Journal of Structural Biology xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

TRDistiller: A rapid filter for enrichment of sequence datasets with proteins containing tandem repeats

François D. Richard [a,c], Andrey V. Kajava [a,b,c,*]

[a] Centre de Recherche de Biochimie Macromoléculaire, UMR5237 CNRS, University of Montpellier 1 and 2, 1919 Route de Mende, 34293 Montpellier Cedex 5, France
[b] University ITMO, 197101 St. Petersburg, Russia
[c] Institut de Biologie Computationnelle, 95 rue de la Galéra, 34095 Montpellier, France

ARTICLE INFO

Article history:
Received 10 December 2013
Received in revised form 14 March 2014
Accepted 17 March 2014
Available online xxxx

Keywords:
Non-globular proteins
Proteomes
Tandem repeats
Sequence analysis
Short string composition

ABSTRACT

The dramatic growth of sequencing data evokes an urgent need to improve bioinformatics tools for large-scale proteome analysis. Over the last two decades, the foremost efforts of computer scientists were devoted to proteins with aperiodic sequences having globular 3D structures. However, a large portion of proteins contain periodic sequences representing arrays of repeats that are directly adjacent to each other (so called tandem repeats or TRs). These proteins frequently fold into elongated fibrous structures carrying different fundamental functions. Algorithms specific to the analysis of these regions are urgently required since the conventional approaches developed for globular domains have had limited success when applied to the TR regions. The protein TRs are frequently not perfect, containing a number of mutations, and some of them cannot be easily identified. To detect such "hidden" repeats several algorithms have been developed. However, the most sensitive among them are time-consuming and, therefore, inappropriate for large scale proteome analysis. To speed up the TR detection we developed a rapid filter that is based on the comparison of composition and order of short strings in the adjacent sequence motifs. Tests show that our filter discards up to 22.5% of proteins which are known to be without TRs while keeping almost all (99.2%) TR-containing sequences. Thus, we are able to decrease the size of the initial sequence dataset enriching it with TR-containing proteins which allows a faster subsequent TR detection by other methods. The program is available upon request.

1. Introduction

Proteins contain a large portion of tandem repeat (TR) regions representing arrays of repeats that are directly adjacent to each other (Björklund et al., 2006; Heringa, 1998; Kajava, 2012; Marcotte et al., 1999). Highly divergent, they range from a single amino acid repetition to domains of 100 or more repeated residues. Over the last decade, numerous studies have demonstrated the fundamental functional importance of such TRs and their involvement in human diseases. Indeed, a number of examples have shown a high incidence of TRs in the sequences of virulence factors of pathogenic agents, toxins and allergens (Kajava et al., 2006). The genetic instability of these regions can cause a rapid response to environmental changes, and thus, can lead to emerging infection threats. This implies that this class of sequences may have a broader role in human diseases than was previously recognized. Thus, TR regions are abundant in proteomes and are related to major health threats. Along this line, the discovery of these domains, understanding of their sequence–structure–function relationship and mechanisms of their evolution promise to be a fertile direction of research.

The growth of proteomic data has led to increasing efforts to develop methods for TR recognition. Protein TRs are frequently not perfect, containing a number of mutations (substitutions, insertions, and deletions) accumulated during evolution, and some of them cannot be easily identified. One of the most sensitive approaches for *ab initio* identification of "covert" TRs, like HHrepID (Biegert and Söding, 2008), relies on HMM–HMM comparisons. However, these methods are relatively slow and, therefore, inappropriate for large-scale analysis of the proteomes. In this situation, algorithms that speed up TR analysis are urgently needed.

Abbreviations: FP, false positive; HMM, hidden Markov model; LRR, Leucine-Reach Repeat; ROC, Receiver Operating Characteristic; SS, short string; TR, tandem repeat; TPR, Tetratrico Peptide Repeat; TP, true positive.

* Corresponding author at: CRBM, CNRS, 1919 Route de Mende, 34293 Montpellier Cedex 5, France.
E-mail address: andrey.kajava@crbm.cnrs.fr (A.V. Kajava).

http://dx.doi.org/10.1016/j.jsb.2014.03.013
1047-8477/

The general idea behind the approach that is presented here was to use, prior to existing TR identification programs, a filter that is able to rapidly pre-select proteins with a high probability of containing TRs, while discarding information about their exact location, sequence pattern, and sequence alignments. The software developed based on these principles allows for the rapid reduction of the initial sequence dataset, enriching it with proteins containing TRs and, thereby, decreasing the time of the subsequent large scale analysis of protein TRs.

2. Methods

2.1. Rationale of the algorithm

TR regions contain similar sequence motifs (repeats) adjacent to each other and repeated several times. The sequence similarity of the repeats supposes that they have more similar (or identical) short words (or short strings) between them as compared to aperiodic sequences. Moreover, if we take into account the relative order of appearance of those short strings (SS) the difference between TR and no-TR regions should be even more pronounced. Fig. 1 illustrates this difference. If we take into account two letter SSs in a TR-containing sequence with a window size equal to the repeat length (Fig. 1A), "IA" is the first SS in the reference window 1, followed by "AG". The SSs "IA" and "AG" are also present in the window 2. Furthermore, they occur in the same order ("IA" is before "AG"). Hereafter, the common pairs of SSs between two windows while being in the same order will be called a "common pair". In total, the two windows in Fig. 1A have six such common pairs. At the same time, when considering two such windows in a sequence without TRs (Fig. 1B), fewer common pairs are unveiled. For example, the sequence in Fig. 1B has only one common pair. Therefore, the number of common pairs of SSs between the adjacent windows may allow distinguishing between TR and no-TR-containing proteins. The complexity of the described algorithm is O(n).

We considered several variables of the algorithm. The first of them is the number of windows in the scanning area (Fig. 2). The score is equal to a number of common pairs between two windows of the same length, as was described above. For this purpose we need at least two windows but their number in the scanning area can also be three, four, etc. The second variable is a way of the scoring procedure. If the number of the windows is more than two, one can imagine different combinations of intermediate scores obtained by the comparison of both adjacent and non-adjacent windows. The score of the area can be an average of all the computed intermediate scores. The third variable is the size of the SS. It can vary from one amino acid residue to the size of half of one repeat. We tested SSs containing either 1, 2, 3 or 4 amino acid residues. Finally, we can use different types of letters when we examine the commonality of the SS pairs. The performance of the program has been tested using either an exact amino acid alphabet or letters reflecting the amino acid groups accordingly to their physicochemical properties (for example, L, V, I, M, as hydrophobic residues; F, Y, W as aromatic residues; D, E, K, R as charged residues; Q, H, N, S, T as hydrophilic residues and A, C, G, P as individual residues that do not belong to any of these groups).

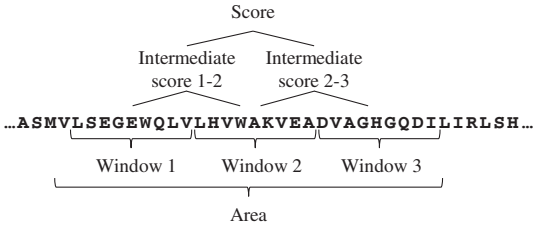

Fig. 2. A scheme of the score calculation used in the SS analysis. A scanning area consists of three windows of the same size. An intermediate score is computed as the number of common pairs of SSs between the adjacent scanning windows. The score of the area is an average of the intermediate scores.

To optimize the algorithm we tested all possible combinations of these variables. The best result was obtained when we scanned sequences with a three window area and counted common pairs of SSs between the first–second windows and the second–third windows. The score of the area represents an average of two intermediate scores (score 1–2 and score 2–3 in Fig. 2). We also established that the best type of the SS has two letters of the exact 20 amino acid alphabet.

2.2. Score and decisions

Based on the algorithm that counts common SS pairs in the adjacent windows we developed a filter (called TRDistiller) to remove no-TR-containing proteins from a set of sequences. The TRDistiller has been implemented under Python version 3.3.2. Here we describe the general scheme of the filtering process (Fig. 3).

We expected that the bias in the amino acid composition (low complexity) of sequences can blur the difference between TR and no-TR regions. By definition, the TR region with repeat length of less than 20 residues should have a low complexity. This bias increases with the decrease of the repeat length. Therefore, we move all proteins with low complexity regions to the TR-containing dataset (Fig. 3). The low complexity regions were identified

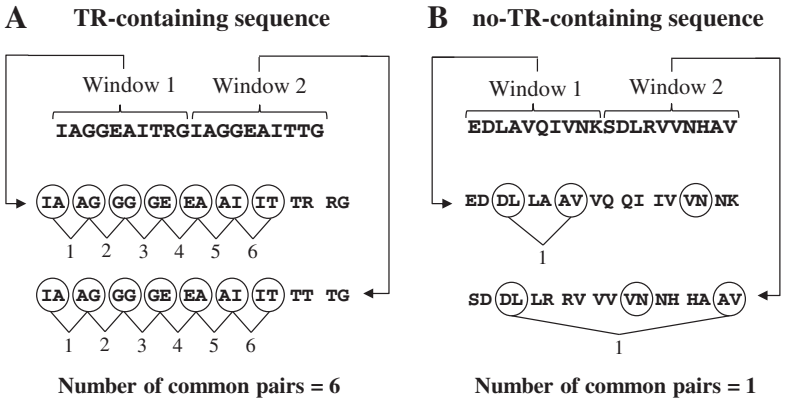

Fig. 1. Scoring algorithm applied to (A) the TR-containing and (B) the no-TR-containing sequences. It is based on examination of common pairs of SSs between the adjacent scanning windows. Common SSs between two windows are represented with circles. The common SSs that have the same order are linked and numbered.

Please cite this article in press as: Richard, F.D., Kajava, A.V. TRDistiller: A rapid filter for enrichment of sequence datasets with proteins containing tandem repeats. J. Struct. Biol. (2014), http://dx.doi.org/10.1016/j.jsb.2014.03.013