BIOINFORMATICS

Vol. 18 no. 2 2002
Pages 298–305

# Extensive Feature Detection of N-Terminal Protein Sorting Signals

*Hideo Bannai[1], Yoshinori Tamada[2], Osamu Maruyama[3], Kenta Nakai[1] and Satoru Miyano[1]*

*[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, JPN, [2]Department of Mathematical Sciences, Tokai University, 1117 Kitakaname, Hiratuka-shi, Kanagawa, 259-1292, JPN and [3]Faculty of Mathematics, Kyushu University, Kyushu University 36, Fukuoka, 812-8581, JPN*

**ABSTRACT**

**Motivation:** The prediction of localization sites of various proteins is an important and challenging problem in the field of molecular biology. TargetP, by Emanuelsson et al. (2000) is a neural network based system which is currently the best predictor in the literature for N-terminal sorting signals. One drawback of neural networks, however, is that it is generally difficult to understand and interpret how and why they make such predictions. In this paper, we aim to generate simple and interpretable rules as predictors, and still achieve a practical prediction accuracy. We adopt an approach which consists of an extensive search for simple rules and various attributes which is partially guided by human intuition.

**Results:** We have succeeded in finding rules whose prediction accuracies come close to that of TargetP, while still retaining a very simple and interpretable form. We also discuss and interpret the discovered rules.

**Availability:** An (experimental) web service using rules obtained by our method is provided at http://hypothesiscreator.net/iPSORT/.

**Contact:** bannai@ims.u-tokyo.ac.jp

## INTRODUCTION

Most proteins are first synthesized in the cytosol, and carried to specified locations, such as mitochondria or chloroplasts. In most cases, the information determining the subcellular localization site is represented as a short amino acid sequence segment called a protein sorting signal (Nakai, 2000). If we could somehow detect the amino acid sequence encoding this information, we would be able to predict the localization sites.

Prediction of localization sites is useful in various ways. Because cellular functions are often localized in specific compartments, the prediction of localization sites of unknown or unannoted proteins may be used to gain some indication of its function. For example, the information may be used to screen candidate genes for drug discovery. Further, if the rules for prediction were biologically interpretable, this knowledge could help in designing artificial proteins with desired properties.

TargetP (Emanuelsson et al., 2000), a neural network based predictor, is known to be the best predictor in the literature for N-terminal sorting signals. However, although neural networks are "readily available" and "often successful in practice", they are also infamous for the difficulty involved in trying to understand and interpret their meaning (Chou, 2001). PSORT (Nakai and Kanehisa, 1992; Nakai and Horton, 1999) and MitoProt (Claros and Vincens, 1996), unlike TargetP, are systems which incorporate existing knowledge about sorting signals, but they use various real numbers as "weights" in their prediction rules which also may not be trivially interpretable. Also, they are somewhat obsolete and their performance is unsatisfactory compared to TargetP.

The aim of this work is to derive simple and interpretable rules which can be used to predict subcellular localization sites, while still achieving a practical prediction accuracy. Through our *discovery oriented approach* to the problem, we managed to find very simple and interpretable rules with prediction accuracies which come fairly close to TargetP.

We will first review the existing knowledge about the N-terminal signals, and then describe the general idea of our approach.

### N-Terminal Sorting Signals

The signals we consider are signals known to be on the N-terminal of the protein. Mitochondrial targeting peptides (mTP), chloroplast transit peptides (cTP), and signal peptides (SP) are the typical N-terminal sorting signals.

Mitochondrial targeting peptides are known to be rich in arginine (R), alanine (A), and serine (S), while negatively charged amino acid residues (aspartic acid (D) and glutamic acid (E)) are rare (von Heijne et al., 1989). Only

weak consensus sequences have been found. Further, they are believed to form an amphiphilic $\alpha$-helix important for import into the mitochondrion.

Chloroplast transit peptides are known to be rare in acidic residues, and also believed to form an amphiphilic $\alpha$-helix (Bruce, 2000).

It has been established that a concrete consensus sequence does not occur in signal peptides. Rather, a three-region structure is conserved: a positively charged n-region, a hydrophobic h-region, and a polar c-region (von Heijne, 1990).

### Overview of Our Approach

Several very important aspects in the process of scientific knowledge discovery are: 1) the generation or discovery of *good* attributes, and ways of looking at the data, which is then used to explain the data, 2) the incorporation of and reflection on existing knowledge, and 3) the trial and error interaction between the expert and the problem. We have been developing a computer software library focusing on these points to speed-up this process, and have been applying it to various problems in the field of bioinformatics (Maruyama et al., 1998, 1999; Bannai et al., 2001).

The overall idea of this approach is to create massive amounts of very simple attributes and their trivial combinations, based on various known attributes. This way, if such rules exist for the data, we can expect to overcome the poor descriptive strength of simple rules, while at the same time control the complexity and structure of the rule to be generated.

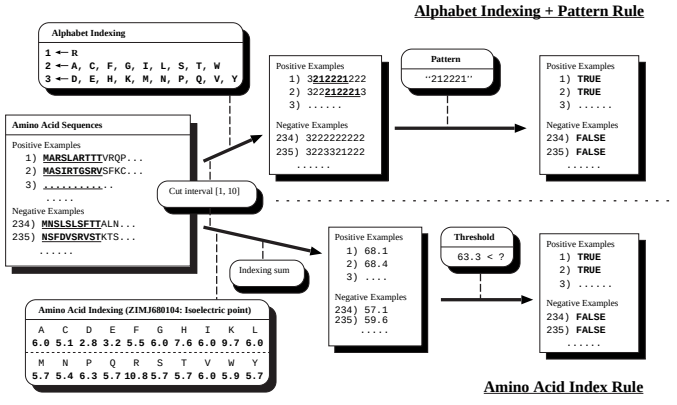


**Fig. 1.** Concept diagram of amino acid index rule and alphabet indexing + approximate pattern rule

Our search for the final hypothesis consists of two main aspects: amino acid index, and alphabet indexing + approximate pattern. The details of each aspect will be presented in the following sections, but we describe them briefly here (See Figure 1).

*Amino acid index* A large amount of experimental and theoretical research has been performed to characterize different kinds of properties of individual amino acids and to represent them in terms of a numerical index. The AAindex database (Kawashima and Kanehisa, 2000) is a compilation of 434 of such indices. As noted in the previous subsection, protein sorting signals have been characterized by the biochemical properties of the amino acids composing them, and it seems reasonable to assume that some kind of characteristic which is important for protein sorting is already contained in the AAindex database. Also, although an amino acid index generally assigns a real number for each amino acid, it should be easy for us to interpret the biological meaning of the rule when we find a "good" amino acid index contained in AAindex, which helps greatly in explaining the data. Therefore, using AAindex as a knowledge base, we generate *amino acid index rules.*

*Alphabet indexing + approximate pattern* Again from previous studies, we know that there is no clear-cut consensus sequence concerning each of the sorting signals. However, since there does seem to be a common structure for the same signals, we wish to somehow capture this knowledge. Our approach here is to consider motifs which allow more ambiguity by using *alphabet indexing* (Shimozono, 1999) and *approximate patterns* (Wu and Manber, 1992) over the indexed sequence, similar to the BONSAI system (Shimozono et al., 1994), which was successful in discovering meaningful knowledge from amino acid sequences.

An alphabet indexing is a classification of characters of an alphabet into a smaller set of characters, and can be viewed as a discrete, unordered version of an amino acid index. For example, we may divide the amino acids into the two classes of hydrophobic amino acids and hydrophilic amino acids. Using this alphabet indexing, we can view the amino acid sequence as a sequence of '0's (hydrophobic) and '1's (hydrophilic), and search for patterns (e.g. '001001100') contained in the sequences.

The outline of this paper is as follows: In the next section, we define the basic concepts used in our methods. We then show the results we have obtained from applying our methods to the data. Finally, we discuss how the rules we discovered may be interpreted.

## SYSTEM AND METHODS

### Amino Acid Index Rule

An *amino acid index* is a mapping from one amino acid to a numerical value, representing various physiochemical

and biochemical properties of amino acids.

**Definition 1 (Amino acid index)**
Let $\mathcal{A}$ denote the set of amino acids, and $\mathcal{R}$ the set of real numbers. For a given amino acid index $I : \mathcal{A} \rightarrow \mathcal{R}$ and amino acid sequence $s = s_1 s_2 \dots s_n$ (for $i : 1 \dots n, s_i \in \mathcal{A}$), let $I(s)$ denote the homomorphism $[|I(s_1); I(s_2); \dots; I(s_n)|]$, where $[|;|]$ denotes a sequence of values. □

The AAindex Database (Kawashima and Kanehisa, 2000) is a compilation of 434 types of amino acid indices, which have appeared in various reports.

**Definition 2 (Amino acid index rule)**
An *amino acid index rule* (R1-rule) is defined by: an amino acid index $I$, a specified region of the amino acid sequence $s$ denoted by $s[u, v]$, a function $f_w \in \{avg, maxavg_w, minavg_w\}$, and a threshold $\tau$, where: $avg(I(s))$ is defined as the average of the values of sequence $I(s)$, $maxavg_w(I(s))$ is the average of a substring of size $w$ in $I(s)$, which gives the maximum value (i.e. $\max\{I(s') \mid s = xs'y, |s'| = w\}$), and $minavg_w(I(s))$ is similarly the average of a substring of size $w$ in $I(s)$ which gives the minimum value (i.e. $\min\{I(s') \mid s = xs'y, |s'| = w\}$). $s[u, v] = s_u \dots s_v$ $(1 \leq u \leq v \leq n)$ for $s = s_1 s_2 \dots s_n$.

A sequence is predicted "positive" by the R1-rule if $f_w(I(s[m, n])) > \tau$ and "negative" otherwise. □

The parameters to be chosen are: $I, u, v, f, w, \tau$, and the task is to look for the best combination of the parameters which can distinguish between sequences of different signals.

With R1-rules, we expect to capture the overall properties of N-terminal sorting signals.

## Alphabet Indexing + Approximate Pattern Rule

An *alphabet indexing* (Shimozono, 1999) is a classification of characters of an alphabet. It is formally defined as follows:

**Definition 3 (Alphabet indexing)**
An *alphabet indexing* $\psi$ is a mapping from one alphabet $\Sigma$ to another alphabet $\Gamma$, where $|\Gamma| \leq |\Sigma|$. For $x = x_1 x_2 \dots x_l \in \Sigma^l$, let $\psi(x)$ denote the homomorphism $\psi(x_1)\psi(x_2)\cdots\psi(x_l) \in \Gamma^l$. We will call $\psi(x)$, the *indexed* sequence. □

**Definition 4 (Approximate pattern)**
An *approximate pattern* (Wu and Manber, 1992) is a string which can match another string, allowing up to $k$ errors (mismatch). The mismatch can consist of up to 3 types: insertion (ins), deletion (del), and substitution (sub). We will call the parameters $k$ and the types of mismatches allowed, the *mismatch allowance* of the approximate pattern. □

**Definition 5**
An *alphabet indexing + approximate pattern rule* (R2-rule) is defined by: an alphabet indexing $\psi$, a specified region of the amino acid sequence $s$ denoted by $s[u, v]$, a pattern $p$, and mismatch allowance $M$. A sequence is predicted "positive" if $p$ matches somewhere in $\psi(s[u, v])$ within mismatch allowance $M$, and "negative" otherwise. □

The parameters to be chosen are: $\psi, u, v, p, M$, and the task again is to find the best combination of the parameters which can distinguish between sequences of different signals.

With R2-rules, we expect to find locally specific characteristics concerning N-terminal sorting signals.

## Data

The data used in our computational experiments was obtained from the TargetP web-site[†]. These data consist of two data sets: plant and non-plant sequences. The plant data set of 940 sequences contained 368 mTP, 141 cTP, 269 SP, and 162 "Other" (consisting of 54 nuclear and 108 cytosolic) sequences. The non-plant data set of 2738 sequences contained 371 mTP, 715 SP and 1652 "Other" (consisting of 1214 nuclear and 438 cytosolic) sequences.

We basically follow the work on TargetP, considering different predictors for plant and non-plant proteins. Also, as in the composition of TargetP, we will first consider binary predictors which just predict whether or not a given sequence contains a specific signal. The knowledge obtained from these binary rules is combined into a *decision list*, to form a final rule. For each binary predictor, we will call the sequences concerning the signal in question *positive examples*, and the sequences concerning the other signals, *negative examples*.

## Search Strategies

We extensively search for various parameters described in the previous section. Since the size of the search space for the combinations of different parameters is huge, an exhaustive search is not feasible even with the powerful computers which were available. We adopted a mixture of heuristics and exhaustive search. Many different combinations of the parameters as well as minute variations in the heuristics were tried.

*Combining the Rules* To create a single rule predicting the sorting signal for a given sequence, we combine the binary rules generated for each sorting signal into a decision list. The structure of the decision list is

[†] http://www.cbs.dtu.dk/services/TargetP/

shown in Figure 2. The structure was determined greedily, according to the "ease" of discrimination by the R1-rules, which was stable for all training/test combinations.

From preliminary experiments, R1-rules seemed to be sufficient for discriminating signal peptide sequences. As for the other signals, neither type of rule seemed good enough for identifying the signals. Therefore, we consider combining both types of rules (R1-rules and R2-rules) into a single rule. The first node of the decision list discriminates signal peptides with a single R1-rule, while the second and third nodes consist of both R1-rule and R2-rule. The two rules (or perhaps their negations) are combined with a logical 'and' where a sequence is judged to have a certain signal if both rules say so.

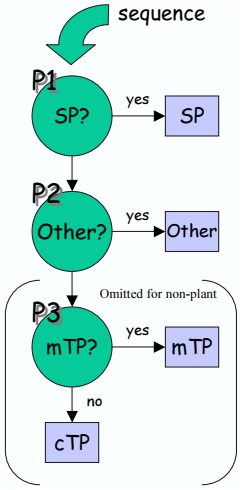


**Fig. 2.** The Structure of the final rule for plant and non-plant data sets. The last node in parentheses concerning cTP and mTP is omitted for the non-plant data set (classifying to mTP). The various parameters such as the amino acid index, substring intervals, alphabet indexing, and patterns which were chosen in the 5 training runs are summarized in Table 1 and 2.

## Evaluation of Prediction Accuracy

The whole search space conducted in our search was enormous, and involved considerable amounts of human intervention, influenced largely by human intuition. To give a *fair* estimate for the prediction accuracy of our methods, we choose a modest range of parameters to search for in the cross validation, and show that the knowledge discovered is fairly stable even in that quite large range.

*Training and Evaluation* We follow the training and evaluation methods used for TargetP. The data were randomly divided into 5 equal sized datasets by dividing each subset of sequences with specific localization sites into 5 datasets. Rules were generated by using 4 of the data sets as training data, and testing was conducted on the remaining data set. This was repeated for the 5 possible pairs of training and test set, and the overall performance is the sum of the 5 results. (All rules are generated by using the training data set only.)

Preliminary experiments showed that R1-rules were fairly stable, but R2-rules seemed to somewhat over-fit the training data. To overcome this problem, in the training phase for R2-rules, the training set was again randomly divided into 5 sets (4 *ttrain* and 1 *ttest* sets), and rules are generated from ttrain and tested with ttest. The arithmetic product of the scores from the ttrain and ttest sets was used to select which alphabet indexing and substring interval to use. The pattern was then trained using all sequences of the original training set, using the alphabet indexing and substring interval.

Rules are evaluated by the Matthews correlation coefficient (MCC) (Matthews, 1975), defined by:

$$\frac{tp \times tn - fp \times fn}{\sqrt{(tp+fn)(tp+fp)(tn+fp)(tn+fn)}}$$

where $tp$ = true positives, $fp$ = false positives, $tn$ = true negatives, and $fn$ = false negatives. Sensitivity, the fraction of correctly predicted positive examples ($\frac{tp}{tp+fn}$), and specificity, the fraction of true positives in the examples predicted as positive ($\frac{tp}{tp+fp}$), were also calculated for reference.

Details of the search is given below:

*Amino acid index rule* Since we know that the signals are located somewhere in the N-terminal region, we look (somewhat) exhaustively at the substring intervals in this region: For amino acid index $I$, all 434 entries in the AAindex Database together with 20 more entries, assigning a value of '1' to one amino acid and '0' to the rest were considered. For these $454$ entries, 72 substring intervals $[u, v] = [5n + 1, 5k]$ (where $n = 0 \ldots 8$ and $k = 1 \ldots 8$) were considered. For these $454 * 72 = 32688$ combinations, $f_w = avg,\ maxavg_w,\ minavg_w$ were considered where $w$ was taken to be 6 to 12, resulting in $32688 * (2 * 7 + 1) = 490320$ combinations. For all these combinations, all possible thresholds are considered for $\tau$: let $f_{w_1}, ..., f_{w_n}$ be all $f_w$ values of $n$ sequences in sorted order. Then, $\tau = (f_{w_i} + f_{w_{i+1}})/2, i = 1, \ldots, n-1$. The combination of $I, u, v, f, w, \tau$ which gives the highest MCC score is recorded.

*Alphabet indexing + approximate pattern rule* The substring intervals $[u, v]$ is taken to be $[5n + 1, 5n + 5k]$ (where $n = 0, 1, 2$ and $k = 2, 3, 4$). For a given alphabet indexing $\psi$, all patterns of length $8$ appearing in the sequences are considered for $p$. The mismatch types were limited to insertion and deletion only (no substitution). The maximum mismatch number was fixed at 2. We started with the alphabet indexing classifying the amino acids into three classes, according to their "charges":

$$\psi_0(x) = \begin{cases} 0 & if \quad x \in \{D, E\}, \\ 1 & if \quad x \in \{K, R\}, \\ 2 & if \quad x \in \mathcal{A} - \{D, E, K, R\} \end{cases}$$

and optimized the indexing by conducting a local search on the alphabet indexing (Shimozono et al., 1994): i.e. consider the alphabet indexings which are obtained by changing the indexing for a single amino acid (40 candidates in this case), and adopt the indexing whose product of the MCC scores for ttrain and ttest is the highest for the best pattern. The process is repeated until a local maximum is reached. A local search strategy was used because an exhaustive search for all possible alphabet indexings would result in $3^{20}$ combinations, which was not feasible. Numerous tries starting from other alphabet indexings, which were chosen randomly, were also conducted, but high scoring indexings seemed to be centered around $\psi_0$.

*Combination of the rules* After determining the parameters for the above rules, the rules and possibly their negations are combined with a logical 'and', but a portion of the parameters are trained again. Namely, the substring intervals, $f$, window sizes, amino acid index, and alphabet indexing are fixed. We retrain the mismatch allowance, pattern, and threshold. Their ranges are expanded in the retraining: The maximum mismatch number of 1 to 3 was allowed, and all patterns of length 5 to 10 appearing in the data were considered. The top 100 R1-rules are combined with all possible R2-rules, and the combination which gives the best MCC score is chosen to be used against the test set.

## IMPLEMENTATION

The software used in our analysis was developed using the Hypothesis Creator Library[‡] (Bannai et al., 2001). Various shared memory multi-processor computers were available for calculation: 2 SGI Origin 2000 with (128, 32) x 195MHZ R10000 processors, 1 Sun Ultra Enterprise 4500 and 2 Sun Ultra Enterprise 3500 with (14, 8, 8) x 400MHz Sun Ultra II processors respectively. Each is equipped with well over 2GB of memory, which was the limit of the software.

## RESULTS AND DISCUSSION

The parameters found for each training set is summarized in Table 1 for the plant data set, and Table 2 for the non-plant data set. The scores of the cross validation is summarized in Table 3, together with the scores for TargetP written in parentheses (The scores for TargetP was taken directly from (Emanuelsson et al., 2000)).

We can see that the MCC scores for our predictor is fairly close to those of TargetP, except for chloroplast transit peptides (cTP). However, it should be noted that our score for cTP (0.64) would rank second, after TargetP (0.72), better than PSORT (0.51), MitoProt (0.44),

[‡] http://hypothesiscreator.net/

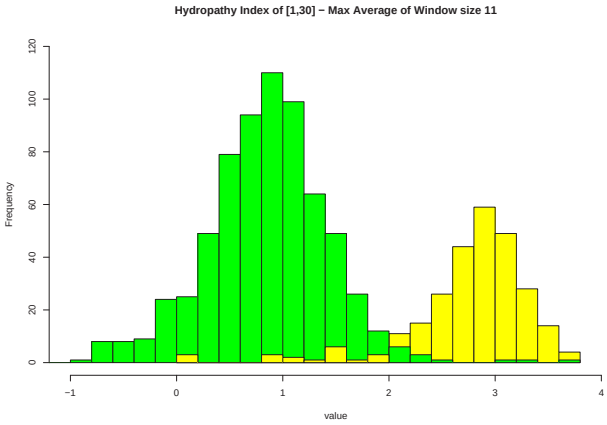


**Fig. 3.** Histograms (light-SP, dark-mTP, cTP, Other) of $maxavg_{11}$ values of hydropathy index (Kyte and Doolittle, 1982) for the substring $[1, 30]$ of the plant data set. The threshold was 2.07727. We can see that there is a clear difference in the distribution.

and ChloroP (0.50) (Emanuelsson et al., 1999), in the comparison of (Emanuelsson et al., 2000). Our predictor scores higher for plant signal peptides, and for the other signals, our scores would again rank second, after TargetP with respect to the other predictors.

### Biological Evaluation of the Rules

#### Amino acid index rules

*SP vs (mTP + cTP + Other):* [Node *P1*, rule *R1* in Table 1] The amino acid index with the highest score was the hydropathy index (Kyte and Doolittle, 1982), and judging from the substring interval $[1, 30]$, and function $maxavg_w$ where $w$ is around 11, we can say this rule corresponds to characteristics known for signal peptides (the hydrophobic h region) (von Heijne, 1990) (Figure 3). What is surprising is that such a simple rule could discriminate signal peptides so well - better than TargetP for plant proteins.

*(mTP + cTP) vs Other:* [Node *P2*, rule *R1* in Table 1] The amino acid index was "negative charge", which assigns a value of 1 to aspartic acid (D) and glutamic acid (E). This also corresponds to known characteristics: mTP and cTP are rare in negatively charged amino acids (von Heijne et al., 1989).

*mTP vs cTP:* [Node *P3*, rule *R1* in Table 1] Various amino acid indices were chosen, with substring regions for a very short region at the N-terminal. However, the amino acid index: Isoelectric point (Zimmerman et al., 1968) can be considered as a more accurate measure of the net amino acid charges. Atom based hydrophobic moment

**Table 1.** The parameters chosen for each training set (threshold $\tau$ is omitted) for the plant data set. The nodes P$n$ corresponds to nodes in Figure 2.

| Node | Trial | R1 Amino Acid Index $[u,v]$ | R1 $f_w, dir$ | R2 Alphabet Indexing $[u,v]$ | R2 Pattern | R2 Mismatch | Combination |
|---|---|---|---|---|---|---|---|
| P1 | 1 | Hydropathy index[1] $[1,30]$ | $maxavg_{11},\uparrow$ | not used | not used | not used | R1→SP |
| | 2 | Hydropathy index $[6,30]$ | $maxavg_{11},\uparrow$ | not used | not used | not used | R1→SP |
| | 3 | Hydropathy index $[1,30]$ | $maxavg_{11},\uparrow$ | not used | not used | not used | R1→SP |
| | 4 | Hydropathy index $[6,30]$ | $maxavg_{11},\uparrow$ | not used | not used | not used | R1→SP |
| | 5 | Hydropathy index $[6,30]$ | $maxavg_{11},\uparrow$ | not used | not used | not used | R1→SP |
| P2 | 1 | Negative Charge $[1,25]$ | $avg,\downarrow$ | DE → 0, AR → 1, Other → 2 $[1,20]$ | 221200020 | 3 ins/del | ¬R1 ∨ R2→Other[†] |
| | 2 | Negative Charge $[1,25]$ | $avg,\downarrow$ | DE → 0, CR → 1, Other → 2 $[1,20]$ | 20002212 | 2 ins/del | ¬R1 ∨ R2→Other |
| | 3 | Negative Charge $[1,25]$ | $avg,\downarrow$ | DE → 0, R → 1, Other → 2 $[1,15]$ | 022120 | 1 ins/del | ¬R1 ∨ R2→Other |
| | 4 | Negative Charge $[1,30]$ | $maxavg_8,\downarrow$ | DE → 0, CR → 1, Other → 2 $[1,20]$ | 2002222222 | 1 ins/del | ¬R1 ∨ R2→Other |
| | 5 | Negative Charge $[1,30]$ | $avg,\downarrow$ | DE → 0, CRF→ 1, Other → 2 $[1,20]$ | 020222222 | 1 ins/del | ¬R1 ∨ R2→Other |
| P3 | 1 | Hyd. mom.[2] $[1,10]$ | $maxavg_6,\uparrow$ | E → 0, KR → 1, Other → 2 $[1,10]$ | 22112221 | 2 ins/del | R1 ∧ R2→mTP |
| | 2 | Isoelectric point[3] $[1,10]$ | $avg,\uparrow$ | E → 0, KRW → 1, Other → 2 $[1,10]$ | 22110 | 2 ins/del | R1 ∧ R2→mTP |
| | 3 | Hyd. mom. $[1,10]$ | $maxavg_9,\uparrow$ | E → 0, ARW → 1, Other → 2 $[1,10]$ | 22110 | 2 ins/del | R1 ∧ R2→mTP |
| | 4 | Net charge[4] $[1,10]$ | $avg,\uparrow$ | E → 0, KR → 1, Other → 2 $[1,10]$ | 1212221 | 2 ins/del | R1 ∧ R2→mTP |
| | 5 | Isoelectric point $[1,15]$ | $maxavg_{12},\uparrow$ | E → 0, DKRW → 1, Other → 2 $[1,10]$ | 11221 | 2 ins/del | R1 ∧ R2→mTP |

[†]: The actual rule was R1 ∧ ¬ R2 → mTP $or$ cTP

**Table 2.** The parameters chosen for each training set (threshold $\tau$ is omitted) for the non-plant data set. The nodes P$n$ corresponds to nodes in Figure 2.

| Node | Trial | R1 Amino Acid Index $[u,v]$ | R1 $f_w, dir$ | R2 Alphabet Indexing $[u,v]$ | R2 Pattern | R2 Mismatch | Combination |
|---|---|---|---|---|---|---|---|
| P1 | 1 | Hydropathy index $[1,30]$ | $maxavg_{12},\uparrow$ | not used | not used | not used | R1→SP |
| | 2 | Hydropathy index $[1,30]$ | $maxavg_{12},\uparrow$ | not used | not used | not used | R1→SP |
| | 3 | Hydropathy index $[1,30]$ | $maxavg_{12},\uparrow$ | not used | not used | not used | R1→SP |
| | 4 | Hydropathy index $[1,30]$ | $maxavg_{12},\uparrow$ | not used | not used | not used | R1→SP |
| | 5 | Hydropathy index $[1,30]$ | $maxavg_{12},\uparrow$ | not used | not used | not used | R1→SP |
| P2 | 1 | Net Charge $[1,25]$ | $minavg_{12},\uparrow$ | DE → 0, R → 1, Other → 2 $[1,25]$ | 202020220 | 3 ins/del | R1 ∧¬ R2→mTP |
| | 2 | Negative Charge $[1,20]$ | $avg,\downarrow$ | DE → 0, R → 1, Other → 2 $[1,25]$ | 2211221222 | 2 ins/del | R1 ∧ R2→mTP |
| | 3 | Negative Charge $[1,20]$ | $maxavg_{12},\downarrow$ | DE → 0, R → 1, Other → 2 $[1,30]$ | 2211221222 | 2 ins/del | R1 ∧ R2→mTP |
| | 4 | Negative Charge $[1,20]$ | $maxavg_{12},\downarrow$ | DE → 0, R → 1, Other → 2 $[1,25]$ | 2212211222 | 2 ins/del | R1 ∧ R2→mTP |
| | 5 | Negative Charge $[1,20]$ | $maxavg_{12},\downarrow$ | DEY → 0,R → 1, Other → 2 $[1,25]$ | 22122212 | 1 ins/del | R1 ∧ R2→mTP |

[1]: Hydropathy index (Kyte and Doolittle, 1982), [2]: Atom based hydrophobic moment (Eisenberg and Mclachalan, 1986),

[3,4]: Net charge, Isoelectric point (Zimmerman et al., 1968).

$f_w \uparrow$ means that rule will answer *yes* if the value of $f_w(I(s[u,v]))$ is above a certain value $\tau$, $f_w \downarrow$ is the opposite.

(Eisenberg and Mclachalan, 1986) is also a similar amino acid index, where the values for arginine (R) and lysine (K) are higher than the other amino acids. Although values for aspartic acid (D) and glutamic acid (E) are also higher for the atom based hydrophobic moment, these amino acids rarely appear in mTP or cTP, and do not effect the average values.

Therefore, together with the interpretation from (mTP+cTP) vs Other, we can see that both mTP and cTP lack negatively charged amino acids, but mTP tend to be more positively charged than cTP for the front end of the signal.

Also seen in the alphabet indexing + approximate pattern rule for mTP vs cTP, the region which was best for distinguishing the two signals seemed to be located in the short portion of the sequences, whereas the best regions for distinguishing the other signals tended to be a longer.

The plain occurrence count of amino acids did not seem to appear in any of the trials. This is perhaps because the number of certain amino acids is too rough an estimate of the overall biochemical properties of the signals.

**Table 3.** The Prediction Accuracy of the Decision Lists (scores of TargetP (Emanuelsson et al., 2000) in parentheses). This represents the sum of the predictions of the 5 hypotheses of Tables 1 and 2 over the test set.

| Data Set | True category | # of seqs | Predicted category | | | | Sensitivity | MCC |
|---|---|---|---|---|---|---|---|---|
| | | | cTP | mTP | SP | Other | | |
| Plant | cTP | 141 | 96 (120) | 26 (14) | 0 (2) | 19 (5) | 0.68 (0.85) | 0.64 (0.72) |
| | mTP | 368 | 25 (41) | 309 (300) | 4 (9) | 30 (18) | 0.84 (0.82) | 0.75 (0.77) |
| | SP | 269 | 6 (2) | 9 (7) | 244 (245) | 10 (15) | 0.91 (0.91) | 0.92 (0.90) |
| | Other | 162 | 8 (10) | 17 (13) | 2 (2) | 135 (137) | 0.83 (0.85) | 0.71 (0.77) |
| | Specificity | | 0.71 (0.69) | 0.86 (0.90) | 0.98 (0.96) | 0.70 (0.78) | | |
| Non-plant | mTP | 371 | – – | 275 (330) | 11 (9) | 85 (32) | 0.74 (0.82) | 0.67 (0.73) |
| | SP | 715 | – – | 8 (13) | 660 (683) | 47 (19) | 0.92 (0.91) | 0.90 (0.92) |
| | Other | 1652 | – – | 119 (152) | 44 (49) | 1489 (1451) | 0.90 (0.85) | 0.78 (0.82) |
| | Specificity | | – – | 0.68 (0.67) | 0.92 (0.92) | 0.92 (0.97) | | |

### Alphabet indexing + approximate pattern rules

*(mTP + cTP) vs Other:* [Node *P2*, rule *R2* in Table 1] The alphabet indexing was stable near $\psi_0$. The best patterns were found to match the 'Other' sequences, rather than patterns matching mTP and cTP signals. Although patterns of the latter type would be of more interest, this is natural since mTP and cTP are different signals and the similarity in their structure may be subtle. Looking at the combination of the rules, a signal is rejected for mTP or cTP if the sequence contains (nearly) consecutive '0's, which is aspartic acid (D) or glutamic acid (E). The occurrence of '1' in each pattern is limited, showing that mTP or cTP signals should contain a number of arginine (R). Lysine (K) is classified to '2' perhaps showing the asymmetry of arginine and lysine (K) in mTP.

*mTP vs cTP:* [Node *P2*, rule *R2* in Table 1] The best patterns were found to match mTP sequences. Some patterns may be too short to judge, but the patterns "22112221' and '1212221' seem to be capturing the periodic occurrence of arginine (R) or lysine (K) ('1') in mTP, which is the characteristic of an amphiphilic $\alpha$-helix (von Heijne et al., 1989).

With the same parameters, we also searched for the best patterns which match cTP and do not match mTP. The patterns found were '022210' for trials 1, 3, and 4, '2222022222' for trial 2, and '220222110' for trial 5, all with a maximum of 1 insertion/deletion. It is interesting that all the patterns contain a '0', which is glutamic acid (E).

### Non-plant

A similar interpretation can be done for rules concerning the non-plant data set. The difference from the plant set being that the alphabet indexing was more stable around $\psi_0$. Also looking at the patterns discovered, the first pattern "202020220" does not match mTP sequences, meaning that mTP sequences are again rare in aspartic acid or glutamic acid. For the other patterns "2211221222", "2212211222", "22122212", we can see again the periodic occurrence of arginine (R) of an amphiphilic $\alpha$-helix. The pattern seems to be more stable in the non-plant data set perhaps the data set is much larger than for the plant set.

Overall, the rules discovered can be interpreted in terms of biological knowledge known for the different signals. The parameters chosen for the rules in each of the training rounds seemed to be fairly stable, suggesting that the rules are capturing relevant characteristics concerning the N-terminal signals.

### Future Work

For the plant data set, looking at the number of classified sequences, the weakness of our predictor seems to lie mainly in the discrimination of mTP and cTP. It would be interesting to find another simple but different form of rule to discriminate the two types of signals.

In the search we conducted, we defined the regions as substring intervals, fixed for all the sequences. Although the N-terminal signals are generally located in a somewhat fixed area, this may not be true for nuclear sorting signals, whose position in the sequence looks arbitrary.

The substring interval may be "simple" for human to understand, but may not be simple for the molecules detecting the signal. It would be desirable to find a way to target the actual location of the signal, and then consider the rules mentioned in this paper. If we are successful, there might also be ways to predict cleavage sites by locating candidate areas, and finding some meaningful amino acid index or alphabet index rule.

## Conclusion

We extensively searched various attributes and their simple combinations and were successful in finding a simple and interpretable rule which could explain the data set well. Despite their simplicities, the prediction accuracy of the rules were still competitive with the prediction scores of TargetP, the best predictor in the literature.

An experimental WWW service for predicting N-terminal sorting signals using a decision list trained on the entire data set is provided at: http://hypothesiscreator.net/iPSORT/. The range of parameters searched to make the rules for the web service is different from that in this paper in that the alphabet indexing was searched in a wider range. Also, only $avg$ was considered for $f_w$. Other parameters were adjusted to give best cross validation scores.

## Acknowledgements

This research was supported in part by Grant-in-Aid for Encouragement of Young Scientists and Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Sports, Science and Technology of Japan.

## REFERENCES

Bannai, H., Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano (2001). Views: Fundamental building blocks in the process of knowledge discovery. In *Proceedings of the 14th International FLAIRS Conference*, pp. 233–238. AAAI Press.

Bruce, B. D. (2000). Chloroplast transit peptides: structure, function and evolution. *Trends Cell Biol. 10*, 440–447.

Chou, K.-C. (2001). Using subsite coupling to predict signal peptides. *Protein Engineering 14*(2), 75–79.

Claros, M. G. and P. Vincens (1996, November). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem. 241*(3), 779–786.

Eisenberg, D. and A. Mclachalan (1986). Solvation energy in protein folding and binding. *Nature 319*(6050), 199–203.

Emanuelsson, O., H. Nielsen, S. Brunak, and G. von Heijne (2000, July). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol. 300*(4), 1005–1016.

Emanuelsson, O., H. Nielsen, and G. von Heijne (1999). Chlorop, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Prot. Sci. 8*, 978–984.

Kawashima, S. and M. Kanehisa (2000). AAindex: Amino Acid index database. *Nucleic Acids Res. 28*(1), 374.

Kyte, J. and R. Doolittle (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105–132.

Maruyama, O., T. Uchida, T. Shoudai, and S. Miyano (1998). Toward genomic hypothesis creator: View designer for discovery. In *Discovery Science*, Volume 1532 of *Lecture Notes in Artificial Intelligence*, pp. 105–116.

Maruyama, O., T. Uchida, K. L. Sim, and S. Miyano (1999). Designing views in HypothesisCreator: System for assisting in discovery. In *Discovery Science*, Volume 1721 of *Lecture Notes in Artificial Intelligence*, pp. 115–127.

Matthews, B. W. (1975). Comparison of predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta 405*, 442–451.

Nakai, K. (2000). Protein sorting signals and prediction of subcellular localization. In P. Bork (Ed.), *Analysis of Amino Acid Sequences*, Volume 54 of *Advances in Protein Chemistry*, pp. 277–344. San Diego: Academic Press.

Nakai, K. and P. Horton (1999). PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci. 24*, 34–35.

Nakai, K. and M. Kanehisa (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics 14*, 897–911.

Shimozono, S. (1999). Alphabet indexing for approximating features of symbols. *Theor. Comput. Sci. 210*, 245–260.

Shimozono, S., A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara, and S. Arikawa (1994). Knowledge acquisition from amino acid sequences by machine learning system BONSAI. *Trans. Information Processing Society of Japan 35*(10), 2009–2018.

von Heijne, G. (1990). The signal peptide. *J. Membr. Biol. 115*, 195–201.

von Heijne, G., J. Steppuhn, and R. G. Herrmann (1989). Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem. 180*, 535–545.

Wu, S. and U. Manber (1992). Fast text searching allowing errors. *Commun. ACM 35*, 83–91.

Zimmerman, J., N. Eliezer, and R. Simha (1968). The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol. 21*, 170–201.