MHR

REVIEW

J. C. BIRO AND A. M. BIRO [2005] MED HYPOTHESES RES 2: 347-355.

PREDICTION OF STABILIZING ELECTROSTATIC INTERACTIONS BETWEEN RESIDUES IN COLLAGENS

J. C. BIRO* AND A. M. BIRO

KAROLINSKA INSTITUTE, STOCKHOLM, SWEDEN (J.C.B.) AND HOMULUS INFORMATICS, 88 HOWARD, #1205, 94 105 CA, USA (A.M.B.)

ABSTRACT. ALL COLLAGEN SEQUENCES have a distinctive signature, described by the X-Y-Gly formula indicating that any amino acid might be present at X and Y positions, in many combinations, while the third position is fixed and invariably glycine. The unique periodic nature of these sequences makes it possible to perform a reliable statistical study on the physico-chemical properties of amino acids at X and Y positions. In this study we have phase separated twenty different main human collagen sequences into three subsequences each. We have found that the X-Y-Gly formula is frequently corrupted by phase shifts caused by deletion of a glycine codon. The overall average charge of amino acids at X positions is always negative, while at Y positions it is always positive. No exception was found. This indicates the periodic nature of collagens even at X and Y positions and predicts a pattern-related interaction in and between collagen triple-helices.

*ADDRESS ALL CORRESPONDENCE TO: DR. JAN C. BIRO, 88 HOWARD, #1205, SAN FRANCISCO, 94 105 CA, USA. PHONE: 1-415-777-1443. E-MAIL: jan.biro@sbcglobal.net

● MEDICAL HYPOTHESES AND RESEARCH ● THE JOURNAL FOR INNOVATIVE IDEAS IN BIOMEDICAL RESEARCH ●

INTRODUCTION

All collagens have a distinctive molecular conformation: a triple-helix composed of three supercoiled polyproline II-like helical chains [1-4]. This triple-helical conformation places strict constraints on amino acid sequence, requiring Gly at every third residue and a high content of proline and hydroxyproline residues. The collagen primary sequence is usually described by the X-Y-Gly formula, where X and Y indicate any amino acid, but the glycine (Gly) is compulsory at every third position. The single collagen strands are very unstable peptides. For mammals and birds, the denaturation temperature appears to be a few degrees higher than body temperature, typically cited as approximately 41.5 Co [5], while the triple-stranded configuration is extremely stable having lifetimes of at least 6 months, and often much longer. This property makes it very interesting to understand the exact organization of the molecular interactions in and between the triple-helical peptide structures. It is widely accepted that the major force which keeps the three strands together are the direct - H - or water mediated bonds between the peptide backbones of the different strands [6,7], however, recent studies indicate that even the amino acid residues play an important role [7,8]. The unique periodic nature of the collagen primary sequence makes it possible to use statistical and bioinformatical methods to examine the physico-chemical properties of the amino acid residues at X and Y positions and make some predictions for the overall behavior of the collagen polymer.

MATERIALS AND METHODS

Protein sequences were selected from the SWALL databank, which is a virtual collection of the SWISSPROT, SPTREMBL and TREMBL-NEW [9].

We developed a JAVA program, called the SeqForm, to find and alter sequence residues. The program makes it possible to predefine residues in optional phases of a sequence (for example every 1st, 2nd or 3rd, even in combinations), replace or remove them and perform the same operation on large files, containing many FASTA sequences. Sequence similarity visualization was performed by DOTLET, which is a dot-plot program [10].

Statistical analyses of protein sequences were performed by the SAPS program [11]. Linear regression analyzes and Students t-tests were used for statistical evaluation of the results.

RESULTS

Twenty human collagen sequences were taken from the SWALL database each belonging to a different A1 class. They are known to form different histological structures, having different physiological roles and interact with different molecules, even if they all are similarly folded and are major components of the extracellular matrix (TABLE 1, FIG. 1).

The sequences were processed by SeqForm and divided into three sub-sequences each containing every third residues. The sub-sequences of an ...abcdefghi... sequence became ...adg..., ...beh... and ...cfi... using period 3 selection and are referred to as sequence phase (SP) 1-2-3 (not to confuse with frame 1-2-3, which is used to define the phases of a nucleic acid translation). The period 3 selection of the human collagens showed the presence of long (>10 residue) poly-Gly sequences which were located in one or different phases. Therefore, the primary SP1-2-3 selections did not directly follow the expected X-Y-Gly formula in each case and for the entire sequences. The poly-Gly sequences often shifted phase (TABLE 1). The average length of the collagen protein sequences was 1444 ± 149 residues, it occurred 7.3 ± 1.2 phase shifts/sequence and the distance between phase shifts was 318 ± 78 residues (means ± S.E.M, N = 20).

It was possible to exactly locate the place of the phase-shifts, in most cases. Phase-shift is always caused by codon deletion (and not a single base deletion) and in our case always by deletion of a Gly codon (FIG. 2).

We have re-arranged the primary SP1-2-3

Human Collagen Classes and Properties

Name	Ac-prot	Ac-na	L	n	L/n+1	Function (a)	Associated disease (a)	Main location (a)
COL1A1	NM_000088	NP_000079	1464	1	732	fibrillar	osteogenesis imperfecta	bone, skin, tendor
COL2A1	NM_001844	NP_001835	1487	0	1487	fibril forming	pr. osteoarthritis	cartilagious tissues
COL3A1	X14420	P02461	1466	1	733	fibril forming	Ehlers-Danlos syndrome	most soft connective tissues
COL4A1	NM_001845	NP_001836	1940	19	97	'chicken-wire' meshwork	Ehlers-Danlos syndrome	glomerular basement membrans
COL5A1	D90279	BAA14323	1838	4	367	fibril forming		ubiquitous distribution
COL6A1	X66405	Q04857	1025	2	341	cell-binding protein		
COL7A1	L02870	AAA75438	2944	16	173		epidermolysis bullosa	epithelial basement membran
COL8A1	X57527	P27658	744	8	82			endothelial Descemet membrane
COL9A1	X54412	CAA38276	921	3	230			fibril associated
COL10A1	NM_000493	NP_000484	680	8	75		chondrodysplasia	hyaline cartillage
COL11A1	J04177	AAA51891	1806	3	451	fibrillogenesis	Stickler syndrome	
COL12A1	NP_004361	NP_004361	3063	4	612			tissu specific
COL13A1	M59217	AAA51685	584	7	73			
COL14A1	BC014640	AAH14640	759	5	126			fibril-associated
COL15A1	L25286	AAA58429	1388	13	99			
COL16A1	M92642	AAA58427	1603	16	94			placenta
COL17A1	NM_000494	NP_000485	1497	10	136		epidermolysis bullosa	
COL18A1	AF018081	AAC39658	1516	14	101		Knobloch syndrome	
COL19A1	NM_001858	NP_001849	1143	9	114	cross-bridge between fibrils		
COL21A1	AF414088	AAL02227	957	3	239			
Mean			1444	7.3	318			
+ / - S.E.M.			149	1.2	78			

Ac: accession number; prot: protein; na: nucleic acid; L: length, number of amino acid residues; n: number of phase-shifts: (a): data restricted for that found in GeneCards.

Different forms of the Collagen Structure and Sequence

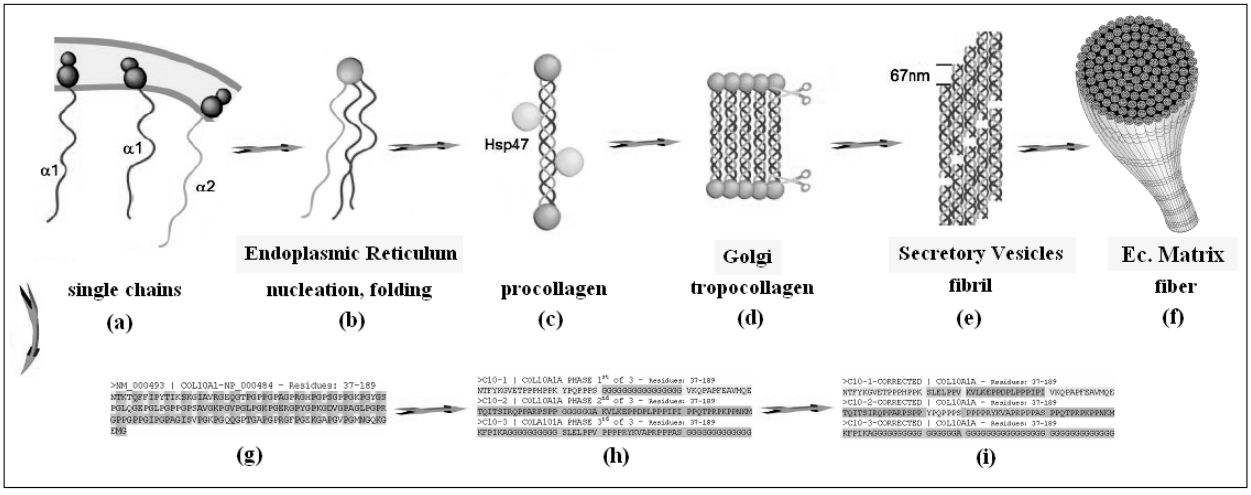

FIGURE 1. DIFFERENT FORMS OF THE COLLAGEN STRUCTURE AND SEQUENCE. Single collagen polypeptide chains (~ 1.400 amino acid long) are synthesized in the endoplasmic reticulum (a). They quickly associate with each other and a nucleation-like process triggers formation of triple-stranded, supercoiled helical structures (b). The process is probably assisted by chaperons, like Hsp47 (c). Cutting of the non-helical ends of pro-collagens leads to tropocollagens, in the Golgi, which is a highly ordered structure of collagen rods (length: 300 nm, diameter: 1.5 nm) and contains 3.3 amino acids/turn (d). The collagen specific 67 nm wide optical pattern occurs in fibrils (diameter: 10-300 nm) (e), which continues to assemble into fibers (diameter: 0.5-3 μm). The lower part of the figure explain the characteristic X-Y-Gly pattern in the primary sequence of single collagen molecules (g) and the data processing used in this article. Phase separation of the sequences into three subsequences displays shifting of poly-Gly sequence between phases (h) which might be easily reversed and corrected (i).

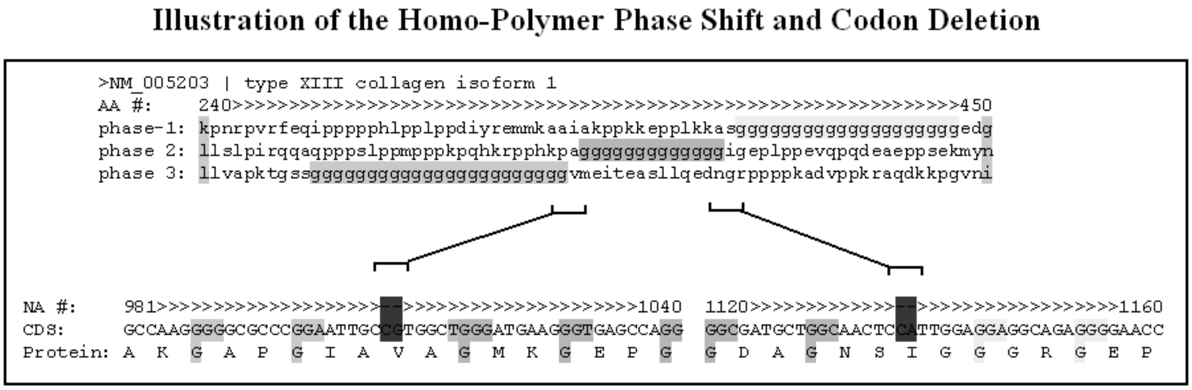

FIGURE 2. ILLUSTRATION OF THE HOMO-POLYMER PHASE SHIFT AND CODON DELETION. Phase separation of a collagen sequence into three subsequences (phase-1-2-3) displays the phase-shifting poly-Gly (...gggg...) homo-polymer. A Gly codon is missing from the coding sequence (CDS) at the place of phase-shifts (red block). Only a short sequence of collagen molecule is shown as indicated by the amino acid and nucleic acid positions numbers (AA # and NA #).

selections by turning the subsequences at every phase shift positions to “shift-back” the poly-Gly strings to only one single phase. This way we could clearly re-format the sequences and the X-Y-Gly formula applied for the entire sequence and in each case.

Statistical analyzes showed that the physico-chemical properties of amino acids are statistically significantly different at X and Y positions (TABLE 2). The primary collagen sequences contain equal number of positively vs. negatively charged, as well as apolar vs. polar amino acids. However the amino acids in the X phase contains significantly more negative than positive charges and the amino acids in the Y phase contains significantly more positive than negative charges (FIG. 3). Both X and Y phases contain significantly more apolar (hydrophobic) than polar (hydrophilic) residues (FIG. 4). The third phase, the poly-Gly phase is not included in the table and figures because it is known to be neutral and highly polar due to the great excess of Gly residues.

A non-collagen protein group, the 6 insulin-like growth factor binding proteins, was used as control. They did not show any separation of physico-chemical properties after similar phase selection.

The distribution of charges and polarity along the X and Y subsequences was studied by dot-plot method. Comparison of X and Y phases from Col1A1 and Col2A2 with each other and a standard showed that the charge and polarity is evenly distributed along the sequences. The distribution of

TABLE 2. PERIODIC DISTRIBUTION OF CHARGE AND POLARITY IN COLLAGENS.

	COLLAGEN 1-20				IGF-BP 1-6				
	P	SP			P	SP			
		X	Y			A	B	C	
POZITIVE %	50.2 ± 0.5	19.1 ± 0.8	34.2 ± 1	p<0.001	56.4 ± 2.2	18 + 2.1	20 + 1.9	19 + 1.8	ns
NEGATIVE %	49.8 ± 0.5	30 ± 0.8	16.6 ± 1	p<0.001	43.6 ± 2.2	13 + 0.7	17 + 1.9	13 + 1.8	ns
	ns	p<0.001	p<0.001		p<0.001	p<0.01	ns	p<0.01	
APOLAR %	52.1 ± 1.3	34.1 ± 0.7	33.2 ± 0.7	ns	51 ± 2.1	19	16	17	ns
POLAR %	47.9 ± 1.3	16.3 ± 0.7	16.4 ± 0.7	ns	49 ± 2.1	15	17	17	ns
	ns	p<0.001	p<0.001		ns	ns	ns	ns	

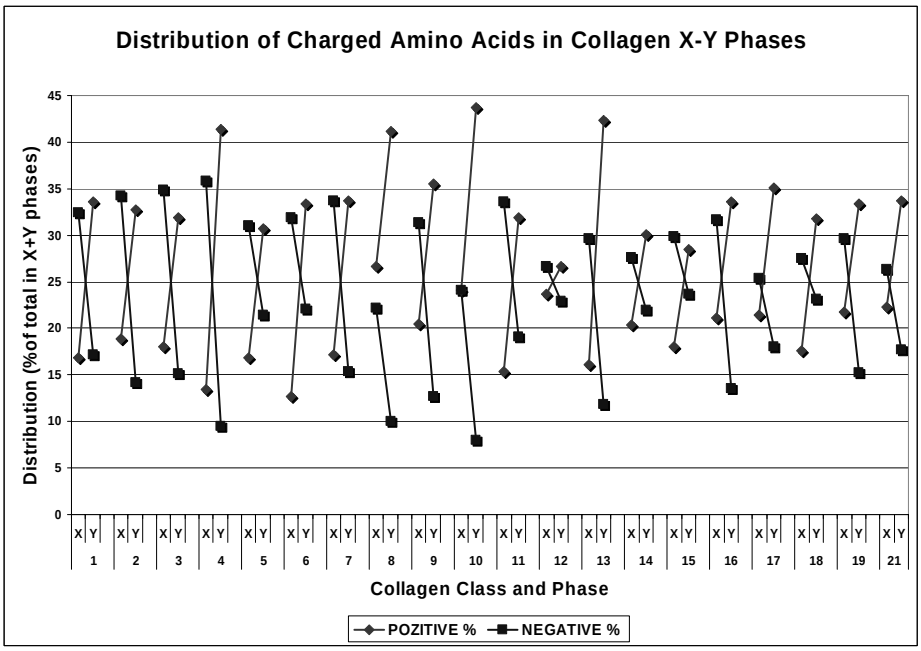

FIGURE 3. DISTRIBUTION OF CHARGED AMINO ACIDS IN COLLAGEN X-Y PHASES. The primary sequence of 20 different A1-class collagens was phase separated into three subsequences and the proportion of positively and negatively charged amino acids in X and Y phases is displayed.

positive and negative charges in the X and Y phases from the same collagen are similar but not exactly matching with each other. The sequence similarity between the original Col1A1 and Col2A1 was maintained in both X and Y sequences even after phase separation (FIG. 5, 6). These rules did apply even for other collagen classes except that they are less similar to each other before and after phase separation (these results are not shown).

DISCUSSION

The basic principles of triple-helix stability is the subject of intensive research and dispute since the first correct model of collagen structure was proposed by Ramachandran and Kartha [1,2] and was immediately declared as "stereochemically unsatisfactory" by Rich and Crick [12,13]. Now it is clear that direct hydrogen and water mediated bonds exist between the Gly molecules in the three neighbor strands, which is possible only because all Gly residues are uniformly, centrally located in the longitudinal axis of the collagen triple-helix. There are no -H- bonds in the same strand, like in the alpha helix. These bonds are positioned between -CN....O=C- or Cα-H....O=C atoms of the peptide backbone [6]. The Gly is a polar amino acid and these interconnected residues form a continuous, hydrophilic, longitudinal core in the collagen. Any interruption of the continuity of this core, for example, replacement of Gly with Ala might alter the structure and function of collagen, as it is known from hereditary disease, like osteogenesis imperfecta [14]. The residues of the non-Gly amino acid are radially located on the surface of the collagen. They were logically suspected to play a role in the inter-collagen interactions, but not in stabilizing the triple-spiral. However, the first high-resolution structure of a triple-helical collagen-mimics 4 indicated a regular network of water molecules surrounding the triple-helices in the crystal lattice. The structure appeared to lend support to a hypothesis that Hyp (hydroxyproline, a very fre-

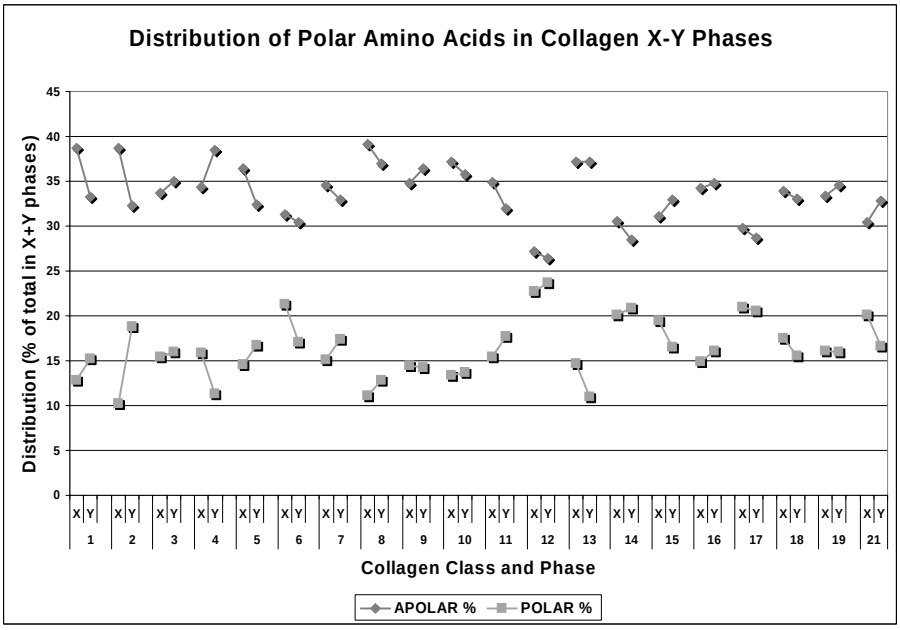

FIGURE 4. DISTRIBUTION OF POLAR AMINO ACIDS IN COLLAGEN X-Y PHASES. The primary sequence of 20 different A1-class collagens was phase separated into three subsequences and the proportion of apolar and polar amino acids in X and Y phases are displayed.

quent amino acid in these molecules) stabilizes collagen by forming hydrogen bonds with water molecules surrounding the triple helix [15]. Confusingly, collagen is stable in anhydrous environment of methanol or propane, [16] and replacement of proline by flouroproline improves collagen stability [17] although fluorine does not form hydrogen bonds [18]. Furthermore, the triple-helix stability is sequence dependent, which is an additional indication of the role of the individual amino acid residues at the X and Y positions [19,20].

The tertiary and quaternary structure of collagens is not a completed capital in the molecular biology. The available studies in the PDB are restricted for short synthetic collagens, and the residue variation in these structures is poor in giving any insight into the role of individual residues in the structure forming. However, we have access to a large amount of primary sequence data and the relatively simple and periodic structure of collagens makes them ideal subjects for bioinformatics studies.

Phase-separation of each major human collagen primary sequences into three subsequences turned out to be a useful sequence visualization method, which clearly displayed and confirmed the X-Y-Gly organization of amino acid residues in collagens. Additionally, we discovered that there are numerous phase-shifts in this organization due to deletion of Gly which is caused by codon deletion. The high frequency of Gly-codon deletions indicate that it is a rule and not an exception and that Gly deletion is an accepted event in contrast to Gly-replacement, which has clear pathological consequences [14,21].

We have observed that the physicochemical properties of X and Y residues are different and this difference increased by "turning back" the subsequences at the positions of phase shifts and "correcting" their effects. The number of negatively charged amino acids in X-phase is about two times as much as the number of positively charged residues. Contrary, the number of positively charged residues in Y-phase is about two times as much as the number of negatively charged residues. No exception was found. This indicates a periodical distribution of charges in collagens. Consequently,

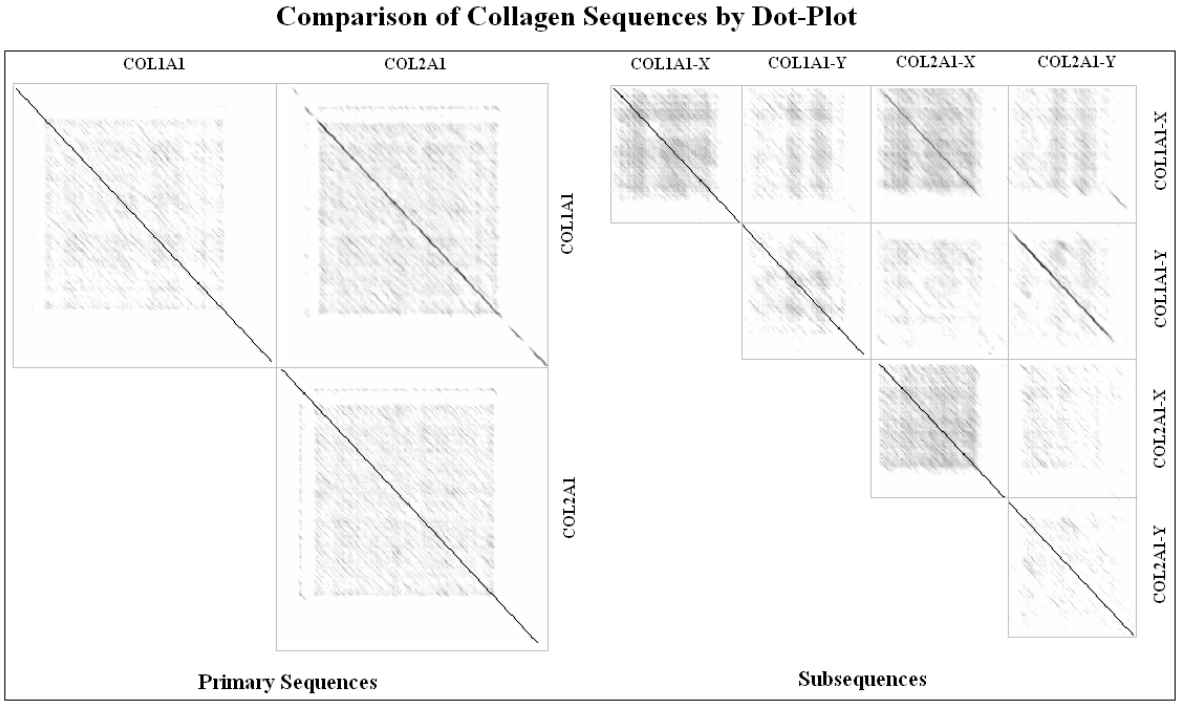

FIGURE 5. COMPARISON OF COLLAGEN SEQUENCES BY DOT-PLOTAmino acid sequences of two main collagens were compared by dot-plot before (primary sequences) and after phase separation into three (X-Y-Gly) subsequences. Identity matrix with window 59 was used.

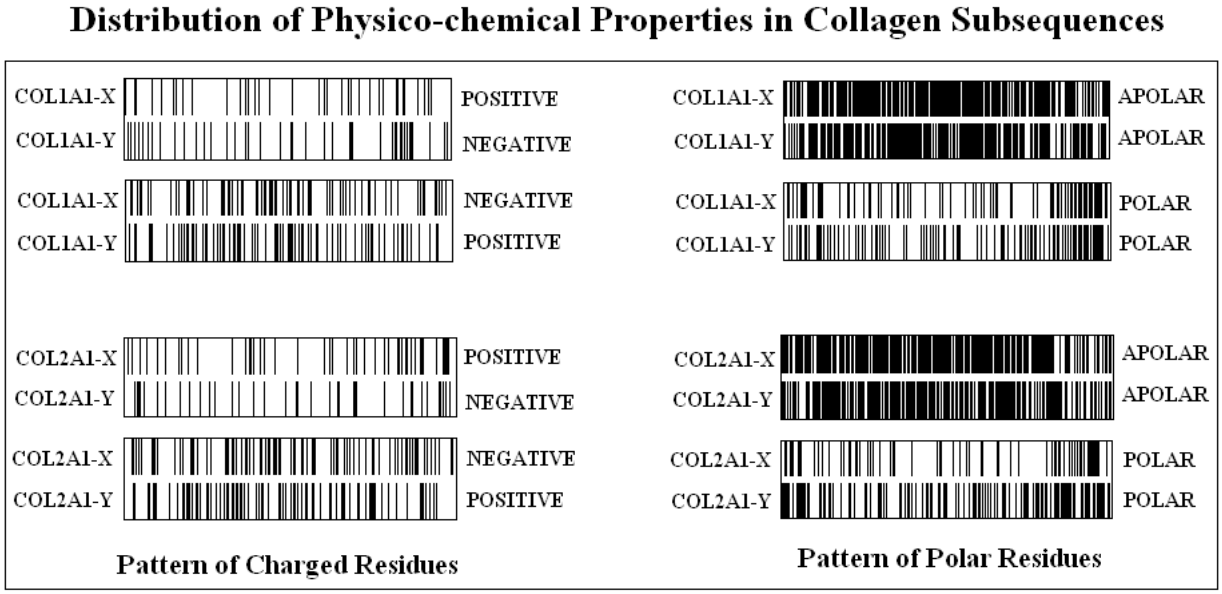

FIGURE 6. DISTRIBUTION OF PHYSICO-CHEMICAL PROPERTIES IN COLLAGEN SUBSEQUENCES. The positions of charges and polar residues in X and Y subsequences of two major collagens are indicated by barrs. The barr-code-like patterns are aligned with each other to reveal the site of potential interaction between residues.

the overall average charge of the collagens is negative at X-, while positive at Y- positions at pH 7.0. This is a simple, general, statistical prediction of stabilizing electrostatic interactions between residues in collagens and is in agreement with the suggestion, which emphasizes the role and influence of charged amino acid residues in forming the collagen structure [20,22]. The distribution pattern of positive vs. negative charges along the X and Y subsequences is not identical. This indicates that these opposite charges are not necessarily located side by side in the same single collagen strand. They would not necessarily be in opposite positions either, when two identical (or highly similar) single strands interact with each other. Opposite charges may, but not necessarily, bind to each other in collagens. It is known, for example, that Glu and Lys display even direct interactions with carbonyl groups and Hyp hydroxyl groups or interactions mediated by water molecules [23]. Therefore, it might be necessary to consider even charge related interactions due to ionization rather than ion pair formation.

Notably, the number of apolar (hydrophobic) residues is about twice as much in both X and Y positions than the number of polar (lipophobic) residues indicating an overall hydrophobic nature of collagen surface. Obviously this hydrophobic surface is interrupted by hydrophilic residues, which might explain the highly ordered hydration network around the collagens [22].

Our results indicate that the periodic nature of collagens is not restricted to the periodic presence of Gly at every third position but it is a distinctive collagen characteristic which applies even for other physicochemical properties. The regular positive/negative charge pattern of the collagen surface might be a simple but strong signature of collagen molecules and it might generally attract different other collagens and assist to establish the collagen network.

REFERENCES

[1] RAMACHANDRAN GN AND KARTHA G [1954] Structure of collagen. Nature 174: 269-270.

[2] RAMACHANDRAN GN AND KARTHA G [1955] Structure of collagen. Nature 176: 593-595.

[3] RICH A AND CRICK FHC [1961] The molecular structure of collagen. J Mol Biol 3: 483-506.

[4] BELLA J, EATON M, BRODSKY B AND BERMAN HM [1994] Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. Science 266: 75-81.

[5] BACHINGER HP, MORRIS NP AND DAVIS JM [1993] Thermal stability and folding of thecollagen triple helix and the effects of mutations in osteogenesis imperfecta on thetriple helix of type I collagen. Am J Med Gen 45: 152-162.

[6] BELLA J AND BERMAN HM [1996] Crystallographic evidence for C alpha-H...O=C hydrogenbonds in a collagen triple helix. J Mol Biol 264: 734-742.

[7] CARA L, JENKINS CL AND RAINES RT [2002] Insights on the conformational stability ofcollagen Nat Prod Rep 19: 49-59.

[8] Persikov AV, Ramshaw JA, Kirkpatrick A and Brodsky B [2000] Amino acidpropensities for the collagen triple-helix. Biochemistry 39: 14960-14967.

[9] BAIROCH A AND APWEILER R [1997] The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25: 31-36.

[10] DOTLET
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html).

[11] BRENDEL V, BUCHER P, NOURBAKHSH I, BLAISDELL BE AND KARLIN S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 89: 2002-2006.

[12] RICH A AND CRICK FH [1955] The structure of collagen. Nature 176: 915-916.

[13] RICH A AND CRICK FH [1955] Structure of polyglycine II. Nature 176: 780-781.

[14] YANG W, BATTINENI ML AND BRODSKY B [1977] Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. Biochemistry 36: 6930-6935.

[15] KRAMER RZ AND BERMAN HM [1998] Patterns of hydration in crystalline collagen peptides. J Biomol Struct Dyn 16: 367-380.

[16] SEMISOTNOV GV, ET AL. [1991] Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe. Biopolymers 31: 119-128.

[17] HOLMGREN SK, TAYLOR KM, BRETSCHER LE AND RAINES RT [1998] Code for collagen's stability deciphered. ture 392: 666-667.

[18] LYR H, FIEDLER HJ AND TRANQUILLINI W (EDS) [1992] Physiologie und Ökologie der Gehölze Fischer, Jena/Stuttgart (1992).

[19] PERSIKOV AV, RAMSHAW JA AND BRODSKY B [2000] Collagen model peptides: Sequence dependence of triple-helix stability. Biopolymers 55: 436-450.

[20] PERSIKOV AV, RAMSHAW JA, KIRKPATRICK A AND BRODSKY B [2002] Peptide investigations of pairwise interactions in the collagen triple-helix. J Mol Biol 316: 385-394.

[21] BECK K, CHAN VC, SHENOY N, KIRKPATRICK A, RAMSHAW JA AND BRODSKY B [2000] Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine. Proc Natl Acad Sci USA. 97: 4273-4278.

[22] BRODSKY B AND RAMSHAW JA [1997] The collagen triple-helix structure. Matrix Biol 15: 545-554.

[23] KRAMER RZ, VENUGOPAL MG, BELLA J, MAYVILLE P, BRODSKY B AND BERMAN HM [2000] Staggered molecular packing in crystals of a collagen-like peptide with a single charged pair. J Mol Biol 301: 1191-1205.

RECEIVED ON 12-2-2004.

ACCEPTED ON 2-30-2005.