AMINO ACIDS, EUCLIDEAN DISTANCE AND SYMMETRIC MATRIX

Matthew X. He[1], Miguel A. Jiménez-Montaño[2], Paolo E. Ricci[3]

[1]Division of Math, Science and Technology
Nova Southeastern University
Ft. Lauderdale, FL 33314, USA
Email: hem@nova.edu

[2]Facultad de Física e Inteligencia Artificial
Universidad Veracruzana, Xalapa, 91000 Veracruz, México
Email: ajimenez@uv.mx

[3]Dipartimento Di Matematica
Università di Roma "La Sapienza", Roma 00185, Italia
Email: riccip@uniroma1.it

Abstract: In this paper we introduce the general notion of matrix associated with basic building blocks of protein amino acid and discuss the fundamental properties of these matrices. We further apply general amino acid matrix to a special amino acid Euclidean distance matrix introduced by Graham [1] and study the basic properties of this matrix and provide statistical discription to amino acid distances.

***Keywords*:** Amino acid, Euclidean distance, genetic code, codons, symmatric matrix.

1. Introduction

It is well known that the genetic code is encoded in combinations of the four nucleotides found in DNA and then RNA. DNA contains the complete genetic information that defines the structure and function of an organism. Proteins are formed using the genetic code of the DNA. Three different processes are responsible for the inheritance of genetic information and for its conversion from one form to another:

- **Replication:** a double stranded nucleic acid is duplicated to give identical copies. This process perpetuates the genetic information.
- **Transcription:** a DNA segment that constitutes a gene is read and transcribed into a single stranded sequence of RNA. The RNA moves from the nucleus into the cytoplasm.
- **Translation:** the RNA sequence is translated into a sequence of amino acids as the protein is formed. During translation, the ribosome reads three bases (a codon) at a time from the RNA and translates them into one amino acid

These processes are called the Central Dogma of Molecular Biology. The genetic code in messenger ribonucleic acid (mRNA) is composed of A, C, G and U (U for uracil). A mathematical view of genetic code is a map

$$\mathbf{g}: C \to A,$$

where $\mathbf{C} = \{(x_1 x_2 x_3): x_i \in \mathbf{R} = \{A, C, G, U\}\}$ = the set of codons and $\mathbf{A} = \{Ala, Arg, Asp, \ldots, Val, UAA, UAG, UGA\}$= the set of amino acids and termination codons. A

codon is three bases in a DNA or RNA sequence which specify a single amino acid.

One noticeable feature of the genetic code is that some amino acids are encoded by several different but related base codons or triplets. There are 64 triplets or codons. Three triplets (UAA, UAG, and UGA) are stop codons-no amino acids corresponds to their code. The remaining 61 codons represent 20 different amino acids. These genetic code triplets of three bases in mRNA that encode for specific acids during the translation process have some interesting and mathematical logic in their organization. An examination of this logical organization may allow us to better understand the logical assembly of the genetic code and life.

In next section, we introduce the general notion of matrix associated with amino acids and discuss the fundamental properties of these matrices. In section 3, we further apply general amino acid matrix to a special amino acid matrix with Euclidean distances introduced by Graham [1] and study the basic properties of this matrix and frequency distributions of the amino acid distances.

2. Symmetric Matrix Associated with Amino Acids

The 20 standard amino acids in the genetic code display a much higher structural diversity than the four nucleibases within 64 codons. Although the occurence of 20 coded amino acids and their contribution to the origin and evolution of the genetic code have been subjected to a wide range of excellent investigations, it has been unclear what principle governs the selection of the 20 amino acids into the genetic code [2, 3]. It was shown in [7] that amino acids distribution within the genetic code is symmetric along the two possible evolutionary axes through the framework of Quasi-28-gon model.

In this section, we arrange the 20 amino acids in a 20x20 square matrix. Abbreviations of the 20 amino acids are represented by the notations summarized in table below.

Table 1. Amino Acid Abbreviations

3-letter notation	**1-letter notation**
Tyr	**Y**
His	**H**
Gln	**Q**
Arg	**R**
Thr	**T**
Asn	**N**
Lys	**K**
Asp	**D**
Glu	**E**
Gly	**G**
Phe	**F**
Leu	**L**
Ala	**A**
Ser	**S**
Pro	**P**
Ile	**I**
Met	**M**
Val	**V**
Cys	**C**
Trp	**W**

Table 2. Amino Acid Matrix

	Y	H	Q	R	T	N	K	D	E	G	F	L	A	S	P	I	M	V	C	W
Y	YY	YH	YQ	YR	YT	YN	YK	YD	YE	YG	YF	YL	YA	YS	YP	YI	YM	YV	YC	YW
H	HY	HH	HQ	HR	HT	HN	HK	HD	HE	HG	HF	HL	HA	HS	HP	HI	HM	HV	HC	HW
Q	QY	QH	QQ	QR	QT	QN	QK	QD	QE	QG	QF	QL	QA	QS	QP	QI	QM	QV	QC	QW
R	RY	RH	RQ	RR	RT	RN	RK	RD	RE	RG	RF	RL	RA	RS	RP	RI	RM	RV	RC	RW
T	TY	TH	TQ	TR	TT	TN	TK	TD	TE	TG	TF	TL	TA	TS	TP	TI	TM	TV	TC	TW
N	NY	NH	NQ	NR	NT	NN	NK	ND	NE	NG	NF	NL	NA	NS	NP	NI	NM	NV	NC	NW
K	KY	**KH**	**KQ**	KR	KT	KN	KK	KD	KE	KG	KF	KL	KA	KS	KP	KI	KM	KV	KC	KW
D	DY	DH	DQ	DR	DT	DN	DK	DD	DE	DG	DF	DL	DA	DS	DP	DI	DM	DV	DC	DW
E	EY	EH	EQ	ER	ET	EN	EK	ED	EE	EG	EF	EL	EA	ES	EP	EI	EM	EV	EC	EW
G	GY	GH	GQ	GR	GT	GN	GK	GD	GE	GG	GF	GL	GA	GS	GP	GI	GM	GV	GC	GW
F	FY	FH	FQ	FR	FT	FN	FK	FD	FE	FG	FF	FL	FA	FS	FP	FI	FM	FV	FC	FW
L	LY	LH	LQ	LR	LT	LN	LK	LD	LE	LG	LF	LL	LA	LS	LP	LI	LM	LV	LC	LW
A	AY	AH	AQ	AR	AT	AN	AK	AD	AE	AG	AF	AL	AA	AS	AP	AI	AM	AV	AC	AW
S	SY	**SH**	**SQ**	SR	ST	SN	**SK**	SD	SE	SG	SF	SL	SA	SS	SP	SI	SM	SV	SC	SW
P	PY	PH	PQ	PR	PT	PN	PK	PD	PE	PG	PF	PL	PA	PS	PP	PI	PM	PV	PC	PW
I	IY	IH	IQ	IR	IT	IN	IK	ID	IE	IG	IF	IL	IA	IS	IP	II	IM	IV	IC	IW
M	MY	MH	MQ	MR	MT	MN	MK	MD	ME	MG	MF	ML	MA	MS	MP	MI	MM	MV	MC	MW
V	VY	VH	VQ	VR	VT	VN	VK	VD	VE	VG	VF	VL	VA	VS	VP	VI	VM	VV	VC	VW
C	CY	CH	CQ	CR	CT	CN	CK	CD	CE	CG	CF	CL	CA	CS	CP	CI	CM	CV	CC	CW
W	WY	WH	RQ	WR	WT	**KN**	WK	WD	WE	WG	WF	WL	WA	WS	WP	WI	WM	WV	WC	WW

It's easy to see that this matrix A is a 20x20 square matrix and A is symmetric since the matrix A is the same as its transpose A^T. The symmetric matrix has a number of properties [5] that we only list a few main results here.

- If A is a square symmetric matrix, then the eigenvalues of A are all real.
- If A is a square symmetric matrix, then the power of matrix A is also symmetric.

3. Amino Acid Distance Matrix

In this section, we consider a square matrix. The entries of this matrix are given by the amino distances. The amino acid distance (physicochemical) was introduced by Granham [1] as follows:

$$D_{ij} = [\, \alpha (c_i - c_j)^2 + \beta (p_i - p_j)^2 + \gamma (v_i - v_j)^2 \,]^{1/2}$$

where c = composition, p = polarity, and v = molecular volume. In a Euclidean space having these properties as axes, D_{ij} would be the distance between amino acids. The properties are not assumed to be mutually independent; the axes are made orthogonal to facilitate distance calculations. Each property is weighted by dividing the mean distance found with it along in the formula. The constants α, β, γ are squares of the inverses of the D's as indicated in [1]. The similarity between any two amino acid may be measured by this distance. It was observed that if the distance between a pair of amino acids is large, the similarity of the two is small and then the corresponding mutational deterioration will be serious. On the contrary, the small distance for a pair of amino acids suggests a weak deterioration in their mutual mutations. Evidently it leads $D_{ij}T = 0$ if i = j. All other distances were determined in [1]. We arrange all the distances in a matrix A as follows:

Table 3. Euclidian Distance of Amino Acids

	Y	H	Q	R	T	N	K	D	E	G	F	L	A	S	P	I	M	V	C	W
Y	0	83	99	77	92	143	85	160	122	147	22	36	112	144	110	33	36	55	194	37
H	83	0	24	29	47	68	**32**	81	40	98	100	99	86	**89**	77	94	87	84	174	115
Q	99	24	0	43	42	46	**53**	61	29	87	116	113	91	**68**	76	109	101	96	154	130
R	77	29	43	0	71	86	26	96	54	125	97	102	112	110	103	97	91	96	180	102
T	92	47	42	71	0	65	78	85	65	59	103	92	58	58	38	89	81	69	149	128
N	143	68	46	86	65	0	94	23	42	80	158	153	111	46	91	149	142	133	139	174
K	85	**32**	**53**	26	78	94	0	101	56	127	102	107	106	**121**	103	102	95	97	202	110
D	160	81	61	96	85	23	101	0	45	94	177	172	126	65	108	168	160	152	154	181
E	122	40	29	54	65	42	56	45	0	98	140	138	107	80	93	134	126	121	170	152
G	147	98	87	125	59	80	127	94	98	0	153	138	60	56	42	135	127	109	159	184
F	22	100	116	97	103	158	102	177	140	153	0	22	113	155	114	21	28	50	205	40
L	36	99	113	102	92	153	107	172	138	138	22	0	96	145	98	**5**	15	32	198	61
A	112	86	91	112	58	111	106	126	107	60	113	96	0	99	27	94	84	64	195	148
S	144	**89**	**68**	110	58	46	**121**	65	80	56	155	145	99	0	74	142	135	124	112	177
P	110	77	76	103	38	91	103	108	93	42	114	98	27	74	0	95	87	68	169	147
I	33	94	109	97	89	149	102	168	134	135	21	**5**	94	142	95	0	10	29	198	61
M	36	87	101	91	81	142	95	160	126	127	28	15	84	135	87	10	0	21	196	67
V	55	84	96	96	69	133	97	152	121	109	50	32	64	124	68	29	21	0	192	88
C	194	174	154	180	149	139	202	154	170	159	205	198	195	112	169	198	196	192	0	215
W	37	115	130	102	128	174	110	181	152	184	40	61	148	177	147	61	67	88	**215**	0

Since D is defined as a Eucleadian distance, we have following properties:

- D (a, a) = 0, reflective property
- D(a, b) = D (b, a), symmetric property
- D(a,c) ≤ D(a, b) + D(b, c), triangle inequality

for any amino acids a, b, and c. The equality may hold true for special amino acids. It's trivial to note that 20 distances hold 0 due to reflective property. Due to symmetric property, we have 190 distances of 20 amino acids. It's also easy to see that the minimum distance occurs at

D(Ile, Leu) = 5 and

maximum distance occurs at

D(Trp, Cys) = 215.

All other distances are in between 5 and 215. The frequency of distance distribution is illustrated below. It shows that 190 distances are approximately divided into two parts between 5 to 121 and then 121 to 215.

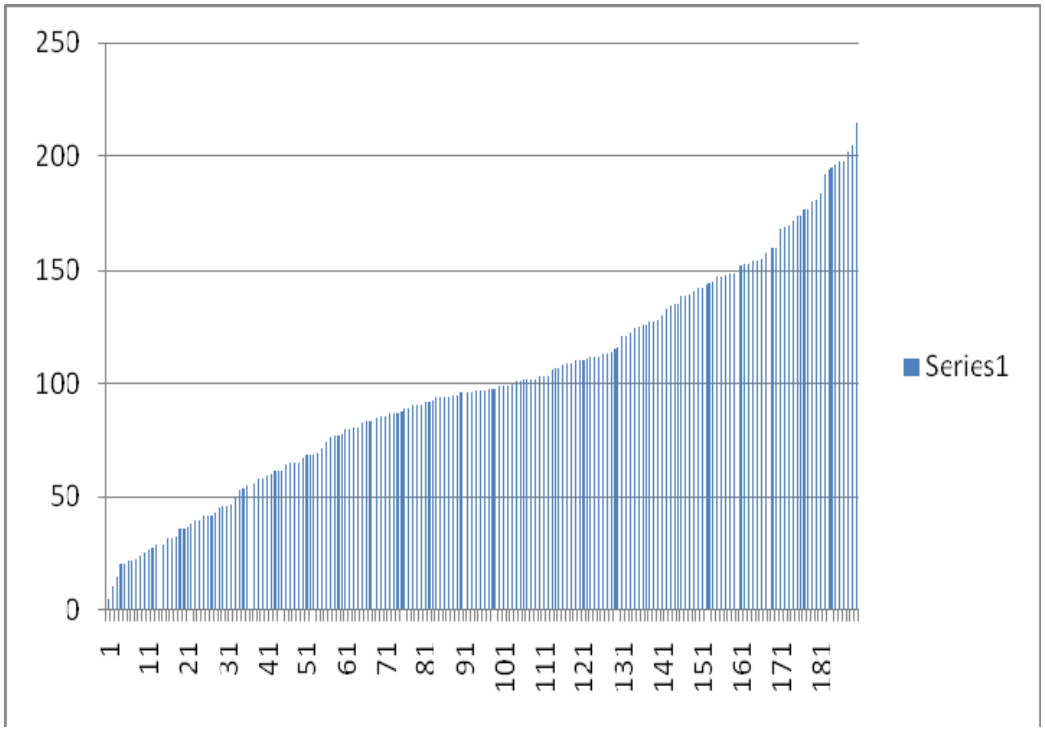

Figure 1. Frequency of Euclidean Distance of Amino Acids

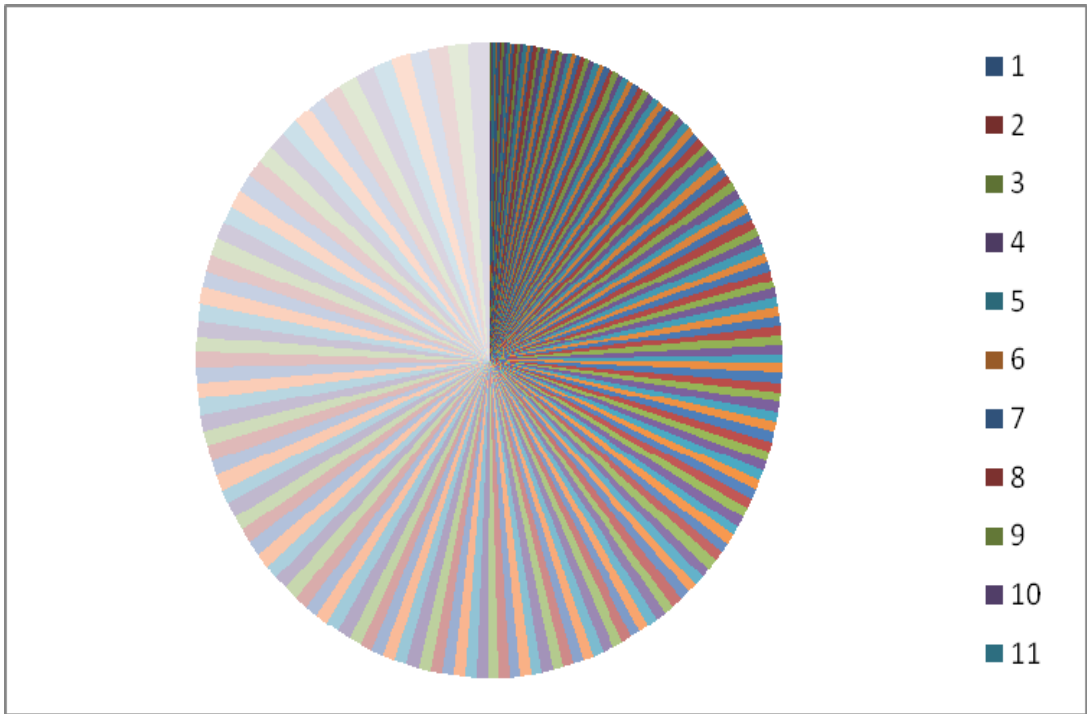

Figure 2. Circular Chart of Euclidean Distance Amino Acids

Furthermore we found that distances of amino acids that the equality of triangle inequality occur at

D(S, K) = D(S, H) + D(H, K),
(121 = 89 + 32)

D(S, K) = D(S, Q) + D(Q, K),
(121 = 68 + 53)

or

D(Ser, Lys) = D(Ser, His) + D(His, Lys),
(121 = 89 + 32)

D(Ser, Lys) = D(Ser, Gln) + D(Gln, Lys),
(121 = 68 + 53)

with an equal sum of 121.

All other distances of three amino acids do not form equality of triangle inequality.

The amino acid distance 121 between **Ser** and **Lys** is located at the centroid position (20/3, 20/3) of lower triangle (0, 0), (20, 0), (0, 20) of amino acid matrix. The centroid of a rigid triangular object is its center of mass: the object can be balanced on its centroid in a uniform gravitational field. The centroid cuts every median in the ratio 2:1, i.e. the distance between a vertex and the centroid is twice the distance between the centroid and the midpoint of the opposite side.

It appears that three amino acids Ser-His-Lys and Ser-Gln-Lys form a pair of interesting tripeptides SHK and SQK.

Our study showed a close relation between amino acid distance and symmetric matrix through a Eucleadian distance. It is hoped that these relationships will help us further explore the protein evolution. Life is based on a repertoire of structured and interrelated molecular building blocks that are shared and passed around. The same and related molecular structures and mechanisms show up repeatedly in the genome of a single species and a cross a very wide spectrum of divergent species. The matrices are storages of digital data. The matrices appear in various dimensions with different shapes. Many literatures on mathematics and biological systems have merged in recent years [5, 6] to further advance our understanding of life and its evolutions. Mathematical rules, physics laws, chemical properties, biological structures and functionalities and environmental impact are the govern bodies of living and nonliving worlds.

Reference

1. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution, Science, 1974, 185; 862.
2. Giolio M D., Capobinanco M, Medugno M. On the Optimization of the Physicochemical Distances Between Amino Acids in the Evolution of the Genetic Code. J. Theor. Biol, 1994, 168: 43.
3. Davydov, O. V. Amino Acid Contribution to the Genetic Code Structure: End-atom chemical rules of doublet composition, J. Theor Biology, 1998, 193: 679-690.
4. Bapat, R.B., Raghavan, T.E.S., *Nonnegative Matrices and Applications*, Cambridge University Press, 1997.
5. Percus, J., Mathematics of Genome Analysis, Cambridge U. Press, New York, 2002.
6. Pevzner, P. Computational Molecular Biology, MIT Press, Cambridge, 2000.
7. Yang, C.M. Chemistry and the 28-Gon Polyhedral Symmetry of the Genetic Code, J. Of Symmetrion, Vol. 12, No. 3-4, 2001, 331-347