DegSampler: Gibbs sampling strategy for predicting E3-binding sites with position-specific prior information

Osamu Maruyama (✉ maruyama@design.kyushu-u.ac.jp)

Kyushu University <https://orcid.org/0000-0002-5760-1507>

Fumiko Matsuzaki

Kyushu University

Research article

Keywords: motif, substrate protein, E3 ubiquitin ligase, prior, posterior, disorder, collapsed, Gibbs sampling, DegSampler

Posted Date: November 6th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-101967/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

DegSampler: Gibbs sampling strategy for predicting E3-binding sites with position-specific prior information

Osamu Maruyama^{1*} and Fumiko Matsuzaki²

Abstract

Background: The ubiquitin-proteasome system is a pathway in eukaryotic cells for degrading polyubiquitin-tagged proteins through the proteasomal machinery to control various cellular processes and maintain intracellular homeostasis. In this system, the E3 ubiquitin ligase (hereinafter E3) plays an important role in selectively recognizing and binding to specific regions of its substrate proteins. The relationship between a substrate protein and its sites bound by E3s is not well understood. Thus, it is challenging to computationally identify such sites in substrate proteins.

Results: In this study, we proposed a collapsed Gibbs sampling algorithm called DegSampler (Degron Sampler) to identify the binding sites of E3s. DegSampler employs a position-specific prior probability distribution, based on the estimated information of the disorder-to-order region bound by any protein.

Conclusions: Our computational experiments show that DegSampler achieved 5 and 3.5 times higher the F-measure values of MEME and GLAM2, respectively. Thus DegSampler is the first model demonstrating an effective way of using estimated information on disorder-to-order binding regions in motif discovery. We expect our results to improve further as higher quality proteome-wide disorder-to-order binding region data become available.

Keywords: motif; substrate protein; E3 ubiquitin ligase; prior; posterior; disorder; collapsed; Gibbs sampling; DegSampler

Background

Eukaryotic cells have two major pathways for degrading proteins in order to control cellular processes and maintain intracellular homeostasis. One of the pathways is autophagy, a catabolic process that delivers intracellular components to lysosomes or vacuoles [1]. The other pathway is the ubiquitin-proteasome system, which degrades polyubiquitin-tagged proteins through the proteasomal machinery [2]. In the ubiquitin-proteasome system, an E3 ubiquitin ligase (hereinafter E3) selectively recognizes and binds to specific regions of the target substrate proteins. These binding sites are called *degrons* [3].

It is important to identify the E3s and substrate proteins that interact with one another to characterize various cellular mechanisms. Information on known degrons has been accumulated in the eukaryotic linear motif (ELM) resource for functional sites of proteins [4]. In the database, degon motifs are grouped into the “DEG” class. In addition, in the E3Net database [5], the relationship between E3s and their associated substrate proteins have been compiled. To date, only 55 of the 837 identified substrate proteins have been annotated with DEG motifs. In other words, 93% of the 837 known substrates are not annotated with any degrons. Thus, it is necessary to identify degrons of substrate proteins bound by specific E3s.

In literature, many sequence motif finders have been presented, such as MEME [6], Gibbs Motif Sampler [7], and GLAM2 [8]. MEME is a popular position-specific scoring matrix (PSSM) finding method based on expectation-maximization (EM) (see <http://meme-suite.org/index.html>). Gibbs Motif Sampler and GLAM2 are Gibbs sampling-based methods. One way of further improving motif identification is to exploit prior information on sequence positions. Narlikar and colleagues proposed a Gibbs sampling-based method called PRIORITY, which showed the effectiveness of the position-specific prior information

* Correspondence: maruyama@design.kyushu-u.ac.jp

¹ Faculty of Design, Kyushu University, Shiobaru, Minami-ku, Fukuoka, Japan

Full list of author information is available at the end of the article

based on nucleosome occupancy for motif discovery in deoxyribonucleic acid (DNA) [9]. In general, prior information can be easily integrated into the posterior probability distribution of motifs by representing it as a prior probability distribution. Inspired by their success, Bailey and colleagues extended the EM-based MEME algorithm to use position-specific priors and reported their effectiveness in identifying transcription factor-binding sites in yeast and mouse DNA sequences [10].

However, to the best of our knowledge, there is no model in which position-specific prior information is integrated in Gibbs sampling algorithms to identify motifs in *protein* sequences. Thus, we propose a Gibbs sampling-based method called DegSampler (Degron Sampler) to identify the locations of degrons in a given set of substrate protein sequences, which are bound by E3s.

One of the candidates of information sources for position-specific prior distributions is the disorderliness of each residue in a protein sequence. This inference is based on the observation that the E3-binding sites of substrate proteins are often located within the disordered regions of the protein structure [11, 12]. The disorderliness for a protein sequence can be calculated by IUPred1 [13] and IUPred2 [14].

Furthermore, we also examined the information on disorder-to-order binding regions in disordered proteins as a source of position-specific prior distribution. The estimated information can be obtained using prediction tools, including ANCHOR [15, 16], ANCHOR2 [14], MoRFpred [17], MFSPSSMpred [18], and DISOPRED3 [19]. Typically, the output of the tools is the probability of each residue being in a disorder-to-order region bound by any protein. These probabilities are expected to work better than disorderliness as position-specific prior information in identifying the locations of degrons.

We have evaluated the efficiency of position-specific prior information related to the disorderliness and disorder-to-order binding affinity of amino acid residues in identifying degrons. We have also compared the performance of DegSampler with those of other popular tools, viz. MEME and GLAM2, on real data sets. We found that DegSampler significantly outperforms the others.

Materials

Sequence data sets and known motifs

The E3Net database [5] identified 491 E3 ubiquitin ligases (E3s) associated with some known substrate proteins. In this research, we selected all E3s with at least three or more known substrate proteins. Consequently, we have extracted 123 E3s that are associated with 837 different substrate proteins from six

species: *Homo sapiens* (human), *S. cerevisiae* (yeast), *Arabidopsis thaliana* (mouse-ear cress), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Drosophila melanogaster* (fruit fly), and three viruses, human herpesvirus 1 (HHV11), human papillomavirus type 16 (HPV16), and human herpesvirus 8 type P (HHV8P). The average number of substrates assigned to an E3 is 8.5.

ELM is a database of short eukaryotic linear motifs that are manually curated from literature; this includes 3,026 experimentally validated motif occurrences on 1,886 proteins. These linear motifs are grouped into six functional classes: cleavage (CLV) sites, degradation (DEG) sites, docking (DOC) sites, ligand (LIG)-binding sites, post-translational modification (MOD) sites, and targeting (TRG) sites. The DEG class motifs are equivalent to degron motifs. Thus, the output of motif finders was evaluated using these DEG motifs. This class has 25 motifs, as shown in Table 1. Some of the motifs are too short to be predicted. Furthermore, only 36 of the 123 E3s in the E3Net database include at least one substrate that is annotated with a DEG motif in ELM (human 29; yeast 3; mouse-ear cress 1; mouse 1; rat 0; fruit fly 0; HHV11 1; HPV16 1; and HHV8P 0). The remaining 87 E3s have no DEG motif annotation on the substrate proteins. Thus, we used these 36 E3-specific sets of substrate proteins as input for motif finders to evaluate the performance of DegSampler.

Although E3Net has 837 E3-specific substrates, only 199 substrates are annotated with some ELM motifs. Furthermore, only 55 substrates are annotated with the DEG motifs. In other words, 93% of the 837 substrates were not annotated with any DEG motif. Therefore, the motif occurrences predicted by reliable motif finders are good candidates for degrons.

Source of position-specific prior information

Here, we describe the source of the position-specific prior information used in this study. We prepared position-specific priors derived from IUPred1 [13] and IUPred2 [14], which provides the probability of each residue belonging to a disordered region. IUPreds have three prediction types, "short disorder", "long disorder", and "structured domain." We selected the short and long disorder types. The outputs are denoted by IUPred1-short, IUPred1-long, IUPred2-short, and IUPred2-long.

Concerning the prior information on disorder-to-order binding regions of protein sequences, we used three prediction tools, ANCHOR [15], which is designated as ANCHOR1 to distinguish from its successor ANCHOR2, ANCHOR2 [14], fMoRFpred [20], and DISOPRED3 [19]. The output of these prediction tools

(a)

(b)

Figure 1 Example of ANCHOR1 outputs. The x-axis shows the position of a protein sequence. The y-axis represents the probability assigned to a sequence position by ANCHOR1. The known occurrence of a DEG motif is indicated by a tall and rectangular broken red line. The graph title has the UniProt accession identifier of the substrate protein and the ID of a DEG motif in the protein. (a) The substrate sequence is DCC_HUMAN (P43146). (b) The substrate sequence is MYC_HUMAN (P01106).

Table 1 DEG motifs in the ELM database. The “RegEx” column represents the regular expression of DEG motifs. The “Instances” column shows the number of known substrate proteins.

ELM Identifier	RegEx	Instances
DEG_APCC_DBOX_1	.R..L..[LIVM].	11
DEG_APCC_KENBOX_2	.KEN.	16
DEG_APCC_TPR_1	.[ILM]R\$	22
DEG_COP1_1	[STDE]1,3,0,2[TSDE].2,3VP[STDE]G0,1[FLIMVYPA]	12
DEG_CRL4_CDT2_1	[NQ]0,1..[ILMV][ST][DEN][FY][FY].2,3[KR]2,3[^DE]	6
DEG_CRL4_CDT2_2	[NQ]0,1..[ILMV]T[DEN][HMFY][FMY].2,3[KR]2,3[^DE]	1
DEG_Kelch_actinfilin_1	[AP]P[MV][IM]V	1
DEG_Kelch_Keap1_1	[DNS].[DES][TNS]GE	13
DEG_Kelch_Keap1_2	QD.DLGV	1
DEG_Kelch_KLHL3_1	E.EE.E[AV]DQH	4
DEG_MDM2_SWIB_1	F[^P]3W[^P]2,3[VIL]	5
DEG_Nend_Nbox_1	^M0,1[FYLIW][^P]	0
DEG_Nend_UBRbox_1	^M0,1[RK][^P].	0
DEG_Nend_UBRbox_2	^M0,1([ED]).	0
DEG_Nend_UBRbox_3	^M0,1([NQ]).	0
DEG_Nend_UBRbox_4	^M0,1(C).	8
DEG_ODPH_VHL_1	[IL]A(P).6,8[FLIVM].[FLIVM]	8
DEG_SCF_COI1_1	..[RK][RK].SL..F[FLM].[RK]R[HRK].[RK].	9
DEG_SCF_FBW7_1	[LIVMP].0,2(T)P..([ST])	6
DEG_SCF_FBW7_2	[LIVMP].0,2(T)P..E	2
DEG_SCF_SKP2-CKS1_1	..[DE].(T)P.K	3
DEG_SCF_TIR1_1	.[VLI]A[VLI]GWPP[VLI]...R.	24
DEG_SCF_TRCP1_1	D(S)G.2,3([ST])	19
DEG_SIAH_1	.P.A.V.P[^P]	9
DEG_SPOP_SBC_1	[AVP].[ST][ST][ST]	8

is the probability of each residue belonging to disorder-to-order binding regions. ANCHOR1 is a biophysics-based energy scoring method, and its successor, ANCHOR2, is an updated version. The fMoRFpred is a faster version of MoRFpred [17] with slightly compromised prediction accuracy, but has the ability to accept up to 2,000 protein sequences at a time, unlike MoRFpred which takes up to five protein sequences. Both methods are available only in their web servers. Since we have as many as 837 sequences, we selected fMoRFpred instead of MoRFpred. DISOPRED3 is a support vector machine (SVM)-based method that uses profile data and several sequence-derived features.

We assigned a probability of 0.5 to all the residues of an input sequence whenever a tool failed to make a prediction.

Here, we show good and bad examples of the ANCHOR1 output in Fig. 1. The substrate protein is shown in Fig. 1 (a) is DCC_HUMAN (P43146). There is a DEG_SIAH_1 occurrence from the position 1331 to 1339. As shown in the graph, the probabilities of ANCHOR1 around that region are approximately 0.9. However, the substrate protein shown in Fig. 1 (b) is MYC_HUMAN (P01106). Although there is a DEG_SCF_FBW7_1 occurrence from the position 55 to 62, the ANCHOR1 probabilities in the region are less than half.

Methods

Preliminaries

Let Σ be an alphabet of 20 distinct letters representing the 20 amino acids. A dataset of N sequences

over Σ is referred to as $\mathbf{X} = \{X_1, \dots, X_N\}$. The j th letter of the i th sequence of X_i is denoted by $X_{i,j}$. $|X_i|$ represents the length of X_i . The collection of all the indexes of the sequence letters is denoted by $U = \{(i, j) | i = 1, \dots, N, j = 1, \dots, |X_i|\}$. The complement of a subset V of U , that is, $U \setminus V$, is denoted by V^c . We introduce a letter-counting function $\mathbf{c}_\mathbf{X}$ for V that returns $\mathbf{c}_\mathbf{X}(V) = (freq_1, \dots, freq_{|\Sigma|})^T$ where $freq_\ell$ is the number of the ℓ th type letter of Σ in the letter set $\{X_{i,j} | (i, j) \in V\}$. For simplicity, we use $\mathbf{c}(V)$ instead of $\mathbf{c}_\mathbf{X}(V)$ because once a set of sequences are given, the set remains fixed.

We used the following notations of basic arithmetic operations as used in [21]. For the two vectors $\mathbf{u}$ and $\mathbf{v}$ of the same dimension n , we define $\mathbf{u} + \mathbf{v} = (u_1 + v_1, \dots, u_n + v_n)$, $\mathbf{u}/\mathbf{v} = (u_1/v_1, \dots, u_n/v_n)$, $\mathbf{u}^\mathbf{v} = u_1^{v_1} \dots u_n^{v_n}$, $x^\mathbf{v} = x^{v_1} \dots x^{v_n}$ for a real x , $|\mathbf{u}| = \sum_{i=1}^n |u_i|$, and $\mathbf{u}_{-i} = (u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_n)$. Note that $x^{|\mathbf{u}|} = x^\mathbf{u}$ if $u_i \geq 0$ for $i = 1, \dots, n$, and that $x^\mathbf{v} \cdot \mathbf{u}^\mathbf{v} = (x\mathbf{u})^\mathbf{v}$. For a gamma function Γ , Let $\Gamma(\mathbf{u}) = \Gamma(u_1) \dots \Gamma(u_n)$.

Motif model

We adopted two models for the distribution of the motif sites: one occurrence per sequence (OOPS) and zero or one occurrence per sequence (ZOOPS). The OOPS model [22] assumes that each sequence has exactly one occurrence of a motif. Let z_i be a hidden random variable that represents the starting position of the occurrence of a motif in X_i ; $z_i \in \{1, \dots, |X_i| - W + 1\}$ for the motif width W where $i = 1, \dots, N$. We denote all z_i

together as the vector $\mathbf{Z} = (z_1, \dots, z_N)$. The ZOOPS model [23] is a generalization of OOPS in which each sequence is permitted to have either zero or one occurrence of a motif. We denote the state where X_i does not have any motif occurrence by $z_i = 0$.

The site of a motif is often modeled as a single contiguous block of residues. However a more discriminative motif can be obtained by excluding some positions within the contiguous block of a motif. Liu *et al.* proposed a way of selecting a constant number of $R (\leq W)$ columns in an aligned block of width W to form a motif [21]. This model is called the fragmentation model of a motif. In this model, R selected positions of a short aligned block of width W are evaluated as parts of the motif (see Fig. 2). DegSampler optimizes R although it is kept constant in [21].

Such R columns selected as parts of the motif are specified by $\lambda = (\lambda_1, \dots, \lambda_W)$, where $\lambda_1 = \lambda_W = 1$, $\lambda_w \in \{0, 1\}$ for $w = 2, \dots, W - 1$, and $\sum_{w=1}^W \lambda_w = R$. $\lambda_w = 1$ (resp. 0) means that the w th column is evaluated as part of a motif (resp. background) model. We refer to W as the motif width. The region of length W covered by a motif is referred to as the occurrence of the motif. Our Gibbs sampling algorithm optimizes W and R by updating λ . After each update, λ is freshly indexed so that $\lambda_1 = \lambda_{W'} = 1$ for a new motif width W' . Thus, both the columns represent the two ends of the new motif. For $w \in \{w | \lambda_w = 1\}$, $\lambda(w)$ denotes the order of the w th element among the elements having the value 1. For example, for $\lambda = (1, 0, 1, 1)$, we have $\lambda(1) = 1$, $\lambda(4) = 2$, and $\lambda(5) = 3$.

We also show that the posterior probability distribution is tempered in the fragmentation model of ZOOPS and OOPS. In other words, the posterior probability distribution is parameterized with a temperature parameter of T . We first describe the untempered version.

Likelihood

The local alignment of the given sequences $\mathbf{X}$ occupied by a motif is represented as a matrix where the i th row corresponds to the i th sequence and the w th column

indicates the w th relative position within a motif occurrence. Our Gibbs sampling algorithm applies several operations to the various subsets of letter indexes of the given sequences. Then, we introduce matrix-like notations representing the subsets of indexes, as shown in Table 2. $I[\mathbf{Z}, \lambda]$ is the collection of the indexes with $z_i > 0$ and $\lambda_w = 1$. $I[\mathbf{Z}, w]$ is a subset of $I[\mathbf{Z}, \lambda]$, in which the elements are restricted to those in the w th column. $I[\mathbf{Z}_{-k}, \lambda]$ is a subset of $I[\mathbf{Z}, \lambda]$, where the indexes of the k th sequence are removed. $I[\mathbf{Z}_{-k}, w]$ is a subset of $I[\mathbf{Z}, w]$ where $(k, z_k + w - 1)$ is excluded. $I[z_k, \lambda]$ is a subset of $I[\mathbf{Z}, \lambda]$ in which the elements are limited to the indexes of the k th sequence. $I[z_k, w]$ is the index corresponding to the w th position of a motif occurrence in the k th sequence, *i.e.*, $(k, z_k + w - 1)$; if the sequence does not have any motif occurrence, the set is empty. Note that all of these notations can be used in both the OOPS and the ZOOPS models.

A position-specific scoring matrix (PSSM) is a *de facto* standard model of a likelihood function of a posterior probability distribution for the sequence motif identification. This model evaluates the conservation of each residue within the selected positions of a motif.

Let L be the number of letters, that is, $L = 20$ for amino acids. A PSSM is formulated as an $L \times R$ matrix $\theta_{1:R} = (\theta_1, \dots, \theta_R)$, where $\theta_r = (\theta_{1,r}, \dots, \theta_{L,r})^T$ is the parameter set of a categorical distribution. For $r = 1, \dots, R$, it holds that $\sum_{\ell=1}^L \theta_{\ell,r} = 1$ and $\theta_{\ell,r} \geq 0$ ($\ell = 1, \dots, L$). These categorical distributions are assumed to be independent of each other.

The letters outside the R selected positions in all the given sequences are assumed to be independently drawn from the same categorical distribution with the parameter θ_0 . This is called the background model.

The unnormalized likelihood function for the ZOOPS model of DegSampler, which is parameterized by $\mathbf{Z}, \lambda$, and $\theta_{0:W} = (\theta_0, \theta_1, \dots, \theta_W)$, is defined as

$$L(\mathbf{Z}, \lambda, \theta_{0:R} | \mathbf{X}) = \theta_0^{c(I[\mathbf{Z}, \lambda]^c)} \cdot \prod_{w: \lambda_w = 1} \theta_{\lambda(w)}^{c(I[\mathbf{Z}, w])}.$$

Thus the probability of $\mathbf{X}$ with $\mathbf{Z}, \lambda$, and $\theta_{0:R}$ can be given as

$$p(\mathbf{X} | \mathbf{Z}, \lambda, \theta_{0:R}) \propto L(\mathbf{Z}, \lambda, \theta_{0:R} | \mathbf{X}),$$

Note that this function is valid even in the OOPS model because the indexes of all the positions in the i th sequence with $z_i = 0$ are excluded in $I[\mathbf{Z}, \lambda]$ and included in $I[\mathbf{Z}, \lambda]^c$.

Prior distributions

Here, we formulate the prior distributions for the parameters $\theta_{1:R}$, $\mathbf{Z}$, and λ .

Table 2 Notations related to $\mathbf{Z}$.

Notation	Definition
$I[\mathbf{Z}, \boldsymbol{\lambda}]$	$\{(i, z_i + w - 1) i = 1, \dots, N, z_i \neq 0, \lambda_w = 1, w = 1, \dots, W\}$
$I[\mathbf{Z}, w]$	$\{(i, z_i + w - 1) i = 1, \dots, N, z_i \neq 0\}$
$I[\mathbf{Z}_{-k}, \boldsymbol{\lambda}]$	$\{(i, j) \in I[\mathbf{Z}, \boldsymbol{\lambda}] i \neq k\}$
$I[\mathbf{Z}_{-k}, w]$	$\{(i, j) \in I[\mathbf{Z}, w] i \neq k\}$
$I[z_k, \boldsymbol{\lambda}]$	$\{(k, j) \in I[\mathbf{Z}, \boldsymbol{\lambda}]\}$
$I[z_k, w]$	$\begin{cases} \emptyset & \text{if } z_k = 0 \\ \{(k, z_k + w - 1)\} & \text{otherwise} \end{cases}$

Prior distribution of $\boldsymbol{\theta}_{1:R}$

Let us recall that $\mathbf{c}(I[\mathbf{Z}, w])$ with $\lambda_w = 1$ is a sample from the categorical distributions with the parameters $\boldsymbol{\theta}_{\lambda(w)}$. A Dirichlet distribution is known to be a conjugate prior to categorical distribution and is defined as $\text{Dir}(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \boldsymbol{\theta}^{\boldsymbol{\alpha}-1}$, where the parameter $\boldsymbol{\alpha}$ is of dimension L ; $\mathbf{1}$ is the L -dimensional vector with all the elements having a value of 1. and the multivariate beta function $B(\boldsymbol{\alpha})$ is defined as follows:

$$B(\boldsymbol{\alpha}) = \frac{\Gamma(\boldsymbol{\alpha})}{\Gamma(|\boldsymbol{\alpha}|)}.$$

In this study, we used the same Dirichlet distribution parameter $\boldsymbol{\alpha}$ for the $R + 1$ random variables $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_R$. For $r = 0, \dots, R$, we define the following: $p(\boldsymbol{\theta}_r | \boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}_r | \boldsymbol{\alpha})$.

Prior distribution of $\mathbf{Z}$

We assume that for each $i = 1, \dots, N$, and $j = 1, \dots, |X_i|$, a prior probability $d_{i,j}$ is given; for example, it will be the outcome of ANCHOR1 for X_i . The probability is used as the likelihood of the position being part of a motif occurrence. We introduce the notation $m(z_i, \boldsymbol{\lambda}) = \{z_i + w - 1 | \lambda_w = 1, w = 1, \dots, W, z_i \neq 0\}$, which represent the positions included in a current motif occurrence specified by z_i and $\boldsymbol{\lambda}$. Then, we suppose that $p(z_i | \boldsymbol{\lambda})$, the probability of z_i given $\boldsymbol{\lambda}$, is proportional to

$$\prod_{j \in m(z_i, \boldsymbol{\lambda})} d_{i,j} \prod_{j=1 \text{ s.t. } j \notin m(z_i, \boldsymbol{\lambda})}^{X_i} (1 - d_{i,j})$$

if $z_i \neq 0$ and is proportional to

$$\prod_{j=1}^{|X_i|} (1 - d_{i,j})$$

if $z_i = 0$. This formulation is almost the same as in [9]. Furthermore, we have applied two modifications to the formulation. Note that the last term $\prod_{j=1}^{|X_i|} (1 - d_{i,j})$ is constant once $\mathbf{X}$ is given. Then, we use the terms that are obtained by dividing both cases by the constant.

Another modification is the exploitation of the coefficient g of this prior term to adjust the effectiveness of the prior information quantity, for which the term is raised to the power of g . The resulting prior probability distribution of z_i , given $\boldsymbol{\lambda}$ is proportional to

$$p(z_i | \boldsymbol{\lambda}) \propto \left(\prod_{j \in m(z_i, \boldsymbol{\lambda})} \frac{d_{i,j}}{1 - d_{i,j}} \right)^{g \cdot \delta_{z_i, 0}},$$

where $\delta_{i,j}$ is the Kronecker delta, equal to 1 if $i = j$ and 0 otherwise. Thus, the right-hand side is 1 when $z_i = 0$. The resulting prior distribution of $\mathbf{Z}$ is formulated as $p(\mathbf{Z} | \boldsymbol{\lambda}) = \prod_{i=1}^N p(z_i | \boldsymbol{\lambda})$ with different normalizing constants in the ZOOPS and the OOPS models.

Hyperprior distribution of $\boldsymbol{\lambda}$

We have discussed the hyperprior distribution of $\boldsymbol{\lambda}$. There are various fragmentation patterns in the ELM resource; therefore, it is reasonable to assume no bias for $\boldsymbol{\lambda}$. Then, we use the uniform distribution of $\boldsymbol{\lambda}$ specified by using the two parameters $R_{\min}$ and $W_{\max}$ such that $R_{\min} \leq R$ and $W \leq W_{\max}$, respectively.

Collapsed posterior distribution

The joint posterior probability distribution of the hidden random variables $\mathbf{Z}, \boldsymbol{\lambda}$, and $\boldsymbol{\theta}_{0:R}$, in the ZOOPS model is formulated as

$$\begin{aligned} & p(\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R} | \mathbf{X}) \\ & \propto L(\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R} | \mathbf{X}) \cdot p(\boldsymbol{\theta}_{0:R} | \boldsymbol{\alpha}) \cdot p(\mathbf{Z} | \boldsymbol{\lambda}) \cdot p(\boldsymbol{\lambda}). \end{aligned}$$

By integrating out $\boldsymbol{\theta}_{0:R}$, as shown in [21], we obtain the collapsed posterior distribution of $\mathbf{Z}$ and $\boldsymbol{\lambda}$ as

$$\begin{aligned} & p(\mathbf{Z}, \boldsymbol{\lambda} | \mathbf{X}) \\ & \propto B(\mathbf{c}(I[\mathbf{Z}, \boldsymbol{\lambda}]^c) + \boldsymbol{\alpha}) \cdot \prod_{w: \lambda_w = 1} B(\mathbf{c}(I[\mathbf{Z}, w]) + \boldsymbol{\alpha}) \\ & \quad \cdot \prod_{i=1}^N \prod_{w: \lambda_w = 1} \left(\frac{d_{i, z_i + w - 1}}{1 - d_{i, z_i + w - 1}} \right)^{g \cdot \delta_{z_i, 0}}. \end{aligned} \quad (1)$$

From this collapsed posterior distribution, we derive conditional distributions of z_k and λ_v to update their values. This collapsed posterior distribution works in both the OOPS and the ZOOPS models.

Update of a motif starting position z_k

To update z_k , we derive the conditional probability $p(z_k | \mathbf{Z}_{-k}, \boldsymbol{\lambda}, \mathbf{X})$ from Eq. (1), which is proportional to

$$\prod_{w: \lambda_w=1} \left\{ \left(\frac{\hat{\theta}_{\lambda(w)[-k]}}{\hat{\theta}_{0[-k]}} \right)^{c(I[z_k, w])} \cdot \left(\frac{d_{k, z_k+w-1}}{1 - d_{k, z_k+w-1}} \right)^{g \cdot \delta_{z_k, 0}} \right\},$$

where

$$\begin{aligned} \hat{\theta}_{\lambda(w)[-k]} &= \frac{c(I[\mathbf{Z}_{-k}, w]) + \alpha}{|c(I[\mathbf{Z}_{-k}, w]) + \alpha|}, \\ \hat{\theta}_{0[-k]} &= \frac{c(I[\mathbf{Z}_{-k}, \boldsymbol{\lambda}]^c) + \alpha}{|c(I[\mathbf{Z}_{-k}, \boldsymbol{\lambda}]^c) + \alpha|}. \end{aligned}$$

Note that the approximation scheme given in [21] has been used for the derivation. In particular, when $z_k = 0$, $p(z_k | \mathbf{Z}_{-k}, \boldsymbol{\lambda}, \mathbf{X})$ is proportional to 1 with the same normalizing constant, all $c(I[z_k, w])$ and $\delta_{z_k, 0}$ become 0.

Update of the motif column selection $\boldsymbol{\lambda}$

We have described how to update $\boldsymbol{\lambda}$. First, the current $\boldsymbol{\lambda}$ is extended by attaching $2 \cdot \Delta$ new columns to the left and right sides of $\boldsymbol{\lambda}$. The columns added to the left are indexed as $-\Delta+1, \dots, -1, 0$, and the columns added to the right are indexed as $W+1, \dots, W+\Delta$. Thus, the columns of the resulting $\boldsymbol{\lambda}$ are indexed continuously from $-\Delta+1$ to $W+\Delta$.

In the second step, the index $v \in \{-\Delta+1, \dots, W+\Delta\}$ was either selected randomly or in a fixed order. In the third step, we performed random sampling according to the conditional probability distribution of $\lambda_v \in \{0, 1\}$ with the current values of $\boldsymbol{\lambda}_{-v}$, $\mathbf{Z}$, and $\mathbf{X}$. The probabilities of the distribution are calculated as

$$\begin{aligned} &p(\lambda_v = 1 | \boldsymbol{\lambda}_{-v}, \mathbf{Z}, \mathbf{X}), \\ &\propto B(c(I[\mathbf{Z}, v]) + \alpha) \cdot \prod_{i=1}^N \left(\frac{d_{i, z_i+v-1}}{1 - d_{i, z_i+v-1}} \right)^{g \cdot \delta_{z_i, 0}}, \\ &p(\lambda_v = 0 | \boldsymbol{\lambda}_{-v}, \mathbf{Z}, \mathbf{X}) \\ &\propto \tilde{\theta}_0^{c(I[\mathbf{Z}, v])} \end{aligned}$$

where $\boldsymbol{\lambda}'$ is the same as $\boldsymbol{\lambda}$ but $\lambda_v = 1$, and

$$\tilde{\theta}_0 = \frac{c(I[\mathbf{Z}, \boldsymbol{\lambda}']^c) + \alpha}{|c(I[\mathbf{Z}, \boldsymbol{\lambda}']^c) + \alpha|}.$$

Both the right-hand side equations share the same normalizing constant.

If $\lambda_v = 1$ is selected in the sampling, and v is an index of the additional columns indexed as $-\Delta+1, \dots, -1, 0$, then the value of z_i is replaced with $z_i + v - 1$ for $z_i \neq 0$. Furthermore, the leftmost and the rightmost contiguous blocks of columns indexed as w with $\lambda_w = 0$ are removed. For example, if the content of $\boldsymbol{\lambda}$ after sampling with $\Delta = 3$ and $v = -1$ is (010110111000), then the updated $\boldsymbol{\lambda}$ is (10110111).

An advantage of this update is that a motif can easily shift to the left or the right on the substrate sequence. In addition, the R and W values are also varied. The default value of Δ is set to 3 in this analysis.

Simulated annealing

We also derived a tempered version of the collapsed posterior probability distribution, which was parameterized with temperature T as follows:

$$\begin{aligned} &p(\mathbf{Z}, \boldsymbol{\lambda} | \mathbf{X}) \\ &\propto B(c(I[\mathbf{Z}, \boldsymbol{\lambda}]^c) / T + \alpha) \\ &\cdot \prod_{w: \lambda_w=1} B(c(I[\mathbf{Z}, w]) / T + \alpha) \\ &\cdot \prod_{i=1}^N \prod_{w: \lambda_w=1} \left(\frac{d_{i, z_i+w-1}}{1 - d_{i, z_i+w-1}} \right)^{g \cdot \delta_{z_i, 0}}. \end{aligned} \quad (2)$$

The corresponding conditional probability distribution of z_k is as follows:

$$\begin{aligned} &p(z_k | \mathbf{Z}_{-k}, \boldsymbol{\lambda}, \mathbf{X}) \\ &\propto \prod_{w: \lambda_w=1} \left(\frac{\hat{\theta}_{\lambda(w)[-k]}}{\hat{\theta}_{0[-k]}} \right)^{c(I[z_k, w]) / T} \\ &\cdot \left(\frac{d_{k, z_k+w-1}}{1 - d_{k, z_k+w-1}} \right)^{g \cdot \delta_{z_k, 0}}, \end{aligned}$$

where

$$\begin{aligned} \hat{\theta}_{\lambda(w)[-k]} &= \frac{c(I[\mathbf{Z}_{-k}, w]) / T + \alpha}{|c(I[\mathbf{Z}_{-k}, w]) / T + \alpha|}, \\ \hat{\theta}_{0[-k]} &= \frac{c(I[\mathbf{Z}_{-k}, \boldsymbol{\lambda}]^c) / T + \alpha}{|c(I[\mathbf{Z}_{-k}, \boldsymbol{\lambda}]^c) / T + \alpha|}. \end{aligned}$$

The conditional probabilities of λ_v are given as follows:

$$\begin{aligned} &p(\lambda_v = 1 | \boldsymbol{\lambda}_{-v}, \mathbf{Z}, \mathbf{X}) \\ &\propto B(c(I[\mathbf{Z}, v]) / T + \alpha) \\ &\cdot \prod_{i=1}^N \left(\frac{d_{i, z_i+v-1}}{1 - d_{i, z_i+v-1}} \right)^{g \cdot \delta_{z_i, 0} / T} \end{aligned}$$

and

$$p(\lambda_v = 0 | \lambda_{-v}, \mathbf{Z}, \mathbf{X}) \propto \tilde{\theta}_0^{c(I[\mathbf{Z}, v])/T}$$

where λ' is the same as λ , but $\lambda_v = 1$ and

$$\tilde{\theta}_0 = \frac{c(I[\mathbf{Z}, \lambda']^c)/T + \alpha}{|c(I[\mathbf{Z}, \lambda']^c)/T + \alpha|}.$$

Algorithm of DegSampler

Here, we provide an overview of the algorithm of DegSampler in Algorithm 1. At the top level of DegSampler, it repeats the main procedure of the function called DegSampler2Core for three times to avoid falling into a local optimum. DegSampler2Core repeats the procedure $M = 300$ times. At $M_{sa} = 240$, the simulated annealing is started. Thus, the last 20% of all iterations are annealed. After each repetition, the value of the temperature parameter T is reduced by $sa\_ratio = 0.999$.

```

Function DegSampler():
  optLogPost  $\leftarrow -\infty$ 
  for  $i \leftarrow 1$  to 3 do
    DegSampler2Core()
  end
  return ( $\mathbf{Z}_{best}, \lambda_{best}, \text{optLogPost}$ )

Function DegSampler2Core():
   $T \leftarrow 1$ 
   $\lambda$  and  $\mathbf{Z}$  are initialized
  UpdateBestState( $\mathbf{Z}, \lambda$ )
  for  $m \leftarrow 1$  to  $M$  do
    if  $m = M_{sa}$  then
      ( $\log\text{Post}, \mathbf{Z}, \lambda$ ) = ( $\text{optLogPost}, \mathbf{Z}_{best}, \lambda_{best}$ )
    if  $m \geq M_{sa}$  then
       $T = sa\_ratio \cdot T$ 
    for  $n \leftarrow 1$  to  $N$  do
       $z_i \sim p(z_k | \mathbf{Z}_{-k}, \lambda, \mathbf{X})$ 
      UpdateBestState( $\mathbf{Z}, \lambda$ )
    end
    choose  $v$  randomly
     $\lambda_v \sim p(\lambda_v | \lambda_{-v}, \mathbf{Z}, \mathbf{X})$ 
    UpdateBestState( $\mathbf{Z}, \lambda$ )
  end

Function UpdateBestState( $\mathbf{Z}, \lambda$ ):
  logPost  $\leftarrow$  the R.H.S. of Eq. (2) for  $\mathbf{Z}$  and  $\lambda$ 
  if  $\text{optLogPost} < \log\text{Post}$  then
    optLogPost  $\leftarrow \log\text{Post}$ 
     $\mathbf{Z}_{best} \leftarrow \mathbf{Z}$ 
     $\lambda_{best} \leftarrow \lambda$ 

```

Algorithm 1: DegSampler algorithm. The three variables optLogPost , $\mathbf{Z}_{best}$, and λ_{best} , are supposed to be global, that is, they can be accessed from any place.

Performance measure

We measure the quality of predicted motif occurrences with respect to the DEG motif sites by precision, recall, and F-measure. In general, the precision score is the ratio of the number of true positives to the predicted positives, and the recall score is the ratio of the number of true positives to the number of real positives. The F-measure is the harmonic mean of precision and recall. In this work we define the positions of the DEG motif sites that are shared by the predicted motif occurrences as true positives. Similarly, for substrate protein sequences annotated by some DEG motifs, the positions of the predicted motif occurrences that are not covered by any of the DEG motif sites are false positives. Finally, the sequence positions of the DEG motif sites that are not covered by the predicted motif occurrences are false negatives. These quantities are calculated only on the protein sequences annotated by some DEG motif occurrences because the other sequences always have no true positives.

Let TP , FP , and FN be the number of true positives, that of false positives, and that of false negatives, respectively. We represent the precision, recall, and F-measure as

$$\begin{aligned}
 \text{precision} &= \frac{TP}{TP + FP}, \\
 \text{recall} &= \frac{TP}{TP + FN}, \\
 F &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.
 \end{aligned}$$

Default parameter set

We set the default value of a parameter as follows: The maximum motif width and the minimum number of selected columns are $W_{\max} = 9$ and $R_{\min} = 8$, respectively. The weight of the position-specific prior distribution, g , is 1. The pseudo count of residues at each position of a motif is set to be $\alpha = 1$, and the site distribution model is ZOOPS.

In this work, with a particular set of parameter values, DegSampler is executed five times.

Results

Performance comparison

We have shown the performance comparison of DegSampler with MEME and GLAM2. DegSampler used default parameter values: $W_{\max} = 9$, $R_{\min} = 8$, $g = 1$, $\alpha = 1$, $d_{i,j}$ is the ANCHOR1 probability, and the motif site distribution model is ZOOPS. We performed GLAM2 using the default values. MEME can optionally accept a prior distribution over the site locations of width $w(\geq 2)$. Thus we set the geometric

Figure 3 Performance comparison of DegSampler with MEME and GLAM2. DegSampler used default parameter values: $W_{\max} = 9$, $R_{\min} = 8$, $g = 1$, $\alpha = 1$, $\tilde{d}_{i,j}$ is the ANCHOR1 probability, and the motif site distribution model is ZOOPS. We performed GLAM2 using the default values: MEME was executed using the default values, except for the specified options: MEME-OOPS (resp. MEME-ZOOPS) represents MEME, where the site distribution model was set as OOPS (resp. ZOOPS); Furthermore, MEME-OOPS-Prior (resp. MEME-ZOOPS-Prior) represents MEME with position-specific prior information, using the specified site distribution model.

mean $\sqrt{\tilde{d}_{i,j} \cdot \tilde{d}_{i,j+1}}$ as the prior probability of the location of width 2 starting at the position j where $\tilde{d}_{i,j} = 0.9 \times d_{i,j} / \sum_j d_{i,j}$. This transformation in the ZOOPS model implies that the prior probability of X_i having no motif site is 0.1. For a motif-width w' different from w , MEME internally transforms a given prior distribution of site width w to the corresponding prior distribution of site width w' , using the geometric mean. Thus, the w value used does not seem to be a critical factor. We have set $w = 2$ in this study. MEME-OOPS-Prior (resp. MEME-ZOOPS-Prior) represents MEME with a position-specific prior using the specified site distribution model of OOPS (resp. ZOOPS). In addition, MEME-OOPS (resp. MEME-ZOOPS) represents MEME without any position-specific prior using the specified site distribution model of OOPS (resp. ZOOPS).

The graph of Fig. 3 shows that DegSampler outperformed the others. The best F-measure among the four methods based on MEME is 0.043 is attained by MEME-OOPS-Prior. The second best, 0.037, is slightly lower than the best and is marked by MEME-OOPS. This result implies that the given position-specific prior information of ANCHOR1 in MEME

cannot be sufficiently utilized. MEME shows the same trend in the ZOOPS model.

GLAM2 is the second best in F-measure. Note that GLAM2 is also a Gibbs sampling-based method as DegSampler. The disadvantage of GLAM2 is that there is no mechanism to utilize position-specific prior information. On the other hand, it is an advantage that the motif can be gapped. Actually, some of the DEG motifs in ELM are regular expressions whose matching sequences are variable in length like $F[\text{^P}\{3\}W[\text{^P}\{2,3\}[\text{VIL}]$ (DEG.MDM2_SWIB_1), in which the part of $[\text{^P}\{2,3\}$ matches any sequence excluding P of length 2 or 3.

DegSampler has 16 wins, 7 draws, and 13 losses against GLAM2 using 36 E3 input instances (see Supplement Table S1). Note that there are 27 instances of the 36 input sets, including substrate proteins annotated by DEG motifs with variable-length patterns. Even after being restricted during the 27 instances, DegSampler has 14 wins, 8 draws, and 5 losses. Under this restriction advantage to GLAM2, DegSampler outperformed GLAM2.

Comparing sources of position-specific prior information

Figure 4 Comparison of various sources of position-specific prior information for DegSampler. Height of the bar represents the mean of precision, recall, and F-measure scores. The error bars represent standard error. IUPred1 and 2 have an option for specifying the range of disordered regions, either short or long (default). The selected option is represented by attaching either "-short" or "-long" to "IUPred1" and "IUPred2."

We examined the predictability of DegSampler using IUPred1, IUPred2, ANCHOR1, ANCHOR2, fMoRFPred, and DISOPRED3, and then identifying the best source of position-specific prior information. These prediction methods are modeled in their own ways, thus the predictability of DegSampler would vary with each method.

Fig. 4 shows the precision, recall, and F-measure of DegSampler for each prior source as well as those with the case where the position-specific prior distribution is uniform, which is used as the baseline for this comparison. The other parameters are set to default.

We could see that ANCHOR1 outperforms the others. Although the successor, ANCHOR2, is reported to be superior to ANCHOR1 in the identification of disorder-to-order binding regions [14], the predecessor works better. The remaining, except the uniform prior cases are better than it. The scores of ANCHOR2 and fMoRFPred are slightly higher than those of IUPreds. This might be caused by the fact that ANCHOR2 and fMoRFPred are predictors for disorder-to-order regions, which are bound by any protein, and IUPreds are only predictors for disorder-to-order regions.

Effectiveness of position-specific prior

In this case, we examine how the position-specific prior of DegSampler affects predictability by controlling the weight of the prior, g . We have executed DegSampler with $g = 0, 0.1, 0.2, \dots, 2$, and $4, 6, \dots, 24$. Figure 5 shows the F-measure values of DegSampler over $g = 0, 0.1, 0.2, \dots, 2$. The other parameters have the default values. In accordance with the increase of g from 0, the F-measure drastically increases and converges around $g = 1.0$. The case where $g = 0$ is equivalent to the case in which the position-specific prior probability distribution is uniform. The mean of the F-measure value of this case is 0.017. This value is quite low; thus, the result indicates the usefulness of the position-specific prior information. The F-measure appears stable over the range from $g = 0.8$ to 2. This plateau continues in the case where $g = 18$ (data not shown), and then the F-measure value falls down to 0.182728, 0.168434, and 0.133904 at $g = 20, 22, 24$, respectively. This decrease is likely due to the biasness of the position-specific prior information.

Example of found motif occurrences

Here, we provide an example of the outputs of DegSampler from the 36 input sets. The E3 ubiquitin ligase of SYVN1_HUMAN is associated with 12 substrate proteins in the E3Net database. Table 3(a)

Figure 5 Effectiveness of position-specific prior information on the F-measure. Height of the bar represents the mean of the F-measure, and the error bar represents the standard error. The coefficient g of the position-specific prior distribution is in the range of $\{0, 0.1, \dots, 2\}$. The parameter values of DegSampler are as follows: $W_{\max} = 9$, $R_{\min} = 8$, $\alpha = 1$, $d_{i,j}$ is the ANCHOR1 probability, the motif site distribution model is ZOOPS.

shows the alignment of the motif occurrences predicted by DegSampler. Substrate protein, P04637, one of the 12 ones, is known to have an occurrence of motif DEG_MDM2.SWIB.1. Its known regular expression, $F[\sim P]\{3\}W[\sim P]\{2,3\}[VIL]$, matches the sequence, FSDLWKLL, of the region from position 19 to 27 of P04637, which are underlined in the table. The remaining substrate proteins do not have any degron annotation in the ELM database.

The ANCHOR1 probability for the 5th substrate protein P04637 (P53_HUMAN) is given in Fig. 6. There are many peaks, specifically five peaks whose probability is higher than 0.9. The site of motif DEG_MDM2.SWIB.1, located in the region from position 19 to 26, is overlapped with one of the five peaks, and its ANCHOR1 probability is approximately 1. DegSampler was able to successfully identify the site.

We found that the regular expression of motif DEG_MDM2.SWIB.1 also matches the sequence, FEETWATLL, of the region from 2464 to 2472 of P42858. However, this occurrence is not recorded as an instance of the DEG motif in the ELM database. In ad-

dition, this site is not predicted by DegSampler. The reason could be that the average of the ANCHOR1 probabilities of within that region is very low (0.249).

We further analyze this local alignment by classifying the 20 amino acid residues into the positively charged residues (RHK), the negatively charged residues (DE), the polar residues (NQSTY), and the non-polar residues (ACGILMFPPWV) [24]. The single-letter amino acid symbols of the categories are replaced with '+', '-', 'x', and 'o' in Table 3(b).

We focus on the indexed symbols of P04637, annotated by the DEG motif DEG_MDM2.SWIB.1. It has the indexed symbol of 'o' in the 1st, 4th, 5th, 7th, and 8th columns, and 'x' in the 2nd, '-' in the 3rd, and '+' in the 6th column, respectively. For each position of the local alignment, a binomial test is applied to evaluate whether the number of the indexed symbol of P04637 in the column is unexpectedly rare under the frequency of each amino acid over all the substrate protein sequences used in this model, as shown in Supplement Table S2. This amino acid frequency distribution

Table 3 (a) Example of an output of DegSampler for E3 SYVN1_HUMAN. The E3 SYVN1_HUMAN in E3Net has the following substrate proteins assigned to it: O00141 (SGK1_HUMAN), O15354 (GPR37_HUMAN), O75460 (ERN1_HUMAN), P04035 (HMDH_HUMAN), P04637 (P53_HUMAN), P10636 (TAU_HUMAN), P10909 (CLUS_HUMAN), P42858 (HD_HUMAN), Q86TM6 (SYVN1_HUMAN), Q8WZ42 (TITIN_HUMAN), Q92542 (NICA_HUMAN), and Q9UKV5 (AMFR_HUMAN). Among them, only the 5th substrate, P04637 (P53_HUMAN), is annotated with a DEG motif, DEG_MDM2_SWIB_1. Column “start” shows the starting position of the motif occurrence. An asterisk on the top of a column indicates that the column was selected as part of a motif. The residues covered by a motif occurrence and the known DEG motifs simultaneously are underlined. (b) The 20 amino acid residues can be classified into the positively charged residues (RHK), the negatively charged residues (DE), the polar residues (NQSTY), and the non-polar residues (ACGILMFPPWV), which are denoted as ‘+’, ‘-’, ‘x’, and ‘o’, respectively. Each residue is replaced with a corresponding indexed symbol.

(a)									
substrate	start	*	*	*	*	*	*	*	*
O00141	21	V	A	I	L	I	A	F	M
O15354	141	A	L	Q	L	F	L	Q	I
O75460	453	L	L	I	G	W	V	A	F
P04035	330	L	L	L	A	V	K	Y	I
P04637	19	<u>F</u>	<u>S</u>	<u>D</u>	<u>L</u>	<u>W</u>	<u>K</u>	<u>L</u>	<u>L</u>
P10636	187	L	L	K	H	Q	L	L	G
P10909	5	L	L	L	F	V	G	L	L
P42858	1	M	A	T	L	E	K	L	M
Q86TM6	519	L	Q	I	N	Q	Y	L	T
Q8WZ42	11823	K	I	F	Q	L	K	A	I
Q92542	21	L	S	F	C	V	L	L	A
Q9UKV5	626	R	R	M	L	A	A	A	A

(b)									
substrate	start	*	*	*	*	*	*	*	*
O00141	21	o	o	o	o	o	o	o	o
O15354	141	o	o	x	o	o	o	x	o
O75460	453	o	o	o	o	o	o	o	o
P04035	330	o	o	o	o	o	+	x	o
P04637	19	<u>o</u>	<u>x</u>	<u>-</u>	<u>o</u>	<u>o</u>	<u>+</u>	<u>o</u>	<u>o</u>
P10636	187	o	o	+	+	x	o	o	o
P10909	5	o	o	o	o	o	o	o	o
P42858	1	o	o	x	o	-	+	o	o
Q86TM6	519	o	x	o	x	x	x	o	x
Q8WZ42	11823	+	o	o	x	o	+	o	o
Q92542	21	o	x	o	o	o	o	o	o
Q9UKV5	626	+	+	o	o	o	o	o	o
p-value		0.0021	0.3781	0.4468	0.0563	0.0139	0.0165	0.0021	1.5311e-04

is almost similar to that of the UniProtKB/Swiss-Prot UniProt release 2020.04 [25].

For example, in the 1st column, the number of observed symbol ‘o’ is 10 and the frequency of non-polar residues under the null hypothesis is 0.4809. The resulting p-value is 0.0021.

The conservation of these symbols is statistically significant in the 1st and 5th to 8th columns, at 0.05 level of significance.

Discussion

Narlikar *et al.* discussed how to formulate a position-specific prior distribution derived from nucleosome occupancy data to identify transcription factor-binding sites, and proposed two types of position-specific prior distributions: *simple* and *discriminative*. These are used in their Gibbs sampling-based motif finder, PRIORITY [9], as the prior distribution for the latent variable representing start positions of the motif occurrences. The simple prior distribution of a DNA sequence is calculated from the sequence like our

position-specific prior distribution. On the other hand, the discriminative one is derived from positive sequences (bound by a profiled transcription factor) as well as negative sequences (not bound by any profiled transcription factors). Inspired by their work, in addition to our position-specific prior probability, $p(z_i|\lambda)$, we have also evaluated an alternative position-specific prior distribution, derived from the positive sequences (E3-specific substrate sequences) and the negative sequences (the other sequences of the E3Net database).

We formulated the distribution as follows: Let $\mathbf{X}$ and $\mathbf{Y}$ be the E3-specific set of substrate protein sequences and the remaining substrate protein sequences in the E3Net database, respectively. We denote the subsequence of X_i (Y_i , resp.) of length W , starting at j , i.e., $X_{i,j} \cdots X_{i,j+W-1}$ by $X_{i,j}^{(W)}$ ($Y_{i,j}^{(W)}$, resp.). Furthermore, average of the ANCHOR1 probabilities in the subsequence is $A(X_{i,j}^{(W)}) = \frac{1}{W} \prod_{k=j}^{j+W-1} d_{i,k}$. In the same way, we define $A(Y_{i,j}^{(W)})$. Then, we define a discriminative position-specific prior probability dis-

Figure 6 The output of ANCHOR1 for the substrate protein P53.HUMAN (P04637). The x-axis shows the position of a protein sequence. The y-axis represents the probability assigned to a sequence position by ANCHOR1. A DEG.MDM2.SWIB.1 site is located in the region from position 19 to 26. The location is indicated by a vertically long rectangle. The entry name of the substrate protein and the DEG motif ID associated with the protein are given above the graph.

tribution:

$$p(z_i) \propto \mathcal{A}_{\mathcal{D}}(z_i) \prod_{j=1, j \neq z_i}^{|X_i|-W+1} (1 - \mathcal{A}_{\mathcal{D}}(j))$$

$$\propto \frac{\mathcal{A}_{\mathcal{D}}(z_i)}{1 - \mathcal{A}_{\mathcal{D}}(z_i)}$$

if $z_i \neq 0$ and $p_{i,z_i} \propto 1$, otherwise with the same normalizing constant, where $\mathcal{A}_{\mathcal{D}}(z_i)$ is

$$\frac{\sum_{(k,\ell): X_{k,\ell}^{(W)} = X_{i,z_i}^{(W)}} A(X_{k,\ell}^{(W)})}{\sum_{(k,\ell): X_{k,\ell}^{(W)} = X_{i,z_i}^{(W)}} A(X_{k,\ell}^{(W)}) + \sum_{(k,\ell): Y_{k,\ell}^{(W)} = X_{i,z_i}^{(W)}} A(Y_{k,\ell}^{(W)})}.$$

Note that in this formulation, the fragmentation model is not used for the sake of simplicity of the formulation.

We used this discriminative prior in DegSampler with the default parameters where $W_{\max} = 9$, $R_{\min} = 8$, $g = 1$, $\alpha = \mathbf{1}$, $d_{i,j}$ is the ANCHOR1 probability, and ZOOPS is the motif site distribution model. Unlike the case of PRIORITY, the performance of DegSampler with the discriminative prior is worse than the original one. The precision, recall, and F-measure with their standard errors are 0.026 (0.009), 0.026 (0.009), and 0.026 (0.009), respectively.

The reason of the unsuccessful result could be as follows. As mentioned in the Materials section, the average number of substrates assigned to an E3 is 8.5.

In addition, the length of those protein sequences are typically several hundreds. Under this situation, many extracted amino acid sequences of length 8 or 9 have one or fewer occurrences over the positive and negative sequences. As a result, the discriminative prior does not work well; probably, because an input protein sequence consists of 20 types of letters, which is significantly higher than the 4 types in a DNA sequence. A larger input set of protein sequences are needed to use the discriminative position-specific prior distribution.

Conclusions

We have proposed a Gibbs sampling-based motif finder, DegSampler, specialized for E3 binding sites. In this model, we confirmed that the probability of each residue being in a disorder-to-order region bound by any protein is a useful source to form a position-specific prior distribution for the degons. DegSampler finds a good combination of input protein sequence sites of with relatively higher position-specific prior probabilities. Simultaneously, we also found that such probabilities of prediction tools are quite different (We used ANCHOR1, ANCHOR2, fMoRFpred, and DISOPRED3 in this study). We evaluated simple and discriminative position-specific prior formulations. DegSampler outperformed MEME, which uses a different formulation of the position-specific prior distribution. However, there is much room for improvement in identifying degons because the F-measure of DegSampler is still low.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials

The datasets used and analyzed during the current study are available in our repository,
<http://www.design.kyushu-u.ac.jp/~maruyama/software.html>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by JSPS KAKENHI Grant Number JP17K00407.

Authors' contributions

OM participated in the conceptualization, formulation of the mathematical model and the algorithm, implementation of the software, computational experiments, analysis of the results, and in the drafting of the manuscript of the project. FM participated in the conceptualization, analysis of the results, and improvement of the manuscript of the project. All the authors read and approved the final manuscript.

Acknowledgements

The authors thank the anonymous reviewers for their careful reading of the submitted version of our manuscript and their many insightful comments and suggestions. We would also like to thank Editage (www.editage.com) for English language editing.

Author details

¹Faculty of Design, Kyushu University, Shiobaru, Minami-ku, Fukuoka, Japan. ²Medical Institute of Bioregulation, Kyushu University, Maidashi, Higashi-ku, Fukuoka, Japan.

References

- Mizushima, N., Komatsu, M.: Autophagy: renovation of cells and tissues. *Cell* **147**, 728–741 (2011)
- Weissman, A.M., Shabek, N., Ciechanover, A.: The predator becomes the prey: regulating the ubiquitin system by ubiquitylation and degradation. *Nat Rev Mol Cell Bio.* **12**, 605–620 (2011)
- Zheng, N., Shabek, N.: Ubiquitin ligases: structure, function, and regulation. *Annu Rev Biochem* **86**, 129–157 (2017)
- Dinkel, H., *et al.*: ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* **44**, 294–300 (2016)
- Han, Y., Lee, H., Park, J.C., Yi, G.-S.: E3Net: A system for exploring E3-mediated regulatory networks of cellular functions. *Mol Cell Proteomics* **11**, 111–014076 (2012)
- Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36. AAAI Press, Menlo Park, California, USA (1994)
- Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., Wootton, J.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993)
- Frith, M.C., Saunders, N.F.W., Kobe, B., Bailey, T.L.: Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol.* **4**, 1000071 (2008)
- Narlikar, L., Gordân, R., Hartemink, A.J.: A nucleosome-guided map of transcription factor binding sites in yeast. *PLOS Computational Biology* **3**(11), 215 (2007)
- Bailey, T.L., Bodén, M., Whittington, T., Machanick, P.: The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* **11**, 179 (2010)
- Guharoy, M., Bhowmick, P., Sallam, M., Tompa, P.: Tripartite degrons confer diversity and specificity on regulated protein degradation in the ubiquitin-proteasome system. *Nat Commun.* **7**, 10239 (2016)
- Guharoy, M., Bhowmick, P., Tompa, P.: Design principles involving protein disorder facilitate specific substrate selection and degradation by the ubiquitin-proteasome system. *J Biol Chem.* **291**, 6723–31 (2016)
- Dosztányi, Z., Csizsók, V., Tompa, P., Simon, I.: The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol.* **347**, 827–839 (2005)
- Mészáros, B., Erdős, G., Dosztányi, Z.: IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**(W1), 329–337 (2018)
- Mészáros, B., Simon, I., Dosztányi, Z.: Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol.* **5**(5), 1000376 (2009)
- Dosztányi, Z., Mészáros, B., Simon, I.: ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009)
- Disfani, F.M., Hsu, W.-L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N., Kurgan, L.: MoRFPred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* **28**(12), 75–83 (2012)
- Fang, C., Noguchi, T., Tominaga, D., Yamana, H.: MFSPSMPred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. *BMC Bioinformatics* **14**, 300 (2013)
- Jones, D.T., Cozzetto, D.: DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**(6), 857–63 (2015)
- Yan, J., Dunker, A.K., Uversky, V.N., Kurgan, L.: Molecular recognition features (morfs) in three domains of life. *Molecular BioSystems* **12**, 697–710 (2016)
- Liu, J.S., Neuwald, A.F., Lawrence, C.E.: Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J Am Stat Assoc.* **90**, 1156–1170 (1995)
- Lawrence, C.E., Reilly, A.A.: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**, 41–51 (1990)
- Bailey, T.L., Elkan, C.: The value of prior knowledge in discovering motifs with MEME. In: *Proc Int Conf Intell Syst Mol Biol*, vol. 3, pp. 21–29. AAAI Press, USA (1995)
- Alberts, B., Johnson, A.D., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P.: *Molecular Biology of the Cell* (Sixth Edition). W. W. Norton & Company, New York (2014)
- UniProtKB/Swiss-Prot UniProt release 2020_04.
<https://www.uniprot.org/statistics/Swiss-Prot>

Figures

Figure 1

Example of ANCHOR1 outputs. The x-axis shows the position of a protein sequence. The y-axis represents the probability assigned to a sequence position by ANCHOR1. The known occurrence of a DEG motif is indicated by a tall and rectangular broken red line. The graph title has the UniProt accession identifier of the substrate protein and the ID of a DEG motif in the protein. (a) The substrate sequence is DCC HUMAN (P43146). (b) The substrate sequence is MYC HUMAN (P01106).

Figure 2

Example of a fragmentation model of a motif. The six rectangles represent the columns selected as parts of the motif. In this example, $\hat{a}$ is (1; 0; 1; 1; 0; 1; 0; 1; 0; 1), which gives $R = 6$ and $W = 10$.

Figure 3

Performance comparison of DegSampler with MEME and GLAM2. DegSampler used default parameter values: $W_{max} = 9$, $R_{min} = 8$, $g = 1$, $a = 1$, d_{ij} is the ANCHOR1 probability, and the motif site distribution model is ZOOPS. We performed GLAM2 using the default values: MEME was executed using the default values, except for the specified options: MEME-OOPS (resp. MEME-ZOOPS) represents MEME, where the site distribution model was set as OOPS (resp. ZOOPS); Furthermore, MEME-OOPS-Prior (resp. MEME-

ZOOPS-Prior) represents MEME with position-specific prior information, using the specied site distribution model.

Figure 4

Comparison of various sources of position-specific prior information for DegSampler. Height of the bar represents the mean of precision, recall, and F-measure scores. The error bars represent standard error. IUPred1 and 2 have an option for specifying the range of disordered regions, either short or long (default). The selected option is represented by attaching either "-short" or "-long" to "IUPred1" and "IUPred2."

Figure 5

Effectiveness of position-specific prior information on the F-measure. Height of the bar represents the mean of the F-measure, and the error bar represents the standard error. The coefficient g of the position-specific prior distribution is in the range of $(0; 0.1; \dots; 2)$. The parameter values of DegSampler are as follows: $W_{\max} = 9$, $R_{\min} = 8$, $a = 1$, d_{ij} is the ANCHOR1 probability, the motif site distribution model is ZOOPS.

Figure 6

The output of ANCHOR1 for the substrate protein P53 HUMAN (P04637). The x-axis shows the position of a protein sequence. The y-axis represents the probability assigned to a sequence position by ANCHOR1. A DEG MDM2 SWIB 1 site is located in the region from position 19 to 26. The location is indicated by a vertically long rectangle. The entry name of the substrate protein and the DEG motif ID associated with the protein are given above the graph.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [table.xlsx](#)