1. Motivation

The proper localization of proteins in a cell is essential to perform their desired function. The 'address' of protein's destination is localised within its sequence in a form called a targeting signal. Ones of them are signal peptides, which are responsible for targeting proteins to the endomembrane system and their export outside the cell. Proteins equipped with signal peptides constitute a substantial fraction of all organismal proteins and play crucial roles in metabolism, maintenance of tissue structure, immune response and regulation of other organismal functions. Moreover, the passing proteins through the endomembrane system is important for their correct folding and posttranslational modification such as glycosylation and phosphorylation. The commonly accepted model of typical signal peptides assumes that they possess three regions with different amino acid composition. Signal peptides are always located at the beginning of proteins and have the length of 20-30 amino acid residues. The common model of signal peptides assumes that they consist of three regions with different amino acid composition. The most important is the region with the large number of hydrophobic residues. The signal peptide ends with a cleavage site, where it is cut off from a mature protein after serving its function.

The current signal predicting software is mainly based on sequences of frequently studied species, such as yeast, higher plants and mammals. However, our analysis of the large number of sequences representing diverse taxonomical groups indicates a great variability of signal peptides. At present, there is no universal signal peptide model. The most popular signal peptide predicting software is based on machine learning methods, which do not reveal attributes of signal peptides important for their prediction. Moreover, such solutions cannot reflect the full variability of signal peptides and improperly recognise atypical cases. They are also not able to answer questions regarding the organization and taxonomic variability of signal peptides, which makes the artificial signal peptide design impossible.

2. Aim of research

Therefore, the goal of this research is to create a statistical model for signal peptides, which will include knowledge about their organization, amino acid composition and variation. The model will be specialized for different taxonomic groups and types of signal peptides. It will be also implemented in software used to the signal peptide prediction and design.

3. Methods

The proposed model will be based on hidden semi-Markov models (HSMMs). Briefly, the hidden Markov model allows the assignment of a given sequence to a particular region of signal peptide based on its amino acid composition. The main advantage of this method is its ability to incorporate intrinsic knowledge about signal peptides. The information about architecture and length of signal peptide regions as well as their physicochemical properties significantly improves decision rules created during the learning phase of the algorithm. The model created in this way is very flexible and can adequately match even atypical signal peptides that have some functional domains or belong to very specific taxonomic groups, for example parasites. Due to these properties, the HSMM model does not have to be periodical relearned because of the constant increase in the number of sequences with signal peptides in databases. The universal decision rules are still similar despite changes in the training set. We also use a novel method of amino acid encoding based on their physicochemical properties. It enables extraction of the most crucial information from the signal peptide sequence, which allows us to create models for even very small or specific groups of proteins. The created model will be compared to state of the art signal peptide predicting software in the accuracy and precision of prediction. To make a fair benchmark test, we will also study the areas of competence of different predictors and create a meta-predictor of signal peptides using the knowledge of all analyzed methods.

4. Preliminary results

The first our model, called signalHsmm, has the biggest AUC value of the other software, equal to 0.98 (AUC is Area Under the Curve, the performance measure ranging from 0 to 1, where 0, 0.5 and 1 are absolutely wrong, random and ideal prediction, respectively). Moreover, our model effectively predicted signal peptides even when it was trained on very small data sets. For instance, the version called signalHsmm1987 trained on very small data sample known in 1987 (2311 sequences) identified peptides with efficiency very similar to the version trained on a much bigger data set (336 sequences).

The universality of the proposed model was check on signal peptides coming from taxonomic groups less known and poorly represented in databases. For example, for signal peptides of malaria parasite Plasmodium and related taxa AUC was 0.92, significantly better than in case of other software (0.84).

4. Impact of results

The final model will be implemented as a computer software and will be useful for both experimental biologists and bioinformaticians willing to study signal peptides. In addition to this, the proposed model speed up the design of artificial signal peptides in laborious and time-consuming laboratory researches. Our model will be also published as a signal peptide predictor. Proteins equipped with signal peptides play important roles in organisms. Therefore, any disorder in their main role, which is protein import into endoplasmic reticulum, can be a reason of many diseases. The efficient signal peptide model will minimize cost and speed up development of new drugs related with peptides. A good example are secretion proteins of malaria parasites Plasmodium, which are not recognized precisely enough by the existing software.