BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 10 2008, pages 1307–1309
doi:10.1093/bioinformatics/btn105

Sequence analysis

CompariMotif: quick and easy comparisons of sequence motifs

Richard J. Edwards[1,2,*], Norman E. Davey[1] and Denis C. Shields[1]

[1]UCD Complex and Adaptive Systems Laboratory and UCD Conway Institute of Biomolecular and Biomedical Sciences, University College Dublin, Dublin 4, Ireland and [2]School of Biological Sciences, University of Southampton, Boldrewood Campus, Southampton SO167PX, UK

Received on February 11, 2008; revised on March 18, 2008; accepted on March 19, 2008

Advance Access publication March 28, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Summary: CompariMotif is a novel tool for making motif–motif comparisons, identifying and describing similarities between regular expression motifs. CompariMotif can identify a number of different relationships between motifs, including exact matches, variants of degenerate motifs and complex overlapping motifs. Motif relationships are scored using shared information content, allowing the best matches to be easily identified in large comparisons. Many input and search options are available, enabling a list of motifs to be compared to itself (to identify recurring motifs) or to datasets of known motifs.

Availability: CompariMotif can be run online at http://bioware.ucd.ie/ and is freely available for academic use as a set of open source Python modules under a GNU General Public License from http://bioinformatics.ucd.ie/shields/software/comparimotif/

Contact: r.edwards@southampton.ac.uk

Supplementary information: Further details are available at http://bioinformatics.ucd.ie/shields/software/comparimotif/

1 INTRODUCTION

Short linear motifs (SLiMs) in proteins are functional microdomains of fundamental importance in many biological systems (Neduva and Russell, 2005). SLiMs typically consist of a 3 to 10 amino acid stretch of the primary protein sequence, of which as few as two sites may be important for activity. SLiMs can usually tolerate a number of alternative amino acids at one or more positions, making precise definitions extremely difficult. Because of this, and the way that SLiMs are commonly represented as regular expressions (e.g. R[SFYW].S.P), it can be hard to judge whether a given motif is similar to another. With the emergence of high-throughput SLiM prediction tools (Davey *et al*., 2006; Edwards *et al*., 2007; Neduva and Russell, 2006; Neduva *et al*., 2005), the need to quickly and easily identify recurring and/or previously described motifs is obvious.

CompariMotif is a novel tool for making motif–motif comparisons that identifies and scores similarities between motifs. When a new SLiM has been predicted computationally or discovered by experimental studies, CompariMotif enables similar motifs to be readily identified from published resources, such as the Eukaryotic Linear Motif (ELM) database (Puntervoll *et al*., 2003), Minimotif Miner (Balla *et al*., 2006) or PhosphoMotif Finder (Amanchy *et al*., 2007). Alternatively, comparing a list of motifs with itself might identify recurring motifs of interest. Although designed for protein motifs, which are the focus of this article, CompariMotif also has an option allowing the comparison of nucleotide motifs expressed as regular expressions. Currently, position-specific scoring matrix (PSSM) representations of motifs are not supported.

2 METHODS

Motifs are reformatted to standardize the regular expressions used and then the two motif sets are compared in a pairwise fashion, with all 'query' motifs compared to all 'search' motifs. First, the pair is assessed for a precise match (one motif is either the same as, or an exact substring of, the other). If a pair of motifs has no exact match but contains enough common amino acids (in any position) to have a potential match, then CompariMotif adopts a sliding window comparison in which every possible overlap between the two motifs are compared against each other (Fig. 1, see Manual and website for details). Matches must meet a minimum match requirement in terms of the numbers of positions that match, as determined by the user. Fixed positions in motifs are often more important than ambiguous ones, especially when the motif has been experimentally determined. For this reason, it is also possible to stipulate that all fixed positions in one or other motif (or both) match exactly to fixed positions in the compared motifs. When motifs have flexible wildcard positions, all variants of the motif are compared separately and the best match (if any) is used. Positions representing sequence termini must match other termini. Note, however, that motifs representing post-translational modifications etc. are not given special treatment and the user should pay special attention to whether specific important residues are included in a match.

For every comparison, each position in each motif is rated according to its relationship with the compared position in the other motif. This determines whether positions are matches, mismatches or some combination of variant/degenerate versions of ambiguous positions. If there are any mismatches—two defined positions that have no common amino acids—then the motif pair comparison is rejected. (This requirement can be relaxed by the user.) Otherwise, each positional comparison is rated for information content:

$$IC_i = -\log_N(f_a)$$

where IC_i is the information content for position i, f_a is the summed frequency for the amino acids (or nucleotides) at position i and N is number of amino acids (or nucleotides) in the alphabet, i.e. $N = 4$ for DNA and $N = 20$ for proteins. This is a modification of Shannon's Information Content (Shannon, 1997) such that a wildcard receives 0.0

[*] To whom correspondence should be addressed.

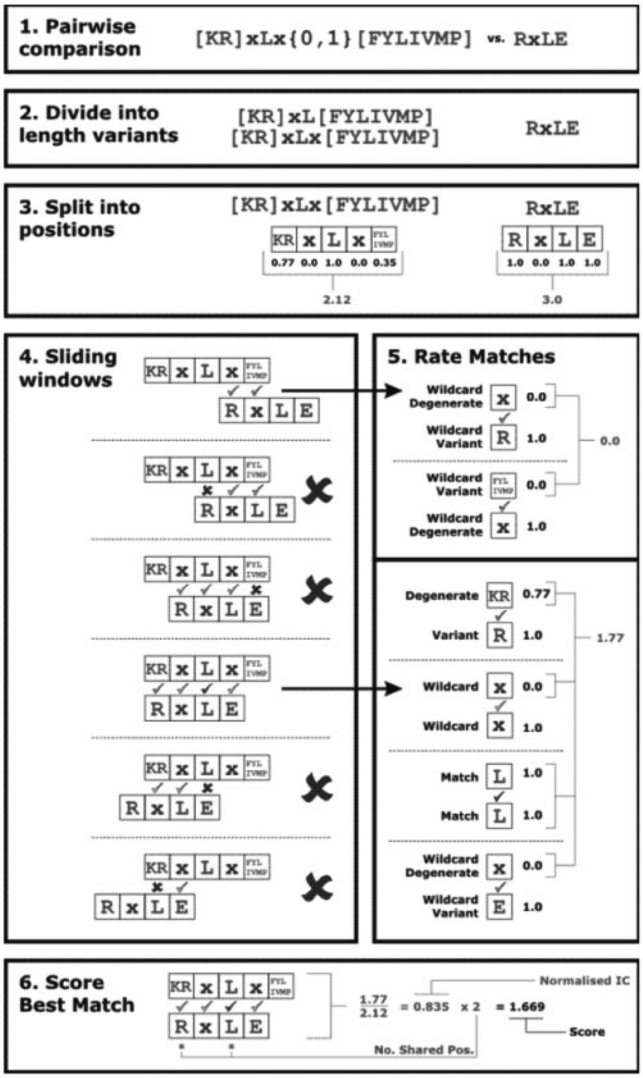

Fig. 1. Overview of CompariMotif. Motifs are first compared for precise matches. If these are not found, then CompariMotif adopts a sliding window comparison in which every possible overlap between (variants of) the two motifs are compared against each other. Matches must meet a minimum match requirement set by the user (see Manual and website for details).

and a fixed position scores 1.0 when a uniform frequency distribution is used. Ambiguous positions score between 0.0 and 1.0. When non-uniform frequencies are used, fixed rare amino acids ($f_a < 1/N$) will score above 1.0, while fixed common amino acids ($f_a > 1/N$) will score >1.0. Termini always get an IC$_i$ score of 1.0. For each comparison, the lower IC$_i$ value is used. For example a fixed variant matching an ambiguity will take the IC$_i$ of the ambiguity.

The IC for the match, IC$_m$ is simply the sum of the component IC$_i$ values. Multiple variants and/or sliding windows can produce multiple matches and so the comparison with the best overall IC$_m$ is selected as the best match for that motif pair. If two or more comparisons have the same IC$_m$, matches are ranked by the total number of matching positions and then by the number of exactly matching fixed positions. The best match (if any) that meets the minimum criteria set by the user is used to define the relationship between the two motifs, which is translated into a text description (Table 1, Fig. 2). These relationships are asymmetrical and comprised of one of four 'match type' keywords plus one of four 'match length' keywords, giving 16 categories in total. Because the raw IC$_m$ score for a given pairwise comparison is highly dependent on both the length and degeneracy of the matching motifs, an additional normalized IC score is calculated which divided the IC$_m$ by the lower IC of the two matching motifs. This reports IC$_m$ as a proportion of the maximum possible IC$_m$ value for that pair of motifs, given their length and degeneracy. This normalized IC is multiplied by the number of matched positions to give a heuristic CompariMotif Score to aid ranking of large results sets.

Table 1. Keywords and codes describing motif relationships

Match types	
Exact [e-]	All the matches in the two motifs are precise. Any ambiguous positions have all amino acids in common.
Variant [v-]	The query motif has variants of degenerate positions in the search motif, in addition to any exact matches.
Degenerate [d-]	The query motif has degenerate versions of positions in the search motif, in addition to any exact matches.
Complex [c-]	Some positions in the query motif are degenerate versions of positions in the search motif, while others are variants of degenerate positions.
Match lengths	
Match [-m]	Both motifs are the same length and match across their entire length.
Parent [-p]	The query motif is longer and entirely contains the search motif.
Subsequence [-s]	The query motif is shorter and entirely contained within the search motif.
Overlap [-o]	Neither motif is entirely contained within the other.

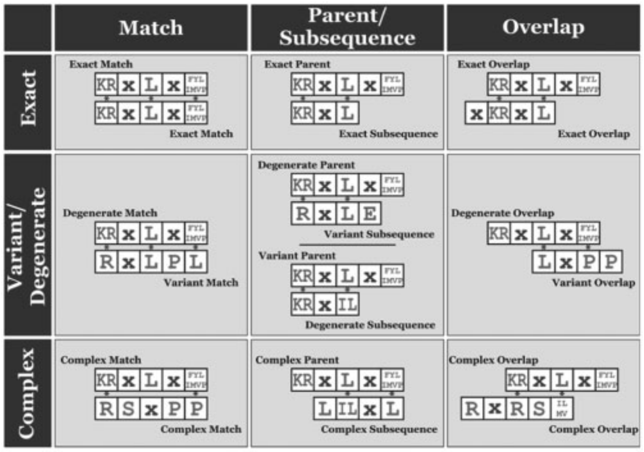

Fig. 2. Example CompariMotif match relationships for each of the 16 match types. In each case, the 'query' motif [KR]xLx[FYLIMVP] is compared to an invented motif for illustration. Because of the natural relationship between parent/subsequence and variant/degenerate matches, these have been grouped in the figure. Matched positions that contribute towards the number of matched positions (i.e. those not involving a wildcard position) are marked with an asterisk. More details can be found in the Manual and at the website.

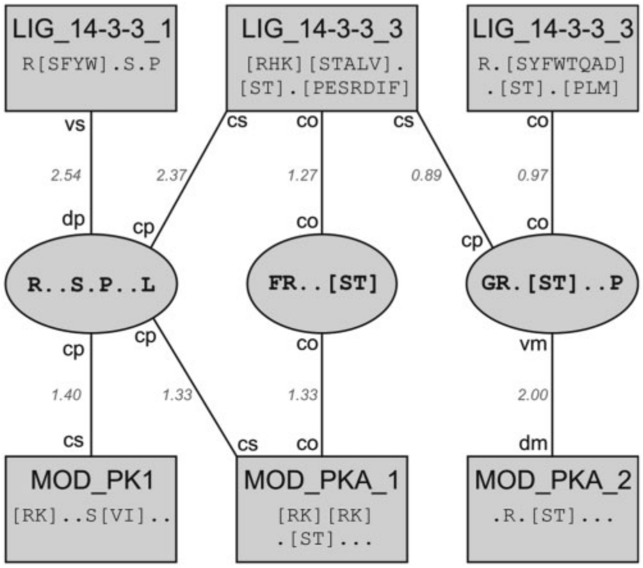

Fig. 3. Partial CompariMotif output of three motifs returned by SLiMFinder analysis of 14-3-3 interaction datasets. SLiMFinder motifs are shown as ellipses, while ELMs for known 14-3-3 ligands and phosphorylation sites are shown as rectangles. Two-letter codes match those given in Table 1. CompariMotif scores are given (grey italics) for each match.

3 RESULTS AND DISCUSSION

A typical application for CompariMotif is given in the SLiMFinder paper (see Example 1 in Edwards *et al.*, 2007), in which HPRD interaction datasets for 14-3-3 proteins (Mishra *et al.*, 2006) were analysed using SLiMFinder, returning several significant motifs ($P < 0.05$, see Table 2 in Edwards *et al.*, 2007). These motifs were compared to the ELM database (Puntervoll *et al.*, 2003) using CompariMotif with a 'normalized IC' cut-off of 0.4. Results were constrained such that fixed positions in an ELM must match a fixed position in the SLiMFinder motif. In total, eight out of 10 SLiMFinder motifs had matches with 17 ELMs. The eight motifs with matches fell into three main clusters: (1) three motifs matching known 14-3-3 motifs (Fig. 3), (2) three motifs matching SH3 binding motifs and (3) two motifs matching the highly degenerate LIG_PCNA_1 motif. In addition to the 14-3-3 and SH3 ELMs, matches to five phosphorylation ELMs were also identified; phosphorylation of the 14-3-3 motif is important for ligand recognition. A full visualization of these results with Cytoscape (Shannon *et al.*, 2003) can be found at the website and in the manual. These comparisons took <2 s to run on an Intel(R) Xeon(TM) dual 3.20GHz processor with 3Gb RAM.

It is beyond the scope of this applications note to discuss these results in detail. They do, however, highlight the ease with which CompariMotif can help to make sense of motif discovery results. As a simple, quick and high-throughput tool, CompariMotif can be an invaluable initial step in making sense of such data. Because of this, CompariMotif is now directly linked to both SLiMDisc and SLiMFinder web implementations (Davey *et al.*, 2007).

ACKNOWLEDGEMENTS

This work was funded by Science Foundation Ireland.

Conflict of Interest: none declared.

REFERENCES

Amanchy,R. *et al.* (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.*, **25**, 285–286.

Balla,S. *et al.* (2006) Minimotif Miner: a tool for investigating protein function. *Nat. Methods.*, **3**, 175–177.

Davey,N.E. *et al.* (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res.*, **34**, 3546–3554.

Davey,N.E. *et al.* (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455–459.

Edwards,R.J. *et al.* (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.

Mishra,G.R. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–414.

Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett*, **579**, 3342–3345.

Neduva,V. and Russell,R.B. (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res.*, **34** W350–355.

Neduva,V. *et al.* (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, e405.

Puntervoll,P. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.

Shannon,C.E. (1997) The mathematical theory of communication. 1963 *MD. Comput.*, **14**, 306–317.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.