HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Patrik Koskinen[1], Petri Törönen[2], Jussi Nokso-Koivisto[2], Liisa Holm[1,2]

[1]Department of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland
[2]Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Contact: patrik.koskinen@helsinki.fi

PANNZER - A high-throughput tool for functional annotation of unknown protein sequences

PANNZER

Introduction

When genome sequencing cost level is at 1000$ or less (3rd and 4th generation sequencing) (Rusk, 2009) and the genome sequencing takes days or hours rather than years or months the number of sequencing projects will grow exponentially. Furthermore, sequencing will be applied by a wider range of scientific fields. Due to the huge amount of various sequence data and diverse methods used in the functional annotation process, some of these sequences and structures are annotated incorrectly (Hadley, 2003; Naumoff, 2004; Punta, 2008). The error rate in public databases has been estimated to be as high as 40% (Schnoes, 2009). Due to the high error level in databases, the traditional nearest neighbour methods are propagating errors through databases. A correctly annotated proteome is the corner stone of a successful genome research project and therefore accurate and reliable tools for this purpose are needed. Here we present a high-throughput tool that uses a weighted k-nearest method in functional annotation (See figure 1).

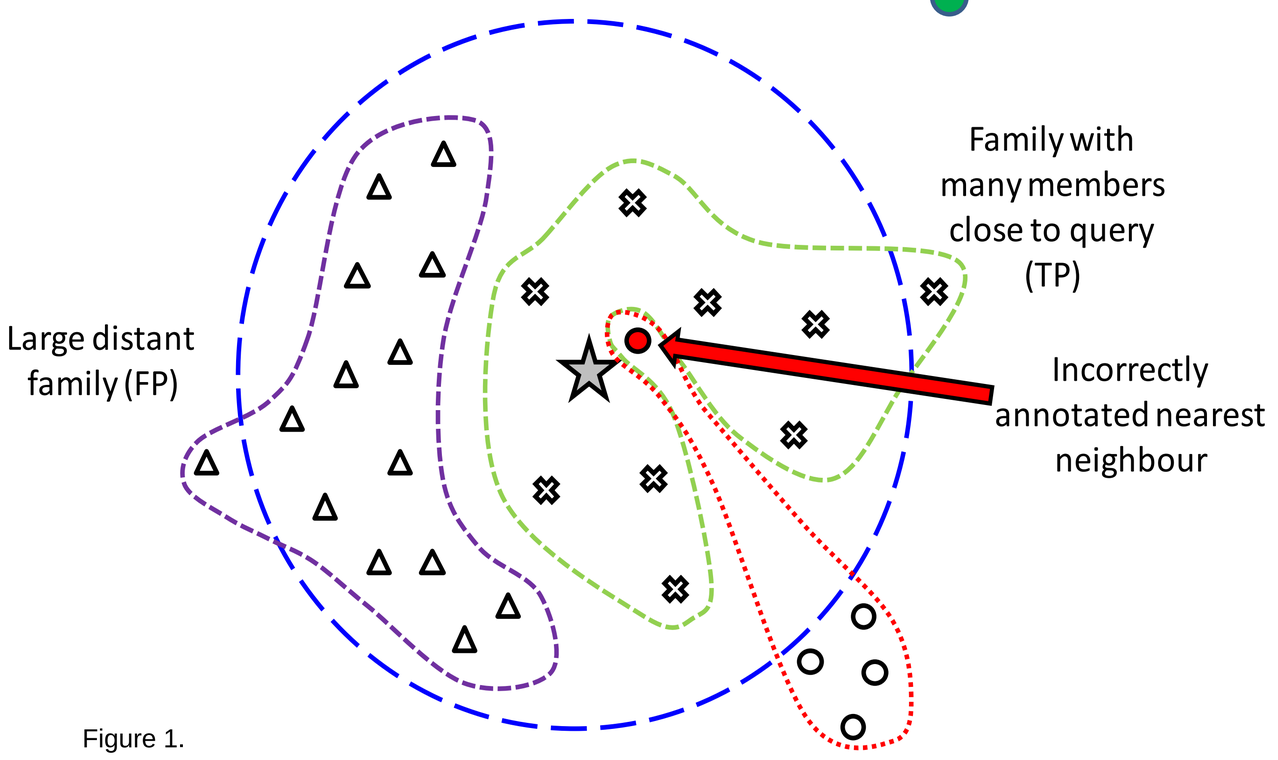

Figure 1.

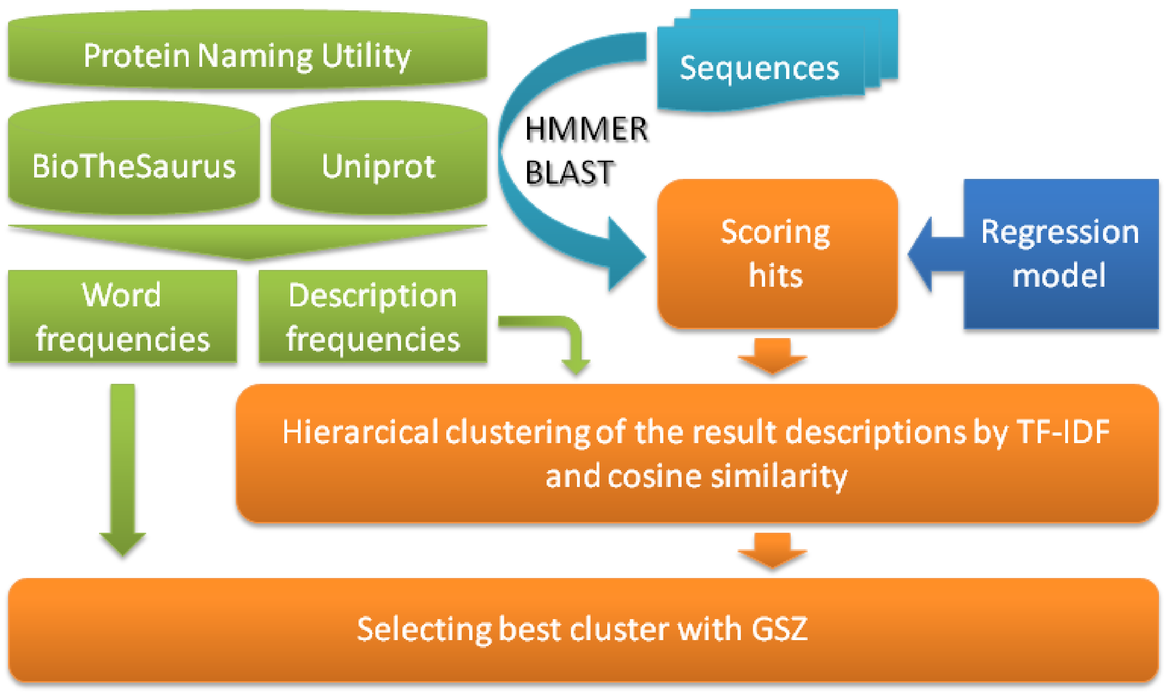

Figure 2. Workflow in PANNZER

Results

The first evaluation set was picked randomly from HAMAP database and annotated by using the UniProt database from which the evaluation dataset was removed. PANNZER was able to recover the original function in 1311 out of 1415 cases (success rate of 93%).

Another evaluation set was a manually annotated proteome of *Leuconostoc gasicomitatum* (unpublished). Comparison was made against RAST, a method which outperformed HAMAP, PGAAP, IMG and IGS annotation pipelines (Blande et al. unpublished). When making TF-IDF comparison against the original description ,PANNZER outperformed RAST in 1152 (60%) cases out of 1913. RAST performed better than PANNZER in 397 (21%) cases and in 364 (19%) cases it was stalemate.

Key concepts

High quality descriptions

PANNZER uses BioTheSaurus (Liu, 2006) to map alternative descriptions for the Uniprot entries and Protein Naming Utility (Goll, 2009) for correct nomenclature.

Regression model

The regression model is constructed by taking the various output values from the BLAST/HMMER search and calculating taxonomic distance between the species. These were used as input to regression analysis which tried to predict the TF-IDF similarity between the descriptions of the target and query sequences. Furthermore, we treated taxonomic distance with a nonlinear model that downweights the similarity to paralogs in the same species, quickly rises to highlight the similarity to orthologs in near relative species and slowly drops to downweight the similarities to sequences in distant species.

TF-IDF and cosine similarity

TF-IDF (Term Frequency–Inverse Document Frequency) is a statistical measure of word importance to a document in a corpus. Common words have lower weights than uncommon ones. After creating TF-IDF vectors for descriptions, we calculate a cosine similarity between these two description vectors.

GSZ (Gene Set Z-score)

GSZ represents a weighted version of the standard hypergeometric Z-score (Törönen, 2009). It monitors the abundance of the descriptions in the background set (database), the size of the selected sequence set and the average and variance of sequence scores in the selected set. GSZ takes into account all the scores of cluster members and also the ratio of cluster members in homology results compared to background.

References:

Goll, J., Montgomery, R., Brinkac, L. M., Schobel, S., Harkins, D. M., Sebastian, Y., Shrivastava, S., Durkin, S. & Sutton G. 2009. The Protein Naming Utility: a rules database for protein nomenclature. Nucleic Acids Research

Hadley, C. 2003. Righting the wrongs. *EMBO Reports*. **Vol 4, NO 9:** 829-831

Liu, H., Hu, Z., Zhang, J. & Wu C. 2006. BioThesaurus: a web-based thesaurus of protein and gene names. BMC Bioinformatics. Vol 22, Issue 1

Naumoff, D. G., Xu, Y., Glansdorff, N. & Labedan B. 2004. Retrieving sequences of enzymes experimentally characterized but erroneously annotated: the case of the putrescine carbamoyltransferase. *BMC Genomics* **Vol 5**:52

Punta, M., Ofran, Y. 2008. The Rough Guide to In Silico Function Prediction, or How To Use Sequence and Structure Information To Predict Protein Function. *PLOS Computational Biology*. **Vol 4, Issue 10**

Rusk, N. 2009. Cheap third-generation sequencing. *Nature Methods* **Vol. 6, No. 4:** 244-245.

Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. 2009. Annotation error in Public Databases: Misannotation of Molecular Funtion in Enzyme Superfamiles. *PLOS Computational Biology*. **Vol 5, Issue 12**

Törönen, P., Ojala P.J., Marttinen P. and Holm, L. 2009. Robust extraction of functional signals from gene set using a generalized threshold free scoring function. *BMC Bioinformatics.* **10**:307.