RESEARCH

DegSampler: An improved Gibbs sampling strategy for predicting E3 binding sites

Osamu Maruyama[1*] and Fumiko Matsuzaki[2]

Abstract

Background: The ubiquitin-proteasome system is a mechanism in eukaryotic cells for degrading polyubiquitin-tagged proteins through the proteasomal machinery to control numerous cellular processes and maintain intracellular homeostasis. In this system, the E3 ubiquitin ligase (hereinafter E3) plays an important role of selectively recognizing and binding specific regions of its substrate proteins. The relationship between substrate proteins and the sites at which they bind to E3s is not yet clear. Therefore, we need to computationally identify such sites from the substrate proteins of E3s. For this motif identification, we proposed Degron Sampler 1 (DegSampler1), a collapsed Gibbs sampling algorithm using position-specific prior information.

Results: In this research, we proposed a new collapsed Gibbs sampling algorithm called DegSampler (Degron Sampler) by designing a new position-specific prior probability distribution and sampling strategies, especially the updating of motif columns.

Conclusions: We have shown that in the computational experiments, DegSampler achieved a 17% higher F-measure value than DegSampler1. This proves the importance of the design of the position-specific prior probability distribution and the sampling strategy of the Gibbs sampling algorithms.

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

Background

Eukaryotic cells have two major pathways for degrading proteins for controlling cellular processes and maintaining intracellular homeostasis. One pathway is autophagy, a catabolic process that delivers intracellular components to the lysosome or the vacuole. The other pathway is the ubiquitin-proteasome system, which is a mechanism for degrading polyubiquitin-tagged proteins through the proteasomal machinery [1]. In the ubiquitin-proteasome system, an E3 ubiquitin ligase (hereinafter E3) selectively recognizes and binds specific regions of its target proteins, called substrate proteins. The characteristic representations compiled from the protein sequences of such regions bound by the same E3 protein are called sequence motifs, especially *degrons* for the E3 binding sites.

It is important to identify which E3s and substrate proteins interact with one another to characterize various cellular mechanisms. Data on known degrons have been accumulated in the eukaryotic linear motif (ELM) resource for the functional sites in proteins [2]. Degrons are grouped into the class called “DEG” in the database. In addition, in the E3Net database [3], the data on the relationships between E3s and their associated substrate proteins have been compiled. To date, of the 837 such substrate proteins that have been identified, only 55 substrate proteins are annotated with DEG motifs. In other words, 93% of the 837 known substrates are not annotated with any degrons. This indicates the need to find degrons of the substrate proteins bound by particular E3s.

In literature many sequence motif finders have been presented, such as Multiple EM for Motif Elicitation (MEME) [4], Gibbs Motif Sampler [5], and Gapped Local Alignment of Motifs (GLAM2) [6]. MEME is a popular PSSM (Position-specific scoring matrix) finding method based on expectation-maximization (EM) (see `http://meme-suite.org/index.html`). Gibbs Motif Sampler and GLAM2 are the Gibbs sampling-based methods. One way of further improving the motif finding is to exploit the prior information on the sequence positions. Narlikar and colleagues proposed a Gibbs sampling-based method called PRIORITY, and they showed the effectiveness of the position-specific prior

[*] 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

information for the motif discovery of deoxyribonucleic acid (DNA) [7]. 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 for identifying the transcription factor binding sites in yeast and mouse DNA sequences [8].

However, to our best knowledge, there is no work in which position-specific prior information is integrated in Gibbs sampling algorithms in sequence motif finding. Thus, we in this paper propose a Gibbs sampling-based method called DegSampler (Degron Sampler) to identify the locations of motif occurrences of a given set of substrate protein sequences bound by an E3.

One of the features of this method is to exploit the degree of disorderness of each residue of a protein sequence as position-specific prior information for motif occurrences. This approach is based on the observation that the E3 binding sites of the substrate proteins are often located within the disordered regions of the protein structure [9, 10].

Furthermore, there are several prediction tools for disorder-to-order binding regions in disordered proteins, including ANCHOR1[11, 12], MoRFpred [13], MFSPSSMpred [14], and DISOPRED3 [15]. Typically, the output of them is the probability of each residue being in a disorder-to-order binding region. These probabilities are expected to work better than disorderness as position-specific prior information in identifying the locations of motif occurrences.

In addition to the feature of the exploitation of position-specific prior, there are the following important features in DegSampler.

- A fragmentation motif means a consequtive pattern of width W such that some specified internal positions within the whole pattern that are evaluated as part of motif, and the remainings are done as the background model. It is assumed that both end positions are always included in the motif model. DegSampler adopted this fragmentation motif model and optimizes the number of such selected positions within a found motif.
- Furthermore, DegSampler can also optimize the width of a motif in fragmentation model.
- The OOPS (one occurrence per sequence) [16] and ZOOPS (zero or one occurrence per sequence) [17] are motif occurrence models. Both are available in DegSampler.
- The posterior probability distribution was also parameterized using a temperature parameter. As a result, simulated annealing was applicable in DegSampler.

We evaluate how efficient position-specific prior information related to disorderness of amino acid residues is in finding degrons. We also compare the performance of DegSampler with those of other popular tools on real data sets. We found that DegSampler drastically outperforms the others.

Materials

Sequence data sets and known motifs

The E3Net database [3] identified 491 E3 ubiquitin ligases (E3s) associated with some known substrate proteins. In this research, we choosed all E3s with three or more known substrate proteins. As a result, we have obtained 123 such E3s, which were 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 the three viruses, human herpesvirus 1 (HHV11), human papillomavirus type 16 (HPV16), and human herpesvirus 8 type P (HHV8P). The number of substrates assigned to an E3 averages 8.5.

ELM is a database of short eukaryotic linear motifs that are manually curated from the 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 degrons. Thus, the output of the motif finders was evaluated by using these DEG motifs. This class has 25 motifs, as shown in Table 1. As seen, some of the motifs looks too short to be predicted. Furthermore, only 36 E3s among 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). Namely, the remaining 87 E3s have no DEG motif annotation on the substrate proteins. Then we used these 36 E3-specific sets of substrate proteins as the input for the motif finders to evaluate 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 the degradation of functional regions, especially degrons.

Table 1 DEG motifs in the ELM database. The "RegEx" column gives the regular expression representation 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	.[VLIA][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

Source of position-specific prior information

We here describe the source of the position-specific prior information used in this work. We prepare a position-specific prior derived from IUPred2 [18], which provides the probability of each residue belonging to a disordered region. IUPred2 has three prediction types, "short disorder", "long disorder", and "structured domain." Thus, we have chosen the short and long disorder types. Those outputs are denoted by IUPred2-short and IUPred2-long.

Concerning the prior information on disorder-to-order binding regions of protein sequences, we use three prediction tools, ANCHOR [11], which we hereafter call ANCHOR1 to distinguish it from the successor ANCHOR2, ANCHOR2 [18], and fMoRFpred [19]. ANCHOR1 is a biophysics based energy scoring method, and the successor, ANCHOR2, is an updated version of the predecessor. fMoRFpred is a fast version of MoRFpred [13] with slightly compromising the prediction accuracy, but has an ability to accept 2,000 protein sequences at a time, though MoRFpred takes up to five protein sequences. Both methods are available only in their web servers. Because we have as many as 837 sequences as mentioned above, we chose fMoRFpred instead of MoRFpred.

The outputs of the four prediction tools we use in this work are all formulated to be the probability of each residue being a disorder-to-ordered region for IUPred2 and being a disorder-to-ordered binding region for the others.

Two examples of the ANCHOR1 output are given in Fig. 1. The substrate protein of (a) is DCC_HUMAN (P43146). There is DEG_SIAH_1 site on the region 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 of (b) is MYC_HUMAN (P01106). Though there is a DEG_SCF_FBW7_1 site on the region from the position 55 to 62, the ANCHOR1 probabilities on 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 $\mathbf{X} = \{X_1, \ldots, X_N\}$. The jth letter of the ith sequence of X_i is denoted by $X_{i,j}$. $|X_i|$ represents the length of X_i. The collection of all indexes of the letters of the sequences is denoted by $U = \{(i,j) | i = 1, \ldots, N, j = 1, \ldots, |X_i|\}$. The complement of a subset V of U, that is, $U \setminus V$, is denoted by V^c. We introduce a counting function $\boldsymbol{c}_{\mathbf{X}}$ that returns $\boldsymbol{c}_{\mathbf{X}}(V) = (\mathit{freq}_1, \ldots, \mathit{freq}_{|\Sigma|})^T$ where freq_ℓ is the number of the ℓth type letter of Σ in the letter set $\{X_{i,j} | (i,j) \in V\}$. For simplicity, we use $\boldsymbol{c}(V)$ instead of $\boldsymbol{c}_{\mathbf{X}}(V)$ because once a set of sequences is given, the set remains fixed.

Motif model

We adopt 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 [16] assumes that each sequence has exactly one

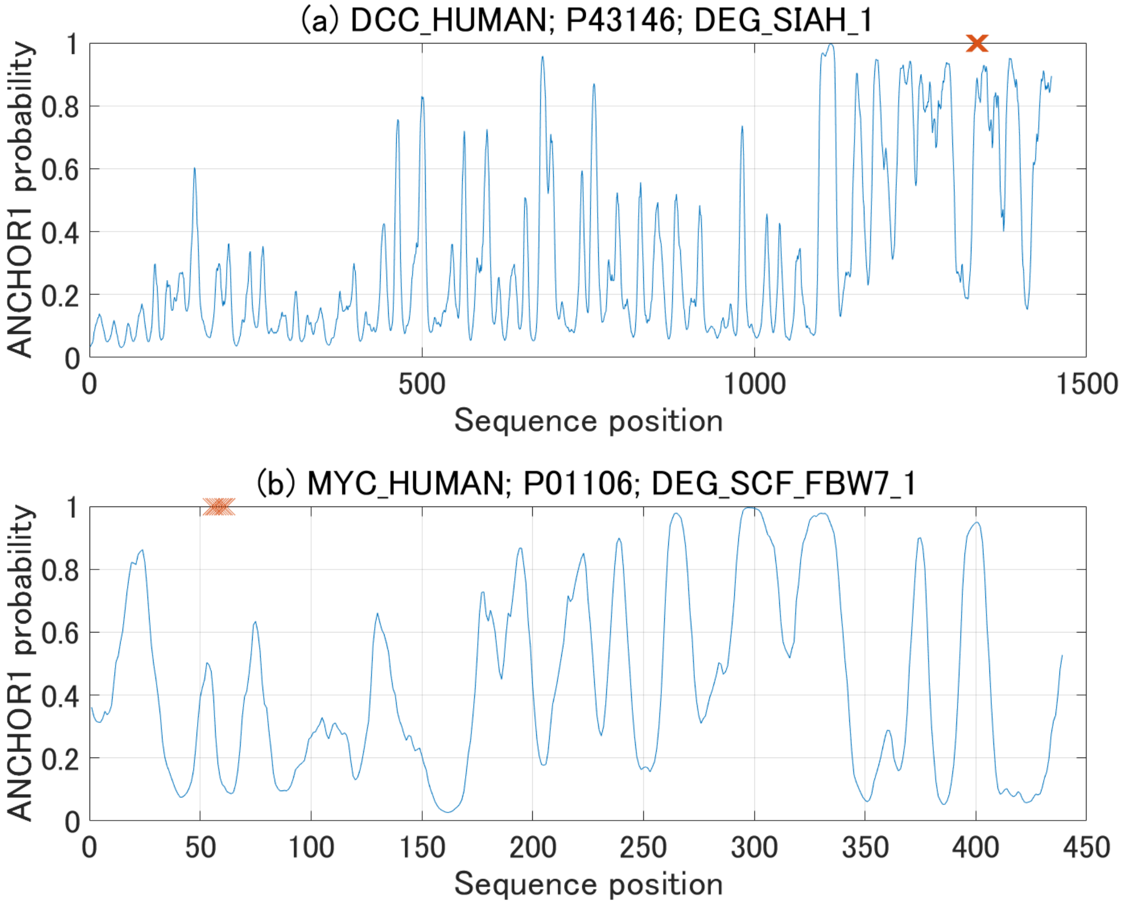

Figure 1 Example of a source information, $pos_{i,j}$ for position prior. The x-axis shows the position of a protein sequence. The y-axis represents the probability assigned to a sequence position by ANCHOR1. The known DEG motif occurrence is specified by the symbol 'x' on their sequence positions at the height of 1. The graph title consists of the entry name and the UniProt accession of the substrate protein and the IDs of the DEG motifs on the protein. (a) The substrate sequence is DCC_HUMAN (P43146). (b) The substrate sequence is MYC_HUMAN (P01106).

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; $z_i \in \{1, \ldots, |X_i| - W + 1\}$ for the motif width W where $i = 1, \ldots, N$. We denote all z_i together as the vector $\mathbf{Z} = (z_1, \ldots, z_N)$. The ZOOPS model [17] 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 $z_i = 0$.

A site of a motif is often modeled as a single contiguous block over residues. However a more discriminative motif can sometimes 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 [20]. This model is called a fragmentation model of a motif.

Such R columns selected as part of a motif are specified by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_W)$, where $\lambda_1 = \lambda_W = 1$, $\lambda_w \in \{0, 1\}$ for $w = 2, \ldots, W - 1$, and $\sum_{w=1}^{W} \lambda_w = R$. $\lambda_w = 1$ (resp. 0) means that the wth 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 $\boldsymbol{\lambda}$. After each update, $\boldsymbol{\lambda}$ 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 $w \in \{w | \lambda_w = 1\}$, $\boldsymbol{\lambda}(w)$ is defined as the order of the wth element among the elements having the value 1. For example, for $\boldsymbol{\lambda} = (1, 0, 0, 1, 1)$, we have $\boldsymbol{\lambda}(1) = 1$, $\boldsymbol{\lambda}(4) = 2$, and $\boldsymbol{\lambda}(5) = 3$.

We used the same notations of the basic arithmetic operations as those used in [20]. For the two vectors $\mathbf{u}$ and $\mathbf{v}$ of the same dimension n, we define $\mathbf{u} + \mathbf{v} =$

$(u_1 + v_1, \ldots, u_n + v_n)$, $\mathbf{u}/\mathbf{v} = (u_1/v_1, \ldots, u_n/v_n)$, $\mathbf{u}^{\mathbf{v}} = u_1^{v_1} \cdots u_n^{v_n}$, $x^{\mathbf{v}} = x^{v_1} \cdots x^{v_n}$ for a real x, $|\mathbf{u}| = \sum_{i=1}^{n} |u_i|$, and $\mathbf{u}_{-i} = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n)$. Note that $x^{|\mathbf{u}|} = x^{\mathbf{u}}$ if $u_i \geq 0$ for $i = 1, \ldots, n$, and that $x^{\mathbf{v}} \cdot \mathbf{u}^{\mathbf{v}} = (x\mathbf{u})^{\mathbf{v}}$. For a gamma function Γ, let $\Gamma(\mathbf{u}) = \Gamma(u_1) \cdots \Gamma(u_n)$.

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

Figure 2 Example of a fragmentation model of a motif. The six rectangles represent the columns selected as the motif part. In this example, $\boldsymbol{\lambda}$ is $(1, 0, 1, 1, 0, 1, 0, 1, 0, 1)$, which gives $R = 6$ and $W = 10$.

We also introduced the fragmentation motif model discussed in [20]. In this model, $R(< W)$ selected positions of a short aligned block of the width W are evaluated as the motif part (see Fig. 2). DegSampler optimizes R although it is kept constant in [20]. We also show that the posterior probability distribution is tempered for the ZOOPS and OOPS models in the fragmentation model. In other words, the posterior probability distribution is parameterized with a temperature parameter of T. We first describe the untempered version.

Likelihood

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 how much of each residue is conserved in each of the selected position 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 $\boldsymbol{\theta}_{1:R} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$, where $\boldsymbol{\theta}_r = (\theta_{1,r} \cdots \theta_{L,r})^T$ is the parameter set of a categorical distribution. For $r = 1, \ldots, R$, it holds that $\sum_{\ell=1}^{L} \theta_{\ell,r} = 1$ and $\theta_{\ell,r} \geq 0$ ($\ell = 1, \ldots, L$). These categorical distributions are assumed to be independent of each other.

The letters outside the R selected positions in all given sequences are assumed to be drawn independently from the same categorical distribution with the parameter $\boldsymbol{\theta}_0$. This is called the background model.

The unnormalized likelihood function for the ZOOPS model of DegSampler, which is parameterized by $\mathbf{Z}, \boldsymbol{\lambda}$, and $\boldsymbol{\theta}_{0:W} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_W)$, is defined as

$$L(\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R} | \mathbf{X}) = \boldsymbol{\theta}_0^{\mathbf{c}(I[\mathbf{Z}, \boldsymbol{\lambda}]^c)} \cdot \prod_{w:\lambda_w = 1} \boldsymbol{\theta}_{\boldsymbol{\lambda}(w)}^{\mathbf{c}(I[\mathbf{Z}, w])}.$$

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

$$p\left(\mathbf{X} \,|\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R}\right) \propto L(\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R} | \mathbf{X}).$$

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

Prior distributions

Here, we formulate the prior distributions for the parameters $\boldsymbol{\theta}_{1:R}$, $\mathbf{Z}$, and $\boldsymbol{\lambda}$.

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

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

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

We use the same Dirichlet distribution parameter $\boldsymbol{\alpha}_m$ for the R random variables $\boldsymbol{\theta}_{1:R} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$. For $r = 1, \ldots, R$, we define the following:

$$p\left(\boldsymbol{\theta}_r \,|\boldsymbol{\alpha}_m\right) = \mathrm{Dir}\left(\boldsymbol{\theta}_r \,|\boldsymbol{\alpha}_m\right).$$

Next is the prior probability distribution for the background model. In general, the prior information of a background model is not more important than that of a motif model because $I[\mathbf{Z}, \boldsymbol{\lambda}]^c$ contains many letters. Thus the prior distribution is set to be the simplest Dirichlet distribution, that is the uniform distribution over the open standard $(|\Sigma - 1)$-simplex.

Table 2 Notations related to **Z**.

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

Prior distribution of **Z**

We assume that for each $i = 1, \ldots, N$, and $j = 1, \ldots, |X_i|$, the prior probability $pos_{i,j}$ is given; this probability represents the likelihood of the position being part of a motif occurrence. We introduce the notation $m(z_i, \boldsymbol{\lambda}) = \{z_i + w - 1 \mid \lambda_w = 1, w = 1, \ldots, W, z_i \neq 0\}$, which represents 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

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

if $z_i \neq 0$ and is proportional to

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

if $z_i = 0$. This formulation is almost the same as in [21]. Furthermore, we apply two modifications to the formulation. Note that the last term $\prod_{j=1}^{|X_i|}(1 - pos_{i,j})$ 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 y of this prior term to adjust the effectiveness of the prior information quantity, for which the term is raised to the power of y. The resulting prior probability distribution of z_i given $\boldsymbol{\lambda}$ is proportional to

$$f_{\mathrm{pos}}(z_i; \boldsymbol{\lambda}) = \prod_{j \in m(z_i,\boldsymbol{\lambda})} \left(\frac{pos_{i,j}}{1 - pos_{i,j}} \right)^{y \cdot \delta_{z_i,0}},$$

where $\delta_{i,j}$ is the Kronecker's delta, equaling 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

$$p(\mathbf{Z}|\boldsymbol{\lambda}) \propto \prod_{i=1}^{N} f_{\mathrm{pos}}(z_i; \boldsymbol{\lambda})$$

with different normalizing constants in the ZOOPS and OOPS models. We describe how to determine the value of $pos_{i,j}$ in the Results section.

Hyperprior distribution of $\boldsymbol{\lambda}$

We discuss the hyperprior distribution of $\boldsymbol{\lambda}$. There are various fragmentation patterns in the ELM resource; therefore, it will be 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

$$\begin{aligned} & p\left(\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R} \middle| \mathbf{X}\right) \\ \propto\ & L\left(\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{0:R} \left| \mathbf{X}\right.\right) \cdot p(\boldsymbol{\theta}_{1:R}|\boldsymbol{\alpha}_m) \cdot p(\boldsymbol{\theta}_0|\boldsymbol{\alpha}_b) \cdot p(\mathbf{Z}|\boldsymbol{\lambda}) \cdot p(\boldsymbol{\lambda}). \end{aligned}$$

By integrating out $\boldsymbol{\theta}_{0:R}$, as shown in [20], we obtain the collapsed posterior distribution of $\mathbf{Z}$ and $\boldsymbol{\lambda}$ as

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

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 ZOOPS models.

Update of a motif starting position z_k

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

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

where

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

Note that the approximation scheme given in [20] has been used for the derivation. In particular, when $z_k = 0$, $p\left(z_k \left| \mathbf{Z}_{-k}, \boldsymbol{\lambda}, \mathbf{X}\right.\right)$ is proportional to 1 with the same normalizing constant because all $\boldsymbol{c}\left(I[z_k, w]\right)$ and $\delta_{z_k,0}$ become 0.

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

We describe 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 on the left are indexed as $-\Delta + 1, \ldots, -1, 0$, and the columns added on the right are indexed as $W+1, \ldots, W+\Delta$. Thus, the columns of the resulting $\boldsymbol{\lambda}$ are indexed continuously from $-\Delta + 1$ to $W + \Delta$.

In the second step, the index $v \in \{-\Delta + 1, \ldots, W + \Delta\}$ was chosen 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

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

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

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

Both two 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 left columns indexed as $-\Delta+1, \ldots, -1, 0$, then the value of z_i is replaced with $z_i + v - 1$ for $z_i \neq 0$. Furthermore, the leftmost and rightmost contiguous blocks of columns indexed by 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 and right on the substrate sequences. In addition, the R and W values are also varied. The default value of Δ is set to 3 in this research.

Simulated annealing

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

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

The corresponding conditional probability distribution of z_k is as follows:

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

where

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

The conditional probabilities of λ_v are given as follows:

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

and

$$\begin{aligned} & p(\lambda_v = 0 | \boldsymbol{\lambda}_{-v}, \mathbf{Z}, \mathbf{X}) \\ \propto\ & \tilde{\boldsymbol{\theta}}_0^{\boldsymbol{c}(I[\mathbf{Z},v])/T} \end{aligned}$$

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

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

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 three times to avoid falling into a local optimmum. DegSampler2Core repeats the procedure $M = 300$ times. At $M_{sa} = 240$, a simulated annealing is started. Thus, the last 20% of all iterations are annealed. After that, at each repetition, the value of the temperature parameter T is reduced by sa_ratio = 0.999.

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

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

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

Algorithm 1: DegSampler algorithm. The three variables optLogPost, $\mathbf{Z}_{best}$, and $\boldsymbol{\lambda}_{best}$, 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 sequence positions of the DEG motif sites that are shared by the predicted motif occurrences as true positives. Similary, the sequence positions of the predicted motif occurrences of the substrate protein sequences with some DEG motif annotations that are not covered by any 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

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

Results

Predictability on 36 E3s

Experimental configuration

In DegSampler, the pseudo count of each position of a motif were set as $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$, and the site distribution model was set as ZOOPS. DegSampler were executed five times for each set of parameter values.

Source of position-specific prior information

We here examine how effective the position-specific prior information derived from

IUPred2, ANCHOR1, ANCHOR2, and fMoRFpred
maxW = 9
minR = 6
$y = 1$

Effectiveness of position-specific prior

One feature of DegSampler is that it can exploit position-specific prior information of the occurrences of E3 motifs. DegSampler can regulate the effect of the prior using the parameter y. The F-measure values of DegSampler with different y values as $y = 0, 0.1, 0.2, \ldots, 2, 3, \ldots, 21$, are shown in Figure 3(a). The other parameter values are the same as the values that help attain the highest F-measure value in the performance comparison section. In accordance with the increase of y, the F-measure drastically increases from $y = 0$ and converges around $y = 1.0$. After that, up to $y = 19$, the F-measure is stable. Then, at $y = 20$ and 21, the F-measure drops by 10%. Probably, this change is caused because the position-specific prior information is too biased. When y is increased to 22, the calculated value proportional to the conditional distribution of z_i becomes zero because y is too large.

The case in which $y = 0$ is equivalent to the case in which the position-specific prior probability distribution is a uniform distribution. The mean of the F-measure value in this case is 0.028. This value is quite

Figure 3 Effectiveness of position-specific prior information on the F-measure. The height of a bar represents the mean of the F-measure, and the error bar shows the standard error. (a) DegSampler. The coefficient y of the position-specific prior distribution is in the range of $\{0, 0.1, \ldots, 2, 3, \ldots, 21\}$. The parameter values of DegSampler are as follows: DegSampler: $W_{\max} = 9$, $R_{\min} = 8$, $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$, $pos_{i,j}$ is the ANCHOR1 probability, the motif site distribution model is ZOOPS. (b) DegSampler1. The coefficient c_d of the position-specific prior distribution is in the range of $\{\{60, 80, \ldots, 160\}\}$. The parameter values of DegSampler1 are as follows: $W = 9$, $R = 7$, $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$; $pos_{i,j}$ is the ANCHOR1 probability. The motif site distribution model is ZOOPS, and the temperature parameter of the second likelihood function $T_{aai} = \infty$.

low; therefore, we can recognize the usefulness of the position-specific prior information.

For reference, we have shown the F-measure values of DegSampler1 with different values of the coefficient of the position-specific prior $c_d = 60, 80, \ldots, 160$ are shown in Figure 3(b). The other parameter values are the same as those by which the highest F-measure value is attained in the performance comparison section. The F-measure is unstable compared with DegSampler. We expected that a higher F-measure value could be achieved with a larger value of c_d because the highest F-measure value on the graph is marked at the right end-point of $c_d = 160$. However, with $c_d = 180$ as DegSampler, the calculated value proportional to the conditional distribution of z_i becomes zero because c_d is too large.

Performance comparison

We searched for the best parameter value set of DegSampler; however, the search space was relatively limited and separated into two spaces. One space was $y = 0, 0.1, \ldots, 2, 3, \ldots, 21$ with $W_{\max}$ and $R_{\min}$ fixed to 9 and 8, respectively. The values of the remaining parameters were the same as the values in the synthetic data. Another search space is $W_{\max} = 6, 7, \ldots, 12$ and $R_{\min} = W_{\max} - 3, \ldots, W_{\max} - 1$ with $y = 1$. The other parameter values are the same as in the synthetic data experiment. Among these parameter values, the configuration $(W_{\max}, R_{\min}, y) = (9, 8, 8)$ marked the highest F-measure of 0.290, as given in Fig. 4. This was 17% higher than that of DegSampler1.

From the graph of Fig. 4, it is clear that DegSampler outperforms DegSampler1 and the other methods. To see details of the F-measure relationship between DegSampler and DegSampler1, we applied a Wilcoxon rank sum test for each E3 input instance. Table 3 shows the results of only those E3s whose p-values were less than 1. A p-value of 1 implies that the mean of the F-measure values of DegSampler is identical or almost the same as that of DegSampler1. Eleven E3s are listed in Tab. 3. Among them, nine instances are statistically significant with the significance level of 0.05. Among the nine instances, DegSampler outperforms DegSampler1 on eight E3 instances.

Example of found motif occurrences

Here, we give an example of the outputs of DegSampler. In E3Net, the E3 ubiquitin ligase of SYVN1_HUMAN is associated with the following 12 substrate proteins: O00141 (SGK1_HUMAN; Serine/threonine-protein kinase Sgk1), O15354 (GPR37_HUMAN; Prosaposin receptor GPR37), O75460 (ERN1_HUMAN; Serine/threonine-protein kinase/endoribonuclease IRE1), P04035 (HMDH_HUMAN; 3-hydroxy-3-methylglutaryl-coenzyme A reductase), P04637 (P53_HUMAN; Cellular tumor antigen p53), P10636 (TAU_HUMAN; Microtubule-associated protein tau), P10909 (CLUS_HUMAN; Clusterin), P42858 (HD_HUMAN; Huntingtin), Q86TM6 (SYVN1_HUMAN; E3 ubiquitin-protein ligase synoviolin), Q8WZ42 (TITIN_HUMAN; Titin), Q92542 (NICA_HUMAN; Nicastrin), and Q9UKV5 (AMFR_HUMAN; E3 ubiquitin-protein ligase AMFR).

Table 4 shows the alignment of the motif occurrences found by DegSampler and DegSampler1 along with the parameter with which the highest F-measure value is marked in the performance comparison on the real data. The motif occurrence of DegSampler on the P04637 (P53_HUMAN) sequence starting at position 18 overlaps with a site of the DEG motif, DEG_MDM2_SWIB_1, starting at 19. The regular expression of the DEG motif is `F[^P]{3}W[^P]{2,3}[VIL]`. The suffix of length 8 of the motif occurrence sequence FSDLWKLL exactly matches the DEG motif. However, the motif occurrence of DegSampler1 on the sequence does not overlap with the DEG site.

We checked the ANCHOR1 probability for the 5th substrate protein P04637 (P53_HUMAN). The graph is shown in Fig. 5. As can be seen, the ANCHOR1 probabilities on the site of DEG_MDM2_SWIB_1 from position 19 to 26 are almost 1. As a result, DegSampler succeeded in using this information, but DegSampler1 failed to do so.

Contributions of the formulation of position-specific prior probability distribution, the utilization of simulated annealing, and the site distribution model

In this section, we examine the important contributions of DegSampler. One contribution is the formulation of the position-specific prior probability distribution. We compare the performance of DegSampler with that of DegSampler in which the formulation was replaced with that of DegSampler1. The second factor

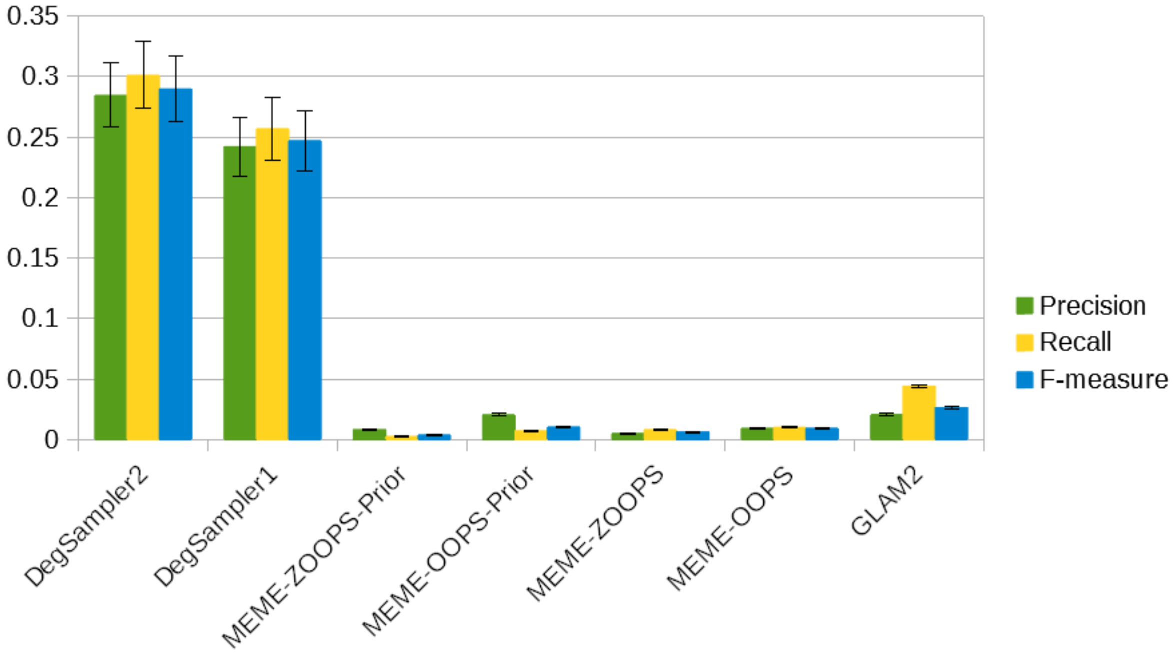

Figure 4 Performance comparison of DegSampler with DegSampler1, MEME, and GLAM2. The parameter values of DegSampler and DegSampler1 are as follows: DegSampler: $W_{\max} = 9$, $R_{\min} = 8$, $y = 1$, $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$, $pos_{i,j}$ is the ANCHOR1 probability, the motif site distribution model is ZOOPS; DegSampler1: $W = 9$, $R = 7$, $c_d = 160$, $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$, $pos_{i,j}$ is the ANCHOR1 probability, the motif site distribution model is ZOOPS, the temperature parameter of the second likelihood function $T_{aai} = \infty$. The parameter settings of MEME-OOPS, MEME-ZOOPS, MEME-OOPS-Prior, MEME-ZOOPS-Prior, and GLAM2 are the same those in the synthetic experiment.

Figure 5 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 with the symbol 'x' on their sequence positions at the height of 1. The graph title consists of the entry name of the substrate protein, its UniProt accession, and the IDs of the DEG motifs on the protein.

is simulated annealing. We compared the performances of DegSampler with and without simulated annealing. Finally, the performance of DegSampler with ZOOPS was compared with the performance of DegSampler with OOPS. The parameter values were set as follows: $W_{\max} = 9$, $R_{\min} = 8$, $y = 8$, $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$; $pos_{i,j}$ is the ANCHOR1 probability. PosPriorVer2 (resp. PosPriorVer1) refers to the formulation of the DegSampler (resp. DegSampler1) position-specific prior probability distribution. SA (resp. NoSA) refers to the utilization (resp. no utilization) of the simulated annealing of DegSampler. ZOOPS and OOPS are the specifications of the site distribution mode. Thus, the leftmost configuration PosPriorVer2-SA-ZOOPS is identical with the parameter values used for the performance comparison on the real data. The second configuration PosPriorVer2-NoSA-ZOOPS is the same but without the simulated annealing. The third configuration PosPriorVer1-SA-ZOOPS is the same as the first one but has the formulation of the DegSampler1 position-specific prior probability distribution. The fourth configuration PosPriorVer1-NoSA-ZOOPS is the same as the first one but has the formulation of the DegSampler1 position-specific prior probability distribution without the simulated annealing. The last configuration PosPriorVer2-SA-OOPS is the same as the first configuration but has the OOPS model.

Fig. 6 presents the results graphically. The two leftmost configurations in Fig. 6 show the contribution of simulated annealing. Although there is no major difference between the two leftmost configurations, a special simulated annealing has been used in DegSampler. There is room for improvement in the simulated annealing. The third and fourth configurations use the formulation of the DegSampler1 pos¡ition-specific prior probability distribution. Clearly, the performance obtained by using the version 2 formulation is superior to that obtained by using the version 1 formulation.

Table 3 E3s with a p-value less than one. The columns show the name of E3, the mean and standard error of the F-measures of DegSampler1, the mean and standard error of the F-measures of DegSampler, and the p-value.

	DegSampler1		DegSampler		
E3	Mean	SE	Mean	SE	p-value
APC(CDH1)_YEAST	0	0	0.0882	0	0.0079
CHIP_HUMAN	0.8941	0.0129	0.9412	0.0000	0.4444
COP1-SPA1_ARATH	0.6667	0	0	0	0.0079
DCX(DET1-COP1)_HUMAN	0.3256	0	0.3721	0	0.0079
DCX(DTL)_HUMAN	0.7000	0	0.8333	0	0.0079
ICP0_HHV11	0.8235	0	0.9412	0.0000	0.0079
MDM2_HUMAN	0.8103	0.0169	0.7692	0	0.4048
MKRN1_HUMAN	0	0	0.7692	0	0.0079
SCF(BTRC)_HUMAN	0	0	0.0735	0	0.0079
SCF(SKP2)_HUMAN	0.2993	0	0.3265	0	0.0079
SYVN1_HUMAN	0	0	0.9412	0.0000	0.0079

Figure 6 Performance with various parameter configurations. The parameter values are as follows: $W_{\max} = 9$, $R_{\min} = 8$, $y = 8$, $\boldsymbol{\alpha}_m = 0.1 \cdot \mathbf{1}$, $pos_{i,j}$ is the ANCHOR1 probability. PosPriorVer2 (resp. PosPriorVer1) means the formulation of the DegSampler (resp. DegSampler1) position-specific prior probability distribution. SA (resp. NoSA) means the utilization (resp. no utilization) of the simulated annealing of DegSampler. ZOOPS and OOPS are the specifications of the site distribution mode. Thus, the leftmost configuration in the graph, PosPriorVer2-SA-ZOOPS is identical with the parameter values used in the performance comparison on the real data. The second configuration PosPriorVer2-NoSA-ZOOPS is the same but does not have simulated annealing. The third configuration PosPriorVer1-SA-ZOOPS is the same as the first configuration but has the formulation of the DegSampler1 position-specific prior probability distribution. The fourth configuration PosPriorVer1-NoSA-ZOOPS is the same as the first configuration but has the formulation of the DegSampler1 position-specific prior probability distribution without the simulated annealing. The last configuration PosPriorVer2-SA-OOPS is the same as the first configuration but has the OOPS model.

The difference between the leftmost and the rightmost configurations is the site distribution model. Their performances were almost the same. However, the result largely depends on the data used.

Discussion

Narlikar *et al.*, also proposed an alternative prior score, which is calculated from negative sequences as well as positive ones.

$$p_{i,z_i} \propto \frac{\mathcal{S}_{\mathcal{D}}}{1 - \mathcal{S}_{\mathcal{D}}}$$

where

$$\mathcal{S}_{\mathcal{D}}(X_i, z_i, \boldsymbol{\lambda}) = \frac{\sum_{(k,\ell):X_{k,\ell,\boldsymbol{\lambda}}=X_{i,z_i,\boldsymbol{\lambda}}} \mathcal{S}_{\mathcal{N}}(X_k, \ell, \boldsymbol{\lambda})}{\sum_{(k,\ell):X_{k,\ell,\boldsymbol{\lambda}}=X_{i,z_i,\boldsymbol{\lambda}}} \mathcal{S}_{\mathcal{N}}(X_k, \ell, \boldsymbol{\lambda}) + \sum_{(k,\ell):Y_{k,\ell,\boldsymbol{\lambda}}=X_{i,z_i,\boldsymbol{\lambda}}} \mathcal{S}_{\mathcal{N}}(Y_k, \ell, \boldsymbol{\lambda})}$$

if $z_i \neq 0$ and $p_{i,z_i} \propto 1$ otherwise with the same normalizing constant.

—

To-do: rename the tool name.

because this method is not limited to the problem of finding degrons.

Gibbs sampler with Position-specific prior

Concluding remarks

Judging from the low predictability of the popular general motif finding tool, MEME, shown in Fig. 4, the identification problem of E3 motifs is challenging. Thus, to resolve the problem, we need to develop specialized motif-finding methods, such as DegSampler1 and DegSampler. In this study, we found the following. First, it is important to determine how to use position-specific prior information. MEME used the same source of the prior information. However, the F-measure of MEME was much lower than those of DegSampler and DegSampler1. Second, the search strategy is important to find motifs. Essentially, the difference between DegSampler and DegSampler1 is their Gibbs updating procedures. The search strategy of DegSampler was superior to that of DegSampler1 because on the eight E3 instances, the F-measure of DegSampler was statistically higher than that of DegSampler1.

One of the future strategies will be to introduce more complex motif models such as the dependency model. The dependency model can deal with the correlation between the residues at remote positions. Then, more accurate predictions will be realized for such degrons.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent to publish
Not applicable.

Table 4 Example of an output of DegSampler. The E3 picked up here is SYVN1_HUMAN and the substrate proteins assigned to the E3 by E3Net are 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.

DegSampler										
substrate	start									
		*	*	*	*	*	*	*	*	*
O00141	20	M	V	A	I	L	I	A	F	M
O15354	140	T	A	L	Q	L	F	L	Q	I
O75460	452	F	L	L	I	G	W	V	A	F
P04035	329	A	L	L	L	A	V	K	Y	I
P04637	18	T	F	S	D	L	W	K	L	L
P10636	188	L	K	H	Q	L	L	G	D	L
P10909	4	T	L	L	L	F	V	G	L	L
P42858	3	T	L	E	K	L	M	K	A	F
Q86TM6	518	M	L	Q	I	N	Q	Y	L	T
Q8WZ42	11822	E	K	I	F	Q	L	K	A	I
Q92542	20	L	L	S	F	C	V	L	L	A
Q9UKV5	625	R	R	R	M	L	A	A	A	A

DegSampler1										
substrate	start									
		*		*	*	*	*	*		*
O00141	35	L	N	D	F	I	Q	K	I	A
O15354	139	P	T	A	L	Q	L	F	L	Q
O75460	562	S	V	V	I	V	G	K	I	S
P04035	416	V	V	G	N	S	S	L	L	D
P04637	40	M	D	D	L	M	L	S	P	D
P10636	187	L	L	K	H	Q	L	L	G	D
P10909	404	E	V	V	V	K	L	F	D	S
P42858	8	M	K	A	F	E	S	L	K	S
Q86TM6	608	L	Q	K	L	E	S	P	V	A
Q8WZ42	0									
Q92542	95	M	V	L	L	E	S	K	H	F
Q9UKV5	624	L	R	R	R	M	L	A	A	A

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 and 18H05214.

Authors' contributions

OM participated in the conception of this project, in the formulation of the mathematical model and the algorithm, in the implementation of the software, in the computational experiment, in the analysis of the results, and in the drafting of the manuscript. FM participated in the conception of this project, in the analysis of the results, and in improvement of the manuscript. All of 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 Edanz (www.edanzediting.co.jp) and Enago (www.enago.jp) for the English language review of the submitted and revised versions of this manuscript, respectively.

Author details

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

References

1. 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)
2. Dinkel, H., *et al.*: ELM 2016–data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. **44**, 294–300 (2016)
3. 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)
4. 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)
5. 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)
6. 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)
7. Gordân, R., Narlikar, L., Hartemink, A.J.: A fast, alignment-free, conservation-based method for transcription factor binding site discovery. In: Proc 12th Annual International Conference, RECOMB 2008 (LNBI 4955), pp. 98–111 (2008)
8. Bailey, T.L., Bodén, M., Whitington, T., Machanick, P.: The value of position-specific priors in motif discovery using MEME. BMC

Bioinformatics **11**, 179 (2010)

9. 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)
10. 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)
11. Mészáros, B., Simon, I., Dosztányi, Z.: Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. **5**(5), 1000376 (2009)
12. Dosztányi, Z., Mészáros, B., Simon, I.: ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics **25**, 2745–2746 (2009)
13. 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)
14. Fang, C., Noguchi, T., Tominaga, D., Yamana, H.: MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics **14**, 300 (2013)
15. Jones, D.T., Cozzetto, D.: DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. Bioinformatics **31**(6), 857–63 (2015)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. Narlikar, L., Gordân, R., Hartemink, A.J.: Nucleosome occupancy information improves *de novo* motif discovery. In: Research in Computational Molecular Biology (RECOMB) 2007, LNBI 4453, pp. 107–121 (2007)