Bulletin UASVM Animal Science and Biotechnologies, 65(1-2)/2008
pISSN 1843-5262; eISSN 1843-536x

CYCLICITY ANALYSIS OF AMINO-ACIDS ON TYPE I COLLAGEN CHAINS

Bolboacă Sorana D.[1], L Jäntschi [1,2]

[1] "Iuliu Haţieganu" University of Medicine and Pharmacy, Depart. of Medical Informatics and Biostatistics, 6 Louis Pasteur Street, 400349 Cluj-Napoca, Romania
[2] Technical University, Depart. of Chemistry, 103-105 Muncii Bvd, 400641 Cluj-Napoca, Romania
email: sbolboaca@umfcluj.ro, lori@academicdirect.org

Key words: Amino Acids Sequence; Type I Collagen; Cyclicity; Rank Correlation.

Abstract: The pattern distribution of standard amino acids on type I collagen chains was investigated in rank correlation and autocorrelation analysis. The alpha 1 and alpha 2 chains of five species (*Bos taurus*, *Canis lupus*, *Danio rerio*, *Homo sapiens*, and *Rattus norvegicus*) were investigated. A series of PHP programs were created in order to accomplish the aim of the research. The rank correlation analysis showed that a moderate to a very good correlation subsist between ranks of position distribution of standard amino acids in the investigated type I collagen chains on all species. The autocorrelation analysis revealed that the amino acid sequences on type I collagen chains have not a repeating patters.

INTRODUCTION

Type I collagen is the most common type of collagens in vertebrates. The main locations over the body are skin, tendons, ligaments, cornea, intervertebral disks, dentine, arteries, granulation tissues, cartilages (Wardale and Duance, 1993). The collagen extracted from animals' connective tissue (bovine, ovine, caprine, deer, elk, mink, and cats) is used in gelatine industry (Venien and Levieux, 2005). Type I collagen is also used biomaterials engineering (Luo et al., 2008; Cummings et al., 2004). Understanding the complex organization of type I collagen could lead to better understanding of its structure and improvement of the products obtained gelatine industry and biomaterial engineering.

The aim of the present research was to identify and analyze the regularities in the amino acid distribution on rank correlation and autocorrelation analysis of the type I collagen chains on five species: *Bos Taurus*, *Canis lupus*, *Danio rerio*, *Homo Sapiens*, and *Rattus Norvegicus*.

MATERIAL AND METHODS

Type I Collagen

The alpha 1 (α1) and alpha 2 (α2) chains of collagen type I (CTI) of five species were taken from the National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/] The following five species were investigated: *Bos taurus* (Shirai *et al.*, 1998), *Canis lupus* (Lowe *et al.*, 2003), *Danio rerio* (Dubois *et al.*, 2002), *Homo sapiens* (Strausberg *et al.*, 2002), and *Rattus norvegicus* (Orjel *at al*., 2006).

The type I collagen chains comprise twenty standard amino acids: alanine (A), arginine (R), asparagine (N), aspartate (D), cysteine (C), glutamine (Q), glutamate (E), glycine (G), histidine (H), isoleucine (I), leucine (L), lysine (K), methionine (M), phenylalanine (F), proline (P), serine (S), threonine (T), tryptophan (W), tyrosine (Y), and valine (V). The α1

type I collagen chain of *Rattus norvegicus* comprises 116 unknown amino acids (out of 1054; abbreviated as X) while the α2 chain comprises 102 unknown amino acids (out of 1026).

The most frequent amino acids in the type I collagen chains of investigated species is glycine (Bolboacă and Jäntschi, 2007).

Statistical Methods of Analysis

Two types of analysis were performed in order to identify the regularities within type I collagen chains on investigated species: a rank correlation and an autocorrelation analysis.

The chains of type I collagen were transformed into a matrix with columns represented by the amino acid (20 columns in our case) and rows represented by the position in the chain (e.g. 1463 rows for investigation of α1 chain of type I collagen for *Bos taurus*). The matrix was filled with 1 and 0 (when the amino acids of interest was present in the place of interest a value of 1 was assigned, otherwise a 0 was placed).

Rank Correlation Analysis

The steps in rank correlation analysis were:

÷ Step 1: matrix representation. The matrix of position of each standard amino acid on each collagen type I chain for every species was obtained. The columns contain the amino acid of interest (SP_α1/2_Z, where SP is the abbreviation of the species - e.g. HS for *Homo sapiens*, α1/2 is the type of chain, and Z is the one-letter abbreviation of standard amino acid). The rows contain the number of amino acid in the chains (from 0 - when the amino acid was not present on the investigated specie to 390 – glycine on *Canis lupus* - α1 chain). For example, the first apparition of the glycine (the most frequent amino acid on all investigated species) appears on position 5 (*Ratus norvegicus* α1 chain), 6 (*Ratus norvegicus* α2 chain), 22 (*Homo sapiens* α1 chain, *Canis lupus* α1 chain, *Bos taurus* α1 chain, *Danio rerio* α1 chain).

÷ Step 2: calculation of Spearman rank correlation (the rank of each amino acid was correlated with the rank of all other amino acids on the chain for the same specie).

Autocorrelation Analysis

Autocorrelation is a mathematical tool for identifying repeating patterns and it is used frequently in signal processing (Broersen, 2006). In statistics, the autocorrelation describes the correlation of a data set with itself, offset by n-values. The autocorrelation analysis of amino acids on type I collagen chains was performed between adjacent entries (an autocorrelation by order k = 1). The autocorrelation with an offset of 1 correlate the data set $\{aa_2, aa_3, aa_4, aa_5,\ldots, aa_n\}$ with the data set $\{aa_1, aa_2, aa_3, aa_4,\ldots, aa_{n-1}\}$. The higher positive value of correlation coefficient was search and identified.

A series of homemade programs were developed in PHP (PHP Hypertext Preprocessor)[http://www.php.net/(viewed June 08, 2008)] for performing the rank correlation and autocorrelation analysis. Note that one limitation of the analysis is given by the possibility of obtaining a positive correlation by 0 (the absence of amino acid of interest).

RESULTS AND DISCUSSIONS

Rank Correlation Analysis

The correlations coefficient varied from 0.2789 (DRα1 L (37 leucine on the chain) - Y (9 tyrosine on the chain)) to 1 (RNα1 V (18 valine on the chain) - H (3 histidine on the chain) & HSα1 S (35 serine on the chain) - Y (3 tyrosine on the chain)). A minimum value of 0.4905

(BTα1 V (42 valine on the chain) - Y (16 tyrosine on the chain)) is obtained when all amino acids with appearance less than 10 are deleted.

Table 1

Frequency apparition on correlations classes

Specie_αchain	Deleted aa	r_{min} (where)	r_{max} (where)	r < 0.5	r ≥ 0.5	r ≥ 0.75	r ≥ 0.95	r ≥ 0.99
BTα1	W	0.4905 (V-Y)	0.9987 (L-R)	2	151	140	77	9
BTα2	C, W	0.6857 (H-L)	0.9989 (G-R)	0	136	133	103	23
CLα1	W	0.5438 (C-Y)	0.9974 (K-R)	0	153	145	90	12
CLα2	W	0.5255 (L-Y)	0.9988 (G-R)	0	153	138	105	17
DRα1	W, Y	0.5959 (H-L)	0.9968 (E-R)	0	153	151	97	11
DRα2	C, W	0.5852 (H-L)	0.9978 (G-P)	0	153	145	100	24
HSα1	H, M, W, Y	0.7363 (C-L)	0.9983 (G-P)	0	120	119	72	9
HSα2	C, W	0.5033 (M-Y)	0.9989 (G-R)	0	153	136	98	21
RNα1	C, I, M, W, Y	0.8953 (A-N)	0.9993 (G-X)	0	105	105	95	25
RNα2	C, M, W, Y	0.8709 (S-T)	0.9983 (G-P)	0	120	120	100	19

BTα1 = *Bos taurus* TICα1; BTα2 = *Bos taurus* TICα2; TIC = type I collagen;
aa = amino acid (one-letter abbreviation of them as it is presented in MATERIAL AND METHODS section)
CLα1 = *Canis lupus* TICα1; CLα2 = *Canis lupus* TICα2;
DRα1 = *Danio rerio* TICα1; DRα2 = *Danio rerio* TICα2;
HSα1 = *Homo sapiens* TICα1; HSα2 = *Homo sapiens* TICα2;
RNα1 = *Rattus norvegicus* TICα1; RNα2 = *Rattus norvegicus* TICα2.

The rank correlation between positions of amino acids in the same chain revealed to be from moderate to good association ($0.5 < r < 0.75$) to very good level of association ($r > 0.75$) (Colton, 1974). The analysis of the obtained rank correlation coefficients leads to the conclusion of similarity distributions of glycine (the most frequent amino acids on all species) and arginine (one of the top-three amino acids as frequency). The maximum value of rank correlation coefficient is always obtained for apparition of amino acids with an absolute frequency greater than or equal with 50. Moreover, the absolute frequencies on amino acids implied in the higher values of correlations coefficients are not the higher values (e.g. glycine, followed by the proline are the most frequent amino acids for BT_α2 (Bolboacă and Jäntschi, 2007)). The rank correlation coefficients shown that there is a moderate to very good relationships between rank distributions of amino acids position on type I collage chains.

Autocorrelation Analysis

Fifty-six out of one hundred positive autocorrelations were identified (see Table 2). The maximum number of amino acids identified on autocorrelation analysis was obtained in investigation of alanine type I collagen α1 chain for *Canis lupus* (nine out of twenty amino acids, 45% $_{95\%\text{-}fa}$CI [5 - 14], where f_a = absolute frequency). The values obtained in investigation of α1 type I collagen of *Bos Taurus* and α2 type I collagen of *Homo sapiens*, respectively (eight out of twenty amino acids, 40% $_{95\%\text{-}fa}$CI [5 - 14], where f_a = absolute frequency) were also closed to this value. The lowest performances were obtained in autocorrelation of type I collagen chains for *Ratus norvegicus* (a positive autocorrelation was obtained for just one amino acid on both chains - alanine for α1 chain and glycine for α2 chain). This could be explained by the presence of a large amount of unspecified/unknown amino acids on these chains (116 on α1 chain and 102 on α2 chain) and the absence of two out of twenty standard amino acids (cysteine and tryptophan).

The dimension of the type I collagen substructures that autocorrelated varied from 7 (*Ratus norvegicus* α2 chain of type I collagen - glycine; r = 0.7300) to 1462 (*Bos Taurus* α1 chain of type I collagen - leucine; r = 0.012) amino acids. The number of simultaneously presence of amino acid of interest in the same position on both substructures varied from 2 to

18 (lower positive correlations were identified, the higher value being of 0.012, *Danio rerio* on 1037 amino acids).

The best performances expressed as higher correlation coefficients obtained in autocorrelation analysis are presented in Table 2. The dimension of the amino acids substructures varied from 7 (*Rattus norvegicus* α2 chain, type I collagen - glycine) to 1449 (Bos taurus α1 chain, type I collagen - glutamine).

Table 2

Autocorrelations results

Chn	Siz	Smi	Sma	Spr	r	Chn	Siz	Smi	Sma	Spr	r
BTα1_A	380	26	27	3	0.0470	DRα1_A	314	26	27	5	0.1140
BTα1_D	271	16	16	2	0.0700	DRα1_D	214	17	17	3	0.1050
BTα1_E	28	3	4	2	0.5190	DRα1_G	161	18	18	3	0.0620
BTα1_L	43	8	8	5	0.5390	DRα1_I	1422	33	34	2	0.0370
BTα1_P	152	25	26	6	0.0810	DRα1_K	1405	55	55	4	0.0350
BTα1_Q	1449	50	50	2	0.0060	DRα1_L	12	4	5	3	0.4780
BTα1_T	1232	26	27	2	0.0550	DRα1_T	1414	51	52	2	0.0030
BTα1_V	1222	29	30	2	0.0450	DRα2_A	216	19	19	3	0.0770
BTα2_A	333	31	32	3	0.0010	DRα2_K	1310	47	47	3	0.0290
BTα2_K	1317	45	46	3	0.0330	DRα2_L	12	4	5	3	0.4780
BTα2_L	12	4	5	3	0.4780	DRα2_N	517	14	15	2	0.1130
BTα2_N	1141	27	28	2	0.0500	DRα2_P	74	12	12	2	0.0050
BTα2_P	49	5	6	2	0.2850	DRα2_S	1205	56	57	3	0.0070
BTα2_V	713	21	21	2	0.0680	HSα1_A	381	27	28	3	0.0400
CLα1_A	326	22	23	2	0.0210	HSα1_D	272	16	16	2	0.0700
CLα1_D	268	15	15	2	0.0820	HSα1_E	465	26	27	2	0.0200
CLα1_E	28	3	4	2	0.5190	HSα1_L	43	8	8	5	0.5390
CLα1_L	39	8	8	5	0.5280	HSα1_P	176	27	27	6	0.0810
CLα1_P	83	5	6	2	0.3200	HSα2_A	104	6	7	2	0.2630
CLα1_Q	1375	47	48	2	0.0080	HSα2_E	377	17	18	2	0.0710
CLα1_T	1229	29	30	2	0.0450	HSα2_K	1319	45	46	3	0.0330
CLα1_V	1219	29	30	2	0.0450	HSα2_L	12	4	5	3	0.4780
CLα1_Y	1211	5	6	2	0.3620	HSα2_N	1143	26	27	2	0.0540
CLα2_A	335	31	32	3	0.0010	HSα2_P	49	5	6	2	0.2850
CLα2_K	1319	45	46	3	0.0330	HSα2_S	1219	43	44	2	0.0110
CLα2_L	12	4	5	3	0.4780	HSα2_V	752	27	28	3	0.0750
CLα2_N	1143	26	27	2	0.0540	RNα1_A	184	16	16	2	0.0420
CLα2_P	49	5	6	2	0.2850	RNα2_G	7	2	3	2	0.7300

Chn = the abbreviation of the species, type I collagen chain (α1/α2), amino acid (one letter abbreviation, see MATERIAL AND METHODS - type I collagen);
BTα1_i = *Bos taurus* TICα1; BTα2_i = *Bos taurus* TICα2; CLα1_i = *Canis lupus* TICα1; CLα2_i = *Canis lupus* TICα2; DRα1_i = *Danio rerio* TICα1;
DRα2_i = *Danio rerio* TICα2; HSα1_i = *Homo sapiens* TICα1; HSα2_i = *Homo sapiens* TICα2; RNα1_i = *Rattus norvegicus* TICα1;
RNα2_i = *Rattus norvegicus* TICα2; i = one letter abbreviation of standard amino acids; TIC = type I collagen;
Siz = the dimension of the collagen type I substructures (number of amino acids) that autocorrelated;
Smi and Sma = number of amino acids present in the two substructures (one being higher than other);
Spr = number of simultaneously presence of amino acid of interest in both substructures (i.e. the same position);
r = correlation coefficient.

The number of simultaneously presence of amino acid of interest in the same position of both substructures on autocorrelation analysis varied from 2 to 6 (*Homo sapiens* α1 chain of type I collagen - proline and *Bos taurus* α1 chain of type I collagen – proline, respectively). None statistically significant value (r = 0.0810) were obtained by both autocorrelations with 6 simultaneously presence of proline in the same position on both α1 chain substructures. These autocorrelations are obtained on a similar dimension of 152 (*Bos Taurus*) and 176 (*Homo sapiens*) amino acids, respectively. This result suggests a similarity of these chains at the level of proline and of regularities among the α1 chains of these two species. Note that a good

similarity on α1 chains was previously identified between *Homo sapiens* and *Bos Taurus* (Bolboacă and Jäntschi, 2007).

A correlation coefficient, greater than 0.5 (it could be considered a significant correlation according with Colton's rules (Colton, 1974)), was obtained in six out of fifty-six cases (see Table 2). The higher correlation coefficient of 0.7300 was obtained in autocorrelation investigation of *Ratus norvegicus* α2 chain in investigation of glycine (the most frequent amino acids in the chain). The dimension of the substructures that autocorrelated was very small (7 amino acids) and on these substructures just a number of 2 glycine were simultaneously present in the same position on both substructures.

Moderate correlations (Colton, 1974) were obtained as follows:

÷ r = 0.5390: *Bos taurus* and *Homo sapines* α1 chain of type I collagen - leucine. The characteristics were identical in both cases: dimension of the type I collagen substructure that autocorrelated - 43 amino acids, with 5 leucine simultaneously in the same position.

÷ r = 0.5280: *Canis lupus* α1 chain of type I collagen - leucine (dimension of the substructures that autocorrelated of 39; 5 leucine simultaneously in the same position).

÷ r = 0.5190: *Bos taurus* and *Canis lupus* α1 chain of type I collagen - glutamate (dimension of the substructures that autocorrelated of 28; 2 glutamate simultaneously in the same position). Thus, the same pattern of glutamate regularity is observed in investigation of α1 chain on *Bos taurus* and *Canis lupus*.

Weak to acceptable degree of association (Colton, 1974) were obtained as follows:

÷ r = 0.4780: *Bos taurus*, *Canis lupus*, *Danio rerio*, and *Homo sapiens* α2 chain of type I collagen - leucine, and *Danio rerio* α1 chain of type I collagen 2 - leucine. The same pattern of regularities is observed on four out of five investigated species regarding the leucine on α2 chain of type I collagen.

÷ r = 0.362: *Canis lupus* α1 chain of type I collagen, 2 - tyrosine.

÷ r = 0.3200: *Canis lupus* α1 chain of type I collagen, 2 - proline.

÷ r = 0.2850: *Bos taurus*, *Canis lupus* and *Homo sapien*s α2 chain of type I collagen - proline. The same pattern of regularities is observed on three out of five investigated species regarding the proline distribution on α2 chain.

÷ r = 0.263: *Homo sapiens* α2 chain of type I collagen, 2 - alanine.

The results obtained in autocorrelation analysis of type I collagen lead to the followings remarks:

÷ The best as well as the weak to acceptable degree of autocorrelation were obtained on lower dimension of the type I collagen substructures. These results showed that the amino acid sequences on type I collagen chains have not a repeating patters.

÷ The presence of autocorrelation is not related with the distribution of amino acids in the type I collagen chains.

÷ Some similarly autocorrelation patterns were identified:
  - Leucine distribution on α1 chain (*Bos taurus* and *Homo sapines*) as well as on α2 chain (*Bos taurus*, *Canis lupus*, *Danio rerio*, and *Homo sapiens*).
  - Glutamate distribution on α1 chain: *Bos taurus* and *Canis lupus*.
  - Proline distribution on α2 chain: *Bos taurus*, *Canis lupus* and *Homo sapiens*.

CONCLUSIONS

The rank correlation analysis revealed the existence of a moderate to a very good correlation between ranks of standard amino acids position in the investigated type I collagen chains on all species.

The autocorrelation is not related with the frequency distribution of amino acids. Moreover, the amino acid sequences on type I collagen chains have not a repeating patters. The investigated ability of autocorrelation is applied just at the extremities of chains.

ACKNOWLEDGEMENTS

UEFISCSU Romania supported the research through project AT/93GR/07.06.2007.

BIBLIOGRAPHY

1. Bolboacă S, L Jäntschi, 2007, Amino Acids Sequences Analysis on Collagen, Bulletin of University of Agricultural Sciences and Veterinary Medicine - Animal Science and Biotechnologies, 63-64, 311-316.
2. Broersen PMT, 2006, Automatic Autocorrelation and Spectral Analysis, Springer, 1st edition.
3. Colton T, 1974, Statistics in Medicine. Little Brown and Company, New York, NY.
4. Cummings CL, D Gawlitta, RM Nerem, JP Stegemann, 2004, Properties of engineered vascular constructs made from collagen, fibrin, and collagen-fibrin mixtures, Biomaterials, 25(17), 3699-3706.
5. Dubois GM, Z Haftek, C Crozet, R Garrone, D Le Guellec, 2002, Structure and spatio temporal expression of the full length DNA complementary to RNA coding for alpha2 type I collagen of zebrafish, Gene, 294, 55-65.
6. Lowe JK, R Guyon, ML Cox, DC Mitchell, AL Lonkar, F Lingaas, C André, F Galibert, EA Ostrander, KE Murphy, 2003, Radiation hybrid mapping of the canine type I and type IV collagen gene subfamilies, Funct Integr Genomics, 3, 112-116.
7. Luo Y-F, S-J Wang, X-F Niu, Y-L Wang, 2008, Synthesis and characterization of collagen type I-modified poly(DL-lactic acid) biomimetic material, Gongneng Cailiao/Journal of Functional Materials, 39(1), 123-126.
8. Orgel JP, TC Irving, A Miller, TJ Wess, 2006, Microfibrillar structure of type I collagen in situ, Proc Natl Acad Sci USA, 103(24), 9001-9005.
9. Shirai T, S Hattori, M Sakaguchi, S Inouye, A Kimura, T Ebihara, S Irie, Y Nagai, H Hori, 1998, The complete cDNA coding sequence for the bovine proalpha2(I) chain of type I procollagen, Matrix Biol, 17(1), 85-88.
10. Strausberg RL, EA Feingold, LH Grouse, JG Derge, RD Klausner, FS Collins, L Wagner, CM Shenmen, GD Schuler, SF Altschul, B Zeeberg, KH Buetow, CF Schaefer, NK Bhat, RF Hopkins, H Jordan, T Moore, SI Max, J Wang, F Hsieh, L Diatchenko, K Marusina, AA Farmer, GM Rubin, L Hong, M Stapleton, MB Soares, MF Bonaldo, TL Casavant, TE Scheetz, MJ Brownstein, TB Usdin, S Toshiyuki, P Carninci, C Prange, SS Raha, NA Loquellano, GJ Peters, RD Abramson, SJ Mullahy, SA Bosak, PJ McEwan, KJ McKernan, JA Malek, PH Gunaratne, S Richards, KC Worley, SHale, AM Garcia, LJ Gay, SW Hulyk, DK Villalon, DM Muzny, EJ Sodergren, XLu, RA Gibbs, JFahey, E Helton, M Ketteman, M Madan, S Rodrigues, A Sanchez, M Whiting, A Madan, AC Young, Y Shevchenko, GG Bouffard, RW Blakesley, JW Touchman, ED Green, MC Dickson, AC Rodriguez, J Grimwood, J Schmutz, RM Myers, YSN Butterfield, MI Krzywinski, U Skalska, DE Smailus, A Schnerch, JE Schein, SJM Jones, MA Marra (Mammalian Gene Collection (MGC) Program Team), 2002, Generation and initial analysis of more than 15,000 full-lengthhuman and mouse cDNA sequences, Proc Natl Acad Sci USA, 99(26), 16899-16903.
11. Venien A, D Levieux, 2005, Differentiation of bovine from porcine gelatines using polyclonal anti-peptide antibodies in indirect and competitive indirect ELISA, J Pharmaceut Biomed, 39(3-4), 418-424.
12. Wardale RJ, VC Duance, 1993, Quantification and immunolocalisation of porcine articular and growth plate cartilage collagens, J Cell Sci, 105, 975-84.