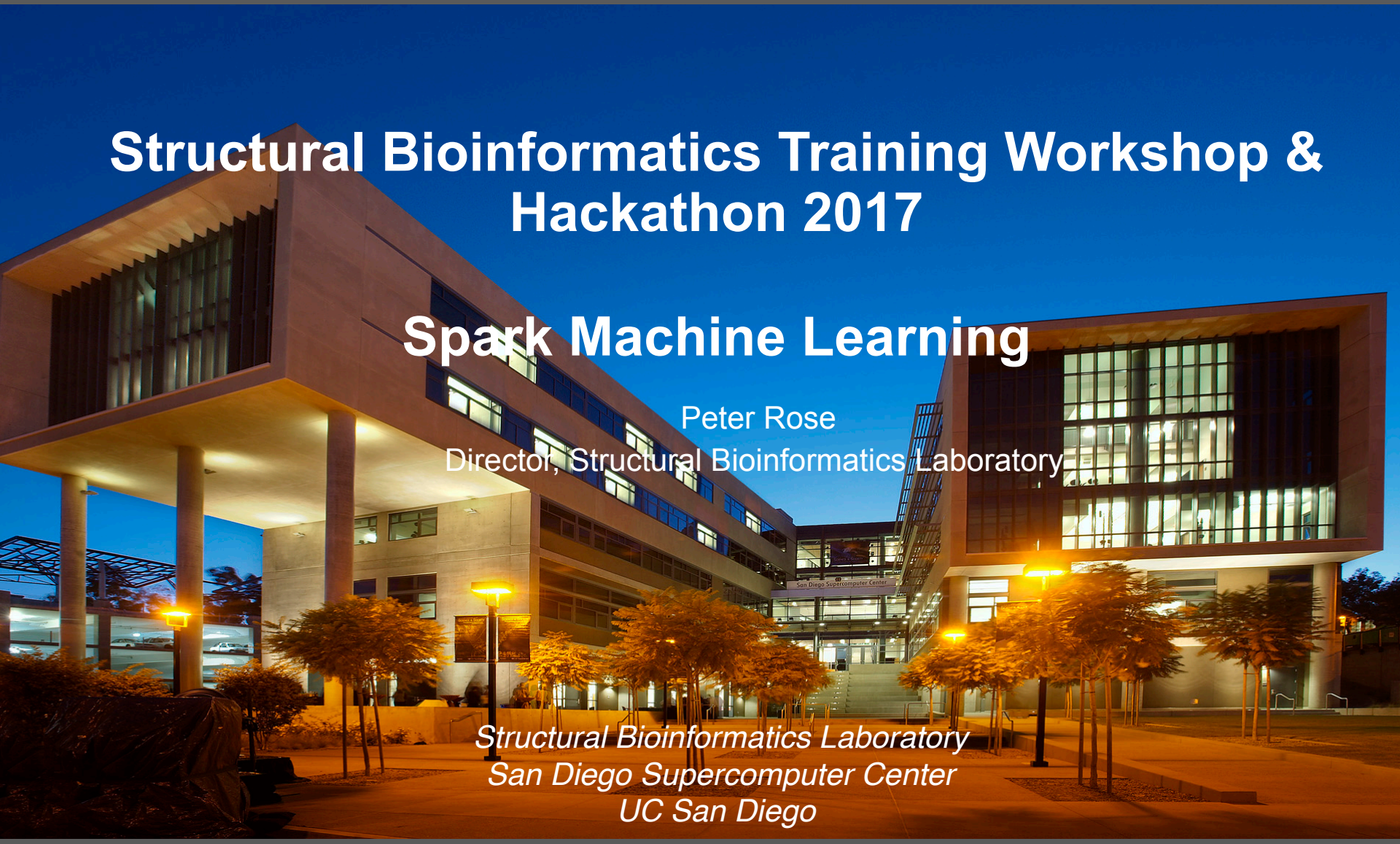
Structural Bioinformatics Training Workshop & Hackathon 2017

Spark Machine Learning

Peter Rose
Director, Structural Bioinformatics Laboratory



Structural Bioinformatics Laboratory
San Diego Supercomputer Center
UC San Diego

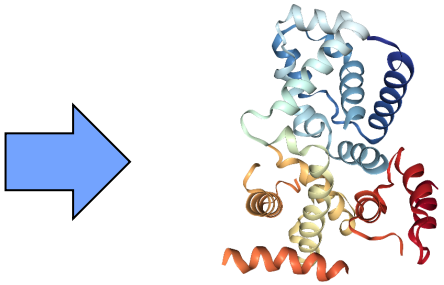
Spark Machine Learning

- **Spark-ML (new developments)**
  - Uses Dataset API
- **Spark-MLlib (in maintenance mode)**
  - Uses RDD API

Example Problem

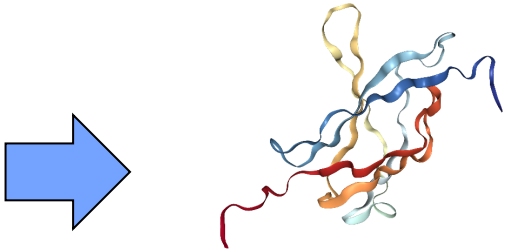
Classify protein fold as either all alpha or all beta given the protein sequence.

MHHHHHHSSGRENLYFQGMTVREKTRLEKFRQLLSSQNTDLDELRKCS
WPGVPREVRPITWRLLSGYLPANTERRKLTLQRKREEYFGFIEQYYDSR
NEEHHQDTYRQIHIDIPRTNPLIPLFQQPLVQEIFERILFIWAIRHPASGYV
QGINDLVTPFFVVFLSEYVEEDVENFDVTNLSQDMLRSIEADSFWCMSK
LLDGIQDNYTFAQPGIQKKVKALEELVSRIDEQVHNHFRRYEVEYLQFAF
RWMNNLLMRELPLRCTIRLWDTYQSEPEGFSHFHLYVCAAFLIKWRKEIL
DEEDFQGLLMLLQNLPTIHWGNEEIGLLLAEAYRLKYMFADAPNHYRR



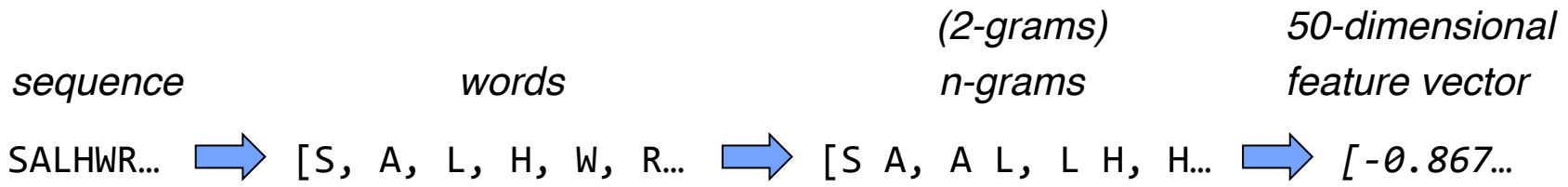
PDB ID: 3DZX – all alpha

GSSGSSGLPQVEAYSPSACSVRGGEELVLTGSNFLPDSKVVFIERGPDG
KLQWEEEATVNRLQSNEVTLTLTVPEYSNKRVSRPVQVYFYVSNGRRK
RSPTQSFRFLPVICKEE



PDB ID: 2YRP – all beta

Feature Vector Creation



Spark-ML Pipeline: ***RegexTokenizer*** ***NGram*** ***Word2Vec***

alpha	beta	coil	foldType	words	ngram	features
0.7194245	0.0	0.28057554	alpha	[S, A, L, H, W, R...	[S A, A L, L H, H...	[-0.8678886349164...
0.4090909	0.0	0.59090906	alpha	[S, H, L, K, S, K...	[S H, H L, L K, K...	[-0.2172588950821...
0.015151516	0.5808081	0.4040404	beta	[M, A, H, H, H, H...	[M A, A H, H H, H...	[-0.6485012823101...
0.7916667	0.0	0.20833333	alpha	[M, G, S, S, H, H...	[M G, G S, S S, S...	[-0.7672874573129...
0.028846154	0.52884614	0.44230768	beta	[M, E, K, A, T, K...	[M E, E K, K A, A...	[-0.7308713050851...
0.6904762	0.011904762	0.29761904	alpha	[P, T, I, H, D, H...	[P T, T I, I H, H...	[-0.7869252321949...
0.8108108	0.0	0.1891892	alpha	[G, A, M, E, P, E...	[G A, A M, M E, E...	[-0.8350848566740...
0.022900764	0.39694658	0.5801527	beta	[I, V, N, G, E, E...	[I V, V N, N G, G...	[-0.6840203473201...
0.8021978	0.0	0.19780022	alpha	[M, T, P, D, V, L...	[M T, T P, P D, D...	[-0.7698143909806...

https://github.com/sbl-sdsc/mmtf-spark/blob/master/src/main/java/edu/sdsc/mmtf/spark/ml/SequenceWord2Vector.java

Create Class Labels

alpha	beta	coil	foldType
0.7194245	0.0	0.28057554	alpha
0.4090909	0.0	0.59090906	alpha
0.015151516	0.5808081	0.4040404	beta
0.7916667	0.0	0.20833333	alpha
0.028846154	0.52884614	0.44230768	beta
0.6904762	0.011904762	0.29761904	alpha
0.8108108	0.0	0.1891892	alpha
0.022900764	0.39694658	0.5801527	beta
0.8021978	0.0	0.19780022	alpha

The Dataset for Classification

We save this dataset as .parquet file:

```
data.write().mode("overwrite").format("parquet").save(filename);
```

class label ***features***

alpha	beta	coil	foldType	words	ngram	features
0.7194245	0.0	0.28057554	alpha	[S, A, L, H, W, R...]	[S A, A L, L H, H...]	[-0.8678886349164...]
0.4090909	0.0	0.59090906	alpha	[S, H, L, K, S, K...]	[S H, H L, L K, K...]	[-0.2172588950821...]
0.015151516	0.5808081	0.4040404	beta	[M, A, H, H, H, H...]	[M A, A H, H H, H...]	[-0.6485012823101...]
0.7916667	0.0	0.20833333	alpha	[M, G, S, S, H, H...]	[M G, G S, S S, S...]	[-0.7672874573129...]
0.028846154	0.52884614	0.44230768	beta	[M, E, K, A, T, K...]	[M E, E K, K A, A...]	[-0.7308713050851...]
0.6904762	0.011904762	0.29761904	alpha	[P, T, I, H, D, H...]	[P T, T I, I H, H...]	[-0.7869252321949...]
0.8108108	0.0	0.1891892	alpha	[G, A, M, E, P, E...]	[G A, A M, M E, E...]	[-0.8350848566740...]
0.022900764	0.39694658	0.5801527	beta	[I, V, N, G, E, E...]	[I V, V N, N G, G...]	[-0.6840203473201...]
0.8021978	0.0	0.1978022	alpha	[M, T, P, D, V, L...]	[M T, T P, P D, D...]	[-0.7698143909806...]

Demo 1

- Create a dataset
- Fit several classification models

Problem 1

- **Change the code to a 3-state classification problem:**
  - alpha, beta, alpha+beta
- **Rerun the classification methods**

Resources

- **Machine Learning Library (MLlib) Guide**
  - https://spark.apache.org/docs/latest/ml-guide.html
- **Extracting, transforming and selecting features**
  - https://spark.apache.org/docs/latest/ml-features.html
  - N-gram
    - https://spark.apache.org/docs/latest/ml-features.html#n-gram
  - Word2Vec example
    - https://spark.apache.org/docs/latest/ml-features.html#word2vec
  - Word2Vec model
    - https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec
- **Classification and regression**
  - https://spark.apache.org/docs/latest/ml-classification-regression.html
- **Parquet files (columnar format)**
  - https://spark.apache.org/docs/latest/sql-programming-guide.html#parquet-files

Funding

This workshop was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.