Weighted decomposition kernels for protein subcellular localization

Menchetti S.(1), Costa F.(1), Frasconi P.(1)

(1) Dipartimento di Sistemi e Informatica, Università di Firenze.

Motivation

Knowledge about the subcellular localization of proteins can provide important information about their function. A reliable automatic classification method for predicting subcellular localization from sequences is therefore a valuable tool to shed light on protein function and may help towards the solution of genomic scale level problems such as the identification of pharmaceutical targets.

Methods

Prediction of protein subcellular localization can be naturally formulated as a supervised learning problem where each protein sequence is classified into one of the main compartements (cytoplasm, nucleus, mithochondrium and extra-cellular for eukariotic organisms or cytoplasm, periplasm and extra-cellular for prokariots). The representation of protein sequences may play a crucial role on the effectiveness of prediction methods. Although widely used, methods that rely on flat representations of sequences (e.g. amino acid composition) cannot exploit the presence of motifs or other regularities at the level of subsequences for determining the compartment in which the mature protein will reside. Here we propose weighted decomposition kernels for comparing sequences based on the co-occurrence of contextualized k-mers. Unlike the spectrum kernel and other kernels based on counting the number of co-occurences of short k-mers, a weighted decomposition kernel can assign different importance to different matching k-mers depending on the surrounding context, that may consist of several flanking residues. To avoid the sparseness problem when matching long contexts, we propose computing a kernel between probability distributions fitted on subsequences localized around the matching k-mer. Product probability kernels or histogram intersection kernels can be used to determine the weight associated with the context.

Results

We compare our method against three other approaches. The first is a connectionist model proposed by Nair & Rost (2003) that uses diverse sources of information such as amino acid composition, solvent accessibility, secondary structure, and evolutionary information. The second one is a support vector machine using amino acid composition developed by Hua and Sun (2002). The third one is the spectrum kernel that estimates the similarity of two proteins counting the co-occurrences of short substrings in the amino acid sequence. Using a set of SWISSPROT protein sequences for empirical evaluation, We found that weighted decomposition kernels can significantly reduce the prediction error compared to the three above methods (we measured a relative prediction error reduction ranging from 15% to 38%).

Contact email: p DASH f AT dsi DOT unifi DOT it
URL: http://www.dsi.unifi.it/neural
Supplementary Information: The prediction program is available on request from the authors.