Supplementary Text

The F-score measure of single feature discriminative power

The F-score is a simple measure of how well a feature separates two classes, defined as:

$$\frac{(\bar{x}^{+}-\bar{x})^{2}+(\bar{x}^{-}-\bar{x})^{2}}{\frac{1}{n_{+}-1}\Sigma_{k=1}^{n_{+}}(x_{k}^{+}-\bar{x}^{+})^{2}+\frac{1}{n_{-}-1}\Sigma_{k=1}^{n_{-}}(x_{k}^{-}-\bar{x}^{-})^{2}}$$

where $\bar{x}^{+}$, $\bar{x}^{-}$, and $\bar{x}$ are the mean values of the feature for the positive, negative and combined examples respectively; and x_k^{+} and x_k^{-} denote the value of the *k*th positive and negative examples respectively. This score roughly reflects the discriminative power of a feature in isolation with a large score indicating a high power.

Supplementary Figure Legends

Figure S1. (A) The length distribution of presequences in yeast, plant (*A.thaliana* and *O.sativa*) and metazoa is shown (above) with p-values from a Kolmogorov–Smirnov test (below) providing a rough estimate of the goodness of fit of parametric modeling using mixture models with 1 to 3 components. (B) Cleavage site prediction accuracy on the *A.thaliana* and *O.sativa* dataset in a 10-fold cross validation test. Error bars indicate the standard mean estimation error over 10 randomized partitionings when performing cross validation. (C) Cleavage site prediction accuracy on human data from the DegraBase dataset.

Figure S2. A PR-curve based on 5-fold cross-validation on the training data is shown for presequence prediction. Detailed performance numbers are listed for two cut-off values.

Figure S3. PR curve and ROC AUC attained by MitoFates when trained on the TargetP or Predotar data set.

Figure S4. Performance comparison between MitoFates and previous predictors using human matrix proteome data as positive test data (226 matrix proteins and 6596 non-presequence proteins obtained after removing sequences sharing more than 25% sequence identity among training and test data of MitoFates and TPpred2).

Figure S5. The 14 statistically significantly hexamer presequence motifs are shown with their p-values, coverage of the positive and negative training examples, PA score and a sequence logo of their matches in the positive data.

Figure S6. The match position distribution and sequence logo of two presequence hexamers (A) ϕϕσβϕϕ, (B)

ϕϕβσϕϕ which often occur at or one residue away from the N-terminus.

Figure S7. The proportion of human genes for which MitoFates prediction is consistent or divergent among the set of isoforms of the given gene is shown.

Supplementary Table Legends

Table S1. Frequency of observing each of 14 hexamer motifs over 100 random trials using scrambled positive example sequences as the negative data.

Table S2. List of presequence containing proteins matching motifs (A) ϕϕσβϕϕ or (B) ϕϕβσϕϕ at or one position away from the N-terminus.

Table S3. List of yeast presequences in each of the three clusters discussed in the main text.

Table S4. List of proteins with MitoFates predicted presequences among a list of human proteins accessible from the mitochondrial intermembrane space (IMS).

Table S5. 1847 proteins predicted by MitoFates to have mitochondrial targeting presequences from 42,217 human protein sequences.

Table S6. List of human protein sequences without mitochondrial localization annotations predicted by MitoFates to have mitochondrial targeting presequences.

Table S7. List of genes predicted by MitoFates to have isoforms both with and without mitochondrial targeting presequences.

Table S8. List of human genes with MitoFates predicted mitochondrial targeting presequences among candidate regulators of parkin translocation.

Table S9. Proteins with MitoFates predicted presequences among a set of human proteins with known disease mutations: with (A) and without (B) mitochondrial annotation.