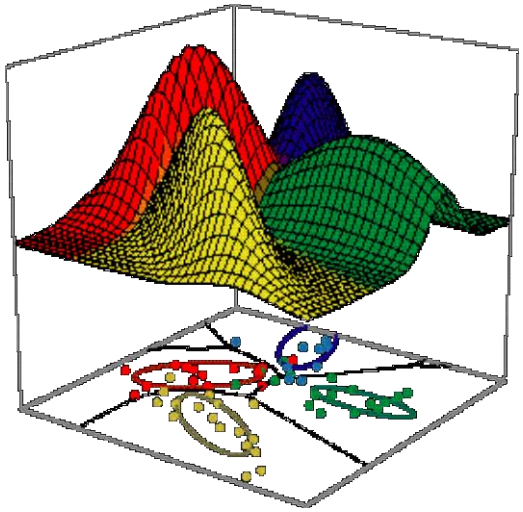
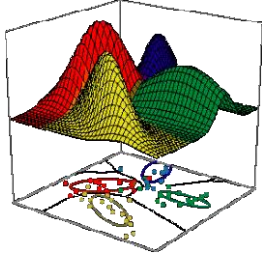
SYSC5405 / BIOM5405



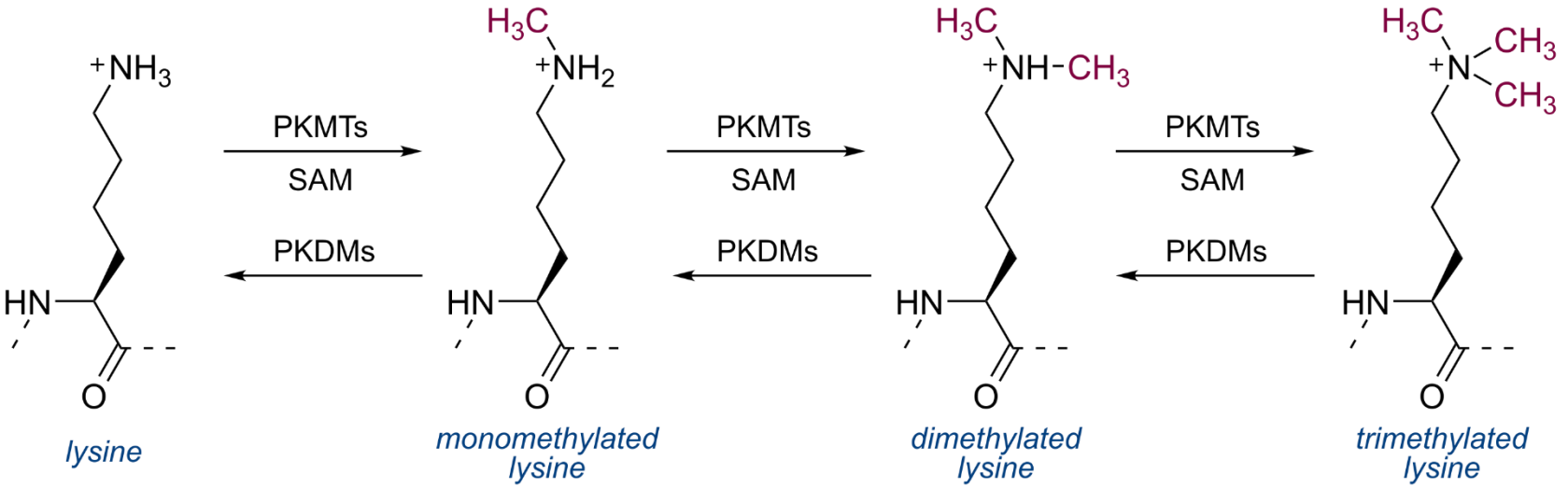
Term Project Launch

19 Nov 2019

Protein Lysine Methylation

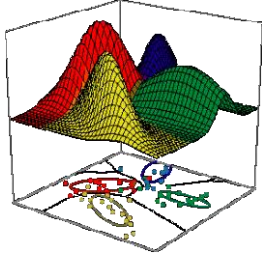


- When a protein is methylated, it can change its function, structure, or stability.
  - An important way for the cell to respond to changes/signals in its environment.



By Hbf878 - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=73213629

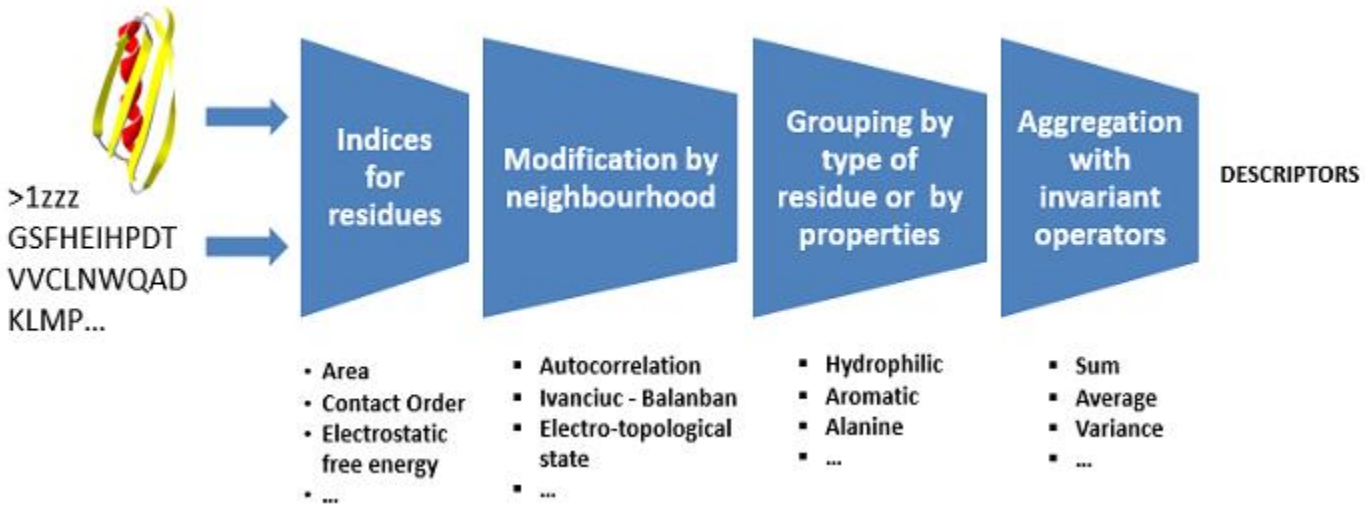
Protein Lysine Methylation



- Predicting which lysine residue will be methylated is very difficulty
  - Examine local sequence window centred on lysine.
  - Derive numerical features/descriptors of that window
  - Train/test a pattern classification approach to distinguish methylated/non-methylated sites

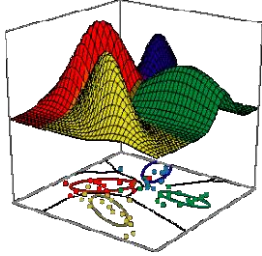
The Dataset

- Each window/site has 29 descriptors
  - Computed using ProtDCal



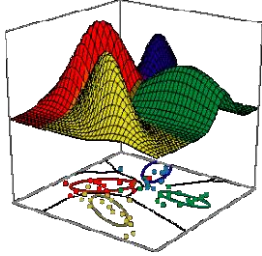
(doi:10.1002/pro.3673)

The Dataset



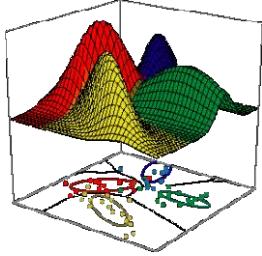
- Actual class denoted by ‘P’ or ‘N’ in final column
- Two CSV data files provided:
  - csv_result-Descriptors_Calibration.csv, containing 4996 rows
  - csv_result-Descriptors_Training.csv, containing 19988 rows
  - Use these however you like…
- We have withheld 5150 rows as a blind test set.
  - The labels of these data will never be released.

Your Goal



- For a site with 29 descriptors, you must predict whether that site will be methylated (‘P’) or not (‘N’)

Additional Project Details



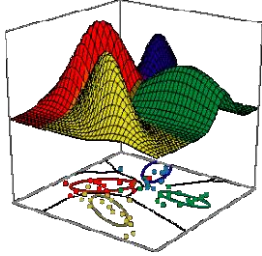
- You will be evaluated on:
  - 1) Prediction accuracy over test data set
    - as measured by maximum achievable recall at a precision of at least 50% (Re@Pr50)

$$Score_{Accuracy} = Re@Pr50$$

  - 2) How close your predicted Re@Pr50 is to your actual test Re@Pr50
    - Provide a mean and standard deviation σ

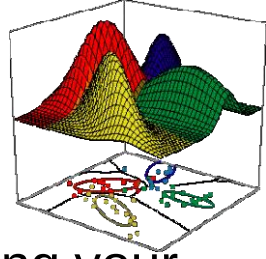
$$Score_{precision} = p(x = Score_{actual}), \text{if } p(x) \sim N\left(Score_{pred}, \sigma^2\right)$$

Project Proposal – 26 Nov



- A **project proposal** presentation detailing:
  - the pattern classification approach that you plan to use,
  - including a source for an implementation of your chosen method.
- This will be a **5 minute** presentation with ~6 slides.
- You will be evaluated on the quality of your presentation and your progress to date (i.e. demonstrate that you've started working, have a software framework in place, understand the problem, etc.)

Project Pitch – 3 Dec

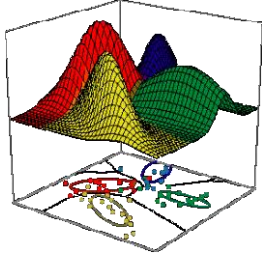


- **The pitch** consisting of a presentation with ~6 slides describing your approach, your predicted accuracy, and how you computed it. Each group will be given **5 minutes** to pitch their method as being the best approach. At the conclusion of this class, all groups will be provided with the blind test data set. Slides should cover:
  a) Quickly review method/implementation
  b) Describe your experiment design
  c) Describe any pre-processing of the data
  d) Describe training/testing protocol
  e) Describe your meta learning strategy (**mandatory**)
  f) Provide your estimated Re@Pr50 (including the standard deviation of your estimate) and describe your methodology for estimating your "true" Re@Pr50

  (i.e. the Re@Pr50 you should expect when applied to new test data).

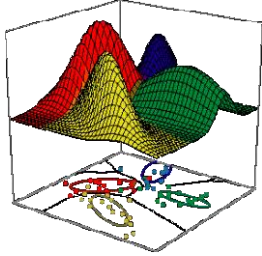
6 slides, 5 minutes each

Order of presentations



#	Approach	Members	
1	**Support vector machines**	Pedro C, Daniel K, Eric M	8:38
2	**Decision Forests**	Andi H, Chanho K, David L	8:43
3	**Linear discriminants**	Kelly B, Pascale J, Shane S	8:48
4	**Convolutional neural networks**	Vishwaa B, Niyati D, Reeham H	8:53
5	**K-nearest-neighbour**	Prathmesh R, Puneet S, Abhinav Y	8:58
6	**Recurrent neural networks**	Joel MK, Maryam TE, Nidheesh V	9:03
7	**Decision trees**	Ben E, Mohamed H, Jason M, Ian S	9:08
8	**Bayesian belief networks**	Anchen L, Zuwen S, Hongzhi Z	9:13
9	**Radial basis function networks**	Anshumaan AA, Ramanjeet K, Navleen KS, Arjun K	9:18
10	**Logistic regression**	Mingfang H, Vishnu R, Yiying Z	9:23
11	**Feed-forward neural networks**	Tarim I, Hamza S, Nizamuddin MS	9:28
12	**K-means clustering**	Kristen B, Victor C, Matthew M	9:33
13	**Probabilistic neural networks**	Ash N, Mohamed Z, ???	9:38
14	**Gradient-boosted decision trees**	Bala PK, Swetha MN, Sreeram S	9:43

Schedule



GOOD LUCK!!!

√

Tuesday 19 Nov:	Competition announced.
Tuesday 26 Nov:	Project proposal presentations
Tuesday 3 Dec:	Pitch presentations given.
3pm Wednesday 4 Dec:	Final classification of blind data submitted to instructor.
Thursday 5 Dec:	Results announced. Winners glorified. Prizes distributed.
Monday 16 Dec:	Final reports submitted **electronically** via CULearn.