MinSet: automatic derivation of maximally representative subsets of protein structural domains

ID - 151

Pandini Alessandro[1], Bonati Laura[1], Fraternali Franca[2], Kleinjung Jens[3]

[1]Dipartimento di Scienze dell'Ambiente e del Territorio, Università degli Studi di Milano-Bicocca, Milano, Italy

[2]Bioinformatics Unit, King's College, London, UK

[3]Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, London NW7 1AA, UK

Motivation

Protein structure databases provide the foundation for the understanding of molecular function at the atomic level. While offering an increasing number of depositions per year, they also include a significant degree of redundancy. When protein structure databases are used for the development or assessment of computational methods, redundancies increase the computational load without providing any new information.

The current size of the protein structure databases is already prohibitively large for computational methods such as molecular dynamics and docking. Even if the chosen 'base set' of structures is a fraction of the entire structure database, it may be advantageous to reduce redundancy by selecting a small and representative subset.

Therefore, the here presented work is aimed to derive protein structure subsets with a maximal representativeness with respect to the range of structural fragments in the originating database (base set) and with a minimal degree of redundancy.

Methods

The MinSet method is a generic procedure for database subsetting that is applicable to any type of string encoded data. The current MinSet implementation focusses on distant homologous protein structures, which are currently domains translated to structure strings by using a structural alphabet. A combination of Genetic Algorithm (GA) and Suffix Tree (ST) data structure allows for an efficient selection of optimised subsets. The selection mechanism of the GA utilises an entropic fitness score that is based on the distribution of all substrings of a given fixed length (k-word dictionary). For example, with k=5, those subsets containing a large number and even distribution of length-5 words in the (collective) structure string are favoured over those with less words and/or uneven distribution. Therefore, when compared to the base set, the optimised subset has a minimal redundancy and maximal information with respect to the k-word dictionary. Structures are encoded as strings with the aid of a structural alphabet. Thus, k-words of the structure string are synonymous of structure fragments of length k.

Results

The MinSet method was applied on the SCOP40 database which includes domains with less than 40% sequence identity. Subsetting was performed on the SCOP 'class' level: the protein domains of each class were employed independently as base set for reduction.

Maximally representative subsets were generated by varying the selection k-word length (3, 5, and 7) and by restraining the subsets to target sizes of 5%, 10%, 15%, and 20% relative to the base set size. The fitness score of the optimised subset was assessed statistically as Z-score by comparison with the scores of random subsets of the same size. A size-invariant subset quality measure was calculated in the form of the coverage: the fraction of k-words in the base set that are preserved in the optimised subset. The resulting subsets and the corresponding statistical analysis are available for download at http://mathbio.nimr.mrc.ac.uk/~jkleinj/MinSet.

Availability: http://mathbio.nimr.mrc.ac.uk/~jkleinj/MinSet

Email: alessandro.pandini@unimib.it