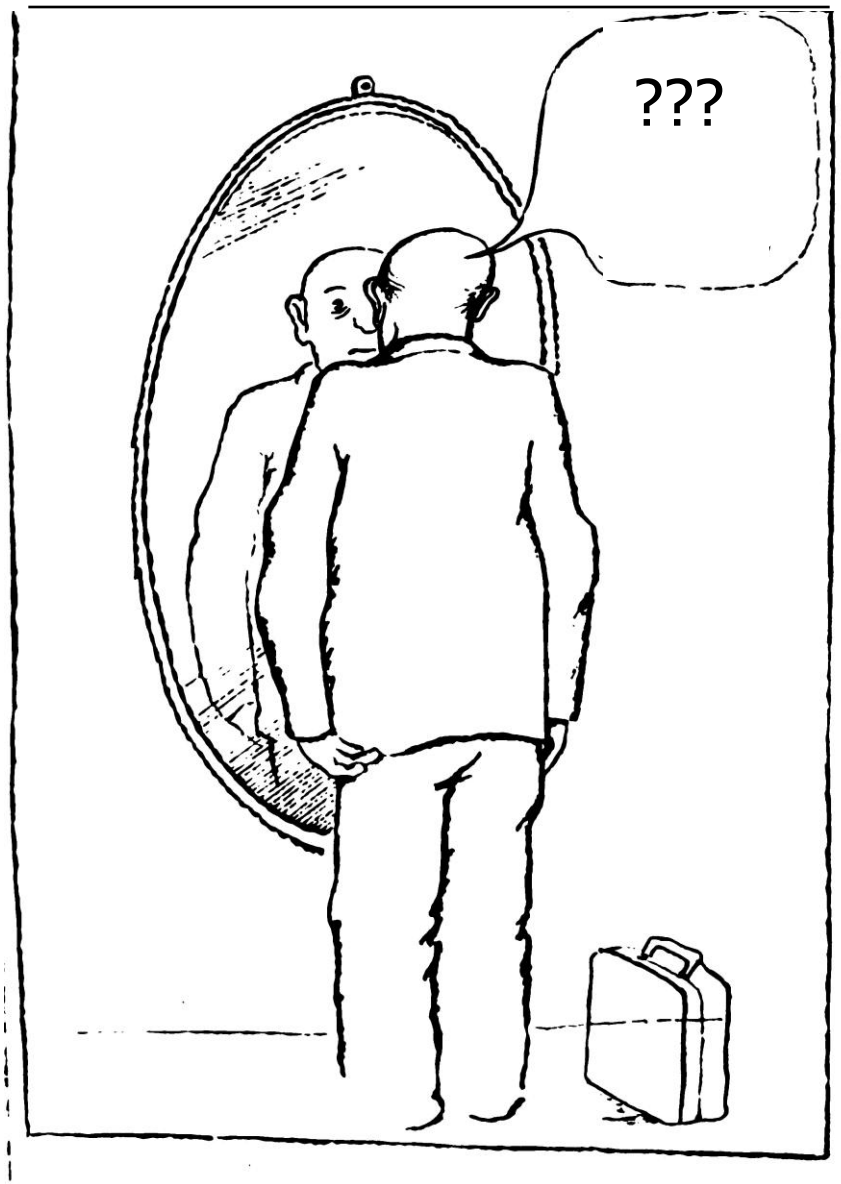
Who am I?

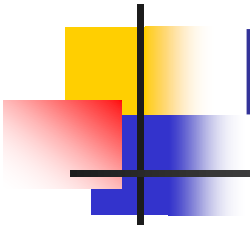


Informational disassembling of biological machines

Alexander Gorban

Department of Mathematics
University of Leicester

With T. Popova and M. Kudryashev



Plan

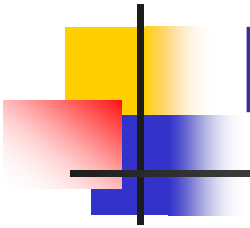
- From reality to schemes: the problem statement;
- Optimal classification of symbols
- Natural language example
- Optimal amino acids classifications for various classes of proteins, comparisons to functional classifications
- What next?

Artificial life: The problem of minimal cell

We should disassemble cell into elementary details, and after that assemble this machine again

What is the minimal set of details sufficient for life creation?

What is the minimal set of amino acids sufficient for life creation?



Minor problems ☺

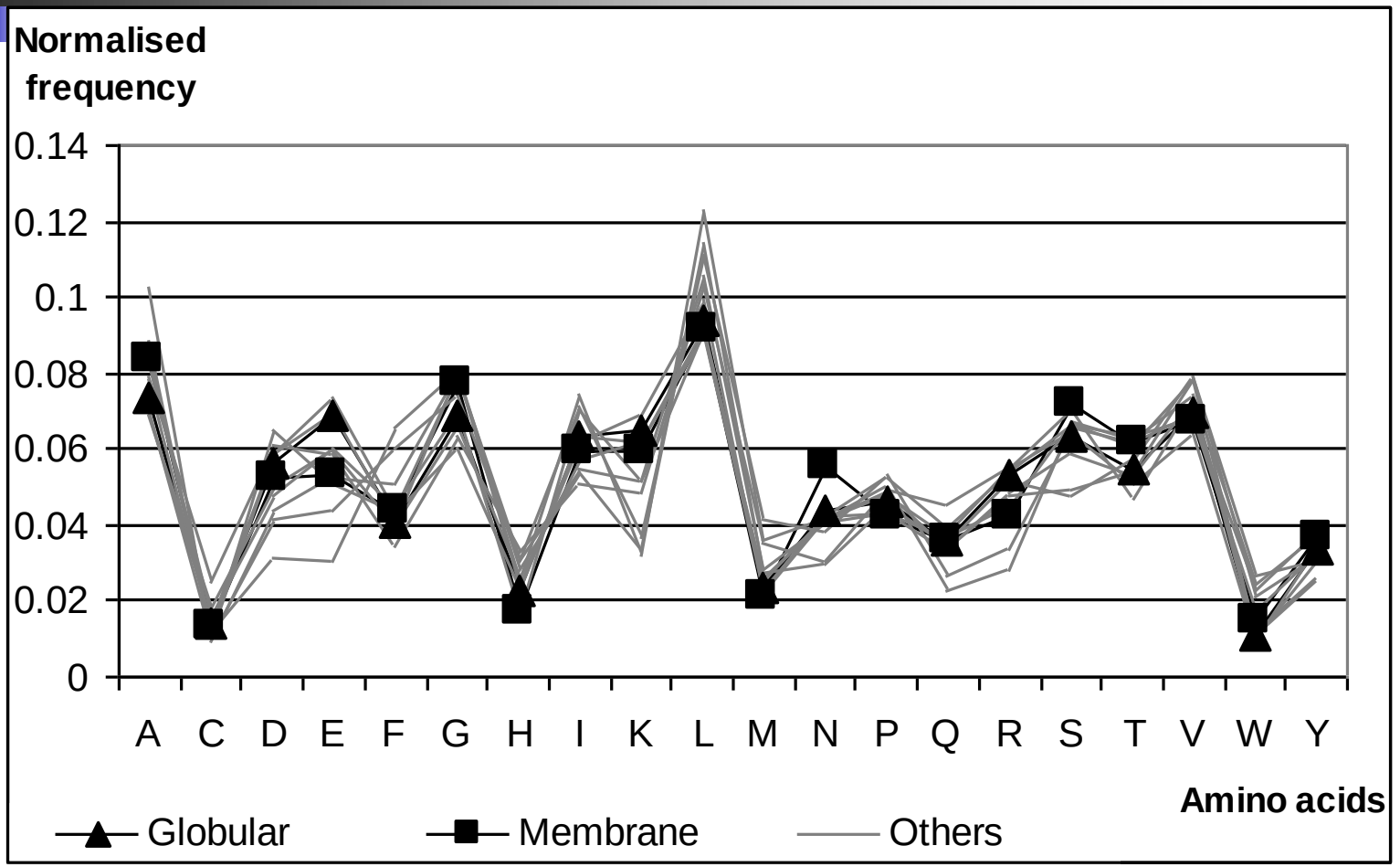
- M. Gromov asked: is there a syntactic difference between Globular and Membrane proteins?

- Are proteins random sequences of amino acids (a long discussion)?

The data sets of 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
	Nitratoreductase	197	Oxidase	500
			ATPase	500
Dataset 2 (SwissProt)	Membrane	10000	Globular	5019

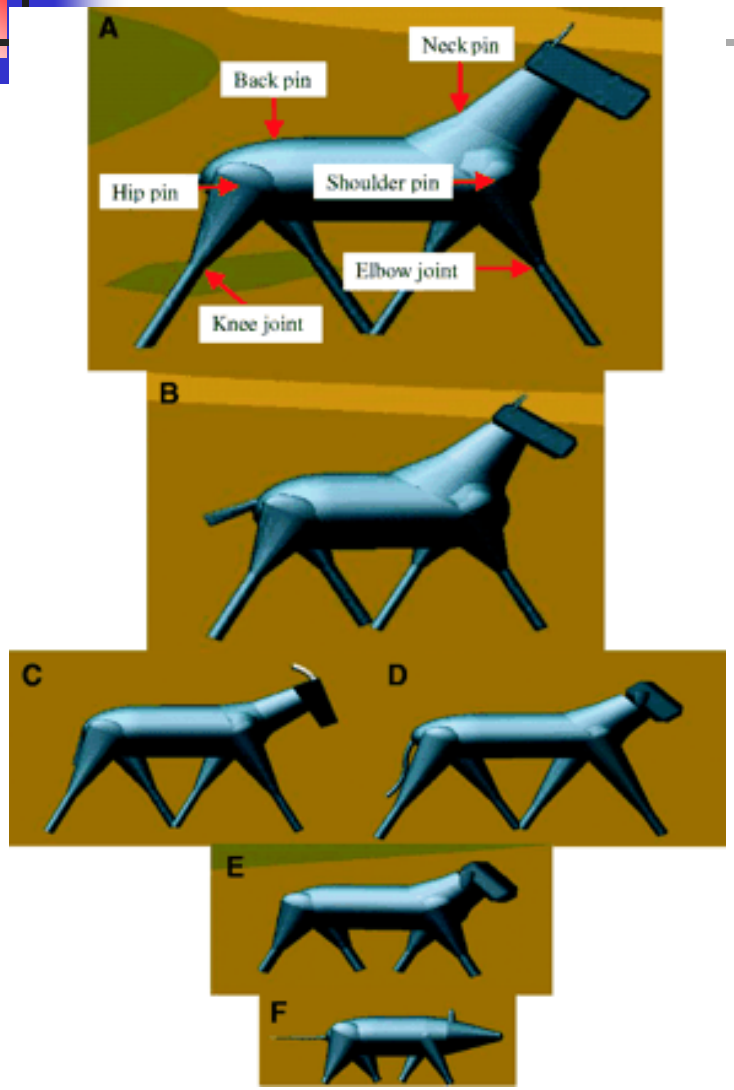
Amino acid frequencies in considered sets of proteins



Why is it difficult to discover non-randomness in protein sequences?

- A string of length 400 in 20-letters alphabet is too short for non-randomness tests;
- Even for random string of such a length we can usually classify letters and reduce alphabet to 0-1 on such a way that the resulting 0-1 string will be obviously non-random.

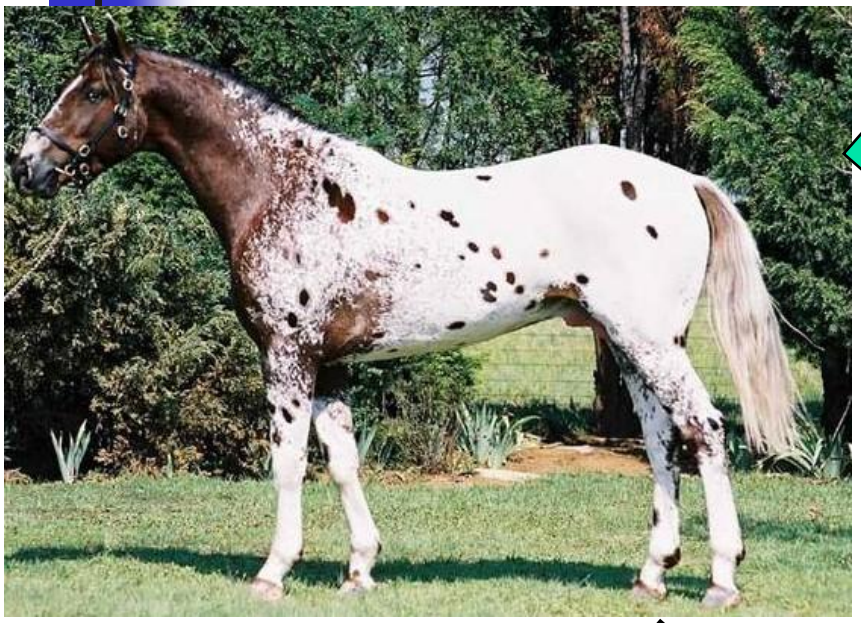
If something is a machine, it should have a scheme



Model structure: (A) large horse, (B) small horse, (C) goat, (D) large dog, (E) small dog and (F) chipmunk. Joint locations, segment dimensions and mass distributions are from photographic, video and anatomical data (Muybridge, 1957; Taylor et al., 1974; Fedak et al., 1982; Alexander, 1985; Farley et al., 1993). All segments are represented as rigid bodies. Pin (rotary) joints are included on the back and neck. Each leg rotates about a pin joint at the shoulder or hip and changes length through a prismatic (telescoping) joint at the elbow or knee. Active hip and shoulder torques control the forward motion from stride to stride. Motions are restricted to the sagittal plane.

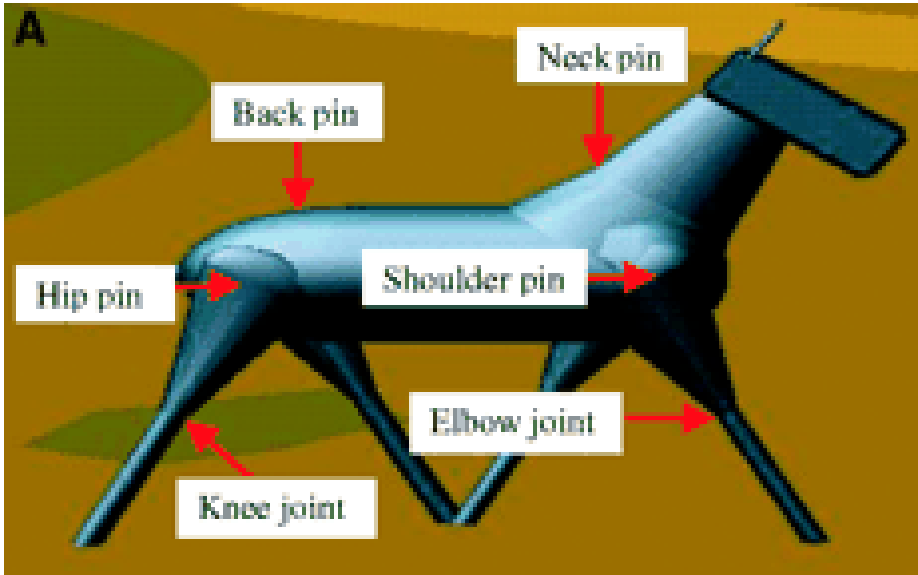
H.M. Herr, G.T. Huang, T.A. McMahon (2002)

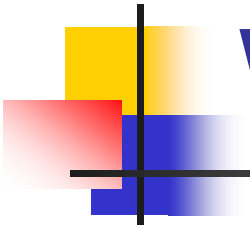
How can we extract scheme from reality?



Functions give us ideas and hints for this extraction

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





What are proteins made from?

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

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 I	1	1	0	0	1	1	0	1	0	1	1	0	1	0	0	0	0/1	1	1	0/1
HP II	1	1	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	1	1	1
HP III	1	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	0	1	1	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

Example: contact energetic classification

Let **M** be a 20·20 matrix of energies of amino acids residuals contact interactions. It appears that

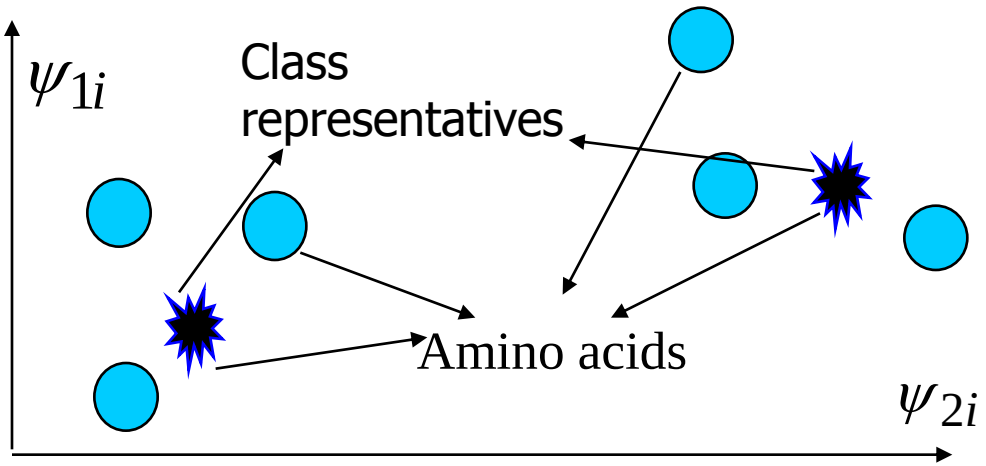
$$\mathbf{M} \approx \lambda_1 |\psi_1\rangle\langle\psi_1| + \lambda_2 |\psi_2\rangle\langle\psi_2|$$

Each amino acid can be represented by a point on the plain

$$i-\text{th amino acid} \mapsto (\psi_{1i}, \psi_{2i})$$

Hypothesis: classification of amino acids is equivalent to classification of these points

Li et al., 1997, Wang et al., 1999, Wang et al., 2000, Cieplak et al., 2001, Wang et al., 2002, Fan et al., 2003,



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, …, *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:

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

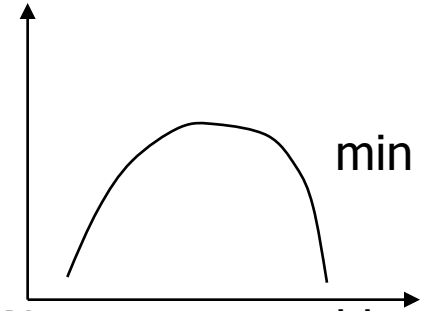
where *P* is real distribution, and *P** is some reference ("random") distribution. That is, *P* is the "most non-random" classification.

Apologies

"Relative entropy" has non-physical sign:

"Relative entropy maximum" means here maximal non-randomness. In physics, the convention about signs is opposite. In that sense, we are looking for the entropy minimum

$$S = -\sum_{X_\varphi} P(X_\varphi) \ln \frac{P(X_\varphi)}{P^*(X_\varphi)}$$



Non-convex problem in the distributions simplex

Frequency dictionary

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

$$i_1 i_2 \ldots i_q$$

its frequency in the text

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

it is a n^q – dimensional real vector,
where n is the number of letters in the alphabet.

What else X_{φ} might be?

- The frequency table of amino acid contacts in folded proteins, for example.

Where should 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,

$$f^{*}_{i_1 \ldots i_q} = f_{i_1} \cdot f_{i_2} \cdot \ldots \cdot f_{i_q} ,$$

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

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

For given q-s-letter word frequencies (for $q - s > 1$)

$$f^{*}_{i_1 \ldots i_{q-s} \ldots i_q} = \frac{f_{i_1 \ldots i_{q-s}} \cdot f_{i_2 \ldots i_{q-s+1}} \cdot \ldots \cdot f_{i_{s+1} \ldots i_q}}{f_{i_2 \ldots i_{q-s}} \cdot f_{i_3 \ldots i_{q-s+1}} \cdot \ldots \cdot f_{i_{s+1} \ldots i_{q-1}}}$$

So, we have a problem:

For word distribution in reduced alphabet

$$\sum_{i_1 i_2 .. i_q} f_{i_1 i_2 .. i_q} \ln \frac{f_{i_1 i_2 .. i_q}}{f^*_{i_1 i_2 .. i_q}} \to \max ,$$

where (the K - formula)

$$f^*_{i_1 i_2 \ldots i_q} = \frac{f_{i_1 i_2 .. i_{q-s}} f_{i_2 i_3 .. i_{q+1-s}} \cdots f_{i_{s+1} i_{s+2} .. i_q}}{f_{i_2 .. i_{q-s}} f_{i_3 .. i_{q+1-s}} \cdots f_{i_{s+1} i_{s+2} .. i_{q-1}}};$$

or

$f_{i_1} f_{i_2} \ldots f_{i_q}$ for $s = q$-1

Entropic classification of letters for English language in Bible text

Relative Entropy	Groups					
	1	2	3	4	5	6
0.767926	a**e**ioudgt	bcfhjklm**n**pqrsvwxyz				
0.934107	a**e**iou	bcdfgklmnpqrs**t**vwxyz	**h**j			
1.096432	a**e**iou	bcfklm**n**pqrsvxz	**h**j	dg**t**wy		
1.171895	a**e**iou	bcfklmpr**s**vxyz	**h**j	dgq**t**w	**n**	
1.227138	a**e**iou	bcfklmpqr**s**vxyz	**h**j	**t**w	**n**	**d**g

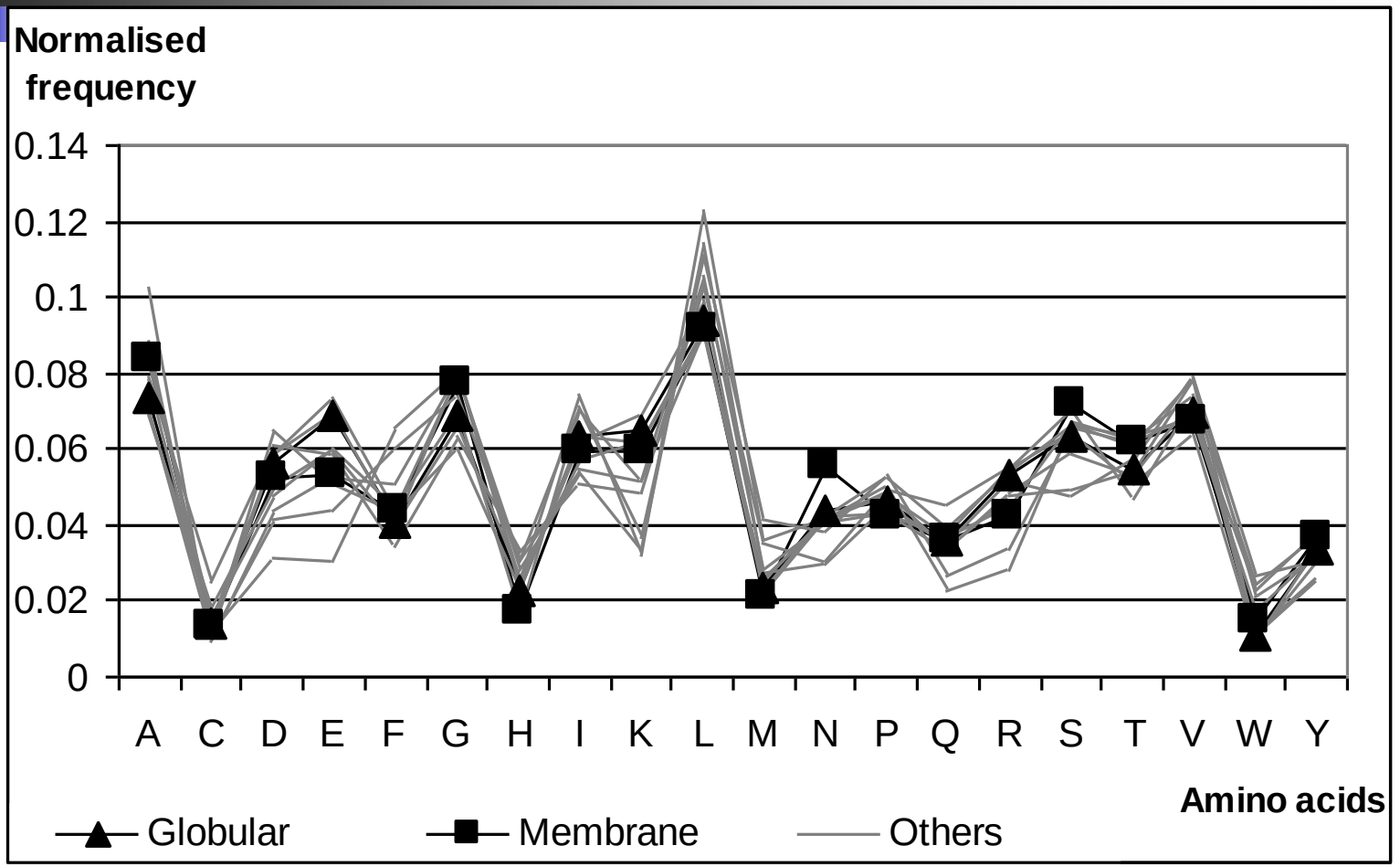
In the beginning was the Word, and the Word was with God, and the Word was God (Jn. 1:1-3)

Num-ber 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

The data sets of 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
	Nitratoreductase	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
Nitratoreductase	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

Globular vs Membrane comparison

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

---

0 0 0 0 0 0 0 1 0/1 1 1 1 1 1 1 1 1 0 0/11

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

G"or"M:

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

L-Leucine	G-Glycine	K-Lysine	D-Aspartic A.
A-Alanin	S-Serine	E-Glutamic A.	N-Asparagin

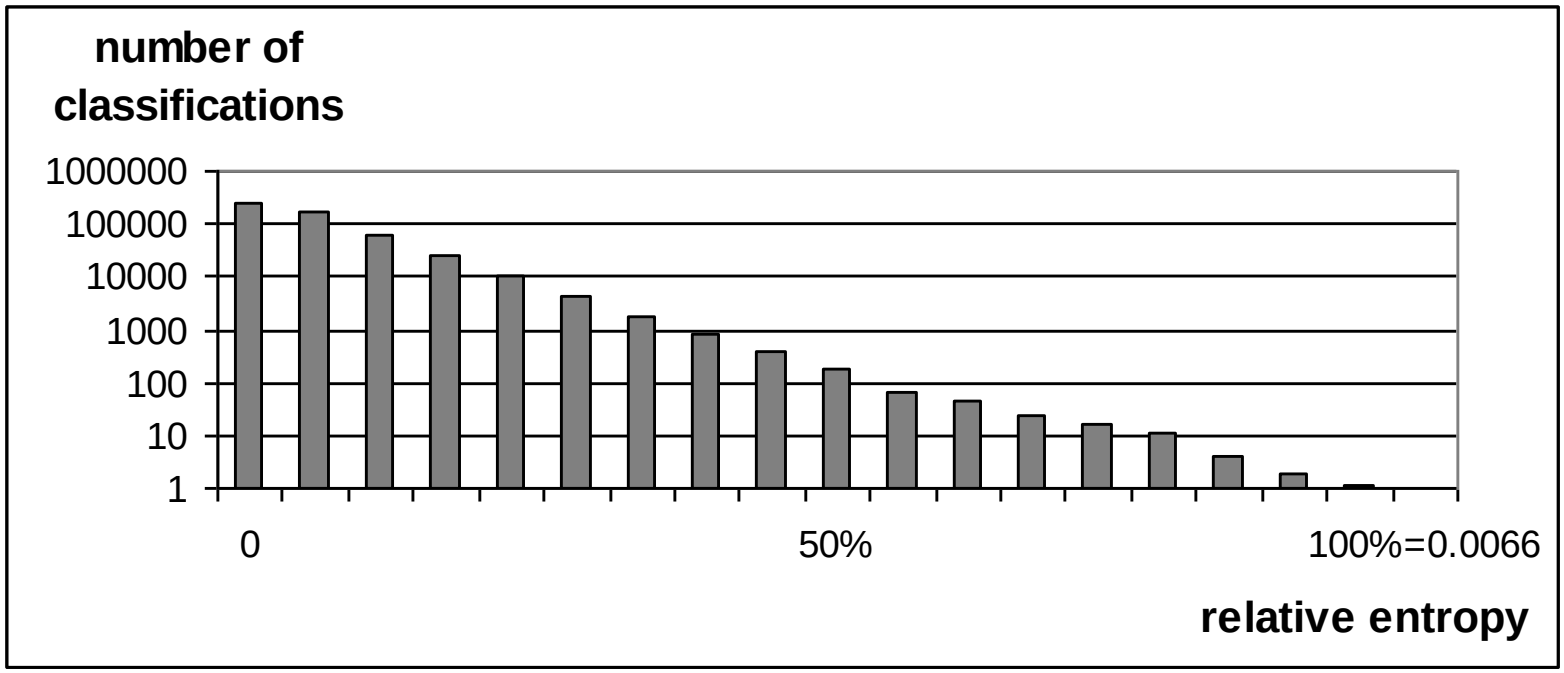
0-hydrophylic, 1-hydrophobic

W-Tryptophan, S-Serine

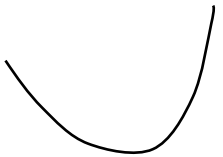
Hamming distances between various binary classifications

	HP I	HP II	HP III	BSl	CU	Membrane	Globular
Membrane	2	5	7	6	4	0	8
Globular	7	9	9	6	6	8	0
(Murphy et al., 2002)	3	3	5	8	2	2	6

Typical distribution of relative entropy for all possible binary classifications of amino acids (Cytochrome dataset)



Informational relative entropy is quadratic near minimum, and has a sharp maximum (disorder is wide, but order is sharp).

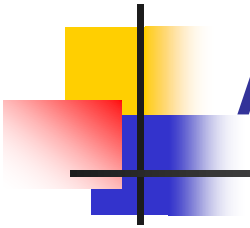


Answer 1.

New 4-class informational classification of amino acids:

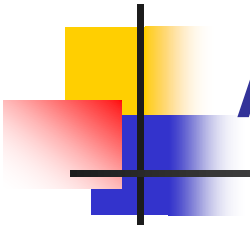
{*A*,**L**,M}U{C,F,**G**,I,P,*S*,T,V}U{*E*,**K**,Q,R}U{**D**,H,*N*,W,Y}

L-Leucine	G-Glycine	K-Lysine	D-Aspartic A.
A-Alanin	S-Serine	E-Glutamic A.	N-Asparagin



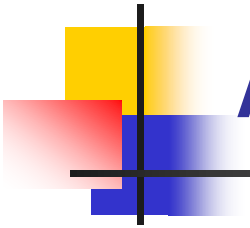
Answer 2.

There exists significant syntactic difference between Globular and Membrane proteins



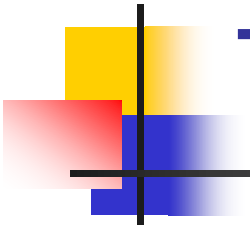
Answer 3.

Amino acid sequences in proteins are definitely not random



Answer 4.

What are proteins made from? We have new pretendents for a minimal set of amino acids. 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!