ELSEVIER

Expert Systems with Applications 28 (2005) 537–545

Expert Systems with Applications

www.elsevier.com/locate/eswa

KSPF: using gene sequence patterns and data mining for biological knowledge management

Hei-Chia Wang[a,*], Hung-Chih Kuo[a], Hong-Hwa Chen[b,c], Yu-Yun Hsiao[b], Wen-Chieh Tsai[b]

[a]*Institute of Information Management, National Cheng Kung University, 1st University Road, Tainan, 701, Taiwan*
[b]*Department of Life Sciences, National Cheng Kung University, 1st University Road, Tainan, 701, Taiwan*
[c]*Institute of Biotechnology, National Cheng Kung University, 1st University Road, Tainan, 701, Taiwan*

Abstract

Most traditional approaches for annotating protein families are not efficient because of high throughput sequences, complex analytic tools and unordered literature and results cannot be reused. Here, we describe a framework, knowledge sharing for protein families (KSPF), that uses sequence pattern data mining and knowledge management to improve upon traditional approaches. It is divided into three modules: automation, retrieval and refinement. This framework builds an environment that allows biological researchers to submit an unknown protein sequence and search for information on its sub-family. Once this sub-family protein category has been found, the related literature and knowledge records provided by previous users can be retrieved. The possible functions of the protein can then be predicted by use of the literature and records. The proposed framework is applicable to all types of protein families. We describe the search for a plant lipid transfer protein (PLTP) with use of the framework. The system KS-PLTP functions to map an unknown sequence to the sub-family of the PLTP knowledge base and predict the sequence's possible function. The prediction rate of KS-PLTP reached 89.6%.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Gene sequence pattern; Data mining; Knowledge management; Function prediction; Literature retrieval

1. Introduction

Gene annotation is an important source for representative functional information, because it reveals many sequence functions and valuable information (Eisenberg, Marcotte, Xenarios, & Yeates, 2000; Hieter & Boguski, 1997; Nowak, 1995). Even though gene annotation is an important reference for researching protein families, annotation for a single gene is not enough to recognize the function of a protein family because protein family annotation is more complex than sequence annotation.

Most traditional methods cannot predict the possible functions of an unknown sequence and to which subfamily a protein belongs. Here, we propose a framework, knowledge sharing for protein families (KSPF) that combines data mining and knowledge management to annotate protein families. The framework builds a decision tree and a knowledge base. It involves collecting domain-related literature and functional terms defined by team researchers. Data for the protein family are downloaded from public databases and organized into a decision tree by use of a C4.5 algorithm. The literature and function terms are divided into different subfamily groups. When an unknown sequence is categorized to a subfamily of the decision tree, the researcher can refer to the literature and use the pre-defined function terms to predict the functions of the unknown sequence.

KSPF is applicable to all kinds of protein families, including lipid transfer proteins. Lipid transfer proteins have many different functions in plant physiology. For example, the surface layers of the cell wall of plants are made up of hydrophobic polyesters (Hincha, Neukamm, Sror, Sieg, & Weckwarth, 2001; Pyee, Yu, & Kolattukudy, 1994) that protect plant organs against biotic and abiotic stresses (Blein, Pierre, Marion, & Ponchet, 2002; Buhot et al., 2001; Wijaya, Neumann, Condron, Hughes, & Polya, 2000). How

* Corresponding author. Tel.: +886-6-2757575 53724; fax: +886-6-2362162.
E-mail addresses: hcwang@mail.ncku.edu.tw (H.-C. Wang), frankie@mail2000.com.tw (H.-C. Kuo), hhchen@mail.ncku.edu.tw (H.-H. Chen), yunhsiao@mail2000.com.tw (Y.-Y. Hsiao), rhingharte@yahoo.com.tw (W.-C. Tsai).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2004.12.015

to determine the exact biological functions of these proteins is thus an important task. Plant lipid transfer proteins (PLTPs) are small, soluble, basic proteins characterized by their ability to catalyze the transfer or exchange of lipids between membranes in vitro. PLTPs are abundant in expressed sequence tag databases of plants (Douliez et al., 2001; Hincha et al., 2001; Pastorello et al., 2000; Segura, Moreno, & Garcia-Olmedo, 1993; Sohal, Pallas, & Jenkins, 1999). However, manually annotating the functions of unknown PLTP genes from a large amount of data is time consuming and unmanageable. KSPF can be an efficient way to predict the functions of PLTP families.

In the following sections, we give a brief background of the research in searching for functions of proteins; introduce the KSPF framework in detail; discuss the implementation of a practical system, knowledge sharing for plant lipid transfer proteins (KS-PLTP); and demonstrate the efficiency and functions supported by this system.

2. Background

To analyze protein families traditionally, researchers download data from public databases such as PubMed or Entrez from the National Center for Biotechnology Information (NCBI) and use software to perform multiple alignment (with CLUSTAL W, Thompson, Higgins, & Gibson, 1994), phylogenetic analysis (with PHYLIP, Felsenstein, 1989; Lim & Zhang, 1999), or motif search (with HMMEer, Durbin, Eddy, Krogh, & Mitchison, 1998). For more information, researchers read these secondary data and study related literature that is not pre-filtered. This process is complicated, and most of the results are not organized and cannot be reused. If researchers' knowledge can be kept and reused, it could facilitate the team member understanding. This idea is widely used in many domains. (Manjarrés, Pickin, & Mira, 2002; Martínez Fernández & García-Serrano, 2000)

The KSPF framework we propose can simplify this process and allow for reusing researchers' knowledge. In many cases, researchers in the same team are doing similar research, so the same documents are viewed repeatedly. If team members can keep their study notes, much time could be saved for other readers in the same team. Knowledge management is a good way of doing this.

3. KSPF

Recently, many literatures have pointed to the need for information technology in automating gene annotation (Bazzan, Engel, Schroeder, & da Silva, 2002; King, Karwath, Clare, & Dehaspe, 2000a,b, 2001; Kretschmann, Fleischmann, & Apweiler, 2001). However, current research rarely predicts or retrieves information on protein families. Thus, we propose an automation framework, KSPF, to classify protein families.

The KSPF workflow, as shown in Fig. 1, is designed to obtain information on the annotation and functions of protein families of unknown sequence. The first step is to download all the data for the protein family in the domain of interest from the PubMed or Entrez databases. The files are in eXtensible Markup Language (XML) and contain a lot of useful information. A parser program is used to separate the contents into two parts: the sequences and some other useful information. The latter is further separated into information sets for subfamilies by use of cluster software and stored into a knowledge base. A pattern recognition program, Pratt (Jonassen, Collins, & Higgins, 1995), is used to generate all possible patterns from all these sequences. The generated

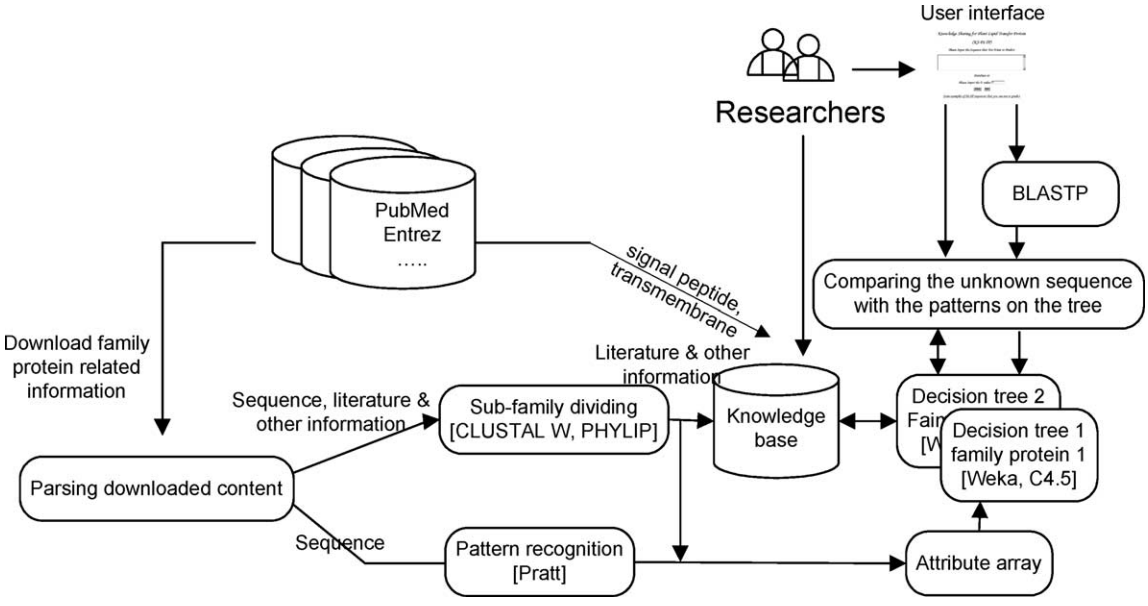

Fig. 1. Workflow of KSPF.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات