PLOS ONE

RESEARCH ARTICLE

Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics

Ehsaneddin Asgari[1], Mohammad R. K. Mofrad[1,2]*

1 Molecular Cell Biomechanics Laboratory, Departments of Bioengineering and Mechanical Engineering, University of California, Berkeley, California 94720, United States of America, **2** Physical Biosciences Division, Lawrence Berkeley National Lab, Berkeley, California 94720, United States of America

* mofrad@berkeley.edu

Classification tasks

Protein family classification

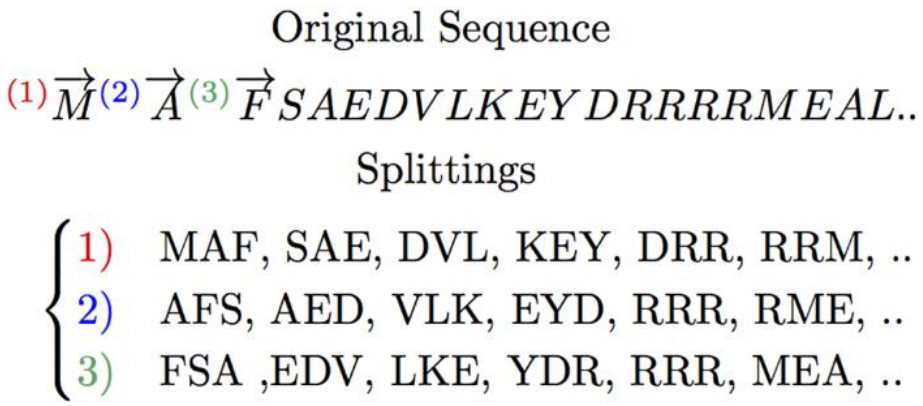
- Pfam database: family, domain repeat, motif

Disordered proteins

- DisProt – experimentally validated disordered proteins (categorizes disordered and ordered regions of a collection of proteins)
- FG-Nups – disordered rehions of phenylalanine-glycine nucleoporins

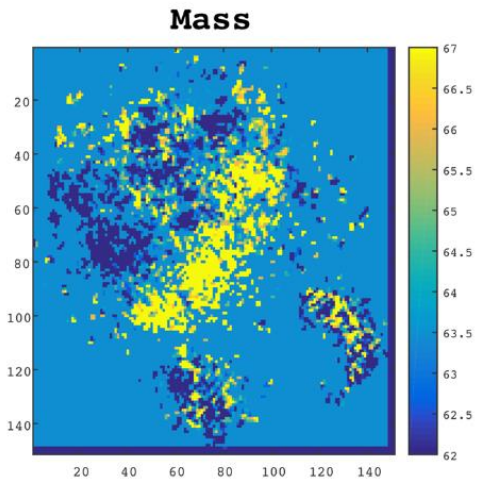
Protein-Space Construction

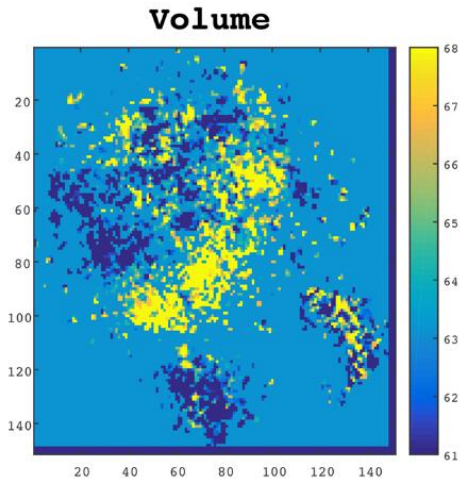
- 546,790 manually annotated and reviewed sequences from Swiss-Prot

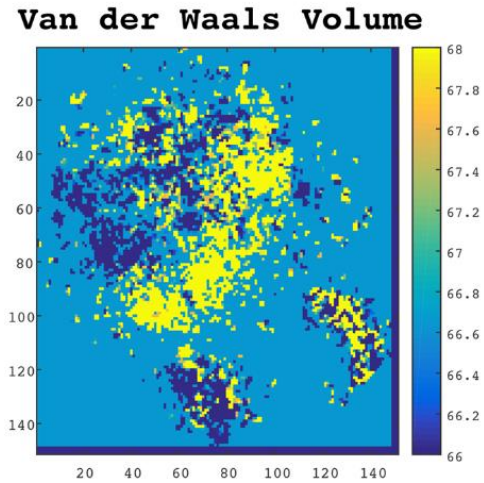


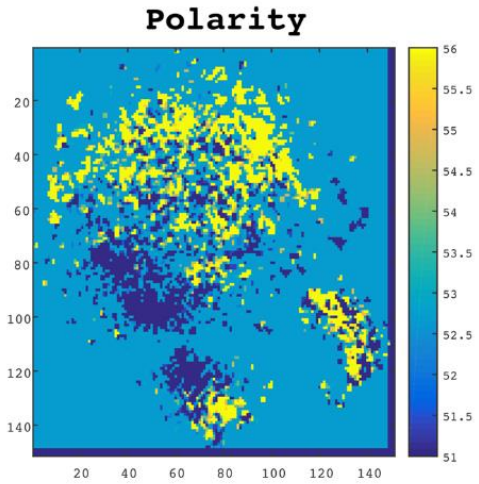
- 1,640,370 3-grams (biological words)
- Training Skip-gram neural network
- Each 3-gram is presented as a vector of size 100

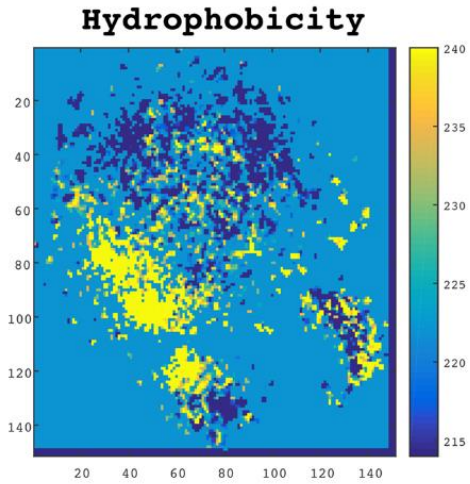
Protein-Space analysis

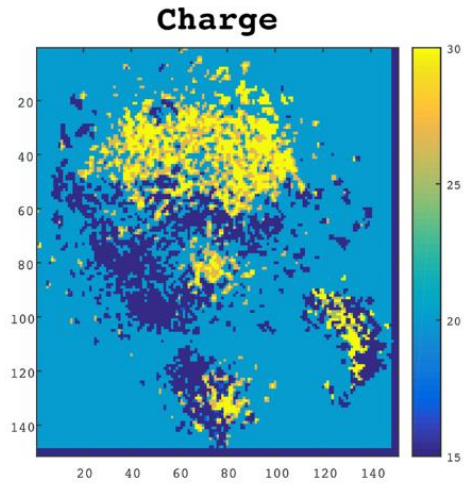












Protein-Space analysis

$$d_f(f_{prop}(w_1), f_{prop}(w_2)) \leq k \times d_w(w_1, w_2)$$

Table 1. Using Lipschitz number to evaluate the continuity of ProtVec with respect to biophysical and biochemical properties.

	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

Protein Family Classification

- Pfam: 324,018 sequences from 7,027 distinct families
- Each sequence is presented as a summation of the vector representation of overlappling 3-grams (a vector of size 100)
- For each family type the same number of sequences from Swiss Prot are selected randomly as negative examples
- Classification using SVM and 10-fold CV

Protein Family Classification

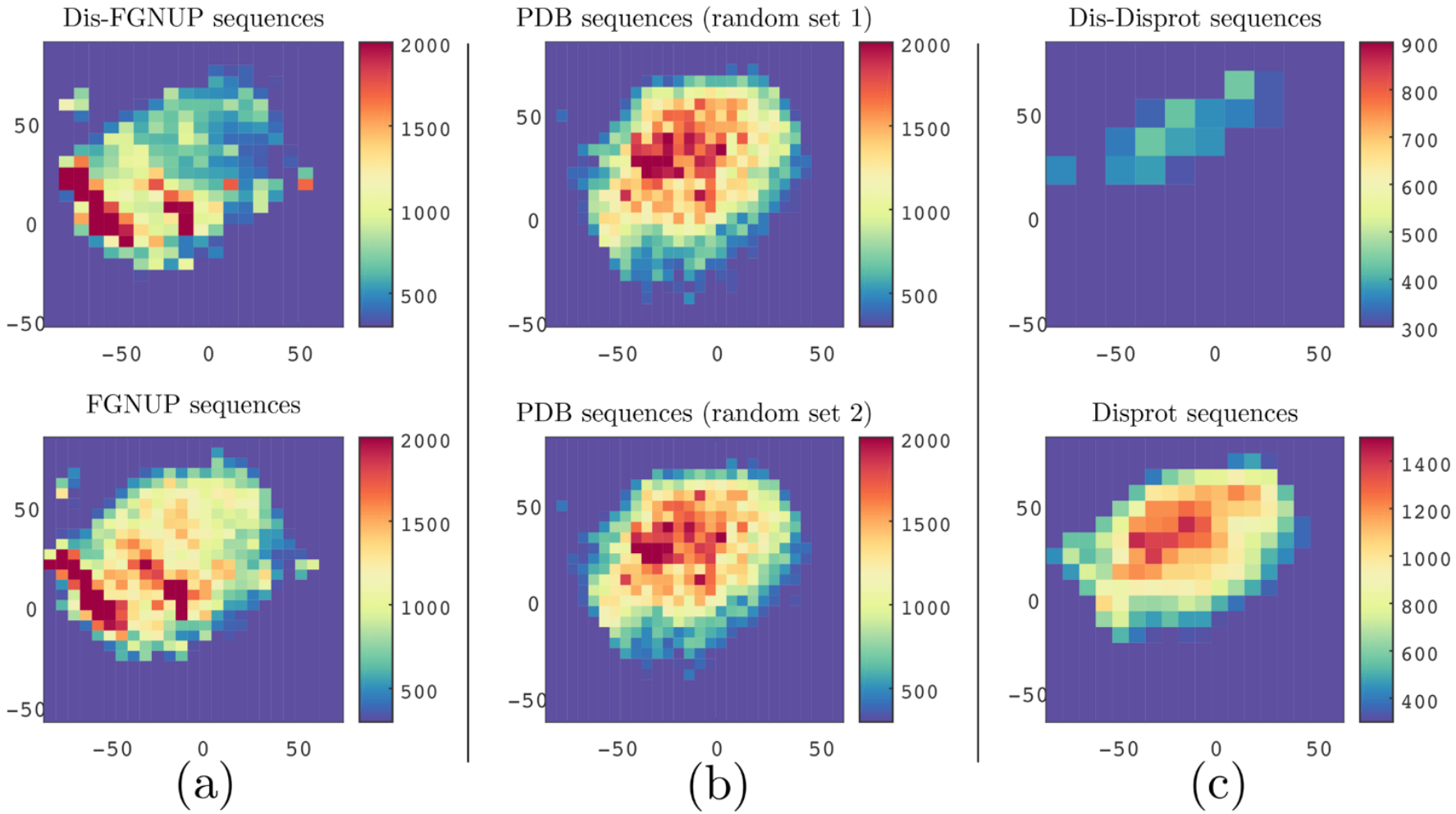
Table 2. Performance of protein family classification using SVM and ProtVec over some of the most frequent families in Swiss-Prot. Families are sorted with respect to their frequency in Swiss-Prot.

Family name	Training instances		Classification Result		
	# 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
Histidine biosynthesis protein	1,248	1,248	0.96	0.97	0.97
TCP-1-cpn60 chaperonin family	1,246	1,246	0.95	0.96	0.95
EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)	1,207	1,207	0.96	0.96	0.96
Aldehyde dehydrogenase family	1,200	1,200	0.93	0.94	0.94
Shikimate–quinate 5-dehydrogenase	1,128	1,128	0.87	0.89	0.88
GHMP kinases N terminal domain	1,120	1,120	0.88	0.92	0.90
Ribosomal protein S2	1,083	1,083	0.95	0.96	0.95
Ribosomal protein S4–S9 N-terminal domain	1,072	1,072	0.95	0.97	0.96

Visualization and Classification of Disordered Proteins

- FG-Nups (1,138 sequences) and two random sets of 1,138 structured proteins from PDB (avg. length of 900 residues)
- DisProt 694 proteins presenting 1539 disordered and 95 ordered regions
- SVM classifier, each protein = sum of 3-gram vectors

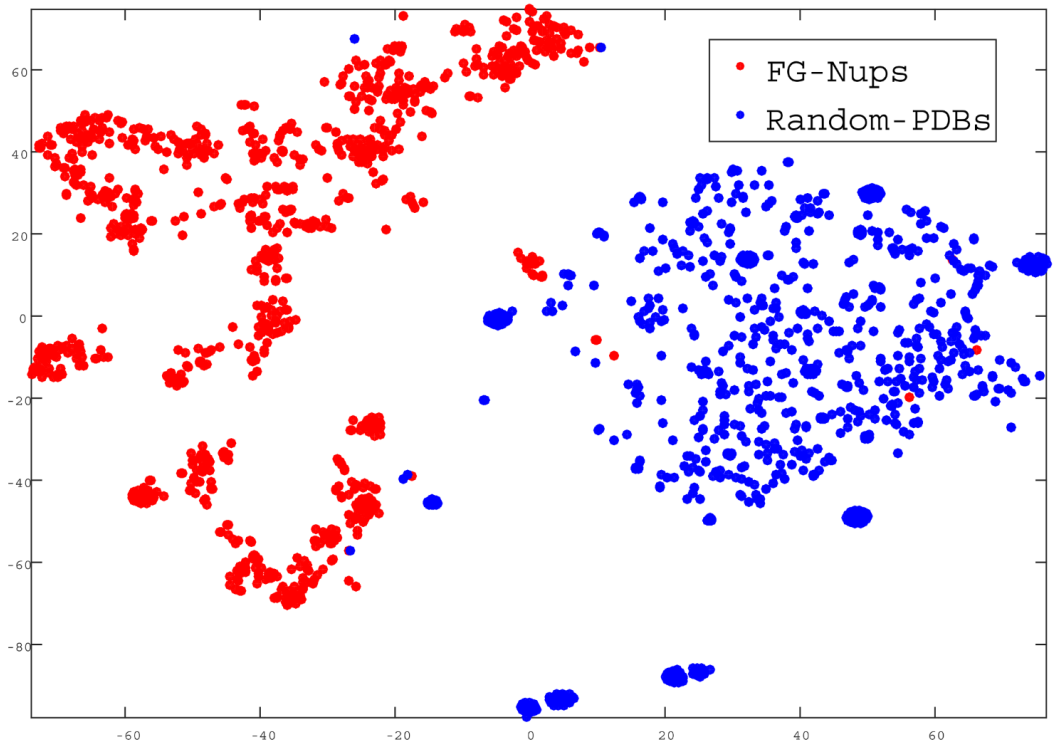
Visualization of Disordered Proteins



Classification of Disordered Proteins

Table 3. The performance of FG-Nups disordered protein classification in a 10xFold cross-validation using SVM.

Sensitivity	Specificity	Accuracy
0.9987	0.9974	0.9981



Conclusions

- ProtVec can be used as an informative and dense representation for biological sequences
- By training this representation solely on protein sequences, feature extraction approach was able to capture a diverse range of meaningful physical and chemical properties
- ProtVec may be used as an approach for protein data visualization
- Embeddings could be trained once and then used to encode biological sequences in any given problem
- This representation can be considered as pre-training for various applications of deep learning in bioinformatics.