

Protein alphabet reduction based on frequency dictionaries

Misha Kudryashev^{1,4}

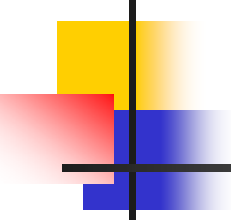
with Tatyana Popova^{2,4} and Alexander Gorban^{3,4}

1. University of Heidelberg, Germany

2. Boeringer GMBH, Germany

3. University of Leicester, UK

4. Institute of Computational Modeling SB RAS, Russia



Plan

- From reality to schemes: the problem statement
- Optimal classification of symbols
- Natural language example
- Optimal amino acids classifications for various classes of proteins, comparisons to functional classifications
- What next?

Studied object

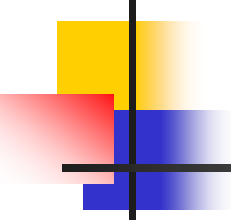


Protein sequence

...ILTGLFLAMHYTSDISTAFSSVI...

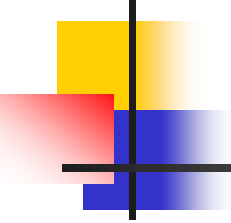
Amino acid alphabet

{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}



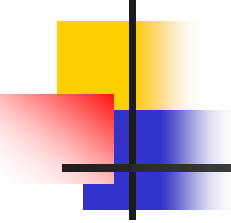
What are proteins made from?

- Amino acids (AAs)?
- Short sequences of AAs?
- Classes of equivalent AAs?
- Short sequences of such classes?
- Anything else?



More questions

- M. Gromov asked: is there a syntactic difference between *Globular* and *Membrane* proteins?
- Are proteins random sequences of amino acids or they have any order?



How to discover non-randomness in protein sequences?

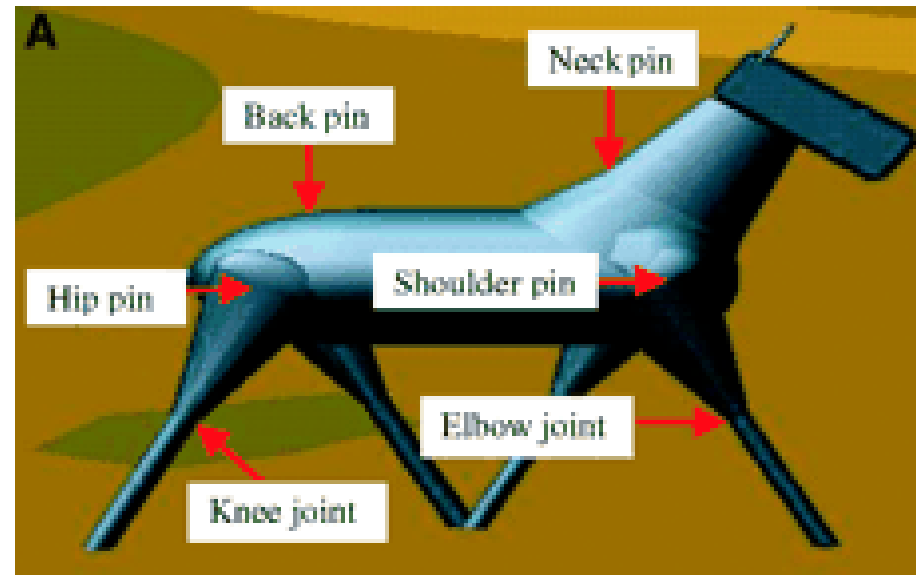
- A string of length 400 in 20-letters alphabet is too short for non-randomness tests;
- We can reduce the alphabet and analyse protein sequences in reduced alphabet.
- With reduced alphabet we should be able to see patterns, schemes in the protein sequence.

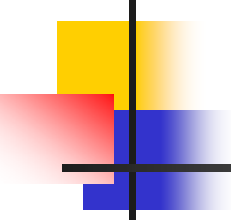
How can we extract scheme from reality?



Functions give us ideas and hints for this extraction

Another source of ideas: let us analyse ensembles and extract **non-random features**



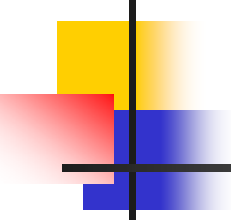


Backgrounds of amino acids classification

The bases of theoretical grouping of amino acids mentioned in literature may be attributed to the following main features:

- physical, chemical properties and amino acids environment in proteins;
- protein alignments and substitution matrices;
- protein spatial structure and contact potential matrix...

Some natural amino acids binary classifications



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
HP	1	1	0	0	1	1	0	1	0	1	1	0	1	0	0	0	0/1	1	1	0/1
B/S	0	0	0	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	1	1
C/U	0	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Optimal informational classification

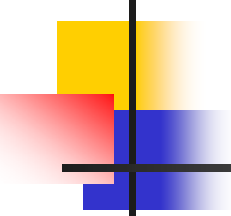
Classification is a map:

$$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \xrightarrow{\varphi} \{1, \dots, k\}$$

We associate with the transformed text a set of objects with some frequency distribution. Optimal informational classification provides maximal relative entropy (information) of distribution of recorded objects:

$$(1) \quad \varphi: \quad D(P | P^*) = \sum_{X_\varphi} P(X_\varphi) \ln \frac{P(X_\varphi)}{P^*(X_\varphi)} \rightarrow \max,$$

where P is real distribution, and P^* is some reference ("random") distribution. That is, φ is the "most non-random" classification.



Frequency dictionary

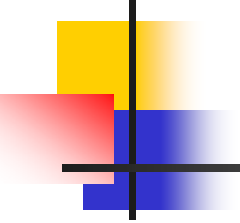
Let X_φ be a " q -letter word ensemble." Then $P(X_\varphi)$ is the q -th frequency dictionary for a text: it is a function that associates with each string of letters

$$i_1 i_2 \dots i_q$$

its frequency in the text

$$f_{i_1 i_2 \dots i_q}$$

it is a kq - dimensional real vector,
where k is the number of letters in the alphabet.



Where do we take the reference distribution?

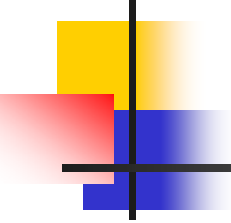
This is the most random distribution MaxEnt (the physical entropy, maximal randomness) for given data.

For example, for given frequencies of symbols,

$$f_{i_1 \dots i_q}^* = f_{i_1} \cdot f_{i_2} \cdot \dots \cdot f_{i_q} ,$$

where $i_1 \dots i_q$ are q -letter words;

$f_{i_1 \dots i_q}$ are frequencies of corresponding words in the symbol sequence.



So, we have a problem:

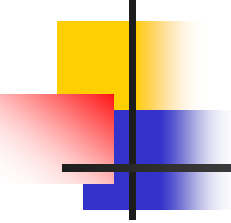
For word distribution in reduced alphabet

$$\sum_{i_1 i_2 \dots i_q} f_{i_1 i_2 \dots i_q} \ln \frac{f_{i_1 i_2 \dots i_q}}{f_{i_1 i_2 \dots i_q}^*} \rightarrow \max$$

where $f_{i_1 \dots i_q}^* = f_{i_1} \cdot f_{i_2} \cdot \dots \cdot f_{i_q}$

Entropic classification of letters for English language in Bible text

Relative Entropy	Groups					
	1	2	3	4	5	6
0.767926	a <u>e</u> ioudgt	bcfhjklm <u>n</u> pqrstvwxyz				
0.934107	a <u>e</u> iou	bcd fgklmnpqrst <u>t</u> vwxyz	<u>h</u> j			
1.096432	a <u>e</u> iou	bcfklm <u>n</u> pqrstvxz	<u>h</u> j	dg <u>t</u> wy		
1.171895	a <u>e</u> iou	bcfklmpr <u>s</u> vxyz	<u>h</u> j	dgq <u>t</u> w	<u>n</u>	
1.227138	a <u>e</u> iou	bcfklmpqr <u>s</u> vxyz	<u>h</u> j	<u>t</u> w	<u>n</u>	<u>d</u> g

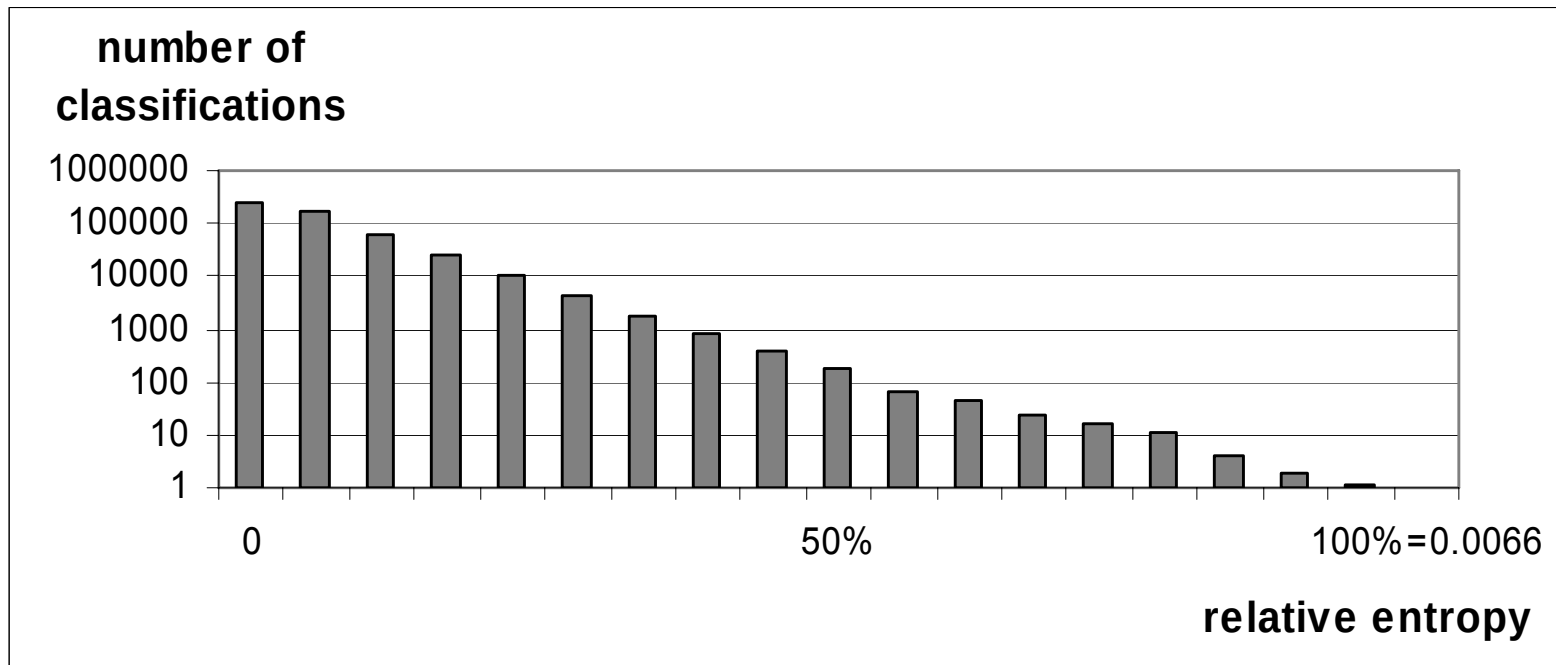


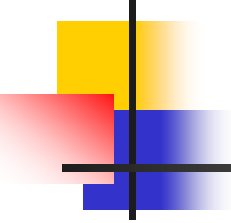
In the beginning was the Word, and the Word was with God, and the Word was God (Jn. 1:1-3)

Number of classes	The coded phrase
2	01 110 101011011 101 110 1011 011 110 1011 101 1011 101 011 110 1011 101 101 En nne nenennenn nen nne Nenn, enn nne Nenn nen nenn Nen, enn nne Nenn nen Nen
3	Et the tetettett tet the Teth, eth the Teth tet teth Teh, eth the Teth tet Teh
4	En the netennent ten the Tent, ent the Tent ten teth Tet, ent the Tent ten Tet
5	En the setennent tes the Test, ent the Test tes teth Set, ent the Test tes Set
6	En the setennend tes the Tesd, end the Tesd tes teth Ded, end the Tesd tes Ded
Initial phrase	In the beginning was the Word, and the Word was with God, and the Word was God

Informational classification finds syllabic structure.

Typical distribution of relative entropy for all possible binary classifications of amino acids.

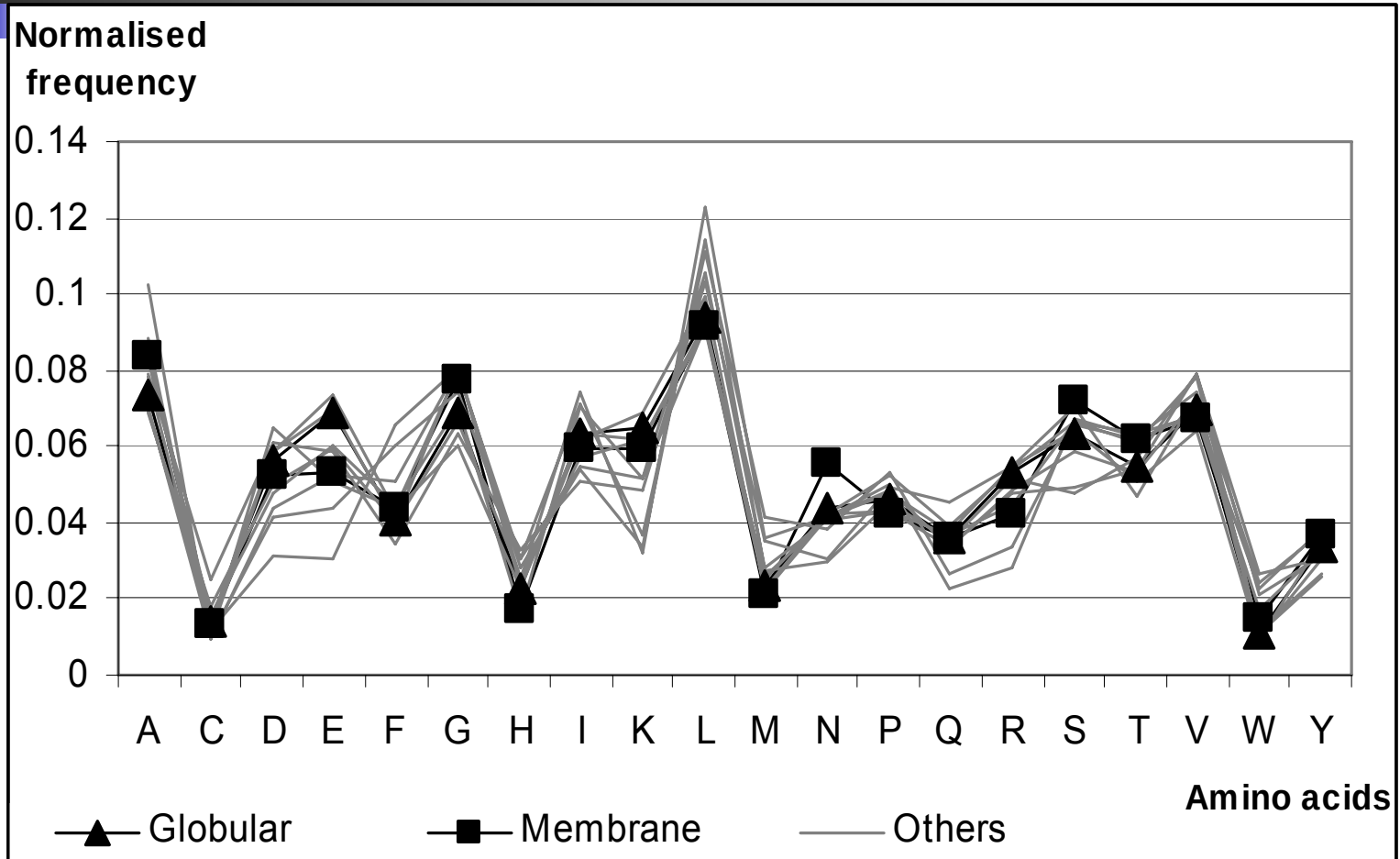




The data sets of used protein sequences

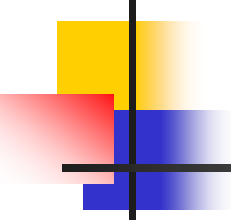
	Keywords	Number of proteins	Keywords	Number of proteins
Dataset 1 (EBI)	Oxidoreductase	452	Transferase	500
	Cytochrome	500	Isomerase	578
	Phytochrome	500	DNA polymerase	500
	Nitrato-reductase	197	Oxidase	500
			ATPase	500
Dataset 2 (SwissProt)	Membrane	10000	Globular	5019

Amino acid frequencies in considered sets of proteins



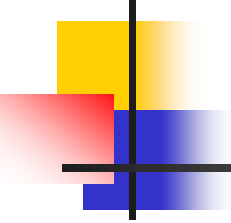
Binary informational classifications for Dataset 1 and 2

Protein dataset	$D_{\max}$	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
ATPase	0.002	1	1	0	0	1	1	1	1	0	1	1	0	0	0	0	1	1	1	1	1
Cytochrome	0.0066	1	1	0	0	1	1	1	1	0	1	1	0	0	1	0	1	1	1	1	1
Nitrato-reductase	0.0027	1	0	0	0	1	1	1	1	0	1	1	0	0	0	0	1	1	1	1	1
Oxidase	0.0029	1	1	0	0	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1	0
DNA polymerase	0.0007	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0
Isomerase	0.0006	1	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	0	1	0	0
Transferase	0.0006	1	0	1	1	0	0	0	1	1	1	1	0	0	0	1	0	0	0	0	0
Phytochrome	0.0074	1	1	1	1	0	0	1	1	1	0	1	1	0	0	0	0	0	1	0	0
Oxidoreductase	0.0024	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	1	0	1
Globular	0.0006	1	0	0	1	0	0	0	0	1	1	1	0	0	1	1	0	0	0	0	0
Membrane	0.0025	1	1	0	0	1	1	0	1	0	1	1	0	1	0	0	1	1	1	0	0



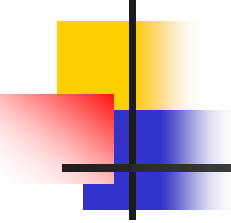
Hamming distances between various binary classifications

	HP	BS	CU	Membrane	Globular
Membrane	2	6	4	0	8
Globular	7	6	6	8	0
(Murphy et al., 2002)	3	8	2	2	6



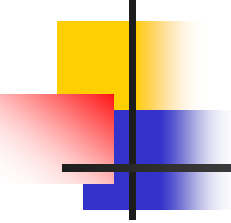
New dataset

	Keywords	Numb. of proteins	Keywords	Numb. of proteins
Dataset 3 (SwissProt)	Chloroplast	5000	Membrane	5000
	Cytoplasmic	2849	Mitochondrial	5000
	Extracellular	5000	Nuclear	5000
	Golgi	748	ATPase	5000



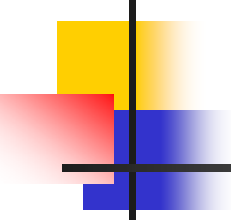
Classifications for new dataset

	$D_{\max}$	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Chloroplast	0,0016	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0
Cytoplasmic	0,0012	1	0	0	1	0	0	0	0	1	1	1	0	0	1	1	0	0	0	0	0
Extracellular	0,0013	1	0	0	1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	0	0
<u>Goldgi</u>	0,0021	1	1	0	0	1	1	1	1	0	0	0	0	1	0	0	1	1	1	1	1
<u>Membrane</u>	0,0031	1	1	0	0	1	1	0	1	0	1	1	0	1	0	0	1	1	1	1	0
Mitochondrial	0,0008	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0
Nuclear	0,0023	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	0	0
<u>ATPase</u>	0,0009	1	1	1	0	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1



Correlations of informational classifications with natural scales

	BS	CU	HP
Chloroplast	0,06	0,01	0,25
Cytoplasmic	0,10	0,30	0,26
Extracellular	0,07	0,29	0,16
Goldgi	0,18	0,06	0,52
Membrane	0,36	0,15	0,72
Mitochondrial	0,06	0,01	0,25
Nuclear	0,69	0,03	0,00
ATPase	0,21	0,38	0,58



Orthogonal classifications

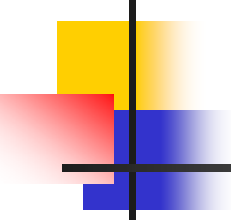
2 nd component	BS	CU	HP
Chloroplast	0,4	0,5	0,32
Cytoplasmic	0,47	0,01	0,05
Extracellular	0,24	0,06	0,1
Goldgi	0,41	0,07	0,03
Membrane	0,32	0,03	0
Mitochondrial	0,11	0,17	0,08
Nuclear	0,28	0,17	0,03
ATPase	0,52	0,07	0,15

3 rd component	BS	CU	HP
Chloroplast	0,1	0,01	0,31
Cytoplasmic	0,12	0,04	0,01
Extracellular	0,48	0,14	0,83
Goldgi	0,21	0,32	0,22
Membrane	0,46	0,23	0,22
Mitochondrial	0,45	0,06	0,27
Nuclear	0	0,28	0,28
ATPase	0,34	0,53	0,2

Frequencies in reduced alphabet

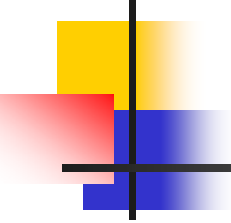
	$f(0)$	$f(1)$	$f^*(00)$	$f^*(01)$	$f^*(10)$	$f^*(11)$	$f(00)$	$f(01)$	$f(10)$	$f(11)$
Chloroplast	0,75	0,25	0,56	0,19	0,19	0,06	0,57	0,18	0,18	0,07
Cytoplasmic	0,57	0,43	0,33	0,24	0,24	0,18	0,34	0,23	0,23	0,19
Extracellular	0,53	0,47	0,28	0,25	0,25	0,22	0,29	0,24	0,24	0,23
Goldgi	0,52	0,48	0,27	0,25	0,25	0,23	0,29	0,23	0,23	0,25
Membrane	0,32	0,68	0,1	0,22	0,22	0,46	0,12	0,2	0,2	0,48
Mitochondrial	0,73	0,27	0,53	0,2	0,2	0,07	0,54	0,19	0,19	0,08
Nuclear	0,67	0,33	0,45	0,22	0,22	0,11	0,46	0,21	0,21	0,12
ATPase	0,78	0,22	0,61	0,17	0,17	0,05	0,62	0,16	0,16	0,05

■ “Block structure”



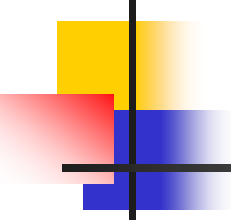
Answer 1.

There exists significant syntactic difference between Globular and Membrane proteins; hydrophobicity is responsible for non-randomness in sequences of membrane proteins.



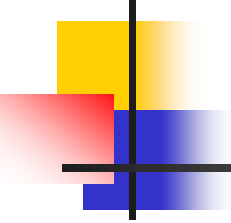
Answer 2.

Amino acid sequences in proteins
are definitely not random



Answer 3

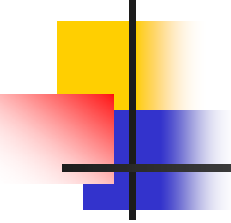
Protein sequences in reduced alphabet have "block" structure as natural texts in English have "syllabic".



Answer 4

What are proteins made from? We have new candidates for a minimal set of amino acids for each protein group.

But, perhaps, it is wiser to classify couples and triples of amino acids. Classes of such couples and triples are, perhaps, the **elementary** details of proteins.



To be continued...

Thank you for your attention!