SLiMDisc Server Accessory Applications

SLiMDisc (Short Linear Motif Discovery) is a motif discovery method for finding putative functional motifs in a group of proteins with a common attribute (1). The SLiMDisc webserver builds on the original SLiMDisc application by adding interactive masking and additional visualisations, as described in the main paper. In addition to these functions, the SLiMDisc server makes use accessory applications for further analysis of the basic SLiMDisc results. This supplementary material summarises the main functions, options and output of these applications. Please see the main software page (http://bioinformatics.ucd.ie/shields/software/) and the SLiMDisc help page (http://bioware.ucd.ie/~slimdisc/example/) for more details.

Three main accessory applications are used in post-processing:

1. SLiMPickings

This is a SLiMDisc post-processing application for filtering and re-ranking results, as well as producing conservation measures and alignments across putative orthologues for motif occurrences. The alignments are generated using GOPHER (below) to identify putative orthologues. SLiMPickings is also used in generating some of the additional SLiMDisc outputs of the main server. Main outputs from SLiMPickings are (a) a revised table of motifs returned from the dataset, with additional calculated statistics, and (b) alignments of the input proteins and putative orthologues with top motifs marked.

2. CompariMotif

CompariMotif is a simple application for comparing two lists of motifs to identify similar (or identical) motifs. In the SLiMDisc pipeline, CompariMotif will compare the top 20 motifs returned by SLiMDisc to a local database consisting of ELM (2), MiniMotif (3) and some additional SLiMs identified from the literature. Output from CompariMotif is a table of top 20 SLiMDisc motifs and the known SLiMs they match, given the chosen settings.

3. GOPHER

GOPHER is a tool for the rapid and automated generation of orthologous protein datasets. GOPHER uses GABLAM, which generates pairwise global alignment statistics from BLAST (4) local alignments. (See the SLiMDisc paper (1) and the GABLAM website (http://bioinformatics.ucd.ie/shields/software/gablam/) for details.) GOPHER attempts to combine the high speed of BLAST mutual best hit methods with the improved accuracy achieved by phylogenetic inference. Final alignments of putative orthologues are generated using MUSCLE (5).

These applications are summarised in the following sections.

1. SLiMPickings

SLiMPickings is a post-processor for SLiMDisc results, enabling generation of extra statistics (including conservation scores), filtering and re-ranking of motifs, and additional outputs. For a full description of SLiMPickings functions, please see the main SLiMPickings website (http://bioinformatics.ucd.ie/shields/software/slimpickings/). Here, the main functions relevant to the SLiMDisc webserver are briefly discussed.

SLiMPickings provides the following additional options to a basic SLiMDisc run:

- **Additional motif pattern statistics**. These are options for calculating more statistics based purely on the motif pattern alone.
- **Motif conservation statistics**. These use GOPHER (below) to identify putative orthologues for the input sequences.
- **Additional motif occurrence statistics**. These are options for calculating more statistics for the returned motifs based on the *occurrences* of the motif, rather than just the motif pattern. These can be returned as means and as percentiles.
- **Filtering and Ranking options**. These options allow results to be filtered by any combination of statistics returned and/or re-ranked on any return statistic.
- **Advanced Options**. These allow the generation of custom statistics (which can be used for ranking and filtering) and the uploading of an INI file containing SLiMPickings options that will over-ride web settings. This allows the user to easily repeat analyses with the same settings.

These options are described in the following table. Some of these options output additional columns in the results table, as indicated. Further information is available in the help file linked from the SLiMPickings page.

Server Option	Description	Column
Additional motif pattern statistics		
Net Charge	Whether to output net charge (positive - negative) of motif, eg. (KR) - (DE).	NetChg
Absolute Charge	Whether to output absolute number of charged positions (KRDE).	AbsChg
Charge Balance	Whether to output the "balance" of charge, *i.e.* the net charge of the N-terminal half of the motif, minus the net charge of the C-terminal half.	BalChg
AILMV	Whether to output if all positions in the motif are A,I,L,M or V.	AILMV
Aromatic	Whether to output count of F+Y+W.	Aromatic
Potential Phosphorylation sites	Whether to output potential phosphorylated residues... X if none or [S][T][Y], whichever are present.	Phos
Expect	Calculate (crude) expected occurrence of motif in search dataset.	Expect
ZScore	Calculate Z-Scores for each motif using the entire dataset.	ZScore
Motif conservation statistics		
Produce alignments of putative orthologues.	Produces an alignment for each protein against its putative orthologues, with motifs also marked	
Calculate mean conservation	Generate putative orthologues using GOPHER and calculate mean motif conservation of motif occurrences	Cons*
Select conservation score used.	This gives a choice of four scoring methods. See SLiMPickings manual for details.	
Conservation weighting	Weight given to global percentage identity for conservation, giving more weight to closer sequences. 0 gives equal weighting to all. Negative values will upweight distant sequences. See SLiMPickings manual for details.	
Incorporate ambiguity	Whether to calculate conservation allowing for degeneracy. Where a SLiMDisc motif has ambiguity, a position is considered conserved if a homologue has a substitution allowed by the ambiguity definition. If this option is disabled, an exact match to the specific fixed variant found at that occurrence is enforced. See SLiMPickings manual for details.	

Motif conservation statistics (ctnd.)		
Weight by information content	Whether to weight positions by information. See SLiMPickings manual for details.	
GOPHER Database selection	This selects from a number of local taxa-specific databases from which GOPHER will try to find orthologues.	
Additional motif occurrence statistics		
Mean Emini Surface Accessibility	Calculate mean Surface Accessibility (SA) using the method Emini *et al.* (6).	SA*
Mean IUPred disorder prediction	Calculate mean disorder using IUPred (7) prediction of intrinsically unstructured regions of proteins based on estimated energy content.	IUP*
Mean Eisenberg Hydrophobicity	Calculate mean Eisenberg Hydrophobicity (8) for occurrences.	Hyd*
Mean Charge calculations	Calculate selected charge statistics (above) for occurrences in addition to pattern.	NetChg*, AbsChg* & BalChg*
Calculate percentiles in addition to Mean values	If <>0 will return each occurrence statistic in percentile steps in addition to mean. E.g. 25 will return the percentiles 25,50,75 and 100. See SLiMPickings manual for details.	STAT_pcX*
Filtering and re-ranking		
Number of top motifs to return, re-ranked using a chosen ranking statistic	Re-ranks according to the chosen ranking statistic and only outputs top X new ranks. If set to 0 will not perform any re-ranking. The chosen ranking statistic should be a SLiMPickings column heading. The original SLiMDisc ranking is output as "OldRank".	OldRank
Perform filtering before re-ranking	Re-ranks the filtered dataset (True) rather than the whole (pre-filtered) dataset (False).	
Perform filtering before Z-score calculation	Calculate the Z-score on the filtered dataset (True) or the whole dataset (False).	
Restrict motifs to these input Proteins	Comma separated list (A,B,C) of proteins from input dataset for which to extract results.	
Restrict motifs to these SLiMs.	Comma separated list (A,B,C) of SLiMs (motif patterns) to extract.	
Custom filtering rules.	Results can be filtered on any output statistics using standard assessment operators >, >=, !=, =, >= , < . Details are available on the website.	
Advanced options		
List of custom scores	New custom scores can be generated using any combination of output statistics and basic mathematical operators. The scores can be used for filtering and ranking data if desired. Details are available on the website.	New Scores
Upload INI file with SLiMPickings options	Upload INI file containing SLiMPickings command-line options. (See SLiMPickings manual for details.) Will over-ride settings on this page.	

*Motif occurrence statistics are output as STAT_mean in the main table (and STAT_pcX if percentiles are used) but may be missing the "_mean" in the PDF output to save space. To use for filtering and ranking, make sure the "_mean" is included. (*e.g.* "Cons_mean" rather than "Cons").

2. CompariMotif

For a full description of CompariMotif functions, please see the main CompariMotif website (http://bioinformatics.ucd.ie/shields/software/comparimotif/) and/or download the full CompariMotif Manual. Here, the main functions relevant to the SLiMDisc webserver are briefly discussed. CompariMotif has only one output (see below), which is for the top 20 motifs returned by SLiMDisc. To analyse different motifs, use SLiMPickings to return the desired motifs and upload the delimited text file into the standalone CompariMotif server (http://bioware.ucd.ie/). By default, the options are quite relaxed and will return a large number of hits. The section below on Output can help interpret which of these may be interesting.

Server Option	Description
Ambiguity cut-off	The number of different amino acids allowed at one site before it is treated as a wildcard position. (Otherwise, it will count as one of the positions shared (see below).)
Minimum motif length	Minimum (non-wildcard) length of motifs to consider
Minimum number of shared positions	Matches between two motifs must share at least this number of non-wildcard positions.
Reverse motifs	Reverses motifs returned from SLiMDisc. For non-palindromic motifs, this option can give a sense of how many hits to expect giving the amino acid composition of the SLiMDisc results and the CompariMotif settings selected.
Must exactly match all fixed positions	By default, all fixed positions of known SLiMs must match fixed positions in the SLiMDisc motif but fixed positions in SLiMDisc motifs are allowed to match ambiguous positions in known SLiMs. This option allows the user to alter this restriction to apply to either, neither or both motifs.

CompariMotif Output

Output of CompariMotif is a sortable text table:

Output Column	Description
Name1	This will give the name of the SLiMDisc dataset (SLiMDisc_temp_X, where X is the job ID) followed by #R, where R is the rank of the motif that has hit a known SLiM.
Name2	This is the name of the known SLiM. This *may* be an ELM/MiniMotif ID.
Motif1	The motif definition (pattern) of the SLiMDisc motif.
Motif2	The motif definition (pattern) of the known SLiM.
Similarity1	The relationship of the SLiMDisc motif to the known SLiM. (See below.)
Similarity2	The relationship of the known SLiM to the SLiMDisc motif. (See below.)
MatchPos	The number of matching non-wildcard positions. (This includes positions that match due to ambiguities.) This must meet the minshare requirement (See above).
MatchIC	The calculated information content of the match. (See the Manual for details.)
NormIC	The calculated information content of the match normalised by the maximum possible MatchIC for that pair of motifs. In essence, higher scores indicate better matches, and a score of 1.0 indicates that no better match was possible between this pair of motifs. (See the Manual for details.)
Info1	The information content of the SLiMDisc motif. (Scaled so that 1.0 is the equivalent of 1 fixed position, no gap penalty.)
Info2	The information content of the known SLiM. (Scaled so that 1.0 is the equivalent of 1 fixed position, no gap penalty.)
Desc1	The description of the SLiMDisc motif, which is the same as Name1.
Desc2	The description of the known SLiM.

CompariMotif Relationships

The best match for any pair of motifs is considered to define the relationship between the two motifs. These relationships are comprised of the following keywords:

- Match type keywords identify the type of relationship seen:
  - **Exact** = all the matches in the two motifs are precise
  - **Variant** = the focal motif contains only exact matches and subvariants of degenerate positions compared to the other motif
  - **Degenerate** = the focal motif contains only exact matches and degenerate versions of positions in the other motif
  - **Complex** = some positions in the focal motif are degenerate versions of positions in the compared motif, while others are subvariants of degenerate positions
- Match length keywords identify the length relationships of the two motifs:
  - **Match** = both motifs are the same length and match across their entire length
  - **Parent** = the focal motif is longer and entirely contains the compared motif
  - **Subsequence** = the focal motif is shorter and entirely contained within the compared motif
  - **Overlap** = neither motif is entirely contained within the other

This gives sixteen possible classifications for each motif's relationship to the compared motif. (See the Manual for details.)

3. GOPHER

Automated orthologue retrieval is difficult and error-prone. Furthermore, orthology relationships depend on the focus of the analysis; lineage-specific duplications mean that several proteins can all be true orthologues of a single protein in another species despite being paralogues of each other. The SLiMDisc server uses the GOPHER algorithm for identifying putative orthologues. In essence, GOPHER removes genome-specific duplicates (in-paralogues) and identifies the closest protein from each species in the database when compared to the query protein. Details can be found at the GOPHER website (http://bioinformatics.ucd.ie/shields/software/gopher/).

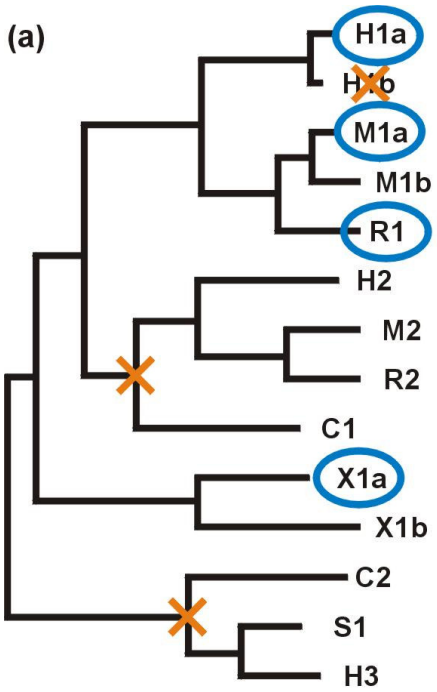
To be classed as an orthologue:

1. The query must be more similar to the hit than to any paralogues (of the query) *and* the hit must be more similar to the query than to any paralogues (of the query). This identifies putative orthologues within the post-duplication clade (Fig. 1, M1a/R1).
2. *Or* If the hit is more similar to any paralogues (of the query) than to the query itself, the query and those paralogues must be more similar to each other than either is to the hit. This identifies putative orthologues that diverged before the most recent duplication event for the query but this divergence event was itself a speciation not a duplication (Fig. 1, X1a)

Although, in reality, any given protein may have multiple true orthologues in another species, GOPHER will only return one per species: the most similar to the query.

Several different databases can be used for the orthology search, depending on the taxonomic group of interest in the input dataset. Details can be found in the help pages for the SLiMPickings options on the website. To be used successfully, GOPHER must be able to recognise the species of the input proteins. For this it uses UniProt species codes and so using UniProt downloads as input is advisable. Details can be found at the website.

(a)



(b)

Seq	Human	Best	Orth
H1b	H1a	-	No
M1a	H1a*	Yes	Yes
R1	H1a	Yes	Yes
M1b	H1a*	No	Yes
C1	H1b	Yes	No
H2	H1a	-	No
M2	H2	No	No
R2	H2	No	No
X1a	H2	Yes	Yes
X1b	H2	No	Yes
S1	H3	Yes	No
H3	H1a	-	No
C2	H3	No	No

Figure 1. Example Orthologue Identification with GOPHER. (a) Example selection of putative orthologous proteins for human protein H1a, circled in blue. Inparalogue H1b is removed. Mouse sequences M1a and M1b are both orthologues of H1a but M1b is arbitrarily removed in favour of the closer M1a. C1 is the closest chicken sequence but is closer to H2. Similarly, sheep sequence S1 is part of the H3 clade. X1a is considered an orthologue, even though H2 is the closest human sequence to it, because H2 is closer to H1a than either is to X1a, while H3 is further from both H1a and X1a. H, human; M, mouse; R, rat; C, chicken; S, sheep; X Xenopus; **(b)** Table summarising relationships. Seq, sequence ID from (a); Human, closest human sequence; Best, whether closest sequence of species to H1a; Orth, whether a true orthologue of H1a.

References

1. Davey, N.E., Shields, D.C. and Edwards, R.J. (2006) *Nucleic Acids Res.*, **34,** 3546-3554.
2. Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B., Costantini, A. *et al.* (2003) *Nucleic Acids Res*, **31,** 3625-3630.
3. Balla, S., Thapar, V., Verma, S., Luong, T., Faghri, T., Huang, C.H., Rajasekaran, S., del Campo, J.J., Shinn, J.H., Mohler, W.A. *et al.* (2006) *Nat Methods.*, **3,** 175-177.
4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J Mol Biol*, **215,** 403-410.
5. Edgar, R.C. (2004) *BMC Bioinformatics*, **5,** 113.
6. Emini, E.A., Hughes, J.V., Perlow, D.S. and Boger, J. (1985) *J Virol.*, **55,** 836-839.
7. Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) *Bioinformatics.*, **21,** 3433-3434.
8. Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) *J Mol Biol*, **179,** 125-142.