ELSEVIER

International Journal of Biological Macromolecules
19 (1996) 131–138

INTERNATIONAL JOURNAL OF
Biological
Macromolecules
STRUCTURE, FUNCTION AND INTERACTIONS

Identification of several types of periodicities in the collagens and their simulation

Didier G. Arquès[a], Jean-Paul Fallot[b], Christian J. Michel*[b]

[a]*Equipe de Biologie Théorique, Université de Marne-la-Vallée, Institut Gaspard Monge, 2 rue de la butte verte, 93160 Noisy le Grand, France*
[b]*Equipe de Biologie Théorique, Université de Franche-Comté, Institut Universitaire de Technologie de Belfort-Montbéliard, BP 527, 90016 Belfort, France*

Received 20 September 1995; revised 26 March 1996; accepted 22 April 1996

Abstract

The collagens constitute an important population of proteins providing the structural support in vertebrate tissues. A collagen is mainly based on a series of tripeptides of the type GX_1X_2 (G = Glycine, X_1 and X_2 being any residues). The nine amino acids occurring with significant frequencies in the X_1 and X_2 residue sites and G form the reduced protein alphabet Q = {A,D,E,G,K,L,P,Q,R,S} (A = Alanine, D = Aspartic acid, E = Glutamic acid, K = Lysine, L = Leucine, P = Proline, Q = Glutamine, R = Arginine, S = Serine). Surprisingly, the method based on the autocorrelation function $w(X)_i w'$ analysing the probability that an amino acid w' in Q occurs any i residues X after an amino acid w in Q (called i-motif $w(X)_i w'$), identifies six types of modulo 3 periodicities in collagens: three basic types 0, 1 and 2 modulo 3 and three combined types 0,1, 0,2 and 1,2 modulo 3. Furthermore, the classification of these 100 i-motifs according to the types of periodicities shows several strong relations between four sub-sets of Q {G}, {A,D,P,S}, {E,L} and {K,Q,R}. Then, these relations allow the construction of a simple automaton for the generation of model collagen sequences. Indeed, this automaton can simulate the six types of periodicities and it retrieves the types of periodicities for almost all i-motifs. Finally, the autocorrelation function based on the sub-set {K,Q,R} identifies segments of 18 amino acids in collagens which may correspond to the exons (segments of genes of 54 nucleotides) coding for those collagens.

Keywords: Identification; Modelling; Autocorrelation function; Automaton; Collagens

1. Introduction

The collagens represent an important population of proteins constituting the major structural components of the extracellular matrix in vertebrate tissues, i.e. bones, tendons, skin, ligaments and blood vessels. The collagens are divided into two main sub-families depending on their spatial structures [1]: the fibrillar collagens and the non-fibrillar collagens which are more heterogeneous and excluded from this study. A (fibrillar) collagen has a structure called α-chain mainly based on a series of tripeptides of the type GX_1X_2 (G = Glycine, X_1 and X_2 being any residues).

In the X_1 and X_2 residue sites, nine amino acids occur with significant frequencies. These nine amino acids with G form the reduced protein alphabet at ten residues Q = {A,D,E,G,K,L,P,Q,R,S}. The distribution of these ten amino acids in collagens is studied in Section 2 with the autocorrelation function $w(X)_i w'$ according to a new definition analysing without bias the occurrence probability of two amino acids w and w' (w and w' in Q) separated by any i residues X (called i-motif $w(X)_i w'$). Indeed, this new definition, unlike the classical one, avoids the decrease of probabilities when the number i of residues X between the two amino acids w and w' increases. The side effect induced by the end of the sequence is corrected, $10^2 = 100$ i-motifs are analysed with the autocorrelation functions. Surprisingly, this approach identifies six types of modulo 3 periodicities in collagens: three basic types 0, 1 and 2 modulo 3 and three combined types 0,1, 0,2 and 1,2

* Corresponding author, Tel.: +33 84 587720; fax: +33 84 222905; e-mail: michel@iut-bm.univ-fcomte.fr

0141-8130/96/$15.00
PII S0141-8130(96)01115-4

modulo 3 representing a mixing of two basic types. Furthermore, the classification of the 100 *i*-motifs according to the types of periodicities identifies several relations associated with four sub-sets of Q {G}, {A,D,P,S}, {E,L} and {K,Q,R}.

These relations allow in Section 3 the construction of a simple automaton for the generation of model collagen sequences. Indeed, the 100 autocorrelation functions applied in these simulated collagens retrieve the types of periodicities for 81 among 100 *i*-motifs.

In Section 4, the periodicities observed in collagens are compared with those in genes. Then, the four sub-sets identified in collagens are connected with the physical and chemical properties of their residues. Earlier amino acid sequence analyses of collagens are compared with the present results. Finally, the autocorrelation function based on the sub-set {K,Q,R} which is specific to the X_2 residue site, identifies segments of 18 amino acids in collagens which may correspond, despite of the genetic code degeneracy, to the exons (segments of genes of 54 nucleotides) coding for those collagens.

2. Identification of several types of periodicities in the collagens

2.1. Method

2.1.1. Definition of the autocorrelation function

This method extends the previous autocorrelation function definition on the gene alphabet [2] to the protein alphabet P at 20 letters called amino acids.

Let C be a collagen population with m(C) collagens. A collagen in C is a word of the type $c = (GX_1X_2)^n$ ($G,X_1,X_2 \in P$, $\in$ being the symbol of membership of a set) with a length $l(c) = 3n$ ($n > 66$). Let the *i*-motif $m_i = w(X)_i w'$ be two letters w and w' ($w,w' \in P$) separated by any i letters X ($i \in [0,29]$). For each collagen c in C, the counter $o_i(c)$ counts the occurrences of m_i in c. In order to count the m_i occurrences in the same conditions for all $i \in [0,29]$, only the first $l(c)-30$ (= $l(c)-(29+2)+1$) letters in c are examined (correction of the side effect induced by the end of the word which would have led to a decrease of probabilities). The occurrence probability $p_i(c)$ of m_i for c is then equal to the ratio of the counter by the total number of current letters read, i.e. $p_i(c) = o_i(c)/(l(c)-30)$. Finally, the occurrence probability $A_{w,w'}(i,C)$ of the *i*-motif m_i for the collagen population C, is equal to

$$A_{w,w'}(i,C) = \frac{1}{m(C)} \sum_{c \in C} p_i(c)$$

For the collagen population C, the function $i \rightarrow A_{w,w'}(i,C)$ giving the occurrence probability that w' occurs any i letters after w, is called autocorrelation function $w(X)_i w'$ (associated with the *i*-motif $w(X)_i w'$) and is represented by a curve as follows: the abscissa shows the number i of letters X between the two letters w and w' by varying i between 0 and 29; the ordinate gives the occurrence probability of $w(X)_i w'$ in the collagen population C.

2.1.2. Data acquisition

The (fibrillar) collagens are extracted from the protein data base SWISS-PROT (release 29 from June 1994). 15 collagens $(GX_1X_2)^n$ (from various fibrillar types: α1(I), α1(II), α1(III), α1(V), α2(I), α2(V) and from different taxonomies: human, mouse, rat, bovin, chick) greater than 200 amino acids ($n > 66$ residues) representing 11 016 amino acids have been obtained to constitute the collagen population C. In order to get significant results with the autocorrelation functions, the frequencies of amino acids occurring in the X_1 and X_2 residue sites are computed. Table 1 shows that several amino acids have a frequency close to 0, W is even absent. Therefore, the statistical analysis will be performed on the reduced protein alphabet at ten residues Q = {A,D,E,G,K,L,P,Q,R,S} (the nine amino acids with a frequency $\geq$4% in the X_1 and X_2 residue sites and G).

2.2. Results

The autocorrelation functions analysing the $10^2 = 100$ *i*-motifs on the alphabet Q identify six types of modulo 3 periodicities in collagens. The three basic types are:

— 0 modulo 3: $A_{w,w'}(i,C) > \{A_{w,w'}(i+1,C), A_{w,w'}(i+2,C)\}$ with $i \equiv 0[3]$ and $i \in [0,26]$ (maximum of the function for $i = 0, 3, 6$, etc), e.g. the autocorrelation function $G(X)_i E$ (Fig. 1a);

— 1 modulo 3: $A_{w,w'}(i,C) > \{A_{w,w'}(i+1,C), A_{w,w'}(i+2,C)\}$ with $i \equiv 1[3]$ and $i \in [0,26]$ (maximum of the function for $i = 1, 4, 7$, etc), e.g. the autocorrelation function $G(X)_i R$ (Fig. 1b);

Table 1
Amino acid occurrence frequencies (%) in the X_1 and X_2 residue sites of GX_1X_2 of collagens.
The nine amino acids with a frequency $\geq$4% in the X_1 and X_2 residue sites and G form the reduced protein alphabet at ten residues Q = {A,D,E,G,K,L,P,Q,R,S}.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
13.1	0.1	4.2	7.1	1.6	1.2	1.0	1.9	5.4	4.0	1.1	2.7	33.3	4.2	7.4	5.4	2.9	3.2	0.0	0.2

— 2 modulo 3: $A_{w,w'}(i,C) > \{A_{w,w'}(i+1,C), A_{w,w'}(i+2,C)\}$ with $i \equiv 2[3]$ and $i \in [0,26]$ (maximum of the function for $i = 2, 5, 8$, etc), e.g. the autocorrelation function $G(X)_iG$ (Fig. 1c).

The three combined types represent a mixing of two basic types, i.e. when two periodicities have high close probabilities. A probability level ($0.75 \times$ (mean probability of the highest periodicity – mean probability of the lowest periodicity)) allows to classify the two higher periodicities:

— 0,1 modulo 3: $A_{w,w'}(i,C) \approx A_{w,w'}(i+1,C) > A_{w,w'}(i+2,C)$ with $i \equiv 0[3]$ and $i \in [0,26]$ (maximum of the function for $i = 0, 1, 3, 4, 6, 7$, etc), e.g. the autocorrelation function $P(X)_iG$ (Fig. 1d);

— 0,2 modulo 3: $A_{w,w'}(i,C) \approx A_{w,w'}(i+1,C) > A_{w,w'}(i+2,C)$ with $i \equiv 2[3]$ and $i \in [0,26]$ (maximum of the function for $i = 0, 2, 3, 5, 6, 8$, etc), e.g. the autocorrelation function $E(X)_iP$ (Fig. 1e);

— 1,2 modulo 3: $A_{w,w'}(i,C) \approx A_{w,w'}(i+1,C) > A_{w,w'}(i+2,C)$ with $i \equiv 1[3]$ and $i \in [0,26]$ (maximum of the function for $i = 1, 2, 4, 5, 7, 8$, etc), e.g. the autocorrelation function $P(X)_iE$ (Fig. 1f).

Table 2 gives the classification of the 100 i-motifs according to the types of periodicities. It shows seven strong relations between the 100 i-motifs where the relation (1–7) is identified by a number in parenthesis:

(1) An i-motif $w(X)_iw'$ has a periodicity 2 modulo 3 with $w = w' \in \{G\}$, $w,w' \in \{A,D,P,S\}$, $w,w' \in \{E,L\}$ or $w,w' \in \{K,Q,R\}$ (except for $D(X)_iS$). In particular, the i-motif $w(X)_iw$ has a periodicity 2 modulo 3.

(2) If an i-motif $G(X)_iw$ has a periodicity 0 modulo 3 then the i-motif $w(X)_iG$ has a periodicity 1 modulo 3 with $w \in \{E,L\}$.

(3) If an i-motif $G(X)_iw$ has a periodicity 1 modulo 3 then the i-motif $w(X)_iG$ has a periodicity 0 modulo 3 with $w \in \{K,Q,R\}$.

(4) If an i-motif $G(X)_iw$ has a periodicity 0,1 modulo 3 then the i-motif $w(X)_iG$ has a periodicity 0,1 modulo 3 with $w \in \{A,D,P,S\}$. It is a combination of the two previous cases.

(5) If an i-motif $w(X)_iw'$ has a periodicity 1,2 modulo 3 then the i-motif $w'(X)_iw$ has a periodicity 0,2 modulo 3 with $w \in \{A,P\}$ and $w' \in \{E,L\}$ (except for $L(X)_iA$).

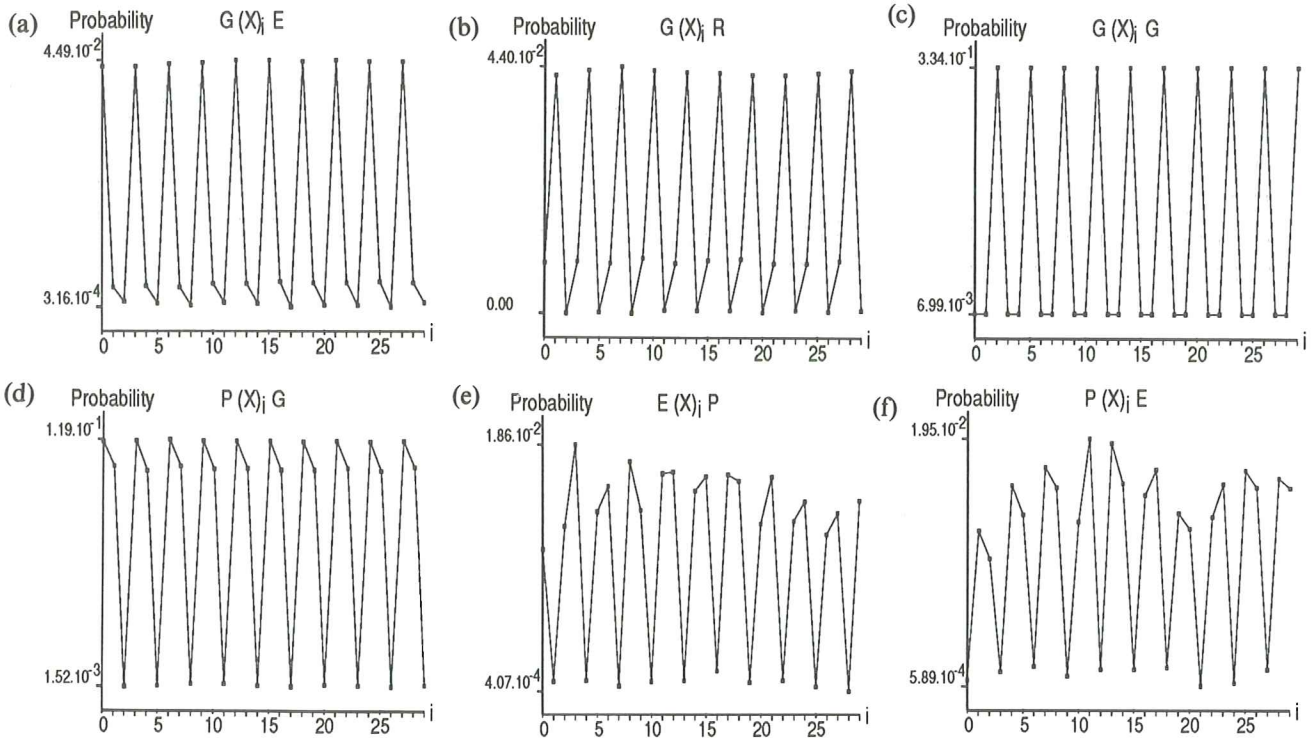

Fig. 1. (a) Periodicity 0 modulo 3 in collagens. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between G and E. The vertical axis represents the autocorrelation function $A_{G,E}(i,C)$ analysing the occurrence probability of $G(X)_iE$ in collagens. (b) Periodicity 1 modulo 3 in collagens. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between G and R. The vertical axis represents the autocorrelation function $A_{G,R}(i,C)$ analysing the occurrence probability of $G(X)_iR$ in collagens. (c) Periodicity 2 modulo 3 in collagens. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between G and G. The vertical axis represents the autocorrelation function $A_{G,G}(i,C)$ analysing the occurrence probability of $G(X)_iG$ in collagens. (d) Periodicity 0,1 modulo 3 in collagens. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between P and G. The vertical axis represents the autocorrelation function $A_{P,G}(i,C)$ analysing the occurrence probability of $P(X)_iG$ in collagens. (e) Periodicity 0,2 modulo 3 in collagens. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between E and P. The vertical axis represents the autocorrelation function $A_{E,P}(i,C)$ analysing the occurrence probability of $E(X)_iP$ in collagens. (f) Periodicity 1,2 modulo 3 in collagens. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between P and E. The vertical axis represents the autocorrelation function $A_{P,E}(i,C)$ analysing the occurrence probability of $P(X)_iE$ in collagens.

Table 2
Classification of the 100 i-motifs according to the six types of periodicities.
The number in parenthesis gives the relation of the i-motif $w(X)_i w'$ (w and w' in $Q = \{A,D,E,G,K,L,P,Q,R,S\}$) defined in Section 2.2.

Periodicity 0 modulo 3						
$E(X)_iK$ (7)	$E(X)_iQ$ (7)	$E(X)_iR$ (7)				
$G(X)_iE$ (2)	$G(X)_iL$ (2)					
$K(X)_iG$ (3)						
$L(X)_iK$ (7)	$L(X)_iQ$ (7)	$L(X)_iR$ (7)				
$Q(X)_iG$ (3)						
$R(X)_iG$ (3)						
Periodicity 1 modulo 3						
$E(X)_iG$ (2)						
$G(X)_iK$ (3)	$G(X)_iQ$ (3)	$G(X)_iR$ (3)				
$K(X)_iE$ (7)	$K(X)_iL$ (7)					
$L(X)_iG$ (2)						
$Q(X)_iE$ (7)	$Q(X)_iL$ (7)					
$R(X)_iE$ (7)	$R(X)_iL$ (7)					
Periodicity 2 modulo 3						
$A(X)_iA$ (1)	$A(X)_iD$ (1)	$A(X)_iK$ (6)	$A(X)_iP$ (1)	$A(X)_iQ$ (6)	$A(X)_iS$ (1)	
$D(X)_iA$ (1)	$D(X)_iD$ (1)	$D(X)_iK$ (6)	$D(X)_iL$	$D(X)_iP$ (1)		
$E(X)_iE$ (1)	$E(X)_iL$ (1)	$E(X)_iS$				
$G(X)_iG$ (1)						
$K(X)_iA$ (6)	$K(X)_iD$ (6)	$K(X)_iK$ (1)	$K(X)_iP$ (6)	$K(X)_iQ$ (1)	$K(X)_iR$ (1)	
$L(X)_iA$	$L(X)_iE$ (1)	$L(X)_iL$ (1)	$L(X)_iS$			
$P(X)_iA$ (1)	$P(X)_iD$ (1)	$P(X)_iK$ (6)	$P(X)_iP$ (1)	$P(X)_iQ$ (6)	$P(X)_iR$ (6)	$P(X)_iS$ (1)
$Q(X)_iA$ (6)	$Q(X)_iK$ (1)	$Q(X)_iP$ (6)	$Q(X)_iQ$ (1)	$Q(X)_iR$ (1)	$Q(X)_iS$ (6)	
$R(X)_iD$ (6)	$R(X)_iK$ (1)	$R(X)_iP$ (6)	$R(X)_iQ$ (1)	$R(X)_iR$ (1)		
$S(X)_iA$ (1)	$S(X)_iD$ (1)	$S(X)_iK$ (6)	$S(X)_iP$ (1)	$S(X)_iQ$ (6)	$S(X)_iS$ (1)	
Periodicity 0,1 modulo 3						
$A(X)_iG$ (4)						
$D(X)_iG$ (4)						
$G(X)_iA$ (4)	$G(X)_iD$ (4)	$G(X)_iP$ (4)	$G(X)_iS$ (4)			
$P(X)_iG$ (4)						
$S(X)_iG$ (4)						
Periodicity 0,2 modulo 3						
$A(X)_iR$						
$D(X)_iQ$	$D(X)_iR$					
$E(X)_iA$ (5)	$E(X)_iD$	$E(X)_iP$ (5)				
$L(X)_iD$	$L(X)_iP$ (5)					
$S(X)_iR$						
Periodicity 1,2 modulo 3						
$A(X)_iE$ (5)	$A(X)_iL$ (5)					
$D(X)_iE$	$D(X)_iS$					
$K(X)_iS$						
$P(X)_iE$ (5)	$P(X)_iL$ (5)					
$Q(X)_iD$						
$R(X)_iA$	$R(X)_iS$					
$S(X)_iE$	$S(X)_iL$					

(6) The i-motifs $w(X)_i w'$ and $w'(X)_i w$ have a periodicity 2 modulo 3 with $w \in \{A,D,P,S\}$ and $w' \in \{K,Q,R\}$ (except for $A(X)_iR$, $D(X)_iQ$, $D(X)_iR$, $S(X)_iR$, $K(X)_iS$, $Q(X)_iD$, $R(X)_iA$ and $R(X)_iS$). The relation (6) is the less significant one with eight among 24 exceptions.

(7) If an i-motif $w(X)_i w'$ has a periodicity 0 modulo 3 then the i-motif $w'(X)_i w$ has a periodicity 1 modulo 3 with $w \in \{E,L\}$ and $w' \in \{K,Q,R\}$.

These seven relations identify four sub-sets of Q {G}, {A,D,P,S}, {E,L} and {K,Q,R}. They also allow to deduce the types of periodicities between these four sub-sets (Table 3).

3. Simulation of periodicities with an automaton

3.1. Method

3.1.1. Construction of a stochastic automaton

A simple automaton can be constructed from the relations between the i-motifs (Section 2.2., Tables 2 and 3). The automaton considered here is stochastic: the choice between the edges issued from a same state is equiprobable. It constructs words having the properties of the automaton and therefore the relations between the i-motifs. The words are constructed by a random path (a series of edges) in the automaton so that the edge crossing concatenates the word associated with the edge to the building word. All edges issued from a same state are associated with a letter $w \in Q$. The automaton is represented as follows: the label of each state is the letter w common to all edges issued from the state and the edge crossing concatenates the letter w to the building word.

The initial state of the automaton A (Fig. 2) is the letter G corresponding to the beginning letter of the word $(GX_1X_2)^n$. The relation (4) demonstrates that the letters A, D, P, S occur 2 times in A, in the X_1 and X_2 letter sites. The relations (2), (3) and (7) explain that the letters E, L, K, Q and R only occur one time in A, E and L in the X_1 letter site, and K, Q and R in the X_2 letter site.

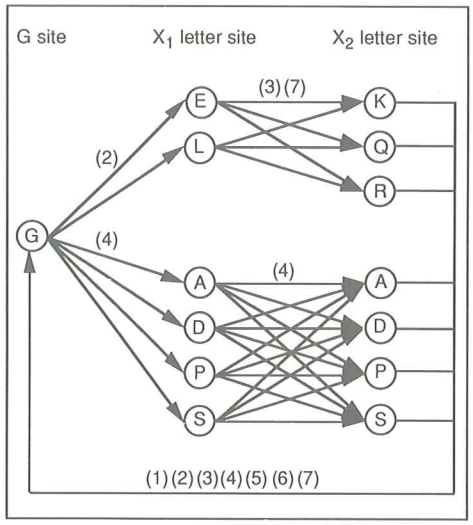

Fig. 2. Automaton A constructed from the relations between the i-motifs (numbers in parenthesis defined in Section 2.2., see also Tables 2 and 3).

The construction of the automaton A is then based on an elementary closed path which constructs the two words $Gww'G$ with $w,w' \in \{A,D,P,S\}$ and with $w \in \{E,L\}$ and $w' \in \{K,Q,R\}$. n closed paths construct the word $(Gww'G)^nG$ containing also the other identified relations (1), (5) and (6).

3.1.2. Simulated collagens constructed with the stochastic automaton

In order to verify that the automaton A can simulate the six types of periodicities and that the types of periodicities are correctly associated with the i-motifs, a population S of 200 simulated collagens of 1000 residue length is generated with the automaton A. Note that the computations obtained with such a sample of 200 000 amino acids are precise and stable (i.e. there is no random fluctuations in the probability calculus of i-motifs: a sample having for example 20 simulated collagens of 1000 residue length leads to similar results). The 100 autocorrelation functions are computed in this simulated population S as defined in Section 2.1.1.

Table 3
Types of periodicities between the four sub-sets of Q {G}, {A,D,P,S}, {E,L} and {K,Q,R} deduced from the relations between the i-motifs defined in Section 2.2.
The number in parenthesis gives the relation of the i-motif $w(X)_i w'$, the sub-sets in rows (resp. columns) representing the residue w (resp. w').

	{G}	{A,D,P,S}	{E,L}	{K,Q,R}
{G}	2 modulo 3 (1)	0,1 modulo 3 (4)	0 modulo 3 (2)	1 modulo 3 (3)
{A,D,P,S}	0,1 modulo 3 (4)	2 modulo 3 (1)	1,2 modulo 3 (5) (in general)	2 modulo 3 (6) (in general)
{E,L}	1 modulo 3 (2)	0,2 modulo 3 (5) (in general)	2 modulo 3 (1)	0 modulo 3 (7)
{K,Q,R}	0 modulo 3 (3)	2 modulo 3 (6) (in general)	1 modulo 3 (7)	2 modulo 3 (1)

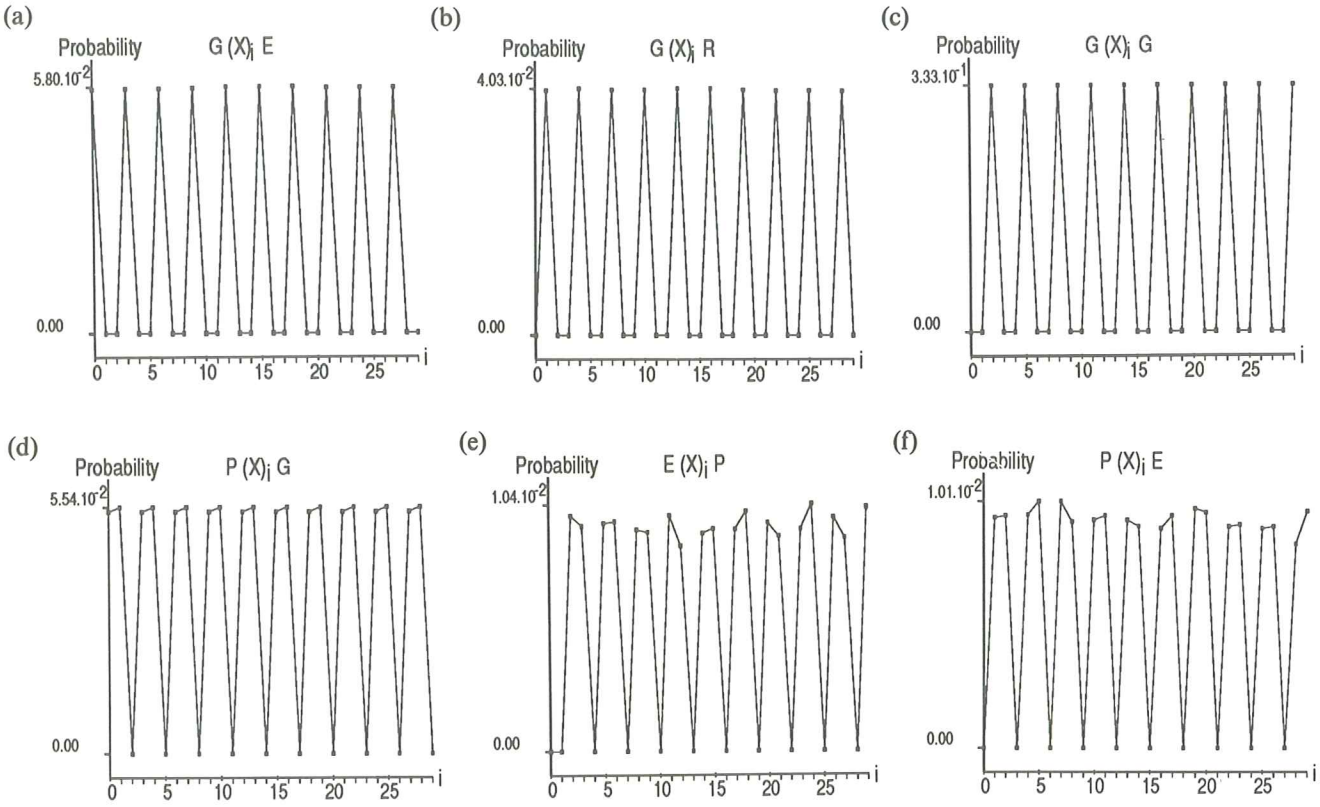

Fig. 3. (a) Simulation of the periodicity 0 modulo 3 identified in collagens (Fig. 1a) with the automaton A. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between G and E. The vertical axis represents the autocorrelation function $A_{G,E}(i,S)$ analysing the occurrence probability of $G(X)_iE$ in the simulated collagens. (b) Simulation of the periodicity 1 modulo 3 identified in collagens (Fig. 1b) with the automaton A. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between G and R. The vertical axis represents the autocorrelation function $A_{G,R}(i,S)$ analysing the occurrence probability of $G(X)_iR$ in the simulated collagens. (c) Simulation of the periodicity 2 modulo 3 identified in collagens (Fig. 1c) with the automaton A. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between G and G. The vertical axis represents the autocorrelation function $A_{G,G}(i,S)$ analysing the occurrence probability of $G(X)_iG$ in the simulated collagens. (d) Simulation of the periodicity 0,1 modulo 3 identified in collagens (Fig. 1d) with the automaton A. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between P and G. The vertical axis represents the autocorrelation function $A_{P,G}(i,S)$ analysing the occurrence probability of $P(X)_iG$ in the simulated collagens. (e) Simulation of the periodicity 0,2 modulo 3 identified in collagens (Fig. 1e) with the automaton A. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between E and P. The vertical axis represents the autocorrelation function $A_{E,P}(i,S)$ analysing the occurrence probability of $E(X)_iP$ in the simulated collagens. (f) Simulation of the periodicity 1,2 modulo 3 identified in collagens (Fig. 1f) with the automaton A. The horizontal axis represents the number i ($i \in [0,29]$) of any residues X between P and E. The vertical axis represents the autocorrelation function $A_{P,E}(i,S)$ analysing the occurrence probability of $P(X)_iE$ in the simulated collagens.

3.2. Results

The Fig. 3a–f show the autocorrelation functions $G(X)_iE$, $G(X)_iR$, $G(X)_iG$, $P(X)_iG$, $E(X)_iP$ and $P(X)_iE$ respectively in the simulated population S generated with the automaton A. For each of these six i-motifs, the autocorrelation function in these simulated collagens is strongly correlated with the autocorrelation function in collagens, for the type as well as for the probability level of periodicity (Fig. 1a–f and 3a–f). Furthermore, this correlation is verified with 81 among 100 i-motifs (data not shown). The 19 remaining i-motifs are moved from a basic type of periodicity to a combined one or reciprocally, but never with a contradiction, i.e. an i-motif with a periodicity 0 (resp. 1; 2; 0,1; 0,2; 1,2) modulo 3 in collagens is never observed with a periodicity 1,2 (resp. 0,2; 0,1; 2; 1; 0) modulo 3 in the simulated collagens.

4. Discussion

The autocorrelation function method applied to collagens identifies three basic types of periodicities 0, 1, 2 modulo 3, and three combined types of periodicities. The three basic types of periodicities are also observed in genes coding for proteins. On the two letter gene alphabet {R,Y} (R = purine = Adenine or Guanine, Y = pyrimidine = Cytosine or Thymine), mainly two types of periodicities are observed: 0 modulo 3 in protein coding genes (eukaryotes, prokaryotes and viruses)

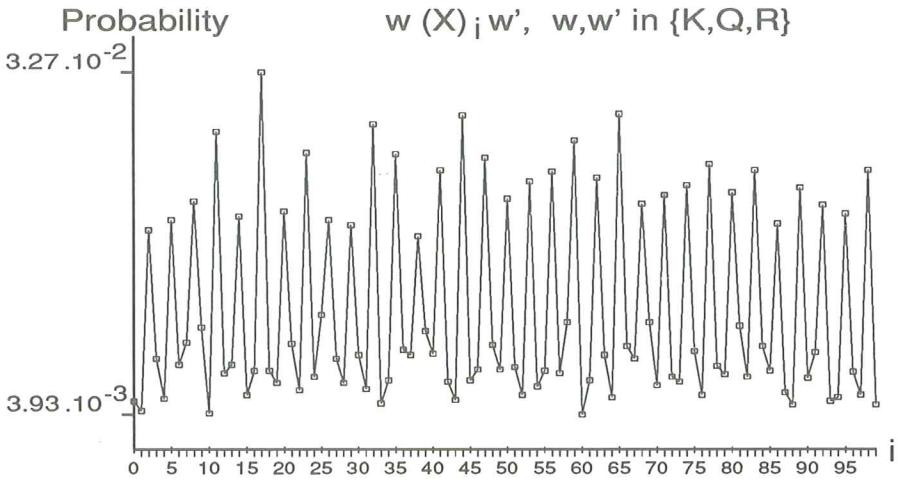

Fig. 4. Maximum of the autocorrelation function for $i = 17$ in collagens which may correspond to the length (54 nucleotides) of exons coding for those collagens. The horizontal axis represents the number i ($i \in [0,99]$) of any residues X between the amino acids w and w' with w and w' in {K,Q,R}. The vertical axis represents the autocorrelation function $A_{w,w'}(i,C)$ analysing the occurrence probability of $w(X)_i w'$ with w and w' in {K,Q,R} in collagens.

and 1 modulo 2 in the eukaryotic non-coding genes (introns, 5′ and 3′ regions) [3,4]. In order to understand the periodicity 0 modulo 3, this approach recently applied to protein coding genes on the four letter gene alphabet {A,C,G,T} revealed the three basic types of periodicities [5]. The results obtained here show unexpectedly that proteins and genes have periodic signals in common. The periodicity 0 modulo 3 in genes was simulated by an independent or markov mixing of three words of small size YRYRYR, YRYYRY and YRYNNNNNN (N = R or Y) (<10 letters and called oligonucleotides) [6,7]. It was also simulated with an automaton at one state generating the word $(YRYNNNNNN)^n$ followed by random insertions (in type and position) and deletions (in position) of words of three letters (called trinucleotides) [8], i.e. followed by an evolutionary process modelling the RNA editing [9,10]. The six types of modulo 3 periodicities in collagens have been simulated here with a stochastic automaton at 14 states (without evolutionary process).

The distribution of the residues X_1 and X_2 in collagens is not random. Indeed, four sub-sets are identified: {G}, {E,L} in the X_1 residue site, {K,Q,R} in the X_2 residue site and {A,D,P,S} both in the X_1 and X_2 residue sites. The constructed automaton reveals in particular that the sub-set {E,L} is followed by the sub-set {K,Q,R} and that the sub-set {A,D,P,S} is followed by itself. There is no direct link between {E,L} or {K,Q,R} and {A,D,P,S} (Fig. 2). These sub-sets may be related to three fundamental physical and chemical properties of their amino acids: the hydrophilicity, the α-helical conformation and the volume. The sub-set {K,Q,R} contains the three among 20 most hydrophilic amino acids (normalized consensus hydrophobicity parameter $H_{nc} \leq -0.85$ [11 p. 90]). The sub-set {E,L} is composed of the first and the fourth α-helical amino acids (α-helical parameter $P_\alpha \geq 1.21$ [11 p. 90]). Both sub-sets {A,D,P,S} and {G} belong (with C) to the six among 20 smallest amino acids (van der Waals volume ≤ 91 Å^3 [12 p. 141]). G is also the less α-helical amino acid ($P_\alpha = 0.57$ [11 p. 90]).

The autocorrelation function method based on the definition given in Section 2.1.1. is a new statistical approach to analyse collagens. This method is more general than the statistical study of motifs (dipeptides, tripeptides, etc) as a motif is a particular i-motif with $i = 0$. Therefore, this method necessary retrieves some previous results obtained with the statistical study of amino acids in collagens. The three identified sub-sets {E,L}, {K,Q,R} and {A,D,P,S} are consistent with the amino acid composition in the X_1 and X_2 residue sites given for example in Table 1 of [13], Table 1 of [14] and in Table 1 of [15]. However, there are mainly two differences between this new method and the classical statistical study of collagens. The autocorrelation function method is applied to several collagens (various fibrillar types and different taxonomies) and not to one sequence, and it does not use the hypothesis of particular sites for the amino acid analysis.

In order to identify a (or several) sub-set of residues having the same distribution in collagens, the classical statistical study has been extended to the statistical study of neighbouring amino acids, e.g. of three tripeptides GX_1X_2 (Table 2 in [14] and Table 2 in [15]). The autocorrelation function method allows the analysis of a large tripeptide series, e.g. ten tripeptides GX_1X_2 can be studied by varying i between 0 and 29 (Fig. 1a–f). Its graphical representation directly allows the identifica-

tion of periodicities, local maxima, etc. The sub-set {E,L} followed by the sub-set {K,Q,R} in the constructed automaton agrees with the observed tripeptide $GX_1^-X_2^+$ where $X_1^- = \{D,E\}$ and $X_2^+ = \{K,R\}$ [16,17]. Furthermore, several tripeptides deduced from the constructed automaton are consistent with the collagen-like X-ray conformation (Table V in [18]): GPP, GPA, GPS, GAP and GSP. The tripeptide GPL which does not belong to the automaton, does not have a collagen-like conformation (Table V in [18]).

Periodicities with respect to the stagger distance D (670 Å or 78 tripeptides) between adjacent triple-helical molecules in the fibril have been searched using the Fourier analysis. Fourier peaks at wavelengths of D/n (n = 5, 6, 11 [13,19,20]) have revealed different periodicities. However, this analysis cannot determine the types of periodicities. In contrast, the autocorrelation function method based on a simpler definition (Section 2.1.1.) allows the identification of the types of periodicities, e.g. for the periodicity modulo 3, the three types 0 modulo 3, 1 modulo 3 and 2 modulo 3 (Fig. 1a–c). These types are important, as seen here, both for the understanding of real collagen sequences, e.g. the relations identified in Table 2, and for the generation of model collagen sequences, e.g. the automaton in Fig. 2.

The simulation based on the constructed automaton shows that the sub-set {K,Q,R} is specific to the X_2 letter site (end of GX_1X_2). A return of the model to the reality consists in investigating the existence in collagens of sub-words including a series of GX_1X_2. Therefore, the autocorrelation function analysing the occurrence probability of $w(X)_i w'$ with w and w' in {K,Q,R}, is applied to collagens in order to identify a significant maximum value associated with the length of sub-words. By varying i between 0 and 99, Fig. 4 shows an obvious maximum for i = 17 among 100 points. This maximum is related to a preferential occurrence of the sub-set {K,Q,R} 17 residues after itself. Therefore, the collagens are divided into sub-words of length equal to 17+1 = 18 residues. Surprisingly, most of the exons (segments of genes) coding for the collagens have a length of 54 nucleotides [1 p. 841,21], i.e. a length corresponding to 54/3 = 18 amino acids. Despite of the genetic code degeneracy, the length of exons may be retrieved (conserved) in collagens. This result states the problem of segments in proteins associated with their exons.

Acknowledgements

We thank Dr Nouchine Soltanifar and the Referee for their advice. This work was supported by GIP GREG grant (Groupement d'Intérêt Public, Groupement de Recherches et d'Etudes sur les Génomes) and by INSERM grant (Contrat de Recherche Externe No 930101).

References

[1] Vuorio, E. and de Crombrugghe, B. The family of collagen genes. *Annu. Rev. Biochem.* 1990; 59: 837–872.

[2] Arquès, D.G. and Michel, C.J. A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. *J. Theor. Biol.* 1987; 128: 457–461.

[3] Arquès, D.G. and Michel, C.J. Study of a perturbation in the coding periodicity. *Math. Biosci.* 1987; 86: 1–14.

[4] Arquès, D.G. and Michel, C.J. Periodicities in coding and non-coding regions of the genes. *J. Theor. Biol.* 1990; 143: 307–318.

[5] Arquès, D.G., Lapayre, J.-C. and Michel, C.J. Deux classes de periodicités non en phase de lecture identifiées dans les gènes codants des eucaryotes et simulées à l'aide d'un automate stochastique. *Tech. Sci. Informat.* 1995; 14: 197–216.

[6] Arquès, D.G. and Michel, C.J. A model of DNA sequence evolution, Part 1 Statistical features and classification of gene populations 743–753, Part 2 Simulation model 753–766, Part 3 Return of the model to the reality 766–770. *Bull. Math. Biol.* 1990; 52: 741–772.

[7] Arquès, D.G., Michel, C.J. and Orieux, K. Analysis of Gene Evolution: the software AGE. *Comput. Applic. Biosci.* 1992; 8: 5–14.

[8] Arquès, D.G. and Michel, C.J. A model of gene evolution based on recognizable languages and on insertion and deletion operations. *Model. Simulat.* 1993; 13:110–113.

[9] Benne, R., Van Den Burg, J., Brakenhoff, J.P.J., Sloof, P., Van Boom, J.H. and Tromp, M.C. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 1986; 46: 819–826.

[10] Covello, P.S. and Gray, M.W. On the evolution of RNA editing. *Trends Genet.* 1993; 9: 265–268.

[11] Gromiha, M.M. and Ponnuswamy, P.K. Relationship between amino acids properties and protein compressibility. *J. Theor. Biol.* 1993; 165: 87–100.

[12] Creighton, T.E. Proteins, Structures and Molecular Properties, W.H. Freeman and Company, New York, 1993.

[13] McLachlan, A.D. Evidence for gene duplication in collagen. *J. Mol. Biol.* 1976; 107: 159–174.

[14] Salem, G. and Traub, W. Conformational implications of amino acid sequence regularities in collagen. *FEBS Lett.* 1975; 51: 94–99.

[15] Traub, W. and Fietzek, P.P. Contribution of the α2 chain to the molecular stability of collagen. *FEBS Lett.* 1976; 68: 245–249.

[16] Katz, E.P. and David, C.W. Energetics of intrachain salt-linkage formation in collagen. *Biopolymers* 1990; 29: 791–798.

[17] Katz, E.P. and David, C.W. Unique side-chain conformation encoding for chirality and azimuthal orientation in the molecular packing of skin collagen. *J. Mol. Biol.* 1992; 228: 963–969.

[18] Traub, W. and Piez, K.A. The chemistry and structure of collagen. *Adv. Protein Chem.* 1971; 25: 243–352.

[19] Hulmes, D.J.S., Miller, A., Parry, D.A.D., Piez, K.A. and Woodhead-Galloway, J. Analysis of the primary structure of collagen for the origins of molecular packing. *J. Mol. Biol.* 1973; 79: 137–148.

[20] Hulmes, D.J.S., Miller, A., Parry, D.A.D. and Woodhead-Galloway, J. Fundamental periodicities in the amino acid sequence of the collagen α1. *Biochem. Biophys. Res. Comm.* 1977; 77: 574–580.

[21] Yamada, Y., Avvedimento, V.E., Mudryi, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. and de Crombrugghe, B. The collagen gene: evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. *Cell* 1980; 22: 887–892.