SLiMEnrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions

Sobia Idrees [1] , **Åsa Pérez-Bercoff** [1] , **Richard J Edwards** Corresp. [1]

[1] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

Corresponding Author: Richard J Edwards
Email address: richard.edwards@unsw.edu.au

Many important cellular processes involve protein-protein interactions (PPIs) mediated by a Short Linear Motif (SLiM) in one protein interacting with a globular domain in another. Despite their significance, these domain-motif interactions (DMIs) are typically low affinity, which makes them challenging to identify by classical experimental approaches, such as affinity pulldown mass spectrometry (AP-MS) and yeast two-hybrid (Y2H). DMIs are generally underrepresented in PPI networks as a result. A number of computational methods now exist to predict SLiMs and/or DMIs from experimental interaction data but it is yet to be established how effective different PPI detection methods are for capturing these low affinity SLiM-mediated interactions. Here, we introduce a new computational pipeline (SLiMEnrich) to assess how well a given source of PPI data captures DMIs and thus, by inference, how useful that data should be for SLiM discovery. SLiMEnrich interrogates a PPI network for pairs of interacting proteins in which the first protein is known or predicted to interact with the second protein via a DMI. Permutation tests compare the number of known/predicted DMIs to the expected distribution if the two sets of proteins are randomly associated. This provides an estimate of DMI enrichment within the data and the false positive rate for individual DMIs. As a case study, we detect significant DMI enrichment in a high-throughput Y2H human PPI study. SLiMEnrich analysis supports Y2H data as a source of DMIs and highlights the high false positive rates associated with naïve DMI prediction. SLiMEnrich is available as an R Shiny app. The code is open source and available via a GNU GPL v3 license at: https://github.com/slimsuite/SLiMEnrich . A web server implementation is available at: http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/ .

1   **SLiMEnrich: computational assessment of protein-protein interaction data as a**

2   **source of domain-motif interactions**

3   Sobia Idrees[1], Åsa Pérez-Bercoff[1] and Richard J Edwards[1]

4   [1]School of Biotechnology and Biomolecular Sciences, University of New South Wales,

5   Sydney, Australia.

7   Corresponding author: Richard J Edwards[1]

9   Email address: richard.edwards@unsw.edu.au

11 **Abstract**

12 Many important cellular processes involve protein-protein interactions (PPIs) mediated by a

13 Short Linear Motif (SLiM) in one protein interacting with a globular domain in another. Despite

14 their significance, these domain-motif interactions (DMIs) are typically low affinity, which

15 makes them challenging to identify by classical experimental approaches, such as affinity

16 pulldown mass spectrometry (AP-MS) and yeast two-hybrid (Y2H). DMIs are generally

17 underrepresented in PPI networks as a result. A number of computational methods now exist to

18 predict SLiMs and/or DMIs from experimental interaction data but it is yet to be established how

19 effective different PPI detection methods are for capturing these low affinity SLiM-mediated

20 interactions. Here, we introduce a new computational pipeline (SLiMEnrich) to assess how well

21 a given source of PPI data captures DMIs and thus, by inference, how useful that data should be

22 for SLiM discovery. SLiMEnrich interrogates a PPI network for pairs of interacting proteins in

23 which the first protein is known or predicted to interact with the second protein via a DMI.

24 Permutation tests compare the number of known/predicted DMIs to the expected distribution if

25 the two sets of proteins are randomly associated. This provides an estimate of DMI enrichment

26 within the data and the false positive rate for individual DMIs. As a case study, we detect

27 significant DMI enrichment in a high-throughput Y2H human PPI study. SLiMEnrich analysis

28 supports Y2H data as a source of DMIs and highlights the high false positive rates associated

29 with naïve DMI prediction. SLiMEnrich is available as an R Shiny app. The code is open source

30 and available via a GNU GPL v3 license at: https://github.com/slimsuite/SLiMEnrich. A web

31 server implementation is available at: http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/.

32

Introduction

34 Proteins interact with their partners through two main classes of functional modules: globular

35 domains and Short Linear Motifs (SLiMs) (Bhattacharyya et al. 2006). SLiMs are short protein

36 regions (typically 3-10 amino acids long) with a small number of key residues that mediate

37 domain-motif interactions (DMIs) with the globular domain of a protein-protein interaction (PPI)

38 partner (Davey et al. 2012). These DMIs underpin critical cellular functions, including cell cycle

39 regulation, cell compartment targeting, post-translational modification, protein degradation, and

40 signal transduction (Van Roey et al. 2014). Knowledge of DMIs can provide molecular details of

41 cellular processes and thus it is important to discover SLiMs and link them to their domain

42 partners (Davey et al. 2012; Neduva & Russell 2005). Despite this, only a small fraction of the

43 likely range of SLiMs, and the DMIs they mediate, have been identified (Tompa et al. 2014) and

44 curated in resources such as the Eukaryotic Linear Motif (ELM) resource (Gouw et al. 2018),

45 Linear Motif mediated Protein Interaction Database (LMPID) (Sarkar et al. 2015), interActions

46 of moDular domAiNs (ADAN) (Encinar et al. 2009), and the database of three-dimensional

47 interacting domains (3did) (Mosca et al. 2014). SLiM-mediated interactions are typically low

48 affinity (Davey et al. 2012) and are thus vulnerable to being overlooked by classical PPI

49 detection methods, such as affinity pulldown mass spectrometry (AP-MS) and yeast two-hybrid

50 (Y2H), where high stringencies are typically employed to reduce false positive interactions.

51 Early analyses of high throughput data revealed that known SLiM-mediated interactions account

52 for less than 1% of interactions (Neduva & Russell 2006). This was used as evidence that many

53 more SLiMs and DMI are yet to be discovered, but also raises concerns that these methods are

54 depleted for DMIs.

55 A range of computational tools now exist for the two main tasks in SLiM prediction: (1)

56 identifying functional instances of known motifs, and (2) *de novo* prediction of new SLiM

57 classes (Edwards & Palopoli 2015). In principle, the task of interrogating a protein sequence for

58 known motif patterns is quite simple. Motif definitions are available from ELM (Gouw et al.

59 2018; Gouw et al. 2017) and PROSITE (Hulo et al. 2006), and various tools exist for searching

60 proteins for these patterns or resource-specific motif definitions (Edwards & Palopoli 2015).

61 Other tools, like Minimotif Miner (MnM) (Lyon et al. 2018), will search sequences for similarity

62 to known SLiMs or post-translational modifications (PTMs), but do not make motif definitions

63   or tools available for proteome-scale searches. The short and degenerate nature of most SLiMs

64   hampers the usefulness of predictions due to the high possibility of false positive results. This is

65   particularly true for SLiMs with very few known occurrences, which will lack the data required

66   for detailed modelling. It is therefore important to improve the specificity of predictions by

67   incorporating contextual information such as evolutionary conservation and/or protein structure

68   (Krystkowiak & Davey 2017; Mi et al. 2012), or knowledge of interaction partners containing

69   relevant SLiM recognition domains (*e.g.* (Encinar et al. 2009; Kelil et al. 2016; Luck et al. 2011;

70   Weatheritt et al. 2012).

71   The *de novo* prediction of SLiMs is inherently more challenging and relies on assembling sets of

72   proteins that share a SLiM. The most widespread approach is to mine PPI data to identify sets of

73   proteins that interact with a common partner (*e.g.* (Edwards et al. 2012; Lieber et al. 2010;

74   Neduva et al. 2005). The success of prediction methods is highly dependent on the signal to

75   noise ratio in these data, in terms of the proportion of proteins likely to contain the SLiM

76   (Edwards et al. 2012; Edwards & Palopoli 2015). Before attempting SLiM discovery, it is

77   therefore useful to know how well the input PPI data is capturing SLiM-mediated interactions.

78   Different experimental parameters will influence how depleted the recovered interactions are for

79   DMIs, and so this assessment is also useful for experimentalists when establishing an appropriate

80   stringency threshold.

81   Here, we introduce a new computational pipeline (SLiMEnrich) that assesses how well PPI data

82   are capturing DMIs and thus, by inference, how useful that data should be for SLiM discovery.

83   SLiMEnrich evaluates DMI enrichment through permutation tests and reports the probability of

84   randomly recovering as many interacting domain-motif pairs as are found in the real PPI data.

85   SLiMEnrich can use known SLiM-mediated interactions for high stringency analysis, or

86   incorporate DMI predictions by using SLiM predictions and/or known SLiM-domain interactions

87   to expand the number of plausible DMIs in the data. Identified/predicted DMIs are returned,

88   along with an estimated false discovery rate based on the mean number of random DMIs

89   generated from the data. Whilst not their primary purpose, SLiMEnrich metrics can also be used

90   to assess SLiM and/or DMI prediction strategies when applied to PPI data that is already known

91   to contain DMIs. SLiMEnrich is therefore of potential use for both DMI prediction and

92   assessment of PPI data. SLiMEnrich has been developed in R and implemented in Shiny to

93   provide easy, user-friendly operation.

Materials and Methods

Algorithm

96 An overview of the SLiMEnrich pipeline is shown in Figure 1. SLiMEnrich uses (known or

97 predicted) SLiM occurrences, domain composition, and known SLiM interactions at the protein

98 or domain level. These are combined to predict SLiM-mediated DMIs within pairwise PPI data

99 supplied by the user. Input data is combined by matching protein, SLiM and Domain IDs from

100 the input data, providing a flexible framework for analysis. PPI data is treated asymmetrically,

101 with specified sets of putative motif- and domain-containing proteins, known as "mProteins" and

102 "dProteins", respectively. First, SLiMEnrich identifies all possible known/predicted DMI links

103 between mProteins and dProteins in the PPI data (Figure 1). DMI mapping can be performed

104 using a number of different strategies depending on the desired balance of quality versus quantity

105 of DMI (Figure 1, Figure 2). At one extreme, analysis can be restricted to mProtein-dProtein

106 pairs known to interact via a DMI (Figure 2, top left). At the other extreme, mProteins with

107 predicted SLiMs can be linked to any dProteins containing a domain known to interact with that

108 SLiM (Figure 2, bottom right). This set of "potential DMIs" represents the overall pool of

109 possible DMIs given the input data and mapping strategy.

110 Next, SLiMEnrich extracts "predicted DMIs" by identifying the subset of potential DMIs that are

111 found in the PPI data, *e.g.* observed PPI pairs where the mProtein is (known or) predicted to

112 interact with the dProtein according to the DMI strategy employed. Finally, SLiMEnrich

113 estimates how well the PPI data is capturing DMIs by comparing the observed DMI predictions

114 to a background distribution of expected DMIs when proteins are randomly assigned interaction

115 partners. For this, the input PPI data is shuffled to generate 1000 random PPI datasets where each

116 protein maintains the same number of interacting partners but the connections are randomly

117 assigned. This is performed by first reducing PPI data to asymmetrical non-redundant protein

118 pairs and then randomly shuffling the dProtein column whilst avoiding the introduction of

119 redundant random PPI pairs. The random PPI datasets are then mapped onto the potential DMIs

120 in the same fashion as the real data. Enrichment is calculated as an empirical *P*-value

121 corresponding to the probability of seeing at least as many DMIs in random PPI data (Figure 3).

122 A False Discovery Rate (FDR) for individual DMIs is also estimated as the proportion of the

123   predicted DMIs explained on average by random associations, using the mean random DMI

124   distribution capped at the observed value.

125   **Requirements and Implementation**

126   **Inputs.** SLiMEnrich requires a delimited pairwise PPI file as input. By default, known ELM

127   instances (ELMi) (Gouw et al. 2018) will be used to define the motif composition of mProteins.

128   This file can be replaced by a SLiM prediction file (generated by *e.g.* SLiMProb (Edwards &

129   Palopoli 2015)), which has predicted SLiMs for the mProteins in the PPI file. DMIs can be

130   predicted by one of three strategies (Figure 2). By default, the DMI file links ELM classes

131   (ELMc) directly to dProteins using known ELM binding partners (Gouw et al. 2018). For more

132   stringent analysis, these binding partners can be linked directly to specific ELM-containing

133   proteins, in which case the DMI file links mProteins and dProteins, and the motif occurrence file

134   is ignored (Figure 1). For more relaxed/flexible analysis, the DMI file will link motifs to binding

135   domains, which are then linked to dProteins via a domain composition file. By default,

136   SLiMEnrich uses Pfam domains (Finn et al. 2016) for reviewed human Uniprot proteins

137   (UniProt Consortium 2017) and links them to ELM-binding domains (Gouw et al. 2018). If

138   alternative data sources are used, users should also provide a file of protein-domain links for the

139   dProteins in the PPI file, and/or a motif-domain file that defines the known domain-motif

140   interactions. Note that this can be used to interrogate PPI data for enrichment of any interaction

141   type. For example, two protein-domain files could be linked through known domain-domain

142   interactions. Alternatively, the ELMi-Protein DMI strategy enables the enrichment analysis of

143   any set of PPIs, allowing SLiMEnrich to examine overlaps between PPI datasets. Default fields

144   for user files ("mProtein", "dProtein", "Motif", "Domain") are show in Figure 1, and can be set

145   to custom values in the SLiMEnrich App.

146   **Example data.** SLiMEnrich comes with example data of Adenoviridae proteins and their human

147   interactors downloaded from the PHISTO database (2017-07-26) (Durmus Tekir et al. 2013).

148   ELM (downloaded 2018-07-17) (Dinkel et al. 2016) regular expression matches in the viral

149   proteins were predicted using SLiMProb v2.5.0 (Edwards & Palopoli 2015) with disorder

150   masking. A table of ELM-binding Pfam domains was downloaded from ELM (2018-07-17)

151   (Dinkel et al. 2016). Pfam domains for human proteins were extracted from Uniprot

152   (downloaded 2017-03-08) (UniProt Consortium 2017).

153   **Outputs.** The primary output of SLiMEnrich is the observed number of known/predicted DMIs

154   compared to the distribution from the randomised PPI data (Figure 3). SLiMEnrich also provides

155   tables of both "potential DMIs" and "predicted DMIs" (Figure 1, see Algorithm for details),

156   summary plots of predicted DMI numbers and an interactive DMI network (Figure 4). Together,

157   these enable the user to explore the data for proteins, SLiMs and/or domains that might be

158   biasing results. This can be seen with the example Adenoviridae analysis, where the Pkinase

159   domain (PF00069) mediates a large proportion of the predicted DMIs via multiple modification

160   ELMs (Figure 4), which will inflate the probability of DMIs in the random PPI data. Tables can

161   be downloaded as comma-separated text files. The summary plots, enrichment histogram and

162   DMI network can be downloaded as PNG files.

163   **Implementation.** SLiMEnrich is a standalone application written entirely in R. It is platform

164   independent and can be launched locally from any R environment (e.g. RStudio). SLiMEnrich

165   takes advantage of the reactive programming feature of Shiny to cache computational steps to

166   avoid unnecessary computing during an interactive session. The code is open source and

167   available via a GNU GPL v3 license at: https://github.com/slimsuite/SLiMEnrich. SLiMEnrich is

168   also implemented as a Shiny webserver at: http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/.

169   Additional details can be found at: https://github.com/slimsuite/SLiMEnrich/wiki.

170   **Case study: Domain-motif resolved yeast-two-hybrid human interactome**

171   Pairwise human PPIs were extracted from a high-throughput human Y2H study that detected

172   ~14,000 binary interactions (Rolland et al. 2014) and converted into a non-redundant,

173   symmetrical PPI dataset of 26,166 mProtein-dProtein PPIs (*i.e.* with each PPI pair present as P1-

174   P2 *and* P2-P1), restricted to reviewed Uniprot proteins. Protein sequences were downloaded

175   from Uniprot (2017-03-01). A list of ELMs and their domain partners was retrieved from the

176   ELM database (2018-07-17) (Dinkel et al. 2016). ELM occurrences in the human proteins were

177   predicted by SLiMProb v2.5.0 (Edwards & Palopoli 2015) with disorder masking (IUPred

178   (Dosztanyi et al. 2005), cut-off 0.2 (Edwards et al. 2007)) to restrict analysis to low stringency

179   predicted disordered regions. Pfam domains were parsed from Uniprot entries using SLiMBench

180   (Palopoli et al. 2015). Splice isoforms for all data were mapped onto their parent Uniprot

181   identifier. SLiMEnrich was used to map known and predicted DMIs onto the Y2H dataset using

182   five strategies of decreasing stringency: (1) known ELM PPIs only, (2) known ELM instances

183   mapped onto proteins known to interact with the ELM class, (3) known ELM instances mapped

184   onto Pfam domains known to interact with the ELM class, (4) SLiMProb predictions mapped

185   onto proteins known to interact with the ELM class, (5) SLiMProb predictions mapped onto

186   Pfam domains known to interact with the ELM class (Figure 2).

Simulation of poor quality SLiM predictions

188   SLiMEnrich is not a DMI prediction tool *per se* and should not require completely accurate

189   SLiM occurrence data to identify enrichment indicative of PPI data that captures DMIs. To

190   investigate the impact of noisy SLiM prediction data, we replaced increasing proportions (25%,

191   50%, 75% and 100%) of the known ELM instances (2018-07-17) (Dinkel et al. 2016) with

192   random occurrences and repeated analysis of the Y2H interactome case study. This was

193   performed by replacing different proportions of the ELM proteins (*i.e.* proteins containing a

194   known ELM) with a protein randomly selected from reviewed human Uniprot proteins (UniProt

195   Consortium 2017). For direct comparison, the distribution of normalised predicted DMIs, D, was

196   calculated as follows:

$$D = \frac{O - R}{\overline{R}},$$

198   where O is the observed predicted DMI count, R is the distribution of random predicted DMIs,

199   and $\overline{R}$ is the mean random predicted DMI count.

Results

Case study: Domain-motif resolved yeast-two-hybrid human interactome

202   SLiMEnrich analysis revealed the case study Y2H data to be enriched for DMIs under all DMI

203   prediction strategies (Figure 5, Table 1). Restricting analysis to known DMIs identified fourteen

204   in the Y2H data, which represented a more than 100-fold enrichment over the random

205   expectation (mean 0.122). Including DMIs where a dProtein was known to interact with the

206   ELM class (Figure 2, centre column), almost doubled the number of predicted DMIs but with

207   nearer six times more random DMIs on average, reducing the enrichment over three-fold.

208   Including DMIs where a dProtein contained Pfam domain known to interact with the ELM class

209   (Figure 2, right column) dramatically increased the numbers of both predicted and random

210   DMIs, with a corresponding drop in enrichment. Using SLiMProb predictions in place of known

211   ELMs (Figure 2, bottom row) similarly increased both predicted and random DMIs, decreasing

212   enrichment. In all cases, none of the 1000 randomised datasets matched or exceeded the

213   observed number of predicted DMI, making the enrichment strongly significant ($P < 0.001$).

214   **Simulation of poor quality SLiM predictions**

215   To directly compare the effects of replacing real ELM-containing proteins with random human

216   proteins in different proportions (25%, 50%, 75%, 100%), the distribution of normalised

217   predicted DMI, D, in the Y2H data was compared for each dataset (Figure 6). D is the

218   distribution of expected true positive predicted DMIs, normalised to units of mean random

219   predicted DMIs, *i.e.* $D = 1$ is equivalent to FDR = 50%; enrichment is $1 + $ mean D. The more

220   permissive domain-based DMI prediction strategy (Figure 2, top right) was used, as the numbers

221   of predicted DMIs for more stringent strategies were very small (Table 1) and this strategy still

222   showed strong (7.6x) DMI enrichment in the data (Figure 5). Despite the decline in enrichment

223   scores with increasing proportions of random motif occurrences, enrichment remained

224   significant even when 75% of the real data was replaced (Figure 6).

225 **Discussion**

226   Using PPI data for SLiM discovery faces something of a contradiction. Due to their scale, data

227   from high throughput PPI detection studies are where the novel interactions are most likely to be

228   found. However, high stringency filters are often applied to high throughput methods to increase

229   confidence in individual interactions, with the concomitant concern that low affinity DMIs will

230   be lost as a consequence. The primary purpose of SLiMEnrich is to address this concern by

231   assessing how well a given PPI dataset is capturing DMIs. Where PPI datasets are large, this

232   assessment can be restricted to a high-quality set of known DMIs. Where the number of known

233   DMIs in the data becomes prohibitively small, predicted DMIs can supplement or replace the

234   known DMIs.

235   A detailed analysis of different PPI data sources is the subject of future study and beyond the

236   scope of this paper. Here, we present a case study to illustrate the use of SLiMEnrich to analyse

237   the DMI enrichment in a single PPI dataset. We have applied five different DMI

238   identification/prediction strategies (Figure 2) to a high-throughput Y2H human PPI study

239   (Rolland et al. 2014) (Table 1, Figure 5). On face value, the ability of the Y2H PPI data to

240   capture known DMIs might be considered disappointing. Only 14 of the 590 known human DMI

241   protein pairs in ELM (2.37%) were found in the 26,166 PPI considered. This is consistent with

242   earlier analyses that have highlighted the rarity of known SLiM-mediated interactions in high

243   throughput PPI data (Neduva & Russell 2006). However, even this modest numbers reflects a

244   massive enrichment (approx. 115-fold) over the expected number of known DMIs to occur in the

245   PPI data by chance. Whilst we cannot rule out unexpected confounding factors, such as

246   additional high affinity interactions between pairs of proteins that also share a DMI, this implies

247   that the low absolute numbers are due to the small number of known DMIs rather than the

248   inability of Y2H methods to detect DMIs. Considered analysis has estimated that the human

249   proteome has in the order of 100,000 SLiMs involved in DMIs (ignoring post-translational

250   modifications) (Tompa et al. 2014), which is orders of magnitude greater than the known DMIs

251   in ELM (Gouw et al. 2018). Overall, SLiMEnrich results indicate that these data are indeed

252   capturing real SLiM-mediated interactions and are therefore suitable for *de novo* SLiM

253   prediction. This, in turn, increases confidence in previous large scale SLiM predictions (Edwards

254   et al. 2012; Lieber et al. 2010; Neduva et al. 2005); these often rely on rediscovery of known

255   motifs as validation, which could be biased by incorporation of literature-based high confidence

256   DMIs in the PPI data.

257   Employing a less stringent DMI identification strategy predictably boosted the numbers of

258   predicted DMIs and continued to reveal significant enrichment in the Y2H data despite the

259   possible incorporation of possible false positive SLiM and/or DMI predictions (Table 1, Figure

260   5). As expected, the enrichment decreased as the noisiness of the data increased, although the

261   enrichment remained highly significant. This was supported by analysis where real ELM-

262   containing proteins were replaced with random human proteins to simulate noise (Figure 6).

263   Taken together, these results indicate a degree of robustness of the SLiMEnrich approach to the

264   quality of the SLiM data. However, they also highlight a lack of robustness in the individual

265   DMI predictions. For the purest known DMI analysis (linking known ELM instances to known

266   ELM-interacting *proteins*), most randomised datasets did not return a single DMI. It is therefore

267   highly likely that the 11 additional DMIs discovered by the ELMc-Protein strategy are real

268   DMIs. The cost is that the low numbers might affect the accuracy with which the mean random

269   DMI count, and thus enrichment, can be calculated. Relaxing the strategy to use SLiMProb

270   predictions and/or allow DMI predictions based on interactions between ELM classes and Pfam

271   domain classes, substantially increased the numbers of predicted DMIs but dramatically reduced

272   the observed enrichment for both known and predicted SLiM occurrences. Using predicted

273   SLiMs, it should be noted that the estimated false positive rate for individual DMI predictions is

274   very high (*FDR*=0.86 when linking predicted SLiMs via ELM-binding Pfam domains). This

275   highlights the need for caution when interpreting naïve large-scale predictions of this nature. As

276   illustrated for the Adenoviridae-human PPI example data (Figure 4), random numbers for the

277   Y2H case study will be inflated by a large over-prediction of kinase domain-mediated DMIs, as

278   well as other domains with a specificity of interaction not captured at the level of Pfam

279   definitions. Users may wish to screen out promiscuous domains and/or motifs if low stringency

280   approaches are required to get sufficient DMI numbers.

281   **Using SLiMEnrich to assess enrichment of different PPI types**

282   Although the focus of SLiMEnrich is on DMIs, the approach is flexible and can be easily

283   adapted to other PPI types. Direct analogues of DMIs can be studied by replacing the motifs with

284   a different interaction feature, *e.g.* replacing motifs with domains to investigate enrichment of

285   DDIs. More simply, SLiMEnrich could be used to study the overlap between two different PPI

286   datasets, accounting for the connectedness of the proteins involved, by replacing the known

287   ELM interactions with any source of pairwise PPIs. Although the PPI data for the case study was

288   made symmetrical, the asymmetrical handling of the PPI data by SLiMEnrich would even allow

289   intra-dataset comparisons, such as examining the overlap between PPIs when proteins are baits

290   versus preys in a Y2H or pulldown experiment.

291   # Conclusion

292   There are many data- and method-specific factors that will determine whether protein-protein

293   interaction (PPI) data are useful for short linear motif (SLiM) prediction. The presence of real

294   domain-motif interactions (DMIs) is a baseline requirement that is generally assumed but rarely

295   tested. SLiMEnrich is an open source R application that will identify known or predicted DMIs

296   in PPI data and estimate how well that PPI data is capturing DMIs compared to randomised PPIs.

297   This estimate is useful for identifying suitable PPI data for *de novo* SLiM prediction.

298   SLiMEnrich statistics also estimate the confidence in individual DMI predictions, enabling

299   assessment of methods that aim to improve the specificity of DMI predictions by filtering SLiM

300   predictions and/or PPI data. Users can run SLiMEnrich online

301 (http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/) or download the code for local use

302 (https://github.com/slimsuite/SLiMEnrich).

Acknowledgements

304 The authors would like to thank the UNSW Science Faculty Computing unit, especially Adrian

305 Plummer, for assistance in setting up the Shiny server. We thank Norman Davey for helpful

306 comments on a draft of the manuscript. We also thank Benjamin Lang and two anonymous

307 reviewers for insightful comments and suggestions on the original manuscript.

References

309 Bhattacharyya RP, Remenyi A, Yeh BJ, and Lim WA. 2006. Domains, motifs, and scaffolds: the

310 role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev

311 Biochem 75:655-680. 10.1146/annurev.biochem.75.103004.142710

312 Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel

313 H, and Gibson TJ. 2012. Attributes of short linear motifs. Mol Biosyst 8:268-281.

314 10.1039/c1mb05231d

315 Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, Milchevskaya V, Schneider

316 M, Kuhn H, Behrendt A, Dahl SL, Damerell V, Diebel S, Kalman S, Klein S, Knudsen AC,

317 Mader C, Merrill S, Staudt A, Thiel V, Welti L, Davey NE, Diella F, and Gibson TJ. 2016. ELM

318 2016-data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids

319 Res 44:D294-300. 10.1093/nar/gkv1291

320 Dosztanyi Z, Csizmok V, Tompa P, and Simon I. 2005. IUPred: web server for the prediction of

321 intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics

322 21:3433-3434. 10.1093/bioinformatics/bti541

323 Durmus Tekir S, Cakir T, Ardic E, Sayilirbas AS, Konuk G, Konuk M, Sariyer H, Ugurlu A,

324 Karadeniz I, Ozgur A, Sevilgen FE, and Ulgen KO. 2013. PHISTO: pathogen-host interaction

325 search tool. Bioinformatics 29:1357-1358. 10.1093/bioinformatics/btt137

326 Edwards RJ, Davey NE, O'Brien K, and Shields DC. 2012. Interactome-wide prediction of short,

327 disordered protein interaction motifs in humans. Mol Biosyst 8:282-295. 10.1039/c1mb05212h

328   Edwards RJ, Davey NE, and Shields DC. 2007. SLiMFinder: a probabilistic method for
329   identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS One
330   2:e967. 10.1371/journal.pone.0000967

331   Edwards RJ, and Palopoli N. 2015. Computational prediction of short linear motifs from protein
332   sequences. Methods Mol Biol 1268:89-141. 10.1007/978-1-4939-2285-7_6

333   Encinar JA, Fernandez-Ballester G, Sanchez IE, Hurtado-Gomez E, Stricher F, Beltrao P, and
334   Serrano L. 2009. ADAN: a database for prediction of protein-protein interaction of modular
335   domains mediated by linear motifs. Bioinformatics 25:2418-2424.
336   10.1093/bioinformatics/btp424

337   Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M,
338   Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, and Bateman A. 2016. The Pfam protein
339   families database: towards a more sustainable future. Nucleic Acids Res 44:D279-285.
340   10.1093/nar/gkv1344

341   Gouw M, Michael S, Samano-Sanchez H, Kumar M, Zeke A, Lang B, Bely B, Chemes LB,
342   Davey NE, Deng Z, Diella F, Gurth CM, Huber AK, Kleinsorg S, Schlegel LS, Palopoli N, Roey
343   KV, Altenberg B, Remenyi A, Dinkel H, and Gibson TJ. 2018. The eukaryotic linear motif
344   resource - 2018 update. Nucleic Acids Res 46:D428-D434. 10.1093/nar/gkx1077

345   Gouw M, Samano-Sanchez H, Van Roey K, Diella F, Gibson TJ, and Dinkel H. 2017. Exploring
346   Short Linear Motifs Using the ELM Database and Tools. Curr Protoc Bioinformatics 58:8 22 21-
347   28 22 35. 10.1002/cpbi.26

348   Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, and
349   Sigrist CJ. 2006. The PROSITE database. Nucleic Acids Res 34:D227-230. 10.1093/nar/gkj063

350   Kelil A, Levy ED, and Michnick SW. 2016. Evolution of domain-peptide interactions to coadapt
351   specificity and affinity to functional diversity. Proc Natl Acad Sci U S A 113:E3862-3871.
352   10.1073/pnas.1518469113

353   Krystkowiak I, and Davey NE. 2017. SLiMSearch: a framework for proteome-wide discovery
354   and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res.
355   10.1093/nar/gkx238

356   Lieber DS, Elemento O, and Tavazoie S. 2010. Large-scale discovery and characterization of
357   protein regulatory motifs in eukaryotes. PLoS One 5:e14444. 10.1371/journal.pone.0014444

358   Luck K, Fournane S, Kieffer B, Masson M, Nomine Y, and Trave G. 2011. Putting into practice
359   domain-linear motif interaction predictions for exploration of protein networks. PLoS One
360   6:e25376. 10.1371/journal.pone.0025376

361   Lyon KF, Cai X, Young RJ, Mamun AA, Rajasekaran S, and Schiller MR. 2018. Minimotif
362   Miner 4: a million peptide minimotifs and counting. Nucleic Acids Res 46:D465-D470.
363   10.1093/nar/gkx1085

364   Mi T, Merlin JC, Deverasetty S, Gryk MR, Bill TJ, Brooks AW, Lee LY, Rathnayake V, Ross
365   CA, Sargeant DP, Strong CL, Watts P, Rajasekaran S, and Schiller MR. 2012. Minimotif Miner
366   3.0: database expansion and significantly improved reduction of false-positive predictions from
367   consensus sequences. Nucleic Acids Res 40:D252-260. 10.1093/nar/gkr1189

368   Mosca R, Ceol A, Stein A, Olivella R, and Aloy P. 2014. 3did: a catalog of domain-based
369   interactions of known three-dimensional structure. Nucleic Acids Res 42:D374-379.
370   10.1093/nar/gkt887

371   Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, and
372   Russell RB. 2005. Systematic discovery of new recognition peptides mediating protein
373   interaction networks. PLoS Biol 3:e405. 10.1371/journal.pbio.0030405

374   Neduva V, and Russell RB. 2005. Linear motifs: Evolutionary interaction switches. Febs Letters
375   579:3342-3345. 10.1016/j.febslet.2005.04.005

376   Neduva V, and Russell RB. 2006. Peptides mediating interaction networks: new leads at last.
377   Curr Opin Biotechnol 17:465-471. 10.1016/j.copbio.2006.08.002

378   Palopoli N, Lythgow KT, and Edwards RJ. 2015. QSLiMFinder: improved short linear motif
379   prediction using specific query protein data. Bioinformatics 31:2284-2293.
380   10.1093/bioinformatics/btv155

381   Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I,
382   Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE,
383   Braun P, Brehme M, Broly MP, Carvunis AR, Convery-Zupan D, Corominas R, Coulombe-
384   Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy

385 MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi

386 A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie JM, Scholz A,

387 Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere JC, Vega K,

388 Walsh J, Cusick ME, Xia Y, Barabasi AL, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J,

389 Calderwood MA, Hill DE, Hao T, Roth FP, and Vidal M. 2014. A proteome-scale map of the

390 human interactome network. Cell 159:1212-1226. 10.1016/j.cell.2014.10.050

391 Sarkar D, Jana T, and Saha S. 2015. LMPID: a manually curated database of linear motifs

392 mediating protein-protein interactions. Database (Oxford) 2015. 10.1093/database/bav014

393 Tompa P, Davey NE, Gibson TJ, and Babu MM. 2014. A million peptide motifs for the

394 molecular biologist. Mol Cell 55:161-169. 10.1016/j.molcel.2014.05.032

395 UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res

396 45:D158-D169. 10.1093/nar/gkw1099

397 Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, and Davey NE.

398 2014. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules

399 Directing Cell Regulation. Chemical Reviews 114:6733-6778. 10.1021/cr400585q

400 Weatheritt RJ, Jehl P, Dinkel H, and Gibson TJ. 2012. iELM--a web server to explore short

401 linear motif-mediated interactions. Nucleic Acids Res 40:W364-369. 10.1093/nar/gks444

402

403

404

Figure 1

A schematic representation of the main SLiMEnrich pipeline.

SLiMEnrich takes four input files: **1.** PPI data provided by the user as a set of pairwise putative motif-containing proteins ("mProteins") and their domain-containing interaction partners ("dProteins"); **2.** A file providing known or predicted motif occurrences within the mProtein sequences (by default, known ELM instances are used); **3.** A DMI file defining Motif-Domain interactions, relating to the DMI Strategy employed (by default, known ELM interactions are used); **4.** A file that links dProteins to their domain composition (by default, human Pfam domains from UniprotKB are used). Input data is combined to establish the complete set of known/predicted "potential DMI" dependent on the DMI strategy selected (see Figure 2 and text for details): **ELMi-Protein** – for highest stringency, the DMI file directly links mProteins to known dProtein DMI partners (Motifs and Domains input not used); **ELMc-Protein** – for medium stringency, the DMI file links Motif classes to known dProtein DMI partners (Domains input not used); **ELMc-Domain** – for lowest stringency, the DMI file links Motif classes to known interacting Domains. Potential DMIs are then mapped on to the input PPI to identify the "Predicted DMIs" in the real data. PPI data is randomised (shuffled) 1000 times and re-mapped to potential DMIs to determine the background distribution of predicted DMIs in the case of random association (see text for details). Finally, the "Random DMI" distribution is compared to the observed "Predicted DMIs" to determine DMI enrichment in the data. Results are output in the form of a tables, a histogram of the Random DMI distribution with the observed count and empirical P-value marked, and an interactive network of the known/predicted DMIs found in the PPI data.

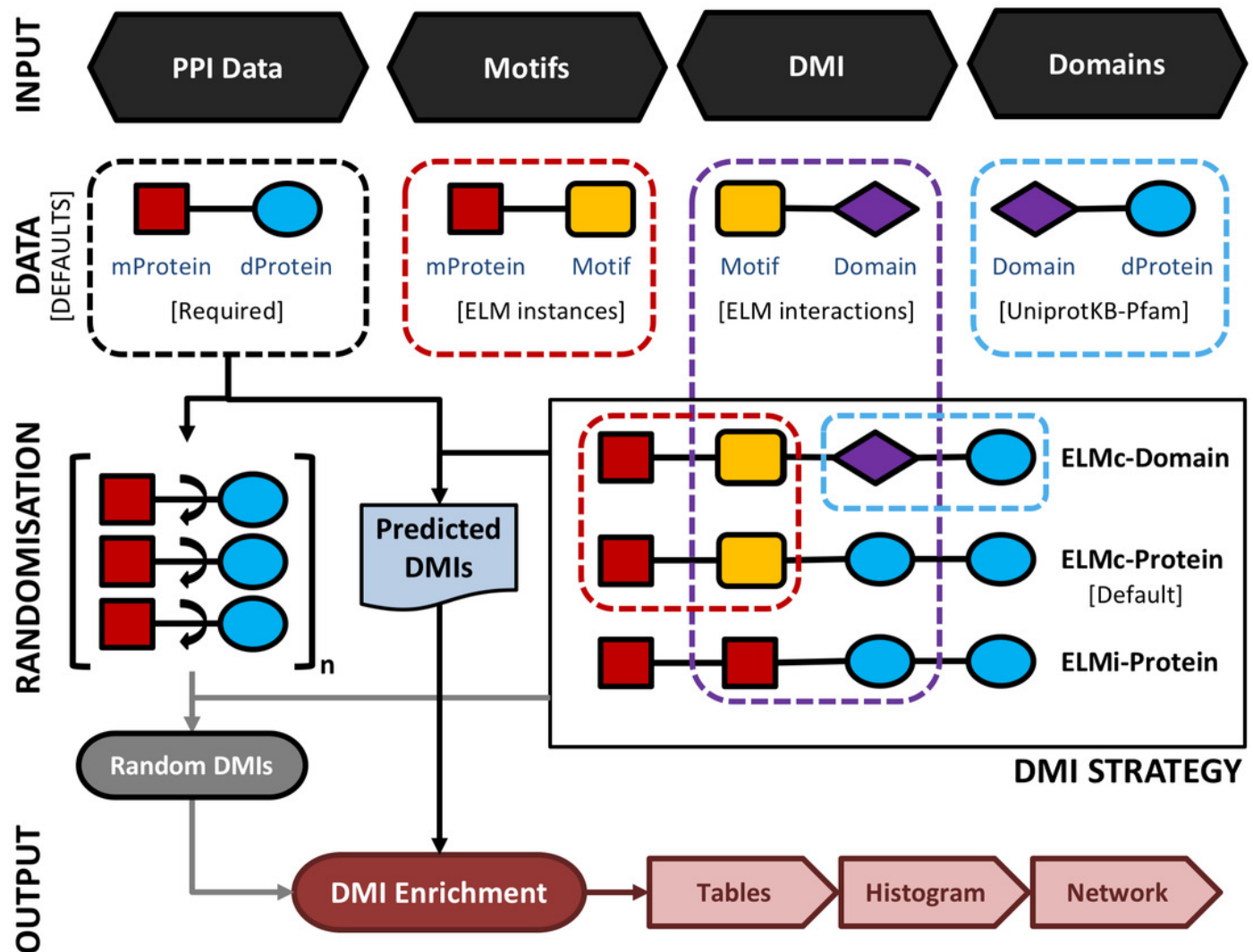



Figure 2

SLiMEnrich DMI prediction strategies.

SLiMEnrich uses known DMI from the ELM database to identify known DMIs or predict DMIs within the supplied PPI data. **(A)** In this example, Motif A is known to interact with Domain B. Motif A has two known occurrences in the data (green circles) and two predicted occurrences (red circles). Domain B is present in four proteins (squares). ELM has two annotated interactions between proteins with Motif A and proteins with Domain B (blue). **(B)** In the simplest and purest strategy, only known ELM interactions (ELMi) are used to assess enrichment (right panel, top left box). For small PPI datasets it might be necessary to increase the number of predicted DMI. This can be done in two ways. Top row: known motif occurrences (green circles) can be connected to all proteins known to interact with that ELM class (ELMc) (blue squares, top centre), or connected to all proteins containing a domain that interacts with that ELM class (all squares, top right). Bottom row: to increase the number of DMI further, known ELM occurrences can be replaced with SLiM predictions (all circles).

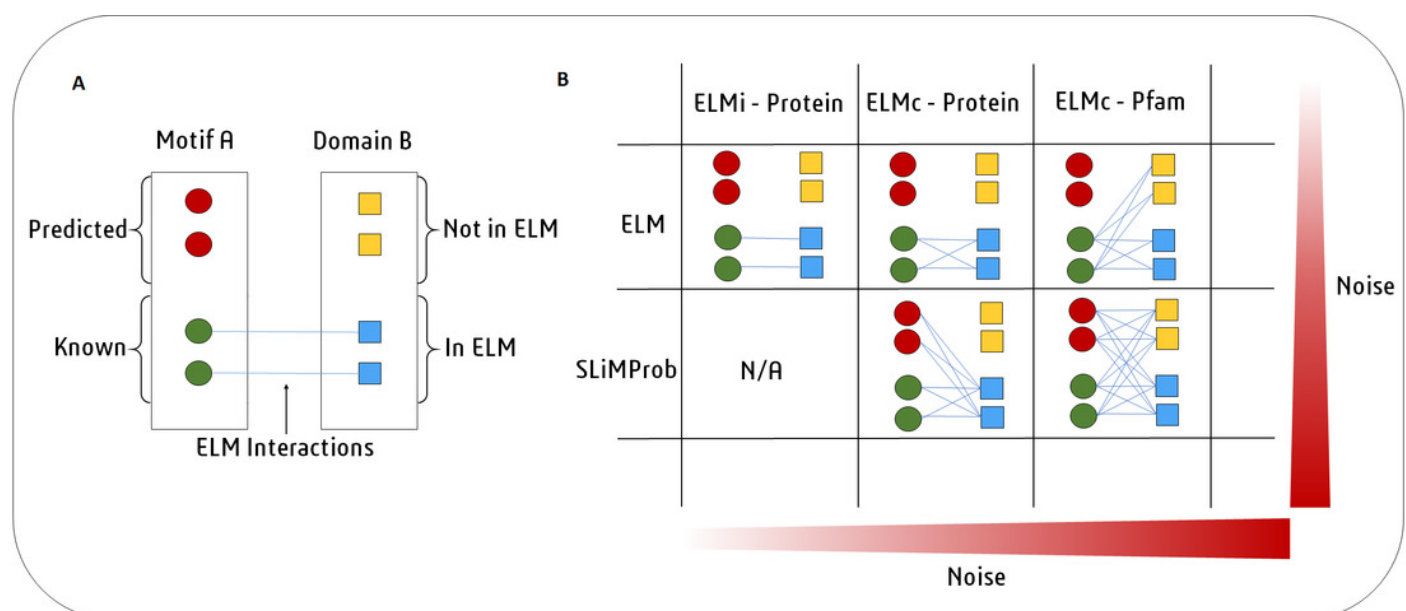



Figure 3

DMI enrichment histogram for SLiMEnrich example data.

Histogram of DMI enrichment in example data for Adenoviridae proteins and their human interactors (see text for details) from the SLiMEnrich app, using the most permissive ELMc-Domain DMI strategy and SLiMProb motif predictions. Frequency bars indicate the number of randomised PPI datasets returning a given number of predicted DMIs. The dotted arrow indicates the observed number of predicted DMIs in the real data.

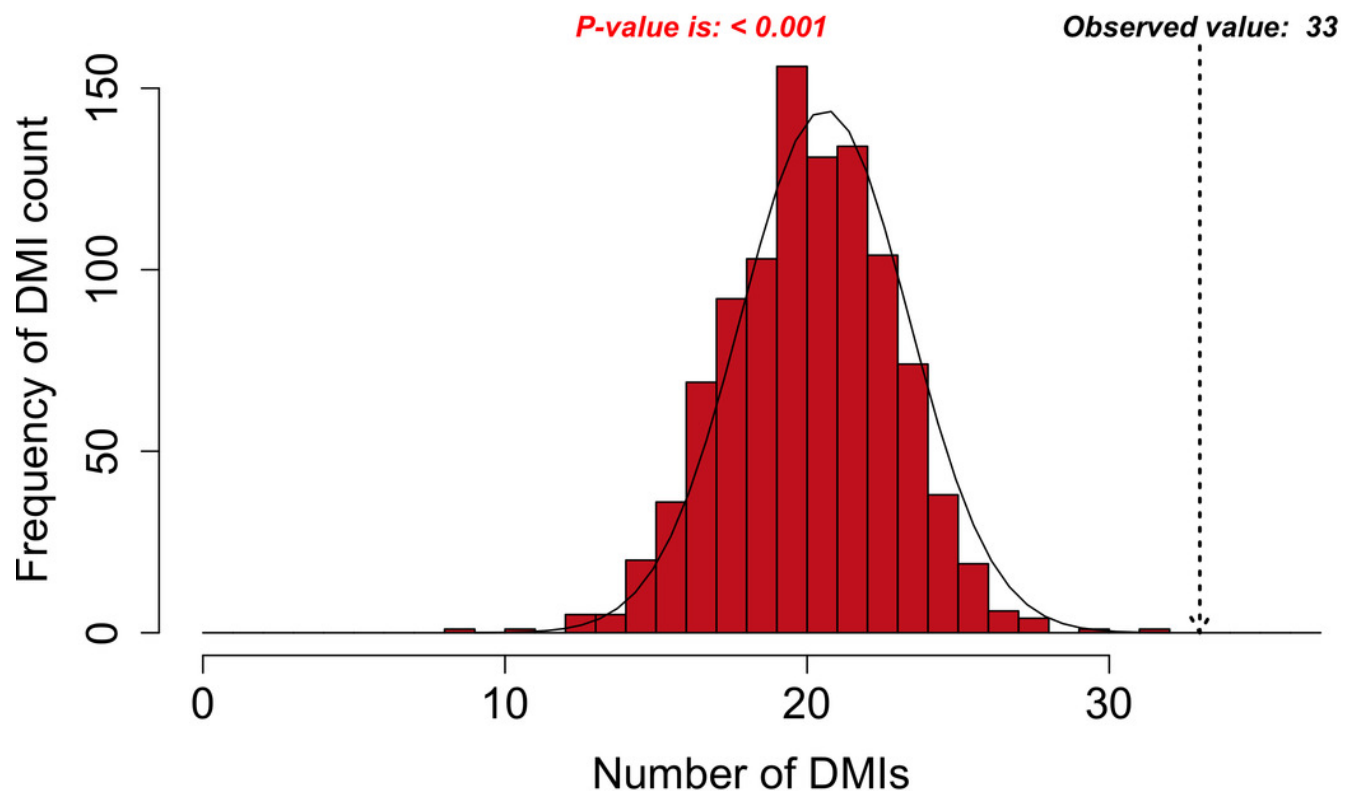



Figure 4

Interactive predicted DMI network for example data.

Predicted DMIs for example Adenoviridae proteins and their human interactors, using the most relaxed strategy (predicted SLiMs connected via domains, see text for details). Several layout options are provided and nodes can be manually positioned. The protein, domain and motif identifiers used in the network are determined by the user input. Using default data, these will be UniprotKB, Pfam and ELM identifiers. For this example, UniprotKB identifiers have been mapped onto HGNC gene symbols and Pfam identifiers onto Pfam domain names. Red square, motif-containing protein ("mProtein"); Yellow box, motif; Purple ellipse, domain; Blue circle, domain-containing protein ("dProtein").

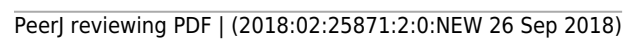



Figure 5

Enrichment statistics and histogram of expected random DMI counts in human Y2H case study data using known and predicted ELM instances.

Frequency bars indicate the number of randomised PPI datasets returning a given number of predicted DMIs. The dotted arrow indicates the observed number of known or predicted DMIs in the Y2H data (see text for details). DMI prediction strategies match those in Figure 2 (see text for details): **(A)** known ELM occurrences connected to interacting proteins; **(B)** known ELM occurrences mapped to proteins known to interact with that motif class; **(C)** known ELM occurrences mapped to proteins containing a domain known to interact with that motif class; **(D)** SLiM predictions mapped to proteins known to interact with that motif class; **(E)** SLiM predictions mapped to proteins containing a domain known to interact with that motif class.

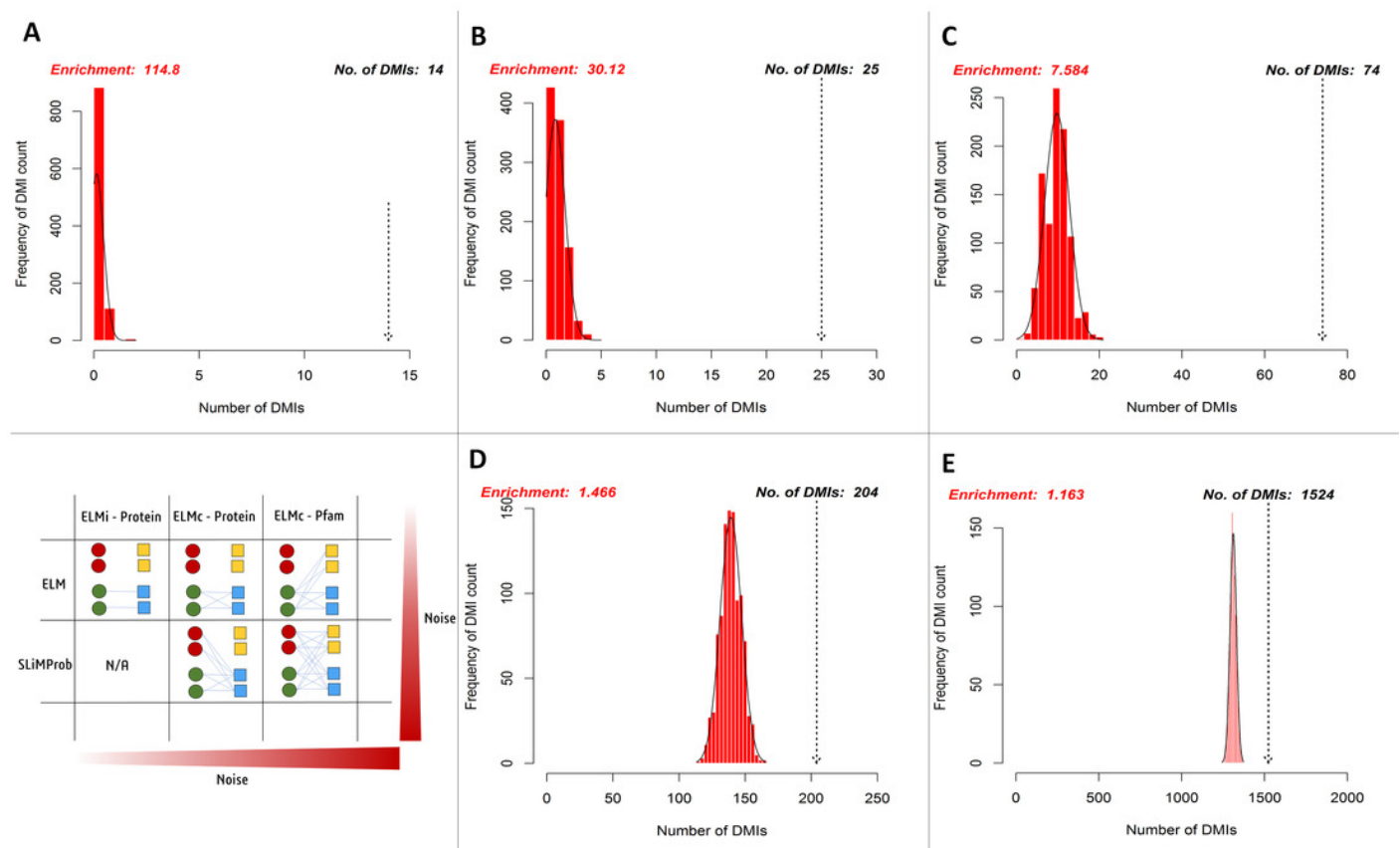
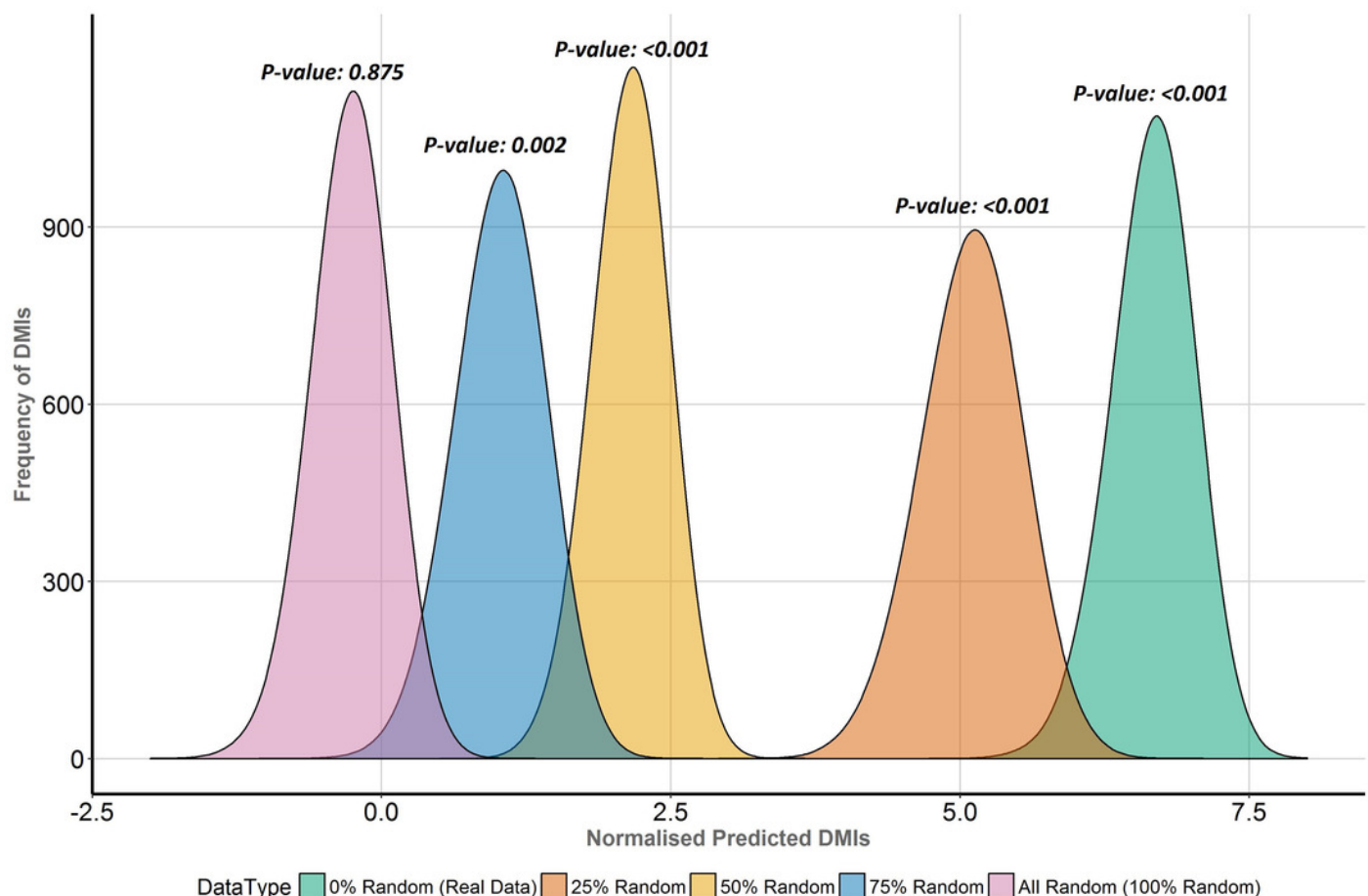



Figure 6

Enrichment analysis of known DMIs in human Y2H case study data with increasing proportions of random motif instances instances.

SLiMEnrich results for known ELMs in the human Y2H case study data mapped using the ELMc-Domain strategy, converted into the normalised number of predicted real DMIs (see text for details). Higher normalised predicted DMI counts indicate greater DMI enrichment, with zero marking no enrichment over random. Green (furthest right) is the real data using all known true positive ELM instances. The other curves (right to left) represent distributions for four randomised datasets where increasing proportions (25%, 50%, 75% and 100%) of ELM proteins were replaced with random human proteins.

SLiMEnrich analysis of Y2H case study using different DMI prediction strategies.

SLiMEnrich strategy	Known: ELMi-Protein	Known: ELMc-Protein	Known: ELMc-Domain	SLiMProb: ELMc-Protein	SLiMProb: ELMc-Domain
Potential DMI (NR)	62	164	6,314	39,572	969,380
Predicted DMI (NR)	14†	25	74	204	1,524
Mean Random DMI (3 s.f.)	0.122	0.830	9.76	139	1,310
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Enrichment (3 s.f.)	115	30.1	7.58	1.47	1.16
FDR (4 d.p.)	0.0087	0.0332	0.1319	0.6820	0.8602
Unique mProteins*	13	22	52	175	768
Unique ELM classes*	N/A	16	40	35	128
Unique Pfam domains*	N/A	N/A	30	N/A	51
Unique dProteins*	10	17	53	36	366

2   * Unique counts correspond to Predicted DMI.
3   † Known DMI from ELM database.