Applied Medical Informatics
Original Research Vol. 22, No. 1-2/2008, pp: 39 - 46

# *Homo Sapiens* Type I Collagen: Patterns Analysis

**Sorana D. BOLBOACĂ [1,*], Lorentz JÄNTSCHI [2]**

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

* Author to whom correspondence should be addressed. Tel.: +4-0264-431697; Fax: + +4-0264-593847.

**Abstract:** The aim of the study was to investigate patterns within and among α1 and α2 chains of *Homo Sapiens* type I collagen. The sequences of amino acids on alpha 1 and alpha 2 chains of type I collagen were taken from the National Center for Biotechnology Information. The CLC Proteine Workbench software was used in the investigation of amino acids patterns on type I collagen (within and among chains). Forty-seven patterns of 2 to 7 amino acids were identified on each chain. The patterns comprised the top-three frequent amino acids: glycine, proline and alanine on both chains. Seventeen out of forty-seven patterns (36%) were common on both chains. The patterns identified on the α2 chain comprise more amino acids compared with the patterns identified on α1 chain. The analysis among type I collagen chains identified patterns of 3 to up to 7 amino acids. Almost 13% of patterns identified on analysis among chains were seen on both chains. Future researcher will be necessary for studying the existence of the same patterns on type I collagen on different species as well as the usefulness of start and end position of the pattern in characterization of repetition and/or cyclicity of amino acids.

**Keywords:** Type I collagen; Homo sapiens; Pattern analysis.

## Introduction

The main protein of the extracellular matrix is collagen, the protein that comprise almost one quarter of all of the protein in the mammals' body [1] and which is continuously synthesized and depredated throughout the lifetime. More than 20 genetically different types of collagen have been identified [2,3].

Type I collagen is composed of two identical alpha 1 polypeptide chains (α1) and one alpha 2 polypeptides chain (α2). Two distinct genes are responsible by production of these polypeptide chains (COL1A1 [4], chromosome 17, and COL1A2 [5], chromosome 7), their expression being inactivated in adult tissues but stimulated after injury.

Type I collagen is involved in many human diseases (osteogenesis imperfecta [6], fibrosis [7,8], osteoporosis [9,10], melanoma [11], lung cancer [12], pancreatic cancer [13], atherosclerosis [14], etc). The degradation products of type I collagen molecule are used as diagnostic [15,16] as well as monitoring tools [17-19] in various pathological conditions due to its widely distribution in the body.

Gelatin, translucent, colourless, brittle, nearly tasteless solid substance extracted from collagen connective tissues of animals is used as an emulsifier in food (E441), pharmaceutical [20,21], photography and cosmetic manufacturing [22,23]. Moreover, the type I collagen is used for developments of: surface for cell interaction [24-26], cartilage biomaterials [27], implant surface microstructures [28], bone graft materials [29], urethral reconstruction biomaterial [30], etc.

The aim of the study was to investigate patterns within and among α1 and α2 chains of type I collagen of *Homo Sapiens*.

## Material and Method

*Amino Acids on Type I Collagen*

The sequences of amino acids on alpha 1 (α1) and alpha 2 (α2) chains of type I collagen (TIC) for *Homo Sapiens* were taken from the National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/] [31].

Twenty standard amino acids are in composition of type I collagen: 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). These amino acids follows a certain distribution, glycine being the most frequent one [32].

*Pattern Analysis*

The CLC Proteine Workbench software was used in the investigation of amino acids patterns on type I collagen. Both chains (α1 and α2) were investigated. Three steps applied in the analysis were as follows:

÷ Step 1: identification of most frequent patterns of 2, 3, 4, 5, 6, and 7 amino acids on each t chain. The maximum of seven amino acids in patters was imposed due to the time needed for searching.

÷ Step 2: identification and analysis of patterns when the range of 1 to up to 9 amino acids is imposed.

÷ Step 3: identification and analysis of most frequent patterns present simultaneously on both chains. In this analysis classes of pattern were search (e.g. class of pattern of 3 amino acids).

## Results

The general sequence information regarding the investigated chains of type I collagen on *Homo Sapiens* is presented in Table 1. The absolute frequency distribution of amino acids on the chains is graphically represented in Figure 1.

**Table 1.** Chains characteristics

| Characteristic | α1 chain | α2 chain |
|---|---|---|
| Sequence type | Protein | Protein |
| Length (amino acids) | 1069 | 1336 |
| N-terminal amino acid | Methionine | Methionine |
| Hydrophobic residue (A, F, G, I, L, M, P, V, W) | 764 | 927 |
| Hydrophilic residue (C, N, Q, S, T, Y) | 115 | 193 |
| Other residue | 190 | 246 |

Forty-seven patterns were identified in investigation of α1 chain (see Table 2), cumulating a total number of five-hundred twenty-seven repetitions. The patterns of four amino acids presented in Table 2 were identified when the range of 1 to 9 amino acids is imposed in searching.

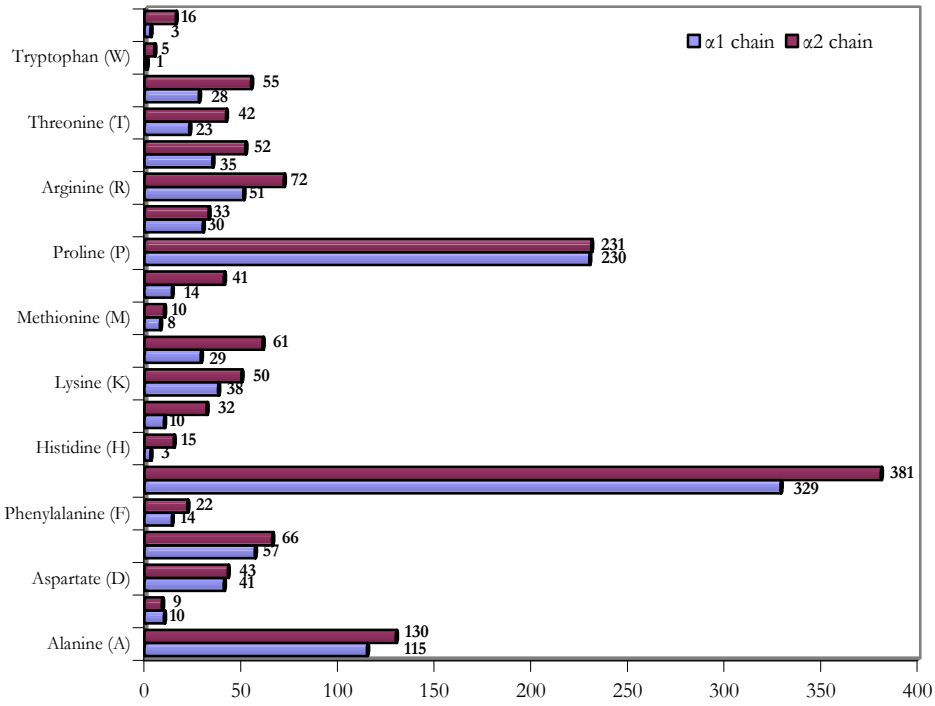


**Figure 1.** Amino acids distribution on type I collagen chains

**Table 2.** Patterns on α1 chain of type I collagen

| No | Pattern | $f_{pattern}$ | Length |
|---|---|---|---|
| 1 | AG | 31 | 2 |
| 2 | GP | 107 | 2 |
| 3 | PG | 70 | 2 |
| 4 | GAP | 19 | 3 |
| 5 | GEP | 11 | 3 |
| 6 | GLP | 11 | 3 |
| 7 | GPP | 38 | 3 |
| 8 | GSP | 9 | 3 |
| 9 | PAG | 20 | 3 |
| 10 | GAPG | 15 | 4 |
| 11 | GEAG | 3 | 4 |
| 12 | GEPG | 6 | 4 |
| 13 | GERG | 4 | 4 |
| 14 | GFPG | 4 | 4 |
| 15 | GLPG | 9 | 4 |
| 16 | GPAG | 8 | 4 |
| 17 | GPPG | 32 | 4 |
| 18 | GPRG | 5 | 4 |
| 19 | GRPG | 4 | 4 |
| 20 | GSPG | 9 | 4 |
| 21 | AGAPG | 4 | 5 |
| 22 | AGPKG | 2 | 5 |
| 23 | GAPGP | 3 | 5 |
| 24 | GLPGP | 4 | 5 |

| No | Pattern | $f_{pattern}$ | Length |
|---|---|---|---|
| 25 | GPAGP | 5 | 5 |
| 26 | GPPGA | 5 | 5 |
| 27 | GPPGE | 3 | 5 |
| 28 | GPPGP | 20 | 5 |
| 29 | PGAKG | 5 | 5 |
| 30 | AGPPGA | 2 | 6 |
| 31 | ANGAPG | 2 | 6 |
| 32 | APGAPG | 2 | 6 |
| 33 | LPGPPG | 2 | 6 |
| 34 | PAGPKG | 2 | 6 |
| 35 | PAGPPG | 5 | 6 |
| 36 | PPGPAG | 8 | 6 |
| 37 | PPGPPG | 6 | 6 |
| 38 | RGPPGP | 2 | 6 |
| 39 | GANGAPG | 2 | 7 |
| 40 | GAPGAPG | 2 | 7 |
| 41 | GERGFPG | 2 | 7 |
| 42 | GERGPPG | 2 | 7 |
| 43 | GLPGPPG | 3 | 7 |
| 44 | GPAGPKG | 2 | 7 |
| 45 | GPAGPPG | 4 | 7 |
| 46 | GPPGPAG | 8 | 7 |
| 47 | GPPGPPG | 5 | 7 |

$f_{pattern}$ = absolute frequency of pattern

Forty-seven pattern were identified in investigation of α2 chain (see Table 3), cumulating a total number of five-hundred and thirty repetitions. The same phenomena as was observed on α1 chain is also seen on α2 chain when the range of pattern searching was from 1 to 9 amino acids.

**Table 3.** Patterns on α2 chain type I collagen

| No | Pattern | $f_{pattern}$ | Length |
|---|---|---|---|
| 1 | GE | 35 | 2 |
| 2 | GP | 120 | 2 |
| 3 | PG | 69 | 2 |
| 4 | AGP | 24 | 3 |
| 5 | EPG | 9 | 3 |
| 6 | LPG | 12 | 3 |
| 7 | PGP | 31 | 3 |
| 8 | PPG | 3 | 3 |
| 9 | RGP | 13 | 3 |
| 10 | SGP | 9 | 3 |
| 11 | VGP | 12 | 3 |
| 12 | GAPG | 9 | 4 |
| 13 | GEPG | 9 | 4 |
| 14 | GLPG | 14 | 4 |
| 15 | GPAG | 25 | 4 |
| 16 | GPPG | 25 | 4 |
| 17 | GPQG | 4 | 4 |
| 18 | GPRG | 3 | 4 |
| 19 | GPSG | 3 | 4 |
| 20 | GPTG | 3 | 4 |
| 21 | GPVG | 4 | 4 |
| 22 | AGPPG | 3 | 5 |
| 23 | GAPGE | 2 | 5 |
| 24 | GAPGP | 6 | 5 |
| 25 | GPAGP | 10 | 5 |
| 26 | GPPGP | 12 | 5 |
| 27 | GPSGP | 4 | 5 |
| 28 | GPVGP | 2 | 5 |
| 29 | RGLPG | 5 | 5 |
| 30 | RGPPG | 3 | 5 |
| 31 | GPAGAR | 3 | 6 |
| 32 | PAGPPG | 2 | 6 |
| 33 | PAGPRG | 4 | 6 |
| 34 | PPGPPG | 7 | 6 |
| 35 | PPGPSG | 2 | 6 |
| 36 | PVGAAG | 3 | 6 |
| 37 | PVGPAG | 2 | 6 |
| 38 | GAPGPAG | 2 | 7 |
| 39 | GARGAPG | 2 | 7 |
| 40 | GNKGEPG | 2 | 7 |
| 41 | GPAGPAG | 2 | 7 |
| 42 | GPAGPPG | 3 | 7 |
| 43 | GPPGPAG | 2 | 7 |
| 44 | GPPGPPG | 4 | 7 |
| 45 | GPRGLPG | 2 | 7 |
| 46 | GPSGPAG | 2 | 7 |
| 47 | GPVGPAG | 3 | 7 |

$f_{pattern}$ = absolute frequency of pattern

The pattern investigation on both type I collagen chains revealed a number of one-hundred and eighteen distinct patterns of 3 to up to 7 amino acids. Almost 13% of the identified patterns were seen on both chains (see Table 4). The patterns identified among type I collagen chains is represented in Figure 2 and 3.

**Table 4.** Patterns among α1 and α2 chains: type I collagen

| No. | Pattern | Length | Frequency | |
|---|---|---|---|---|
| | | | α1 | α2 |
| 1 | DLRL | 4 | 1 | 1 |
| 2 | GAPGPAG | 7 | 1 | 2 |
| 3 | GAPGPQG | 7 | 1 | 1 |
| 4 | GEPGAPG | 7 | 1 | 1 |
| 5 | GFPGAAG | 7 | 1 | 1 |
| 6 | GPAGPPG | 7 | 4 | 3 |
| 7 | GPPGPAG | 7 | 10 | 2 |
| 8 | GPPGPPG | 7 | 3 | 4 |
| 9 | GPPGPQG | 7 | 1 | 1 |
| 10 | GPPGPSG | 7 | 1 | 2 |
| 11 | GPRGLPG | 7 | 1 | 2 |
| 12 | GPSGPAG | 7 | 1 | 2 |
| 13 | GPSGPQG | 7 | 1 | 1 |
| 14 | GPVGPAG | 7 | 1 | 3 |
| 15 | GSAGPPG | 7 | 1 | 1 |

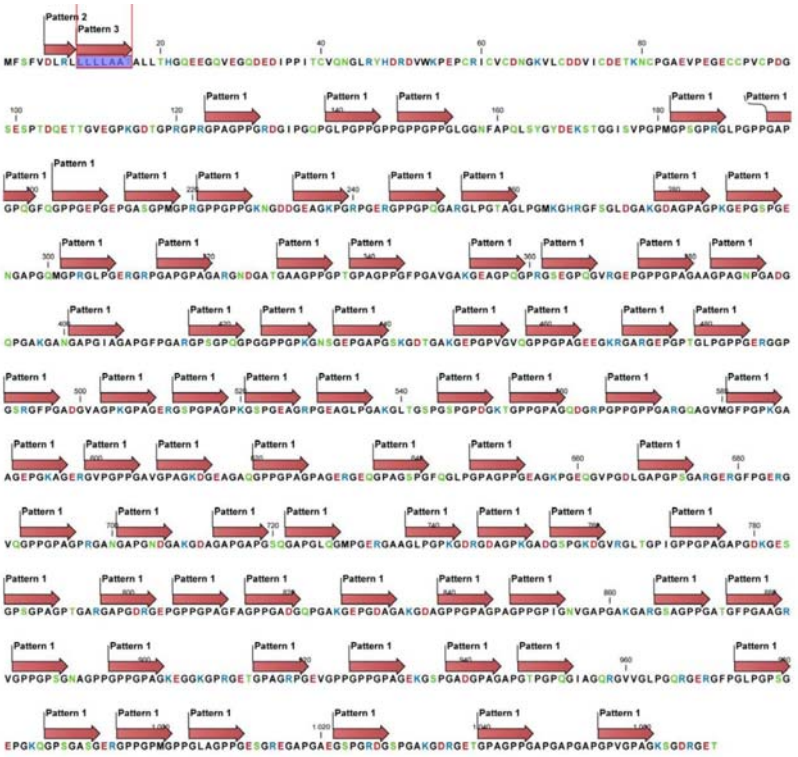

**Figure 2.** Patterns of amino acids on α1 chain

**Figure 3.** Patterns of amino acids on α2 chain

## Discussion

The present study was intended to be the first step in an approach to analyze the distribution of amino acids within and among type I collagen chains. The aim of the research was accomplished, the most frequent patterns of amino acids on length of 2 to 7 were identified and analyzed within and among chains of type I collagen. Note that, the similarity between α1 and α2 chains is of 40.49% [32] even if a similar distribution of amino acids in the chains is present (see Figure 1).

The number of distinct patterns was the same on both chains (forty-seven, see Table 2 and 3), even if the length of α2 chain contains more amino acids (with 267) compared α1 chain.

The top three most frequent amino acids on both chains are glycine, proline, and alanine as it can be observed from the graphical representation presented in Figure 1. Thus, it was expected to see these three amino acids on the patterns. With few exceptions (pattern no. 1 - α1 chain (Table 2), pattern no. 1 - α3 chain (Table 3)), the sequences patterns contains glycine and proline. The presence of all three amino acids in the same pattern varied from 38% (Table 3) to 53% (Table 4).

Seventeen out of forty-seven patterns (36%) are common on both chains: GP, PG, GAPG, GEPG, GLPG, GPAG, GPPG, GPRG, GAPGP, GPAGP, GPPGP, PPGPAG, PPGPPG, GERGPPG, GPAGPPG, GPPGPAG, and GPPGPPG. The most frequent 2 amino acids pattern on both chains was GP (107 apparitions on α1 and 120 apparitions on α2 chain), followed by the mirror pattern PG (70 apparitions on α1 and 69 apparitions on α2 chain).

Glycine and proline amino acids are found into the composition of the most frequent 3 amino acids pattern, but their order is different on the α1 (GPP) and α2 (PGP) chains. The two most frequent amino acids had also been found on the most frequent 4 amino acid pattern (GPPG - 32 apparitions on α1 chain and 25 apparitions on α2 chain).

The most frequent 5 amino acids pattern on the investigated chains was GPPGP (20 apparitions on α1 and 12 apparitions on α2 chain). A combination of glycine, proline, and alanine is also found on the most frequent pattern with 7 amino acids (GPPGPAG with 10 apparitions on α1 chain and 2 apparitions on α2 chain).

Beside the top three most frequent amino acids, the patterns identified on both chains also contain: glutamate (E), lysine (K), leucine (L), arginine (R), and serine (S). The similar distribution at the level of proportion of these amino acids could explain the presence of them into the patterns of both chains. Even if similar proportions of phenylalanine (F) are seen on both chains, this amino acid appears twice just on the α1 chain. Valine (V), threonine (T) and glutamine (Q) are found just in the patterns of α2 chain. The relative differences of the distribution of V, T, and Q varied from 0.33% to 1.50 %. This lead to the conclusion that the similar distribution of amino acids in the type I collagen is not related with the apparition of amino acids into the patterns.

A series of questions arises from the pattern investigation on type I collagen: "Is any explanation of apparition on patterns of certain amino acid?", "The start and end position of pattern could be a useful variable in characterization of repetition and/or cyclicity of identified patterns?", "If the α1 and/or α2 type I collagen chains of other species are analyzed, will the patterns be the same?". These will require future research.

## Conclusions

A similar number of patterns with 2 to up to 7 amino acids were identified on both type I collagen chains. The main amino acids on these patterns were represented by the top three ones: glycine, proline and alanine.

The most frequent pattern on both chains was one with 2 amino acids (GP). The patterns identified on the α2 chain comprise more amino acids compared with the patterns identified on α1 chain.

Future researcher will be necessary for studying the existence of the same patterns on type I collagen on different species as well as the usefulness of start and end position of the pattern in characterization of repetition and/or cyclicity of amino acids.

**Acknowledgements**

CNCSIS Romania supported the research thought AT/GR93/2007 grant.

**References**

1. Fraztzl P. Collagen: Structure and Mechanisc, an Introduction. In: Fratzl P (Editor), Collagen: Structure and Mechanics. 1st ed. Springer; 2008;1-10.
2. Boot-Handford RP, Tuckwell DS, Plumb DA, Rock CF, Poulsom R. A novel and highly conserved collagen (pro(alpha)1(XXVII)) with a unique expression pattern and unusual molecular characteristics establishes a new clade within the vertebrate fibrillar collagen family. J Biol Chem 2003;278:31067-31077.
3. Veit G, Kobbe B, Keene DR, Paulsson M, Koch M, Wagener R. Collagen XXVIII, a novel von Willebrand factor A domain-containing protein with many imperfections in the collagenous domain. J Biol Chem 2006;281(6):3494-504.
4. Chu ML, de Wet W, Bernard M, Ramirez F. Fine structural analysis of the human pro-alpha 1 (I) collagen gene. Promoter structure, AluI repeats, and polymorphic transcripts. J Biol Chem 1985;260(4):2315-2320.
5. Runyan CE, Schnaper HW, Poncelet AC. Smad3 and PKCdelta mediate TGF-beta1-induced collagen I expression in human mesangial cells. Am J Physiol Renal Physiol 2003;285(3):F413-F422.
6. Pope FM, Nicholls AC, McPheat J, Talmud P, Owen R. Collagen genes and proteins in osteogenesis imperfecta. J Med Genet 1985;22:466-478.
7. Hibino I, Okita M, Inoue T, Banno Y, Hoso M. Effect of immobilization on insoluble collagen concentration and type I and type III collagen isoforms of rat soleus muscle. Journal of the Japanese Physical Therapy Association 2008;11(1):1-6.
8. Barisic-Dujmovic T, Boban I, Clark SH. Regulation of collagen gene expression in the Tsk2 mouse. Journal of Cellular Physiology 2008;215(2):464-471.
9. Oxlund H, Mosekilde Li, Ørtoft G. Reduced concentration of collagen reducible cross links in human trabecular bone with respect to age and osteoporosis. Bone 1996; 19(5):479-484.
10. Wang X, Shen X, Li X. Mauli Agrawal C. Age-related changes in the collagen network and toughness of bone. Bone 2002;31(1):1-7.
11. Van Kempen LCLT, Rijntjes J, Mamor-Cornelissen I, Vincent-Naulleau S, Gerritsen MJP, Ruiter DJ, Van Dijk MCRF et al. Type I collagen expression contributes to angiogenesis and the development of deeply invasive cutaneous melanoma. Int J Cancer 2008;122(5):1019-1029.
12. Shintani Y, Maeda M, Chaika N, Johnson KR, Wheelock MJ. Collagen I promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming growth factor-β signaling. Am J Respir Cell Mol Biol 2008;38(1):95-104.
13. Grzesiak JJ, Ho JC, Moossa AR, Bouvet M. The integrin-extracellular matrix axis in pancreatic cancer. Pancreas 2007;35(4):293-301.
14. Hollander NC, Mulder DJ, Graaff R, Thorpe SR, Baynes JW, Smit GPA, Smit AJ. Advanced glycation end products and the absence of premature atherosclerosis in glycogen storage disease Ia. J Inherit Metab Dis 2007;30(6):916-923.
15. Villar J, Arenas MI, MacCarthy CM, Blánquez MJ, Tirado OM, Notario V. PCPH/ENTPD5 expression enhances the invasiveness of human prostate cancer cells by a protein kinase Cδ-dependent mechanism. Cancer Res 2007;67(22):10859-10868.
16. Van den Abbeele A, De Corte V, Van Impe K, Bruyneel E, Boucherie C, Bracke M, Vandekerckhove J, Gettemans J. Downregulation of gelsolin family proteins counteracts cancer cell invasion in vitro. Cancer Lett 2007;255(1):57-70.
17. Wang J, Pei F, Tu C, Zhang H, Qiu X. Serum bone turnover markers in patients with primary bone tumors. Oncology 2008;72(5-6):338-342.
18. Brown JE, McCloskey EV, Dewar JA, Body JJ, Cameron DA, Harnett AN, Ruutu M, Coleman RE. The use of bone markers in a 6-week study to assess the efficacy of oral clodronate in patients with metastatic bone disease. Calcif Tissue Int 2007;81(5):341-351.

19. Lipton A, Cook RJ, Coleman RE, Smith MR, Major P, Terpos E, Berenson JR. Clinical utility of biochemical markers of bone metabolism for improving the management of patients with advanced multiple myeloma. Clin Lymphoma Myeloma 2007;7(5):346-353.
20. Wong H-L, Wang M-X, Cheung P-T, Yao K-M, Chan BP. A 3D collagen microsphere culture system for GDNF-secreting HEK293 cells with enhanced protein productivity. Biomaterials 2007;28(35):5369-5380.
21. Isales CM, McDonald JM. Future developments in therapy. Ann N Y Acad Sci 2007;1117:258-263.
22. Hsu TSJ, Zelickson B, Dover JS, Kilmer S, Burns J, Hruza G, Brown DB, Bernstein EF. Multicenter study of the safety and efficacy of a 585 nm pulsed-dye laser for the nonablative treatment of facial rhytides. Dermatol Surg 2005;31(1):1-9.
23. Kist D, Flor M, Zelickson B. Vibradermabrasion - New technique for superficial exfoliation. Cosmet Dermatol 2005;18(2):131-135.
24. Bilic G, Hall H, Bittermann AG, Zammeretti P, Burkhart T, Ochsenbein-Kolble N, Zimmermann R. Human preterm amnion cells cultured in 3-dimensional collagen I and fibrin matrices for tissue engineering purposes. Am J Obstet Gynecol 2005;193(5):1724-1732.
25. Porjazoska A, Yilmaz OK, Baysal K, Cvetkovska M, Sirvanci S, Ercan F, Baysal BM. Synthesis and characterization of poly(ethylene glycol)-poly(D,L-lactide- co-glycolide) poly(ethylene glycol) tri-block co-polymers modified with collagen: A model surface suitable for cell interaction. J Biomater Sci Polym Ed 2006;17(3):323-340.
26. Berthod F, Germain L, Li H, Xu W, Damour O, Auger FA. Collagen fibril network and elastic system remodeling in a reconstructed skin transplanted on nude mice. Matrix Biol 2001;20(7):463-473.
27. Chajra H, Rousseau CF, Cortial D, Ronzière MC, Herbage D, Mallein-Gerin F, Freyria AM. Collagen-based biomaterials and cartilage engineering. Application to osteochondral defects. Biomed Mater Eng 2008;18(SUPPL.1):S33-S45.
28. Jia MZ, Tsuru K, Hayakawa S, Osaka A. Modification of Ti implant surface for cell proliferation and cell alignment. J Biomed Mater Res A 2008;84(4):988-993.
29. Bernhardt A, Lode A, Boxberger S, Pompe W, Gelinsky M. Mineralised collagen - An artificial, extracellular bone matrix - Improves osteogenic differentiation of bone marrow stromal cells. J Mater Sci Mater Med 2008;19(1):269-275.
30. Kanatani I, Kanematsu A, Inatsugu Y, Imamura M, Negoro H, Ito N, Yamamoto S, Tabata Y, Ikada Y, Ogawa O. Fabrication of an optimal urethral graft using collagen-sponge tubes reinforced with copoly(L-lactide/ε-caprolactone) fabric. Tissue Eng 2007;13(12):2933-2940.
31. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM et al. Generation and initial analysis of more than 15,000 full-lengthhuman and mouse cDNA sequences. Proc Natl Acad Sci USA 2002;99(26):16899-16903.
32. Bolboacă S, Jäntschi L. Amino Acids Sequences Analysis on Collagen, Bulletin of University of Agricultural Sciences and Veterinary Medicine - Animal Science and Biotechnologies 2007;63-64:311-316.