REPRESENTATION OF PROTEIN SEQUENCES USING NLP TECHNIQUES

Team 19

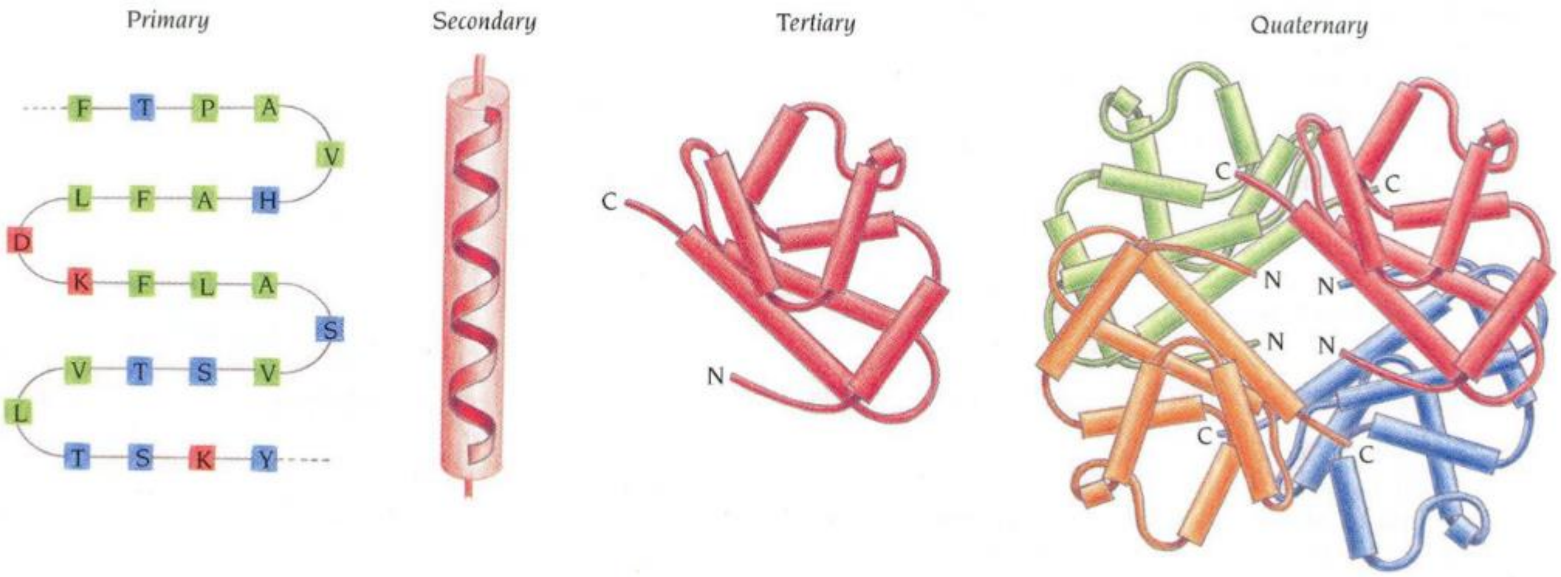
Protein Sequence Language

- Like strings of letters, protein sequences are linear chains of amino acids
- The amino acids are fundamental building blocks in protein sequences
- They can be one of 20 different types, corresponding to different chemical structures labeled as

  (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y)
- There are hundreds of different scales of properties of amino acids, including size, hydrophobicity, electronic properties, aromaticity, polarity, flexibility, secondary structure propensity and charge to name just a few
- Thus, although the 20 amino acids are a reasonable starting point to define building blocks in protein sequences, smaller, larger or uniquely encoded units may often be functionally more meaningful
- Motifs

Protein Sequence Language

- Most functions in biological systems are carried out by proteins mainly transmission of information, for example in signaling pathways, enzymatic catalysis and transport of molecules
- Proteins are synthesized from small building blocks, amino acids, of which there are 20 different types. The amino acids are connected to form a string (Primary Structure)
- The three-dimensional structure has certain repetitive elements known as secondary structures. α-helices, β-strands are highly occurring secondary structures and are connected by different types of loops



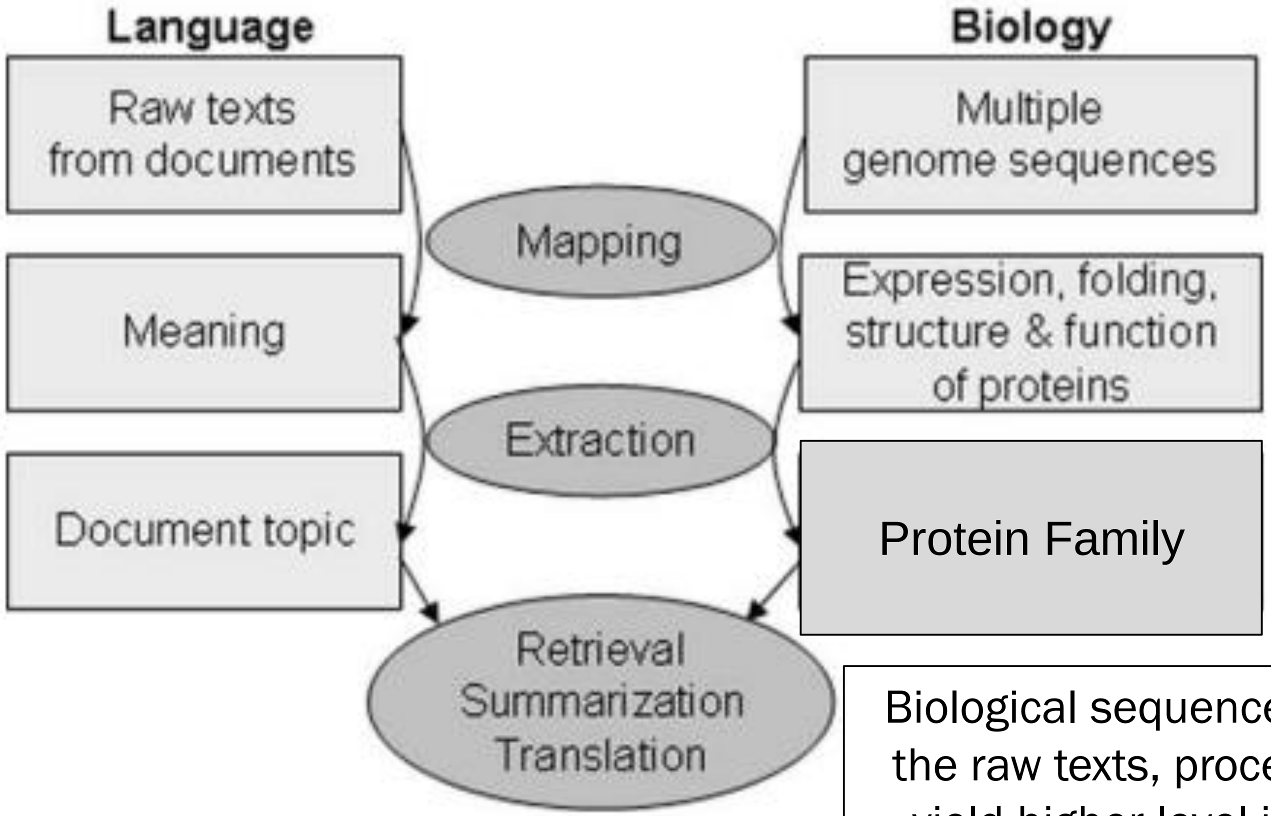
Parallels of protein sequences and words

- Understanding the structure, dynamics and function of proteins strongly parallels the mapping of words to meaning in natural language
- The words in a text document map to a meaning and convey rich information pertaining to the topic of the document. Similarly, protein sequences also represent the “raw text” and carry high-level information about the structures, dynamics and functions of proteins
- Many of the hallmarks of statistical analysis of biological sequences are similar to those of human languages
  - *Large data bodies need to be analyzed in both cases*
  - *Fundamental units of human languages include higher order structures, paralleled by domains, sub units or functionally linked proteins*
  - *Computer-based derivation of meaning from text is analogous to the prediction of structure and function from primary sequence data*
- This has been the motivation for our application of NLP techniques to the protein classification problem

Analogy between “Natural Language” and “Protein Sequence Language”

Words in sentences combine to convey meaningful content

Processing words in a document can convey the topic of the document and its meaningful content

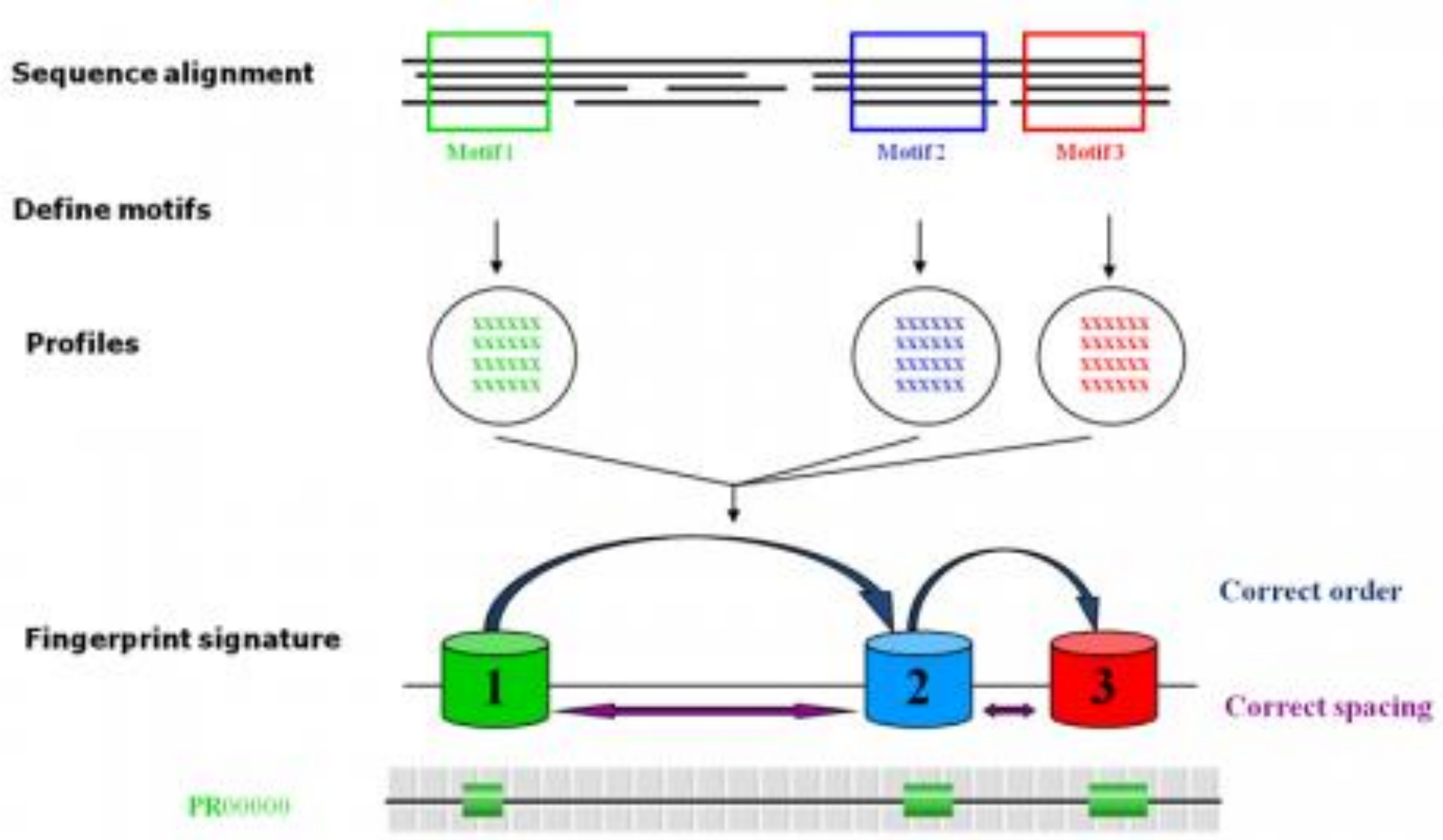


Biological sequences are analogous to the raw texts, processing which would yield higher level information on the behavioral properties of the physical entities corresponding to the sequences

Why a word2vec based approach?

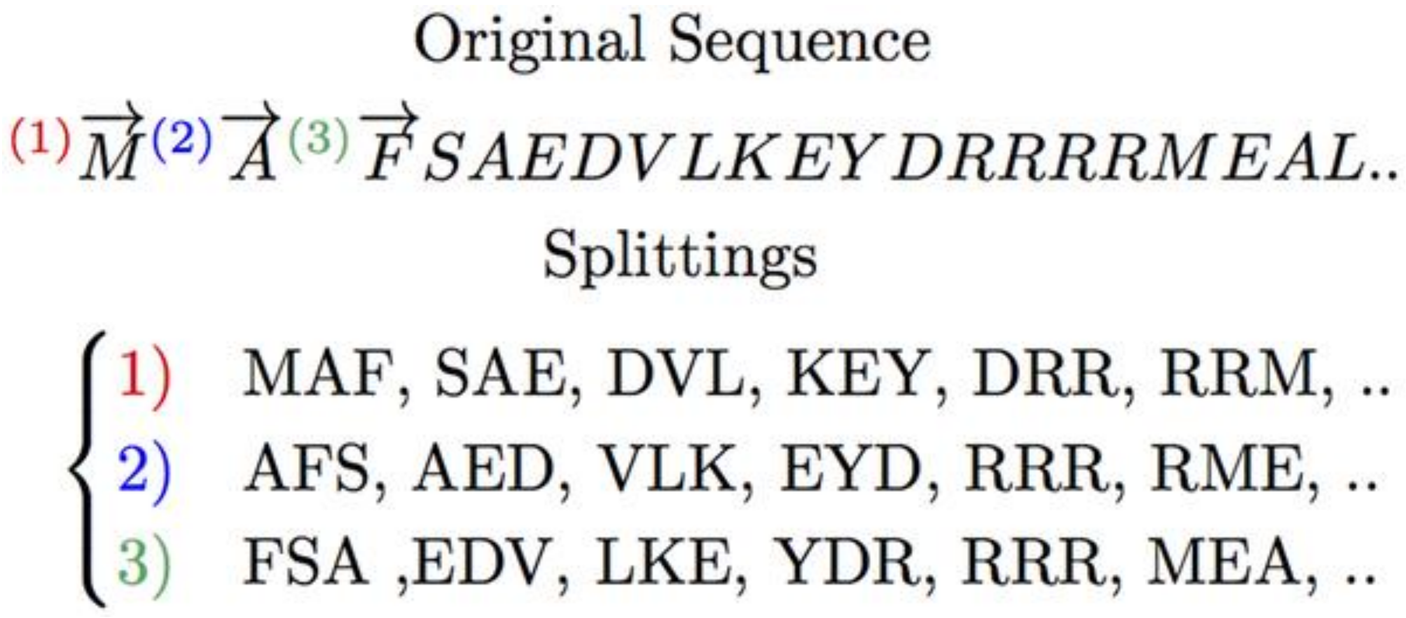
- Single amino acid propensities have limited ability to predict secondary structure elements
- It was therefore investigated if larger segments composed of several amino acids, so-called k-mers or n-grams of amino acids are more appropriate units of protein sequence language with respect to their meaning for secondary structure[2]
- However, this study found that n-grams do not capture secondary structure propensity of protein segments well
- Therefore, it was investigated if a compact representation of position specific n-grams as x{−|+}N, where x is the n-gram, {−|+} indicates whether it occurs before or after the residue under question, and N is the distance from this residue to the n-gram, may be a better representation of the protein sequence
- The analogy to language can be found when classifying documents into possible topics. This task also requires identification of crucial words that can discriminate between possible topics. For example, the word ‘ball’ can discriminate between “science” and “sports” topics but cannot distinguish between “cricket” and “football” topics

Context in a Protein Sequence



Prot-Vec Explained

Protein sequence splitting



Original Sequence

$$^{(1)}\overrightarrow{M}\,^{(2)}\overrightarrow{A}\,^{(3)}\overrightarrow{F}SAEDVLKEYDRRRRMEAL..$$

Splittings

$$\begin{cases} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{cases}$$

In n-gram modeling of protein informatics, usually an overlapping window of 3 to 6 residues is used.
Instead of taking overlapping windows, 3 lists of shifted non-overlapping words were generated, as shown above.

Creating the protein vector space

- In training word vector representations, Skip-gram attempts to maximize the probability of observed word sequences (contexts)
- In other words, for a given training sequence of words we would like to find their corresponding n-dimensional vectors maximizing the following average log probability function. Such a constraint allows similar words to assume a similar representation in this space

$$\underset{v,v}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{i+j}|w_i)$$

$$p(w_{i+j}|w_i) = \frac{\exp\left(v'^{T}_{w_{i+j}} v_{w_i}\right)}{\sum_{k=1}^{W} \exp\left(v'^{T}_{w_k} v_{w_i}\right)},$$

Where N is the length of the training sequence
2c is the window size we consider as the context
w_i is the center of the window
W is the number of words in the dictionary
v_w and v'_w are input and output n-dimensional representations of word w
The probability $p(w_{i+j}|w_i)$ is defined using a softmax function

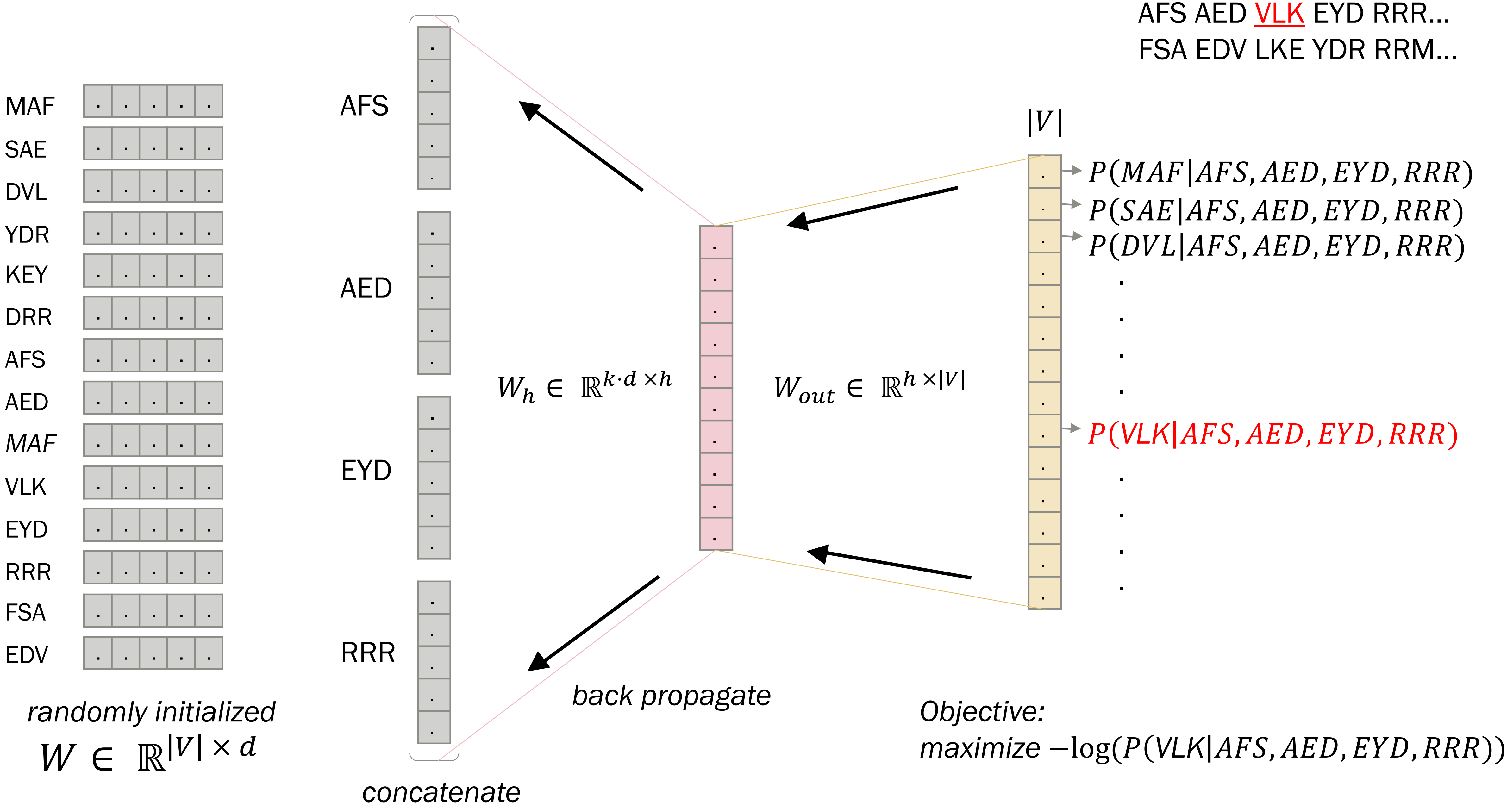
Note: To train the embedding vectors, we consider a vector size of 100 and a context size of 25.
Thus each 3-gram is presented as a vector of size 100.

Training instance: MAFSAEDVLKEYDRRR...

MAF SAE DVL KEY DRR...

AFS AED VLK EYD RRR...

FSA EDV LKE YDR RRM...



MAF
SAE
DVL
YDR
KEY
DRR
AFS
AED
MAF
VLK
EYD
RRR
FSA
EDV

randomly initialized

$W \in \mathbb{R}^{|V| \times d}$

AFS
AED
EYD
RRR

concatenate

$W_h \in \mathbb{R}^{k \cdot d \times h}$

back propagate

$W_{out} \in \mathbb{R}^{h \times |V|}$

$|V|$

$P(MAF|AFS, AED, EYD, RRR)$

$P(SAE|AFS, AED, EYD, RRR)$

$P(DVL|AFS, AED, EYD, RRR)$

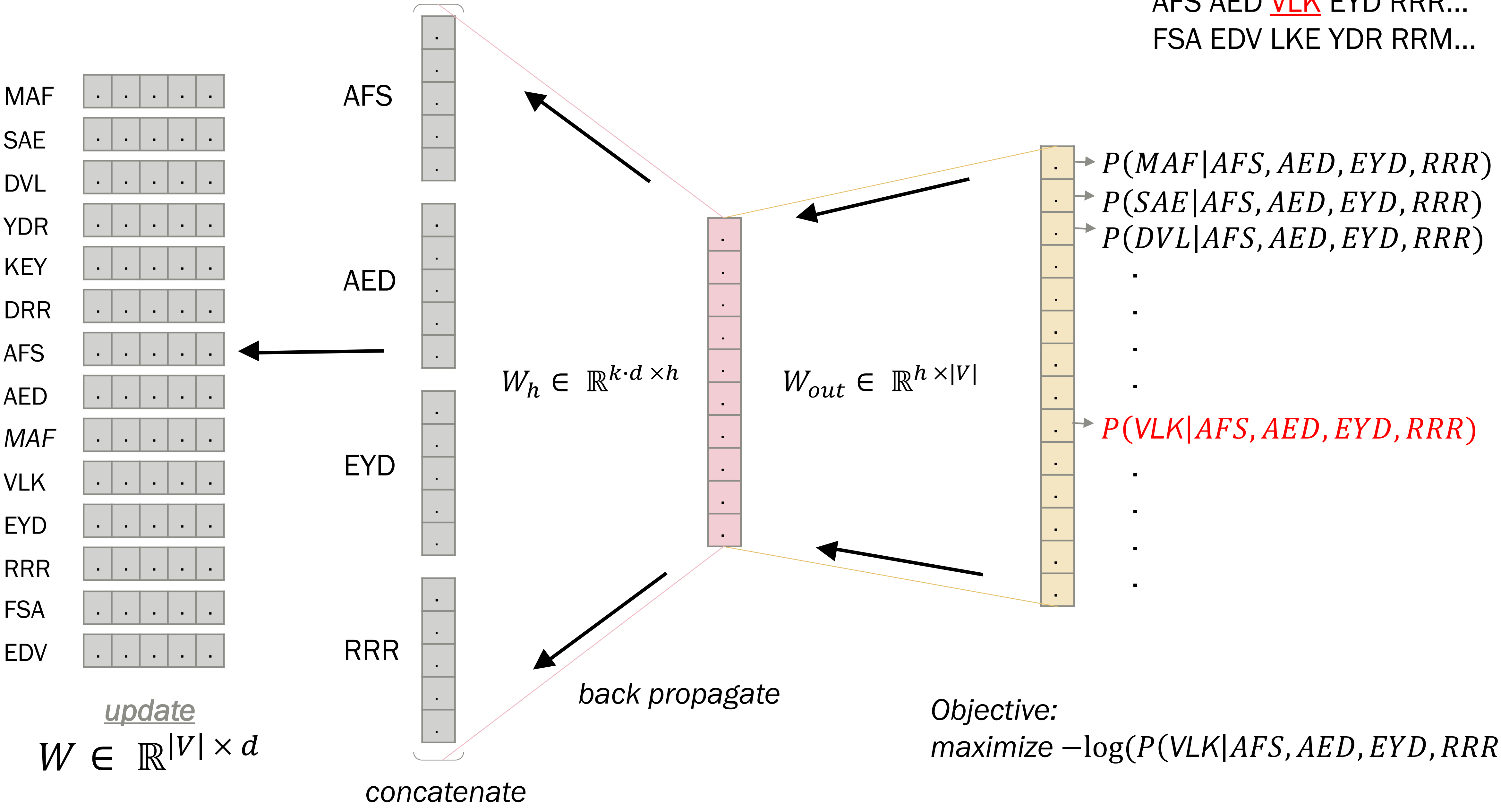
$P(VLK|AFS, AED, EYD, RRR)$

Objective:

maximize $-\log(P(VLK|AFS, AED, EYD, RRR))$

Training instance: MAFSAEDVLKEYDRRR...

MAF SAE DVL KEY DRR...
AFS AED VLK EYD RRR...
FSA EDV LKE YDR RRM...

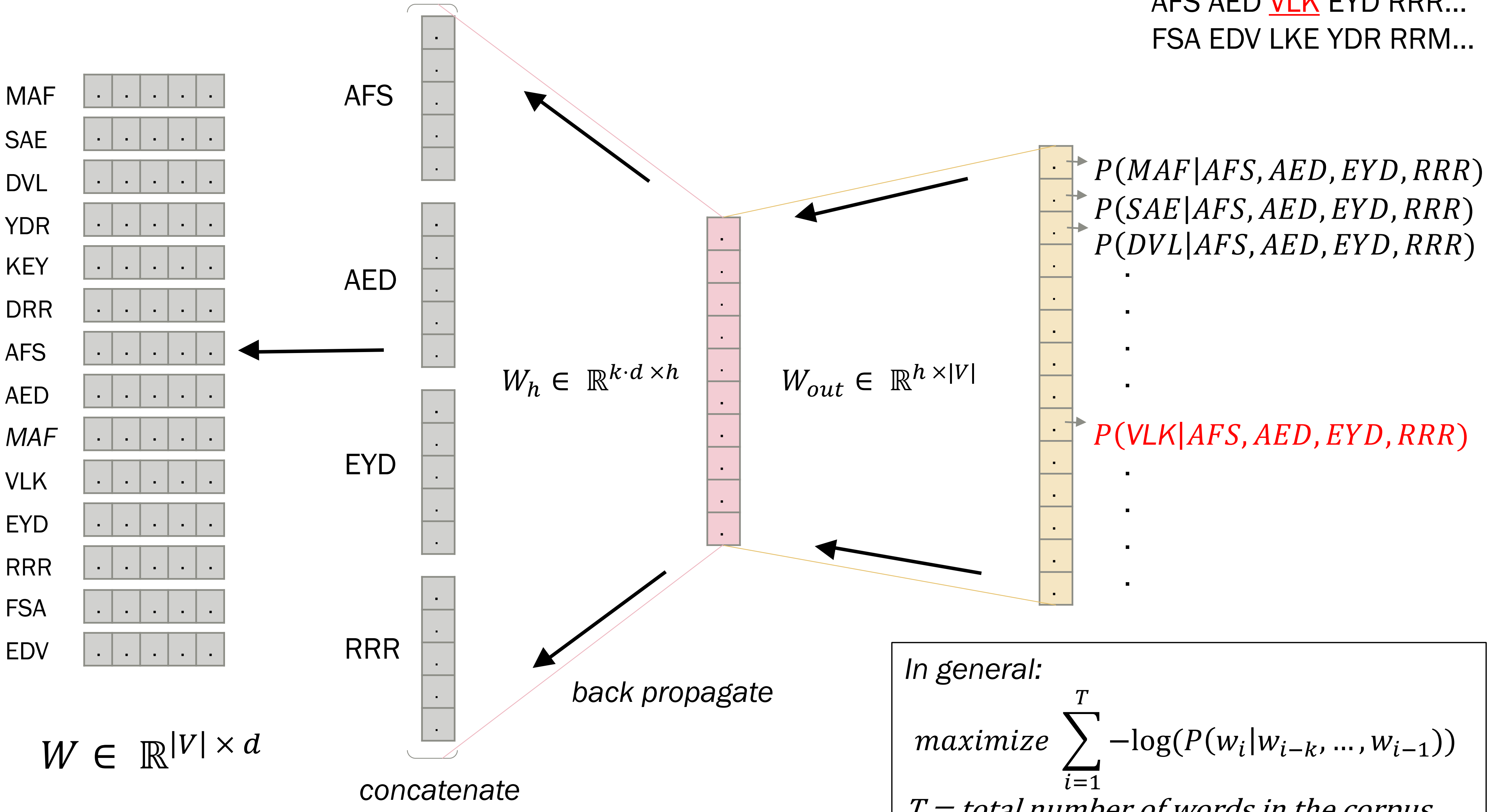


MAF
SAE
DVL
YDR
KEY
DRR
AFS
AED
MAF
VLK
EYD
RRR
FSA
EDV
update
$W \in \mathbb{R}^{|V| \times d}$
AFS
AED
EYD
RRR
concatenate
$W_h \in \mathbb{R}^{k \cdot d \times h}$
back propagate
$W_{out} \in \mathbb{R}^{h \times |V|}$
$P(MAF|AFS, AED, EYD, RRR)$
$P(SAE|AFS, AED, EYD, RRR)$
$P(DVL|AFS, AED, EYD, RRR)$
$P(VLK|AFS, AED, EYD, RRR)$

Objective:
maximize $-\log(P(VLK|AFS, AED, EYD, RRR))$

Training instance: MAFSAEDVLKEYDRRR...

MAF SAE DVL KEY DRR...
AFS AED VLK EYD RRR...
FSA EDV LKE YDR RRM...


MAF
SAE
DVL
YDR
KEY
DRR
AFS
AED
MAF
VLK
EYD
RRR
FSA
EDV
$W \in \mathbb{R}^{|V| \times d}$
AFS
AED
EYD
RRR
concatenate
$W_h \in \mathbb{R}^{k \cdot d \times h}$
back propagate
$W_{out} \in \mathbb{R}^{h \times |V|}$
$P(MAF|AFS, AED, EYD, RRR)$
$P(SAE|AFS, AED, EYD, RRR)$
$P(DVL|AFS, AED, EYD, RRR)$
$P(VLK|AFS, AED, EYD, RRR)$

In general:

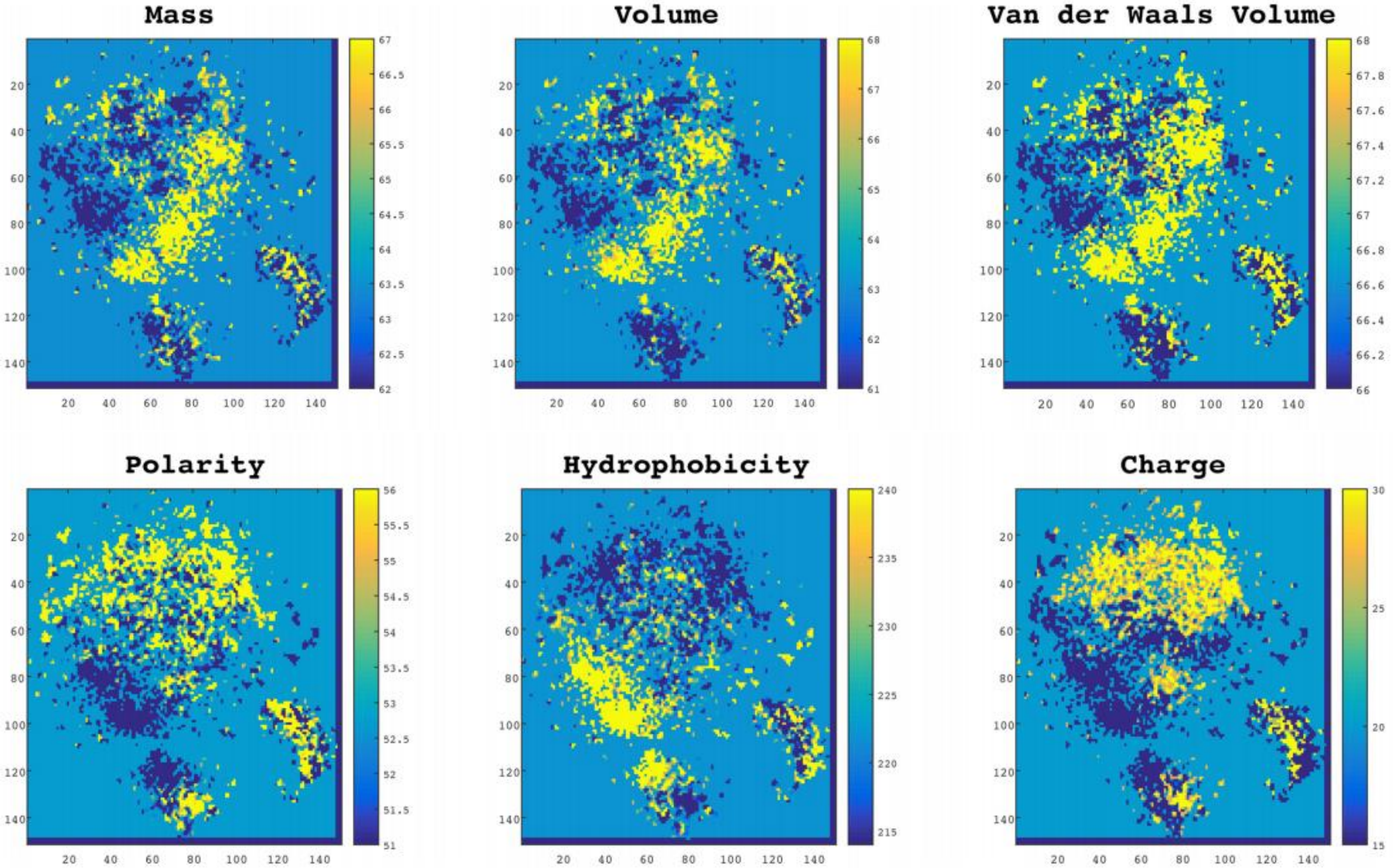
$$\text{maximize} \sum_{i=1}^{T} -\log(P(w_i|w_{i-k}, \dots, w_{i-1}))$$

$T = \textit{total number of words in the corpus}$

Evaluation of the protein space

Qualitative analysis:

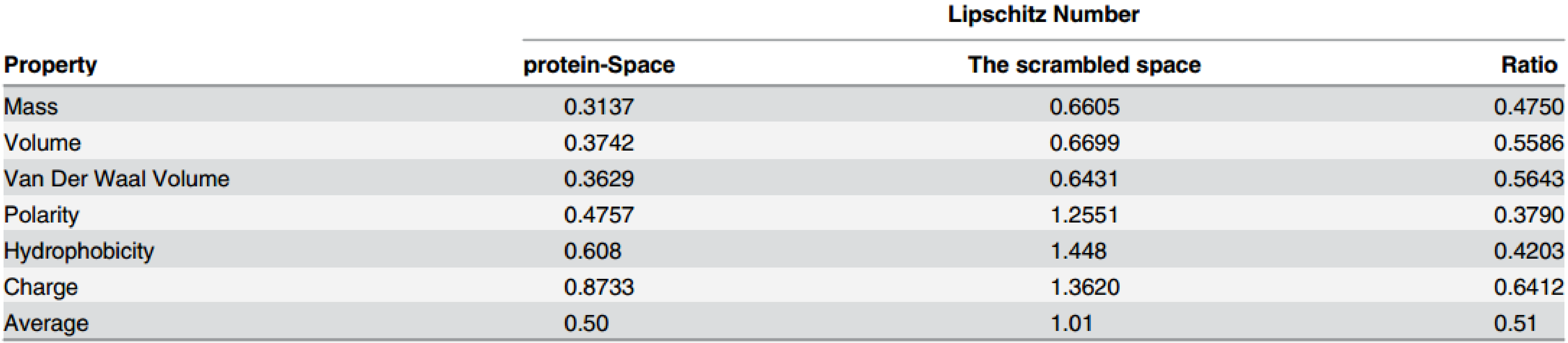
In order to visualize the distribution of the aforementioned properties,
All 3-gram embeddings from 100-dimensional space were projected to a 2D space using Stochastic Neighbor Embedding (t-SNE)



Interestingly, as can be seen in the figure,
3-grams with the same biophysical and biochemical properties were grouped together. This observation suggests that the proposed embedding not only encodes protein sequences in an efficient way
that proved useful for classification purposes,
but also reveals some important physical and chemical patterns in protein sequences.

Quantitative analysis:

We calculated the best Lipschitz constant. For all 6 properties presented previously, we calculated the minimum k. To evaluate this result we made an artificial space called “scrambled space” by randomly shuffling the labels of 3-grams in the 100 dimensional space.

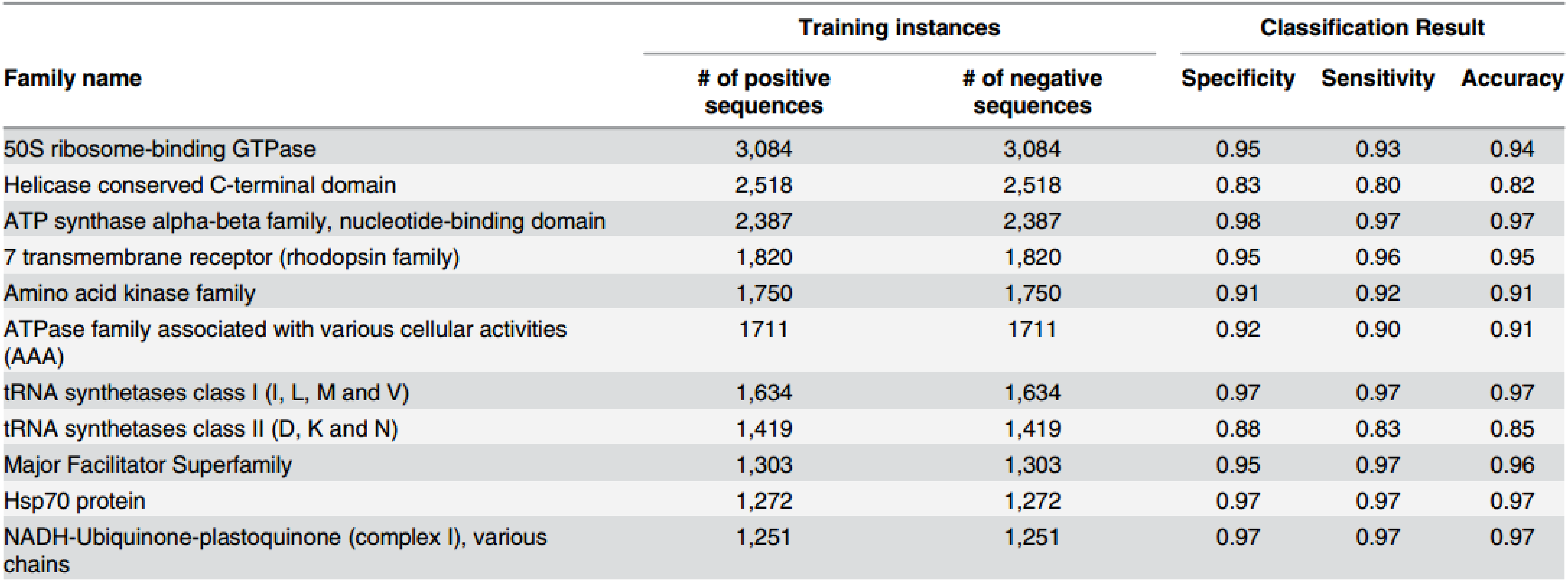


	Lipschitz Number		
Property	**protein-Space**	**The scrambled space**	**Ratio**
Mass	0.3137	0.6605	0.4750
Volume	0.3742	0.6699	0.5586
Van Der Waal Volume	0.3629	0.6431	0.5643
Polarity	0.4757	1.2551	0.3790
Hydrophobicity	0.608	1.448	0.4203
Charge	0.8733	1.3620	0.6412
Average	0.50	1.01	0.51

- Normally if k = 1 the function is called a short map, and if 0 k < 1 the function is called a contraction.
- The results suggest that the protein-space is on average 2-times smoother in terms of physical and chemical properties than a random space.
- This quantitative result supports our qualitative observation of the space structure and suggests that the training space encodes, 3-grams in an informative manner.

Performance of ProtVec

- By training on primary sequence data alone the following results where obtained.
- In order to evaluate the strength of ProtVec, classifications of 7,027 protein families and obtained a weighted average accuracy of 93 ± 0.06%, which exhibits a more reliable result than the existing methods



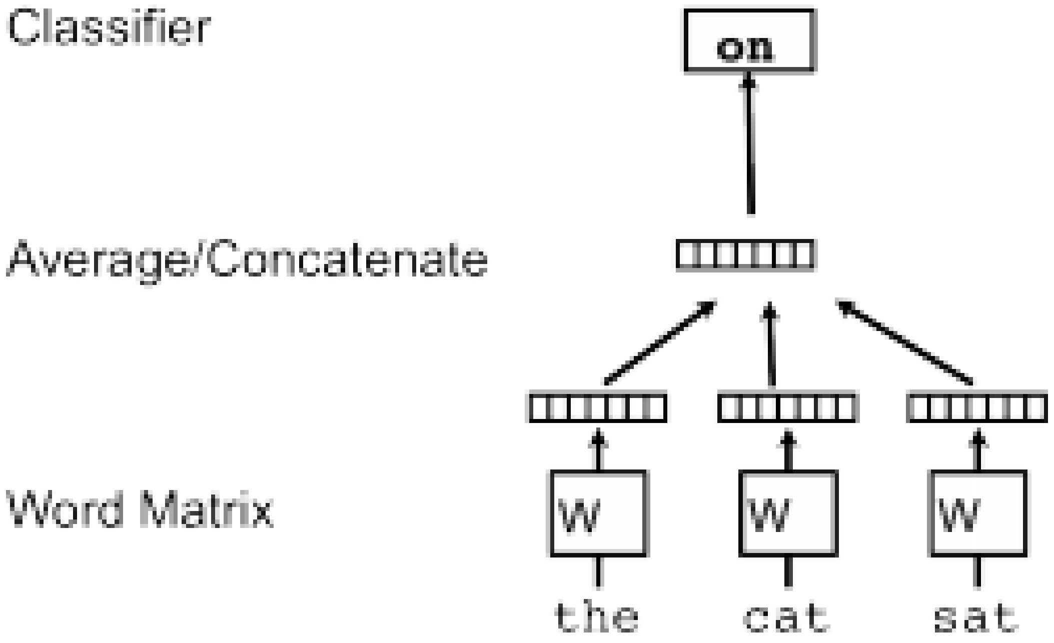
	Training instances		Classification Result		
Family name	# of positive sequences	# of negative sequences	Specificity	Sensitivity	Accuracy
50S ribosome-binding GTPase	3,084	3,084	0.95	0.93	0.94
Helicase conserved C-terminal domain	2,518	2,518	0.83	0.80	0.82
ATP synthase alpha-beta family, nucleotide-binding domain	2,387	2,387	0.98	0.97	0.97
7 transmembrane receptor (rhodopsin family)	1,820	1,820	0.95	0.96	0.95
Amino acid kinase family	1,750	1,750	0.91	0.92	0.91
ATPase family associated with various cellular activities (AAA)	1711	1711	0.92	0.90	0.91
tRNA synthetases class I (I, L, M and V)	1,634	1,634	0.97	0.97	0.97
tRNA synthetases class II (D, K and N)	1,419	1,419	0.88	0.83	0.85
Major Facilitator Superfamily	1,303	1,303	0.95	0.97	0.96
Hsp70 protein	1,272	1,272	0.97	0.97	0.97
NADH-Ubiquinone-plastoquinone (complex I), various chains	1,251	1,251	0.97	0.97	0.97

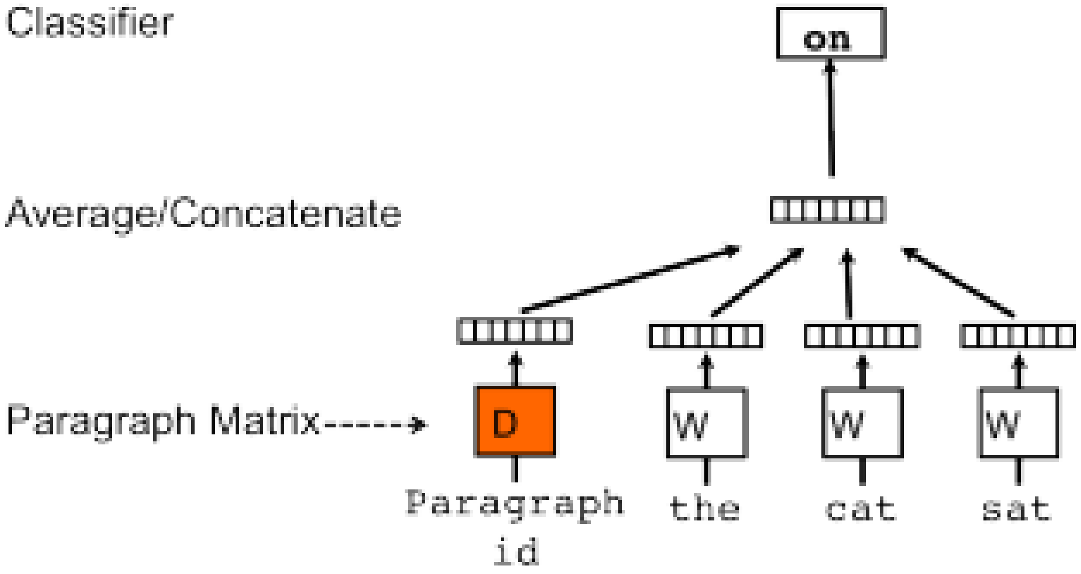
A few of the results are displayed above

Steps forward for consideration

- Paragraph to Vector
- Different sequence segmentation techniques
  - *Graph Theory based approach vertices as segments and best segmentation is chosen based on minimal edge costs*
  - *Using domain knowledge to segment based on secondary structure*
  - *Using domain knowledge to segment based on profiles and patterns in a sequence*

Paragraph Vector





References

- Computational Biology and Language Madhavi Ganapathiraju , N. Balakrishnan , Raj Reddy , and Judith Klein-Seetharaman
- Liu, Y and Carbonell, J and Klein-Seetharaman, J and Gopalakrishnan, V: “Comparison of Probabilistic Combination Methods for Protein Secondary Structure Prediction.” Bioinformatics, volume 16, issue 4. (2004) 376-82
- Asgari, Ehsaneddin; Mofrad, Mohammad R.K. (2015). "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics". *PloS one*. **10** (11): e0141287. Bibcode:2015PLoSO..1041287A. doi:10.1371/journal.pone.0141287