# 論 文 の 内 容 の 要 旨

論文題目　Research on the amino sequence characteristics determining the transport, membrane topology and peptidase processing of mitochondrial proteins
(ミトコンドリア蛋白質の輸送、膜トポロジー及びペプチダーゼによる切断を決定するアミノ酸配列の特徴に関する研究)

氏　名　深沢　嘉紀

It is well known that mitochondria function as the essential power plants of most eukaryotic cells. They also make important contributions to many other vital cell functions such as lipid metabolism and calcium homeostasis. Moreover, mitochondrial dysfunction has been implicated in numerous medical conditions such as Parkinson's and Alzheimer's disease.

Although many mitochondrial proteins have been experimentally identified, a complete list is not available even for intensely studied model organisms. Thus bioinformatics tools to predict mitochondrial proteins from their amino acid sequences are widely used to complement experimental data; but their accuracy is far from perfect and they have not improved significantly for roughly a decade.

Existing prediction tools already employ sophisticated machine learning techniques so the key to progress seems to lie in the utilization of new proteomics data and the identification or refinement of sequence features that reflect the underlying molecular biology. Moreover, the development of such sequence features may be useful not only for more accurate prediction but also provide useful biological hints. Here I report features of local sequences in mitochondrial proteins which regulate their transport, membrane insertion, and peptidase processing.

In chapter 1, I summarize the necessary background to understand my thesis work. This chapter reviews the known biology of mitochondrial proteins in terms of their import, cleavage and membrane spanning domains.

In chapter 2, I report on the sequence divergence of N-terminal sorting signals, and show that divergence is a promising novel feature for signal prediction. For yeast, mammal and plant datasets, evolutionary sequence divergence alone has significant power to identify sequences with N-terminal sorting sequences. First I utilized YGOB, a curated database for orthologs between budding yeast and its related species, for calculation of sequence divergence of yeast proteins. I then demonstrate that sequence divergence is nearly as effective when computed on automatically defined orthologs sets for yeast, mammal, and plant datasets as on the hand curated ones. Unfortunately, sequence divergence did not necessarily increase classification performance when combined with some traditional sequence features such as amino acid composition. However, a post-hoc analysis of the proteins in which sequence divergence changes the prediction yielded some proteins with atypical (i.e. not MPP-cleaved) matrix targeting signals as well as a few misannotations.

In chapter 3, I introduce MitoFates, a prediction system for mitochondrial presequences

(N-terminal regions cleaved upon translocation into the mitochondria) designed with the knowledge of mitochondrial intermediate peptidases in mind and trained on recent proteomics data. MitoFates achieves better performance in both signal and cleavage site prediction. To obtain this performance, I revisited classical features for predicting this signal and searched for novel specific sequence motifs in the mitochondrial N-terminal presequence. Among the classical features, I revisited a detector of local sequences with the potential to form an amphiphilic α-helix, with a hydrophobic and hydrophilic face, inspired by the structure of the presequence recognizer Tom20 and Tom22. In previous applications, this feature has not been very effective, but I noted that the formulation used did not distinguish between negative and positive charge. By introducing a new term rewarding helices with positive charges opposite the hydrophobic face, I greatly increased the discriminant power of this feature. Employing recently developed techniques for sensitive multiple hypothesis testing, I discovered several novel and significant motifs from presequence, most of which show a positively charged amphiphilicity (possibly indicating recognition by TOM complex) or matching the consensus sequence of presequence cleavage sites. I also refined cleavage site of presequence by utilizing recent proteomics data and taking into consideration recent experimental results such as the discovery of Icp55. This leads to greatly improved performance of presequence cleavage site prediction (reducing misprediction of cleavage site position from ≈ 48% to ≈ 29%, addressing the longstanding and often discussed lack of accurate tools for this task. In addition, in the light these refined and novel presequence features, I cluster and discuss classes of presequences.

In chapter 4, I present sequence features of transmembrane domains (TMD) of proteins in the mitochondrial inner membrane, improving the discrimination between those regions and spuriously similar regions in soluble cytosolic proteins. The difficulty of predicting the TMDs of mitochondrial membrane proteins has been noted anecdotally, but the distinct characteristics of mitochondrial TMDs had not been analyzed from the viewpoint of computational biology. Therefore, I analyzed the problem, starting with a previous model which calculates the free energy of TMD membrane insertion using positional amino acid profiles, based on parameters measured for the TMDs of E.R. membranes. As expected, TMDs of the mitochondrial inner membrane show characteristics distinct from either E.R. TMDs or spurious hydrophobic regions of cytosolic globular proteins. However, in terms of free energy distribution, the mitochondrial TMDs overlap with those two distributions, leading to difficult prediction. My statistical analysis surprisingly shows that glycine is significantly enriched in the center of mitochondrial TMDs and negatively charged residues show an asymmetric distribution, consistent with pioneering experimental work on mitochondrial TMDs. I employ these different characteristics to discriminate mitochondrial TMDs from spurious regions of cytosolic proteins using the sequences of those proteins and their homologs in other organisms, leading to much improved prediction of mitochondrial TMDs in comparison to general predictors for TMDs. I examined the position of predicted TMDs in proteins from a mitochondrial presequence dataset, especially the presequences with atypical features discussed in the chapter 3, and found some interesting cases of non-annotated cleavage sites that locate downstream of TMDs.

Finally, in chapter 5, I summarize, discuss, and conclude my thesis work.