Protein Prediction II
Winter 18/19

for Computer Scientists
for Bioinformaticians

8.11.2018

Exercise timeslots

- Every Thursday
- 12:00 - 13:00
- Rostlab seminar room (01.09.034)
- Computer Scientists and Bioinformaticians

- Generally check the websites for changes:
  - https://www.rostlab.org/ws1819/pp2compscient
  - https://www.rostlab.org/wise201819/pp2bioinf

Timeline

25.10.	Introduction
1.11.	No exercise
8.11.	Build a machine learning pipeline
15.11	
22.11.	
29.11.	
6.12.	
13.12.	
20.12.	Maybe no exercise

10.1.	
17.1.	
24.1.	
31.1.	
...	

Exercise topic

- Evaluate sequence representations from the field of natural language processing for bioinformatics applications
- Start with word2vec / ProtVec and improve models, parameters, data, … throughout the term's exercise
- Prediction task: Given a protein sequence, predict protein interaction sites

Bonus: 0.3 on your final exam grade

Requires active participation in your group as well as working code and a final presentation at the end of the semester

Exercise groups

- Keep for the rest of the exercise
- No new groups from this point onwards

1	2	3	4	5	6	7	8
Nabil	Sofie	Kayalvizi	Marla	Jinlong	Felix	Muhammed	Francesco
Vanessa	Ghalia	Amrei	Vagram	Reza	Elisabeth	Mustafa	Muhammad
Lukas	Tobias	Corinna	Michaela	Chris	Issar	Abdulrahman	Silvia
Nathalie	Marco	Daniel	Rinita	Julian	Omar		Aynesh

Protein-protein interaction?

- ~20,000 protein coding genes in human
- ~200,000 transcripts of those genes (alternative splicing, ….)
- For many of these, interactions between transcripts are what drives protein function
- Disturbing interactions leads to issues up to diseases on the organism level
- Exact location of binding sites is important for understanding of how function is performed and might be affected by changes

Protein-protein interaction?

- Potentially multiple binding partners
- Binding sites may be reused for different partners
- Interactions can be permanent or just temporary

Here: Which residues are binding sites (independent of time, partner, …)

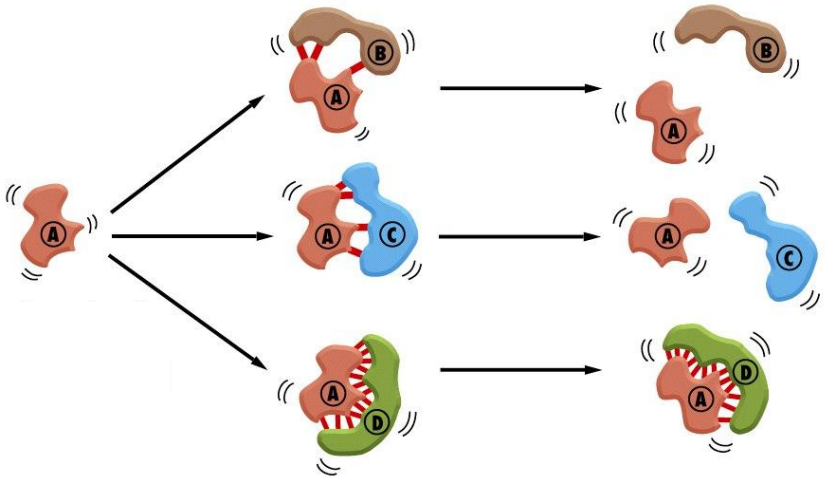
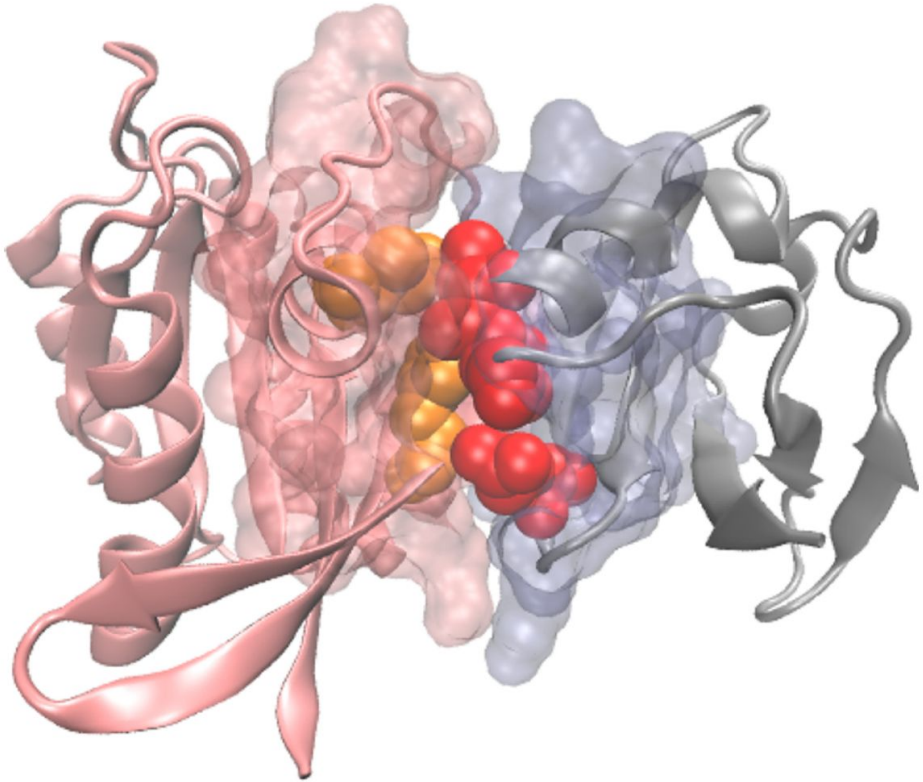


Figure 3-42 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Protein-protein interaction?



Keskin et al., 2016, *Chem Rev*

Tasks from the last two weeks

- Any general questions about the concepts of NLP or ProtVec?
- Everyone able to create ProtVecs with the provided framework?

PPI Dataset

```
>Q72KL7
MVWLNGEPRPLEGKTLKEVLEEMGVELKGVAVLLNEEAFLGLEVPDRPLRDGDVVEVVALMQGG
---++++-+---------------+++-++++++--+++++--+-----------+++++-+++
```

- 540 proteins involved in protein-protein binding (download)
- Each protein is represented by three lines
  - Header with protein name
  - Amino acid sequence
  - Per-residue label (“+” is binding, “-” is non-binding)

Window-based prediction

- How to represent one residue as an input vector?
  - Sliding window approach with residue of interest at the center
  - Sum over all n-grams within window
  - We use window size=7 (i.e. 3 residues to each side)
    - Cannot predict first/last 3 residues (for now)

- Example (with 3-grams):
  Sequence: M**VWLNGEP**RPLEG…
  Vec(**VWLNGEP**) = Vec(**VWL**) + Vec(**WLN**) + Vec(**LNG**) + Vec(**NGE**) + Vec(**GEP**)

Tasks for next week

- Implement a ML pipeline with scikit-learn to predict-protein binding residues
  - Use stratified 10-fold cross-validation (`StratifiedKFold`)
  - Use an artificial neural network (`MLPClassifier`)
  - Optimize parameters (`GridSearchCV`)
- Use the ProtVecs you've created last week and our control (download)
- Next week, report performance
  - Use the `sklearn.metrics` module where possible
  - Accuracy, Precision, Recall, AUC (ROC)

Tasks for next week

- Parameters to optimize (`MLPClassifier`)
  - Hidden layer size (25, 50, 100); only one hidden layer!
  - Initial learning rate (0.001, 0.01)
  - Number of iterations/epochs (100, 200, 500)
  - For everything else, use the default
- Use `random_state=42` wherever possible (reproducibility)
  - `StratifiedKFold`
  - `MLPClassifier`

Contact

- Questions and discussions should happen during weekly exercise meetings
- In case of severe issues that block you from continuing to work, write an email to:
  - reeb@rostlab.org
  - bernhoferm@rostlab.org

Questions?