### Real World Applications

**Bioinformatics**: Protein localization?

- PKLHPODQP

- **Data Representation**:
  - rich enough to capture distinctions that are relevant from the standpoint of learning
  - but not so rich as to make the task of learning harder due to everything

### Approaches to Feature Construction

- **Sequence Data**:

  - **Challenge**: Data representation provided to a learner
    - **Super-structuring**:
      - The operation of generating all the contiguous sub-sequences of a certain length k from a “super-structure” or k-gram
      - Let $(x_1, x_2, \ldots, x_k)$ be a sequence over a finite set $\mathcal{X}$, $x \in \mathcal{X}$.
      - Helps model dependencies between neighboring elements in a sequence
      - **Abstraction**:
        - In the operation of grouping “similar” entities to generate more abstract entities

### From Data to Knowledge

- **Axioms**:
  - **Antisense**
  - **Amino Acids**
  - **Aromatic**
  - **Basic**
  - **Charged**
  - **Hydrophobic**
  - **Hydrophilic**
  - **Polar**

### DESIGNING ABSTRACT FEATURES FOR SEQUENCE CLASSIFICATION TASKS

#### Constructing Abstractions over K-Grams

- **Greedy agglomerative procedure**
- **Initially map each abstraction to a k-gram**
- **Recursively group pairs of abstractions until m abstractions are obtained**

#### Distance between two abstractions $d_{\alpha}$ and $d_{\beta}$

Let $A$ denote a random variable that takes values in a set of abstractions $A = \{a_1, \ldots, a_n\}$.

- **Coal**: find a set of abstractions $A$ that reduce the mutual information between $A$ and the class variable $Y$, $I(A; Y)$, is minimized at each step of the greedy procedure.

#### Feature Selection

- **Alternative approach to reducing the number of k-grams to m k-grams**
- **We used mutual information between the class variable and k-grams to rank the k-grams**

### Task: Protein Subcellular Localization Prediction

- **Plant data set** [Emanuelsson et al., 2000]
- 940 protein sequences classified into: chloroplast, mitochondrial, secretory pathogen signal peptide, and other
- **Non-plant data set** [Emanuelsson et al., 2000]
- 2,378 protein sequences classified into: mitochondrial, secretory, pathogen signal peptide, and other

### Experiments

We compare Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers trained using:

- **Unigrams**: a bag of letters representation of protein sequences, no super-structuring, abstraction or feature selection (UNIGRAM);
- **Super-structuring and feature selection**: a bag of k-grams (k = 3) chosen using feature selection (from the bag of k-grams obtained by super-structuring (See Section 3 for details)) (SS+SS+SEL);
- **Super-structuring and abstraction**: a bag of m abstractions over k-grams (k = 3) obtained using the combination of super-structuring and abstraction (See Section 2 for details) (SS+FSEL);
- **Combining super-structuring and abstraction**: yields better performing models than those obtained by feature selection in combination with super-structuring.

### Results

![Comparison of super-structuring and abstraction (SS-+[AB]B) with super-structuring alone (SS), super-structuring and feature selection (SS+FSEL), and UNIGRAM on the plant and non-plant data sets using Naïve Bayes (NB) (left column), and Support Vector Machines (SVM) with linear kernel (right column). The plots show the accuracy as a function of the number of features used in the classification model, ranging from 1 to ~8,000 on both data sets. The x-axis shows the number of features on a logarithmic scale.](image)

### Analysis of Abstractions

![Class probability distributions induced by one of the m abstractions, namely $a_i$, and by three 3-grams, namely “I/L”, “SS”, and “PSF”, on the plant data set, where n = 10 and m = 100 and k = 3 (right). The three 3-grams are initially sampled from $a_i$ (when n = 100). The number of classes in the data set is 4.](image)