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: ecoli 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: 55% for Yeast data 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unpublished results.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: 1484 for the Yeast dataset.

6. Number of Attributes.

for Yeast dataset: 9 ( 8 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. alm: Score of the ALOM membrane spanning region prediction program.

5. mit: Score of discriminant analysis of the amino acid content of

the N-terminal region (20 residues long) of mitochondrial and

non-mitochondrial proteins.

6. erl: Presence of "HDEL" substring (thought to act as a signal for

retention in the endoplasmic reticulum lumen). Binary attribute.

7. pox: Peroxisomal targeting signal in the C-terminus.

8. vac: Score of discriminant analysis of the amino acid content of

vacuolar and extracellular proteins.

9. nuc: Score of discriminant analysis of nuclear localization signals

of nuclear and non-nuclear proteins.

8. Missing Attribute Values: None.

9. Class Distribution. The class is the localization site. Please see Nakai &

Kanehisa referenced above for more details.

CYT (cytosolic or cytoskeletal) 463

NUC (nuclear) 429

MIT (mitochondrial) 244

ME3 (membrane protein, no N-terminal signal) 163

ME2 (membrane protein, uncleaved signal) 51

ME1 (membrane protein, cleaved signal) 44

EXC (extracellular) 37

VAC (vacuolar) 30

POX (peroxisomal) 20

ERL (endoplasmic reticulum lumen) 5