

Fast Homology Search using Categorisation Profiles

Abhijit Chattaraj¹

Hugh E. Williams¹
{abhijit,hugh,cannane}@cs.rmit.edu.au

Adam Cannane¹

¹ School of Computer Science & IT, RMIT University, GPO Box 2476V, Melbourne, 3001, Australia.

Keywords: automatic categorisation, profiles, SCOP, BLAST, homology search

1 Introduction

Homology search is an important step in discovering evolutionary relationships in modern molecular biology. In particular, it is key to analysing data from large-scale sequencing initiatives: by establishing homology between a newly-discovered sequence and well-understood sequences in curated repositories, it is possible to infer structure and function without the need for costly, time-consuming wet laboratory work.

Biologists almost always search protein collections in preference to nucleotide collections because, in general, proteins are better annotated and, therefore, more useful in discovering evolutionary relationships. It is possible to deduce the function of a protein based on its similarity to sequences that have been previously characterised by comparing the primary or linear structure of a protein. Indeed, more than half of the data derived from newly-sequenced genomes can be characterised based on its similarity to other well-understood organisms [5]. As new genomes are sequenced and annotated, function elucidation for this new data becomes increasingly possible.

2 Categoriser Directed Homolog-family Prediction: CHOMP

We propose a novel technique, CHOMP, that uses categorisation techniques [4] adapted from text information retrieval to address the limitations of BLAST [1]. CHOMP works as follows: first, a training collection of pre-categorised proteins is used to train a categoriser to develop sequence profiles; second, for each unknown query sequence, the sequence is compared to the profiles and the best-matching profiles determined; third, the original sequences that were used to derive the profiles are retrieved to form a small collection; and, last, the unknown sequence is compared to the small collection using homology search techniques. The benefit of our approach is that only a small fraction of the original collection is retrieved in response to each query, resulting in faster homology searching.

3 Results

We have found that a version of CHOMP that includes a Rocchio-based categoriser [4], a simple inverse document frequency (IDF) similarity measure [6], and uses BLAST for homology search in the final stage is twice as fast as BLAST alone and around 1% more accurate for searching SCOP data [2]. The collection and queries we use are derived from the SCOP database [2]: 32,126 sequences are used in training the categoriser and as the target search collection, and the remaining 14,855 sequences are used as queries.

Table 1: Overall effectiveness and efficiency of homology search techniques. Asterisks indicate that the results are statistically significant relative to the performance of BLAST.

Technique	Total time (minutes)	Average Precision (%)	Average MRR (%)
BLAST	86.09	73.85	99.1
PSI-BLAST	569.26	77.14*	98.4
FASTA	551.35	75.94*	99.3
CHOMP-BLAST	43.14	74.78*	98.2
CHOMP-FASTA	74.37	78.04*	98.2

Table 1 shows the overall performance of CHOMP compared to the BLAST [1], PSI-BLAST [1], and FASTA [3] tools. The accuracy of our categorisation techniques are reported using the standard measures of recall and precision [6], and mean reciprocal of rank (MRR). We measure success of the search task by rewarding a scheme that correctly identifies the sequences in the superfamily from which a query is derived, and ranks these as highly as possible; recall measures the fraction of correct answers retrieved, precision measures the fraction of answers that are correct, and mean average precision is the average precision at each answer recalled. MRR calculates the reciprocal of the rank at which the first correct answer appears. If a correct answer appears at rank 4, the $MRR \frac{1}{4} = 0.25$. The overall scores are the mean of individual scores for each query.

Table 1 shows that BLAST is over six times faster for searching SCOP than FASTA, with an average precision that is less than 2% lower. PSI-BLAST is almost an order of magnitude slower than BLAST but over 3% more accurate in average precision. CHOMP is both fast and accurate. Our results show that combining CHOMP categorisation with BLAST searching improves on BLAST searching alone by almost 1% in accuracy with almost exactly twice the speed. When paired with FASTA, the results are even more impressive: CHOMP-FASTA is 14% faster than BLAST and over 4% more accurate. Indeed, CHOMP-FASTA is the most accurate scheme we tested, surpassing the other popular profile-based scheme, PSI-BLAST.

All results are statistically significant improvements verified using the Wilcoxon signed ranked test. We conclude CHOMP is a valuable new method for protein homology search of well-curated collections.

References

- [1] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, Volume 25, Number 17, pages 3389–3402, 1997.
- [2] A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia and A.G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, Volume 32, pages D226–D229, 2004.
- [3] W.R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, Volume 183, pages 63–98, 1990.
- [4] J. Rocchio. *The SMART Retrieval System—Experiments in Automatic Document Processing*, Chapter Relevance Feedback in Information Retrieval, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [5] L. Stein. Genome annotation: From sequence to biology. *Nature Reviews Genetics*, pages 493–503, 2001.
- [6] I.H. Witten, A. Moffat and T.C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, 2nd edition edition, 1999.