```
1. Title: Protein Localization Sites


2. Creator and Maintainer:
             Kenta Nakai
             Institue of Molecular and Cellular Biology
             Osaka, University
             1-3 Yamada-oka, Suita 565 Japan
             nakai@imcb.osaka-u.ac.jp
             http://www.imcb.osaka-u.ac.jp/nakai/psort.html
     Donor: Paul Horton (paulh@cs.berkeley.edu)
     Date:  September, 1996
     See also: yeast database

3. Past Usage.
Reference: "A Probablistic Classification System for Predicting the Cellular
           Localization Sites of Proteins", Paul Horton & Kenta Nakai,
           Intelligent Systems in Molecular Biology, 109-115.
           St. Louis, USA 1996.
Results: 81% for E.coli with an ad hoc structured
           probability model. Also similar accuracy for Binary Decision Tree and
           Bayesian Classifier methods applied by the same authors in
           unpublished results.

Predicted Attribute: Localization site of protein. ( non-numeric ).


4. The references below describe a predecessor to this dataset and its
development. They also give results (not cross-validated) for classification
by a rule-based expert system with that version of the dataset.

Reference: "Expert Sytem for Predicting Protein Localization Sites in
           Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa,
           PROTEINS: Structure, Function, and Genetics 11:95-110, 1991.

Reference: "A Knowledge Base for Predicting Protein Localization Sites in
           Eukaryotic Cells", Kenta Nakai & Minoru Kanehisa,
           Genomics 14:897-911, 1992.


5. Number of Instances:  336 for the E.coli dataset and


6. Number of Attributes.
        for E.coli dataset:  8 ( 7 predictive, 1 name )


7. Attribute Information.

   1.   Sequence Name: Accession number for the SWISS-PROT database
   2.   mcg: McGeoch's method for signal sequence recognition.
   3.   gvh: von Heijne's method for signal sequence recognition.
   4.   lip: von Heijne's Signal Peptidase II consensus sequence score.
            Binary attribute.
   5.   chg: Presence of charge on N-terminus of predicted lipoproteins.
            Binary attribute.
   6.   aac: score of discriminant analysis of the amino acid content of
            outer membrane and periplasmic proteins.
   7.  alm1: score of the ALOM membrane spanning region prediction program.
   8.  alm2: score of ALOM program after excluding putative cleavable signal
            regions from the sequence.



8. Missing Attribute Values: None.


9. Class Distribution. The class is the localization site. Please see Nakai &
                       Kanehisa referenced above for more details.
```

```
cp  (cytoplasm)                                   143
im  (inner membrane without signal sequence)       77
pp  (perisplasm)                                   52
imU (inner membrane, uncleavable signal sequence)  35
om  (outer membrane)                               20
omL (outer membrane lipoprotein)                    5
imL (inner membrane lipoprotein)                    2
imS (inner membrane, cleavable signal sequence)     2
```