БИООРГАНИЧЕСКАЯ ХИМИЯ
том 3 * № 7 * 1977

УДК 547.96 : 519.2

ИСПОЛЬЗОВАНИЕ ДАННЫХ О ПЕРВИЧНОЙ СТРУКТУРЕ ГОМОЛОГИЧНЫХ БЕЛКОВ ПРИ УСТАНОВЛЕНИИ СТРОЕНИЯ НОВОГО ПРЕДСТАВИТЕЛЯ ТОЙ ЖЕ ГРУППЫ *

Костецкий П. В.

Институт биоорганической химии им. М. М. Шемякина Академии наук СССР, Москва

Предложена программа для ЭВМ, позволяющая получать важную в практическом отношении информацию при установлении первичной структуры нового представителя семейства гомологичных белков. Для применения программы достаточно наличия ограниченного количества экспериментальных данных по аминокислотному составу пептидных фрагментов исследуемого белка. Использование настоящей программы позволяет делать заключение о порядке чередования имеющихся пептидов, а также о числе и характере аминокислотных замен в них. Для успешного применения предлагаемого метода необходимо, чтобы аминокислотные составы пептидов исследуемого белка и соответствующих участков гомологичного белка в сумме различались не более чем на 10—15%.

В настоящее время при установлении первичной структуры белков используются трудоемкие, но надежно отработанные экспериментальные методы. С помощью различных ферментов молекулу белка расщепляют на пептидные фрагменты, получая в конечном итоге блоки с перекрывающимися аминокислотными последовательностями [2]. Рядом авторов были предложены математические приемы для машинного перебора возникающих при этом вариантов с целью уменьшения количества необходимых экспериментальных данных [3—5].

В настоящее время уже известны аминокислотные последовательности больших семейств белков, выполняющих одинаковые функции в различных организмах (цитохромы, гемоглобины, иммуноглобулины, нейротоксины ядов змей и др.) [6]. Многие представители этих семейств обладают высокой степенью гомологии и иногда отличаются друг от друга заменой только одного или нескольких аминокислотных остатков. При установлении первичной структуры нового представителя какого-либо семейства важно максимально использовать информацию, содержащуюся в уже известных гомологичных структурах. В случае замены одного аминокислотного остатка обычно достаточно знать только аминокислотную последовательность соответствующего пептидного фрагмента [7]. Если различий несколько, то требуется анализ, включающий в себя предварительную расстановку пептидов в последовательность, гомологичную уже известной, с последующим определением аминокислотных замен.

В настоящей работе предложена программа для ЭВМ, позволяющая по аминокислотному составу и N-концевым аминокислотам пептидов, по-

* Отдельные результаты данной работы были сообщены ранее [1].

лученных в результате гидролиза белковой молекулы, выбрать среди гомологичных белков с известной структурой представителя с наибольшим подобием. В случае достаточной степени подобия с помощью выбранной гомологичной последовательности удается сделать вывод о порядке чередования имеющихся пептидов и о последовательности аминокислот в них. Приложимость метода исследована на примере установления строения представителей двух семейств белков: цитотоксинов ядов змей и леггемоглобинов из клубеньков ряда бобовых растений.

Для применения предлагаемого метода необходим ряд экспериментальных данных. С одной стороны, это аминокислотный состав и N-концевые аминокислоты пептидов, полученных в результате гидролиза белковой молекулы, строение которой требуется определить, с другой — известные последовательности гомологичных белков семейства, к которому принадлежит исследуемый белок.

Алгоритм программы, блок-схема которой представлена на рис. 1, заключается в следующем. Первоначально вводится одна из известных структур семейства, к которому принадлежит исследуемый белок. Вслед за этим сравниваются аминокислотные составы и N-концевые аминокислоты первого из имеющихся пептидов и равного ему по длине N-концевого участка гомологичного белка. В результате подсчитывается и хранится в памяти машины количество аминокислотных различий (R) в сравниваемых структурах и местоположение участка (L1—L2) гомологичной последовательности, после чего в ней выбирается следующий участок сравнения, начинающийся со второй аминокислоты. Если число различий не увеличивается, то вместо прежних результатов запоминаются новые. Этот процесс повторяется до тех пор, пока пептид не займет позицию напротив С-концевого участка гомологичного белка. После этого выводятся на печать структура пептида (A$) и отвечающий ему по аминокислотному составу участок последовательности гомологичного белка (B$). Различающиеся аминокислоты отмечаются звездочками. На рис. 2 показан результат расстановки триптического пептида цитотоксина СМ-8 из *Naja haja annulifera* на аминокислотной последовательности гомологичного белка СМ-II из *N. h. annulifera*.

Подобная операция осуществляется с каждым из имеющихся пептидов. Сумма аминокислотных различий по всем пептидам (ΣR), отнесенная к сумме длин пептидов, принимается за меру различия ($R1$) исследуемого и известного белков. Тот из гомологичных белков, для которого величина $R1$ достаточно мала (см. ниже), выбирается для предсказания структуры исследуемого белка.

Программа написана на языке BASIC для ЭВМ Hewlett-Packard 9830 A. Время работы программы зависит от количества и длины известных гомологичных белков и пептидов исследуемого белка. Для сравнения структур полного набора пептидов длиной от 2 до 20 аминокислотных остатков с гомологичной последовательностью из 150 аминокислотных остатков требуется $\sim$ 3 ч.

Правомочность предложенного алгоритма для определения порядка чередования пептидов, образующихся при расщеплении белковой молекулы, была опробована на модели, основанной на применении псевдослучайных чисел. Для получения последних использовался программный датчик, генерирующий числа, равномерно распределенные на интервале [0, 1]. Успешное применение псевдослучайных чисел в работе с аминокислотными последовательностями описано в ряде работ (см., например, [9—12]).

В предлагаемой модели какая-либо из 24 известных последовательностей цитотоксинов, содержащих 60 или 61 аминокислотный остаток (табл. 1), разбивалась произвольным образом на фрагменты длиной от 2 до 20 аминокислот. Затем из исходной молекулы получали искусственную гомологичную последовательность путем введения заданной доли аминокислотных замен. Необходимое количество аминокислотных замен обра-

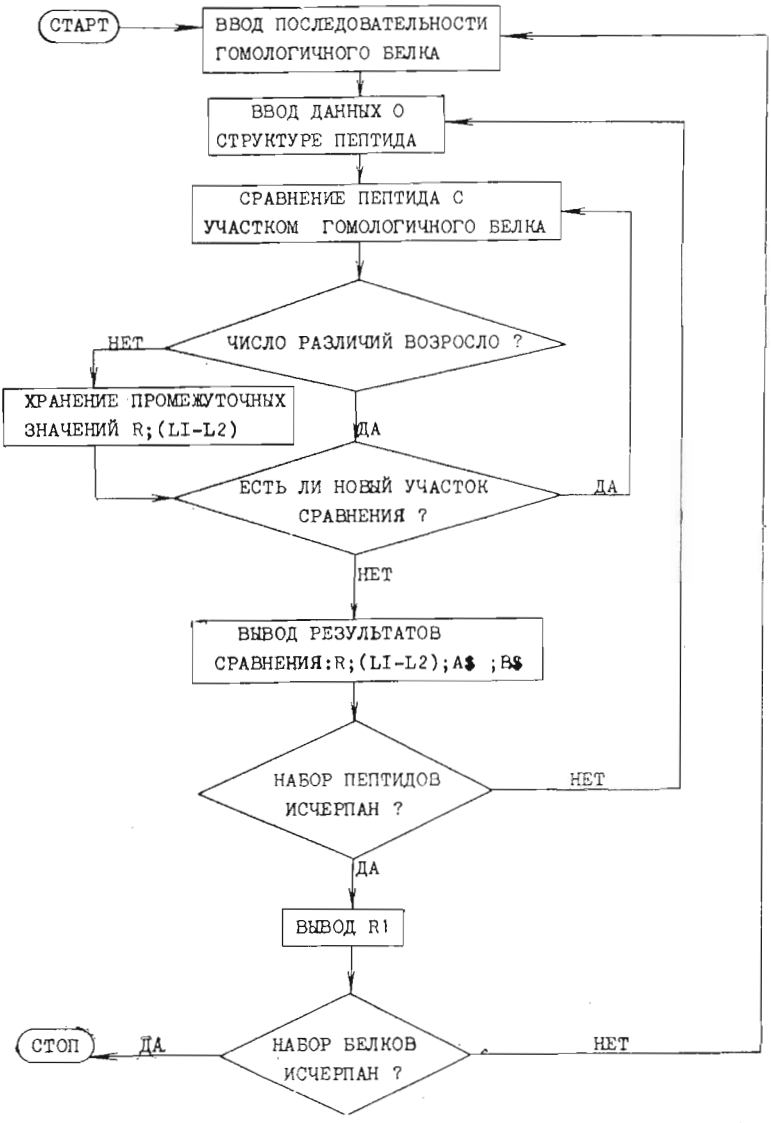

Рис. 1. Блок-схема программы сравнения пептидов исследуемого белка с известными последовательностями гомологичных белков. Подробности и обозначения приведены в тексте

зовывалось посредством накопления требуемого числа точечных мутаций [6], при которых новая и заменяемая аминокислоты обладали максимальным сходством, имея два одинаковых нуклеотидных основания в соответствующих триплетных кодонах [13].

С помощью полученной гомологичной последовательности по рассмотренному выше алгоритму производилось определение порядка чередования фрагментов исходной молекулы. Количественной мерой результата служила величина суммы длин сместившихся пептидов, отнесенная к сумме длин всех имеющихся пептидов $R2$. Для определенного уровня аминокислотных замен проводилось по 50 испытаний и подсчитывалось среднее значение доли сместившихся пептидов.

Согласно табл. 2, при уровне аминокислотных замен до 10% средняя доля сместившихся фрагментов невелика, что позволяет легко определить порядок чередования пептидов. Возрастание уровня замен до 25—28% приводит к резкому увеличению доли сместившихся фрагментов, что де-

```
    5   10   15   20   25   30   35   40   45   50   55   60
LKCYKLVPPFWKTCPEGKNLCYKMYMVSTLTVPVKRGCIDVCPKNSALVKYVCCNTDKCN  - цитотоксин СМ-II из
                                                                N.h.annulifera

(L1-L2)        R
  6-12         1


  *
L(V,P,P,F,W,K) - A$

  *
L I P P F W K  - B$
```

Рис. 2. Нахождение места с наибольшим подобием (L1 — L2) по аминокислотному составу для триптического пептида цитотоксина СМ-8 из *N. h. annulifera* (A$) на аминокислотной последовательности гомологичного белка — цитотоксина СМ-II из *N. h. annulifera*. Для названия аминокислот здесь и далее используются принятые однобуквенные обозначения [8]. B$ — участок последовательности 6—12 цитотоксина СМ-II из *N. h. annulifera*, *R* — число несовпадений в аминокислотных составах сравниваемых структур

лает практически невозможным установление порядка чередования пептидов. В табл. 2 приведены также частоты появления случаев с небольшим значением доли сместившихся пептидов ($R2 \leqslant 10\%$), которые соответствуют величине вероятности правильного определения порядка чередования пептидов при данном уровне аминокислотных замен. Видно, что при уровне замен до 10% эта вероятность достаточно высока, тогда как при уровне

Таблица 1

Результаты определения порядка чередования триптических пептидов цитотоксина II из *N. n. oxiana* с помощью известных гомологичных цитотоксинов *

№	Цитотоксин	ΣR	$R1$, %	$R2$, %
1	*N. naja*, кобрамин A	13	17	26
2	*N. naja*, кобрамин B	7	9	7
3	*N. naja*, F8	14	18	32
4	*N. n. atra*	6	8	7
5	*N. n. atra*, кардиотоксин	6	8	4
6	*N. n. oxiana*, I	15	20	22
7	*N. n. oxiana*, II	0	0	0
8	*N. melanoleuca*, $V^{II}1$	14	18	42
9	*N. melanoleuca*, $V^{II}1A$	16	21	42
10	*N. melanoleuca*, $V^{II}2$	25	33	50
11	*N. melanoleuca*, $V^{II}3$	22	29	50
12	*N. m. mossambica*, $V^{II}1$	14	18	39
13	*N. m. mossambica*, $V^{II}2$	15	20	24
14	*N. m. mossambica*, $V^{II}3$	15	20	24
15	*N. m. mossambica*, $V^{II}4$	18	24	47
16	*N. nigricollis*	13	17	39
17	*H. haemachatus*, DLF	15	20	30
18	*N. h. annulifera*, $V^{II}1$	15	20	32
19	*N. h. annulifera*, $V^{II}2$	13	17	28
20	*N. h. annulifera*, $V^{II}2A$	13	17	28
21	*N. h. annulifera*, СМ-8	13	17	12
22	*N. h. annulifera*, СМ-11	16	21	32
23	*N. h. annulifera*, СМ-13A	25	33	87
24	*N. n. atra*, аналог I	14	18	24

* Последовательности 7, 5 и 24 взяты из работ [14], [15] и [16] соответственно, остальные — из работы [7].

Таблица 2

Зависимость результатов модельных опытов по определению порядка чередования фрагментов белка с помощью известных гомологичных последовательностей от уровня аминокислотных замен $R1$ (%)

$R1$	Без учета инвариантных позиций		С учетом инвариантных позиций	
	$\overline{R2}$ *	f	$\overline{R2}$	f
1–2	0,8	95	0,3	100
8–10	11	62	11	60
25–28	51	3	46	6

* $\overline{R2}$ — средняя доля (%) сместившихся пептидов; f — частота появления случаев (%) с правильным порядком чередования пептидов ($R2 \leqslant 10\%$).

Таблица 3

Различия аминокислотных составов триптических пептидов цитотоксина из *N. n. oxiana* и соответствующих участков кардиотоксина из *N. n. atra*

Интервалы последовательности, в которых имеются аминокислотные замены	Изменяемые аминокислоты	
	в кардиотоксине из *N. n. atra*	в пептидах цитотоксина из *N. n. oxiana*
4	K	N
7–12	S	Y
25–35	A, H	T, K
52–60	R	K

более 20% она резко падает до 3%. Учет 33 инвариантных позиций в последовательностях цитотоксинов [7] не сказывается существенным образом на полученных результатах.

На рис. 3 приведен результат установления порядка чередования триптических пептидов цитотоксина II из *N. n. oxiana* с помощью гомологичной последовательности кардиотоксина из *Naja naja atra*. В результате работы программы все пептиды, кроме Т-2, располагаются непротиворечивым образом в порядке, совпадающем с известным [14]. Так, пептиды Т-1, Т-5, Т-6, Т-8, Т-10 занимают позиции 1—2, 13—18, 19—23, 36—44, 45—50 и не имеют различий в аминокислотных составах с соответствующими участками гомологичной последовательности. Пептиды Т-3 и Т-11 заняли позиции 5—12 и 51—60 и имеют по одной аминокислотной замене. Пептиды Т-4 и Т-9 оказались фрагментами пептидов Т-3 и Т-8 соответственно. В связи с этим отпадает необходимость дальнейшего исследования пептидов Т-4 и Т-9. Смещенным оказался только пептид Т-2, содержащий 3 аминокислотных остатка. В итоге доля смещенного участка последовательности в суммарной длине имеющихся фрагментов составляет 4%, что находится в согласии с невысоким значением уровня аминокислотных замен ($R1 = 8\%$) в сравниваемых структурах. Легко видеть также, что имеется достаточно оснований расположить пептид Т-2 в позициях 3—5 вместо позиций 42—44, где его присутствие противоречит строению пептидов Т-8 и Т-9.

В результате машинного сравнения структур триптических пептидов цитотоксина II из *N. n. oxiana* с помощью последовательности кардиотоксина из *N. n. atra* удается определить характер и число аминокислотных замен в соответствующих фрагментах (табл. 3). Следует, однако, иметь в виду, что часть аминокислотных замен может взаимно компенсировать друг друга. Поэтому для определения точного числа и позиций аминокислотных замен необходимо располагать полными данными о последовательности аминокислот в пептидах исследуемого белка.

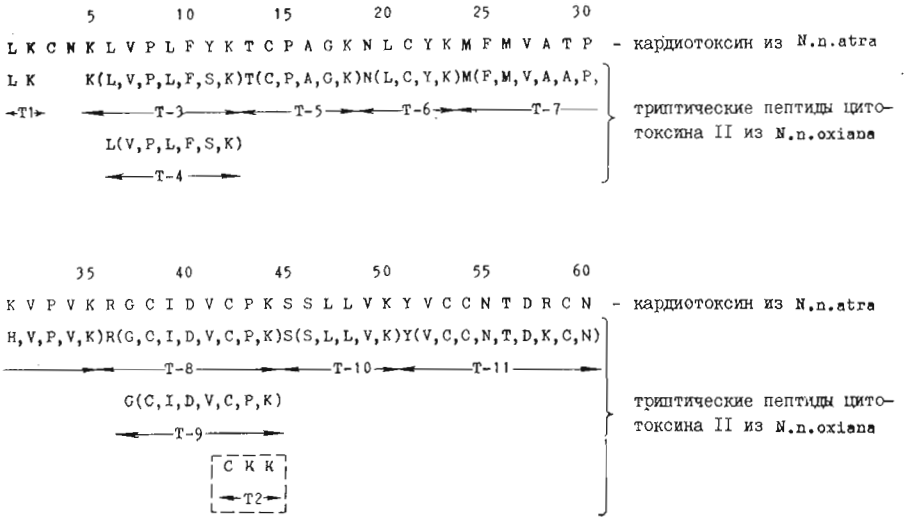

Рис. 3. Определение структуры цитотоксина II из *N. n. oxiana* на основе данных о его триптических пептидах и последовательности кардиотоксина из *N. n. atra*. Пунктирными линиями обозначен сместившийся пептид Т-2

Вместе с тем важно отметить, что для установления полной структуры цитотоксина II из *N. n. oxiana* с помощью предлагаемого метода дополнительное расщепление молекулы исследуемого белка каким-либо другим ферментом излишне.

Результаты применения структур других цитотоксинов в качестве гомологичных последовательностей при установлении строения цитотоксина II из *N. n. oxiana* приведены в табл. 1. Как и следовало ожидать, сборка пептидов в последовательностях с небольшим числом замен (до 10%) осуществляется почти полностью, тогда как в случаях с высоким уровнем замен ($> 20\%$) большое число пептидов оказывается смещенным.

Сходные результаты были получены и при определении структуры цитотоксина СМ-II на основании данных о его химотриптических пептидах длиной от 3 до 21 аминокислотного остатка [7]. Этот факт свидетельствует о независимости предлагаемого метода от способа расщепления исследуемого белка на фрагменты. В рассмотренных примерах пептиды собирались в правильном порядке ($R2 < 10\%$) в ряде случаев и при уровне аминокислотных замен $R1$ до 15—17%.

Метод, изложенный выше, был применен также при установлении структуры леггемоглобина II из клубеньков желтого люпина *, принадлежащего к семейству растительных кислородсвязывающих гемопротеидов, состоящих подобно миоглобину из одной полипептидной цепи длиной ~ 150 аминокислотных остатков. Исходными данными служили аминокислотные составы и N-концевые аминокислоты 15 триптических пептидов леггемоглобина I-I из люпина длиной от 2 до 21 аминокислотного остатка (рис. 4). В качестве гомологичных последовательностей были взяты 3 известные структуры леггемоглобинов и миоглобин кашалота. Ближайшим гомологом оказался леггемоглобин I люпина, при сравнении с которым только 19 из 152 аминокислот, содержащихся в имеющихся триптических пептидах, составляют расхождение с аминокислотным составом соответствующих участков гомологичной последовательности.

---

* Данные любезно предоставлены Ц. А. Егоровым (Институт биоорганической химии АН СССР).

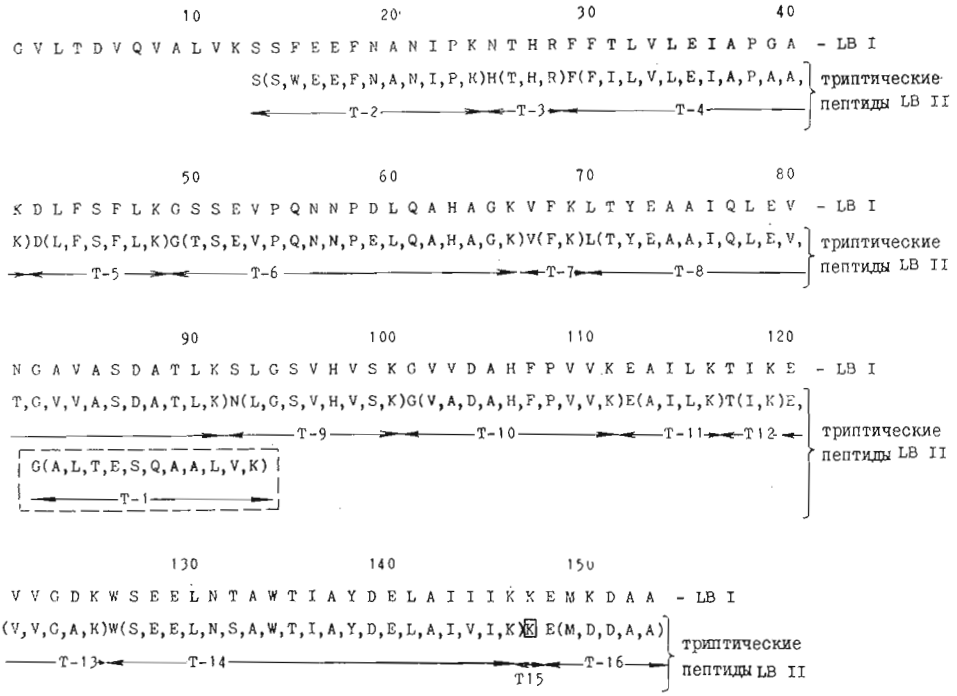

Рис. 4. Определение структуры леггемоглобина II из люпина (LBII) на основе данных о его триптических пептидах и последовательности леггемоглобина I из люпина (LBI). Пунктирными линиями выделен сместившийся пептид Т-1, сплошными линиями — пептид Т-15, данные о котором в ЭВМ не вводились

Низкое значение уровня аминокислотных замен ($R1 = 12\%$) дает основание считать, что доля сместившихся фрагментов леггемоглобина II при их сборке на последовательности леггемоглобина I окажется небольшой. Действительно, из 15 пептидов только один (Т-1) занял неправильную позицию, где его присутствие противоречит строению пептидов Т-8 и Т-9 (рис. 4). Легко видеть, что истинное местоположение пептида Т-1 находится на N-концевом участке молекулы. Оставшейся незакрытой позиции 147 соответствует пептид Т-15, состоящий из остатка лизина. В итоге сборки образуется непрерывная пептидная последовательность длиной в 153 аминокислотных остатка.

Отличие в аминокислотных составах триптических пептидов леггемоглобина I люпина и соответствующих участков структур леггемоглобина фасоли, леггемоглобина сои и миоглобина кашалота превышает 20% (табл. 4). Это делает невозможным использование данных последовательностей для определения структуры леггемоглобина II люпина, исходя из аминокислотных составов его триптических пептидов, что находится в полном согласии с рассмотренной выше моделью.

В заключение следует отметить, что применение гомологичных последовательностей белков при установлении строения исследуемого белка на основании аминокислотного состава его пептидных фрагментов позволяет получить практически важные результаты.

Во-первых, удается определить порядок чередования имеющихся пептидов, а также установить, какие из них являются избыточными. В случае достаточно полного набора пептидов можно ограничиться одним видом расщепления исследуемой молекулы на фрагменты.

Во-вторых, удается предсказать характер и местоположение аминокислотных замен в изучаемом белке по сравнению с известными гомологичными структурами.

Таблица 4

Результаты определения порядка чередования триптических пептидов леггемоглобина II из люпина с помощью известных последовательностей леггемоглобинов и миоглобина кашалота

обозначения в тексте

№	Белок	Лит-ра	ΣR	$R1$, %	$R2$, %
1	Леггемоглобин I люпина	[17]	19	12	8
2	Леггемоглобин фасоли	[18]	47	27	70
3	Леггемоглобин сои	[19]	45	26	73
4	Миоглобин кашалота	[20]	55	31	100

Для успешного применения предлагаемого метода необходимо, чтобы аминокислотные составы пептидов исследуемого белка и соответствующих участков гомологичного белка в сумме различались не более чем на 10—15%.

Автор выражает признательность Ц. А. Егорову и Е. В. Гришину за участие в постановке задачи и О. С. Шереметьеву за помощь в составлении программ.

ЛИТЕРАТУРА

1. Kostetsky P. V. (1976) Abstracts of USSR — FRG Symposium on Chemistry of Peptides and Proteins, p. 87, Dushanbe.
2. Hirs C. H. W., Moore S., Stein W. H. (1960) J. Biol. Chem., **235**, 633—647.
3. Bradley D. F., Merril C. R., Shapiro M. B. (1964) Biopolymers, **2**, 415—444.
4. Dayhoff M. O. (1964) J. Theor. Biol., 8, 97—112.
5. Mosimann J. E., Shapiro M. B., Merril C. R., Bradley D. F., Vinton J. E. (1966) Bull. Math. Biophys., **28**, 236—260.
6. Dayhoff M. O. (1972) Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, University Medical Center, Washington D. C.
7. Joubert F. J. (1976) Eur. J. Biochem., 64, 219—233.
8. (1968) Eur. J. Biochem., 5, 151—153.
9. Krzywicki A., Slonimski P. P. (1967) J. Theor. Biol., **17**, 136—158.
10. Поройков В. В., Еспипова Н. Г., Туманян В. Г. (1976) Биофизика, **21**, 397—400.
11. Wittman-Liebold B., Dzionara M. (1976) FEBS Lett., 65, 281—283.
12. Ohta T. (1976) J. Mol. Evol., 8, 1—12.
13. Fitch W. M. (1966) J. Mol. Biol., **16**, 9—16.
14. Grishin E. V., Sukhikh A. P., Adamovich T. B., Ovchinnikov Yu. A., Yukelson L. Ya. (1974) FEBS Lett., 48, 179—183.
15. Hayashi K., Takechi M., Kaneda N., Sasaki T. (1976) FEBS Lett., **66**, 210—214.
16. Hayashi K., Takechi M., Sasaki T., Lee C. Y. (1975) Biochem. and Biophys. Res. Communs., 58, 117—122.
17. Егоров Ц. А., Фейгина М. Ю., Казаков В. К., Шахпаронов М. И., Миталева С. И., Овчинников Ю. А. (1976) Биоорган. химия, **2**, 125—128.
18. Lehtovaara P., Ellfolk N. (1975) Eur. J. Biochem., 54, 577—584.
19. Ellfolk N., Sievers G. (1971) Acta chem. scand., 25, 3532—3534.
20. Edmundson A. B. (1965) Nature, **205**, 883—887.

Поступила в редакцию
20.XII.1976

USE OF THE DATA ON PRIMARY STRUCTURE OF HOMOLOGOUS PROTEINS FOR SEQUENCE DETERMINATION OF A NEW REPRESENTATIVE OF THE SAME GROUP

KOSTETSKY P. V.

M. M. Shemyakin Institute of Bioorganic Chemistry, Academy of Sciences of the USSR, Moscow

A computer program has been devised which provides valuable information for the primary structure determination of a new member of a homologous protein family. The program may be applied provided even limited experimental data are available, such as amino acid composition of any peptide fragments obtained from the protein under study. Basing on the program developed, a conclusion can be drawn on the alignment and amino acid sequence of the fragments. The results thus obtained facilitate the planning further experiments.