MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **57** (2007) 143-155

ISSN 0340 - 6253

Numerical characterization and similarity analysis of neurocan gene

Wen Zhu, Bo Liao *, Jiawei Luo, Renfa Li
School of Computer and Communication, Hunan University
Changsha Hunan 410082,China

(Received May 15, 2006)

Abstract. In terms of the classifications of the amino acids , we reduce an amino-acid sequence of neurocan gene to three binary sequences. With a so-generated binary sequences representation of neurocan gene, we associate two 2×2 matrices whose elements are independent of the lengths of neurocan, the elements of which are given by the frequency of occurrence of all $(0,1)$ triplets in the binary sequences. Also, we define an information entropy for the amino-acid sequence of neurocan gene. The eigenvalues of the so-constructed matrices and the information entropies are used to characterize individual neurocan sequences.

1 Introduction

Previously, most sequence comparisons are based on the alignment of the strings: a distance function is used to represent insertion, deletion, and substitution of letters of the compared strings. Using the distance function, one can compare DNA, RNA or protein primary sequences and resolve the questions of homology of macromolecules.

Christa K.Prange[1] described the complete coding sequence of human neurocan mRNA, known as CSPG3, as well as mapping data, expression analysis,and genomic structure. Sequence homology searches indicated close homology to the mouse and Rat proteoglycan, neurocan(GenBank access Nos X84727 and M97101).[2-6]

In this paper, based on the classification of the amino acids, we reduce an amino-acid sequence of neurocan gene to three $(0,1)$ sequences. Each $(0,1)$ sequence may be regarded as a coarse-grained description of the amino-acid sequences. Via comparisons of the reduced sequences it will be easier to understand the biological function of various kinds of the amino acids. We construct a set of 2×2 matrices for the $(0,1)$ sequence of a amino acid sequence of neurocan gene and introduce a set of novel invariants to characterize the amino acid sequences of neurocan genes. We construct 12-component vectors consisting of f^X-components, 6-component vectors consisting of the normalized leading eigenvalues and 3-component vectors consisting of the information entropies. We will make a comparison for human neurocan gene(AAC80576), Rattus brevican(Rattus norvegicus, NP_037048), Gallus neurocan(AAD24546), mouse neurocan(S52781), Mus brevican (Mus musculus, NP_031555), rat neurocan(S28764), Rattus neurocan (Rattus norvegicus, AAC15766), Versican core protein precursor ($Q9ERB4\_1$), versican -Rattus norvegicus(AAC40166). The similarities are computed by calculating the Euclidean distance between the end point of the vectors or calculating the correlation angle of two vectors.

*Corresponding author E-mail: dragonbw@163.com; Fax:+86-731-8821715

2 Construction of the binary sequences and a matrix for Neurcan gene

In a protein sequence associated to a neurocan gene, the 20 amino acids can be divided into four classes, i.e, hydrophilic, which are polar $HP = \{D, N, S, H, T, C\}$; hydrophobic which are apolar $HA = \{Y, F, V, I, W, M, L\}$; apolar and small volume $AS = \{G, P\}$; and others $O = \{R, K, E, A, Q\}$

Let $S = s_1 s_2 \cdots$ be an arbitrary amino acid sequence of neurocan gene. Then we have three map $\phi_i, i = 1, 2, 3$, which maps S into (HP, HA), (HP, AS), (HP, O)-characteristic sequences , respectively. Explicitly $\phi_i(s) = \phi_i(s_1)\phi_i(s_2)\cdots$, where

$$\begin{aligned}
\phi_1(g_i) = 1, if \quad g_i \in \{D, N, S, H, T, C, Y, F, V, I, W, M, L\};\\
and \quad \phi_1(g_i) = 0, if \quad g_i \in \{R, K, E, A, Q, G, P\}\\
\phi_2(g_i) = 1, if \quad g_i \in \{D, N, S, H, T, C, R, K, E, A, Q\};\\
and \quad \phi_2(g_i) = 0, if \quad g_i \in \{Y, F, V, I, W, M, L, G, P\}\\
\phi_3(g_i) = 1, if \quad g_i \in \{D, N, S, H, T, C, G, P\};\\
and \quad \phi_3(g_i) = 0, if \quad g_i \in \{Y, F, V, I, W, M, L, R, K, E, A, Q\}
\end{aligned}$$

So, we can obtain three $(0, 1)$ sequences corresponding to the same amino acid sequence of neurocan gene . We call them (HP,HA)-,(HP,AS)-, and (HP,O)-characteristic sequences of the amino acid sequences.

For a neurocan gene, using the three classifications, respectively, we reduce the sequence into three $(0, 1)$ sequences. In this representation, some information of the amino acid sequence of neurocan gene structure may be lost, however, it does make it easier to compare long neurocan gene sequences and get some conclusions that cannot be obtained from direct comparison of neurocan gene sequences.

For a (0,1) sequence, there are only eight possible triplets in a characteristic sequence. Using the frequencies of eight possible triplets, we construct a $2 \times 2 \times 2$ cubic matrix with eight entries $f^X_{ijk} = m^X_{ijk}/(N-2)$, where m^X_{ijk} is the number of $(0, 1)$-triplets ijk in X, and N is the length of X, obviously, $\Sigma m^X_{ijk} = N - 2$ and f^X_{ijk} represents the frequency of occurrence of the $(0, 1)$-triplet ijk in X. The elements of the matrix are independent of the lengths of neurocan gene sequences when different length of neurocan gene sequence are compared.

In Tables 1-3, the condensed matrices are constructed for nine amino-acid sequences of neurocan gene, where the headers of the two 2×2 matrices represent the first entry i of a triplet (i, j, k) and the j, k entries of the triplets consist of j(row) and k(column) entries of the 2×2 matrices.

Table 1: Frequency of Triplets ijk for the ten (HP, HA)-characteristic sequence

Human					
0	0	1	1	0	1
0	0.1077	0.1160	0	0.1160	0.1387
1	0.1213	0.1334	1	0.1327	0.1342
Brevican rattus					
0	0	1	1	0	1
0	0.1299	0.1198	0	0.1198	0.1369
1	0.1446	0.1120	1	0.1120	0.1322
Gallus					
0	0	1	1	0	1
0	0.1289	0.1165	0	0.1165	0.1281
1	0.1157	0.1289	1	0.1289	0.1366
Mouse					
0	0	1	1	0	1
0	0.1051	0.1082	0	0.1082	0.1438
1	0.1098	0.1422	1	0.1414	0.1414
Brevican Mus musculus					
0	0	1	1	0	1
0	0.1237	0.1201	0	0.1203	0.1249
1	0.1305	0.1146	1	0.1146	0.1510
Rat					
0	0	1	1	0	1
0	0.0924	0.1067	0	0.1067	0.1434
1	0.1092	0.1410	1	0.1402	0.1602
Rattus					
0	0	1	1	0	1
0	0.0934	0.1057	0	0.1044	0.1499
1	0.1093	0.1462	1	0.1450	0.1462
Versican core protein precursor					
0	0	1	1	0	1
0	0.0578	0.0867	0	0.0896	0.1618
1	0.0983	0.1503	1	0.1532	0.2023
Versican Rattus					
0	0	1	1	0	1
0	0.0657	0.0934	0	0.0938	0.1428
1	0.0879	0.1482	1	0.1482	0.2198

Table 2: Frequency of Triplets ijk for the ten (HP, AS)-characteristic sequence

Human					
0	0	1	1	0	1
0	0.0773	0.1031	0	0.1024	0.1440
1	0.1137	0.1237	1	0.1237	0.1941
Brevican rattus					
0	0	1	1	0	1
0	0.1042	0.1011	0	0.0995	0.1493
1	0.1135	0.1369	1	0.1353	0.1602
Gallus					
0	0	1	1	0	1
0	0.0831	0.1025	0	0.1025	0.1491
1	0.1196	0.1320	1	0.1320	0.1793
Mouse					
0	0	1	1	0	1
0	0.0711	0.1082	0	0.1074	0.1422
1	0.1153	0.1342	1	0.1342	0.1872
Brevican Mus musculus					
0	0	1	1	0	1
0	0.1022	0.1010	0	0.0999	0.1453
1	0.1010	0.1453	1	0.1453	0.1600
Rat					
0	0	1	1	0	1
0	0.0709	0.1067	0	0.1060	0.1418
1	0.1052	0.1426	1	0.1426	0.1841
Rattus					
0	0	1	1	0	1
0	0.0602	0.1106	0	0.1106	0.1474
1	0.1093	0.1486	1	0.1486	0.1646
Versican core protein precursor					
0	0	1	1	0	1
0	0.0838	0.1040	0	0.1012	0.1503
1	0.1127	0.1416	1	0.1387	0.1676
Versican Rattus					
0	0	1	1	0	1
0	0.0536	0.0808	0	0.0808	0.1629
1	0.0934	0.1503	1	0.1499	0.2282

Table 3: Frequency of Triplets ijk for the ten (HP, O)-characteristic sequence

Human					
0	0	1	1	0	1
0	0.1569	0.1312	0	0.1312	0.1114
1	0.1228	0.1198	1	0.1198	0.1069
Brevican rattus					
0	0	1	1	0	1
0	0.2255	0.1400	0	0.1384	0.0762
1	0.1089	0.1073	1	0.1058	0.0980
Gallus					
0	0	1	1	0	1
0	0.1522	0.1258	0	0.1258	0.1071
1	0.1118	0.1211	1	0.1211	0.1351
Mouse					
0	0	1	1	0	1
0	0.1461	0.1342	0	0.1342	0.1177
1	0.1342	0.1177	1	0.1177	0.0979
Brevican Mus musculus					
0	0	1	1	0	1
0	0.2077	0.1351	0	0.1339	0.0931
1	0.1237	0.1044	1	0.1044	0.0976
Rat					
0	0	1	1	0	1
0	0.1394	0.1315	0	0.1315	0.1171
1	0.1275	0.1211	1	0.1211	0.1108
Rattus					
0	0	1	1	0	1
0	0.1118	0.1204	0	0.1192	0.1278
1	0.1155	0.1327	1	0.1314	0.1413
Versican core protein precursor					
0	0	1	1	0	1
0	0.1763	0.1503	0	0.1503	0.1098
1	0.1474	0.1127	1	0.1127	0.0405
Versican Rattus					
0	0	1	1	0	1
0	0.1223	0.1336	0	0.1340	0.1260
1	0.1294	0.1302	1	0.1307	0.0938

Observing Table 1-3, we can obtain some common features of nine neurocan gene sequences, that are not easily visible in primary sequences. In each of the three characteristic sequences, either $m^X_{001} = m^X_{100}$ or $|m^X_{001} - m^X_{100}| = 1$, and also either $m^X_{110} = m^X_{011}$ or $|m^X_{011} - m^X_{110}| = 1$, so the differences $f^X_{001} - f^X_{100}$, and $f^X_{011} - f^X_{110}$ are not more than $\frac{1}{N-2}$ in magnitude. Obviously, $lim_{N\longrightarrow\infty}(f^X_{001} - f^X_{100}) = 0$, and $lim_{N\longrightarrow\infty}(f^X_{011} - f^X_{110}) = 0$; On the other hand, $f^X_{001}+f^X_{101}$ and $f^X_{010}+f^X_{011}$ give fractions of 2-member sequences 01 either precedes for f_{i01} or succeeded for f_{01k} by some other index and so must be asymptotically equal for $N \to \infty$.

Next, we will prove the conjecture: $m^X_{001} = m^X_{100}$ or $|m^X_{001} - m^X_{100}| = 1$. Suppose DNA sequence is $g = g_1g_2 \dots g_s$ and corresponding (01) sequence is $\phi(g) = \phi(g_1)\phi(g_2) \dots \phi(g_s)$. we prove it by induction on s.

Proof : Clearly, when $s = 1, 2, 3$, the conjecture holds. Suppose the conjecture holds for all $s < n$, and when $s = n$ the first occurrence of 100 is at i.

Case 1: if $i = 1$, i.e. $\phi(g) = 100\phi(g_4) \dots \phi(g_n)$

If $\phi(g_4) = 1$, then the occurrence number of100 in $100\phi(g_4)$ is equal to the occurrence number of 001 in $100\phi(g_4)$. Because the length of the sequence $\phi(g_5) \dots \phi(g_n)$ is less then n, the conjecture is holed by assumption of the induction.

If $\phi(g_4) = 0$, and if there is a $j, 4 < j < n$, such that $\phi(g_5) = \dots = \phi(g_{j-1}) = 0, \phi(g_j) = 1$, then the number of 100 and the number of 001 in the sequence $100\phi(g_4)\phi(g_5) \dots \phi(g_j)$ are the same. Because the length of sequence $\phi(g_{j+1}) \dots \phi(g_n)$ is less then n, the conjecture is holed by the assumption of the induction. If there not exist such a j, such that $\phi(g_j) = 1$,i.e. $\phi(g_4) = \phi(g_5) \dots = \phi(g_n) = 0$,then difference of number of 100 and the number of 001 is one. The conjecture is proved.

Case2: If $i = 2$, $\phi(g) = \phi(g_1)100\phi(g_5) \dots \phi(g_n)$, we just consider the sequence $100\phi(g_5) \dots \phi(g_n)$, the conclusion is true.

Case3: If $i = 3$, $\phi(g) = \phi(g_1)\phi(g_2)100\phi(g_6) \dots \phi(g_n)$, and if $\phi(g_1)\phi(g_2) = 00$, then the number of 001 and the number of 100 in sequence $\phi(g_1)\phi(g_2)100$ are the same, and the length of the sequence $\phi(g_6) \dots \phi(g_n)$, is less than n, the conclusion is true by the induction. If $\phi(g_1)\phi(g_2) \neq 00$, we consider the sequence $100\phi(g_6) \dots \phi(g_n)$, because the length of the sequence is less than n, the conjecture is right by the induction.

Case4: If $i > 3$, i.e., the sequence is $\phi(g_1) \dots \phi(g_{i-1})100\phi(g_{i+3}) \dots \phi(g_n)$, and if $\phi(g_{i-2})\phi(g_{i-1}) = 00$, we just consider the sequence $\phi(g_1) \dots \phi(g_{i-3})$, if $\phi(g_{i-3}) = 1$, then $\phi(g_{i-3})\phi(g_{i-2})\phi(g_{i-1}) = 100$. This is contrary to that the first occurrence of 100 is at i, so $\phi(g_{i-3}) = 0$, and $\phi(g_1) = \phi(g_2) = \dots = \phi(g_{i-3}) = 0$ follows by the same reason, the number of 001 and the number of 100 are the same in sequence $\phi(g_1) \dots \phi(g_{i-1})100$, and the length of $\phi(g_{i+3}) \dots \phi(g_n)$ is less than n, the conclusion is true by the assumption of the induction. If $\phi(g_{i-2})\phi(g_{i-1}) \neq 00$, we consider the sequence $\phi(g_1) \dots \phi(g_{i-1})$ again. If 001 appears in the sequence $\phi(g_1) \dots \phi(g_{i-1})$ and the last occurrence of 001 is at $j < i - 2$, that is $\phi(g_j)\phi(g_{j+1})\phi(g_{j+2}) = 001$. If $\phi(g_{j-1}) = 1$, $\phi(g_{j-1})\phi(g_j)\phi(g_{j+1}) = 100$ is contrary, so $\phi(g_{j-1}) = 0$, and $\phi(g_1) = \dots = \phi(g_{i-1}) = 0$ follows the same reason. It follows that the number of 001 and the number of 100 in the sequence $\phi(g_1) \dots \phi(g_{i-1})100$ are the same, and

the length of $\phi(g_{i+3})\dots\phi(g_n)$ is less than n, the conjecture is true by the induction.
If 001 do not appear in the sequence $\phi(g_1)\dots\phi(g_{i-1})$, we consider the sequence $100\phi(g_{i+3})\dots\phi(g_n)$. Because the length of it is less than n, the conclusion is true by the induction.
Summing up Case1 , Case2, Case3 and Case4, the conjecture is proved.

We can also prove $m^X_{110} = m^X_{011}$ or $|m^X_{011} - m^X_{110}| = 1$ using the similar method.

In order to illustrate the numerical characterization of neurocan gene, we construct four independent f^X-components: $(f^X_0)^3, (f^X_0)^2 f^X_1, f^X_0 (f^X_1)^2, (f^X_1)^3$, where $f^X_0 = \sum_{j,k=0or1} f^X_{0jk}$, $f^X_1 = \sum_{j,k=0or1} f^X_{1jk}$. From Table 4-6, we present all f^X-components of nine neurocan gene sequences.

Table 4: f^X-components: $(f^X_0)^3, (f^X_0)^2 f^X_1, f^X_0 (f^X_1)^2, (f^X_1)^3$ based on patterns HP(HA)

Components	Human	B-Rattus	Gallus	Mouse	B-Mus	Rat	Rattus	V	V-Rattus
$(f^X_0)^3$	0.1095	0.1244	0.1176	0.1007	0.1171	0.0908	0.0939	0.0607	0.0618
$(f^X_0)^2 f^X_1$	0.1194	0.1248	0.1224	0.1157	0.1222	0.1112	0.1127	0.0938	0.0945
$f^X_0 (f^X_1)^2$	0.1302	0.1252	0.1275	0.1331	0.1276	0.1362	0.1352	0.1448	0.1445
$(f^X_1)^3$	0.1419	0.1256	0.1327	0.1530	0.1333	0.1669	0.1623	0.2235	0.2211

Table 5: f^X-components: $(f^X_0)^3, (f^X_0)^2 f^X_1, f^X_0 (f^X_1)^2, (f^X_1)^3$ based on patterns HP(AS)

Components	Human	B-Rattus	Gallus	Mouse	B-Mus	Rat	Rattus	V	V-Rattus
$(f^X_0)^3$	0.0777	0.0946	0.0835	0.0789	0.0908	0.0770	0.0788	0.0865	0.0541
$(f^X_0)^2 f^X_1$	0.1044	0.1130	0.1075	0.1051	0.1112	0.1040	0.1050	0.1091	0.0889
$f^X_0 (f^X_1)^2$	0.1402	0.1350	0.1385	0.1399	0.1362	0.1404	0.1399	0.1376	0.1462
$(f^X_1)^3$	0.1883	0.1613	0.1784	0.1863	0.1668	0.1896	0.1865	0.1376	0.2405

Table 6: f^X-components: $(f^X_0)^3, (f^X_0)^2 f^X_1, f^X_0 (f^X_1)^2, (f^X_1)^3$ based on patterns HP(O)

Components	Human	B-Rattus	Gallus	Mouse	B-Mus	Rat	Rattus	V	V-Rattus
$(f^X_0)^3$	0.1495	0.1967	0.1334	0.1509	0.1861	0.1402	0.1108	0.2020	0.1370
$(f^X_0)^2 f^X_1$	0.1322	0.1415	0.1277	0.1325	0.1399	0.1296	0.1199	0.1423	0.1288
$f^X_0 (f^X_1)^2$	0.1169	0.1171	0.1222	0.1164	0.1051	0.1199	0.1297	0.1002	0.1210
$(f^X_1)^3$	0.1034	0.0732	0.1170	0.1022	0.0790	0.1109	0.1404	0.0706	0.1137

By F^{HP}, F^{AS} and F^O, we denote the cubic matrices for the $(HP, HA), (HP, AS)$ and (HP, O) characteristic sequences, respectively. We partition each of the cubic matrices into a

pair of 2×2 condensed matrices F_0^X and F_1^X, whose elements are 100 times the overall frequencies, in other word, $F_0^X = 100\begin{pmatrix} (f_0^X)^3 & (f_0^X)^2 f_1^X \\ f_0^X f_1^X f_0^X & f_0^X (f_1^X)^2 \end{pmatrix}$ and $F_1^X = 100\begin{pmatrix} f_1^X (f_0^X)^2 & f_1^X f_0^X f_1^X \\ (f_1^X)^2 f_0^X & (f_1^X)^3 \end{pmatrix}$ with X being HP, AS or O. For example, for human neurocan gene $F_0^{HP} = \begin{pmatrix} 10.95 & 11.94 \\ 11.94 & 13.02 \end{pmatrix}$.

Since the pioneering work of Shannon(1948) entropy is regarded as a measure of information in a probability distribution $\overrightarrow{P}$[7-10]. Here, let $H_X = \sum_{i,j,k=0or1} f_{ijk}^X log_2 f_{ijk}^X$, we call H_X be the information entropy of neurocan sequence X. In Table 7, the information entropies of three characteristic sequences are listed for nine neurocan sequences.

Table 7: The information entropies of the coding sequences

Patterns	Human	B-Rattus	Gallus	Mouse	B-Mus	Rat	Rattus	V	V-Rattus
HP(HA)	2.9949	2.9947	2.9977	2.9864	2.9945	2.9762	2.9770	2.9015	2.9039
HP(AS)	2.9517	2.9778	2.9636	2.9537	2.9725	2.9506	2.9488	2.9662	2.8665
HP(O)	2.9907	2.9247	2.9923	2.9901	2.9496	2.9966	2.9962	2.9176	2.9925

Among all eigenvalues the leading eigenvalue of a matrix, λ_1, often plays a special role. In the case of the adjacency matrix of trees, Lovasz and Pelikan[11] suggested the leading eigenvalue λ_1 as an index of molecular branching. More recently it was shown that the leading eigenvalue of a substituted path matrix, $\lambda\lambda_1$, gives an even better characterization of molecular branching[12-15]. Since the eigenvalues of a matrix are one of the well-known matrix invariants, we consider the leading eigenvalues of the six matrices $(F_0^{HP}, F_1^{HP}, F_0^{AS}, F_1^{AS}, F_0^O, F_1^O)$ to acquire more compact information of the six matrices for each neurocan sequence. In each

Table 8: Leading Eigenvalues of the 6 Matrices F_0^X, F_1^X, (X being HP,AS,O),for the nine neurocan

	F_0^{HP}	F_1^{HP}	F_0^{AS}	F_1^{AS}	F_0^O	F_1^O
Human	23.9698	26.1335	21.9727	29.2692	26.6401	23.5584
B-Rattus	24.9601	25.0401	22.9591	27.4293	30.3891	22.9328
Gallus	24.5050	25.5154	22.1962	28.5915	25.5623	24.4667
Mouse	23.3729	26.8750	21.8836	29.1372	26.7268	23.4732
B-Mus	24.4663	25.5471	22.6993	27.8008	29.1244	21.8872
Rat	22.6993	27.8068	21.7424	29.3579	26.0047	24.0514
Rattus	22.9126	27.4956	21.8704	29.1464	24.0522	26.0254
V	20.5544	31.7309	22.4102	28.2679	30.2229	21.2870
V-Rattus	20.6301	31.5557	20.0269	32.9382	25.8048	24.2485

row of Table 8, the leading eigenvalues of F_1^{HP} , F_1^{AS} and F_0^O are large, Corresponding to neurocan sequence, this means that the values of D, N, S, H, T, C, Y, F, V, I, W, M, L; D, N, S, H, T, C, R, K, E, A, Q; Y, F, V, I, W, M, L, R, K, E, A, Q are large, respectively.

3 Similarities and Dissimilarities

In order to compare similarity, we construct 12-component vectors consisting of the f^X-components, 6-component vectors consisting of the leading eigenvalue of the six matrices $(F_0^{HP}, F_1^{HP}, F_0^{AS}, F_1^{AS}, F_0^O, F_1^O)$, and 3-component vectors consisting of the information entropies. All the vectors rooted at the $(0,0,0)$ position. For example, in figure 1, vector $\overrightarrow{oa}$ and vector $\overrightarrow{ob}$ are correspond the coding sequence of Human and B-Rattus, respectively. $\|\overrightarrow{ba}\|$ is the Euclidean distance between the end points of the vectors $\overrightarrow{oa}$ and $\overrightarrow{ob}$. α is the the correlation angle of two vectors $\overrightarrow{oa}$ and $\overrightarrow{ob}$.

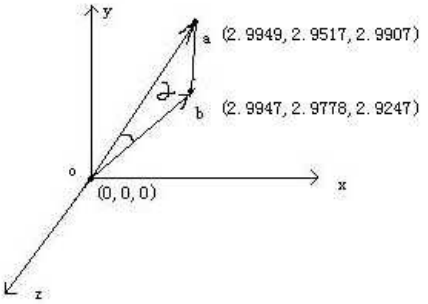

Figure 1

The similarities among such vectors can be computed in two ways: (1) we calculate the Euclidean distance between the end points of the vectors; (2) we calculate the correlation angle of the two vectors. The smaller Euclidean distance between the end points of two vectors, the more similar the corresponding neurocan genes. And the more small the the correlation angle between two vectors, the more similar the neurocan gene. Obviously, if $\alpha = 0$ and $\|\overrightarrow{ba}\| = 0$, then the corresponding structure of vector $\overrightarrow{oa}$ is the same as the corresponding structure of vector $\overrightarrow{ob}$. If $\alpha \to 90^o$ or $\|\overrightarrow{ba}\| \to \infty$, then the corresponding structure of $\overrightarrow{oa}$ and the corresponding structure of vector $\overrightarrow{ob}$ have little similarity.

Let V be the Versican core protein precursor ($Q9ERB4\_1$). In Tables 9 and 10, we list the Euclidean distances between the end points of the 12-component vectors and the angle between the 12-component vectors,respectively. In Tables 11 and 12, we list the Euclidean distances between the end points of the 3-Component vectors and the angle between the 3-component vectors, respectively. In Tables 13 and 14, we list the Euclidean distances between the end point of the 6-component vectors and the angle between the 6-component vectors, respectively.

Table 9: Similarity/Dissmilarity Table for the nine neurocan sequences based on Euclidean between the End point of the 12-Component Vectors

	B-Rattus	Gallus	Mouse	B-mus	Rat	Rattus	V	V-Rattus
Human	0.069885	0.028341	0.015248	0.054552	0.035197	0.062579	0.120183	0.114983
B-Rattus		0.081900	0.073727	0.021611	0.095522	0.126555	0.124689	0.168314
Gallus			0.038042	0.069786	0.048630	0.053503	0.142095	0.130951
Mouse				0.057114	0.023407	0.059574	0.106749	0.104620
B-mus					0.079025	0.112057	0.113330	0.152272
Rat						0.044417	0.104207	0.086736
Rattus							0.141797	0.101414
V								0.112749

Table 10: Similarity/Dissmilarity Table for the nine neurocan sequences based on the Angle between the 12-Component Vectors

	B-Rattus	Gallus	Mouse	B-mus	Rat	Rattus	V	V-Rattus
Human	0.154433	0.062736	0.034010	0.121879	0.078165	0.139921	0.252273	0.233200
B-Rattus		0.180961	0.163064	0.046239	0.211488	0.281118	0.263834	0.354200
Gallus			0.084407	0.156187	0.107452	0.119387	0.300579	0.267446
Mouse				0.127518	0.051903	0.133068	0.222258	0.209926
B-mus					0.176084	0.250570	0.237363	0.318081
Rat						0.098850	0.217341	0.169818
Rattus							0.301092	0.202776
V								0.234212

Table 11: Similarity/Dissmilarity Table for the nine neurocan sequences based on Euclidean the End point of the 3-Component Vectors

	B-Rattus	Gallus	Mouse	B-mus	Rat	Rattus	V	V-Rattus
Human	0.07097	0.01233	0.00875	0.04607	0.01964	0.01895	0.11949	0.12467
B-Rattus		0.06914	0.07019	0.02546	0.07907	0.07916	0.09419	0.0.15884
Gallus			0.01518	0.04374	0.02549	0.02574	0.12183	0.13501
Mouse				0.04538	0.01249	0.01223	0.11234	0.12007
B-mus					0.05499	0.05513	0.09855	0.14589
Rat						0.00201	0.10984	0.11098
Rattus							0.11037	0.11014
V								0.12472

Table 12: Similarity/Dissmilarity Table for the nine neurocan sequences based on the Angle between the 3-Component Vectors

	B-Rattus	Gallus	Mouse	B-mus	Rat	Rattus	V	V-Rattus
Human	0.01304	0.00155	0.00150	0.00863	0.00348	0.00325	0.01588	0.01458
B-Rattus		0.01186	0.01314	0.00446	0.01509	0.01514	0.01342	0.02725
Gallus			0.00135	0.00740	0.00361	0.00353	0.01439	0.01549
Mouse				0.00868	0.00230	0.00219	0.01469	0.01416
B-mus					0.01065	0.01069	0.01235	0.02280
Rat						0.00036	0.01480	0.01217
Rattus							0.01512	0.01211
V								0.02450

Table 13: Similarity/Dissmilarity Table for the nine neurocan sequences based on Euclidean the End point of the 6-Component Vectors

	B-Rattus	Gallus	Mouse	B-mus	Rat	Rattus	V	V-Rattus
Human	4.626	1.8103	0.97292	3.5402	2.2520	3.9697	7.8978	7.6359
B-Rattus		5.2931	4.8646	1.8437	6.2000	8.0024	8.1412	11.0858
Gallus			2.4226	4.4970	8.1101	3.3989	9.2856	8.6721
Mouse				3.6978	1.4988	3.7758	7.0398	6.9841
B-mus					5.0976	7.1786	7.4439	10.0474
Rat						2.8136	6.8606	5.8460
Rattus							9.2268	6.7632
V								7.4712

Table 14: Similarity/Dissmilarity Table for the nine neurocan sequences based on the Angle between the 6-Component Vectors

	B-Rattus	Gallus	Mouse	B-mus	Rat	Rattus	V	V-Rattus
Human	0.07214	0.02877	0.01563	0.05696	0.03609	0.06389	0.12157	0.11446
B-Rattus		0.08214	0.07627	0.02573	0.09812	0.12705	0.12896	0.17255
Gallus			0.03861	0.07226	0.04946	0.05445	0.14340	0.13039
Mouse				0.05949	0.02400	0.06075	0.10768	0.10384
B-mus					0.08193	0.11552	0.11444	0.15449
Rat						0.04519	0.10517	0.08525
Rattus							0.14345	0.10025
V								0.11607

From Tables 9-14, one can find that the neurocan gene of human, mouse, gallus and rat are more similar with each other. And Brevican-Rattus is more similar to Brevican-Mus. Also we find versican -Rattus norvegicus (AAC40166) and Versican core protein precursor ($Q9ERB4\_1$) is very dissimilar to others among the 9 species because its corresponding row has larger entries.

The Euclidean distance measure between vector end points and the correlation angle between vectors are different measures of the similarity of neurocan genes. Observing Table 9 and Table 10, Table 11 and Table 12, Table 13 and Table 14, we find that there exists an overall qualitative agreement among similarities. In general, the correlation angle is the best measure for the similarities. On the other hand, there exists an overall qualitative agreement among similarities based on different descriptors despite some variations among them.

4 Conclusion

In this paper, we have outlined an approach that gives a numerical characterization and similarity analysis of the coding sequences for amino acid sequences of the neurocan gene. As mathematical invariants, 12-component vector, 6-component vector , 3-component vector, leading eigenvalue , and the information entropy were used to characterize the neurocan genes and to analyze the similarities. Our approach considers not only sequences themselves but also chemical structures of the amino acids.

References

[1] Christa K.Prange, Len A.Pennacchio, Kimberly Lieuallen, Wufang Fan, Gregory G.Lennon, Characterization of the human neurocan gene, CSPG3, Gene 221(1998)199-205.

[2] Seares, M.B., Bonaldo, M.Jelene, P.Su, L.Lawton, L.efstratiadis, A., Construction and characterization of a normalized cDNA library,Proc.Natl.Acad.Sci.USA 91(1994), 9228-9232.

[3] Snow, D.Steindler, D.Silver, J., Molecular and cellur characterization of the glial roof plate of the spinal cord and optic tectum: a possible role for a proteoglycan in the development of an axon barrier, Dev.Biol.138(1990b)359-376.

[4] Rauch, U., Karthikeyan, L.,Maurel, P., Margolis, R., Margolis, R., 1992. Cloning and primary structure of neurocan, a developmentally regulated, arregating, chondroitin sulfate and proteoglycan of the brain. J.Biol.Chem.267, 19536-19547.

[5] Rauch, U., Grimpe, B., Kulbe, G., Arnold-Ammer, I., Beier, D., Fassler, R., 1995. Structure and chromosomal location of the mouse neurocan gene. Genomics 28, 405-410.

[6] Rauch, U,.Gao, P., Janetzko,A., flaccus,A., Kilgenberg, L., Tekotte, H., Margolis, R., Margolis, R., 1991. Isolation and characterization of developmentally regulated chon-

droitin sulfate and chondroitin/keratan sulfate proteoglycans of brain identified with monoclonal antibodies. J.Biol.Chem. 266, 14785-14801.

[7] Weiss, O., Miguel, A., Jimenez-Montano, Hanspeter Herzel, Information content of Protein sequences, J.Theor.Biol., 2000,206, 379-386.

[8] Armin, O., Schmit, Hanspeter Herzel, Estimating the entropy of DNA sequences, J.Theor.Biol., 1997,1888, 369-377.

[9] J.L.Oliver, P.Bernaola-Galvan, J.Guerrero-Garcia, R.Roman-Roldan, Entropic Profiles of DNA sequences through Chaos-game-derived Images, J.Theor.Biol., 1993,160, 457-470.

[10] Strait B.J., Dewey, T.G, The shannon information entropy of protein sequences, J. Biophys., 1996, 71(1), 148-155.

[11] Lovasz, L.; Pelikan, J.I, On the eigenvalues of trees, Period. Math. Hung. 1973,3, 175-182

[12] Randic,M., On structural ordering and branching of acyclic saturated hydrocarbons, J.Math.Chem., 1998,24, 345-358

[13] Randic,M.; Plavsic,D.; Razinger, M. ,Double invariants. MATCH Commun. Math. Comput. Chem., 1997,35, 243-259

[14] Randic,M.;Guo, X.;Bobst, S., Use of matrices for characterization of molecular structures. Discrete Mathematical Chemistry DIMACS Workshop on Discrete Mathematical Chemistry;(Hansen,P.;Folwer, P.; Zheng, M. Eds.), Amer., Math. Soc.; Providence, RI,2000,pp 305-322

[15] Randic,M., On molecular branching, Acta Chim.Sloven.1997,44, 57-77