Prediction of Protein Localization for Specialized Compartments using Time Series Kernels

Marco Mernberger[1,3], Daniel Moog[2,3], Simone Stork[2,3], Stefan Zauner[2,3], Uwe G. Maier[2,3] and Eyke Hüllermeier[1,3]
[1]FB Mathematik/Informatik, [2]FB Biologie, Philipps-Universität Marburg
[3]LOEWE-Zentrum für Synthetische Mikrobiologie (SynMikro)

Abstract: The prediction of sub-cellular localization of proteins based on sequence information is a central task in bioinformatics. So far, many approaches have been developed, utilizing machine learning approaches such as neural networks and support vector machines. While being successful in general, predictive performance of these methods varies between different organisms. Moreover, standard methods are usually not well equipped to produce meaningful results in the case of specialized compartments, which are typically not covered by standard approaches. Yet, organisms exhibiting such compartments can be very interesting for the production of pharmaceuticals. In this paper, we present a rather unorthodox alternative approach to the prediction of sub-cellular protein localization, utilizing kernel-based machine learning methods combined with time-series metrics as distance measures on physicochemical models of amino acid sequences.

1 Introduction

Identifying the sub-cellular localization of proteins is an important task in bioinformatics. However, depending on the organism of interest, an experimental verification of a large number of proteins can be difficult, and in the case of high-throughput experiments outright infeasible. Hence, there is a need for reliable prediction tools capable of inferring protein localization from sequence. Thus, it comes at no surprise that many tools for the prediction of protein localization already exist. These tools exploit different machine learning techniques, such as hidden Markov Models (HMM) [JWVH$^+$03, BNvHB04], neural networks [BNvHB04, ENBvH00], support vector machines [SCLK05, BGR05, SCL$^+$07] or nearest neighbor classification [HPO$^+$07].

Predictions are typically based on the presence of targeting signals, such as signal peptide (SP) or mitochondrial targeting peptide (mTP). Alternatively, more general amino acid sequence features are exploited, e.g., amino acid composition [ENBvH00, BR04, HPO$^+$07], compartment-specific motifs and Pfam domains [ENBvH00], dipeptide composition [BR04] or averages of hydropathy index, charge and isoelectric point [BTM$^+$02].

Regardless of the classification method, the performance of prediction tools depends on the available training data, namely the set of proteins for which the sub-cellular localization is known. While ample data exists for many model organisms and standard compartments (e.g. mitochondria), this is not necessarily the case for specialized compartments occurring

only in certain organisms, such as complex plastids of diatoms or the apicoplast of *P. falciparum*, the causative agent of malaria, and other *Apicomplexa*.

Generally speaking, standard prediction tools roughly distinguish between basic groups of organisms, e.g., plant or animal cells [BNvHB04, ENBvH00]. Therefore, prediction focuses on standard compartments omnipresent in all cells, for which targeting signals are mostly known. As a consequence, standard tools are less suitable for specialized compartments for which this information is not available. In fact, it is not even possible to obtain a prediction for such cases, since they are generally not considered by the prediction tool. Alternatively, one can resort to motif searches and protein sequence comparison to predict the sub-cellular localization of proteins of interest. Yet, motif searches may also turn out unsuccessful, if the targeting signal is not highly conserved.

Here, we present an alternative approach using time series kernels to predict protein localization for a "special case" compartment of that kind: the secondary plastid of the diatom *Phaeodactylum tricornutum*. Instead of focusing on the amino acid sequence itself, we obtain a more general physicochemical model by mapping different amino acid properties to the sequence, thereby deriving a sequence of feature vectors. As distance measures on these models, we propose the use of time series metrics which can be used in conjunction with distance-based classifiers.

2 Complex plastids in *Pheaodactylum tricornutum*

In synthetic biology, eukaryotic organisms such as *P. tricornutum* are especially interesting, since they offer the potential to constitute synthetic biochemical pathways in distinct reaction compartments separated from the cytoplasm. This enables the production of potentially harmful products, e.g., pharmaceuticals, that could interfere with cytoplasmic proteins. Algae, such as *P. tricornutum*, have the additional advantages of a high reproduction cycle and cost-efficient cultivation, making them attractive for industrial scale biotechnology. To exploit these advantages, the protein import mechanism must be understood to ensure an efficient transport of engineered proteins.

According to the actual view, diatoms evolved by secondary endocytobiosis, i.e., by engulfing a primary photosynthetic red alga which was subsequently reduced to an organelle. As a consequence, diatoms contain complex plastids surrounded by four instead of two membranes. In the case of *P. tricornutum*, the outermost membrane of the complex plastid is in continuum with the endoplasmic reticulum (ER) [Gib81]. This is followed by the periplastidal membrane (PPM), which separates the periplastidal compartment (PPC) from the ER (see Fig. 1). Obviously, a more complex protein targeting system is required to import nucleus-encoded proteins into the diatom plastid.

Proteins destined for the complex plastid are imported into the ER via a Signal Peptide (SP) prior to reaching their destination, along with secretory proteins. PPC-resident and plastid proteins will then get transported into the PPC via a transit peptide, while the secretory proteins follow the default secretory pathway. It has been shown that an aromatic or bulky residue at the +1 position of a predicted SP cleavage site is necessary for plastid import

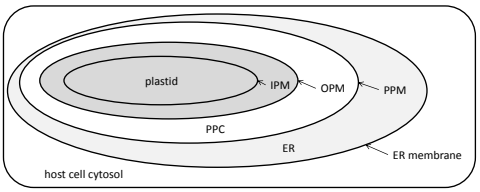

Figure 1: Secondary plastid of *P. tricornutum*. ER = endoplasmic reticulum, PPC = periplastid compartment, PPM = periplastid membrane, OEM = outer envelope membrane, IEM = inner envelope membrane

$$f_{AA} = \begin{pmatrix} pK_i(AA) \\ \dots \\ \text{hydropathy}(AA) \end{pmatrix}$$

$$\text{MTMRPRTKSL} \rightarrow (f_M, f_T, f_M, f_R, f_P, f_R, f_T, f_K, f_S, f_L)$$

Figure 2: Example of the modeling approach for the sequence MTMRPRTKSL. Each amino acid (AA) is modeled by a feature vector f_{AA}.

[GVH$^+$07]. Despite this knowledge, differentiating between secretory, PPC-resident and plastid proteins is a problem that has not yet been solved effectively.

3 Modeling of protein sequences

We represent amino acid sequences as a sequences of d-dimensional feature vectors f_{AA}, containing amino acid properties (e.g., hydropathy indexes, charges, etc.). This allows the simultaneous integration of different levels of information, as opposed to a single sequence. An example is given in Fig. 2. This is especially interesting if the sequence alone is insufficient to detect similarities among the proteins, simply because no strongly conserved motif can be found. Instead, focusing on patterns of general properties of the amino acids might be more successful, as these could be the result of certain important secondary structures, e.g., amphiphilic helices, which play an important role in mitchondrial protein import [LNS00]. For an amphiphilic α-helices, the charge of the amino acid is obviously more relevant than the actual amino acid type. Our more general model might help to detect such characteristics more efficiently.

4 Time series metrics as distance measures

While mostly known from other fields (e.g., economics, mathematical finance or signal processing), times series analysis has been used in bioinformatics mainly for the analysis

of microarray data [OV10]. In time series analysis, several distance metrics are known that internally derive an alignment of the time series, very similar to deriving a sequence alignment of protein sequences. In fact, many of these metrics are calculated via dynamic programming, similar to the well-known Needleman-Wunsch algorithm [NW70], the major difference being the recursive equation defining the alignment cost.

Here, we will focus on four of these measures: Dynamic Time Warping (DTW), [BC96], Edit Distance on Real sequences (EDR) [CÖO05], Edit Distance with Real Penalty (ERP) [CN04] and Longest Common Sub-Sequence (LCSS) [VGK02] [1], defined as follows:

$$\delta_{DTW}(x^d_{[1,n]}, y^d_{[1,m]}) = \begin{cases} 0 & \text{if } m = n = 0 \\ \infty & \text{if } m = 0 \underline{\vee} n = 0 \\ \|x^d_{[1]} - y^d_{[1]}\|_2 + \min\{\delta_{DTW}(x^d_{[2,n]}, y^d_{[2,m]}), & \text{else} \\ \delta_{DTW}(x^d_{[2,n]}, y^d_{[1,m]}), \delta_{DTW}(x^d_{[1,n]}, y^d_{[2,m]})\} & \end{cases}$$

$$\delta_{ERP}(x^d_{[1,n]}, y^d_{[1,m]}) = \begin{cases} \sum_{i=1}^{n} |x^d_{[i]} - g| & \text{if } m = 0 \\ \sum_{i=1}^{m} |y^d_{[i]} - g| & \text{if } n = 0 \\ \min\{\delta_{ERP}(x^d_{[2,n]}, y^d_{[2,m]}) + |x^d_{[1]} - y^d_{[1]}|, & \text{else} \\ \delta_{ERP}(x^d_{[2,n]}, y^d_{[1,m]}) + |x^d_{[1]} - g|, & \\ \delta_{ERP}(x^d, y^d_{[2,m]}) + |y^d_{[1]} - g|\} & \end{cases}$$

$$\delta_{EDR}(x^d_{[1,n]}, y^d_{[1,m]}) = \begin{cases} m & \text{if } n = 0 \\ n & \text{if } m = 0 \\ \min\{\delta_{EDR}(x^d_{[2,n]}, y^d_{[2,m]}) + dist_{EDR}(x^d_{[1]}, y^d_{[1]}), & \text{else} \\ \delta_{EDR}(x^d_{[2,n]}, y^d_{[1,m]}) + dist_{EDR}(x^d_{[1]}, g), & \\ \delta_{EDR}(x^d_{[1,n]}, y^d_{[2,m]}) + dist_{EDR}(y^d_{[1]}, g)\} & \end{cases}$$

$$\delta_{LCSS}(x^d_{[1,n]}, y^d_{[1,m]}) = \begin{cases} 0 & \text{if } m = 0 \vee n = 0 \\ \delta_{LCSS}(x^d_{[2,n]}, y^d_{[2,m]}) + 1 & \text{if } \forall d : |x^d_{[1]} - y^d_{[1]}| \leq \epsilon \\ \max\{\delta_{LCSS}(x^d_{[2,n]}, y^d_{[1,m]}), & \text{else} \\ \delta_{LCSS}(x^d_{[1,n]}, y^d_{[2,m]})\} & \end{cases}$$

with x^d, y^d denoting d-dimensional time series, g denoting a d-dimensional gap vector and $dist_{EDR} = 0$ if $|x^d_{[i]} - y^d_{[i]}| \leq \delta$ and 1 otherwise. In order to use support vector machines, the distance metrics have to be converted into a similarity measure s in order to yield a kernel function on time series. This is achieved by using

$$s(x^d, y^d) = e^{-\delta(x^d, y^d)} \tag{1}$$

with δ denoting a distance metric on time series.

[1] Note that LCSS is in fact a similarity rather than a distance measure

Using time series metrics to compare protein sequences might seem unorthodox at first, since we are actually not analyzing time series data. Yet, the problems of comparing time series and comparing protein sequences are similar, since both come down to comparing linearly ordered sequences of observations. In our case, the order information has a spatial (positional) rather than a temporal interpretation.

One advantage of times series measures is the fact that they can be multidimensional. Hence it becomes possible to compare sequences of feature vectors, regardless of the dimension d, allowing to incorporate multiple layers of information. More importantly, time series metrics account for shifts and stretches in the sequences, which allows for considering patterns that are only approximately similar. Again for the example of amphiphilic helices, the exact position of a charged amino acid might not be crucial, i.e., a residue position may vary slightly or a charged residue may even be missing at some positions, without disrupting the amphiphilic character of the helix. Allowing for tolerances is therefore mandatory to capture such approximate similarities.

5 Results

To assess the performance of the time series approach on the example of *P. tricornutum*, we used a test data set consisting of 133 protein sequences, whose location within the diatom are known due to function or experimentally verified. This data set contains 37 verified PPC-resident proteins, 53 plastid proteins and 43 secretory proteins.

For each sequence, an SP was predicted using SignalP [BNvHB04], further supporting the conjecture that each protein is imported into the ER. In each case, the remaining sequences after the predicted SP cleavage site were converted into a sequence of 2-dimensional feature vectors, using pKI and the hydrophobicity values according to the Hopp-Woods scale [HW83] as attributes. In this initial study, we limited ourselves to only two features to avoid the risk of including too many potentially irrelevant features that might obscure discriminating information with background noise. Hydrophobicity and pKI value were chosen since both are already considered relevant features in the literature [BTM$^+$02]. All sequences were truncated after the first 15 residues, since the transit peptide responsible for further protein targeting was assumed to be located after the SP cleavage site.

Tables 1 and 2 summarize the classification results using leave-one-out cross validation for different two-class classification scenarios. Results were obtained by k-nearest neighbor classification and an SVM using the LibSVM package [CL11]. Additionally, the Euclidean distance (ED) is included as a baseline.

Table 1 shows the accuracy for separating plastid proteins from non-plastid proteins as a proof-of-concept. Table 2a shows the classification accuracy for predicting PPC-directed secretory vs. (plastid and ppc-resident) proteins, Table 2b for predicting secretory vs. PPC-resident proteins.

As additional baselines, we included predictions by three of the most widely used methods for localization prediction: MultiLoc2 [BBK09], TargetP [ENBvH00] and WoLF PSORT [HPO$^+$07]. MultiLoc2 employs an SVM approach utilizing sequence motifs and

	ED	DTW	EDR	ERP	LCSS
1-NN	69.9	85.0	80.5	72.2	75.9
3-NN	73.7	89.5	78.2	72.9	82.0
5-NN	71.4	90.2	82.0	78.2	82.0
7-NN	75.2	90.2	82.7	82.7	76.7
SVM	76.7	89.5	82.0	84.2	78.9

Table 1: Percent of correct classifications for plastid vs. non-plastid proteins.

	ED	DTW	EDR	ERP	LCSS
1NN	65.4	72.2	72.9	82.0	82.0
3NN	69.9	72.2	77.4	83.5	79.7
5NN	68.4	75.9	80.5	80.5	81.2
7NN	68.4	77.4	79.7	77.4	80.5
SVM	75.9	76.7	69.2	69.2	82.7

(a) PPC-directed vs. secretory

	ED	DTW	EDR	ERP	LCSS
1NN	62.2	81.1	68.9	64.4	68.9
3NN	71.1	86.7	76.7	73.3	77.8
5NN	74.4	88.9	77.8	76.7	76.7
7NN	73.3	86.7	78.9	80.0	71.1
SVM	76.7	88.9	73.3	75.6	63.3

(b) PPC-resident vs. secretory

Table 2: Predictive accuracy (percent of correct classifications) for separating secretory proteins from PPC-directed and PPC-resident proteins.

amino acid composition in conjunction with gene ontology, TargetP uses a neural network and hidden Markov models to check for targeting signals, while WoLF PSORT utilizes weighted nearest neighbor classification based on known targeting signals and correlative sequence features.

Note that these tools in principle return more information than our binary classifiers, being predictors for multiple locations. For the pairwise classification, the majority of these classes are irrelevant, since we are interested in a separation of plastid and non-plastid proteins, respectively, secretory and non-secretory proteins. Obviously, this will result in an overestimation of the accuracy of these approaches, since, for example, a secretory protein that is predicted to be mitochondrial by TargetP will still be considered correctly classified as non-plastid, although the actual prediction is wrong. The results for the three pairwise classifications are given in Table 3.

In each classification scenario, the time series kernels achieve higher accuracy than the ED, indicating the benefit of allowing for a certain tolerance regarding stretches and shifts.

	plastid vs. non-plastid	sec vs. PPC-directed	sec. vs. PPC
TargetP	65.4	56.4	63.8
WoLF PSORT	65.4	68.4	50.0
MultiLoc2	67.7	61.7	55.0

Table 3: Classification accuracy for MultiLoc2, TargetP and WoLF PSORT, using standard parametrization.

Moreover, the time series metrics outperform all three competitors.

In the first experiment, one could also classify plastid proteins based on the presence of a bulky or aromatic amino acid at the +1 cleavage site position of the signal peptide. In the case of *P. tricornutum*, this yields an accuracy of 88.7%. Without using this explicit knowledge, the time series approach still yields reasonable accuracies, and in the the case of δ_{DTW} also surpasses this value. However, such information is known for some but not all algae species, such information is not available.

6 Discussion

In this paper, we proposed an intriguing new application of time series kernels for subcellular localization prediction. The results showed that the methodology is capable of producing reasonable results, even outperforming widely used baseline methods.

In preliminary experiments, we found that the transit peptide could not be reliably detected by standard approaches or motif searches (data not shown). Yet, we are still able to separate the secretory proteins from the proteins imported into the PPC to some extend. Plastid-resident proteins could be distinguished from secretory and PPC-resident proteins as well. In the case of *P. tricornutum*, this could in principle also be achieved based on the signal peptide cleavage site. Yet, such information is generally not available for other species, for which the identification of plastid proteins is still an open problem.

In our opinion, the promising results justify further investigation, for example by using related time series kernel methods or including additional features, such as alternative hydrophobicity scales, charge or bulkiness of the residues. Moreover, it could be interesting to expand the methodology to other species. This will be addressed in future work.

Literatur

[BBK09] T. Blum, S. Briesemeister und O. Kohlbacher. MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, 10(1):274–285, 2009.

[BC96] D.J. Berndt und J. Clifford. Finding patterns in time series: a dynamic programming approach. In *Advances in Knowledge Discovery and Data Mining*, Kapitel 9, Seiten 229–248. American Association for Artificial Intelligence, 1996.

[BGR05] M. Bhasin, A. Garg und G.P.S. Raghava. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21(10):2522–2524, 2005.

[BNvHB04] J.D. Bendtsen, H. Nielsen, G. von Heijne und S. Brunak. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, 340(4):783–795, 2004.

[BR04] M. Bhasin und G.P.S. Raghava. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, 32(Suppl. 2):W414–W419, 2004.

[BTM+02] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai und S. Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305, 2002.

[CL11] C. Chang und C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[CN04] L. Chen und R. Ng. On the marriage of Lp-norms and edit distance. In *VLDB'04: 30th International Conference on Very Large Data Bases. Proceedings*, Seiten 792–803, Toronto, Canada, September 2004.

[CÖO05] L. Chen, M. T. Özsu und V. Oria. Robust and fast similarity search for moving object trajectories. In *SIGMOD'05: ACM SIGMOD International Conference on Management of Data. Proceedings*, Seiten 491–502, Baltimore, USA, June 2005.

[ENBvH00] O. Emanuelsson, H. Nielsen, S. Brunak und G. von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, 300(4):1005–1016, 2000.

[Gib81] S.P. Gibbs. The chloroplast endoplasmic reticulum: structure, function, and evolutionary significance. *International Review of Cytology*, 72:49–99, 1981.

[GVH+07] A. Gruber, S. Vugrinec, F. Hempel, S.B. Gould, U.G. Maier und P.G. Kroth. Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Molecular Biology*, 64(5):519–530, 2007.

[HPO+07] P. Horton, K.J. Park, T. Obayashi, N. Fujita, H. Harada, CJ Adams-Collier und K. Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, 35(suppl 2):W585–W587, 2007.

[HW83] T.P. Hopp und K.R. Woods. A computer program for predicting protein antigenic determinants. *Molecular Immunology*, 20(4):483–489, 1983.

[JWVH+03] A.S. Juncker, H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen und A. Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, 12(8):1652–1662, 2003.

[LNS00] C.M. Lee, W. Neupert und R.A. Stuart. Mitochondrial targeting signals. In *Protein, Lipid and Membrane Traffic: Pathways and Targeting*, Jgg. 322, Seiten 151–159. IOS Press, 2000.

[NW70] S.B. Needleman und C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[OV10] C. Orsenigo und C. Vercellis. Time series gene expression data classification via L 1-norm temporal SVM. *LNCS. Pattern Recognition in Bioinformatics*, 6282/2010:264–274, 2010.

[SCL+07] E.C.Y. Su, H.S. Chiu, A. Lo, J.K. Hwang, T.Y. Sung und W.L. Hsu. Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics*, 8(330):12, 2007.

[SCLK05] D. Sarda, G.H. Chua, K.B. Li und A. Krishnan. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*, 6(1):152–164, 2005.

[VGK02] M. Vlachos, D. Gunopoulos und G. Kollios. Discovering similar multidimensional trajectories. In *ICDE'02: 18th International Conference on Data Engineering. Proceedings*, Seiten 0673–0685, San Jose, USA, February 2002.