Bioinformatics Advance Access published January 3, 2007

BIOINFORMATICS

MinSet : A general approach to derive maximally representative database subsets by using fragment dictionaries and its application to the SCOP database

Alessandro Pandini,[a] Laura Bonati,[a] Franca Fraternali,[b] and Jens Kleinjung [c,*]

[a]Dipartimento di Scienze dell'Ambiente e del Territorio, Università degli Studi di Milano-Bicocca, Milano, Italy. [b]Bioinformatics Unit, King's College, London, UK. [c]Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, London NW7 1AA, UK.

Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: The size of current protein databases is a challenge for many Bioinformatics applications, both in terms of processing speed and information redundancy. It may be therefore desirable to efficiently reduce the database of interest to a maximally representative subset.

Results: The MinSet method employs a combination of a Suffix Tree and a Genetic Algorithm for the generation, selection and assessment of database subsets. The approach is generally applicable to any type of string-encoded data, allowing for a drastic reduction of the database size whilst retaining most of the information contained in the original set. We demonstrate the performance of the method on a database of protein domain structures encoded as strings. We used the SCOP40 domain database by translating protein structures into character strings by means of a structural alphabet and by extracting optimised subsets according to an entropy score that is based on a constant-length fragment dictionary. Therefore, optimised subsets are maximally representative for the distribution and range of local structures. Subsets containing only 10% of the SCOP structure classes show a coverage of $> 90\%$ for fragments of length 1-4.

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

Contact: jkleinj@nimr.mrc.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Protein sequence and structure databases contain a substantial amount of redundant information. For many Bioinformatics applications it is advantageous to reduce the base set of the databank to a subset by elimination of proteins whilst retaining as much information as possible. The MinSet method presented here is a novel and holistic approach that operates simultaneously on a very large collection of strings. Thus, the method is applicable to databases containing protein sequences, string-encoded protein structures or any other string-encoded information. The subset generation is an optimisation process aimed at creating a reduced collection of strings, whose overall composition in terms of short substrings (words) contains the maximal amount of information about the original large set. Depending on the type of data, substrings represent, for example, motifs (sequence data) or structure fragments (structure data). The employed entropic target function favours subsets with a diverse and evenly distributed substring composition. The application of the MinSet method on the SCOP40 protein domain structure database is presented here. Several subsets with maximal representativeness of structure fragment conformations were derived. The choice of the SCOP database was motivated by the availability of a hierarchical structure classification scheme and structure quality measures.

Structures of the SCOP40 (v1.69) database (Chandonia *et al.*, 2004), which includes domains with less than 40% sequence identity, were transformed into sequences by using a structural alphabet (Camproux and Tuffery, 2004). Only X-ray structures with complete backbone chains and SPACI quality scores greater than 0.4 were used. Subsetting was performed on the SCOP 'class' level, *i.e.* on the protein domains of a each 'class' category (α, β, α/β, $\alpha+\beta$) as base set. Each subset was generated by randomly assigning to all protein domains of the base set either 0 (='excluded') or 1 ='included'), yielding a binary array as minimal subset descriptor. Additionally, an explicit subset descriptor was derived by concatenating the 'included' structure strings (with protein domain delimiters). The minimal subset descriptor was used by a Genetic Algorithm, the explicit subset descriptor served for a Suffix Tree data structure (see below).

In subset selection and assessment we employed dictionaries of constant-length words (k-words), the latter being simply substrings of structure strings. For example, with $k = 5$ the k-word dictionary is the collection of all words of length 5 in the concatenated structure string, or synonymously, all penta-peptide fragments of that structure collection. Transformation of the subset structure string to a Suffix Tree data structure allowed for very fast (linear time complexity) dictionary matching in the computation of entropy and coverage as explained below. The fitness score of each subset j was calculated as $S_j^k = H_j^k\,(1 - D_{k,1})$. The term $H_j^k = \sum_i - p_{i|k}^j\, log(p_{i|k}^j)$ is the Shannon entropy with $p_{i|k}^j$ specifying the probability of dictionary word i in subset j when using a dictionary of word length k. Maximising the Shannon entropy is equivalent to maximising the information content of the subset with regard to the k-word dictionary. Practically the required frequencies were derived by searching the entire subset dictionary against the Suffix Tree of the subset. The term $D_{k,1} = \sum_i p_{i|k}\, log(p_{i|k}/q_{i|1})$ is the Kullback-Leibler divergence, serving here to restrain the subset string composition to the expected composition as reported in Table 1 of Camproux and Tuffery (2004). A Genetic Algorithm was employed to search for the subset with highest fitness score S_j^k. The Genetic Algorithm was run with a population size of 2500 genomes over 100 generations, cross-over breeding of the fittest 10% of genomes and elitism (fittest genomes survive). Subsets of different target size were derived by restraining the maximal number of included proteins to a fixed ratio t relative to the base set, for example 10% ($t = 10$). Restraining was achieved by random exclusion of proteins during the optimisation process. Due to the nature of the scoring function, in some cases the optimal subset size turned out to be smaller than the target size.

*to whom correspondence should be addressed

Table 1. Statistics of subsets for four structure classes derived with parameters k_s = 3, 5 and 7 and t = 10 (see text). k_s: subset selection word length; C^c_{cum}: cumulative coverage, summed over k-word lengths 1 to c.

k_s	class	base set	subset	C^3_{cum}	C^5_{cum}	C^7_{cum}	Z-score
3	α	1069	58	2.93	4.24	5.15	8.6
3	β	1506	62	2.92	3.83	4.20	9.9
3	α/β	1183	111	2.98	4.37	5.22	9.9
3	α+β	1269	126	2.98	4.42	5.34	9.2
5	α	1069	106	2.97	4.52	5.68	11.2
5	β	1506	150	2.97	4.29	4.98	13.3
5	α/β	1183	118	2.98	4.49	5.47	10.9
5	α+β	1269	126	2.98	4.50	5.50	9.2
7	α	1069	106	2.97	4.55	5.75	11.4
7	β	1506	150	2.97	4.28	4.98	13.8
7	α/β	1183	118	2.98	4.49	5.48	10.1
7	α+β	1269	126	2.98	4.50	5.52	8.9

Subsets were derived for four SCOP classes (α, β, α/β, $\alpha+\beta$), using selection on three k-word lengths (k_s = 3, 5, and 7) and five target subset sizes (t = 5, 10, 15 and 20), yielding 48 optimised subsets in total.

The statistical significance of the fitness score of the optimal subset was analysed by standardisation to the Z-score of a background distribution derived from 2500 random subsets. Independent from the fitness score analysis, the quality of the optimal subset j was also assessed by calculating its coverage $C^j_k = q^j_k/r_k$, with q^j_k the number of k-words in the base set with a match in subset j and r_k the number of all k-words in the base set. Here the frequencies were obtained by searching the entire base set dictionary against the Suffix Tree of the subset.

Computational time for subsetting an SCOP class containing in the order of 1000 proteins is about 2 hours on a Linux machine with a single 64bit 2.8 GHz CPU.

Subsets were optimised on an entropic fitness score S^k_j (see Methods) as quantitative measure of representativeness, combining maximal k-word diversity with closeness to a given native composition. By using a structural alphabet, k-word diversity is synonymous with diversity of local structural fragments. Therefore, the selected subset is a collection of domains with maximal representativeness of local conformations of the protein backbone. This is not equivalent with a representative subset of the structural classification of the SCOP database, because local structure diversity does not transcend directly to fold space diversity.

Subset properties are illustrated here on the example of subsets selected on k-words of length 5 (k_s = 5) and target size 10% (t = 10), each comprising about 10000 to 20000 different k-words. The statistical significance of the optimised subset scores *versus* a distribution of random subset scores is shown in Supplementary Figure 1a. Random subset scores form Gaussian curves, while the corresponding optimised scores are indicated by an arrow of the same line type. Z-values of about 10 σ (Table 1) demonstrate the high effectiveness of the Genetic Algorithm selection. Furthermore, being implicitly an information metric, the employed fitness score maximises the information content of the subset with respect to the k-word dictionary.

Subset coverages indicate the amount of preserved information as well as the redundancy of the base set. The coverage of our example subset with respect to the base set is greater than 90% for fragments of length k = 1 to 4 and decreases to about 30% at length k = 9 (Supplementary Figure 1b). The coverage decreases with increasing k because the number of k-words grows with i^k for small k (i being the number of structure alphabet symbols). Although this subset was selected on k_s = 5, words of length k greater than 5 show a reasonably high coverage. Statistics for subsets of target size 10% (t = 10) are shown in Table 1. The cumulative coverages as well as the Z-scores indicate an improvement when selecting on k_s = 5 instead of k_s = 3, but similar results for selection on k_s = 5 and k_s = 7.

It is apparent from the presented data that the subsets of different SCOP classes show qualitatively the same picture, but quantitatively differences are quite large. Naturally the two 'mixed' classes α/β and $\alpha+\beta$ are structurally more diverse than the 'pure' classes α and β, which is reflected in a higher entropic fitness score (Supplementary Figure 1a). It seems that the higher diversity also accounts for a slightly better coverage of the mixed classes for shorter k-word lengths, probably because the fragment diversity within each protein is comparatively large and their subsets are therefore more likely to achieve high coverage than subsets of the more uniform 'pure' classes.

It should be noted that in this MinSet application the inclusion/exclusion acts at the domain level. This imposes an interdependence between fragments occuring within the same domain. Therefore, the derived subsets achieve maximal representativeness within the constraints of this interdependence. In general, rare fragments are likely to be included, but some may be excluded if belonging to domains that contain otherwise highly redundant fragments. To give an account of these competing effects, domain lists and sequence lists for base set and subsets are reported on the MinSet website as well as lists of included or excluded k-words.

The usage of a constant-length word dictionary and the above definition of representativeness are both choices that have proved successful, but the MinSet method is not limited to these. The combination of a Genetic Algorithm and a Suffix Tree is a fast and extensible approach to the general problem of extracting representative subsets from large collections of proteins.

Resulting from the application on the SCOP database, the optimised subsets are a collection of domains selected on the basis of local conformations, without being biased by constraints on fold diversity. With an appropriate choice of structural alphabet and fragment dictionary, the MinSet method can be easily adapted to achieve representativeness on more global structural features.

Optimisation on local conformations was motivated by the fact that protein structure fragments are gaining increasing importance in structure prediction, conformational searching and docking procedures. The derived subset can be of specific interest for all knowledge-based parametrisations of methods for prediction of local features. Other potential applications of the MinSet method include the fast analysis of secondary structure segmentation, the identification of ordered/disordered regions and the detection of functionally important fragments.

REFERENCES

Camproux, A C, G. R. and Tuffery, P. (2004). A hidden markov model derived structural alphabet for proteins. *J. Mol. Biol.*, **339**, 591–605.

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–92.