Bioinformatics, 2022, 1–3
https://doi.org/10.1093/bioinformatics/btac437
Advance Access Publication Date: 30 June 2022
Applications Note

Sequence analysis

HSMotifDiscover: identification of motifs in sequences composed of non-single-letter elements

Vinod Kumar Singh[1], **Rohan Misra**[1], **Steven C. Almo**[2], **Ulrich G. Steidl**[3], **Hannes E. Bülow**[1,4] **and Deyou Zheng**[1,4,5,*]

[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA, [2]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA, [3]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY 10461, USA, [4]Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY 10461, USA and [5]Department of Neurology, Albert Einstein College of Medicine, Bronx, NY 10461, USA

*To whom correspondence should be addressed.

Associate Editor: Tobias Marschall

Received on December 6, 2021; revised on April 24, 2022; editorial decision on June 24, 2022; accepted on June 28, 2022

Abstract

Summary: The functional sub-string(s) of a biopolymer sequence defines the specificity of its interaction with other biomolecules and is often referred to as motifs. Computational algorithms and software have been broadly developed for finding such motifs in sequences in which the individual elements are single characters, such as those in DNA and protein sequences. However, there are more complex scenarios where the motifs exist in non-single-letter contexts, e.g. preferred patterns of chemical modifications on proteins, DNAs, RNAs or polysaccharides. To search for those motifs, we describe a new method that converts the modified sequence elements to representative single-letter codes and then uses a modified Gibbs-sampling algorithm to define the position specific scoring matrix representing the motif(s). As a proof of principle, we describe the implementation and application of an R package for discovering heparan sulfate (HS) motifs in glycan sequences, which are important in regulating protein–protein interactions. This software can be valuable for analyzing high-throughput glycoprotein binding data using microarrays with HS oligosaccharides or other biological polymers.

Availability and implementation: HSMotifDiscover is freely available as an open source R package released under an MIT license at https://github.com/bioinfoDZ/HSMotifDiscover and also available in the form of an app at https://hsmotifdiscover.shinyapps.io/HSMotifDiscover_ShinyApp/.

Contact: deyou.zheng@einsteinmed.edu

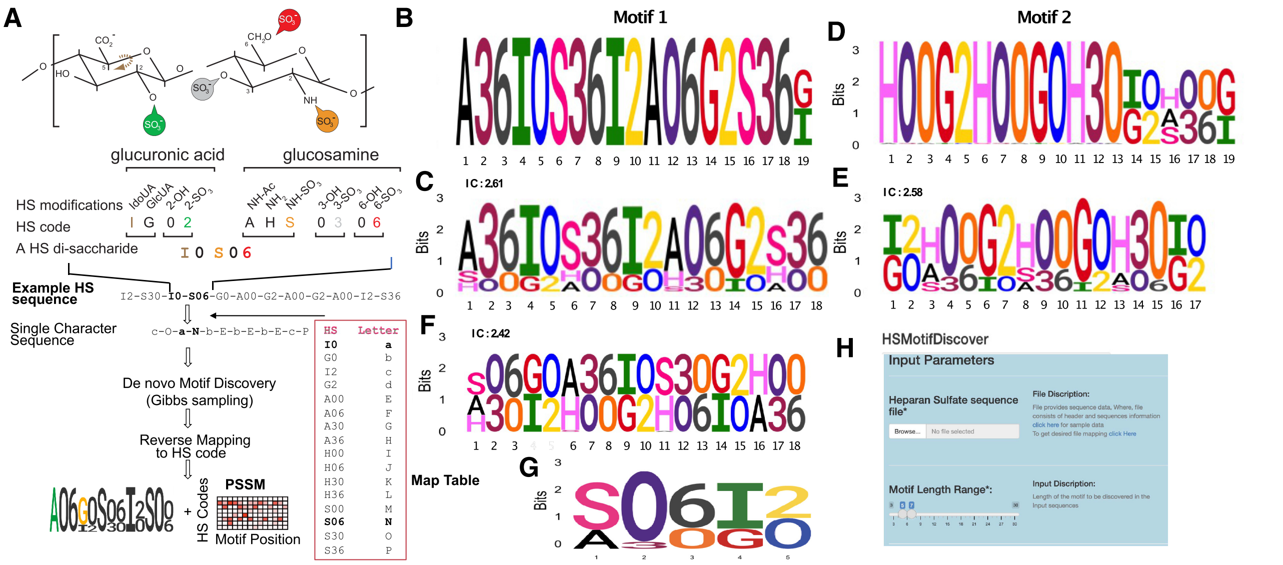
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

De novo motif discovery from a set of nucleotide or protein sequences, such as the cognate DNA motifs for transcription factors, has long been an area of computational interests (D'Haeseleer, 2006; Lawrence *et al.*, 1993; Tompa *et al.*, 2005). Currently available tools, however, are limited to DNA, RNA and protein sequences whose composition elements (i.e. units) are single-letter alphabet (e.g. A/T/G/C for DNA) (Bailey *et al.*, 2009). Chemical modifications on the fundamental units, however, can occur, resulting in two or more letters (e.g. 5mC for methylation on C in DNA) for representing the functional elements.

Heparan sulfate (HS) sequences, as one type of glycosaminoglycans with important roles in mediating extracellular protein and virus–host interactions, fall into this class. Its basic units are dimers composed of a hexuronic acid and glucosamine, with epimerizations as well as *O*- and *N*-sulfate modifications at alternate positions (Fig. 1A). As a result of these modifications, HS glycans are amongst the most diverse molecules in nature (Bülow and Hobert, 2006; Esko and Lindahl, 2001; Sarrazin *et al.*, 2011) and play critical roles in determining the interaction specificity and functions of proteoglycans (Kjellén and Lindahl, 2018; Townley and Bülow, 2018; Xu and Esko, 2014). The identification of HS sequence motifs that confer functional specificity remains a great experimental challenge, but new technologies such as microarray platforms have been developed to systematically evaluate binding selectivity of HS-interacting proteins, most recently for the receptor binding domain of the spike of severe acute respiratory syndrome-related coronavirus 2 (SARS-Cov-2) (Arungundram *et al.*, 2009; Liu *et al.*, 2021; Zong *et al.*, 2017). Computational identification of the HS motifs from such high-throughput data, however, has not been developed.

HS sequences can be described in a reduced format referred as modified 'Lawrence code', which is a combination of letters and digits ('I' and 'G' for the two types of glucuronic acids and '0/2' for the SO_3



HS	Letter
I0	a
G0	b
I2	c
G2	d
A00	E
A06	F
A30	G
A36	H
H00	I
H06	J
H30	K
H36	L
S00	M
S06	N
S30	O
S36	P

HSMotifDiscover

Input Parameters

Heparan Sulfate sequence file*

Browse... No file selected

File Discription:
File provides sequence data, Where, file consists of header and sequences information
click here for sample data
To get desired file mapping click Here

Motif Length Range*: 5 7

Input Discription:
Length of the motif to be discovered in the Input sequences

Fig. 1. Algorithm outline and results. (**A**) Scheme of the key information and methods. Two simulated motifs (**B**, **D**) and the motifs discovered by HSMotifDiscover (**C**, **E**, **F**). (**G**) HS motif for a receptor binding domain of SARS-Cov-2. (**H**) GUI interface for the Shiny App

modifications in the second positions of the sugar ring; ‘A’, ‘H’ and ‘S’ for the three types of glucosamines and ‘0/3’ or ‘0/6’ for the SO_3 modifications in the third and sixth positions of the sugar ring, respectively) (Lawrence *et al.*, 2008; Townley and Bülow, 2018). As such, the individual saccharide units of an HS sequence can be represented by two or three digits/letters (i.e. I0 and S06; Fig. 1A), rendering typical sequence analysis software unsuitable for their analysis. Thus, new software is needed for motif discovery in HS sequences where the individual units are represented by two or more letters and digits of non-uniform lengths. These tools also need to directly consider the affinity/interaction scores as they are valuable for defining binding specificity.

2 Algorithm and implementation

To address this need, we developed a new software (‘HSMotifDiscover’) that is based on Gibbs sampling in its core. It includes three key features that are specifically designed to find motifs in non-single-letter sequences. The first is to map the input sequences of multi-letter units to a single-letter representation; in the case of HS sequences 16 standard letters are used (Fig. 1A), but the software can accommodate up to 62 characters. A reverse mapping process is also needed for plotting the discovered motif logos using the original HS codes and the ggseqlogo package (Wagih, 2017). The second algorithmic feature is to weight the input sequences during Gibbs sampling, which allows users to directly include affinity values in motif discovery, when data from microarray platforms or other high-throughput technologies are used. The third feature is to group units representing different sulfate modifications on the same saccharide. While the first two features may be implemented as an extension to existing software like MEME (Bailey *et al.*, 2009), the last feature needs modifications of the core Gibbs sampling and it is essential for HSMotifDiscover function. This is because the two saccharide units are in different lengths, representing different extents of modification patterns that can occur, and thus should occupy mutually exclusive positions in a position specific scoring matrix (PSSM).

More details about our specific modification of the Gibbs algorithm, including PSSM calculation and how sequence weights are used, are provided in Supplementary Section S1. In brief, the Gibbs sampler takes an input of *N* sequences and generates as output a set of *N* subsequences, one in each input sequence, that represent the ‘most similar’ substrings in the *N* sequences. Here, our measure of ‘most similar’ is a likelihood function derived from the PSSM. For simplicity, our current implementation assumes that all the sequences share an ungapped pattern of length *W* (defined by users) and that each sequence contains exactly one instance of this pattern. Importantly, since some letters in the mapped HS sequences represent different chemical modifications to the same saccharide unit, e.g. sulfation on the third and sixth positions of glucosamines, we account for such distinct modifications by multiplying the PSSM score of the basic units by the proportion of the chemical group associated to the basic unit. This is critical because it ensures the PSSM to start from all units representing different groups with equal probability. As such, our software allows users to group units representing different modifications of the same saccharide (or basic elements of other molecules or sequences) and provide the information as an input to HSMotifDiscover.

3 Results

We first tested HSMotifDiscover and compared with the MEME software (Bailey *et al.*, 2009) on simulated DNA sequences possessing CArG (SRF-binding sites, JASPER ID: MA0083.2) or SOX4 (JASPER ID: MA0867.1) motifs and a subset of SRF-binding sequences (i.e. peaks containing the CArG motifs) from the ENCODE ChIP-seq data (ENCFF909FRA) (Dunham *et al.*, 2012). The results demonstrated that both tools successfully discovered the known motifs in their expected locations (Supplementary Section S2), as expected for two software implementing the same core algorithm.

To demonstrate the performance of HSMotifDiscover on HS sequences, we tested it on simulated datasets containing 200 HS sequences of 30 monosaccharide units (Supplementary Section S3), half of which contained ‘Motif 1’ and the other half with ‘Motif 2’ at random locations (Fig. 1B and D). When HS sequences with ‘Motif 1’ were given higher weight over the sequences with ‘Motif 2’, the discovered motif resembled the ‘Motif 1’ (Fig. 1C). Conversely, the opposite weighting scheme yielded a motif like the ‘Motif 2’ (Fig. 1E). When all the sequences were given equal weight; however, the discovered motif had resemblance to both (Fig. 1F), but the information content (IC) of the motifs was reduced, as expected. To test on experimental data, we applied HSMotifDiscover to microarray data that measured the HS binding specificity of the SARS-Cov-2 spike protein S1 using 93 synthetic HS oligosaccharides. The top motif is shown in (Fig. 1G), which is highly similar to what was reported previously (Liu *et al.*, 2021), S06I2 for GlcNS6S-IdoA2S.

To demonstrate the value of grouping function, we compared HSMotifDiscover with MEME on simulated HS sequences and showed that it is needed for recovering the targeted motifs consistently under different scenarios (Supplementary Section S4).

4 Discussion

Although HSMotifDiscover is motivated by HS motif discovery, it is implemented with a flexibility that allows its usage for motifs in any kind of biological sequences (e.g. other glycans), or sequences of predefined words or even non-words like shapes, with uniform- or non-uniform-length elements, because it converts a given non-single-letter sequence to a string of single-letter characters, as defined by users, and group them when appropriate. In the current implementation, we provide a standalone R package and a Shiny server (Fig. 1H). We have also introduced many customization options to enable analysis of most types of biopolymers (see GitHub release for details). The current model assumes that each sequence has exactly *o*ne motif *o*ccurrence *p*er *s*equence (OOPS model). So, its performance may drop if the input set of sequences include a high proportion of sequences without the motif. To address this issue, future release will introduce *z*ero or *o*ne *o*ccurrences *p*er *s*equence (ZOOPS) model and parallel computing steps in order to account for more complex scenarios (Peng *et al.*, 2018).

In this report, we focused on evaluating our software with simulated data. Conceivably the performance of the Gibbs sampler could be affected by the length and diversity of the input sequences, when applied to experimental data. Data from the current HS oligosaccharide synthesis technology (Arungundram *et al.*, 2009; Liu *et al.*, 2021; Zong *et al.*, 2017) may not be optimal for HSMotifDiscover, e.g. the HS oligosaccharide diversity used for Figure 1G is relatively low. Nevertheless, we believe that our package will pave the way for a more systematic evaluation of data for molecules binding to HS sequences (or other biopolymers) from microarray or other platforms.

Funding

This study was supported by an NIH [U01CA241981].

Conflict of Interest: none declared.

References

Arungundram,S. *et al.* (2009) Modular synthesis of heparan sulfate oligosaccharides for structure−activity relationship studies. *J. Am. Chem. Soc.*, **131**, 17394–17405.

Bailey,T.L. *et al.* (2009) MEME suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Bülow,H.E. and Hobert,O. (2006) The molecular diversity of glycosaminoglycans shapes animal development. *Annu. Rev. Cell Dev. Biol.*, **22**, 375–407.

D'Haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.

Dunham,I. *et al.* (2012) An integrated encyclopaedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Esko,J.D. and Lindahl,U. (2001) Molecular diversity of heparan sulfate. *J. Clin. Invest.*, **108**, 169–173.

Kjellén,L. and Lindahl,U. (2018) Specificity of glycosaminoglycan-protein interactions. *Curr. Opin. Struct. Biol.*, **50**, 101–108.

Lawrence,C. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lawrence,R. *et al.* (2008) Disaccharide structure code for the easy representation of constituent oligosaccharides from glycosaminoglycans. *Nat. Methods*, **5**, 291–292.

Liu,L. *et al.* (2021) Heparan sulfate proteoglycans as attachment factor for SARS-CoV-2. *ACS Cent. Sci.*, **7**, 1009–1018.

Peng,S. *et al.* (2018) Efficient computation of motif discovery on intel many integrated core (MIC) architecture. *BMC Bioinformatics*, **19**, 101–110.

Sarrazin,S. *et al.* (2011) Heparan sulfate proteoglycans. *Cold Spring Harb. Perspect. Biol.*, **3**, 1–33.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Townley,R.A. and Bülow,H.E. (2018) Deciphering functional glycosaminoglycan motifs in development. *Curr. Opin. Struct. Biol.*, **50**, 144–154.

Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.

Xu,D. and Esko,J.D. (2014) Demystifying heparan sulfate-protein interactions. *Annu. Rev. Biochem.*, **83**, 129–157.

Zong,C. *et al.* (2017) Heparan sulfate microarray reveals that heparan sulfate–protein binding exhibits different ligand requirements. *J. Am. Chem. Soc.*, **139**, 9534–9543.

HSMotifDiscover: identification of motifs in sequences composed of non-single-letter elements

Supplemental Information

1. Calculation of PSSM (Position specific scoring matrix, $\mathcal{P}$):

Let us consider a set of N sequences, $S = \{s_n: 1 \leq n \leq N\}$ of lengths $L = \{l_n: 1 \leq n \leq N\}$, and associated normalized weights, $R = \{r_n: 1 \leq n \leq N\}$, each sequence s_n is incorporated with a motif sequence m_n= $s_n[a_n, (a_n + W - 1)]$ of width W at an initial random location $a_n \in \{k : 1 \leq k \leq L_n - W\}$.

Suppose there are J number of basic saccharide units $U = \{u_j: 1 \leq j \leq J\}$ for constructing the *N* sequences and their associated groups $G = \left\{ \mathcal{G}_{u_j} : 1 \leq j \leq J \right\}$, where $\mathcal{G}_{u_j} \epsilon \{g_1, g_2, \cdots, g_x\}$ and x is the number of groups present; a group represent units that can occur at a position, e.g., exchangeable variants of glucosamines. Number of variable units forms the weight vector η for the groups. For example, in HS sequences alternate positions are occupied by 4 (two letters/digits) and 12 (three letters/digits) different units. Hence, the latter group has higher weight than the former.

The weighted count c_{i,u_j} of unit u_j at position i of the motif sequences $M = \{m_n: 1 \leq n \leq N\}$ is obtained by incorporating sequence weights as

$$c_{i,u_j} = \sum_{n=1}^{N} \boldsymbol{r_n}[\, m_n(i) = u_j] + \alpha$$

where $m_n(i)$ is the unit at position i of motif sequence m_n and belongs to U. The Iverson bracket "$[Condition]$" is equals to 1 if "$Condition$" is true and 0 if it is false.

Weighted counts are used to obtain PSSM $\mathcal{P}$ entries as

$$p_{i,u_j} = \frac{c_{i,u_j}}{\sum_{j=1}^{J} c_{i,u_j}}, \qquad 1 \leq i \leq W, \quad 1 \leq j \leq J$$

$$-- eq\ (1)$$

The $\mathcal{P}$, is then used to estimate probability $P(\mathcal{M}|\mathcal{P})$ of an arbitrary sequence $\mathcal{M}$ as

$$P(\mathcal{M}|\mathcal{P}) = \sum_{i=1}^{W} \log\left(p_{i,\ u_i^{\mathcal{M}}} \times \frac{\mathbf{1}}{\boldsymbol{\eta}\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{u}_i^{\mathcal{M}}}\right)} \right) \qquad --eq\ (2)$$

where, $u_i^{\mathcal{M}}$ is the unit at i^{th} position of the sequence $\mathcal{M}$ and $\eta\left(\mathcal{G}_{u_i^{\mathcal{M}}}\right)$ is the weight of the group having unit at i^{th} position of the sequence $\mathcal{M}$.

Similarly, the probability $P(\mathcal{M}|\mathcal{Q})$ is the motif sequence $\mathcal{M}$ in the background sequence model $\mathcal{Q}$.

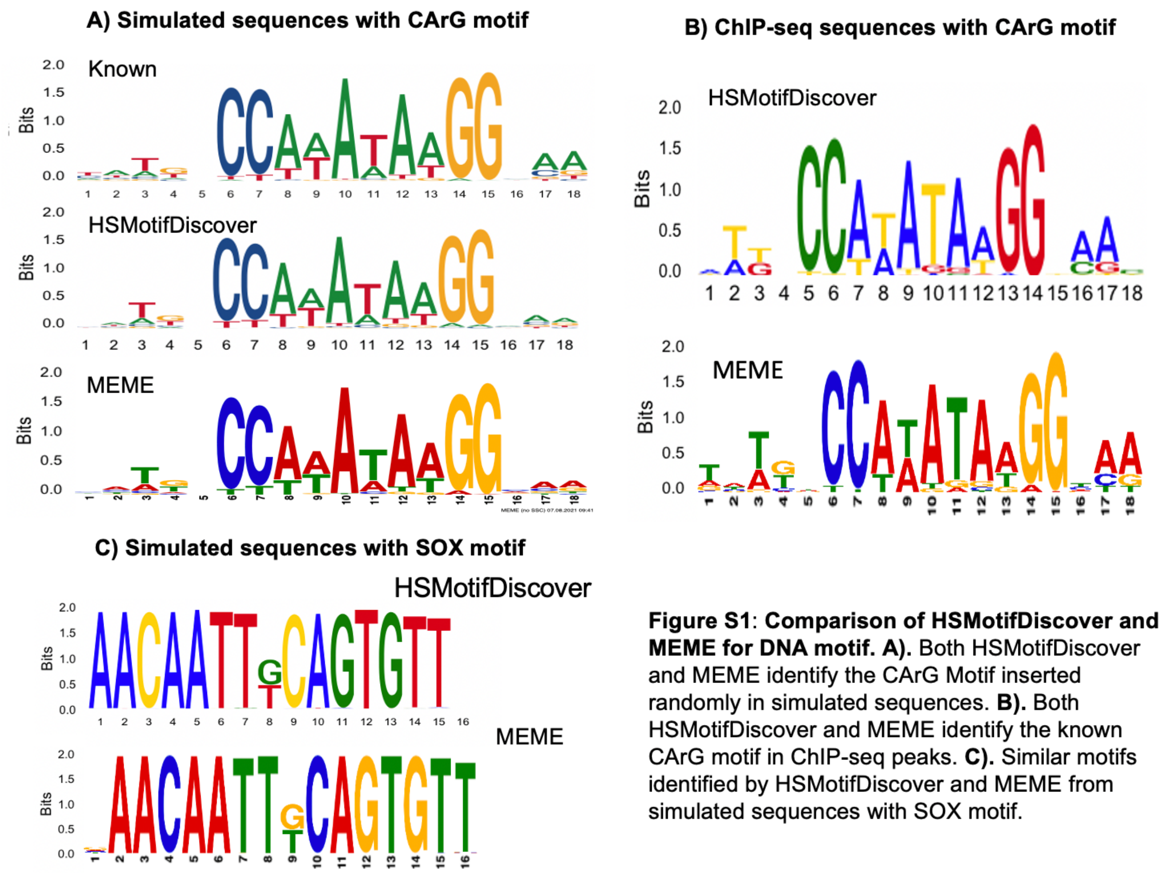
The likelihood that $\mathcal{M}$ is the true motif of interest for the obtained model $\mathcal{P}$ against the background $\mathcal{Q}$ can be represented by the equation

$$L(\mathcal{M}|\ \mathcal{P}, \mathcal{Q}) = \frac{P(\mathcal{M}|\mathcal{P})}{P(\mathcal{M}|\mathcal{Q})} \qquad --eq\ (3)$$

The quality of PSSM model $\mathcal{P}$ over the background model $\mathcal{Q}$ is obtained by using Kullback-Leibler divergence (KL_{Div}) measure (Kullback and Leibler, 1951). KL_{Div} was also used as objective function of Gibbs sampling algorithm optimization. How well a match of the given sequence to the motif PWM is quantified by p-value (Tremblay, 2021)

2. **Test HSMotifDiscover and MEME on DNA motifs:**

We compared the performance of HSMotifDiscover to the widely used MEME tool, on identifying motifs in randomly simulated DNA sequences of variable length between 50 to 100 bp, embedded with the CArG or SOX motif sequences (**Figure S1**). Note that MEME also uses Gibbs sampling. The result indicates that HSMotifDiscover was able to discover the target motifs, as MEME. Likewise, both tools successfully identified the known CArG motif in a randomly selected subset of serum response factor (SRF) binding ChIP-seq peak sequences, obtained from the ENCODE database (**Figure S1B**).


A) Simulated sequences with CArG motif

B) ChIP-seq sequences with CArG motif

C) Simulated sequences with SOX motif

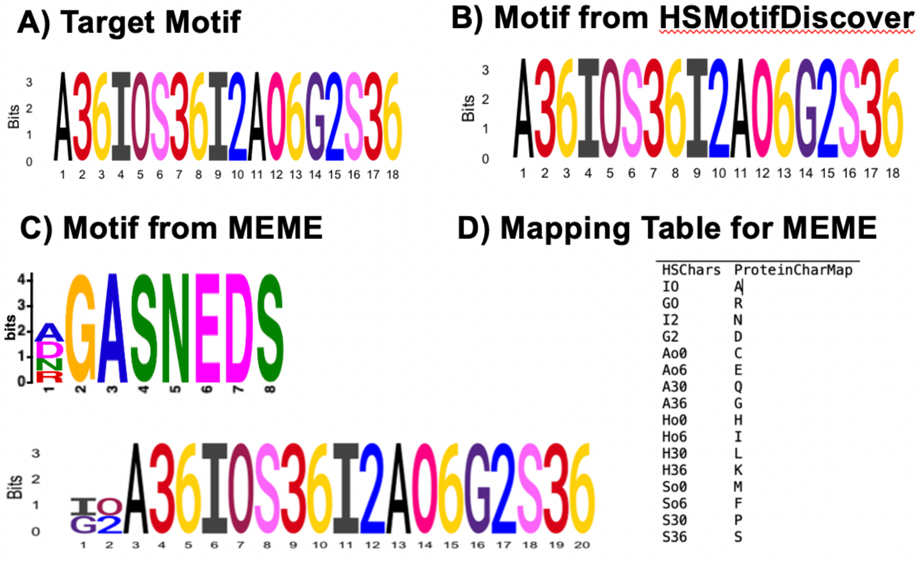
Figure S1: **Comparison of HSMotifDiscover and MEME for DNA motif. A).** Both HSMotifDiscover and MEME identify the CArG Motif inserted randomly in simulated sequences. **B).** Both HSMotifDiscover and MEME identify the known CArG motif in ChIP-seq peaks. **C).** Similar motifs identified by HSMotifDiscover and MEME from simulated sequences with SOX motif.

3. **Simulation of HS sequences with embedded motifs:**

HS motif sequences (n = 200), each of the length equal to the number of PSSM columns, were simulated, where the HS residue at a given position was sampled from the probability distribution of HS residues at the given position in PSSM. These motif sequences were then inserted at random locations in 200 randomly generated HS sequences of length 22 to 23 saccharide residues. When two PSSM motifs were involved, 100 sequences with one motif and another 100 with the other motif were generated and combined as inputs, and furthermore the resultant HS sequences were associated with equal weights or weights that were favor for one of the two motifs.

4. **Compare HSMotifDiscover and MEME on finding HS motifs:**

   4.1. To illustrate that grouping function in HSMotifDiscover is important, which is not in standard MEME, we applied both tools on 100 simulated HS sequences (containing motif 1 in **Figure 1B**) after converting the basic HS units to amino acid codes for MEME. As shown in the **Figure S2** below, our tool can detect the true motif (**Figure S2A/B**), while motif from MEME is slightly off, starting with an additional unit (**Figure S2C**). In this analysis, HS sequences were not weighted, and our tools were used to make the back-and-forth translation and plot the final motif logos, suggesting our tools may be used as extensions to MEME and other sequence analysis tools.



A) Target Motif

B) Motif from HSMotifDiscover

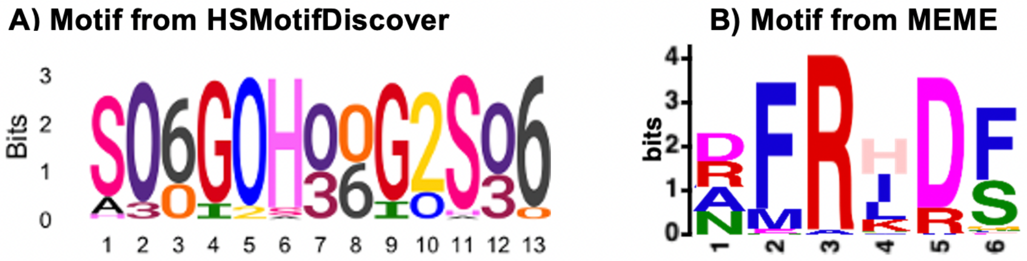
C) Motif from MEME

D) Mapping Table for MEME

HSChars	ProteinCharMap
IO	A
GO	R
I2	N
G2	D
Ao0	C
Ao6	E
A30	Q
A36	G
Ho0	H
Ho6	I
H30	L
H36	K
So0	M
So6	F
S30	P
S36	S

Figure S2: **HS motifs from HSMotifDiscover and MEME . A).** Target motif embedded in 100 simulated HS sequences. **B).** Motif from HSMotifDiscover. **C).** Motifs from MEME in single letter amino acid codes (top) and HS codes (bottom). D). Map between HS code and amino acid codes used to run MEME

4.2 In more complicated cases, the two types of saccharides in HS sequences may be inappropriately put in the same position by MEME but not by HSMotifDiscover. In this example run, HSMotifDiscover discovered the motif (**Figure S3A**), where glucuronic acid (representing by one letter and one digit) and glucosamine (represented by one letter and two digits) are separated without misalignment. However, the motif discovered by MEME (**Figure S3B**) from the same input HS sequences showed that at the 2nd position (character "D"), which is glucuronic acid, is aligned with character "F" and "M," which are glucosamine in the original HS sequences (**Figure S2D**). To what extent this is important will need substantially more tests in the future with a large number of HS sequences from real experimental data. This example nevertheless indicates that we need software like HSMotifDiscover to find motifs that are beyond MEME capability. Also note that the HS sequences from the SARS-Cov-2 are too short for MEME.



A) Motif from HSMotifDiscover

B) Motif from MEME

Figure S3: HS motifs from HSMotifDiscover and MEME where two types of HS saccharides were misaligned by MEME.

References:

Kullback,S. and Leibler,R.A. (1951) On Information and Sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Tremblay,B.J.-M. (2021) universalmotif: Import, Modify, and Export Motifs with R., R package version 1.12.1, https://bioconductor.org/packages/universalmotif/.