Bull. Inst. Chem. Res., Kyoto Univ., Vol. 66, No. 4, 1989

Repeat Sequences of Amino Acids Suggest the Origin of Protein

Yasuhiko SETO* and Minoru KANEHISA**

Received November 1, 1988

In order to test our oligoglycine hypothesis for the generation of primordial proteins, we examined amino acid sequence repeats in present proteins stored in the PRF/SEQDB database. We found high occurrences of glycine containing repeat patterns such as GX, GGX, GGGX, GGXY, GGGGX, GGGXY, where G is glycine and X and Y are other amino acids. This is consistent with our hypothesis that short oligoglycines reacted with aldehydes and alkenes to form various oligopeptides as basic building blocks of protein molecules.

KEY WORDS: Oligoglycine Hypothesis/ Repeat Sequence/ Molecular Evolution

I. INTRODUCTION

The earth was thought to be born 4.6 billion years ago. Since then it took 0.6 billion years for the creation of life. The first 0.6 billion years and the following 4 billion years seem to have principally different meanings for proteins. Before the appearance of life primordial forms of proteins may have been produced. There are several hypotheses concerning the production of amino acids and nucleotides. Ohno's idea is an interesting one which states that nucleic acids have evolved from oligonucleotides by duplication (1). Akabori proposed that polyglycine is the primordial protein and it reacted with aldehydes and unsaturated compounds to form polypeptides containing various amino acids (2). Akabori's hypothesis explains well how the amino acids of proteins have been produced and why the amino acids are alpha amino acids. However, it is unlikely that functional proteins containing tens to hundreds of amino acids were derived directly from polyglycines as Akabori originally proposed. It is hard to imagine that various reactions occurred at various locations on a long polyglycine chain. In fact, it is simpler to think that short oligoglycines reacted with aldehydes and alkenes as in Akabori's hypothesis to form various oligopeptides. Short oligopeptide chains may condense with each other to form long protein chains. The idea implies that the structural unit of a protein is an oligopeptide. With this hypothesis in mind we are trying to find a clue for the evolution of protein molecules. In this report we analyze amino acid sequence repeats in present functional proteins.

II. METHODS OF CALCULATION

For the analysis of amino acid sequences we used version 87.5 of the protein

* 瀬戸保彦 : Protein Research Foundation, Peptide Institute, Ina, Minoh, Osaka 562
** 金久　實 : Institute for Chemical Research, Kyoto University, Uji, Kyoto 611

sequence database of the Protein Research Foundation (PRF/SEQDB). The database SEQDB has been produced by the Protein Research Foundation since 1979, along with the literature database LITDB and its printed version "Peptide Information". The PRF/SEQDB database contains full and partial sequences of natural peptides and proteins with more than 50 amino acid residues. PRF/SEQDB version 87.5 contains the sequence data extracted from literature published before December 1986, totaling 2,351,308 residues in 10,021 entries.

An amino acid sequence is represented by 20 letters: A(Ala), C(Cys), D(Asp), E(Glu), F(Phe), G(Gly), H(His), I(Ile), K(Lys), L(Leu), M(Met), N(Asn), P(Pro), Q(Gln), R(Arg), S(Ser), T(Thr), V(Val), W(Trp), and Y(Tyr). We searched repetitions of all possible amino acid patterns with lengths 1 through 15. An amino acid sequence repeat is characterized by two indices n and m, where n is the length of an oligopeptide repeat unit and m is the number of repetitions. We searched both continuous and discontinuous repeats against the PRF/SEQDB detabase. Repeat units are perfectly aligned in continuous repeats, while repeat units may be separated by any number of other amino acids in discontinuous repeats. We extracted continuous repeats with the following threshold values for the minimum number of repetitions: 4 for n=1, 3 for n=2, and 2 for n=3–15. The threshold values for discontinuous repeats were: 6 for n=3, 4 for n=4, 3 for n=5, and 2 for n=6–15.

III. RESULTS

3.1. Amino acid composition

The average amino acid composition in all the entries of PRF/SEQDB was the following:

L 9.10%, A 7.55%, G 7.35%, S 7.20%, V 6.45%, E 6.06%, K 5.94%,
T 5.92%, P 5.18%, R 5.15%, I 5.15%, D 5.13%, N 4.31%, Q 4.17%,
F 3.96%, Y 3.30%, H 2.30%, M 2.24%, C 2.09%, W 1.38%

Thus, glycine was the third most abundant amino acid component. When the composition was examined individually, 1,009 entries out of 10,021 had glycine as the most abundant amino acid component, while 2522 and 1099 entries, respectively, had leucine and serine as most abundant.

3.2. Repeats of single amino acids

We first searched continuous repeats of single amino acids with four or more residues. The longest repeat for each amino acid was:

P 29, E 27, G 20, Q 19, N 15, A 14, S 14, D 13, H 14, R 11,
L 10, K 8, T 7, F 7, I 6, V 5, Y 5, C 4, M 4

Repeats of 4 or more residues were not found for tryptophan. Repeats of up to only 4 residues were found for cystein and methionine. Proteins with long repeats of single amino acids were as follows:

P 29:	Epstein-Barr virus protein
E 27:	high mobility group protein HMG
G 20:	Ca dependent protease S
Q 19:	glucocorticoid receptor

The total number of occurrences of repeats with four or more residues for each amino acid was:

L 420, P 279, R 259, A 250, S 250, G 190, E 177, Q 164, K 88, N 65,
T 63, D 52, H 41, V 36, I 16, F 13, C 13, Y 4, M 2, W 0

Thus, repeat sequences appeared frequently for eight amino acids: leucine, proline, arginine, serine, alanine, glycine, glutamic acid, and glutamine. The proteins which were mainly composed of repeating amino acids or rich in specific amino acids were:

S:	phosvitin
A:	antifreeze protein
H:	Plasmodium antigen
G:	petunia structural protein

3.3 Repeats of oligopeptide units

Repeating oligopeptide units of sizes 2–15 were identified from the PRF/SEQDB database with the threshold values given in the Methods section. The results are shown in Tables I–III. Table I summarizes the patterns found for dipeptide repeats.

Table I. Dipeptide repeat patterns found

2 nd / 1 st	G	S	P	E	D	A	V	K	R	L	Q	T	F	Y	N	H	I	C	M	W
G	+	+	+	+	+	+	+	+	+		+	+				+		+		
S	+	+	+	+	+		+		+	+	+	+	+			+				
P	+	+	+	+		+	+	+	+		+	+								
E	+	+	+	+	+	+	+	+	+						+					
D	+	+		+	+		+		+	+		+		+		+				
A	+		+	+		+		+		+	+		+		+					
V	+	+	+	+	+					+							+	+		
K	+		+	+		+		+	+					+	+					
R	+	+	+	+	+			+	+	+										
L		+			+	+	+		+	+			+				+			
Q	+	+	+			+					+			+						
T	+	+	+		+							+								
F		+				+				+			+	+						
Y					+			+			+		+							
N				+		+		+							+					
H	+	+			+											+				
I							+			+							+			
C	+						+													
M																				
W																				

Table II. List of tripeptide repeat patterns found

XYZ type								XXY type			
ACG	AML	CPK	DTH	FLG	GSK	KTS	PVT	GGA	EEA	RRA	YYG
R	Q	CQK	V	I	P	KVS	PYQ	C	D	F	I
ADG	R	L	DVL	R	R	T	QST	D	F	G	T
L	ANC	CRL	T	S	Y	LNR	V	E	L	K	CCG
AEG	L	CSN	DWS	FPG	GTP	LPQ	QYS	F	P	L	S
K	APE	DEL	DYE	R	S	R	RTV	I	Q	M	NNI
L	G	P	N	FSK	GVI	S	STV	K	R	S	Y
P	R	DFI	EFQ	M	L	V		L	S	T	FFS
Q	AQK	V	EGL	V	M	LQN		M	T	V	
R	L	DGI	EIK	FTL	N	P		N	V	PPA	
V	ARF	K	Q	FVL	P	T		P	SSA	G	
AFL	I	L	EKG	FVR	GWN	Y		S	E	K	
AGD	K	N	L	GDV	GYS	LRS		V	F	L	
E	L	P	ELF	GHP	HQP	V		Y	G	R	
L	V	R	G	GIL	HVL	LSM		AAD	L	S	
N	ASC	DIL	K	P	IKN	P		E	N	T	
P	L	P	Q	T	Q	T		F	P	V	
Q	Q	S	R	GKI	S	LTP		G	R	TTA	
R	R	V	S	L	V	S		I	T	L	
T	T	DKF	T	N	ILS	W		K	V	Q	
V	V	G	EML	P	T	LVP		L	QQA	S	
AHQ	ATG	L	ENL	GLP	Y	R		N	D	V	
AIE	K	Q	EPG	Q	INK	S		P	E	Y	
G	P	DLK	L	R	R	LYN		Q	F	DDE	
K	R	R	EQL	S	IPL	MFT		R	G	K	
L	S	S	N	T	IQL	MST		S	I	M	
Q	V	V	P	GML	IRS	MYN		T	L	T	
AKE	AVE	Y	ERG	P	ISK	R		V	M	Y	
R	G	DNS	L	V	L	NST		LLA	P	VVA	
S	K	DPE	S	GNS	R	V		D	KKA	G	
V	L	I	ESL	GPK	ITN	NTQ		E	E	L	
ALD	P	L	R	L	S	S		F	G	R	
E	Q	R	ETW	Q	IVN	V		G	L	HHA	
G	S	DQE	EVL	R	IYV	NVS		I	M	P	
H	AWQ	G	EYG	S	KLM	NYT		M	P	T	
K	AYI	L	FGP	V	R	PRS		P	R	IIG	
N	CDK	S	S	GRI	KNS	PSR		S	S	T	
Q	CGK	DRG	FHR	S	V	V		T	Y	V	
R	L	L	FIV	T	KPQ	PTQ		V			
S	P	DSE	FKH		S	R					
V	S	V			T	S					

Table III. List of oligopeptide repeat patterns found

4-PEPTIDE				5-PEPTIDE			6-PEPTIDE	
G3A	FILL	GAGQ	ANSL	G4E	GSGAA	AVSQD	G5L	TPAPAA
F	IGLL	GLGY	ARID	F	GGNPP	CAERQ	M	TPATAA
K	AGLL	AGSG	ASLP	M	AGAGG	DLSTP	E5D	AEAAVD
L	LLAA	SGLG	AVKN	S	CGCGG	FNSTW	G	AHHAAD
M	LLVV	AEAD	AVPF	A4P	LGSGG	RKYSI	A5P	AHHAAN
R	LLCV	ASAL	CAER	G3DT	LGVGG	SPRKG	A4QQ	DGVSKK
S	ALLV	IAQA	DVLK	SE	LGYGG	SPTKR	E4VQ	EPVTTQ
V	GGSS	KAKR	EVTA	SF	DSGYG		R4G2	GGYGGA
Y	GGVP	KAKS	FIRK	SS	GQRAA		G3MGM	GQGQQS
E3A	DTGG	CFIF	FLND	L3AS	EPTCC		GFG3S	GQQPGQ
D	SGGK	DPDH	FSLA	WV	DDEPV		AGAG3	GSSAFA
G	EEGQ	EQDQ	GQSR	A3G2	EEGQQ		GGLS3	KEPTPP
I	EERA	FNFI	IYLS	N3SP	EEKKD		P3VHL	KKDDEG
L	EEKK	GTFT	KGLA		FRLPP			LAQQAS
A3E	SSEE	IYIS	LDHV		PHGHH		LFNSTW	AEGAPP
G	EQKK	LGLV	LDIC		IRRPR			PGLPGS
K	ETKK	PKPA	LSVY		YLSII			GPAGPP
P	TPKK	NANP	LTER		AAKPK			QQPFPP
Q3E	QQIL	PQPH	NKIT		KKPAA			PRSPRE
P	QQPF	SIFI	QGLN		GYQNN			PGQWQQ
R	ELQQ	WRWG	QTLD		NTNSS			GGSVAS
S	AAIG	YIYL	RAEP		PASAA			TGSCVV
L3A	SEAA		SLVQ		PPIHK			PHQPLQ
F	ESST		SPAF		PLTRR			PSPTAS
P	RASS		SPRK		RRAWI			TGTAKV
R3G	IIGL		VCMY		IYLSS			
H	RRAK		VSLY		TCCPD			
S3I	LAFF		WGEN		WIRRS			
T	NNTS		WPLG		AEAKT			
C3G	TTNS		WPVG		AHHDG			
F3M					RPERP			
I3C								
N3M								
T3S								

The number of times a given amino acid occupied either or both of the positions in the dipeptide repeat unit was:

G 13, S 12, P 10, E 10, D 10, A 9, V 8, K 8, R 8, L 8,
Q 6, T 5, F 5, Y 4, N 4, H 4, I 3, C 2, M 0, W 0

Table II shows all the patterns in tripeptide repeats. The tripeptide patterns are classified into two types, those containing three different amino acids such as GCA, and those containing an internal repeat of one amino acid such as GGA. The usage of amino acids in the first type was determined, for example, by counting

Table III. (continued)

7-PEPTIDE		8-PEPTIDE		9-PEPTIDE		10-PEPTIDE	
R6C	AEGAAPP	AG7	GGCGCGCC	H7DA	AHHAHHVAD	GH9	SYGGSSG4
S6K	CGGCGCG	D7E	GGCGGCGC	KPQGP4	CCGGCGGCG	LQ9	MGMG3MG3
A4KRK	NTTDNNT	Q7L	GGAGGAGA	FG3YGGSS	DRADGQPAG	APH8	GGFGGRG3R
SP4EH	PQPQLPY	R7C	GGSGGRGR	G3SYGGSS	EQPAAGAGG	DAH8	KPQGP3QGG
SP4VH	PQPQQPF	R6CC	PQPQPQPE	G3SYG3S	DRAAGQPAG		CGCGCGGCGG
	PSTTIPA	R4SQSP	NANPNVDP	G3S3GGY	KKARKRPKC		GTGTGTGSGS
	PSSPSYS	Q4PPFS	EDNNKPGK	GGAGAG3A	NNVFQPSQQ		HHDDAHHDGA
	PTSPSYS	G3LGAGF	AKVTGTGT		GQQGQQPGQ		HKPPVYTPPV
	PTSPNYS	G3SGSGF	KPGKEDGN		SGDSEDKKE		PEKRSKSGSR
	PTSPKYS	G3WGQPH	KSDEAEAL		SGTSGTSAQ		TSCCQPTSIQ
	YGSGGSS	N3MNHNM	KSDEAEAR				AKVTGTGTGT
	YGSGGGS		REAAEQAK				
	LGYGGSS		SPRKSPKK				
	KRFMRFG		TPPTPSPS				
	KRYGGFM						
	NEGLKTE						
	SILEYIH						

the patterns of GXY, XGY, and XYG for glycine where X and Y represent two different amino acids.

L 99, G 80, A 77, S 69, P 57, V 54, K 49, R 49, D 49, E 47,
T 44, I 43, Q 41, N 32, F 25, Y 16, M 15, C 14, H 8, W 5,

The total was 873 out of possible 20×19×18=6840. The patterns shown in Table II are unique patterns obtained by considering the difference in the phase at which the repeat unit begins; for example, XYZ, YZX, and ZXY are considered identical repeat nuits. The total number of unique patterns belonging to the first type was 291.

The usage of amino acids in the second type was determined by counting the patterns of GGX for glycine, for example.

G 14, A 14, L 11, E 10, S 10, Q 9, K 9, R 9, P 8, T 6,
D 5, V 4, H 3, I 3, Y 3, C 2, N 2, F 1, M 0, W 0

The total was 123 out of possible 20×19=380. It is therefore noteworthy that the second type containing the repetition of the same amino acid within a repeat unit appears to be relatively more abundant.

Table III shows repeat patterns of unit sizes 4–10 classified by the internal repetition within each repeat unit. In 2 to 5-peptide repeat units the types of GX, G2X, G3X, G2X2, G4X, and G3X2 appeared to be more frequent, where in this notation X is any amino acid other than glycine and a numeral denotes the number of consecutive runs. Types of X2Y, X3Y, and X4Y are also found many times, where X and Y are amino acids other than glycine.

IV. DISCUSSION

If primordial repeating units were produced from oligoglycines, we may find repeat units with glycine as a basic component. This was our working hypothesis in the present analysis. In terms of possible mechanisms of producing different amino acids, it is more probable that an oligoglycine is modified to give the same amino acid at a few positions than to give all different amino acids as shown below:

GGGGG ⟶ GGGXX	more probable
GGGGG ⟶ GABCD	less probable

The abundance of repeat patterns of the types of GX, G2X, G3X, G2X2, G4X, and G3X2 are, at least, consistent with the hypothesis.

Even if the production probability of repeat units containing different amino acids are low, such units, once produced, may become important as functional units to interact with various ligands such as other peptides, nucleotides, metal ions, sugars, lipids, and so on. Examples of biologically active functional repeat units in present proteins, which are extracted from the PRF/SEQDB database, are shown below:

peptide precursor	size	repeat sequence
FMRF amide	309	KRFMRFGK (20 times)
enkephalin	239	KRYGGXM (7 times, X=F or L)
thyroliberin	255	KRQHPGXR (5 times, X=K or R)
mating factor alpha	165	WHWLQLKPGQPKR (4 times)

These functional units in proteins are often reported as common functional motifs or conserved sequences.

Proteins with repeat sequences are found to be mainly structural proteins, storage proteins, and viral proteins. These proteins may be considered primordial from the standpoint of function. In dipeptide repeats, frequently appeared amino acids are so called primitive amino acids, such as glycine, serine, and proline, which may originate older than the other amino acids.

Biological organisms have developed various processes and systems for the efficiency and simplicity of their life during 4 billion years. Most significant is the mechanism of constructing a protein molecule from amino acids following genetic information. It appears less likely that polypeptides are produced by combining amino acids one by one for each protein. Protein molecules may have evolved from shuffling and combining of peptide segments. Recently there have been suggestions that exons and modules are elementary units of proteins (3). Modules are defined from the analysis of three-dimensional protein structures and usually composed of about 20 amino acid residues, while the median of observed exon lengths is about 40 residues.

In addition to these structurally defined units, there are functionally important units in the amino acid sequences. Many proteins have peptide fragments exhibiting function of native proteins and sometimes activity which has no direct relation to mother proteins. Examples are shown below:

Peptide	Protein	Activity
YGGFM	beta lipotropin fragment	morphine like activity
WMDF	gastrin fragment	hormonal activity
RGDS	fibronectin fragment	cell adhesion activity
PTKKGKG	nucleoplasmin fragment	translocate to cell nucleus

A special region in a protein may have acquired one function during the long history of chemical evolution. Alternatively, a special oligopeptide with one function may have grown up to a protein molecule adding various fragments for efficiency. If the latter process is the principal one, we may be able to find such primordial functional unit peptides in the current proteins. We presume that they are composed of five to ten amino acid residues from the consideration of occurrence probability. We find present proteins to be usually multifunctional, but by analyzing repeating sequences we may be able to obtain insights into the character of prebiotic proteins.

ACKNOWLEDGMENT

We thank Dr. Tatuso Ooi for his continued support to the PRF/SEQDB database and Dr. Shiro Akabori for inspiring us to undertake this work.

REFERENCES

(1) S. Ohno, "Evolution by gene duplication", Springer-Verlag, 1979.
(2) S. Akabori, IUB Symp. I., The Origin of Life on the Earth, I, 189 (1959).
(3) M. Go, Nature 291, 90 (1981): *Proc. Natl. Acad. Sci. USA,* **80**, 1964 (1983).
(4) Y. Seto, *Mol. Design* **8**, 2 (1987).