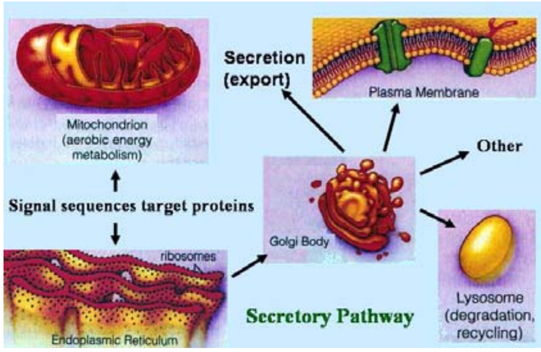
For written notes on this lecture, please read Chapter 9 of *The Practical Bioinformatician*

# Knowledge Discovery Techniques for Bioinformatics, Part V-1: Applications to Protein Subcellular Localization Prediction

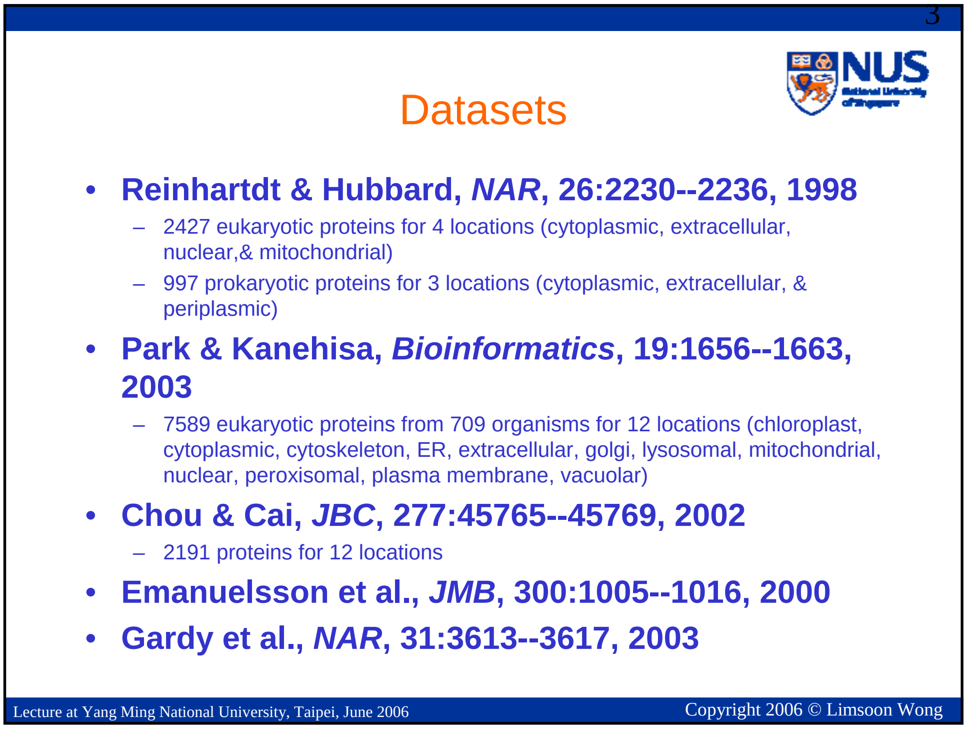
**Limsoon Wong**

## Compartments and Sorting


Mitochondrion (aerobic energy metabolism)
Secretion (export)
Plasma Membrane
Other
Signal sequences target proteins
Golgi Body
ribosomes
Endoplasmic Reticulum
Secretory Pathway
Lysosome (degradation, recycling)

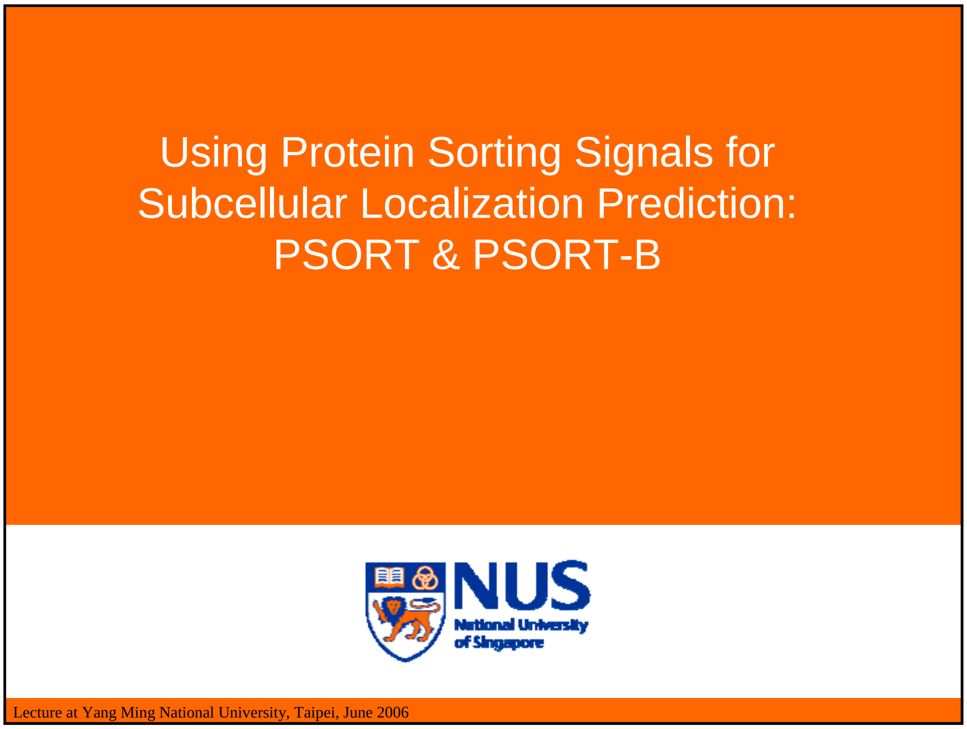
- **Protein sorting is determined by specific amino acid sequences, or "signals", within the protein**
- **Secretory pathway targets proteins to plasma membrane, some membrane-bound organelles such as lysosomes, or to export proteins from the cell**

- **Eukaryotic cells requires proteins be targeted to their subcellular destinations**



## Datasets

- **Reinhartdt & Hubbard, *NAR*, 26:2230--2236, 1998**
  - 2427 eukaryotic proteins for 4 locations (cytoplasmic, extracellular, nuclear,& mitochondrial)
  - 997 prokaryotic proteins for 3 locations (cytoplasmic, extracellular, & periplasmic)
- **Park & Kanehisa, *Bioinformatics*, 19:1656--1663, 2003**
  - 7589 eukaryotic proteins from 709 organisms for 12 locations (chloroplast, cytoplasmic, cytoskeleton, ER, extracellular, golgi, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane, vacuolar)
- **Chou & Cai, *JBC*, 277:45765--45769, 2002**
  - 2191 proteins for 12 locations
- **Emanuelsson et al., *JMB*, 300:1005--1016, 2000**
- **Gardy et al., *NAR*, 31:3613--3617, 2003**



## Using Protein Sorting Signals for Subcellular Localization Prediction: PSORT & PSORT-B

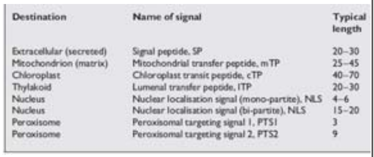
# Common Eukaryotic Protein Sorting Signals

| Destination | Name of signal | Typical length |
|---|---|---|
| Extracellular (secreted) | Signal peptide, SP | 20–30 |
| Mitochondrion (matrix) | Mitochondrial transfer peptide, mTP | 25–45 |
| Chloroplast | Chloroplast transit peptide, cTP | 40–70 |
| Thylakoid | Lumenal transfer peptide, ITP | 20–30 |
| Nucleus | Nuclear localisation signal (mono-partite), NLS | 4–6 |
| Nucleus | Nuclear localisation signal (bi-partite), NLS | 15–20 |
| Peroxisome | Peroxisomal targeting signal 1, PTS1 | 3 |
| Peroxisome | Peroxisomal targeting signal 2, PTS2 | 9 |

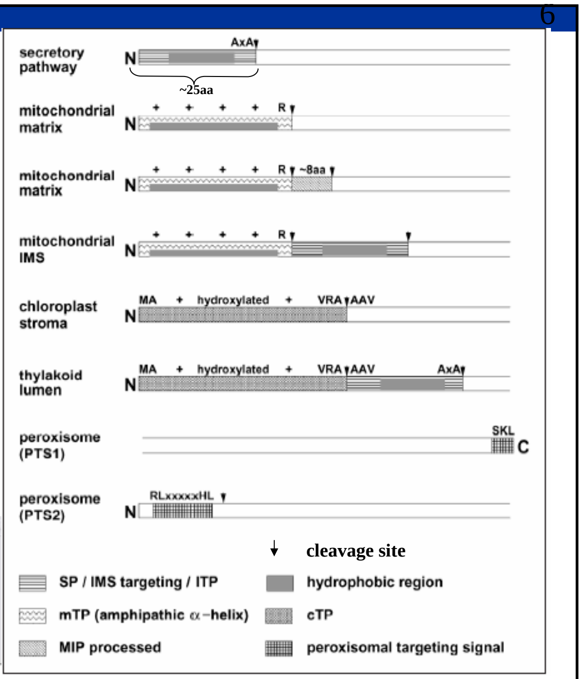
For a comprehensive list of cellular localization sites, see

http://mendel.imp.univie.ac.at/CELL_LOC/index.html

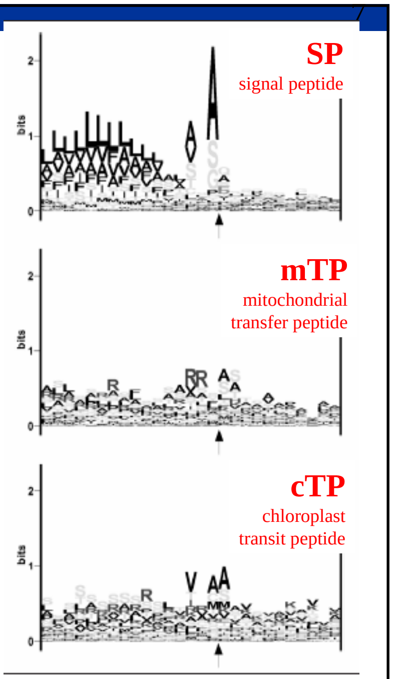
# Schematic View of Sorting Signals

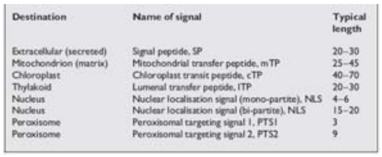


| Destination | Name of signal | Typical length |
|---|---|---|
| Extracellular (secreted) | Signal peptide, SP | 20–30 |
| Mitochondrion (matrix) | Mitochondrial transfer peptide, mTP | 25–45 |
| Chloroplast | Chloroplast transit peptide, cTP | 40–70 |
| Thylakoid | Lumenal transfer peptide, ITP | 20–30 |
| Nucleus | Nuclear localisation signal (mono-partite), NLS | 4–6 |
| Nucleus | Nuclear localisation signal (bi-partite), NLS | 15–20 |
| Peroxisome | Peroxisomal targeting signal 1, PTS1 | 3 |
| Peroxisome | Peroxisomal targeting signal 2, PTS2 | 9 |



## Sequence Logos of SP, mTP, & cTP



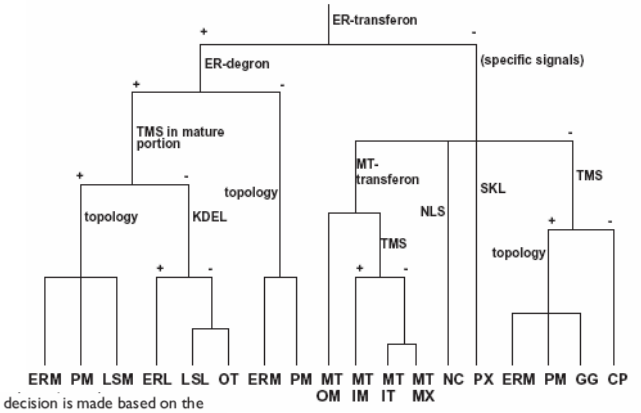


| Destination | Name of signal | Typical length |
|---|---|---|
| Extracellular (secreted) | Signal peptide, SP | 20–30 |
| Mitochondrion (matrix) | Mitochondrial transfer peptide, mTP | 25–45 |
| Chloroplast | Chloroplast transit peptide, cTP | 40–70 |
| Thylakoid | Lumenal transfer peptide, lTP | 20–30 |
| Nucleus | Nuclear localisation signal (mono-partite), NLS | 4–6 |
| Nucleus | Nuclear localisation signal (bi-partite), NLS | 15–20 |
| Peroxisome | Peroxisomal targeting signal 1, PTS1 | 3 |
| Peroxisome | Peroxisomal targeting signal 2, PTS2 | 9 |

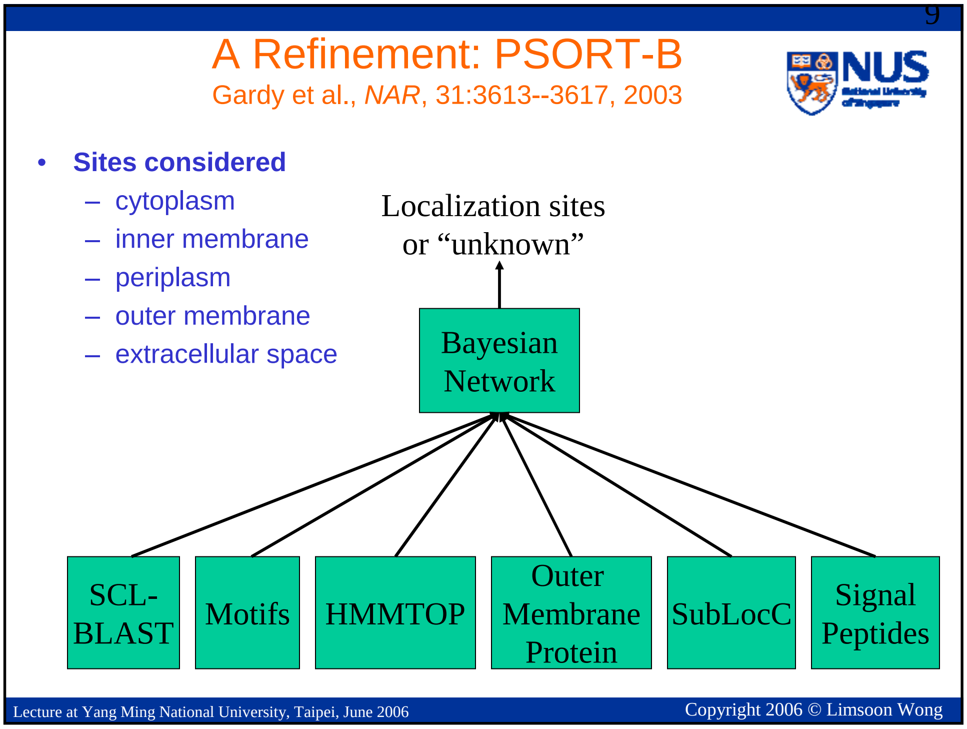
## Expert System Approach: PSORT

Horton & Nakai, *ISMB*, 1997

A simplified version of the decision tree that PSORT uses to check and reason over various sorting signals



At each node, a decision is made based on the result of the corresponding calculation. (+), yes; (−), no; *ER*, endoplasmic reticulum; *TMS*, transmembrane segment; *KDEL*, ER retention signal; *NLS*, nuclear localisation signal; *SKL*, peroxisomal location signal; *PM*, integral plasma membrane; *LSM*, lysosome membrane; *ERL*, endoplasmic reticulum lumen; *LSL*, lysosome lumen; *OT*, extracellular; *MT*, mitochondrion (*OM*, outer membrane; *IM*, inner membrane; *IT*, intermembrane space; *MX*, matrix); *NC*, nuclear; *PX*, peroxisomal; *GG*, Golgi complex; *CP*, cytoplasmic



## A Refinement: PSORT-B

Gardy et al., *NAR*, 31:3613--3617, 2003

- **Sites considered**
  - cytoplasm
  - inner membrane
  - periplasm
  - outer membrane
  - extracellular space

Localization sites
or "unknown"

Bayesian Network

SCL-BLAST

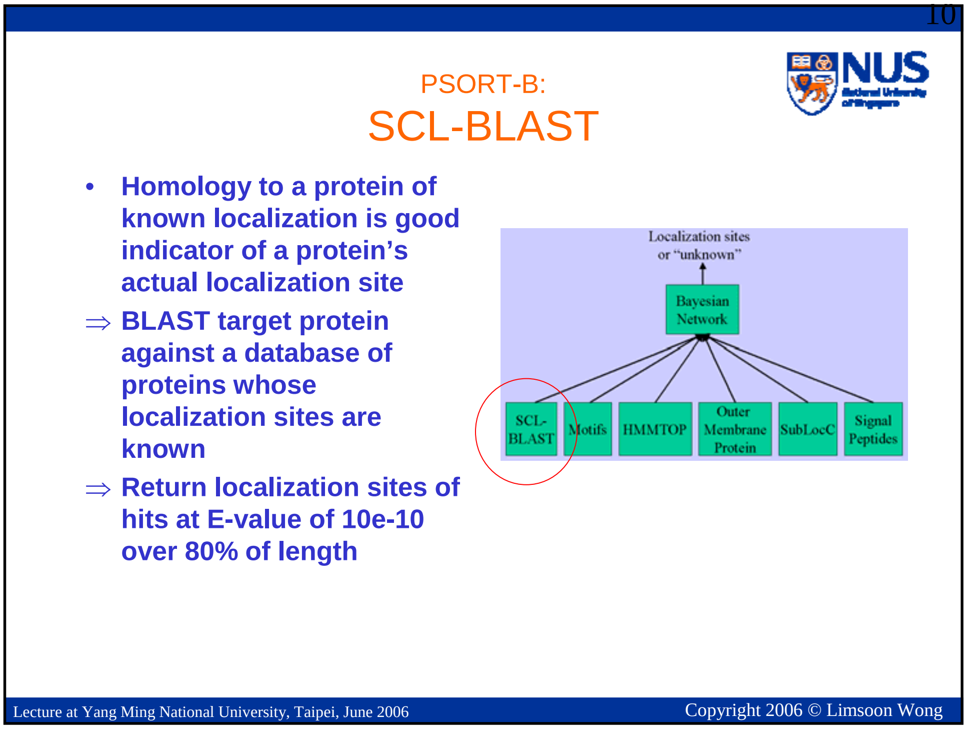
Motifs

HMMTOP

Outer Membrane Protein

SubLocC

Signal Peptides

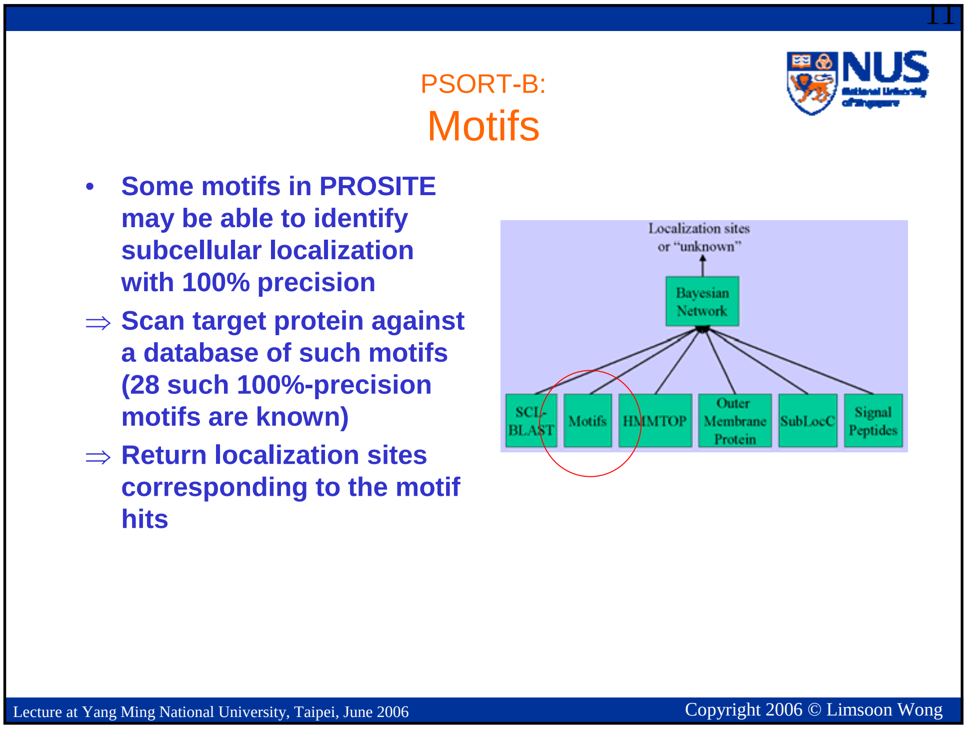


## PSORT-B:
## SCL-BLAST

- **Homology to a protein of known localization is good indicator of a protein's actual localization site**

⇒ **BLAST target protein against a database of proteins whose localization sites are known**

⇒ **Return localization sites of hits at E-value of 10e-10 over 80% of length**



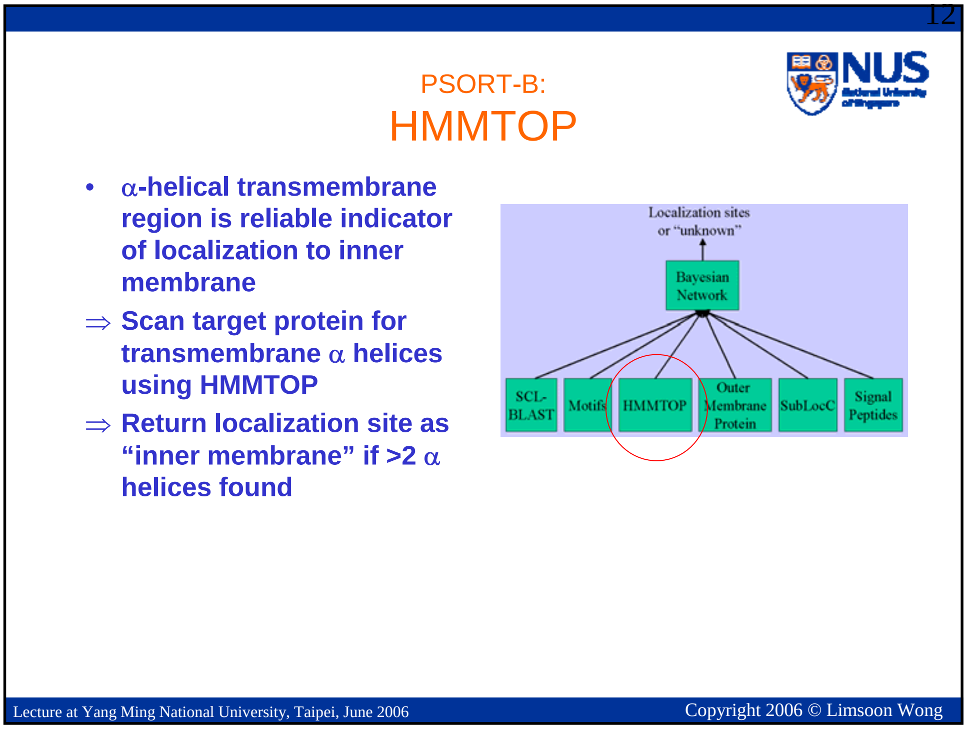
## PSORT-B:
# Motifs

- Some motifs in PROSITE may be able to identify subcellular localization with 100% precision

⇒ Scan target protein against a database of such motifs (28 such 100%-precision motifs are known)

⇒ Return localization sites corresponding to the motif hits

Localization sites
or "unknown"
Bayesian Network
SCL-BLAST
Motifs
HMMTOP
Outer Membrane Protein
SubLocC
Signal Peptides



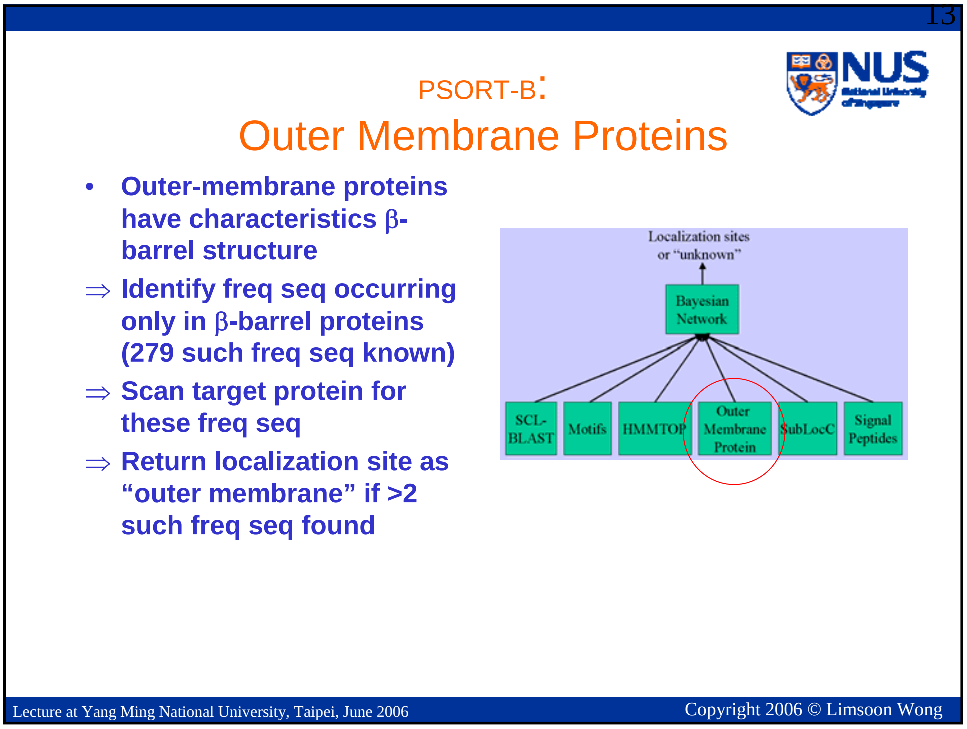
## PSORT-B:
# HMMTOP

- $\alpha$-helical transmembrane region is reliable indicator of localization to inner membrane

⇒ Scan target protein for transmembrane $\alpha$ helices using HMMTOP

⇒ Return localization site as "inner membrane" if >2 $\alpha$ helices found

Localization sites
or "unknown"
Bayesian Network
SCL-BLAST
Motifs
HMMTOP
Outer Membrane Protein
SubLocC
Signal Peptides



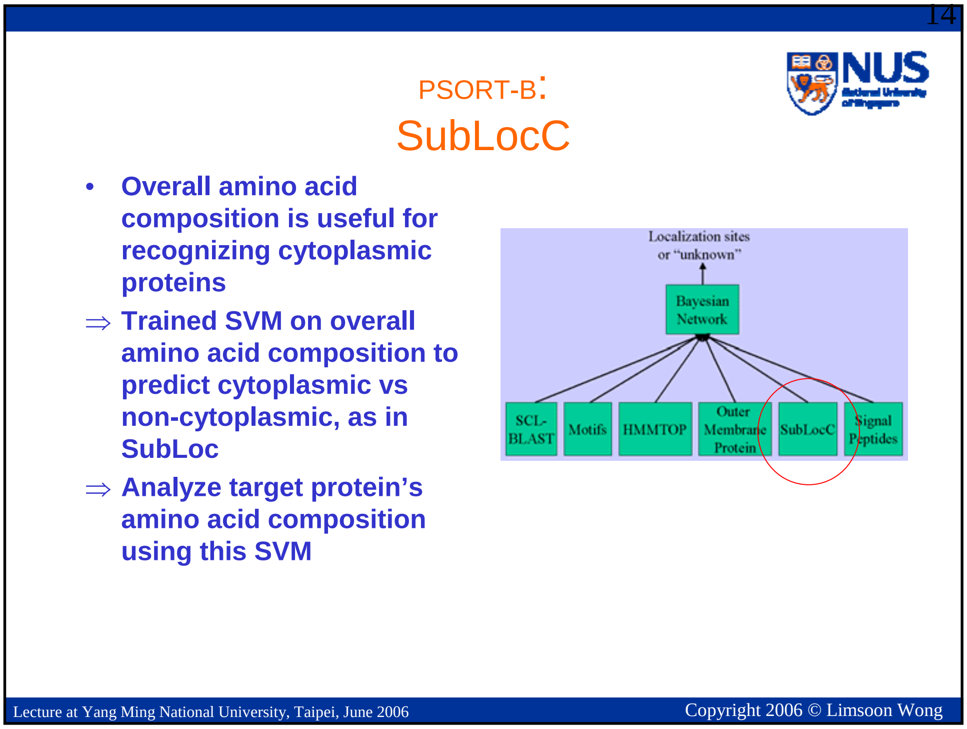
## PSORT-B: Outer Membrane Proteins

- Outer-membrane proteins have characteristics β-barrel structure

⇒ Identify freq seq occurring only in β-barrel proteins (279 such freq seq known)

⇒ Scan target protein for these freq seq

⇒ Return localization site as "outer membrane" if >2 such freq seq found



## PSORT-B: SubLocC

- Overall amino acid composition is useful for recognizing cytoplasmic proteins

⇒ Trained SVM on overall amino acid composition to predict cytoplasmic vs non-cytoplasmic, as in SubLoc

⇒ Analyze target protein's amino acid composition using this SVM

## PSORT-B:

# Signal Peptides

- Presence of signal peptide at N-terminal means protein not cytoplasmic

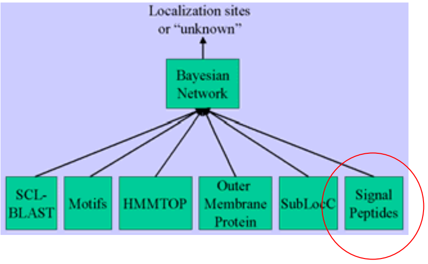
⇒ Train HMM & SVM to recognize signal peptides and their cleavage sites

⇒ If high-confidence cleavage site found by HMM in first 70aa of target protein, then “non-cytoplasmic”

⇒ If low-confidence cleavage site found, pass candidate signal peptide to SVM to confirm

⇒ If confirmed, then “non-cytoplasmic”

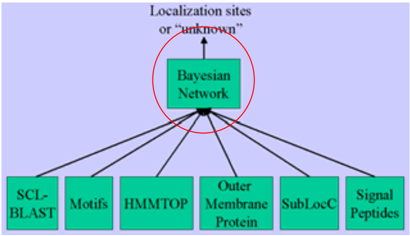
⇒ Otherwise, “unknown”

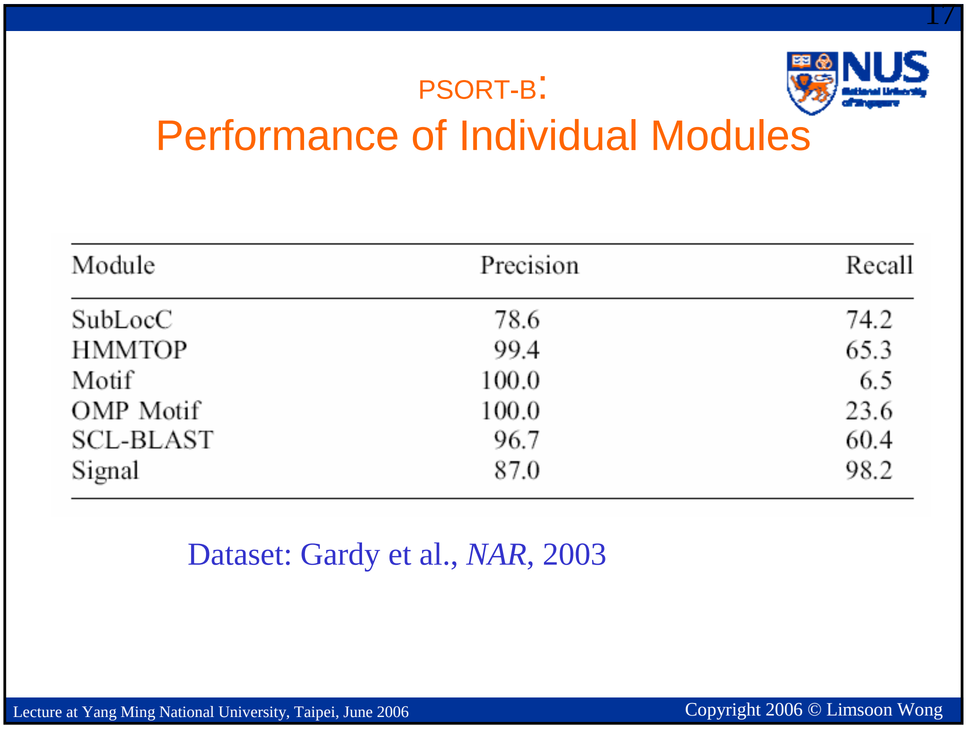


## PSORT-B:

# Bayesian Network

- Bayesian Network integrates results from the 6 modules
- Produces a score for each of the 5 possible localization sites
- If a site scores >7.5, then predicts as a localization site of the target protein
- If no site scores >7.5, then makes no prediction



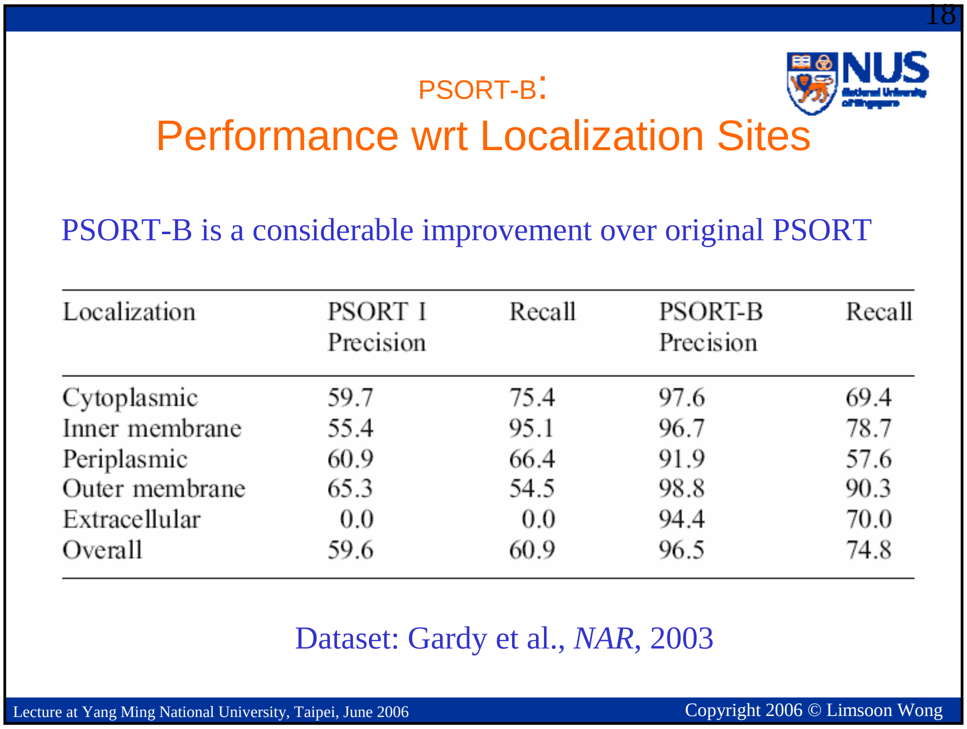


PSORT-B:

## Performance of Individual Modules

| Module | Precision | Recall |
|---|---|---|
| SubLocC | 78.6 | 74.2 |
| HMMTOP | 99.4 | 65.3 |
| Motif | 100.0 | 6.5 |
| OMP Motif | 100.0 | 23.6 |
| SCL-BLAST | 96.7 | 60.4 |
| Signal | 87.0 | 98.2 |

Dataset: Gardy et al., *NAR*, 2003



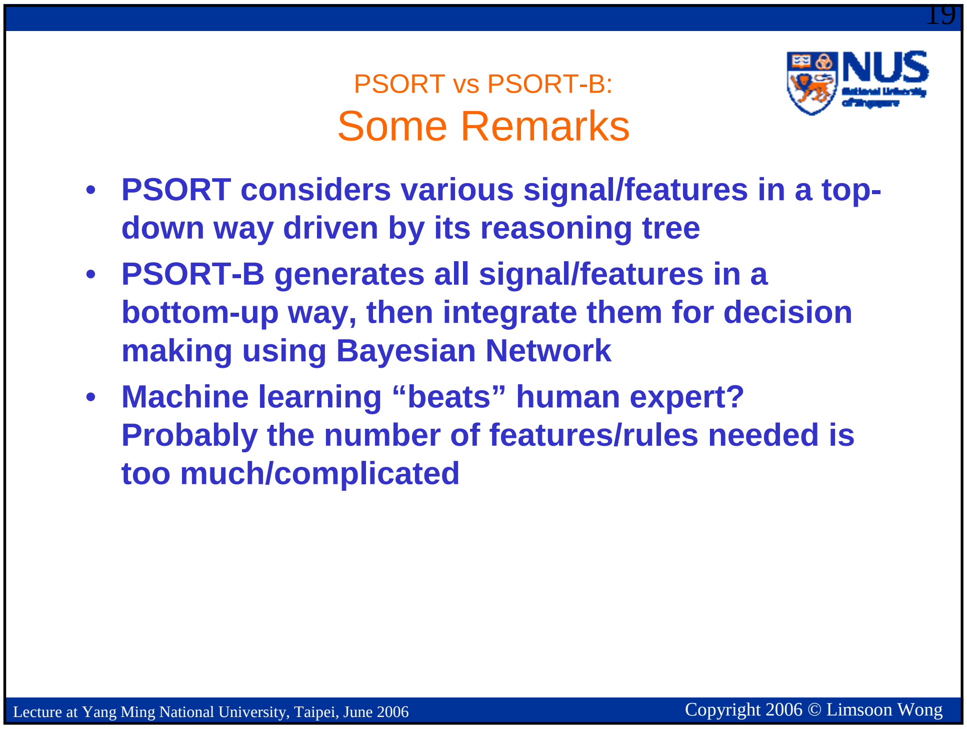
PSORT-B:

## Performance wrt Localization Sites

PSORT-B is a considerable improvement over original PSORT

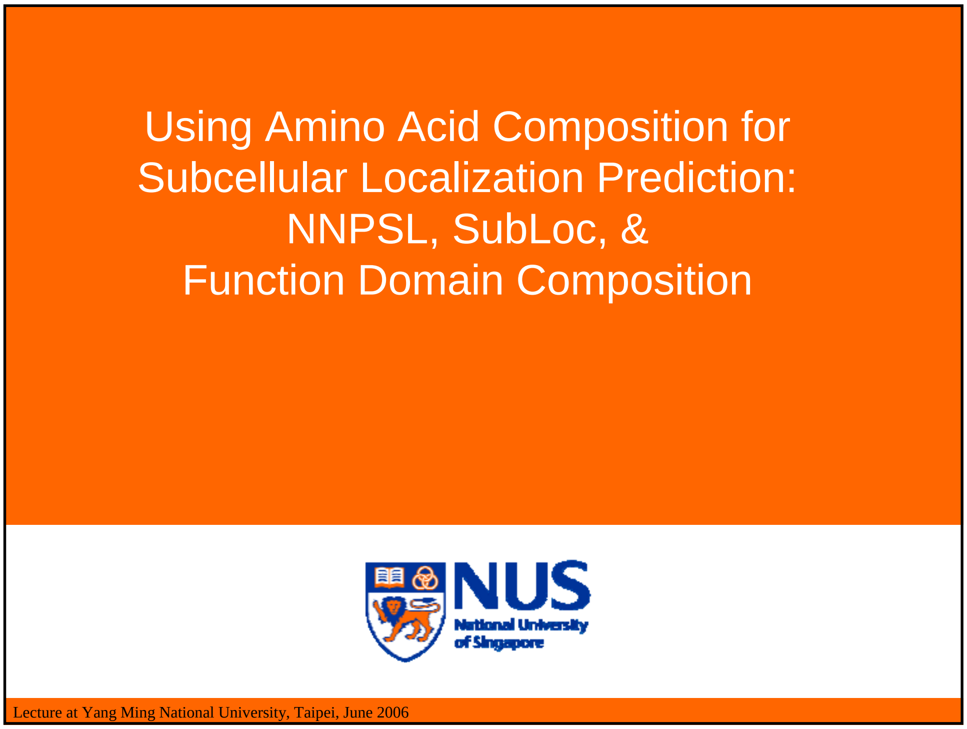
| Localization | PSORT I<br>Precision | Recall | PSORT-B<br>Precision | Recall |
|---|---|---|---|---|
| Cytoplasmic | 59.7 | 75.4 | 97.6 | 69.4 |
| Inner membrane | 55.4 | 95.1 | 96.7 | 78.7 |
| Periplasmic | 60.9 | 66.4 | 91.9 | 57.6 |
| Outer membrane | 65.3 | 54.5 | 98.8 | 90.3 |
| Extracellular | 0.0 | 0.0 | 94.4 | 70.0 |
| Overall | 59.6 | 60.9 | 96.5 | 74.8 |

Dataset: Gardy et al., *NAR*, 2003



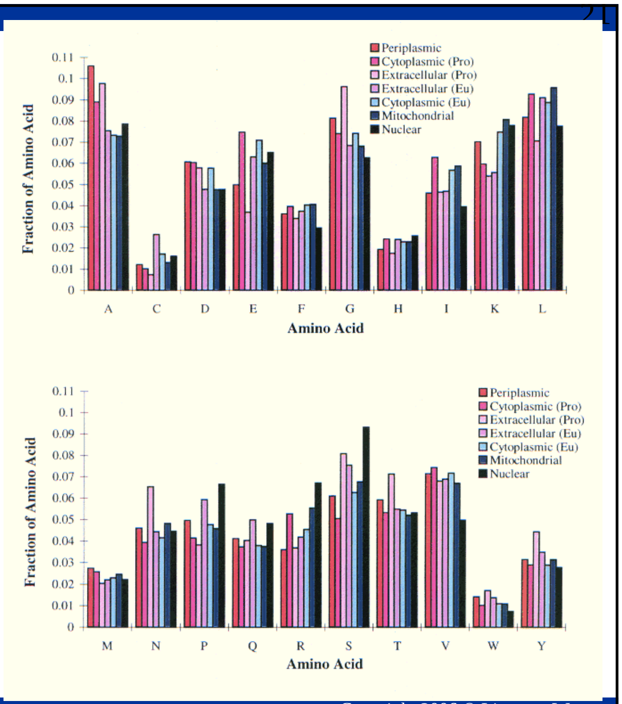
## PSORT vs PSORT-B: Some Remarks

- PSORT considers various signal/features in a top-down way driven by its reasoning tree
- PSORT-B generates all signal/features in a bottom-up way, then integrate them for decision making using Bayesian Network
- Machine learning “beats” human expert? Probably the number of features/rules needed is too much/complicated



## Using Amino Acid Composition for Subcellular Localization Prediction: NNPSL, SubLoc, & Function Domain Composition

## Amino acid composition of proteins residing in different sites are different



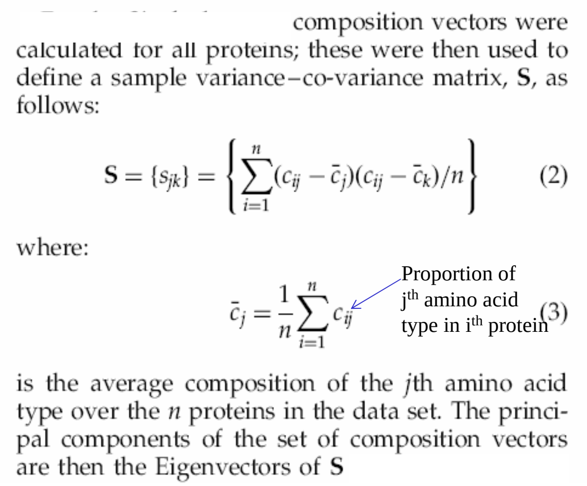
## Amino Acid Composition Differences

- each cellular location has own characteristic physio-chemical environment
- proteins in each location have adapted thru evolution to that environment
- thus reflected in the protein structure and amino acid composition
- If the above is true, the amino acid composition differences wrt cellular location sites should be more pronounced on protein surfaces than protein interior
- Exercise: Why?

# Adaptation of Protein Surfaces

Andrade et al., *JMB*, 1998

- To test the theory of adaptation of protein surfaces to subcellular localization, we do a plot of 3 types of composition vectors along their first two principal components



composition vectors were calculated for all proteins; these were then used to define a sample variance–co-variance matrix, **S**, as follows:

$$\mathbf{S} = \{s_{jk}\} = \left\{ \sum_{i=1}^{n} (c_{ij} - \bar{c}_j)(c_{ij} - \bar{c}_k)/n \right\} \qquad (2)$$

where:

$$\bar{c}_j = \frac{1}{n} \sum_{i=1}^{n} c_{ij} \qquad (3)$$

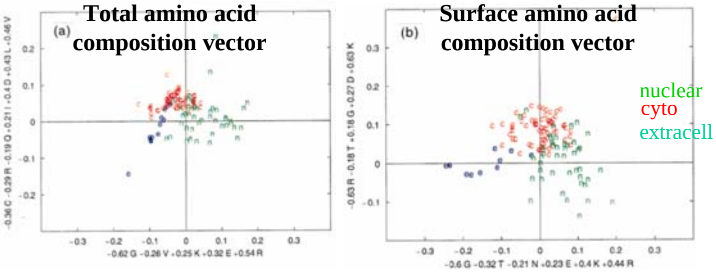
Proportion of $j^{th}$ amino acid type in $i^{th}$ protein

is the average composition of the $j$th amino acid type over the $n$ proteins in the data set. The principal components of the set of composition vectors are then the Eigenvectors of **S**

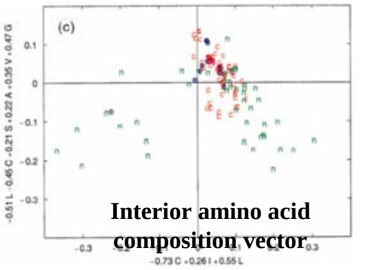
# Adaptation of Protein Surfaces

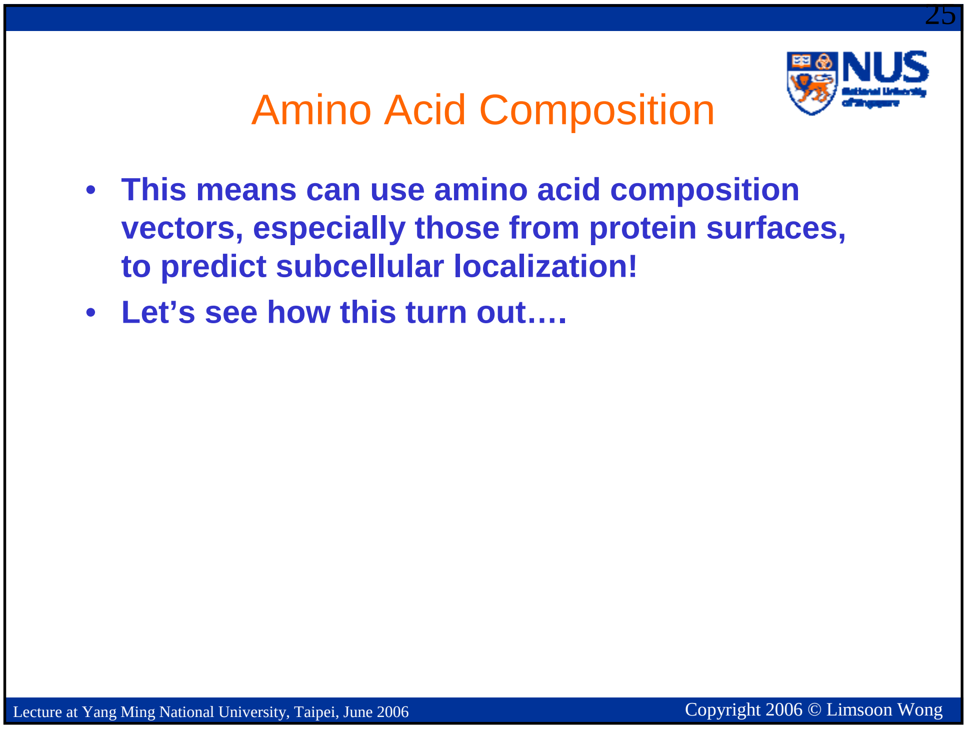
Andrade et al., *JMB*, 1998

EXTRA-CELLULAR SPACE
NUCLEUS
CYTOPLASM



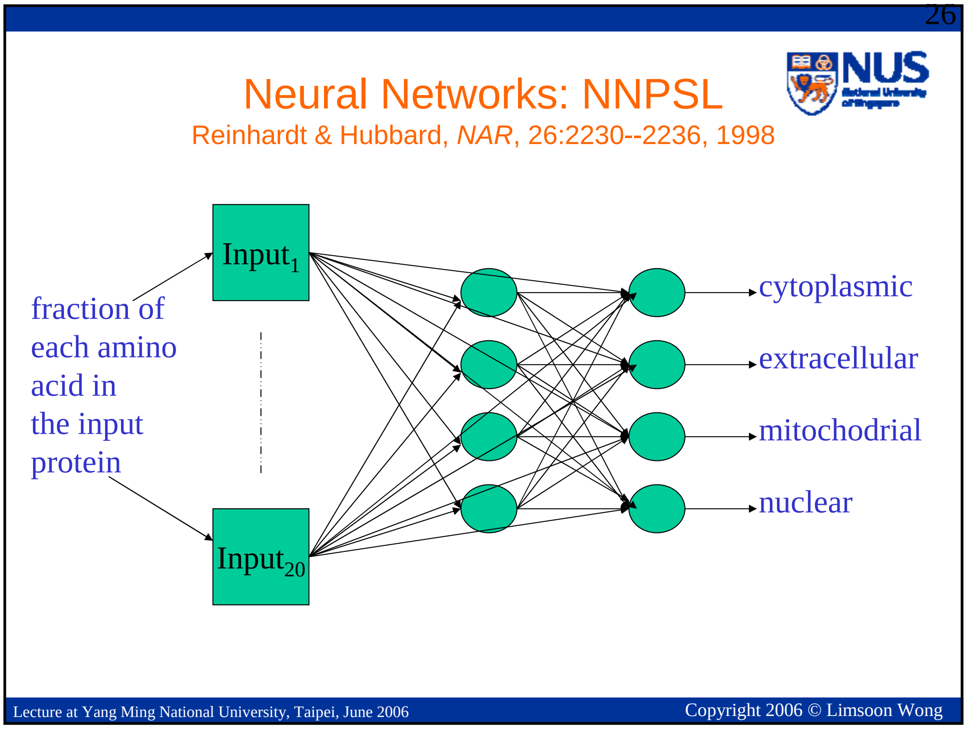
- Clearly total & surface composition vectors show better separation than interior composition vectors





## Amino Acid Composition

- This means can use amino acid composition vectors, especially those from protein surfaces, to predict subcellular localization!
- Let’s see how this turn out….



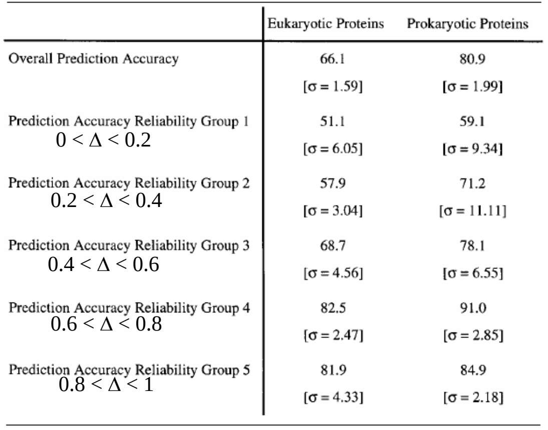
## Neural Networks: NNPSL

Reinhardt & Hubbard, *NAR*, 26:2230--2236, 1998

## NNPSL: Performance

- **Outputs of NNPSL have values 0 to 1. The difference (Δ) between the highest and the next highest nodes can be used as a reliability index**

Dataset:
Reinhardt & Hubbard, *NAR*, 1998

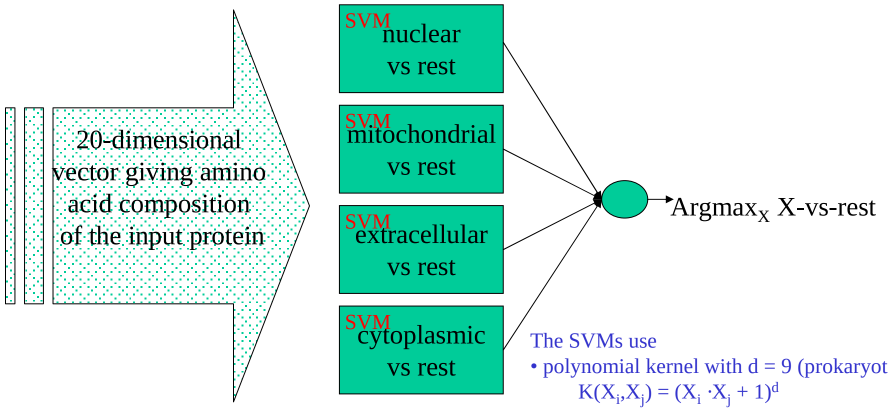


| | Eukaryotic Proteins | Prokaryotic Proteins |
|---|---|---|
| Overall Prediction Accuracy | 66.1<br>[σ = 1.59] | 80.9<br>[σ = 1.99] |
| Prediction Accuracy Reliability Group 1<br>$0 < \Delta < 0.2$ | 51.1<br>[σ = 6.05] | 59.1<br>[σ = 9.34] |
| Prediction Accuracy Reliability Group 2<br>$0.2 < \Delta < 0.4$ | 57.9<br>[σ = 3.04] | 71.2<br>[σ = 11.11] |
| Prediction Accuracy Reliability Group 3<br>$0.4 < \Delta < 0.6$ | 68.7<br>[σ = 4.56] | 78.1<br>[σ = 6.55] |
| Prediction Accuracy Reliability Group 4<br>$0.6 < \Delta < 0.8$ | 82.5<br>[σ = 2.47] | 91.0<br>[σ = 2.85] |
| Prediction Accuracy Reliability Group 5<br>$0.8 < \Delta < 1$ | 81.9<br>[σ = 4.33] | 84.9<br>[σ = 2.18] |

Summary of the prediction accuracy achieved by the neural networks for eukaryotic and prokaryotic sequences. Shown is the overall accuracy and the accuracy for the various reliability groups together with the standard deviation σ as yielded by cross validation tests.

## Support Vector Machines: SubLoc

Hua & Sun, *Bioinformatics*, 17:721--728, 2001



20-dimensional vector giving amino acid composition of the input protein

SVM nuclear vs rest

SVM mitochondrial vs rest

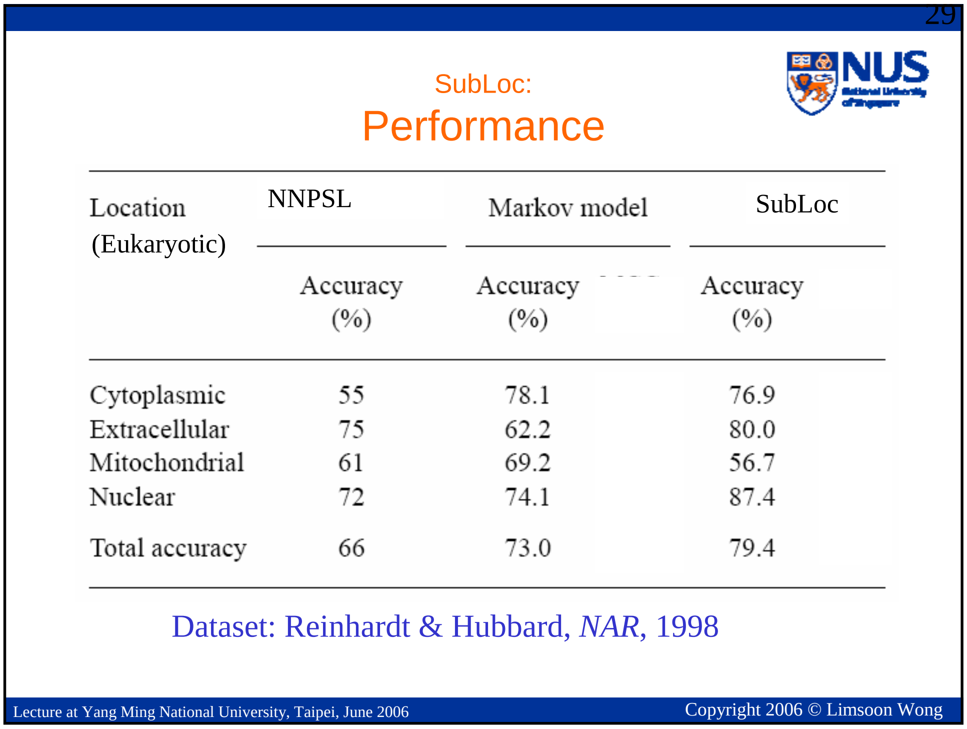
SVM extracellular vs rest

SVM cytoplasmic vs rest

$\text{Argmax}_X$ X-vs-rest

The SVMs use

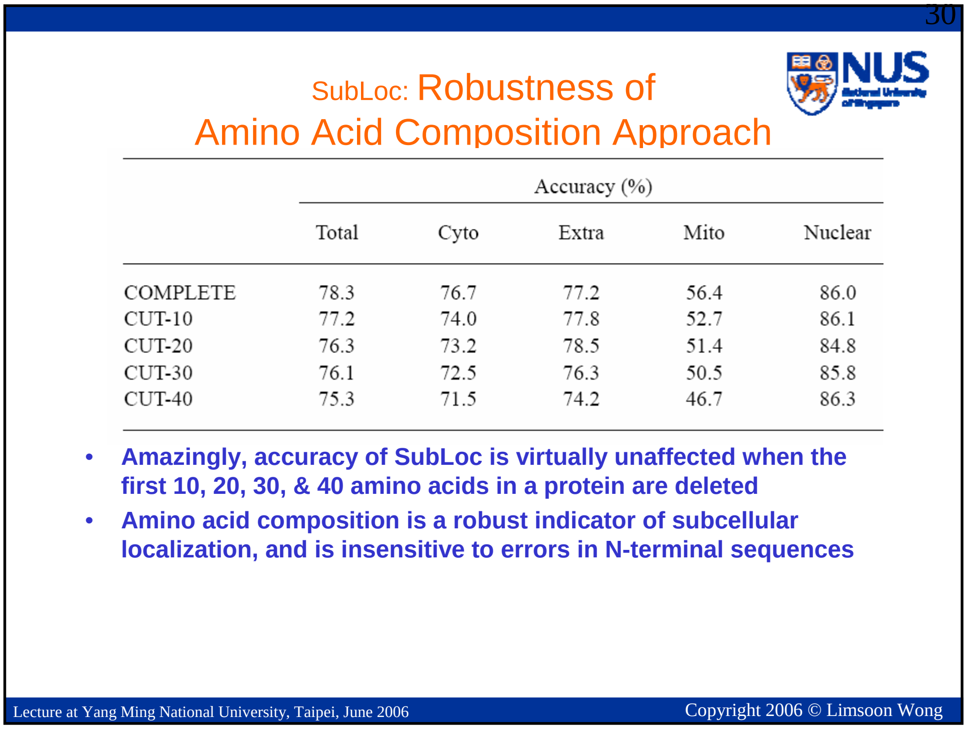
- polynomial kernel with d = 9 (prokaryotic),
  $K(X_i,X_j) = (X_i \cdot X_j + 1)^d$
- RBF kernel with γ=16 (eukaryotic),
  $K(X_i, X_j) = \exp(- \gamma |X_i - X_j|^2)$



## SubLoc: Performance

| Location (Eukaryotic) | NNPSL | Markov model | SubLoc |
|---|---|---|---|
| | Accuracy (%) | Accuracy (%) | Accuracy (%) |
| Cytoplasmic | 55 | 78.1 | 76.9 |
| Extracellular | 75 | 62.2 | 80.0 |
| Mitochondrial | 61 | 69.2 | 56.7 |
| Nuclear | 72 | 74.1 | 87.4 |
| Total accuracy | 66 | 73.0 | 79.4 |

Dataset: Reinhardt & Hubbard, *NAR*, 1998



## SubLoc: Robustness of Amino Acid Composition Approach

| | Accuracy (%) | | | | |
|---|---|---|---|---|---|
| | Total | Cyto | Extra | Mito | Nuclear |
| COMPLETE | 78.3 | 76.7 | 77.2 | 56.4 | 86.0 |
| CUT-10 | 77.2 | 74.0 | 77.8 | 52.7 | 86.1 |
| CUT-20 | 76.3 | 73.2 | 78.5 | 51.4 | 84.8 |
| CUT-30 | 76.1 | 72.5 | 76.3 | 50.5 | 85.8 |
| CUT-40 | 75.3 | 71.5 | 74.2 | 46.7 | 86.3 |

- **Amazingly, accuracy of SubLoc is virtually unaffected when the first 10, 20, 30, & 40 amino acids in a protein are deleted**
- **Amino acid composition is a robust indicator of subcellular localization, and is insensitive to errors in N-terminal sequences**

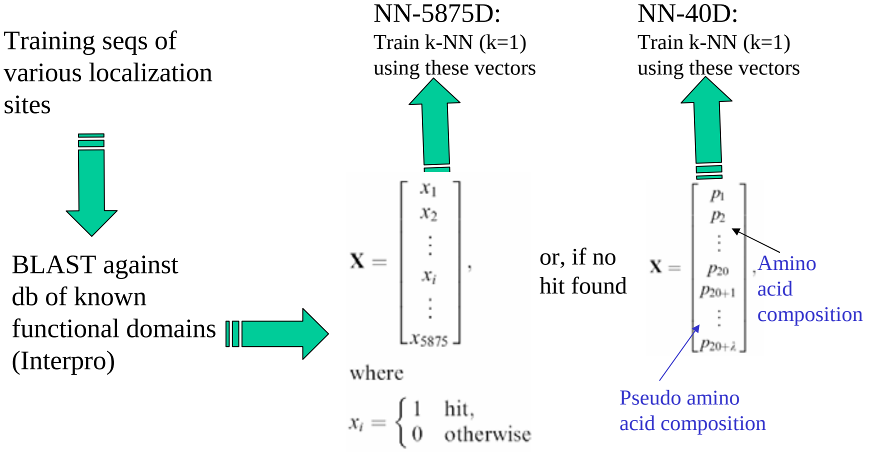
## Amino Acid Composition: Taking it Further

- **How about pairs of consecutive amino acids? (a.k.a 2-grams) How about 3-grams, …, k-grams?**
- **How about pseudo amino acid composition?**
- **How about presence of entire functional domains?** **(I.e. think of the presence/absence of a functional domain as a summary of amino acid sequence info...)**

## Functional Domain Composition

Cai & Chou, *BBRC*, 305:407--411, 2003

If a protein got a hit in Interpro, use NN-5875D; else use NN-40D



Training seqs of various localization sites

BLAST against db of known functional domains (Interpro)

NN-5875D: Train k-NN (k=1) using these vectors

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_{5875} \end{bmatrix},$$

where

$$x_i = \begin{cases} 1 & \text{hit,} \\ 0 & \text{otherwise} \end{cases}$$

or, if no hit found

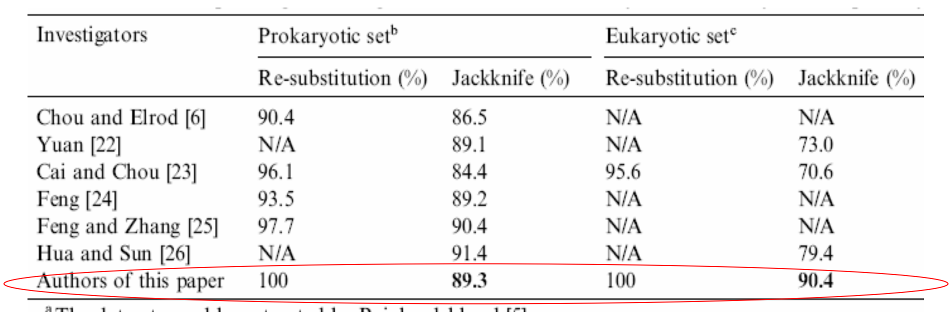
NN-40D: Train k-NN (k=1) using these vectors

$$\mathbf{X} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix},$$

Amino acid composition

Pseudo amino acid composition

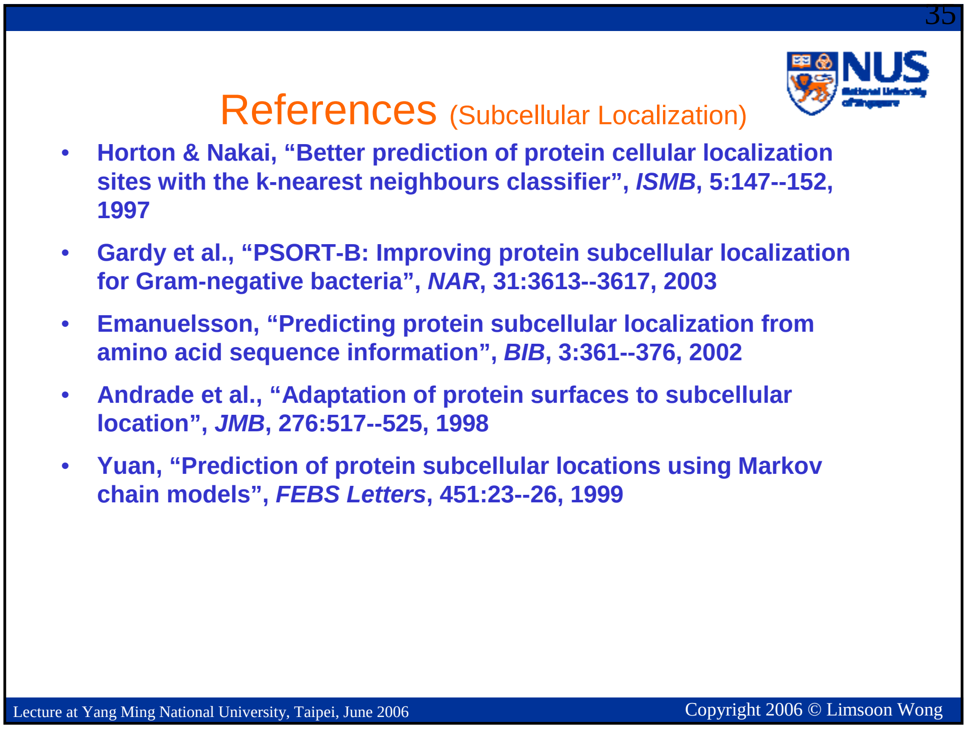
## Functional Domain Composition: Performance



| Investigators | Prokaryotic set[b] | | Eukaryotic set[c] | |
|---|---|---|---|---|
| | Re-substitution (%) | Jackknife (%) | Re-substitution (%) | Jackknife (%) |
| Chou and Elrod [6] | 90.4 | 86.5 | N/A | N/A |
| Yuan [22] | N/A | 89.1 | N/A | 73.0 |
| Cai and Chou [23] | 96.1 | 84.4 | 95.6 | 70.6 |
| Feng [24] | 93.5 | 89.2 | N/A | N/A |
| Feng and Zhang [25] | 97.7 | 90.4 | N/A | N/A |
| Hua and Sun [26] | N/A | 91.4 | N/A | 79.4 |
| Authors of this paper | 100 | **89.3** | 100 | **90.4** |

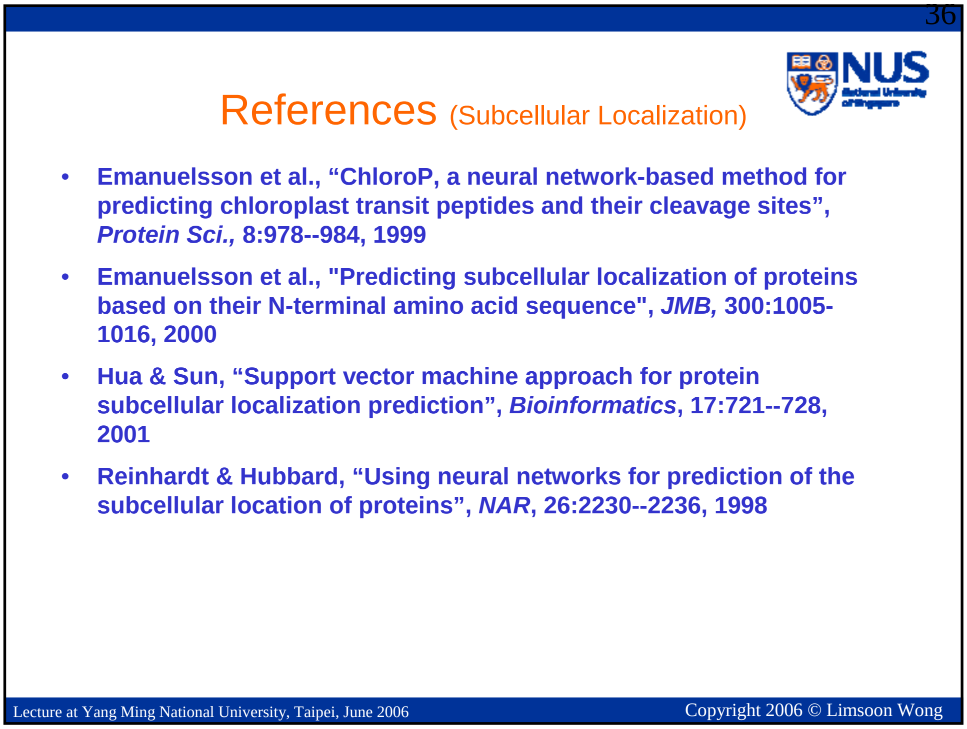
Dataset: Reinhardt & Hubbard, *NAR*, 1998

## Any Question?



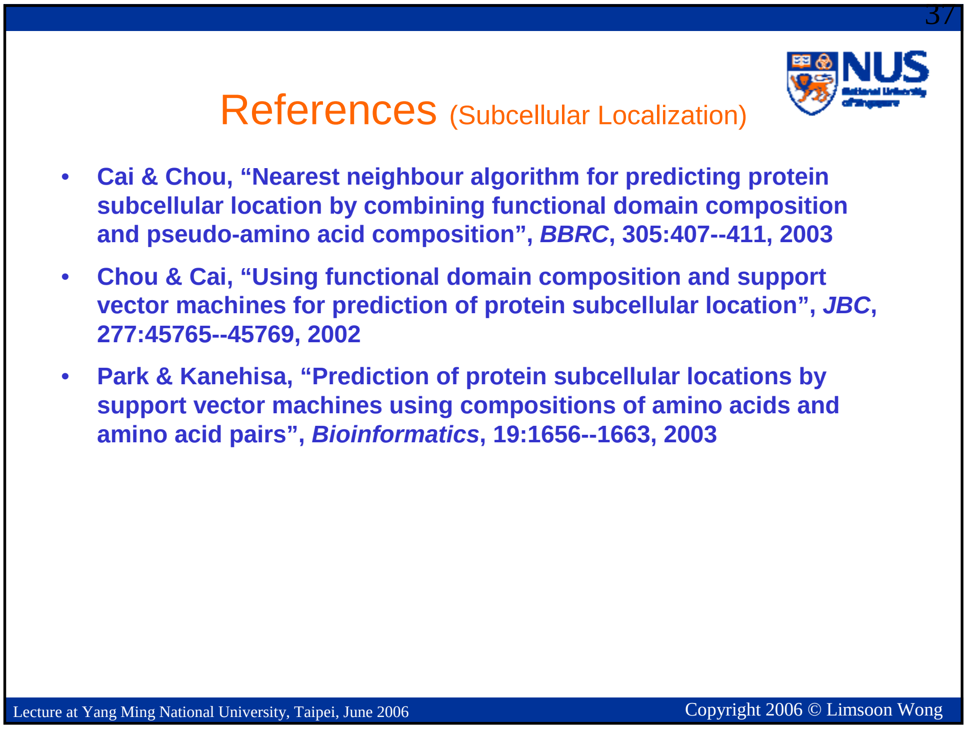
# References (Subcellular Localization)

- Horton & Nakai, "Better prediction of protein cellular localization sites with the k-nearest neighbours classifier", *ISMB*, 5:147--152, 1997
- Gardy et al., "PSORT-B: Improving protein subcellular localization for Gram-negative bacteria", *NAR*, 31:3613--3617, 2003
- Emanuelsson, "Predicting protein subcellular localization from amino acid sequence information", *BIB*, 3:361--376, 2002
- Andrade et al., "Adaptation of protein surfaces to subcellular location", *JMB*, 276:517--525, 1998
- Yuan, "Prediction of protein subcellular locations using Markov chain models", *FEBS Letters*, 451:23--26, 1999



# References (Subcellular Localization)

- Emanuelsson et al., "ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites", *Protein Sci.,* 8:978--984, 1999
- Emanuelsson et al., "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence", *JMB,* 300:1005-1016, 2000
- Hua & Sun, "Support vector machine approach for protein subcellular localization prediction", *Bioinformatics*, 17:721--728, 2001
- Reinhardt & Hubbard, "Using neural networks for prediction of the subcellular location of proteins", *NAR*, 26:2230--2236, 1998



## References (Subcellular Localization)

- **Cai & Chou, “Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition”, *BBRC*, 305:407--411, 2003**
- **Chou & Cai, “Using functional domain composition and support vector machines for prediction of protein subcellular location”, *JBC*, 277:45765--45769, 2002**
- **Park & Kanehisa, “Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs”, *Bioinformatics*, 19:1656--1663, 2003**