. Dayhoff's Atlas [6]. They are stored in the data bank of the US National Biomedical Research Foundation, which contains about four thousand texts with a total length of 500 thousand amino acids, and are grouped into individual isofunctional families.

In [6] the primary structures of proteins are separated into several taxonomic groups according to the degree of their difference: the superfamily (<85-90% differences), the family (<50%), and the subfamily (<20%). According to this publication, proteins are subdivided into 181 superfamilies, 314 families, and 537 subfamilies. The superfamilies are combined into 24 groups of proteins with a complex biochemical function. Among the best-studied superfamilies are the cytochromes C (>90 representatives).

We have considered the structures that are combined into a single class (taxonomic group) by functional characteristics. The amino acid sequences of the exotic proteins considered in this paper, to which the toxic polypeptides from the venoms of a number of snakes and scorpions belong, are far from completely represented in data banks. Some of them have been deciphered in the Institute of Biochemistry of the Uzbek SSR Academy of Sciences together with the M. M. Shemyakin Institute of Bioorganic Chemistry of the USSR Academy of Sciences. To reveal, compare, and generalize the information concealed in genetic texts of these proteins is impossible without the use of a computer.

It must be mentioned that, while earlier, theoretical investigations had a general nature, at the present time the significance of concrete models deliberately affecting the tactics of experiments and possessing predictive possibilities and economic efficacy has increased. Banks of experimental results are continuously being supplemented, which makes it possible to check theoretical laws established beforehand. All these investigations performed with the aid of computers require the development of special programs constituting their mathematical provisions; in the final account, the work amounts to the solution of identical general problems out of the field of comparing amino acid sequences and presupposes the discovery of homologous sections of different lengths with a predetermined percentage of accuracy, the optimum comparison of two sequences according to predetermined criteria, and their equalization with the aid of deletions. All these operations differ from one another only by the objects of investigation and the characteristics through which the best classification is achieved.

In the comparison of programs for the systematic analysis of the initial information, we have made use of the methods of graph theory to construct phylogenetic trees [3, 7, 8], the methods of graph theory of probability and mathematical statistics to determine the generalized characteristics of families [9], the algorithms of pattern recognition theory [10, 17] for the solution of the problems of classification and prediction, and the method of quantitative structure-activity relationships (QSAR) in the construction of amino acid sequences of hypothetical compounds [11, 12]. The primary structures of four families of toxic polypeptides [13, 14] have been investigated: short and long neurotoxins (SNTs and LNTs), the cytotoxins (CTs) of cobra venom, and the toxins from scorpion venom (SCOs). They have been combined into the corresponding homologous groups according to the principle of functional closeness, but there are differences in the writing of their structures. By arranging the sequences of the different toxins one above the other we obtain a table the elements of which are symbols coding amino acids [1, 6, 14, 15].

All the primary structures are written in the IUPAC one-letter code. Table 1 also gives, in addition to this code, the triplets of the four nucleotides A, G, U, and C coding each amino acid and show their functional affinities according to their physicochemical properties.

The four families of homologously arranged amino acid sequences (Table 2) served as the starting material for comparative analysis.

In the comparison of proteins, great importance is attached to the calculation of the distances between them, i.e., between their primary structures [5, 7, 8, 15, 17]. Such an analysis of the measure of closeness is important for a multiplicity of tasks in which the initial material is specified by various descriptions.

To solve the problems connected with evolutionary biology use is frequently made of the idea of the minimum mutational distance, i.e., the smallest number of nucleotides that must

TABLE 1. Coding of the Amino Acids and Their Functional Affinities

Amino acid	Three-letter code	One-letter code	Coding triplets	Functional affinity
Alanine	Ala	A	GC (−)*	Small nonpolar
Cysteine	Cys	C	UG (U,C)	Small monpolar
Aspartic	Asp	D	GA (U,C)	Small polar
Glutamic	Glu	E	GA (A,G)	Large polar
Phenylalanine	Phe	F	UU (U,C)	Large nonpolar
Glycine	Gly	G	GG (−)	Small polar
Histidine	His	H	CA (U,C)	Medium polar
Isoleucine	Ile	I	AU (U,C)	Large nonpolar
Lysine	Lys	K	AA (A,G)	Large polar
Leucine	Leu	L	UU (A,G)	
			CU (−)	Large nonpolar
Methionine	Met	M	AU (A,G)	Large nonpolar
Asparagine	Asn	N	AA (U,C)	Small polar
Proline	Pro	P	CC (−)	Small nonpolar
Glutamine	Gln	Q	CA (A,G)	Large nonpolar
Arginine	Arg	R	CG (−)	
			AG (A,G)	Large nonpolar
Serine	Ser	S	AG (U,C)	
			UC (−)	Small polar
Threonine	Thr	T	AC (−)	Small nonpolar
Valine	Val	V	GU (−)	Large nonpolar
Tryptophan	Trp	W	UGG	Medium polar
Tyrosine	Tyr	Y	UA (U,C)	Medium polar

*In the "coding triplets column, (−) means that any of the four nucleotides may be present in the third position.

TABLE 2. Generalized Structure of the Four Families of Toxic Polypeptides

I. LONG NEUROTOXINS

```
ITCFXXXITPDITSKTCPPGENICYTKTWCDAFCSSRGKRVDLGCAATCPKVKPGVDIKCCSTDNCNIFPTXPKKP
RR Y   K  SVK        XHV        GW        E        T  T  E                   P
```

II. SHORT NEUROTOXINS

```
L                                            Q
MICHNQQSSQPPTTKTCPXGETSCYKKQWSDHRGTIIERGCGCPSVKPGIKLNCCTTDKCNN
R                T      NN   T R      R      T           SN
```

III. CYTOTOXINS

```
LKCXNKLVPPFWKTCPEGKNLCYKMFMVSXTPTVPVKRGCIDVCPKSSLLVKYVCCNTDKCN
       I   Y   A                           N A        R
```

IV. SCORPION VENOM TOXINS

```
XVKDGYIVDDKNXCVYFCXXXGRNAYCNGECKKKXXGGSSGYCQWLGPYGNACWCYKLPDNVPIKXXLPGKXCHX
 XR   LA P  G T H   L      DDL T N   AE     FAX S F      D      T      NR  N
         K    K     P       E
```

be substituted to conver the codon of one amino acid into another. The minimum mutational distances (MMDs) are expressed numerically by values of from 0 to 3 and, with the aid of this measure, problems connected with the structure of the evolutionary tree for various families of proteins and the restoration of the amino acid sequences of their ancestral forms are being solved [3, 6, 8, 15]. Evolutionary trees for the families that we have considered have been given a number of publications [3, 6-8, 15, 18].

Sometimes, a matrix taking into account the functional closeness of the amino acids is included in the analysis. It is based on a comparison of their main physicochemical properties such as hydrophobicity, polarity, charge, etc. Different authors consider the corresponding numerical characteristics differently and then, of course, they obtain different matrices.

Nevertheless, trees constructed from other matrices have proved to be topologically similar, and, consequently, the general pattern of the process of divergence is reproduced fairly stably and they accurately reflect the objective laws of the process of evolution. Each tree is its graphical model, and here constancy of the accumulation of mutations is postulated.

Information on primary structures can be written compactly with some threshold of disregard in the form of a branched structure reflecting the percentage content of all the amino acids in each position. The advantage of such a description is obvious, since with its aid a family of amino acid sequences difficult to analyze is replaced by a single "generalized class" into which, as new structures are decoded, only small corrections will be introduced. The comparison of different isofunctional families is considerably simplified. This decription gives the initial information for the "computer streamlining" of the hypothetical primary structures of proteins. The method is based on algorithms for the construction of drugs [11, 12].

Particular interest is presented by invariant positions the combination of which forms part of the structure which is invariable in the process of evolution and resistant to mutations. Various authors give it different names: acceptor [5], consensus [14], prototype [3], and signature [16].

In positions not present in a consensus, the diversity is greater than unity. If an amino acid is rarely found in a position, i.e., if it is unique, it is excluded from further consideration by the introduction of a threshold of disregard. As the result of such a procedure, a multiplicity of hypothetical amino acid sequences compactly written in the form of branched structures is obtained (Table 2). From them a computer program "generates" the primary structures of new compounds close to the existing proteins of each family. As a rule, a very large number of structures is obtained which must be decreased with the aid of a selection procedure that consists in screening for primary structures and their amino acid compositions. The former sets demands for the observance of a number of statistical laws in the new set of structures close to the initial one. Let us consider screening based on limitations of the amino acid composition, which is one of the main biochemical parameters.

Corresponding to each protein are set 20 numbers – percentages of all 20 amino acids – and this forms a point in a 20-dimensional space. In it the points corresponding to similar proteins are generally located in compact fashion, forming a number of clusters characterized by a definite basic function.

To each class corresponds a 20-dimensional parallelepiped given by the intervals of variation of the coordinates – from the minimum to the maximum values – by which the amino

TABLE 3. Reference Parallelepipeds of the Classes

Class	1	2	3	4	5	6	7	8	9	10
	A	C	D	E	F	G	H	I	K	L
LNTs	$\frac{2,8^*}{6,8}$	$\frac{13,7}{15,2}$	$\frac{2,7}{9,6}$	$\frac{0}{6,9}$	$\frac{0}{4,2}$	$\frac{4,2}{8,3}$	$\frac{0}{2.7}$	$\frac{2,7}{7,0}$	$\frac{5,6}{13,9}$	$\frac{1,4}{4,2}$
SNTs	$\frac{0}{1,7}$	$\frac{12,9}{15,0}$	$\frac{1.6}{5,0}$	$\frac{3,2}{6,7}$	$\frac{0}{3,2}$	$\frac{6,7}{11,3}$	$\frac{1.6}{6,7}$	$\frac{1,6}{9.8}$	$\frac{4,8}{11,7}$	$\frac{0}{3\ 3}$
CTs	$\frac{0}{6,7}$	$\frac{12,9}{13,3}$	$\frac{0}{5.0}$	$\frac{0}{3.3}$	$\frac{0}{4,9}$	$\frac{3,2}{10,9}$	$\frac{0}{1,7}$	$\frac{1,7}{6,7}$	$\frac{6,5}{19,7}$	$\frac{2,2}{11,5}$
SCOs	$\frac{0}{9,2}$	$\frac{12,1}{12,7}$	$\frac{3\ 0}{9,4}$	$\frac{0}{9,2}$	$\frac{0}{4,7}$	$\frac{9,1}{16,7}$	$\frac{0}{4,5}$	$\frac{0}{6,2}$	$\frac{6,1}{12,5}$	$\frac{1.6}{9,1}$

Class	11	12	13	14	15	16	17	18	19	20
	M	N	P	Q	R	S	T	V	W	Y
LNTs	$\frac{0}{2,8}$	$\frac{2,7}{7,0}$	$\frac{5,6}{11,3}$	$\frac{0}{4,2}$	$\frac{2,7}{8,5}$	$\frac{4,1}{9,1}$	$\frac{7\ 0}{12\ 7}$	$\frac{2,7}{7,2}$	$\frac{1,4}{4,3}$	$\frac{1,4}{5,6}$
SNTs	$\frac{0}{1,7}$	$\frac{4,8}{9.8}$	$\frac{3,2}{9.7}$	$\frac{3,3}{8,3}$	$\frac{4,8}{9,8}$	$\frac{3.3}{12,9}$	$\frac{8,1}{14,8}$	$\frac{1,6}{3,3}$	$\frac{1,6}{3.3}$	$\frac{1,6}{6,7}$
CTs	$\frac{0}{6.7}$	$\frac{3,3}{10.0}$	$\frac{5,0}{10,9}$	$\frac{0}{1,7}$	$\frac{1,6}{6,5}$	$\frac{1,7}{15,2}$	$\frac{3.3}{11,5}$	$\frac{0}{11,7}$	$\frac{0}{1,7}$	$\frac{1,6}{6.7}$
SCOs	$\frac{0}{0}$	$\frac{3,1}{7,8}$	$\frac{1,5}{9,5}$	$\frac{0}{4,6}$	$\frac{0}{7,7}$	$\frac{3\ 1}{9.5}$	$\frac{0}{7,8}$	$\frac{1.5}{9,4}$	$\frac{1\ 5}{4,6}$	$\frac{3.0}{10,9}$

*The numerator gives the minimum and the denominator the maximum values.

acid resources of each class are determined (Table 3). To identify the hypothetical proteins, use is made of the algorithms of pattern recognition theory [10] which had not hitherto been applied to protein structures. We are the first to have used them [17].

The procedure of pattern recognition theory is as follows. A priori information is given on a number of substances belonging to the corresponding classes — a teaching process. By means of a definite algorithm, the program reveals the relationship between the combination of characteristics and each class and deduces a law which is called the deciding rule.

After such teaching, the computer is capable of recognizing a new, previously unpresented, substance and assigning it to one of the classes known to it, i.e., of performing identification. In addition, it can perform the separation of groups of substances into clusters according to their closeness to classes, i.e., classification.

A necessary condition for reliable recognition is the successful choice of characteristics for describing the substances and the compactness of the teaching substances. The quality of recognition is evaluated by the method of sliding control, according to which out of n objects one is selected for examination, and teaching is carried out on the remaining n — 1 objects. This procedure is repeated n times, until all the objects have been screened, after which estimates are averaged.

As a result of the work of the program the most informative characteristics are revealed — in the case of the toxins that we are comparing, the percentage contents of the following amino acids: glycine, cysteine, tryptophan, and glutamine. The addition of other characteristics leads to a sharp rise in the percentage of errors.

The main result of the calculations is the production of scores for classification (they are given in Table 4 together with the corresponding gradations of characteristics, after which classification and identification are carried out with the aid of a computer.

In the identification of unknown proteins from the percentages of the amino acids mentioned, the intervals of the gradations are determined and the corresponding scores are summed over the columns. A substance is assigned to the class with the maximum sum.

As an example is given the toxin of N. mos. mossambica 1 (Nmm 1) with the amino acid sequence LECHNQQSEPPTTRCSGGETNCYKKRWRDHRCYRTERCGCCPTVKKGIELNCCTTDRCNN, which was not present in the teaching set [14].

The scores for the identification of Nmm 1 are given in Table 5. Thus, Nmm 1 belongs to the class of short neurotoxins. According to [14], this toxin is in fact a short neurotoxin, which shows the correctness of the classification scores obtained. A similar examination has been carried out on 15 toxins.

Let us show, using the family of cytotoxins (CTs) as an example, how the computer method described is used for screening in the process of prediction. As can be seen from Table 2,

TABLE 4. Scores for the Classification of Toxins

Amino acid	Gradation of the characteristic, %	Class of toxins			
		LNTs	SNTs	CTs	SCOs
Glycine	$\bar{G} < 5{,}4$	12	0	26	2
	$5{,}4 < \bar{G} < 8{,}8$	23	23	1	0
	$G > 8{,}8$	0	12	0	23
Cysteine	$\bar{G} \leqslant 12{,}7$	6	3	0	29
	$12{,}7 < \bar{C} \leqslant 13{,}3$	1	22	25	0
	$G\ 13{,}3$	29	20	0	5
Tryptophan	$\bar{W} \leqslant 1$	2	0	23	2
	$1 < \bar{W} \leqslant 1{,}5$	22	3	0	16
	$1{,}5 < \bar{W} \leqslant 1{,}7$	0	24	16	17
	$\bar{W} > 1{,}7$	25	8	0	23
Glutamine	$\bar{Q} \leqslant 3{,}1$	26	0	27	26
	$\bar{Q}\ 3{,}1$	15	30	0	14

TABLE 5. Identification of the Toxin Nmm 1

Amino acid	Characteristic, %	LNTs	SNTs	CTs	SCOs
Glycine	$\bar{G} = 8,05$	23	23	1	0
Cysteine	$\bar{C} = 14,5$	20	20	0	5
Tryptophan	$\bar{W} = 1,7$	0	24	16	17
Glutamine	$\bar{Q} = 3,2$	15	30	0	14

in the corresponding branched structure six bifurcations were obtained, which generate 2^6 = 64 primary structures with a threshold of disregard of 30%. The amino acid compositions were calculated for each of the hypothetical structures and these were checked for their inclusion in the initial reference parallelepiped. In our case, after this procedure all 64 structures that had not been included in the initial set remained. Then for each of them identification was carried out according to the scores of Table 4. They all proved to be cyclotoxins.

Thus, on the basis of the method of quantitative structure-activity relationships and pattern recognition theory, the initial set of 35 amino acid sequences of cytotoxins has permitted the "prediction" of 64 primary structures of presumed polypeptides with a cytotoxic function.

The possibility of computer analysis based on results concerning the amino acid compositions of proteins are also demonstrated by an example with the toxins from scorpion venoms.

We have previously [18] constructed an evolutionary tree for the family of scorpion toxins according to the mutational distances of their primary structures. In it the toxins proved to be grouped into clusters each of which originated from one ancestral protein. The compositions of these groups practically coincided with the clusters obtained by recognition from amino acid composition, from neuro- and cytofunctions, and also from an immunological classification [19]. Thus, correlation was detected between mutational distance, closeness in amino acid composition, and similarity of basic function.

At the present time, interest in the computer analysis of information included in primary structures (amino acid sequences) and genetic text is rising and, accordingly, various mathematical models and programs are being brought in and new prospects are being opened up in the development of molecular biology, bioorganic chemistry, and pharmacology.

The necessity for the rapid identification of certain proteins for which the amino acid sequences are unknown but the amino acid composition are readily determined is also increasing in clinical diagnostics [20].

The method proposed in the present paper is simple in use and, if classification scores have been obtained previously, it is extremely convenient and economical, since it is based on readily determinable magnitudes (amino acid compositions).

SUMMARY

1. Generalized structures of four classes of toxic peptides with clearly expressed invariant and variable fragments characteristic for each class have been obtained by computer methods.

2. Identification scores for these classes have been calculated by the methods of pattern recognition.

LITERATURE CITED

1. R. F. Doolittle, Science, 214, 149 (1981).
2. G. G. Kneale and M. J. Bishop, Comput. Appl. Biosci., 1, No. 1, 11 (1985).
3. V. A. Ratner and N. A. Kolchanov, The Role of Data Banks and Complexes of Research Programs in the Development of the Theory of Molecular-Genetic Control Systems [in Russian], Proceedings of the ITsiG SO AN SSSR [Institute of Cytology and Genetics, Siberian Division of the USSR Academy of Sciences], Novosibirsk (1983), p. 42.
4. A. A. Aleksandrov, Zh. Vses. Khim. Obshch. im. D. I. Mendeleeva, 29, No. 2, 64 (1984).

5. Yu. A. Pankov, Vestn. Akad. Med. Nauk. SSSR, 2, 9 (1983).
6. M. O. Dayhoff, Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, DC., Vol. 5 (1972) Suppl. 1 (1973); Suppl. 2 (1976); Suppl. 3 (1979).
7. W. M. Fitch and E. Margoliash, Science, 155, 279 (1967).
8. F. Dzh. Aiala, Zh. Obshch. Biol., 17, No. 4, 479 (1986).
9. Yu. P. Adler, E. V. Markova, and Yu. V. Granovskii, The Planning of Experiments in the Search for Optimum Conditions [in Russian], Nauka, Moscow (1976), p. 279.
10. V. N. Vapnik, Algorithms and Programs for the Restoration of Relationships [in Russian], Nauka, Moscow (1984), p. 816.
11. A. J. Stuper, W. E. Brugger, and P. C. Jurs, Computer Assisted Studies of Biological Activity Relations and Biological Function, Wiley, New York (1978).[Russian translation, Mir, Moscow (1982), p. 235].
12. V. E. Golender and A. B. Rozenblit, Computer Methods of Constructing Drugs [in Russian], Zinatne, Riga (1987),. p. 238.
13. D. J. Strydom, Snake Venoms (Volumes 52 of Handbook of Experimental Pharmacology), Springer, New York (1979), No. 4, p. 258.
14. M. J. Dufton and R. C. Hider, CRC Crit. Rev. Biochem., 14, No. 2, 113 (1983).
15. I. I. Parilis, G. L. Bussel', L. Ya. Yukel'son, and D. Kh. Khamidov, Dokl. Akad. Nauk UzSSR, No. 10, 47 (1984).
16. G. I. Chipens, Vestn. Akad. Med. Nauk SSSR, 2, 18 (1983).
17. I. I. Parilis, G. L. Bussel', L. Ya. Yukel'son, and D. Kh. Khamidov, Mol. Biol. (Moscow), 22, No. 6, 1697 (1988).
18. G. L. Bussel', I. I. Parilis, L. Ya. Yukel'son, and D. Kh. Khamidov, in: Abstracts of Lectures of the VIth All-Union Symposium on Molecular Mechanism of Genetic Processes [in Russian], Moscow (1987),, p. 65.
19. M. J. Dufton, A. F. Drake, and H. Rochat, Biochim. Biphys. Acta, 869, No. 1, 16 (1986).
20. W. C. Barker and M. O. Dayhoff, in: Proceedings of the 7th Annual Symposium of Computing Applied to Medical Care, Washington, Silver Springs, MD (1983), p. 584.
21. G. I. Chipens, L. K. Polevaya, N. I. Veretennikova, and A. Yu. Krikis, The Structure and Functions of Small Peptides [in Russian], Zinatne, Riga (1980), p. 328.

POTENTIOMETRIC TITRATION OF PEPTIDES AND THE STARTING MATERIALS AND INTERMEDIATES OF THEIR SYNTHESIS.

I. ACIDIMETRIC DETERMINATION OF N-ACETYLAMINO ACIDS

A. Ya. Veveris and B. A. Stintse

UDC 543.257.155.47.466

The potentiometric titration of N-acetylamino acids and their sodium salts in nitromethane—acetic anhydride (2:1) with a nitromethane solution of perchloric acid has been investigated. A procedure has been developed for the quantitative determination of N-acetylamino acids in the presence of acetic acid in aqueous solutions. A differentiation determination of N-acetylamino acids and their salts in the presence of sodium acetate has been carried out.

The development of technological methods of peptide synthesis requires an improvement in the methods of analytical control, including those that are intended for determining the quantitative composition of the starting materials, the auxiliary reagents, and the intermediates of the various technological processes. Our preceding investigations [1, 2] have shown that extremely reliable information on the quantitative content of substances can be obtained from the results of potentiometric acid-base titration. The use of modern technology for the performance of potentiometric titration permits the consumption of samples undergoing analysis to be restricted to a few milligrams with the retention of an adequately high accuracy and reliability of the results.

All-Union Scientific-Research Institute of Applied Biochemistry, Biolar Scientific-Industrial Association, Olaine. Translated from Khimiya Prirodnykh Soedinenii, Vol. 6, pp. 830-834, November-December, 1989. Original article submitted January 31, 1989.

0009-3130/89/2506-0704$12.50