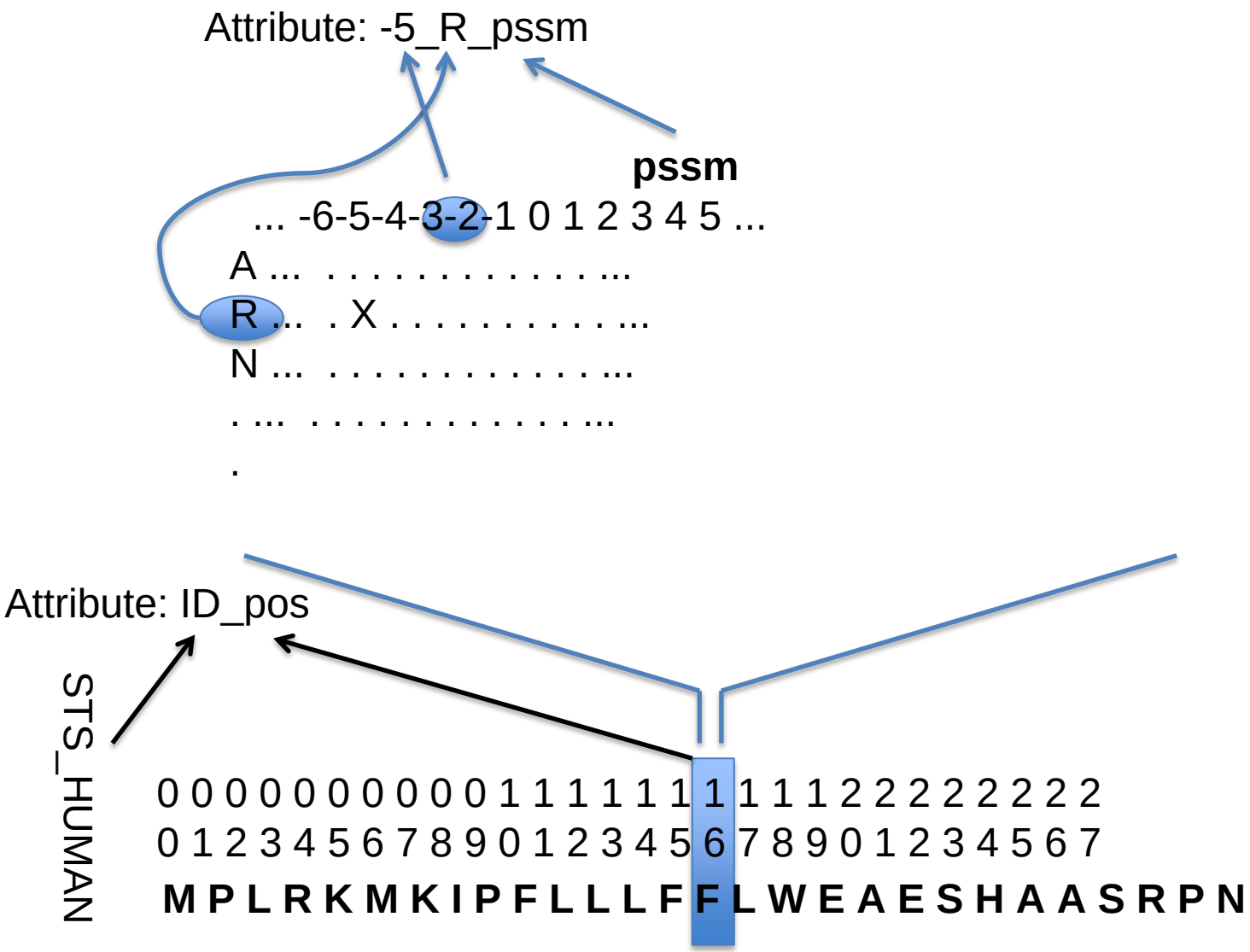
# Protein Prediction I Exercise

# Milestone 2

- You have to provide a stand-alone executable that works on the student machines
- Your program should be executable by any other user without changing environment settings
- Do not hard-code any paths
- Provide a short and easy documentation
- You stand-alone program should talke an “.arff” file as input (same as the training data)
- For every instance your program should return a prediction
- The output has to look like this:

  >Identifier
  ++++++-----++++++++++++++++++++++++++--++++++++----++++++

# Representation of Complex Descriptors as Features in .arff-Format



# Strategic Considerations

- appr. 80 descriptors result in 2886 features that gives possible $2^{2886}$ possible combinations  multiplied with various parameter values => exhaustive search is impossible
- do on the training set:
- start with compulsory features, determine optimal window size
- select an additional feature and determine prediction improvement (forward selection) on the test set, then add another feature
- monitor your time consumption

# Strategic Considerations II

- start with probing the search space to identify meaningful combinations
- with this optimized set of features you can start to optimize your parameter settings using the cross training set
- rotate the folds: in the end every fold should have served once as training, cross training and test set
- average your performance results

Next meetings will be

- 06.06 Group A
- 13.06 Group B

(Presentation of Milestone 2)