DoReMi: Context-based prioritization of linear motif matches

Heiko Horn

NNF Center for Protein Research, University of Copenhagen, Denmark.

Niall Haslam

Complex and Adaptive Systems Laboratory, University College Dublin, Dublin, Ireland

Conway Institute of Biomolecular and Biomedical Science, University College Dublin, Dublin,

Ireland.

Lars Juhl Jensen

NNF Center for Protein Research, University of Copenhagen, Denmark.

Email: heiko.horn@cpr.ku.dk or lars.juhl.jensen@cpr.ku.dk

Introduction

Physical interactions among proteins play a crucial role in cellular signaling. There are two types of binding: domain–domain and domain–motif. For the latter type, a conserved domain interacts with a short peptide sequence (linear motif), which may be subject to post-translational modification (PTM) (Linding, 2010). Many proteins carry more than one domain, which allows them to bind to multiple different proteins and thereby act, for example, as scaffolds. Moreover, recent work has shown interdependence of PTM sites both with each other and with linear motifs (Minguez et al., 2012). Studying these binding sites and the domains that recognize them can thus help decipher the complex signalling networks in the cell.

The computational task of dealing with motifs can be split into two separate tasks: the discovery of novel, not known motifs and the search for new instances of already known motifs. The discovery of novel motifs is challenging, nonetheles a number of techniques have been developed to address this problem (Chou & Schwartz, 2011; Davey et al., 2010; Lam et al., 2010). More recently structural approaches have been applied to the problem(Pathak et al., 2013), which similar to other tools rely on filtering out false positives in the discovery phase of motif identification. Physicochemical approaches have been developed to uncover patterns in the primary sequence that are characteristic of short peptide binding region(Dosztányi et al., 2009), and machine learning techniques employed to learn motifs from literature curated datasets(Lieber et al., 2010; Mooney et al., 2012; Nguyen Ba et al., 2012). Also, several resources exist that attempt to catalogue instances of linear motifs, incorporating manually curated and computational predictions of motifs (Dinkel et al., 2011; Mi et al., 2012; Sigrist et al., 2010). The discovery of both novel motifs and new instances is hindered by the problem of finding statistical significant matches. Many motifs are short peptide sequences and thereby appear in many different proteins by chance without representing a functional site (Obenauer et al., 2003).

When searching for new instances, all occurrences of a motif should thereby be screened for their biological validity. Based on the protein sequence, predictors for protein folding, structure and disorder can be used to filter for potential motifs (Davey et al., 2011). As functionally important sections of proteins tend to be conserved, further improvement can be gained by incorporating

40   orthology information. Additionally, keyword annotations can help to distinguish the interesting
41   proteins for the investigated processes (Ramu, 2003).

42   Further improvements can be achieved by integrating the cellular context of the interactions. As
43   we try to filter for binding-associated motifs, the information about protein-protein interactions
44   between the substrates and a protein containing the binding domain is an obvious approach. As
45   long as this information is available, the question of biological relevance can be coupled to the
46   question of co-localization and association; a binding site that will never be in proximity to the
47   corresponding domain is biological not relevant in the studied context.

48   Previous work has shown that contextual information in the form of known physical interactions
49   (Mi et al., 2012) or functional associations (Linding et al., 2007; Weatheritt & Jehl, 2012) can
50   improve motif based approaches. These methods are publicly available, but limit the query to a
51   pre-defined set of motifs and interacting partners, which prevents their use for analysis of not
52   already known motifs.

53   Here, we present DoReMi (Domain aided Regular Expression Mining), a tool that allows the user
54   to specify any motif and the associated binding domain(s). By integrating the protein-protein
55   association data into the scoring schema, we extend the usage from being a subsequently applied
56   filter to a factor for ranking found potential instances. The algorithm is publicly available using
57   the web interface at doremi.jensenlab.org.

Materials & Methods

58
59   The algorithm takes a motif representation as input and scans a selection of representative
60   sequences to find matching instances. To filter these hits, the score of a motif search gets
61   combined with the protein-protein association scores derived from the STRING database to rank
62   the occurrences (Franceschini et al., 2013). Previous work has demonstrated the utility of
63   allowing both regular expressions (RegEx) and position specific scoring matrices (PSSM) to
64   generate motif definitions (Haslam & Shields, 2012); therefore we allow both of these as input as
65   well as the result of any more advanced machine learning approach with a qualitative result.

66   The association scores are generated from the STRING confidence association scores. As the
67   database only provides direct associations, indirect paths have to be calculated to cover also
68   indirect interactions like e.g. scaffolding. Indirect interactions are computed by multiplying
69   connecting edges on the path between the proteins of interest. To reduce the effect of over-
70   connected nodes (hubs), we penalize for the connectivity by normalizing for the sum of all
71   connecting edges of the node. The final score of an interaction can be described by the formula:

$$\frac{C\left(n_n \rightarrow n_x\right) - \left(C\left(n_{n-1} \rightarrow n_n\right) + C\left(n_n \rightarrow n_n\right)\right)}{1 + \sum ¿}$$

$$¿$$
$$¿^{\beta}$$
$$¿$$
$$¿$$

72   with $C\left(n_{n-1} \rightarrow n_n\right)$ being the confidence score of

$$\frac{C\left(n_{n-1} \rightarrow n_n\right)}{¿}$$

$$\prod_{n=2}^{n_{last}} ¿$$

73 two nodes interacting and $\sum C(n_n \rightarrow n_x)$ being the sum of the confidence scores of all

74 connecting edges of n_n. The parameter β was set based on our experience in the

75 NetworKIN project to a value of 0.34.

76 After extending the network to cover indirect paths, the dataset covers a wide range of
77 associations between single proteins, each represented by a single score. This score can be
78 interpreted as the likelihood of two proteins interacting with each other.

79 The potential interaction partners are recruited by selecting all proteins containing the interacting
80 domain(s). To find these proteins, the Hidden-Markov-Model (HMM) representations provided by
81 the PFAM database are used to scan the whole STRING sequence database for matching
82 structures (Punta et al., 2012). For the scan, we use the HMMER3 package with default
83 parameters as stated by PFAM. These domain-protein associations are pre-calculated to later
84 increase the performance of the web service. All proteins carrying at least one of the interacting
85 domain(s) of interest are treated as potential interactors. In the case of multiple potential
86 interactors for one substrate, the highest scoring one is considered the most probable and its score
87 is used in the subsequent calculations.

88 The result from the PSSM search and the derived contextual network score are combined using a
89 weighting factor to be able to adjust the algorithm to different biological cases.

$$Final\ Score = Motif^{1-\alpha} * Association^{\alpha}$$

90 This weighting factor is used to account for the different impact the two factors have in the
91 biological context. For example, the SUMO binding domain has a strongly conserved motif,
92 leading to the assumption that the motif is sufficient to achieve good predictive performance.
93 Conversely, the process of ubiquitination is well known for not having any preference for certain
94 amino acids surrounding the modified residue, thus it is favorable to put more weight on the
95 context information. If the motif search method only returns a binary value (e.g. regular
96 expression returns only matches but no scores), DoReMi sorts the matches by the context network
97 score alone.

98 To make the algorithm easily accessible and applicable, we developed a web service
99 implementing the described workflow. The interface of this webservice allows for multiple
100 different inputs for the motif finding part, which are subsequently enhanced with the pre-
101 computed network information.

Results and Discussion
103 To estimate the performance of our method, we selected two sets of binding sites, which have
104 been previously been used for training of SUMOylation predictors (Matic et al., 2010; Ren et al.,
105 2009). The family of SUMO (Small Ubiquitin-like Modifier) proteins comprises of multiple
106 small members that are covalently attached to the proteins. As the name suggests, the proteins are
107 similar to ubiquitin, and the whole enzymatic cascade is alike to the process of ubiquitination.
108 SUMOylation is involved in functions like protein stability, nuclear-cytosolic transport, and
109 transcriptional regulation. The SUMO binding motif can be split into 2 groups: The first motif is
110 a strongly conserved one whereas the second is less restrictive.

111 The older dataset from Ren et al. was used to calculate the PSSM and set the weighting factor α.
112 We subsequently benchmarked the resulting predictor on the newer dataset by Matic et al.,
113 removing all sites that have been already used in the training. To compare the approach of
114 creating a PSSM to a regular expression, we used the SUMOylation motif

$$[VILMAFP](K).E$$

115 from the Eukaryotic Linear Motif (ELM) database (Dinkel et al., 2011).

116 In addition, we trained a predictor based on the combined datasets and benchmarked its
117 performance by five-fold cross validation. The results of the partitioning were slightly better than
118 the results from training on one dataset and benchmarking on the other. This suggests that cross
119 validation gives a realistic performance estimate.

120 Given this, we also used cross validation to test the ability of DoReMi to predict binding partners
121 for PDZ domains (Kim et al., 2012). PDZ domains are abundant in eukaryotes and eubacteria and
122 commonly bind to the C-terminal end of proteins (Ponting, 1997). They are often associated with
123 other domains like SH3 and therefor play an important role in the cell signaling(Zhang et al.,
124 2011). The ELM database lists three different motifs, although they can be easily joined as they
125 differ only in position four:

$$\left.\begin{array}{l}...[DE].[ACVILF]\$ \\ ...[VLIFY].[ACVILF]\$ \\ ...[ST].[ACVILF]\$\end{array}\right\}...[DEFILSTVY].[ACVILF]\$$$

126 Overall, the motif is relatively loose and the C-terminal location may be the strongest indicator for
127 a real occurrence.

128 The combined set for SUMOylation contained 1063 sites while the publication by Kim et al.
129 provided 367 instances. For the benchmarking, we pre-defined α to 0.4 and 0.8 for the SUMO
130 and PDZ test sets respectively. We selected these values based on the conservation of the known
131 sequences, as one can expect a greater influence of the cellular context, given the higher
132 flexibility of the motif. Cross validation was repeated five times and the reported numbers are the
133 median performance.

134 When selecting an appropriate score cut off by fixing the false positive rate (0.3 for SUMO, 0.15
135 for PDZ) we can show an improvement of the true positive rate by up to 15 percentage points for
136 the PDZ domain and 10 for the SUMO binding domain. The smaller increase for the SUMO
137 domain is explained by the already good performance of the motif-based predictor, which leaves
138 less room for improvement. These examples use simple motif predictors; for comparison, we also
139 applied the context information to an already existing SUMO predictor (SUMOsp) to investigate
140 the possibility of improving on more advanced methods. We acknowledge that the result will be
141 influenced by the fact that this benchmark uses true positives on which the predictor got trained.
142 The shown improvement by using contextual information should be valid. Comparable to the first
143 examples, we see an improvement of the true positive rate by up to 10 percentage points, which
144 motivated the inclusion of this possibility into the web interface. A more detailed listing of the
145 performance numbers can be seen in Table 1. Although the overall performance is not significant

146 better for some of the examples, in a real world scenario, the method performs better. The stated
147 improvements mean, that if we would have a list of results and sort it by the final score, true
148 positives are enriched in the beginning of this list. This is important, as the search of new
149 instances of known motifs is mainly dominated by the task of filtering out false positives.

150 These results show that, although advanced methods like SUMOsp perform better than the basic
151 approach of combining a RegEx or PSSM with contextual information, it can also be improved by
152 the same approach. This shows the flexibility and versatility of this approach. Especially under
153 the current state of lack of data that we see in the field of linear motif, it can be of great help in
154 discriminating between true and false positive predictions. It also adds a non-sequence specific
155 layer to many methods.

156 The web interface allows the user to easily apply the algorithm in a simple two-step process. The
157 first step allows the user to define the potential binding sites. The first input type is specifying a
158 regular expression; further details on the allowed format options are shown in Table 2.
159 Alternatively, the user can provide a set of known binding peptides, which will be used to
160 calculate a PSSM. Both the regular expression and the PSSM will be used to scan the human
161 proteome for occurrences. As last option, the user can submit predictions from other motif search
162 engines by providing identifier, position and score. These identifiers and positions have to be
163 prior been mapped to the background sequence set used in DoReMi (STRING database version
164 9.05).

165 The second step allows for the selection of one or more potential binding domains. The interface
166 allows the user to search for Pfam domains by their accession, id or description. From the list of
167 search results, the user can select the relevant domains (Figure 1A). If necessary, multiple
168 searches can be performed to select differently named domains.

169 The output page shows a brief summary of the analysis (Figure 1B). This includes plots of score
170 distributions for each score (motif, network and combined score) to aid in the selection of an
171 appropriate score cut-off for downstream analysis of the results after downloading.

Conclusions

173 Our work focused on making the integration of contextual information to enhance the search for
174 motif occurrences into an easy to execute task. We show, that the addition of protein-protein
175 interaction knowledge can improve the predictive performance of existing methods, even when
176 doing it in a fully automated manner. This can be especially useful in the field of linear motifs; as
177 for these, due to lack of data, advanced methods are not easy to develop. Further performance
178 enhancements can potentially be reached when improving the domain associations in a more
179 confined way.

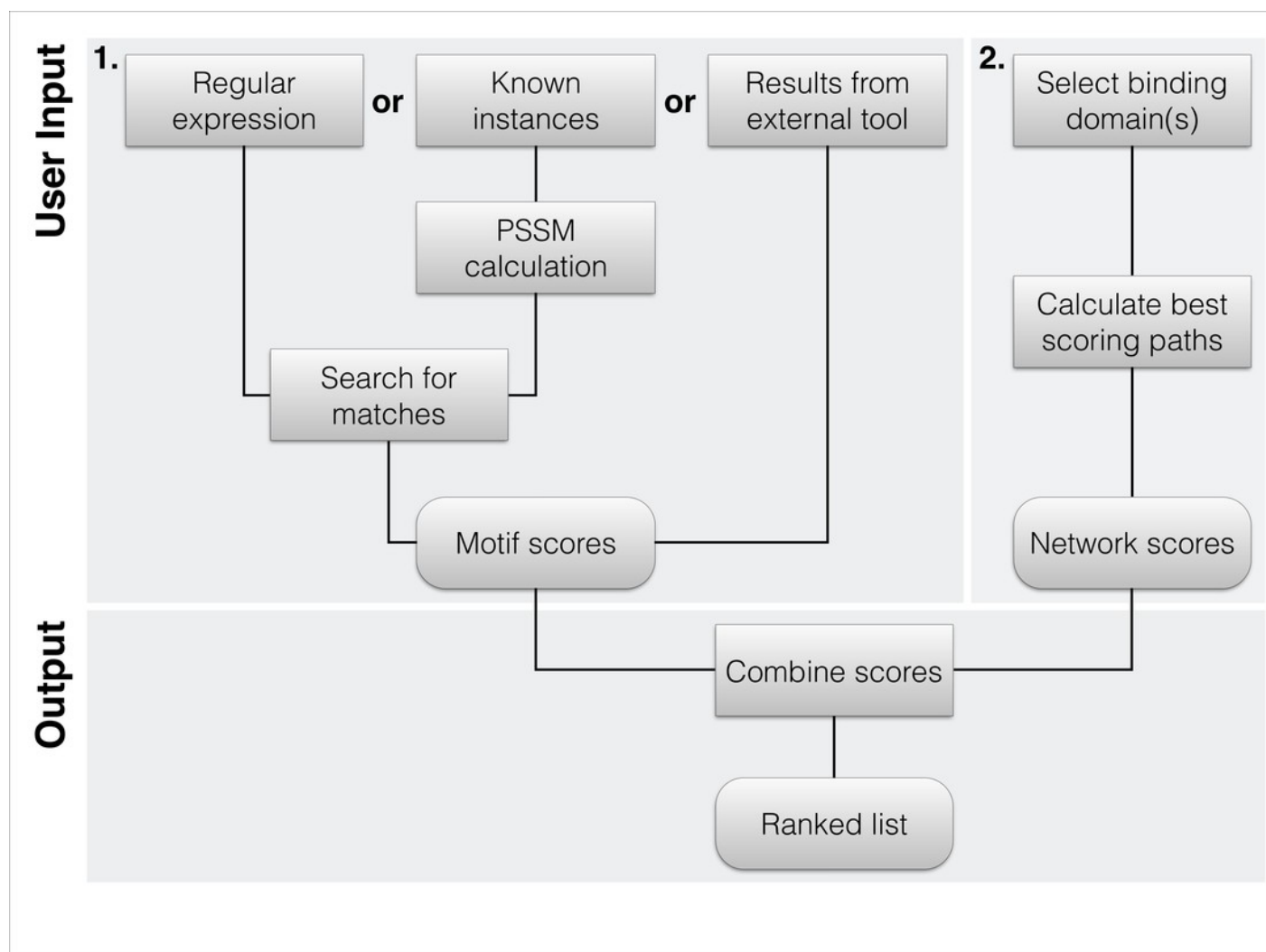
180 For future enhancements, we plan to expand the usability of the method and web interface.
181 Currently, only the human proteome is available; however, the method is in principle applicable to
182 any fully sequenced organism. We thus consider expanding it to major eukaryotic model
183 organisms. We are also working on integrating DoReMi with existing motif resources such as the
184 ELM database.

185 Chou, M. F., & Schwartz, D. (2011). Biological sequence motif discovery using motif-x. *Current*
186   *Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *Chapter 13*,
187   Unit 13.15–24. doi:10.1002/0471250953.bi1315s35
188 Davey, N. E., Haslam, N. J., Shields, D. C., & Edwards, R. J. (2010). SLiMFinder: a web server to
189   find novel, significantly over-represented, short protein motifs. *Nucleic Acids Research*,
190   *38*(Web Server issue), W534–9. doi:10.1093/nar/gkq440
191 Davey, N. E., Haslam, N. J., Shields, D. C., & Edwards, R. J. (2011). SLiMSearch 2.0: biological
192   context for short linear motifs in proteins. *Nucleic Acids Research*, *39*(Web Server issue),
193   W56–60. doi:10.1093/nar/gkr402
194 Dinkel, H., Michael, S., Weatheritt, R. J., Davey, N. E., Van Roey, K., Altenberg, B., Toedt, G.,
195   Uyar, B., Seiler, M., Budd, A., Jödicke, L., Dammert, M. a, Schroeter, C., Hammer, M.,
196   Schmidt, T., Jehl, P., McGuigan, C., Dymecka, M., Chica, C., Luck, K., Via, A., Chatr-
197   Aryamontri, A., Haslam, N., Grebnev, G., Edwards, R. J., Steinmetz, M. O., Meiselbach, H.,
198   Diella, F., & Gibson, T. J. (2011). ELM--the database of eukaryotic linear motifs. *Nucleic*
199   *Acids Research*, *49*(2), 1–10. doi:10.1093/nar/gkr1064
200 Dosztányi, Z., Mészáros, B., & Simon, I. (2009). ANCHOR: web server for predicting protein
201   binding regions in disordered proteins. *Bioinformatics (Oxford, England)*, *25*(20), 2745–6.
202   doi:10.1093/bioinformatics/btp518
203 Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J.,
204   Minguez, P., Bork, P., von Mering, C., & Jensen, L. J. (2013). STRING v9.1: protein-protein
205   interaction networks, with increased coverage and integration. *Nucleic Acids Research*,
206   *41*(Database issue), D808–15. doi:10.1093/nar/gks1094
207 Haslam, N. J., & Shields, D. C. (2012). Profile-based short linear protein motif discovery. *BMC*
208   *Bioinformatics*, *13*(1), 104. doi:10.1186/1471-2105-13-104
209 Kim, J., Kim, I., Yang, J.-S., Shin, Y.-E., Hwang, J., Park, S., Choi, Y. S., & Kim, S. (2012).
210   Rewiring of PDZ domain-ligand interaction network contributed to eukaryotic evolution.
211   *PLoS Genetics*, *8*(2), e1002510. doi:10.1371/journal.pgen.1002510
212 Lam, H. Y. K., Kim, P. M., Mok, J., Tonikian, R., Sidhu, S. S., Turk, B. E., Snyder, M., &
213   Gerstein, M. B. (2010). MOTIPS: automated motif analysis for predicting targets of modular
214   protein domains. *BMC Bioinformatics*, *11*, 243. doi:10.1186/1471-2105-11-243
215 Lieber, D. S., Elemento, O., & Tavazoie, S. (2010). Large-scale discovery and characterization of
216   protein regulatory motifs in eukaryotes. *PLoS One*, *5*(12), e14444.
217   doi:10.1371/journal.pone.0014444
218 Linding, R. (2010). (R)evolution of complex regulatory systems. *Science Signaling*, *3*(127), eg4.
219   doi:10.1126/scisignal.3127eg4
220 Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. a T. M., Jørgensen, C., Miron, I. M.,
221   Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin,
222   J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., &
223   Pawson, T. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, *129*(7),
224   1415–26. doi:10.1016/j.cell.2007.05.052
225 Matic, I., Schimmel, J., Hendriks, I. a, van Santen, M. a, van de Rijke, F., van Dam, H., Gnad, F.,
226   Mann, M., & Vertegaal, A. C. O. (2010). Site-specific identification of SUMO-2 targets in
227   cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif.
228   *Molecular Cell*, *39*(4), 641–52. doi:10.1016/j.molcel.2010.07.026
229 Mi, T., Merlin, J. C., Deverasetty, S., Gryk, M. R., Bill, T. J., Brooks, A. W., Lee, L. Y.,
230   Rathnayake, V., Ross, C. a, Sargeant, D. P., Strong, C. L., Watts, P., Rajasekaran, S., &

231   Schiller, M. R. (2012). Minimotif Miner 3.0: database expansion and significantly improved
232   reduction of false-positive predictions from consensus sequences. *Nucleic Acids Research*,
233   *40*(Database issue), D252–60. doi:10.1093/nar/gkr1189
234 Minguez, P., Parca, L., Diella, F., Mende, D. R., Kumar, R., Helmer-Citterich, M., Gavin, A.-C.,
235   van Noort, V., & Bork, P. (2012). Deciphering a global network of functionally associated
236   post-translational modifications. *Molecular Systems Biology*, *8*(599), 599.
237   doi:10.1038/msb.2012.31
238 Mooney, C., Pollastri, G., Shields, D. C., & Haslam, N. J. (2012). Prediction of short linear
239   protein binding regions. *Journal of Molecular Biology*, *415*(1), 193–204.
240   doi:10.1016/j.jmb.2011.10.025
241 Nguyen Ba, A. N., Yeh, B. J., van Dyk, D., Davidson, A. R., Andrews, B. J., Weiss, E. L., &
242   Moses, A. M. (2012). Proteome-wide discovery of evolutionary conserved sequences in
243   disordered regions. *Science Signaling*, *5*(215), rs1. doi:10.1126/scisignal.2002515
244 Obenauer, J. C., Cantley, L. C., & Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of
245   cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, *31*(13),
246   3635–41. doi:10.1093/nar/gkg584
247 Pathak, S., Kundeti, V. K., Schiller, M. R., & Rajasekaran, S. (2013). A Structure Based
248   Algorithm for Improving Motifs Prediction. In A. Ngom, E. Formenti, J.-K. Hao, X.-M.
249   Zhao, & T. Laarhoven (Eds.), *Pattern Recognition in Bioinformatics* (Vol. 7986, pp. 242–
250   252). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-39159-0
251 Ponting, C. P. (1997). Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Science :*
252   *A Publication of the Protein Society*, *6*(2), 464–8. doi:10.1002/pro.5560060225
253 Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund,
254   K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R.,
255   Bateman, A., & Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids*
256   *Research*, *40*(Database issue), D290–301. doi:10.1093/nar/gkr1065
257 Ramu, C. (2003). SIRW: A web server for the Simple Indexing and Retrieval System that
258   combines sequence motif searches with keyword searches. *Nucleic Acids Research*, *31*(13),
259   3771–4. doi:10.1093/nar/gkg546
260 Ren, J., Gao, X., Jin, C., Zhu, M., Wang, X., Shaw, A., Wen, L., Yao, X., & Xue, Y. (2009).
261   Systematic study of protein sumoylation: Development of a site-specific predictor of
262   SUMOsp 2.0. *Proteomics*, *9*(12), 3409–3412. doi:10.1002/pmic.200800646
263 Sigrist, C. J. a, Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., &
264   Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and
265   annotation. *Nucleic Acids Research*, *38*(Database issue), D161–6. doi:10.1093/nar/gkp885
266 Weatheritt, R., & Jehl, P. (2012). iELM—a web server to explore short linear motif-mediated
267   interactions. *Nucleic Acids Research*, *40*(Web Server issue), W364–9.
268   doi:10.1093/nar/gks444
269 Zhang, J., Petit, C. M., King, D. S., & Lee, A. L. (2011). Phosphorylation of a PDZ domain
270   extension modulates binding affinity and interdomain interactions in postsynaptic density-95
271   (PSD-95) protein, a membrane-associated guanylate kinase (MAGUK). *The Journal of*
272   *Biological Chemistry*, *286*(48), 41776–85. doi:10.1074/jbc.M111.272583

Figure 1

Flowchart of the typical workflow of DoReMi

1) The user typically provides the motif description in 3 different ways: Regular Expression(s) define a motif by the basic format allowed in most implementations. Amino acids are defined by their one-letter code with "." standing for any. Multiple potential residues can be encoded by using square brackets, e.g. "[DE]". To quantify selected residues, we allow the basic operators "*", "+","{2}" or "{3,4}". To uses a PSSM for the motif search, the user simply provides a set of known binding motifs; these are used to calculate the amino acid distribution at each position, correcting for the overall amino acid distribution of the proteome. As a last option, users can provide results from other tools like SLiMsearch.
2) The second required input is the set of interacting domains. We provide the domains in PFAM-A. Proteins carrying any of the selected domains are defined as potential interaction partners. The highest scoring interacting protein for each motif instance is selected as potential binding partner.
3) The two scores are combined to rank each instance of the found motif.

Detailed performance numbers for the benchmarking.

The first number states the performance without contextual information while the second
number includes contextual information.

	Sensitivity	Specificity	AROC
SUMO – Split set - RegExp	0.50/0.50	0.98/0.99	-/0.46
SUMO – Split set - PSSM	0.59/0.62	0.96/0.96	0.82/0.84
SUMO - Partitioning	0.75/0.77	0.90/0.90	0.82/0.84
PDZ - Partitioning	0.43/0.49	0.90/0.90	0.70/0.72

Figure 2

Webinterface of DoReMi

A) The interface allows the user to search for Pfam domains by their accession, id or description. From the list of search results, the user can select the relevant domains. If necessary, multiple searches can be performed to select differently named domains.
B) The output page shows a brief summary of the analysis. This includes plots of score distributions for each score (motif, network and combined score) to aid in the selection of an appropriate score cut-off for downstream analysis of the results after downloading.