Cyclicity Analysis of Amino-Acids on Type I Collagen Chains

Sorana D. BOLBOACĂ
and Lorentz JÄNTSCHI

Intro

- Type I collagen is the most common type of collagens in vertebrates.
- It is used in
  - gelatine industry
    - (Venien and Levieux, 2005)
  - biomaterials engineering
    - Cummings et al., 2004)
    - (Luo et al., 2008)

Aim

- 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*
  - *Rattus Norvegicus*

Material

- α 1 & α 2 chains of collagen type I (CTI)
- 5 species:
  - *Bos taurus* (Shirai *et al.*, 1998)
  - *Canis lupus* (Lowe *et al.*, 2003)
  - *Danio rerio* (Dubois *et al.*, 2002)
  - *Homo sapiens* (Strausberg *et al.*, 2002)
  - *Rattus norvegicus* (Orjel *at al*., 2006)
- Source:
  - NCBI database: http://www.ncbi.nlm.nih.gov/

Data

Ala (A)	Arg (R)	Asn (N)	Asp (D)	Cys (C)
Glu (E)	Gln (Q)	Gly (G)	His (H)	Ile (I)
Leu (L)	Lys (K)	Met (M)	Phe (F)	Pro (P)
Ser (S)	Thr (T)	Trp (W)	Tyr (Y)	Val (V)

- On α 1 type I collagen chain of *Rattus norvegicus* were assigned a number of 116 to unknown amino acids (out of 1054; abbreviated as X; out of 20 standard); on α 2 chain 102 were unknown amino acids (out of 1026).
- The most frequent amino acids in the type I collagen chains of investigated species (Bolboacă and Jäntschi, 2007) is glycine (Gly).

Methods: rank correlation analysis

N	390	389	382	382	381	381	380	345	344	329	279	278	236	235	235	231	230	223	162	143	138	137	130	126	126	125	123	116	115	113	108	102	8
0	C1g	B1g	D1g	D2g	C2g	H2g	B2g	R2g	R1g	H1g	B1p	C1p	B2p	C2p	D1p	H2p	H1p	D2p	D1a	B1a	C1a	D2a	H2a	R1p	B2a	R1a	C2a	R1x	H1a	R2p	R2a	R2x	D
1	22	22	22	25	33	33	33	6	5	22	36	32	34	34	48	34	36	29	10	14	14	14	14	16	14	33	14	28	14	9	4	21	2
2	26	26	24	28	36	36	36	7	12	26	37	33	40	40	50	40	37	32	15	15	15	20	20	18	20	51	20	31	15	11	23	27	4
3	40	30	30	31	39	39	39	8	17	30	55	51	46	46	75	46	55	38	20	17	17	54	35	21	28	73	28	34	17	17	26	36	6
4	62	44	31	34	42	42	42	10	20	44	57	78	47	47	77	47	57	41	80	84	53	67	78	24	30	90	30	43	84	20	48	39	6
5	79	66	37	37	45	45	45	13	23	66	82	83	49	49	86	49	82	45	114	126	80	68	79	30	78	97	78	46	126	29	50	42	8
6	85	98	59	40	48	48	48	16	26	83	87	84	50	50	89	50	87	48	138	160	96	73	104	36	79	114	79	49	160	35	60	54	9
7	94	109	100	43	51	51	51	19	29	89	93	89	59	59	96	58	93	51	139	194	122	97	105	42	83	118	104	61	195	47	83	57	9
8	105	112	103	46	54	54	54	22	32	98	96	92	61	61	101	61	96	53	160	212	156	100	107	54	102	121	107	64	213	53	113	66	1
9	108	115	106	49	57	57	57	25	35	109	102	98	62	62	105	62	102	56	173	234	191	107	117	57	105	135	117	79	235	56	122	72	1
10	111	118	109	52	60	60	60	28	38	112	113	109	64	64	113	64	113	57	195	251	209	116	129	60	127	153	129	85	252	69	134	87	1
11	114	121	112	55	63	63	63	31	41	115	119	115	65	65	116	65	119	59	197	258	231	118	131	63	129	157	131	94	259	78	146	90	1

- Step 1: matrix representation
  - Column: amino acid of interest on a specie
  - Rows: ranks of positions in chains
  - Elements: positions of amino acids in chains
- Step 2: ranks correlations
  - Spearman correlation coefficient

Methods: autocorrelation analysis

Pair»	Correlation
B1cH1c»	1.000
B1wH1w»	1.000
B1yH1y»	1.000
C2cH2c»	1.000
C2kH2k»	1.000
C2wH2w»	1.000
C2gH2g»	0.993
C2nH2n»	0.988
C2rH2r»	0.978
C2eH2e»	0.976
C2qH2q»	0.970
C2yH2y»	0.968
C2dH2d»	0.952
C2lH2l»	0.948
C2pH2p»	0.938
C2fH2f»	0.929
B2gR1g»	0.877
C2iH2i»	0.876
B1gR2g»	0.875
C2aH2a»	0.874
H2gR2g»	0.871
D1gR2g»	0.870
C2gR2g»	0.868
B1qD1q»	0.860

- 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,…, aa_n} with the data set {aa_1, aa_2, aa_3, aa_4,…, aa_{n-1}}
- programs were developed in PHP
- limitation of the analysis: possibility of obtaining a positive correlation by 0 (the absence of amino acid of interest)

Results: frequency apparition on correlations classes

Chain	Deleted	r_{min} (where)	r_{max} (where)	r<.5	r⩾.	r⩾.7	r⩾.9	r⩾.9
BT α 2	W	0.4905 (V-Y)	0.9987 (L-R)	2	151 5	140 5	77 5	9 9
BT α 1	C, W	0.6857 (H-L)	0.9989 (G-R)	0	136	133	103	23
CL α 2	W	0.5438 (C-Y)	0.9974 (K-R)	0	153	145	90	12
CL α 1	W	0.5255 (L-Y)	0.9988 (G-R)	0	153	138	105	17
DR α 2	W, Y	0.5959 (H-L)	0.9968 (E-R)	0	153	151	97	11
DR α 1	C, W	0.5852 (H-L)	0.9978 (G-P)	0	153	145	100	24
HS α 2	H, M, W, Y	0.7363 (C-L)	0.9983 (G-P)	0	120	119	72	9
HS α 1	C, W	0.5033 (M-Y)	0.9989 (G-R)	0	153	136	98	21
RN α 2	C, I, M, W, Y	0.8953 (A-N)	0.9993 (G-X)	0	105	105	95	25
RN α 1	C, M, W, Y	0.8709 (S-T)	0.9983 (G-P)	0	120	120	100	19

BT α 2 = *Bos taurus* α 1; BT α 2 = *Bos taurus* α 2; CL α 1 = *Canis lupus* α 1;
CL α 2 = *Canis lupus* α 2; DR α 1 = *Danio rerio* α 1; DR α 2 = *Danio rerio* α 2;
HS α 1 = *Homo sapiens* α 1; HS α 2 = *Homo sapiens* α 2; RN α 1 = *Rattus norvegicus* α 1;
RN α 2 = *Rattus norvegicus* α 2.

Discussion: frequency apparition on correlations classes

- The correlations coefficient varied from 0.2789 to 1
  - 0.2789: (DR α 1 L (37 leucine on the chain) - Y (9 tyrosine on the chain))
  - 1: (RN α 1 V (18 valine on the chain) - H (3 histidine on the chain) & HS α 1 S (35 serine on the chain) - Y (3 tyrosine on the chain)).
- A minimum value of 0.4905 is obtained when all amino acids with appearance less than 10 are deleted.
  - (BT α 1 V (42 valine on the chain) - Y (16 tyrosine on the chain))
- It is a similarity between distributions of glycine (the most frequent on all species) and arginine (top-three frequency).
- The max. of rank correlation coefficient is always obtained for apparition of amino acids with an frequency $\geqslant$ 50.
- The rank correlation between positions in the same chain
  - from moderate to good association ($0.5 < r < 0.75$)
  - to very good level of association ($r > 0.75$)

Results and discussion: autocorrelation analysis (1/4)

- 56/100 autocorrelations > 0
- α 1 *Canis lupus*:
  - Max. no. of aa (9/20), 45% $[5\text{-}14]^{95\%}$
- α 1 *Bos Taurus* & α 2 *Homo sapiens*:
  - (8/20), 40% $[4\text{-}13]^{95\%}$
- *Ratus norvegicus*
  - Lowest autoc > 0 just for α 1Ala & α 2Gly
  - Nota bene – a large amount of unknown/unexpected aa (116 on α 1, 102 on α 2)
  - Absence of 2/20 aa (Cys and Trp)

Results and discussion: autocorrelation analysis (2/4)

- Autocorrelated substructures sizes:
  - 7 (*Ratus norvegicus* α 2 Gly), r = 0.73
  - 1462 (*Bos Taurus* α 1 Leu), r = 0.012
  - α 1 vs. α 2 on same: from 2 to 18; max.r=0.012 *Danio rerio*
  - α 1 vs. α 1 on other: from 2 to 6 (*Homo sapiens* vs. *Bos taurus*) aa. None statistically significant value were obtained
- 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 α 1 chains of these two species. A good similarity on α 1 chains was previously identified between *Homo sapiens* & *Bos Taurus* (Bolboacă and Jäntschi, 2007).

Results and discussion: autocorrelation analysis (3/4)

- r ≥ .5: 6/66 cases
- Max. r: *Ratus norvegicus* α 2 Gly (0.73)
- r = 0.539: *Bos taurus* α 1 Leu & *Homo sapines* α 1 Leu (5/43 on same position)
- r = 0.528: *Canis lupus* α 1 Leu (5/39 on same position)
- r = 0.519: *Bos taurus* α 1 & *Canis lupus* α 1 on Glu (2/28 on same position)
- r = 0.478: *Bos taurus*, *Canis lupus*, *Danio rerio*, and *Homo sapiens* α 2 on Leu; *Danio rerio* α 1 on Leu

Results and discussion: autocorrelation analysis (4/4)

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.

Autocorrelation analysis 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 α 1 chain (*Bos taurus* and *Homo sapines*) as well as on α 2 chain (*Bos taurus*, *Canis lupus*, *Danio rerio*, and *Homo sapiens*).
  - Glutamate distribution on α 1 chain: *Bos taurus* and *Canis lupus*.
  - Proline distribution on α 2 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.
- Thank you for attention.