Chem Biol Drug Des 2008; **71:** 173–179

Research Letter

doi: 10.1111/j.1747-0285.2007.00619.x

A Structural Informatics Study on Collagen

Sorana D. Bolboaca[1,*] and Lorentz Jäntschi[2]

[1]Department of Medical Informatics and Biostatistics, 'Iuliu Hatieganu' University of Medicine and Pharmacy Cluj-Napoca, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania
[2]Technical University of Cluj-Napoca, 103–105 Muncii Bvd, 400641 Cluj-Napoca, Romania
**Corresponding author: Sorana D. Bolboaca, sbolboaca@umfcluj.ro*

The study integrates knowledge resulting from structure–activity relationships analysis of amino acids with respect to the characterization of α1 and α2 type I collagen chains. Specifically, 15 amino acids and 14 properties were investigated and their structure–activity relationship models were obtained. The models were integrated into a web application and were used to predict the properties of a set of six amino acids. The similarities in α1 and α2 type I collagen chains has been investigated starting from the observed and predicted properties of amino acids by using two-step cluster analysis.

Key words: amino acids, structure–activity relationships, type I collagen

Received 17 July 2007, revised 21 November 2007 and accepted for publication 8 December 2007

Molecular biology has made significant progress in the analysis and characterization of the essential roles of amino acids, and this has led to development of therapeutically useful biopharmaceutical agents (1,2).

The relationships between biophysical properties and the amino acid structure have been investigated for more than three decades (3). Quantitative structure–activity relationship (SAR) analysis of amino acids in this regard remains a very important topic for many researchers (4–6). Mathematical investigations, from a topological perspective, have provided important contributions in the field (7,8).

Collagen, the main protein of connective tissues in animals, and the most abundant in mammals, is found within connective tissues from heart, vessels, skin, cornea, cartilage, ligaments, tendons, bone, and teeth. Twenty-eight types of collagens are known to date (9). The structural arrangement of type I collagen was an early focus of researchers (10) with its structure being more recently determined (11). Structural type I collagen characterization has reached a new era to now include x-ray crystallography and three-dimensional (3D) mapping (12–14). The importance of such collagen research is illustrated by its implication in numerous diseases such as osteogenesis imperfecta[a], osteoporosis[b], Ehlers-Danlos syndrome (15), Caffey disease (16), and bone metastasis diagnosis (17).

The main objective of the present study was to investigate similarities in α1 and α2 type I collagen chains relative to a molecular descriptors family (MDF) on SAR investigations on amino acids and by using the amino acids observed and the predicted properties.

Methods and Materials

The properties of interest of a set of 21 amino acids were first investigated by using MDF SAR approach (18). Fifteen amino acids were used to generate the MDF SAR models and the properties of a set of six amino acids were predicted based on the obtained equations. In the second step of the analysis, using the properties observed and predicted by the models, the similarities in α1 and α2 type I collagen chains were investigated. In addition, one of the properties of amino acids [the hydrophobicity on the Hessa *et al.* scale – Hyd(19)] was investigated using two multivariate analysis methods.

Molecular modeling

Twenty-one amino acids were investigated: alanine (Ala), arginine (Arg), asparagine (Asn), aspartate (Asp), cysteine (Cys), glutamine (Gln), glutamate (Glu), glycine (Gly), histidine (His), hydroxyproline (Hyp), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), proline (Pro), serine (Ser), threonine (Thr), tryptophan (Trp), tyrosine (Tyr), and valine (Val).

Fifteen compounds (including Ala, Asn, Asp, Cys, Gln, Glu, Gly, Ile, Leu, Lys, Met, Phe, Ser, Thr, and Val) were used to identify the relationship between amino acids and their 14 properties. Six properties and five quantum mechanics calculated parameters, which confer properties to the protein molecule as a whole, were investigated. The values of the investigated parameters were taken from previously reported studies (note that the reference is given following the parameter abbreviation) as appropriate. HYPERCHEM software was used to calculate the quantum mechanic parameters based on the 3D structure of the amino acids (noting that the use of HYPERCHEM software is indicated by a superscript letter[a] following the abbreviated parameter).

The amino acid parameters investigated are listed below:

- *Dipole moment*: The measure of polarity in a molecule [abbreviated as DM(20), unit of measurement (Debye)].

- *Molar refraction*: The measure of the volume occupied by an atom or group being dependent on temperature, index of refraction, and pressure [abbreviated as MR(20), unit of measurement (cm^3/mol)].

- *Molar Magnetic Susceptibility*: The degree of magnetization in response to an applied magnetic field [abbreviated as CHI(20), unit of measurement (m^3/mol)]. Positive values revealed its paramagnetic property; alternatively, a negative value revealed a diamagnetic property.

- *Solubility* (*in water*): The chemical property referring to the ability to dissolve in a solvent [abbreviated as Slb(20), unit of measurement (M)].

- *Hydrophobicity*: The physical property of a molecule that is repelled from a mass of water [abbreviated as Hyd(20), Hyd(19), Hyd(21), unit of measurement (dimensionless)]. Three different scales were included into analysis; several scales being known to date (22). A consensus hydrophobicity scale has not been identified yet (23). Two of the three scales included into analysis were chosen as scales with extreme values while one had middle values.

- *Logarithm of the activity coefficient*: The factor used to account for deviations from the ideal behavior in a mixture of chemical substances [abbreviated as Lac(20), unit of measurement (dimensionless)].

- *Partition coefficient for n-octanol/water in logarithmic scale*: The ratio of the molar concentrations of a chemical in n-octanol and water, in dilute solution [abbreviated as log P(20) and log P[c], unit of measurement (dimensionless)]. Two different values were used because of the differences between values.

- *Hückel energy*: The determination of pi electron energies on molecular orbitals [abbreviated as EHu[c], unit of measurement (kcal/mol)].

- *Hydration energy*: The reaction enthalpy for the dissolution of a compound into aqueous solution for peptides and proteins [abbreviated as HyE[c], unit of measurement (kcal/mol)].

- *Molar refractivity*: The measure of the volume occupied by an atom or group of atoms [abbreviated as Ref[c], unit of measurement ($Å^3$)].

- *Polarizability*: The ease of distortion of the electron cloud of a molecular entity by an electric field [abbreviated as Pol[c], unit of measurement ($Å^3$)].

The above properties were modeled by using the MDF on the SAR method (18). This approach has shown success in its estimation and prediction abilities for a series of biological active compounds (24). A detailed presentation of the MDF SAR approach can be found in Ref. (18).

Based on the information extracted strictly from the structure of the 15 amino acids investigated, a set of descriptors were generated for each property. The best performing monovariate models have been identified for each property. Starting from these models, the properties of a set of six amino acids (i.e. Arg, His, Hyp, Pro, Trp, and Tyr), which were not included into the generation of descriptors, were then predicted[d] and the values were used in the analysis of type I collagen similarities.

Molecular modeling – multivariate analysis

Multivariate analysis techniques were applied on molecular descriptors obtained on the 15 amino acids considering the hydrophobicity calculated on Hessa *et al.* scale [abbreviated as Hyd(19)]. Multiple linear regression and factor analysis techniques were applied on a set of 200 descriptors by using SPSS 12.0 software. The factor analysis technique was used to reveal simpler patterns within the 200 molecular descriptors to discover whether the amino acid hydrophobicity on Hessa *et al.* scale [abbreviated as Hyd(19)] could be explained in terms of one or more than one factor.

Cluster analysis on type I collagen

The Rattus Norvegicus type I collagen (25) was investigated. Starting from the amino acid sequence of $\alpha1$ and $\alpha2$ chains, 14 properties were added to each amino acid and the obtained data were investigated. The two-step cluster analysis technique was used (SPSS 12.0 software) as it included a specific feature of automatic selection of the best number of clusters as well as an ability to create cluster models simultaneously based on categorical and continuous variables.

Results

Structure–activity relationships for amino acids

Fourteen properties were investigated relative to a sample of 15 amino acids. The SAR models with one descriptor, the name of the descriptor, the dominant atomic property, the type of interaction, the model interaction, and the structure on the activity scale are presented in Table 1.

Multivariate analysis on hydrophobicity

The multivariate analysis techniques were applied on hydrophobicity on the Hessa *et al.* scale [abbreviated as Hyd(19)] starting from a number of 200 molecular descriptors generated with respect to structural information. A significant model with two descriptors was identified in the multivariate linear regression analysis. The characteristics of the model are presented in Table 2. The 3D graphical representation of the model is presented in Figure 1.

The factor analysis technique identified a number of four factors based on 200 molecular descriptors included in the analysis. The eigenvalues were as follows: 227 (Factor 1, 91% total variance), 6.9 (Factor 2, 2.7% total variance), 5.7 (Factor 3, 2.3% total variance), and 3.5 (Factor 4, 1.4% total variance). Upon further investigation,

Table 1: SAR models for amino acids

Amino acid property	Hyd(20)	DM(20)	Slb(20)	Log P^c
MDF SAR equation	$\hat{Y} = -160X - 0.065$	$\hat{Y} = -8.7X - 0.19$	$\hat{Y} = -25X + 4$	$\hat{Y} = -1.4X - 0.87$
SAR determination (%)	65	79	87	90
MDF descriptor (X)	AbmrEQg	IiDRLQt	IiDRLQt	IGDdKQg
Dominant atomic property	Charge (Q)	Charge (Q)	Charge (Q)	Charge (Q)
Interaction via	Space (geometry)	Bonds (topology)	Bonds (topology)	Space (geometry)
Interaction model	Qd^2	$\bar{Q}d$	$\bar{Q}d$	Q^2d
Structure on activity scale	Proportional	Proportional	Proportional	Logarithmic
Amino acid property	Hyd(19)	CHI(20)	Log P(20)	Lac(20)
MDF SAR equation	$\hat{Y} = 8.5X - 0.58$	$\hat{Y} = -93X + 84$	$\hat{Y} = -4.9X + 6.4$	$\hat{Y} = -45X + 18$
SAR determination (%)	90.5	91	93	93
MDF descriptor (X)	iMDRoQg	iHMRqQg	IHMrqQg	iGMmLQt
Dominant atomic property	Charge (Q)	Charge (Q)	Charge (Q)	Charge (Q)
Interaction via	Space (geometry)	Space (geometry)	Space (geometry)	Bonds (topology)
Interaction model	Q^{-1}	$\bar{Q}^{-1}$	$\bar{Q}^{-1}$	$\bar{Q} \cdot d$
Structure on activity scale	Inversed	Inversed	Logarithmic	Inversed
Amino acid property	HyE^c	Hyd(21)	Ref^c	Pol^c
MDF SAR equation	$\hat{Y} = 18X - 19$	$\hat{Y} = -21X + 12$	$\hat{Y} = 94X - 13$	$\hat{Y} = 37X - 4.8$
SAR determination (%)	93	95	97	98
MDF descriptor (X)	iGPmLQt	IGDROQg	iIMdWEg	iIMdWEg
Dominant atomic property	Charge (Q)	Charge (Q)	Electronegativity (E)	Electronegativity (E)
Interaction via	Bonds (topology)	Space (geometry)	Space (geometry)	Space (geometry)
Interaction model	$\bar{Q}d$	Q	Q^2d^{-1}	Q^2d^{-1}
Structure on activity scale	Inversed	Proportional	Inversed	Inversed
Amino acid property	MR(20)	EHu^c		
MDF SAR equation	$\hat{Y} = -0.89X + 6.7$	$\hat{Y} = 87X - 1400$		
SAR determination (%)	98	99.7		
MDF descriptor (X)	IFMMwQg	IfPdkEg		
Dominant atomic property	Charge (Q)	Electronegativity (E)		
Interaction via	Space (geometry)	Space (geometry)		
Interaction model	Q^2d^{-1}	Q^2d^{-1}		
Structure on activity scale	Logarithmic	Logarithmic		

Hydrophobicity: Hyd(20) – Bumble (1999), Hyd(19) – Hessa *et al.* (2005), Hyd(21) – Kyte and Doolittle (1982); dipole moment: DM(20) – Bumble (1999); solubility: Slb(20) – Bumble (1999); logarithm of the partition coefficient: log P(20) – Bumble (1999), log P^c – HYPERCHEM^c; magnetic susceptibility: CHI(20) – Bumble (1999); log activity coefficient: Lac(20) – Bumble (1999); hydration energy: HyE^c – HYPERCHEM^c; refractivity: Ref^c – HYPERCHEM^c; polarizability: Pol^c – HYPERCHEM^c; molar refraction: MR(20) – Bumble (1999); Hückel energy: EHu^c – HYPERCHEM^c; $\hat{Y}$: property estimated by the MDF model; SAR, structure–activity relationship; MDF, molecular descriptors family.

Table 2: Multivariate MDF SAR for hydrophobicity on the Hessa *et al.* scale

Amino acid property	Hydrophobicity – Hessa *et al.* [abbreviated as Hyd(19)]
MDF SAR equation	$0.08X_1 + 6.03X_2 - 1.36$
SAR determination (%)	95.8
MDF descriptor (X_i)	ISPDwQg (i = 1), iMDRoQg (i = 2)
Dominant atomic property	Charge (Q)
Interaction via	Space (geometry)
F (significance p-value)	124 (0.002)
Standard error of the estimate	0.31
Sample size	15

the stepwise linear regression analysis was applied on the obtained factors. Two models proved to be statistically significant.

- Model with one variable ($F = 96$, $p < 0.001$): Factor 1 ($r^2 = 0.88$, $r^2_{adj} = 0.87$, $s = 0.49$) and
- model with two variables ($F = 69$, $p < 0.001$): Factors 1 and 4 ($r^2 = 0.92$, $r^2_{adj} = 0.91$, $s = 0.42$),

where r^2 = squared correlation coefficient, r^2_{adj} = adjusted squared correlation coefficient, s = standard error of the estimate.

Cluster analysis on type I collagen

The two-step cluster analysis technique was applied to identify similarities in α1 and α2 type I collagen chains. Three clusters were identified for each chain.

- α1 chain: cluster 1 (Ala, Gly, and Hyp), cluster 2 (Ile, Leu, Met, Phe, and Pro), and cluster 3 (Arg, Asn, Asp, Gln, Glu, His, Lys, Ser, Thr, and Tyr).
- α2 chain: cluster 1 (Ile, Leu, Met, Phe, Pro, and Val), cluster 2 (Arg, Asn, Asp, Gln, Glu, His, Hyp, Lys, Ser, Thr, and Tyr), and cluster 3 (Ala and Gly).

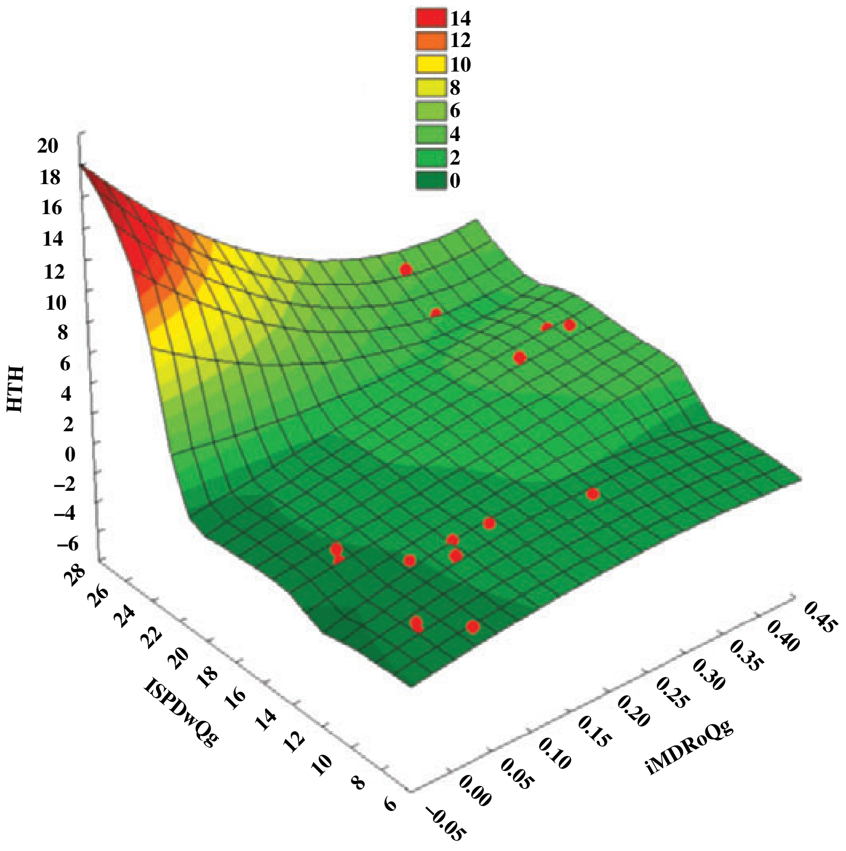

Figure 1: Quadratic surface plot: iMDRoQg (*X*-axis) versus ISPDwQg (*Y*-axis) versus HTH (*Z*-axis) hydrophobicity (*Z* = distance weighted least squares).

The parameters of descriptive statistics for each property according to chain and cluster are presented in Table 3. With one exception, the clustering of the amino acids (as abbreviation and properties) was statistically significant. For the α1 chain, the molecular refraction [abbreviated as MR(20)] had a significant statistical importance in the first and third cluster. In the α2 chain, the property abbreviated as log P(20) had statistical importance in the first and second cluster.

Discussion

Fourteen amino acid properties (11 distinct ones) were investigated by using the MDF on the SAR approach. A linear regression model with one variable was obtained for each property. The lowest performance in terms of squared correlation coefficient was obtained using an investigation of hydrophobicity calculated on the Bumble scale [abbreviated as Hyd(20), see Table 1]. The values published by Bumble [abbreviated as Hyd(20)] could be considered biased data based on the observation of the other two MDF SAR models obtained for the hydrophobicity published by Hessa *et al.* [abbreviated as Hyd(19)] and Kyte and Doolittle [abbreviated as Hyd(21)]. In these two models the squared correlation coefficients were >0.90. The ability of the MDF SAR approach in the investigation of the amino acids hydrophobicity will be comparatively analyzed with other hydrophobicity scales (26), in future studies. Except for the hydrophobicity calculated on the Bumble scale [abbreviated as Hyd(20), see Table 1], the MDF SAR models with one variable provided good performances for each investigated amino acid property. The ability of the model to determine the investigated properties varied from 79% [for the property abbreviated as DM(20)] to 99.7% (for the property abbreviated as EHu^c). Two different properties represented by dipole moment [abbreviated as DM(20)] and solubility [abbreviated as Slb(20)] were estimated by using the same molecular descriptor (IiDRLQt). These properties revealed to be related to amino acid topology and atomic partial charges. The MDF SAR models for refractivity (abbreviated as Ref^c) and polarizability (abbreviated as Pol^c) also use the same descriptor (iIMdWEg), showing that these properties are related to the geometry of compounds. Electronegativity proved to be the dominant atomic property according to SAR models. Ten of 14 properties were related to the geometry of amino acids: Hyd(20), Hyd(19), Hyd(21), log P^c, log P(20), CHI(20), Ref^c, Pol^c, MR(20), and EHu^c. The charges were the dominant atomic property for 11 of the 14 properties: Hyd(20), Hyd(19), Hyd(21), DM(20), Slb(20), CHI(20), log P(20), log P^c, Lac(20), HyE^c, and MR(20).

The multivariate analysis on hydrophobicity calculated on the Hessa *et al.* scale [abbreviated as Hyd(19)] has provided important information. As expected, the multivariate model provided a better estimation compared with the model with one descriptor (an SAR determination of 95.8% versus 90.5% as found in Table 2 and Table 1, respectively). As shown in Tables 1 and 2 that the MDF SAR model with two variables used the molecular descriptors identified

Table 3: Two-step cluster analysis on type I collagen: results

	α1 chain			α2 chain		
	Cluster			Cluster		
	1 (*n* = 581)	2 (*n* = 198)	3 (*n* = 275)	1 (*n* = 224)	2 (*n* = 340)	3 (*n* = 462)
DM(20)						
m	1.75	0.60	3.52	0.79	3.18	1.71
SD	0.05	0.72	1.50	0.71	1.32	0.25
Log P(20)						
m	3.32	1.70	4.35	1.71	4.19	3.24
SD	0.22	0.34	0.63	0.22	0.67	0.51
Log P^c						
m	−0.94	0.10	−1.11	0.21	−1.12	−0.92
SD	0.22	0.43	0.39	0.43	0.38	0.26
Lac(20)						
m	7.82	7.43	13.16	7.28	13.02	6.53
SD	2.39	1.33	3.04	0.89	1.66	0.92
Hyd(20)						
m	−2.77	−0.91	−6.76	−0.89	−5.42	−2.80
SD	0.21	0.69	2.40	0.67	2.93	0.42
Hyd(19)						
m	0.47	−0.14	2.40	−0.21	1.73	0.58
SD	0.33	0.21	1.04	0.24	1.37	0.28
Hyd(21)						
m	0.36	3.90	−3.82	4.03	−2.53	0.12
SD	0.95	0.80	2.36	0.49	3.12	0.93
Slb(20)						
m	8.67	5.84	14.12	6.30	13.27	8.27
SD	0.50	1.80	3.37	1.55	3.66	1.17
CHI(20)						
m	32.31	40.43	21.51	41.40	22.33	32.79
SD	2.67	5.96	5.91	1.17	5.73	4.75
MR(20)						
m	14.90	20.71	28.02	22.90	25.41	13.42
SD	2.84	8.89	6.50	8.31	6.64	2.68
EHu^c						
m	-1.57×10^4	-2.06×10^4	-2.51×10^4	-2.13×10^4	-2.45×10^4	-1.37×10^4
SD	3584	3604	3550	2037	3105	2186
HyE^c						
m	−11.89	−7.26	−16.42	−7.43	−15.70	−11.43
SD	0.84	1.44	3.49	0.95	3.53	1.74
Ref^c						
m	19.55	30.03	32.94	31.03	32.16	16.74
SD	5.13	6.39	7.27	4.55	6.12	3.03
Pol^c						
m	7.95	12.07	13.35	12.51	13.02	6.82
SD	2.06	2.56	2.86	1.82	2.40	1.23

Dipole moment: DM(20) – Bumble (1999); logarithm of the partition coefficient: log P(20) – Bumble (1999), log P^c – HYPERCHEM^c; log activity coefficient: Lac(20) – Bumble (1999); hydrophobicity: Hyd(20) – Bumble (1999); Hyd(19) – Hessa *et al.* (2005); Hyd(21) – Kyte and Doolittle (1982); solubility: Slb(20) – Bumble (1999); magnetic susceptibility: CHI(20) – Bumble (1999); molar refraction: MR(20) – Bumble (1999); Hückel energy: EHu^c – HYPERCHEM^c; hydration energy: HyE^c – HYPERCHEM^c; refractivity: Ref^c – HYPERCHEM^c; polarizability: Pol^c – HYPERCHEM^c; m, arithmetic mean; SD, standard deviation; SAR, structure–activity relationship; MDF, molecular descriptors family.

by the model with one descriptor (the two descriptors evolved from a total number of 200 descriptors). Interestingly, the interaction was made via geometry and the dominant atomic property was the charge both in the model with two descriptors and in the model with one descriptor. Four factors were identified in the analysis of 200 molecular descriptors by using the factor analysis approach. Two models (one with one factor and the other with two factors) proved to be statistically significant. The squared correlation coefficient of the model with two factors has shown that this investigation is not useful for characterizing hydrophobicity on the Hessa *et al.* scale [abbreviated as Hyd(19)], the value being lower than the value of the squared correlation coefficient obtained by the model with two molecular descriptors (see Table 2). Furthermore, it may be concluded that the decrease in the number of descriptors through factors creation is not a useful approach compared with the linear regression approach.

By using the MDF SAR models, the 14 properties of six amino acids were predicted and the determined values were included into the cluster analysis of α1 and α2 type I collagen chains. Three clusters were obtained on each chain. By analyzing the α1 or the α2 type I collagen chains using two-step cluster analysis (based on similarities of amino acid sequences) the following findings are highlighted.

- Some amino acid groups were identified to be in the same cluster:
- Ala and Gly (cluster 1 on α1 chain and cluster 3 on α2 chain);
- Ile, Leu, Met, Phe, and Pro (cluster 2 on α1 chain and cluster 1 on α2 chain) and
- Arg, Asn, Asp, Gln, Glu, His, Lys, Ser, Thr, and Tyr (cluster 3 on α1 chain and cluster 2 on α2 chain).
- Similar values for the mean and standard deviations of the investigated properties were obtained according to the cluster amino acid composition (see Table 3).

Based on these results it can be concluded that the two-step cluster analysis technique is a useful statistical instrument in the characterization and identification of similarities in α1 and α2 type I collagen chains. Moreover, other cluster analysis techniques may also be applied to similarity studies on type I collagen.

A compelling question that arises from the investigation of type I collagen is as follows: 'Is the obtained clusterization specific to the investigated species or is it a characteristic of the type I collagen chains?' This will require future research.

Regarding the ability of the MDF SAR approach to characterize amino acid properties, the following question exists: 'If the α1 and/or α2 type I collagen chains are analyzed, will the model be the same? This will also require future research.

The present study was intended to be the first step in an approach to understand the relationship among chemical structure, physical–chemical properties, and biological role.

Once the MDF models have been constructed, the MDF methodology is able to provide useful information related to the structural nature of the physical–chemical and biological properties of such peptides. Because amino acids are structural components of the collagen, the MDF methodology provides parameters for physical models. These models can then serve for further investigations on the entire structure of collagen, known to be triple chain of amino acids, and for which parameters of the physical model exist.

The modeling process for a small molecule (such as an amino acid) is a relatively easy task for computational chemists. However, difficulty occurs upon extending calculations on large molecules, such as collagen, when all quantum calculations become strongly affected by the extension to the larger scale molecular system (as well as the execution time). The concept proposed in this research was to use knowledge collected at small scale (i.e. on amino acids contained in collagen) such as more specific parameters of the physical model (dominant atomic property, interaction via, interaction model, and structure on activity scale) to reduce the complexity of such calculations.

The similarity analysis on type I collagen also aimed to identify common parameters for physical–chemical and biological properties.

Conclusions

Fourteen amino acid properties (11 distinct ones) were modeled by using the MDF on the SAR approach. In one of 14 case, the MDF SAR determination was ≤65%. In almost 79%, the MDF SAR determination was ≥90%, which proved the ability of the method to characterize amino acid properties. Sixty-four percent of the investigated amino acid properties proved to be strongly related to the geometry of compounds.

The linear regression approach was shown to be the best solution compared with factor analysis in the characterization of hydrophobicity using a sample of 200 molecular descriptors.

Two-step cluster analysis techniques exemplified their usefulness in similarity analysis of α1 and α2 type I collagen chains. Future research will be necessary for studying the usefulness of other cluster analysis techniques in the characterization of type I collagen chains.

Acknowledgment

The research was partly supported by the UEFISCSU Romania through research grants (AT93/2007, ID_458/2007, 10_1051/2007).

References

1. Koyack M.J., Cheng R.P. (2006) Design and synthesis of beta-peptides with biological activity. Methods Mol Biol;340:95–109.
2. Benavides M.A., Oelschlager D.K., Zhang H.-G., Stockard C.R., Vital-Reyes V.S., Katkoori V.R., Manne U. *et al.* (2007) Methionine inhibits cellular growth dependent on the p53 status of cells. Am J Surg;193:274–283.
3. Grantham R. (1974) Amino acid difference formula to help explain protein evolution. Science;185:862–864.
4. Dosztanyi Z., Torda A.E. (2001) Amino acid similarity matrices based on force fields. Bioinformatics;17:686–699.
5. Ivanisenko V.A., Eroshkin A.M., Kolchanov N.A. (2005) WebProAnalyst: an interactive tool for analysis of quantitative structure-activity relationships in protein families. Nucleic Acids Res;33(Web Server Issue):W99–W104.
6. Mulakala C., Lambris J.D., Kaznessis Y. (2007) A simple, yet highly accurate, QSAR model captures the complement inhibitory activity of compstatin. Bioorg Med Chem;15:1638–1644.

7. Restrepoa G., Villaveces J.L. (2005) From trees (dendrograms and consensus trees) to topology. Croat Chem Acta;78:275–281.
8. Bolboaca S.D., Jäntschi L. (2007) How good the characteristic polynomial can be for correlations? Int J Mol Sci;8:335–345.
9. Veit G., Kobbe B., Keene D.R., Paulsson M., Koch M., Wagener R. (2006) Collagen XXVIII, a novel von Willebrand factor A domain-containing protein with many imperfections in the collagenous domain. J Biol Chem;281:3494–3504.
10. Hulmes D.J., Miller A. (1979) Quasi-hexagonal molecular packing in collagen fibrils. Nature;282:878–880.
11. Kadler K.E., Holmes D.F., Trotter J.A., Chapman J.A. (1996) Collagen fibril formation. Biochem J;316:1–11.
12. Orgel J.P., Wess T.J., Miller A. (2000) The in situ conformation and axial location of the intermolecular cross-linked non-helical telopeptides of type I collagen. Structure;8:137–142.
13. Orgel J.P., Miller A., Irving T.C., Fischetti R.F., Hammersley A.P., Wess T.J. (2001) The in situ supermolecular structure of type I collagen. Structure;9:1061–1069.
14. Orgel J.P.R.O., Miller A., Irving T.C., Wess T.J. (2002) Recent insights into the three-dimensional molecular packing structure of native type I collagen. Fibre Diffraction Rev;10:40–49.
15. Cabral W.A., Makareeva E., Colige A., Letocha A.D., Ty J.M., Yeowell H.N., Pals G. *et al.* (2005) Mutations near amino end of alpha1(I) collagen cause combined osteogenesis imperfecta/Ehlers-Danlos syndrome by interference with N-propeptide processing. J Biol Chem;280:19259–19269.
16. Gensure R.C., Makitie O., Barclay C., Chan C., Depalma S.R., Bastepe M., Abuzahra H. *et al.* (2005) A novel COL1A1 mutation in infantile cortical hyperostosis (Caffey disease) expands the spectrum of collagen-related disorders. J Clin Invest;115:1250–1257.
17. Fukumitsu N., Uchiyama M., Mori Y., Kishimoto K., Nakada J. (2003) A comparative study of prostate specific antigen (PSA), C-terminal propeptide of blood type I procollagen (PICP) and urine type I collagen-crosslinked N telopeptide (NTx) levels using bone scintigraphy in prostate cancer patients. Ann Nucl Med;17:297–303.
18. Jäntschi L. (2005) Molecular descriptors family on structure activity relationships: 1. Review of the methodology. Leonardo Electronic J Pract Technol;6:76–98.
19. Hessa T., Kim H., Bihlmaier K., Lundin C., Boekel J., Andersson H., Nilsson I. *et al.* (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature;433:377–381.
20. Bumble S. (1999) Computer Generated Physical Properties. LLC, Boca Raton: CRC Press.
21. Kyte J., Doolittle R.F. (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol;157:105–132.
22. Wolfenden R. (2007) Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins. J Gen Physiol;129:357–362.
23. Andersen O.S. (2007) Perspectives on membrane protein insertion, protein-bilayer interactions, and amino acid side hydrophobicity. J Gen Physiol;129:351–352.
24. Jäntschi L., Bolboaca S. (2007) Results from the use of molecular descriptors family on structure property/activity relationships. Int J Mol Sci;8:189–203.
25. Orgel J.P., Irving T.C., Miller A., Wess T.J. (2006) Microfibrillar structure of type I collagen in situ. Proc Natl Acad Sci U S A; 103:9001–9005.
26. Nakai K., Kidera A., Kanehisa M. (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng;2:93–100.

Notes

[a] Pollitt R., McMahon R., Nunn J., Bamford R., Afifi A., Bishop N., Dalton A. (2006) Mutation analysis of COL1A1 and COL1A2 in patients diagnosed with osteogenesis imperfecta type I-IV. Human mutation #901 [online]. Available at: http://www3.interscience.wiley.com/homepages/38515/pdf/901.pdf

[b] Ralston S.H., Uitterlinden A.G., Brandi M.L., Balcells S., Langdahl B.L., Lips P., Lorenc R. *et al.* (2006) Large-scale evidence for the effect of the COLIA1 Sp1 polymorphism on osteoporosis outcomes: the GENOMOS study. PLoS Med;3:e90 [online]. Available at: http://medicine.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pmed.0030090

[c] HYPERCHEM, Molecular Modelling System (online); ©2003, Hypercube (cited 2007 June). Available at: http://hyper.com/products/

[d] MDF SAR (online software); ©2003, L (cited 2007 June). Available at: http://l/Chemistry/SARs/MDF_SARs/