The data set (and description) can be downloaded here:
http://archive.ics.uci.edu/ml/machine-learning-databases/ecoli/ecoli.data

Description:

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 robability 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 Also similar accuracy for Binary Decision Tree and Bayesian Classifier methods applied by the same authors in unpublished results.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 (non cross-validated) for classification by a rule-based expert system with that version of the dataset.

Reference: "Expert System 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,
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

Citation Request:

Please refer to the repository http://archive.ics.uci.edu/ml (see citation policy).
See also Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].
Irvine, CA: University of California, School of Information and Computer Science.

Descriptive statistics:

Dataset= ecoli_imvspp : n= 129 , d= 5

Class1: n= 77

Covariance matrix:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.0378	-0.0001	0.0032	0.0028	0.0099
[2,]	-0.0001	0.0078	-0.0001	-0.0009	-0.0019
[3,]	0.0032	-0.0001	0.0130	-0.0005	0.0078
[4,]	0.0028	-0.0009	-0.0005	0.0107	0.0088
[5,]	0.0099	-0.0019	0.0078	0.0088	0.0278

Correlation matrix:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.0000	-0.0086	0.1426	0.1418	0.3063
[2,]	-0.0086	1.0000	-0.0107	-0.0940	-0.1323
[3,]	0.1426	-0.0107	1.0000	-0.0439	0.4074
[4,]	0.1418	-0.0940	-0.0439	1.0000	0.5108
[5,]	0.3063	-0.1323	0.4074	0.5108	1.0000

Median:	0.494	0.4923	0.551	0.7627	0.7638
Mean:	0.4784	0.4966	0.5361	0.7575	0.7304
MCD-estimated:					
MDC-0.975-Mean:	0.4982	0.495	0.5596	0.767	0.7929
MDC-0.750-Mean:	0.492	0.496	0.5611	0.768	0.7938
MDC-0.500-Mean:	0.492	0.496	0.5611	0.768	0.7938

```
Class2: n= 52

Covariance matrix:
        [,1]    [,2]    [,3]    [,4]    [,5]
[1,]  0.0082  0.0049 -0.0017  0.0035 -0.0011
[2,]  0.0049  0.0167 -0.0046  0.0008 -0.0012
[3,] -0.0017 -0.0046  0.0073 -0.0006  0.0012
[4,]  0.0035  0.0008 -0.0006  0.0102  0.0040
[5,] -0.0011 -0.0012  0.0012  0.0040  0.0140

Correlation matrix:
        [,1]    [,2]    [,3]    [,4]    [,5]
[1,]  1.0000  0.4169 -0.2151  0.3791 -0.1011
[2,]  0.4169  1.0000 -0.4196  0.0649 -0.0761
[3,] -0.2151 -0.4196  1.0000 -0.0658  0.1182
[4,]  0.3791  0.0649 -0.0658  1.0000  0.3309
[5,] -0.1011 -0.0761  0.1182  0.3309  1.0000

Median:           0.6617 0.7065 0.4372 0.4572 0.3642

Mean:             0.6521 0.6998 0.4367 0.4681 0.3744
MCD-estimated:
MDC-0.975-Mean:   0.6741 0.7227 0.4312 0.4595 0.3571
MDC-0.750-Mean:   0.6724 0.7213 0.4292 0.4584 0.3584
MDC-0.500-Mean:   0.6726 0.7238 0.4279 0.4633 0.3636


Measures:
Mah.Dist:                        3.9434
Mah.Dist-MCD-0.975:              4.7943
Mah.Dist-MCD-0.750:              4.7653
Mah.Dist-MCD-0.500:              4.7943

All the MCD estimates have been obtained after a slight perturbation of the
data set
```