Codon-Based Distance Matrix using a Modified Empirical Codon Mutation Matrix

Attiya Mahmood and Werner Henkel

Transmission Systems Group (TrSyS)
Jacobs University Bremen
Germany

DNA from a Channel-Coding Perspective

Motivation...

- If there are code properties, what is the underlying "channel" that requires protection?
- Take into account natural selection pressure, favoring some mutations, suppressing others.
- What is the "modulation alphabet"?
- Understand the mapping of 64 codons to 20 amino acids.

Codon Usage Table*

First Position	Second Position: U	Second Position: C	Second Position: A	Second Position: G	Third Position
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop (och)	UGA Stop (opal)	A
	UUG Leu	UCG Ser	UAG Stop (amb)	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val (Met)	GCG Ala	GAG Glu	GGG Gly	G

* Bases are given as ribonucleotides. GUG usually codes for valine, but it can code for methionine to initiate an mRNA chain. Stop (och) refers to the ochre termination triplet and Stop (amb) refers to the amber.

DNA from a Channel-Coding Perspective

Problem: Correct mutation transition probabilities are difficult to achieve, since...

- evolution will suppress unfavorable mutation, hence, they do not become visible now;
- repair mechanisms fix some of the mutational changes.

Contents

- Mechanistic vs. Empirical Models
- Point Accepted Mutations (PAM) Matrix
- Empirical Codon Mutation (ECM) Matrix
- Chemical Distance Matrix
- Codon-based Distance Matrix
- Multidimensional Scaling
- Taylor Classification
- 2D and 3D plots of Codon-based Distance Matrix

Mechanistic vs. Empirical Models

Evolutionary models are generally used for

- Sequence alignment
- Phylo-genetic tree reconstruction
- Database searches

Two kinds of Markov evolutionary models to describe protein sequence evolution:

- Mechanistic Models
- Empirical Models

Mechanistic models take into account features of protein evolution such as selective pressure and consider biological factors that shape protein evolution.
Empirical models attempt to summarize the substitution patterns observed from large quantities of data.

Point Accepted Mutations (PAM) Matrix

First empirically-based probabilistic model of amino acid substitution 1978 by Dayhoff et al.

- A Markov model of protein sequence evolution, which estimated the accepted mutations between closely related proteins from 34 super-families grouped into 71 evolutionary trees.
- To suppress the effect of a dominant occurrence frequency, the relative mutability of each amino acid is then computed, which is the ratio of modified amino acids relative to the overall number of appearance of this amino acid.

O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, "A model of evolutionary change in proteins," Natl Biomedical Research Foundation, Washington, D.C., vol. 5, pp. 345-352, 1978.

Point Accepted Mutations (PAM) Matrix

The non-diagonal elements of the PAM matrix are expressed as

$$M_{ij} = \frac{\lambda m_i A_{ij}}{\sum_i A_{im}}$$

where

M_{ij}	Mutation probability of each amino acid
λ	Proportionality constant
A_{ij}	Element of accepted point mutation matrix
m_i	Mutability of the ith amino acid

The diagonal elements have the values $M_{jj} = 1 - \lambda m_i$.
Any K-evolutionary distance matrix can be generated by taking the Kth power of the 1-PAM matrix.

With higher order of evolutionary distances, the matrix contains less information.

Point Accepted Mutations (PAM) Matrix

		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

REPLACEMENT AMINO ACID

Empirical Codon Mutation (ECM) Matrix

Proposed by Schneider et al. in 2005.
17,502 alignments of orthologous sequences from five vertebrate genomes leading to 8.3 million aligned codons for counting the mutations.
Effective for modeling the evolutionary changes, because they give the transversion/transition bias which is invisible at the amino acid level.
Widely used for finding ancestral DNA sequences.
64×64 matrix describing the substitutions between all the codons.
Substitutions from sense codons to stop codons not considered, resulting in block diagonal matrices consisting of 61x61 entries for the sense codons and 3×3 entries for the stop codons.

Empirical Codon Mutation (ECM) Matrix

AAA 0.40849 0.01506 0.26198 0.01904 0.01527 0.00650 0.01149 0.00774
0.09164 0.00886 0.06529 0.01076 0.00660 0.00143 0.00546 0.00199 0.03077
0.00710 0.01811 0.00826 0.00398 0.00212 0.00365 0.00235 0.04568 0.02826
0.03891 0.02978 0.00271 0.00126 0.00142 0.00159 0.01551 0.00399 0.00889
0.00437 0.00663 0.00287 0.00496 0.00357 0.00548 0.00272 0.00424 0.00339
0.00423 0.00159 0.00221 0.00206 0.00000 0.00093 0.00000 0.00119 0.00521
0.00275 0.00405 0.00343 0.00000 0.00141 0.00127 0.00163 0.00240 0.00044
0.00188 0.00054
AAC 0.01100 0.41485 0.00959 0.25289 0.00916 0.01865 0.00960 0.01375
0.00594 0.06004 0.00586 0.04124 0.00198 0.00335 0.00212 0.00227 0.00651
0.02338 0.00584 0.01640 0.00191 0.00320 0.00236 0.00226 0.00323 0.00648
0.00382 0.00479 0.00094 0.00154 0.00076 0.00105 0.00487 0.02458 0.00468
0.01611 0.00379 0.00645 0.00491 0.00461 0.00562 0.01425 0.00697 0.01034
0.00173 0.00277 0.00133 0.00205 0.00000 0.00490 0.00000 0.00365 0.00484
0.00759 0.00499 0.00589 0.00000 0.00411 0.00048 0.00314 0.00098 0.00121
0.00101 0.00096
...

Chemical Distance Matrix

- Relating amino acids by identifying chemical factors
- Estimation of chemical differences between amino acids regarding three chemical properties: composition, polarity, and molecular volume, which correlate to residual substitution frequencies
- The 3 properties define axes in Euclidean space, leading to distances D_{ij} between the ith and the jth amino acid.

Chemical Distance Matrix

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	Thr
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	100	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	215	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

Table 2. Difference D for each amino acid pair (*10*). The mean chemical distance from the three-property formula (see text) $\bar{D}_{cpv} = 100$ (D_{ij} values have been multiplied by 50.723 to make this mean possible). Linear regression of *RSF* and log *RSF* on these D values gives correlation coefficients of −.66 and −.72, respectively. Previous difference indexes give correlation coefficients against *RSF* of −.34 (minimum base changes), −.42 (Sneath difference), and −.49 (Epstein formula). In each case, correlation is between the two sets (difference and *RSF*) of 190 values (*3*, *4*, *7*).

R. Grantham, "Amino Acid Difference Formula to Help Explain Protein Evolution," Science, vol. 185.

Distances from ECM Matrix

- Assuming the error process to be Gaussian in nature.
- Pairwise error probability (PEP) assuming some constant standard deviation

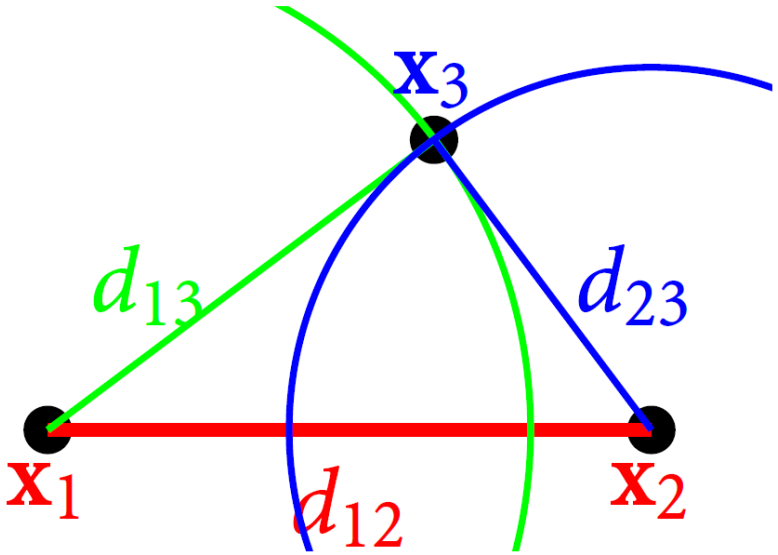
$$P_{ij} = \frac{1}{2}\text{erfc}\left(\frac{D_{ij}}{\sqrt{2}\sigma_{ij}}\right) \Longrightarrow D_{ij} = \sqrt{2}\sigma_{ij} \cdot \text{inverfc}(2P_{ij})$$

P_{ij}	Mutation probability
D_{ij}	Euclidean distance
σ_{ij}	Standard deviation

Multidimensional Scaling

For illustration, the 64×64 codon distance matrix needs to be projected to 2 or 3 dimensions.
Classical Multidimensional Scaling (CMD) is a suitable tool based on eigenvalue decomposition reducing the problem to the major eigenvalues.
CMD transfers an $n \times n$ distance matrix into n points in p-dimensional space.

Multidimensional Scaling



- Sources:
  1. Dietmar Maringer, "Datenanalyse," Uni Basel, 2010
  2. Andreas Handl, *Multivariate Verfahren*, Springer, 2002

Multidimensional Scaling

$$d_{ij}^2 = \sum_{m=1}^{p} (x_{im} - x_{jm})^2 = \underbrace{\sum_{m=1}^{p} x_{im}^2}_{b_{ii}=\mathbf{x}_i\mathbf{x}_i'} + \underbrace{\sum_{m=1}^{p} x_{jm}^2}_{b_{jj}=\mathbf{x}_j\mathbf{x}_j'} - \underbrace{\sum_{m=1}^{p} (x_{im}x_{jm})}_{b_{ij}=\mathbf{x}_i\mathbf{x}_j'}$$

with

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & & & \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} = \mathbf{X}\mathbf{X}'$$

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}' = \mathbf{Y}\mathbf{Y}'$$

Multidimensional Scaling

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij} \Longrightarrow b_{ij} = -\frac{1}{2}(d_{ij}^2 - b_{ii} - b_{jj})$$

Handl, pp. 154-171:
With $\mathbf{A} = [a_{ij}]$ and $a_{ij} = -\frac{1}{2}d_{ij}^2$

$$b_{ij} = a_{ij} - \frac{1}{n}\sum_{k=1}^{n} a_{ik} - \frac{1}{n}\sum_{k=1}^{n} a_{kj} + +\frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n} a_{km}$$

Using only the 2 (or 3) biggest eigenvalues in

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

together with its, e.g., two eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$, one obtains

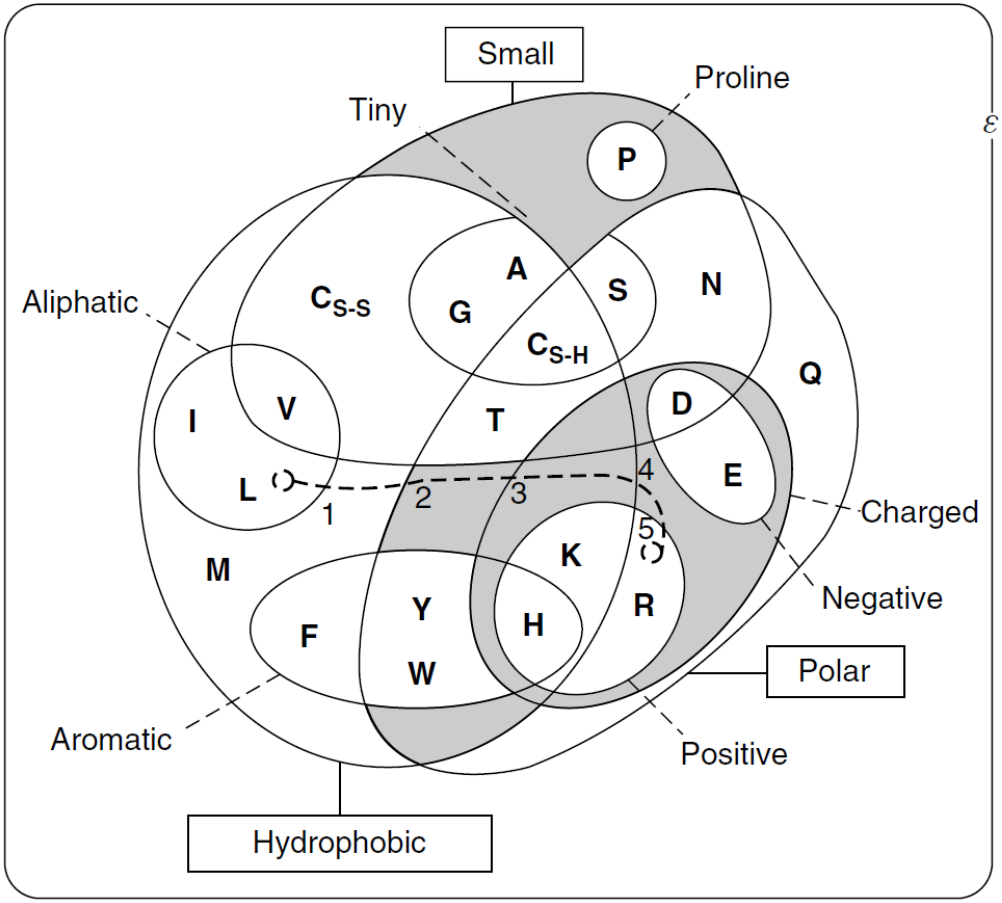
$$\mathbf{Y} = \mathbf{U}_1\mathbf{\Lambda}_1^{1/2} \text{ with } \mathbf{\Lambda}_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \text{ and } \mathbf{U}_1 = [u_1\, u_2]$$

<Matlab:CMDSCALE>

Taylor Classification

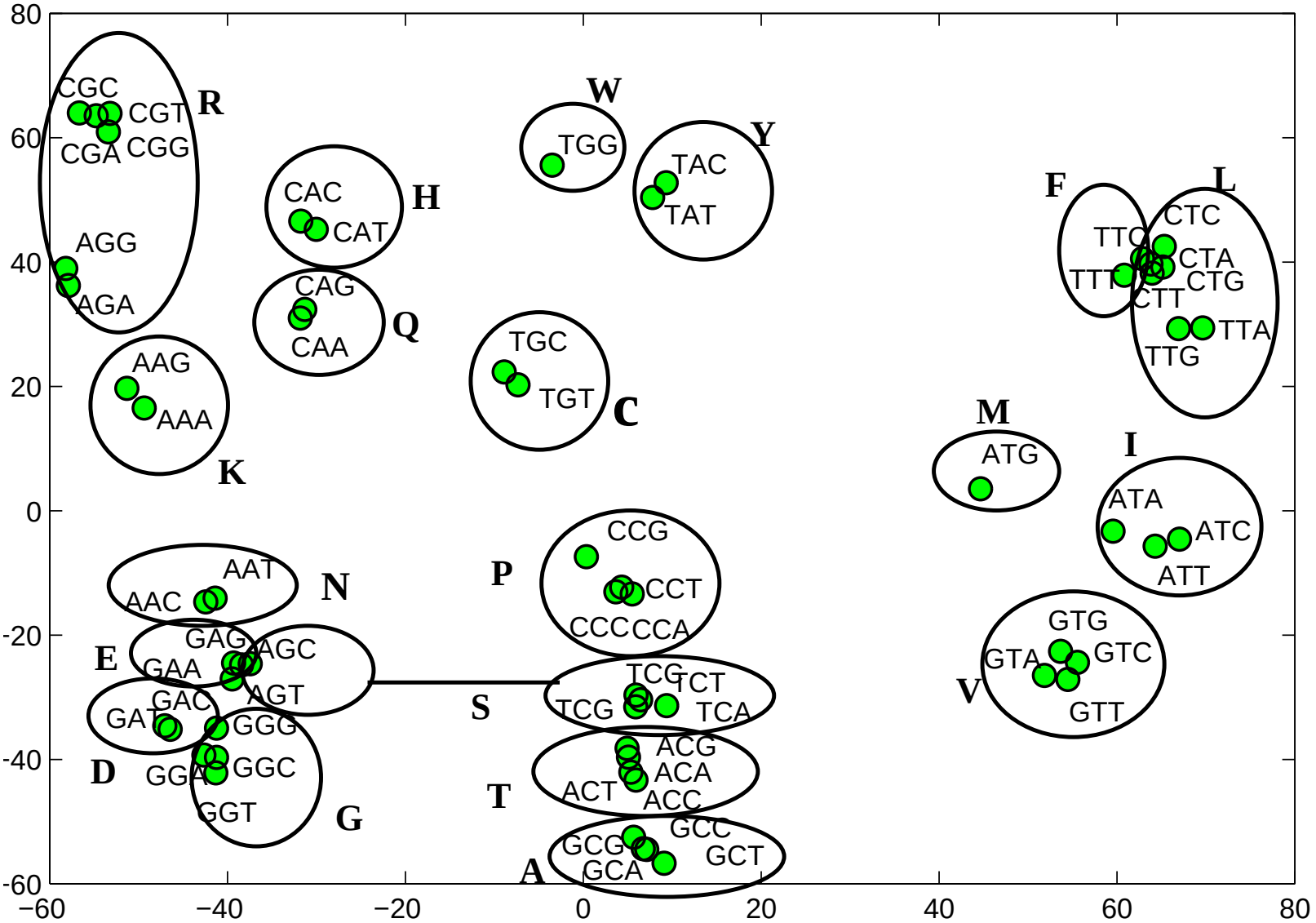
Classification of amino acids based on a synthesis of physico-chemical properties such as hydrophobicity, hydrophilic, and size. Provides grouping of amino acids in a Venn diagram:

Hydrophilic:	A P G T S
	D E Q N
Basic:	K R H
Aromatic:	W Y F
Aliphatic:	M L I V
Sulfhydryl:	C



http://123bioinformatics.com/...

2D-View of Codon Distance Matrix



Differences of ECM Mutation Matrix with Chemical Distance Matrix

The codons which belong to amino acid **'S'(AGC and AGT)** overlap with codons of amino acid **'E'**.

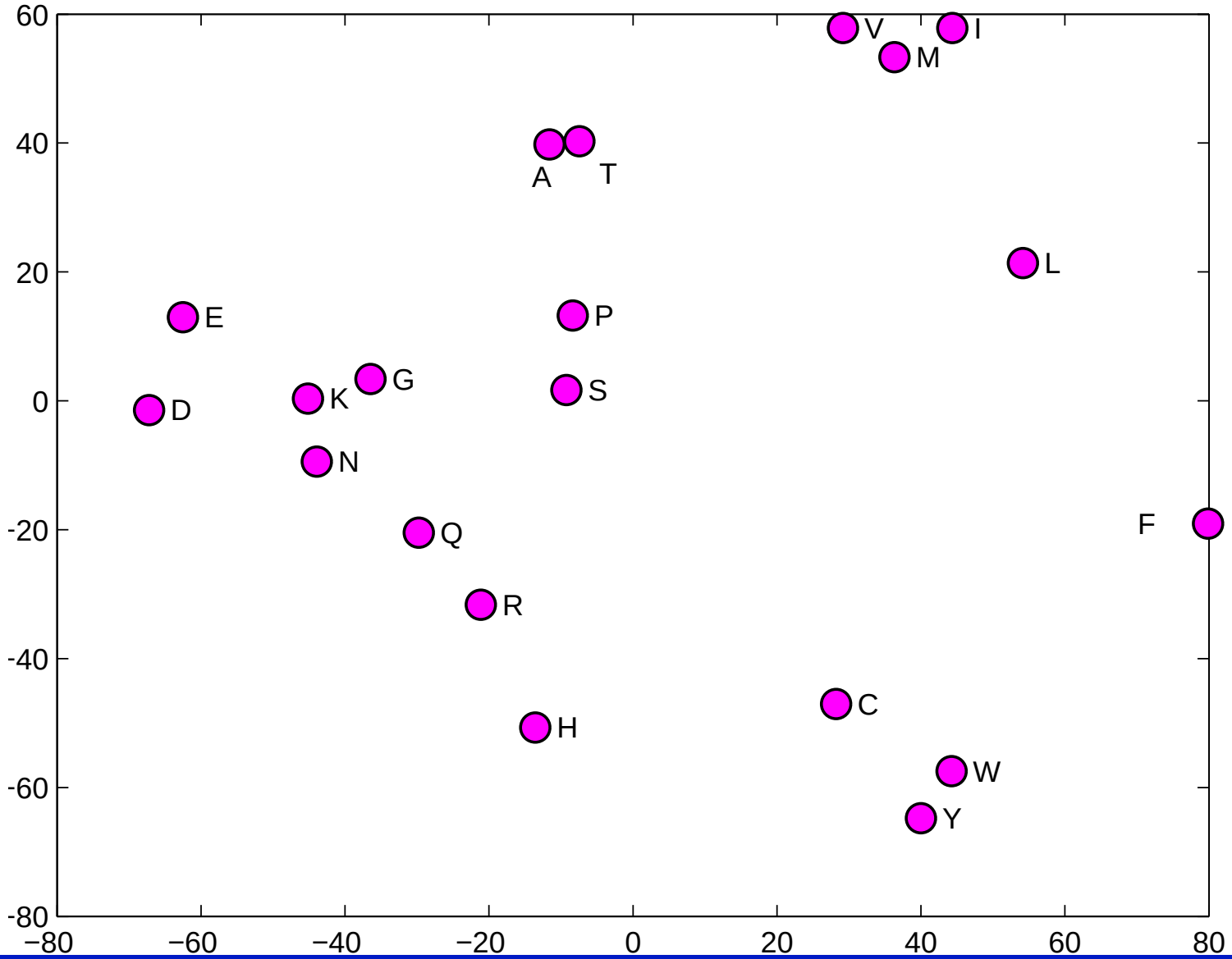
Larger chemical distance but smaller mutation distance

- **C** with **all others**
- **G** with **E**
- **S** with {**P,T,A**}
- {**D,N**} with **E**
- {**D,N**} with **G**
- {**Q,H**} with {**W,Y**}
- **K** with **N**

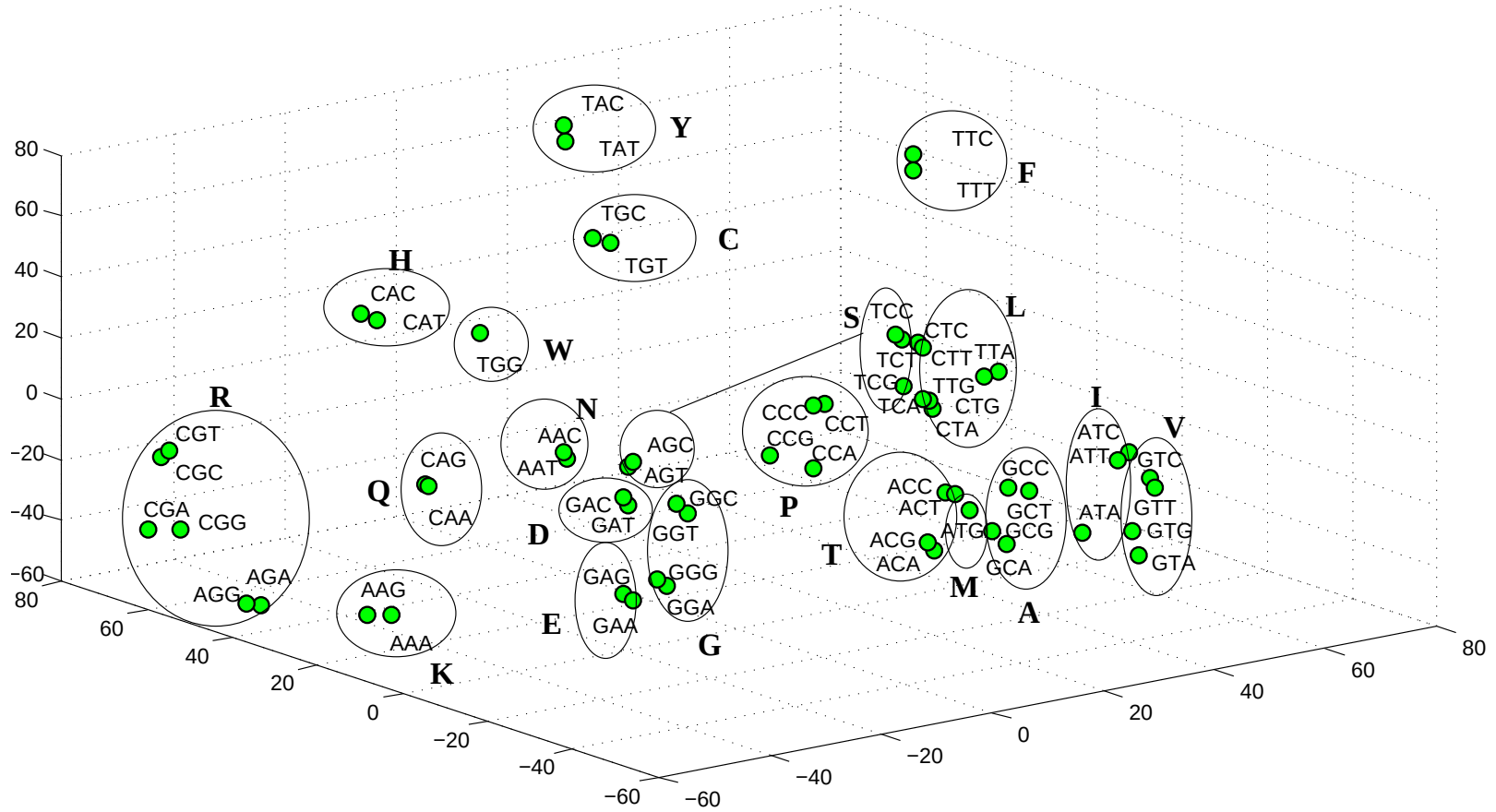
Smaller chemical distance but larger mutation distance

- {**W,Y**} with {**F,L,M,I,V**}
- {**P,T,A**} with {**Q,H,R**}

Amino Acids Mutation Distances



3D-View of Codon Distance Matrix



What is this all about?