UNIVERSITY OF Southampton

CompariMotif: Motif-Motif comparison software

Contents

Figures

Tables

1. Introduction

1.1. Version

1.2. Using this Manual

written in Courier New

written in bold Courier New **red** **dark red**

italicised Courier New [ ]

1.3. Why use CompariMotif?

1.4. Getting Help

help

```
python comparimotif_V3.py help
```

1.4.1. Something Missing?

1.5. Citing CompariMotif

- *Bioinformatics* **24(10):**

1.6. Availability and Local Installation

2. Fundamentals

2.1. Running CompariMotif

2.1.1. The Basics

```
python comparimotif_V3.py motifs=FILENAME
```

```
python comparimotif_V3.py motifs=comparimotif_eg.motifs
```

searchdb=FILENAME

```
python comparimotif_V3.py motifs=file1 searchdb=file2
```

IMPORTANT:
motifs="example file"

2.1.2. Options

2.1.3. Running in Windows

win32=T

2.2. Input

searchdb=FILE motifs=FILE

2.2.1. Motif Input Formats

```
Name Sequence # Comments
```

- A

- [AB] A B [ABC] A B C [^A]
  A
- <R:m:n> m n R R
- X .
- X{m,n} .{m,n} m n
- R{n} n R R
- (AB|CD) AB CD
- (ABC) ABC BAC CAB
- ^
- $
- ? {0,1}

E.g. [IL][^P]X{3}RG

E.g. (2) ^<KR:3:5>P

2.2.2. Motif Splitting

<R:m:n> (AB|CD)

etc. *etc.*

E.g. Motif [IL]{1,2}[^P].(RG|K)

```
Motif_a [IL][^P].RG
Motif_b [IL][^P].K
Motif_c [IL][IL][^P].RG
Motif_d [IL][IL][^P].K
```

2.2.3. Advanced Input Options

reverse=T

reverse=T

2.2.4. SLiM Database Files

Table 2.1. SLiM Databases provided for CompariMotif searches.

ELM	The Eukaryotic Linear Motif database provides a number of high quality annotated SLiMs with known occurrences. **Note.** Some motifs have been split into _a and _b to be compatible with CompariMotif input formats. Such motifs are marked ***Modified*** in their descriptions.	(Puntervoll, Linding et al. 2003)	ELM.motifs
MiniMotif	Another database of SLiMs from all organisms. This has less annotation than ELM but more motifs. These motifs have been reformatted to conform to standard regular expressions.	(Balla, Thapar et al. 2006)	MnM.motifs
Phospho-MotifFinder	Motifs from the PhosphoMotif Finder database of HPRD. Motifs are labelled KIN for Kinase / Phosphatase motifs or BIND for binding motifs. _Y indicates a tyrosine motif, while _ST indicated serine/threonine. The number part of the motif identifier is arbitrary and has no link to the website. **Note**. All these motifs are phosphorylation motifs and, as such, have a key Ser, Thr or Tyr position. These are not given special treatment in CompariMotif and the user should pay special attention to whether the appropriate residue is included in the match.	(Amanchy, Periaswamy et al. 2007)	phosphomotif.motifs
Misc. Literature	Miscellaneous motifs collected from the literature. (These include pubmed links to the relevant paper but we cannot guarantee the accuracy of the motifs or their descriptions.)	See file.	literature.motifs
Combined Literature	A combined database of the above sources. The source database is indicated in the motif description: [ELM] for ELM, [MnM] for MiniMotif and [PMF] for PhosphoMotif Finder.	See above.	combined.motifs
Neduva & Russell	Predicted interaction SLiMs from the high-throughput study of Neduva *et al.* (2005). The motif name indicates what part of the study it is from. All names begin NR, followed by a two-letter code for the species and a one-letter code denoting Domain-level datasets or Protein-level datasets. (See paper for details.) Species codes are: Ce, *C. elegans*; Dm, *D. melanogaster*; Hs, *H. sapiens*; Sc, *S. cerevisiae*.	(Neduva, Linding et al. 2005)	Ned2005_Sig.motifs

2.2.5. Use of DNA Motifs

dna=T/F dna=T

2.2.6. Amino acid frequencies

dna=T

aafreq=FILE FILE

```
AA	FREQ
A	0.074
C	0.033
…
Y	0.033
```

2.3. Output

Table 2.2. Field headings for main CompariMotif output file.

File1	Name of motif file 1 (motifs=FILE). [outstyle=multi only]
File2	Name of file 2 (searchdb=FILE). [outstyle=multi only]
Name1	Name of motif from motif file 1.
Name2	Name of motif from motif file 2.
Motif1	Motif (pattern) from motif file 1.
Motif2	Motif (pattern) from motif file 2.
Sim1	Description of motif1's relationship to motif2.
Sim2	Description of motif2's relationship to motif1.
Match	Regular expression of match between motifs. Upper case positions indicate an exact match, while lower case positions have some degree of degeneracy difference between the two motifs. Mismatches are marked with an asterisk.
MatchPos	Number of matched positions between motif1 and motif2 (>=misshare=X).
MatchIC	Information content of matched positions.
NormIC	MatchIC as a proportion of the maximum possible MatchIC. If this is 1.0 then the match is a good as could possibly be achieved.
CoreIC	MatchIC as a proportion of the maximum possible IC in the matched region only (coreic=T).
Score	Heuristic score (MatchPos x NormIC) for ranking motif matches.
Info1	Information Content of motif1 (if motific=T).
Info2	Information Content of motif2 (if motific=T).
Desc1	Description of motif1 (if motdesc=1 or motdesc=3).
Desc2	Description of motif1 (if motdesc=1 or motdesc=2).

motific=T

motdesc=X

2.3.1. Output Styles

outstyle=multi

outstyle=single

searchdb

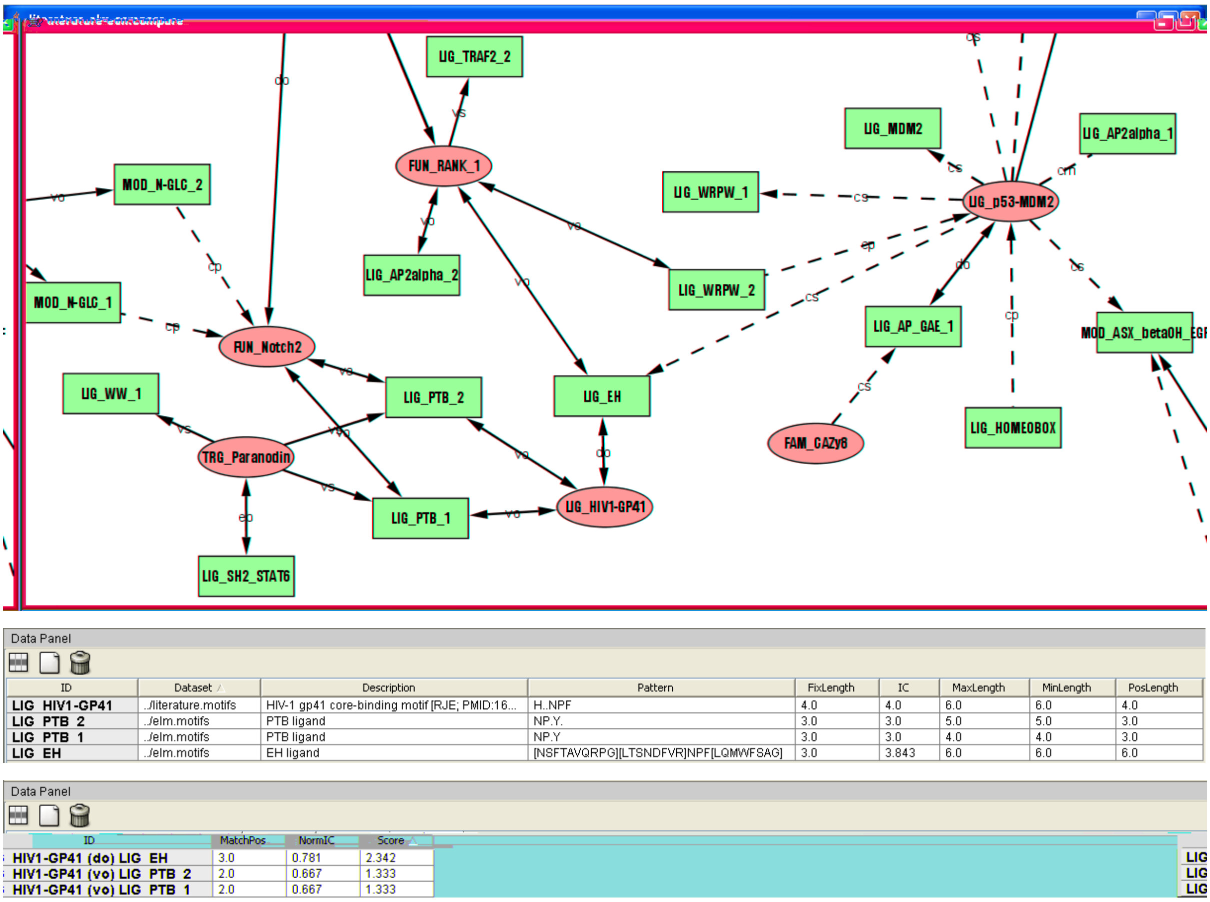
outstyle=normalsplit outstyle=multisplit

[File1,]
Name1, Motif1, Sim1, [Info2,] [Desc1,] [File2,] Name2, Motif2, Sim2, [Info2,] [Desc2,] Match, MatchPos, MatchIC, NormIC, Score

2.3.2. Optional Motif Information File

motinfo=FILE

Motif	The name of the motif.
Pattern	The regular expression pattern of the motif (see **2.2.1Motif Input Formats**).
Description	The description of the motif.
MaxLength	Maximum length of the motif in terms of non-wildcard positions.
MinLength	Minimum length of the motif in terms of non-wildcard positions.
FixLength	Maximum Length of motif in terms of fixed positions.
FullLength	Maximum length of the motif, including wildcard positions.
Expect	The expected number of times the motif will occur in the search database given (searchdb=FILE).
IC	Information Content of motif (if motific=T).



ID	Dataset	Description	Pattern	FixLength	IC	MaxLength	MinLength	PosLength
LIG_HIV1-GP41	../literature.motifs	HIV-1 gp41 core-binding motif [RJE; PMID:16...	H..NPF	4.0	4.0	6.0	6.0	4.0
LIG_PTB_2	../elm.motifs	PTB ligand	NP.Y.	3.0	3.0	5.0	5.0	3.0
LIG_PTB_1	../elm.motifs	PTB ligand	NP.Y	3.0	3.0	4.0	4.0	3.0
LIG_EH	../elm.motifs	EH ligand	[NSFTAVQRPG][LTSNDFVR]NPF[LQMWFSAG]	3.0	3.843	6.0	6.0	6.0

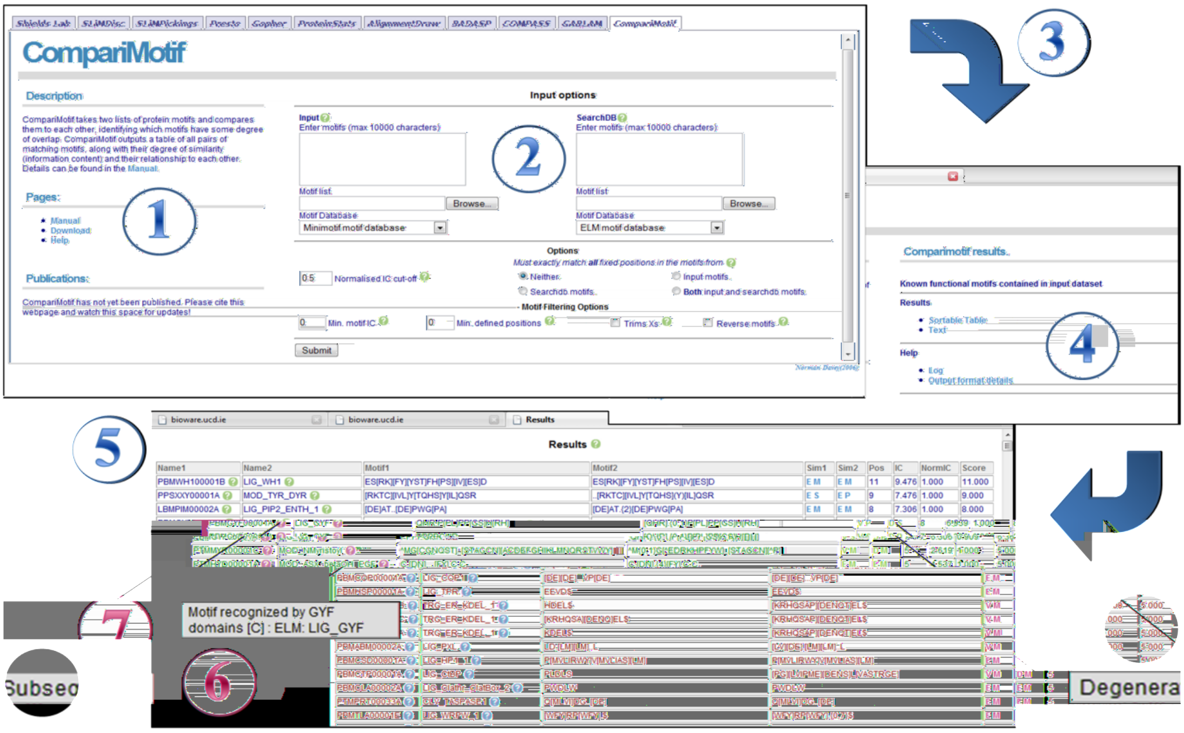
ID	MatchPos	NormIC	Score
HIV1-GP41 (do) LIG_EH	3.0	0.781	2.342
HIV1-GP41 (vo) LIG_PTB_2	2.0	0.667	1.333
HIV1-GP41 (vo) LIG_PTB_1	2.0	0.667	1.333

Figure 2.1. Partial Cytoscape visualisation of CompariMotif relationships.
Motifs from the query and search database file are displayed as blue ellipses and red rectangles, respectively. Edges represent relationships identified by CompariMotif. Arrows indicate the direction of any parent/subsequence relationship. Relationship codes match those output by the webserver. Motif and match details are included in the XGMML file and can be viewed in the Cytoscape Data Panel.

2.3.3. Cytoscape XGMML File

Table 2.3. CompariMotif Commandline Options.

	Basic Input/Output Options	
motifs=FILE		[None]
searchdb=FILE		[None]
dna=T/F		[False]
resfile=FILE		[*]
motinfo=FILE		[None]
motific=T/F		[False]
coreic=T/F		[True]
unmatched=T/F		[False]
	Motif Comparison Parameters	
minshare=X		[2]
normcut=X		[0.5]
matchfix=X		[0]
ambcut=X		[10]
overlaps=T/F		[True]
	Advanced Input Parameters	
minic=X		[2.0]
minfix=X		[0]
minpep=X		[2]
trimx=T/F		[False]
nrmotif=T/F		[False]
reverse=T/F		[False]
mismatches=X		[0]
aafreq=FILE		[None]
	Advanced Input Parameters	
xgmml=T/F		[True]
pickle=T/F		[False]
motdesc=X		[3]
outstyle=X	normal multi single reduced normalsplit multisplit	[normal]



Shields Lab | SLiMDisc | SLiMPickings | Presto | Gopher | ProteinStats | AlignmentDraw | BADASP | COMPASS | GABLAM | CompariMotif
CompariMotif
1
2
3
4
5
6
7
Norman Davey(2006)
bioware.ucd.ie
Results
Subseq
Degenera

Description

CompariMotif takes two lists of protein motifs and compares them to each other, identifying which motifs have some degree of overlap. CompariMotif outputs a table of all pairs of matching motifs, along with their degree of similarity (information content) and their relationship to each other. Details can be found in the Manual.

Pages:

- Manual
- Download
- Help

Publications:

CompariMotif has not yet been published. Please cite this webpage and watch this space for updates!

Input options

Input
Enter motifs (max 10000 characters)
Motif list [ ] Browse...
Motif Database: Minimotif motif database

SearchDB
Enter motifs (max 10000 characters)
Motif list [ ] Browse...
Motif Database: ELM motif database

0.5 Normalised IC cut-off

Options
*Must exactly match **all** fixed positions in the motifs from*
● Neither ○ Input motifs
○ Searchdb motifs ○ **Both** input and searchdb motifs

- Motif Filtering Options
0 Min. motif IC   0 Min. defined positions   □ Trims Xs   □ Reverse motifs

Submit

Comparimotif results

Known functional motifs contained in input dataset

Results

- Sortable Table
- Text

Help

- Log
- Output format details

Results

Name1	Name2	Motif1	Motif2	Sim1	Sim2	Pos	IC	NormIC	Score
PBMWH100001B	LIG_WH1	ES[RK][FY][YST]FH[PS][IV][ES]D	ES[RK][FY][YST]FH[PS][IV][ES]D	E M	E M	11	9.476	1.000	11.000
PPSXXY00001A	MOD_TYR_DYR	[RKTC][IVL]Y[TQHS]Y[IL]QSR	..[RKTC][IVL]Y[TQHS](Y)[IL]QSR	E S	E P	9	7.476	1.000	9.000
LBMPIM00002A	LIG_PIP2_ENTH_1	[DE]AT..[DE]PWG[PA]	[DE]AT.{2}[DE]PWG[PA]	E M	E M	8	7.306	1.000	8.000

Motif recognized by GYF domains [C] : ELM: LIG_GYF

Figure 2.2. CompariMotif webserver.
1. Links to help pages and further information. 2. The main input form. Users can enter motifs directly into the text boxes, upload motif files, or use databases provided. In this example, Minimotif is being compared to ELM. Further options can be set to customise search parameters. The help buttons by each opti
the search progress. Upon completion, the CompariMotif results page (4) will open. 4. The webserver returns results as a simple unformatted text and also a sortable table (5). In addition, the program log is returned, which is useful for identifying formatting errors etc. 5. The main results table opens in a

with high Scores tend to be of better quality. 6. Mousing over the ? by each motif name will reveal the motif description (dependent on input format). 7. Match similarities are reduced to a code, which is expanded by positioning the mouse over them.

3. Motif Comparisons

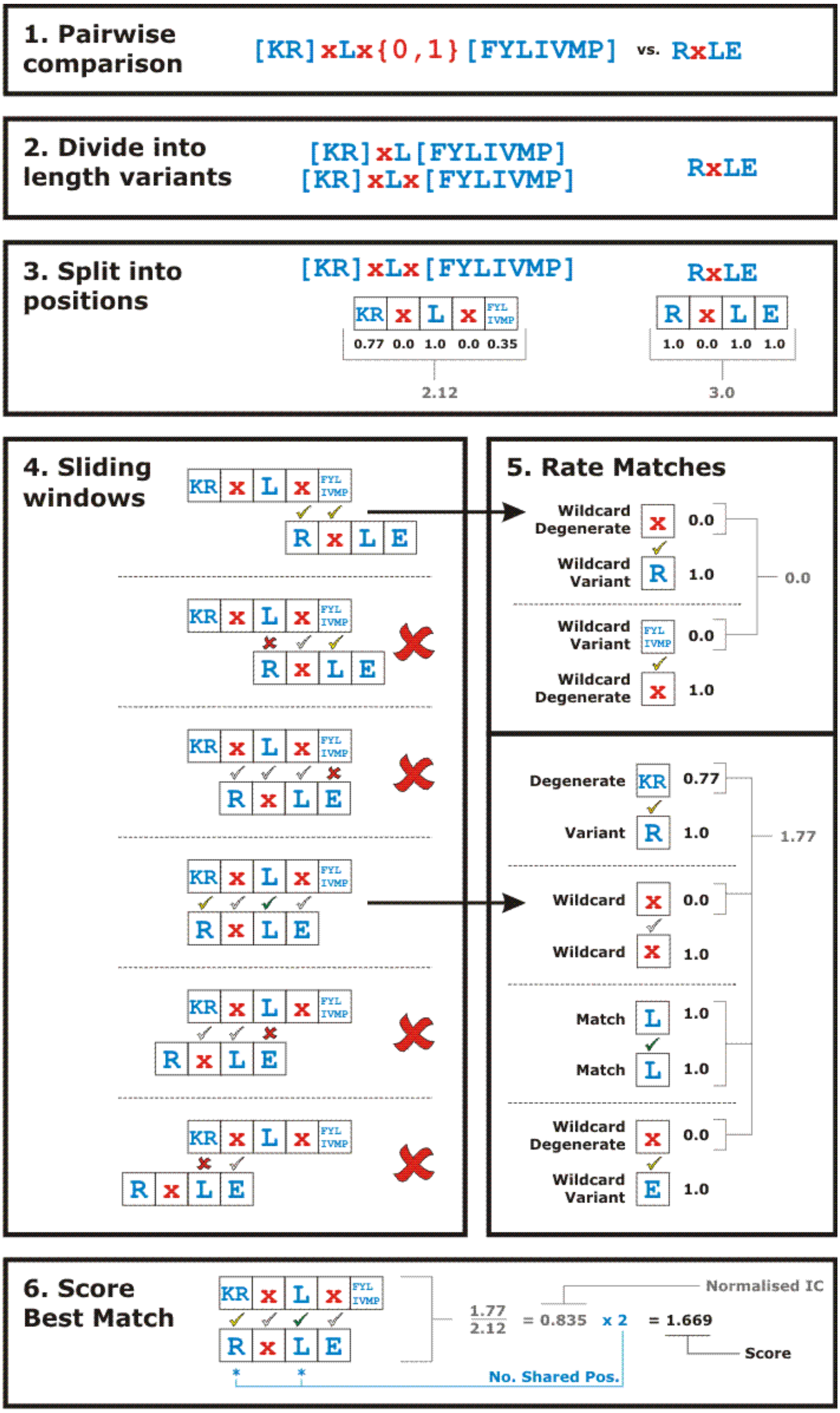


Figure 3.1. Overview of CompariMotif.
Details can be found in the text.

3.1. How CompariMotif Works

minshare=X normcut=X matchfix=X

matchfix=X

3.1.1. Single Position Comparisons

-
-
-
-
-
-
-
-
-
-
-
-
-

E.g.

3.1.2. Selecting Pairwise Matches

-
-
-
- *e.g.*
- mismatches=X
- minshare=X
- matchfix=X

- normcut=X

3.1.3. Defining Motif Relationships

-
    - **Exact**
    - **Variant**
    - **Degenerate**
    - **Complex**
    - **Ugly =**
      overlaps=F
-
    - **Match**
    - **Parent**
    - **Subsequence**
    - **Overlap**

3.2. Information Content

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

IC_i i f_a

i N *i.e.* N

N

f_a N f_a N

lowest

IC_m IC_i

IC_m

normcut=X

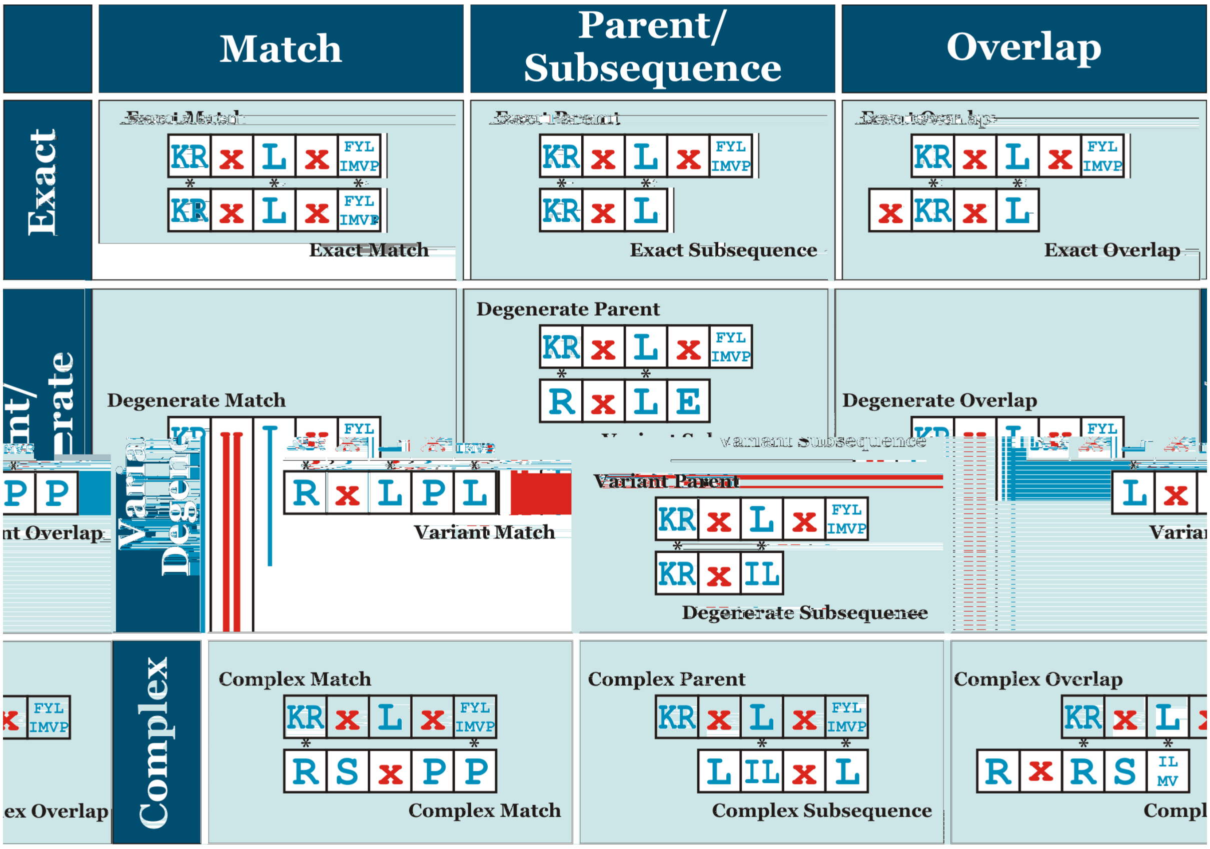


Figure 3.2. CompariMotif Match Type Examples.

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. **Exact Match:** All positions are identical and the match spans the full length of both motifs; **Variant/Degenerate Match:** The match spans the full length of both motifs. All of the positions of the query are either the same as the match (X v. X and L v. L) or more degenerate ([KR] v. R, X v. P & [FYLIMVP] v. L) and so it is classed degenerate. Likewise, all positions in the other motif, RxLPL, are either identical to the query or variants of the query positions, so it is classed as variant; **Complex Match:** Again, the match spans the full length of both motifs. This time, each motif has some positions that are more degenerate than in the other motif. *i.e.* The query is a variant for the L v. X position but more degenerate at all other positions; **Exact Parent/Subsequence:** The [KR]xL motif is entirely and exactly contained within the query; **Degenerate Parent/Variant Subsequence:** The RxLE motif is entirely contained within the query. At two positions, however, ([KR] v. R & X v. E) the quer

Variant Parent/Degenerate Subsequence: This time it is the query that is the variant in one position (L v. IL) and so the variant/degenerate labels are swapped; **Complex parent/subsequence:** the L[IL]xL motif is less degenerate at two positions (X v. L

Exact overlap: Neither motif is entirely contained within the other but the positions overlapping match exactly; **Degenerate/variant overlap:** Neither motif is entirely contained within the other. The first P of LxPP is a variant of the [FYLIMVP] in the query, while the other two matches (an L and an X) are exact, therefore th

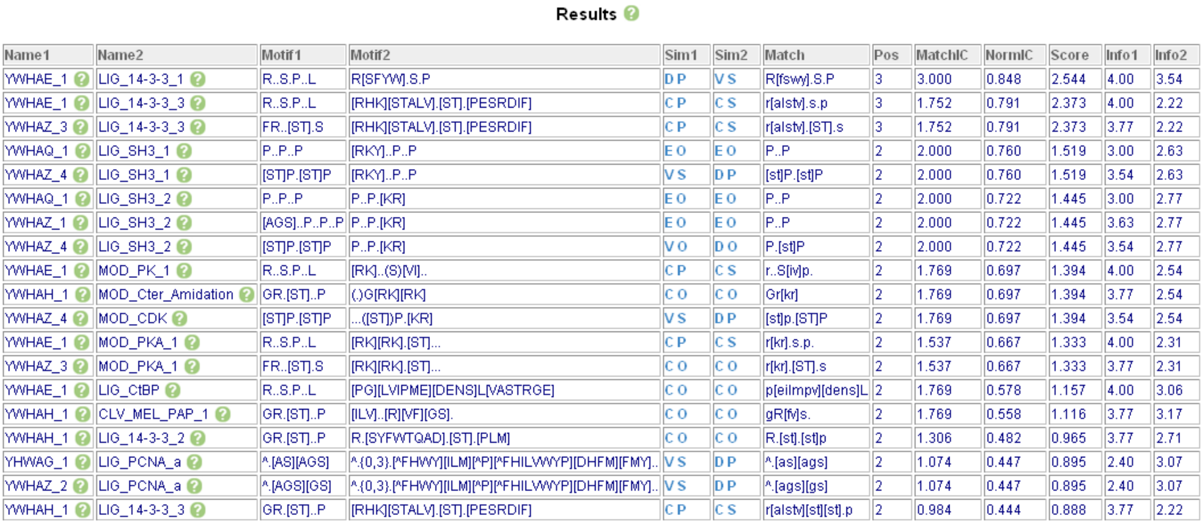
Complex overlap: Neither motif is entirely contained within the other and both contain positions that are degenerate when compared to the matching position in the other motif ([KR] v. R & X v. S are degenerate in the query, L v. [ILMV] is degenerate in RxRS[ILMV]). **"Ugly"** matches are not show. These are like **degenerate/variant** or **complex** pairings except they feature 1+ positions where ambiguities partially overlap (see text).

3.3. Score

3.4. Example Application

p

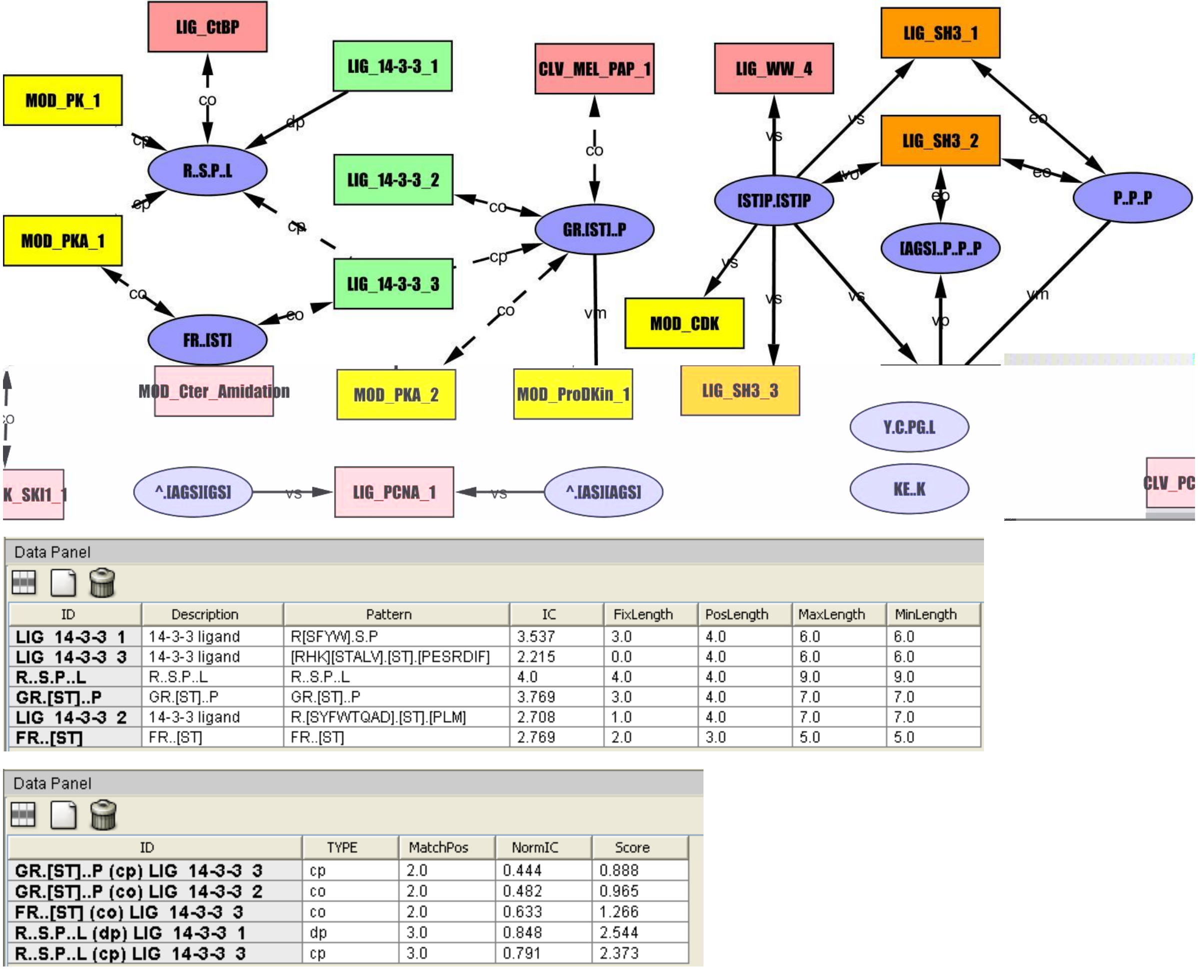
```
YWHAE_1  R..S.P..L      # Sig. SLiMFinder motif for HPRD 14-3-3 Epsilon interactors
YWHAH_1  GR.[ST]..P     # Sig. SLiMFinder motif for HPRD 14-3-3 Eta interactors
YHWAG_1  ^.[AS][AGS]    # Sig. SLiMFinder motif for HPRD 14-3-3 Gamma interactors
YHWAG_2  KE..K          # Sig. SLiMFinder motif for HPRD 14-3-3 Gamma interactors
YWHAQ_1  P..P..P        # Sig. SLiMFinder motif for HPRD 14-3-3 Theta interactors
YWHAZ_1  [AGS]..P..P..P # Sig. SLiMFinder motif for HPRD 14-3-3 Zeta interactors
YWHAZ_2  ^.[AGS][GS]    # Sig. SLiMFinder motif for HPRD 14-3-3 Zeta interactors
YWHAZ_3  FR..[ST].S     # Sig. SLiMFinder motif for HPRD 14-3-3 Zeta interactors
YWHAZ_4  [ST]P.[ST]P    # Sig. SLiMFinder motif for HPRD 14-3-3 Zeta interactors
YWHAZ_5  Y.C.PG.L       # Sig. SLiMFinder motif for HPRD 14-3-3 Zeta interactors
```



Results

Name1	Name2	Motif1	Motif2	Sim1	Sim2	Match	Pos	MatchIC	NormIC	Score	Info1	Info2
YWHAE_1	LIG_14-3-3_1	R..S.P..L	R[SFYW].S.P	D P	V S	R[fswy].S.P	3	3.000	0.848	2.544	4.00	3.54
YWHAE_1	LIG_14-3-3_3	R..S.P..L	[RHK][STALV].[ST].[PESRDIF]	C P	C S	r[alstv].s.p	3	1.752	0.791	2.373	4.00	2.22
YWHAZ_3	LIG_14-3-3_3	FR..[ST].S	[RHK][STALV].[ST].[PESRDIF]	C P	C S	r[alstv].[ST].s	3	1.752	0.791	2.373	3.77	2.22
YWHAQ_1	LIG_SH3_1	P..P..P	[RKY]..P..P	E O	E O	P..P	2	2.000	0.760	1.519	3.00	2.63
YWHAZ_4	LIG_SH3_1	[ST]P.[ST]P	[RKY]..P..P	V S	D P	[st]P.[st]P	2	2.000	0.760	1.519	3.54	2.63
YWHAQ_1	LIG_SH3_2	P..P..P	P..P.[KR]	E O	E O	P..P	2	2.000	0.722	1.445	3.00	2.77
YWHAZ_1	LIG_SH3_2	[AGS]..P..P..P	P..P.[KR]	E O	E O	P..P	2	2.000	0.722	1.445	3.63	2.77
YWHAZ_4	LIG_SH3_2	[ST]P.[ST]P	P..P.[KR]	V O	D O	P.[st]P	2	2.000	0.722	1.445	3.54	2.77
YWHAE_1	MOD_PK_1	R..S.P..L	[RK]..(S)[VI]..	C P	C S	r..S[iv]p.	2	1.769	0.697	1.394	4.00	2.54
YWHAH_1	MOD_Cter_Amidation	GR.[ST]..P	(.)G[RK][RK]	C O	C O	Gr[kr]	2	1.769	0.697	1.394	3.77	2.54
YWHAZ_4	MOD_CDK	[ST]P.[ST]P	...([ST])P.[KR]	V S	D P	[st]p.[ST]P	2	1.769	0.697	1.394	3.54	2.54
YWHAE_1	MOD_PKA_1	R..S.P..L	[RK][RK].[ST]...	C P	C S	r[kr].s.p.	2	1.537	0.667	1.333	4.00	2.31
YWHAZ_3	MOD_PKA_1	FR..[ST].S	[RK][RK].[ST]...	C O	C O	r[kr].[ST].s	2	1.537	0.667	1.333	3.77	2.31
YWHAE_1	LIG_CtBP	R..S.P..L	[PG][LVIPME][DENS]L[VASTRGE]	C O	C O	p[eilmpv][dens]L	2	1.769	0.578	1.157	4.00	3.06
YWHAH_1	CLV_MEL_PAP_1	GR.[ST]..P	[ILV]..[R][VF][GS].	C O	C O	gR[fv]s.	2	1.769	0.558	1.116	3.77	3.17
YWHAH_1	LIG_14-3-3_2	GR.[ST]..P	R.[SYFWTQAD].[ST].[PLM]	C O	C O	R.[st].[st]p	2	1.306	0.482	0.965	3.77	2.71
YHWAG_1	LIG_PCNA_a	^.[AS][AGS]	^.{0,3}.[^FHWY][ILM][^P][^FHILVWYP][DHFM][FMY]..	V S	D P	^.[as][ags]	2	1.074	0.447	0.895	2.40	3.07
YWHAZ_2	LIG_PCNA_a	^.[AGS][GS]	^.{0,3}.[^FHWY][ILM][^P][^FHILVWYP][DHFM][FMY]..	V S	D P	^.[ags][gs]	2	1.074	0.447	0.895	2.40	3.07
YWHAH_1	LIG_14-3-3_3	GR.[ST]..P	[RHK][STALV].[ST].[PESRDIF]	C P	C S	r[alstv][st][st].p	2	0.984	0.444	0.888	3.77	2.22

Figure 3.3. CompariMotif webserver output for 14-3-3 HPRD SLimFinder analysis vs. ELM.



Data Panel

ID	Description	Pattern	IC	FixLength	PosLength	MaxLength	MinLength
LIG 14-3-3 1	14-3-3 ligand	R[SFYW].S.P	3.537	3.0	4.0	6.0	6.0
LIG 14-3-3 3	14-3-3 ligand	[RHK][STALV].[ST].[PESRDIF]	2.215	0.0	4.0	6.0	6.0
R..S.P..L	R..S.P..L	R..S.P..L	4.0	4.0	4.0	9.0	9.0
GR.[ST]..P	GR.[ST]..P	GR.[ST]..P	3.769	3.0	4.0	7.0	7.0
LIG 14-3-3 2	14-3-3 ligand	R.[SYFWTQAD].[ST].[PLM]	2.708	1.0	4.0	7.0	7.0
FR..[ST]	FR..[ST]	FR..[ST]	2.769	2.0	3.0	5.0	5.0

Data Panel

ID	TYPE	MatchPos	NormIC	Score
GR.[ST]..P (cp) LIG 14-3-3 3	cp	2.0	0.444	0.888
GR.[ST]..P (co) LIG 14-3-3 2	co	2.0	0.482	0.965
FR..[ST] (co) LIG 14-3-3 3	co	2.0	0.633	1.266
R..S.P..L (dp) LIG 14-3-3 1	dp	3.0	0.848	2.544
R..S.P..L (cp) LIG 14-3-3 3	cp	3.0	0.791	2.373

Figure 3.4. CompariMotif XGMML output visualized with Cytoscape(Shannon, Markiel et al. 2003) (recoloured).

Motifs returned by SLiMFinder analysis of 14-3-3 interaction datasets are shown as blue ellipses. ELMs with CompariMotif matches are shown as rectangles. These are pale red by default but the following groups have been manually recoloured: 14-3-3 ligands, green; SH3 ligands, orange, phosphorylation motifs, yellow. Arrows proceed from parent to subsequence motifs (bidirectional where equal); Data Panel details for 14-3-3 ELMs and connected nodes and edges are shown.

4. Appendices

4.1. Troubleshooting & FAQ

4.2. SLiM Definitions

- *Short*
- *Linear*
- *Motifs*

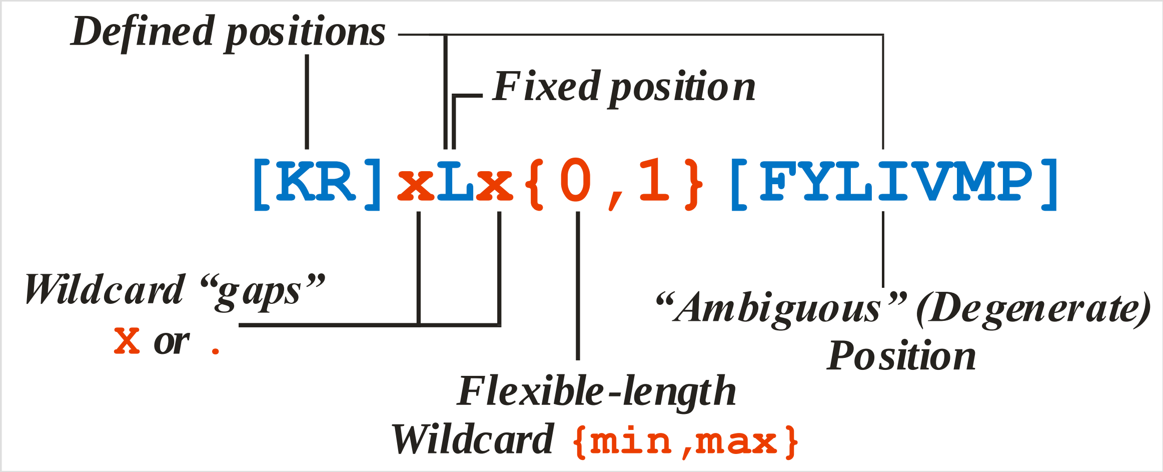


Figure 4.1. Anatomy of a SLiM.
Definitions of different properties of SLiM have been marked on the example ELM, LIG_CYCLIN_1 (Puntervoll, Linding et al. 2003). This motif has three defined positions (one fixed and two degenerate) and two wildcard spacers (one fixed, one flexible-length) for a total length of 4-5aa.

4.3. References

Nat Biotechnol. **25(3)**

Nat Methods. **3(3)**

Nucleic Acids Res **35(Web Server issue)**

Nucleic Acids Res. **34(12)**

Nucleic Acids Res **40(Database issue)**

Mol Biosyst **8(1)**

PLoS ONE **2(10)**

Bioinformatics **24(10)**

Nucleic Acids Res. **34(Database issue)**

PLoS Biol. **3(12)**

FEBS Lett. **579(15)**

Nucleic Acids Res. **34(Web Server issue)**

Nucleic Acids Res **31(13)**

Nucleic Acids Res. **31(13)**

MD. Comput. **14**

Genome Res **13(11)**