Greedy motif-based approach to parsing large and diverge coiled-coil proteins into domains

Hamed Khakzad[1,2], Johan Malmström[3] and Lars Malmström[2*]

[1]Faculty of Science, Institute for Computational Science, University of Zurich, Switzerland
[2]Service and Support for Science IT (S3IT), University of Zurich, Switzerland
[3]Division of Infection Medicine, Department of Clinical Sciences, Lund University, Sweden

*E-mail address of the corresponding author: lars.malmstroem@uzh.ch

Introduction

The rise of superbugs resistant to all FDR-approved antibiotics poses a massive health issue predicted to kill 10 million people per year in 2050, more than heart disease and cancer combined [1]. A potential solution is to target surface proteins for a given species or strain to decrease the viability of that bacterium in the host; one needs to know the identity of the infectious agent for this to work, something that the recent development of fast and accurate diagnostics might enable. Bacterial surfaces are complex, built of from membranes, peptide-glycans and, importantly, proteins. The proteins play crucial roles as the key regulator of how the bacterium interacts with its environment. Unsurprisingly, these proteins play key roles in the infectious process, and the number of protein-protein interactions is surprisingly high. Over 150 surface proteins found in S. pyogenes, an important human pathogen, binds about 250 human plasma proteins in significant amounts (as measured by enrichment over background plasma) [2]. High-resolution models of the protein-protein interactions enable the design of inhibitors, for example, large L/D-mixed synthetic circular peptides [3]. The majority of these models were created using high-resolution data such as NMR or X-ray crystallography, but these datasets remain challenging to produce. One of the key surface proteins of S. pyogenes is the M protein, a coiled-coil homodimer that extends over 500 A from the cell wall; the M protein is thought to bind several plasma proteins such as fibrinogen [4] and albumin [5]. The crystal structure of M and fibrinogen was published in 2011, and the authors postulate that the M and fibrinogen form a complex structure on the surface of the bacterium. Further, the M protein is composed of several repeats that are present a variable number of times; some of these repeats overlap with protein-protein interactions. A full catalog of the motifs in coiled-coil proteins and their relative conservation grade is a pre-requisite to target the protein-protein interaction that bacterial surface protein makes to host proteins [6]. In this paper, we present a greedy approach to iteratively identify conserved motifs in large sequence collections, identify all occurrences of these motifs and mask them. Remaining unmasked sequences are subjected to the second round of motif detection until no more significant motifs can be found or all protein segments have been assigned to a motif. We present the results for the S. pyogenes M protein.

Methods

We collected a large sequence collection of M proteins from four sources: Patric BRC, genomes we have previously sequenced and assembled, the M database from CDC and TrEMBL. Any protein sequence without anchor or signal peptide was discarded, and the remaining sequences were reduced to a 98% sequence identity using CD-HIT [7]. As Figure 1 shows, MEME [8] was applied to the sequence collection, restricting the number of found motifs to a single motif. Motif occurrences were discovered in the sequence collection using FIMO [9] and only occurrences with e-values of 1e-6 or lower were considered. Proteins were split using the occurrences and remaining parts longer than ten amino acids were carried

forward to create a new sequence collection, mixed with full-length and partial proteins. This sequence collection was the input iterative rounds of MEME, FIMO, split until no more significant motifs could be discovered, or all remaining sub-sequences were below ten amino acids. All the motif occurrences with corresponding features were stored in an SQLite database. Known motifs were integrated from [5] and the InterPro database and visualized using pViz.js [10].

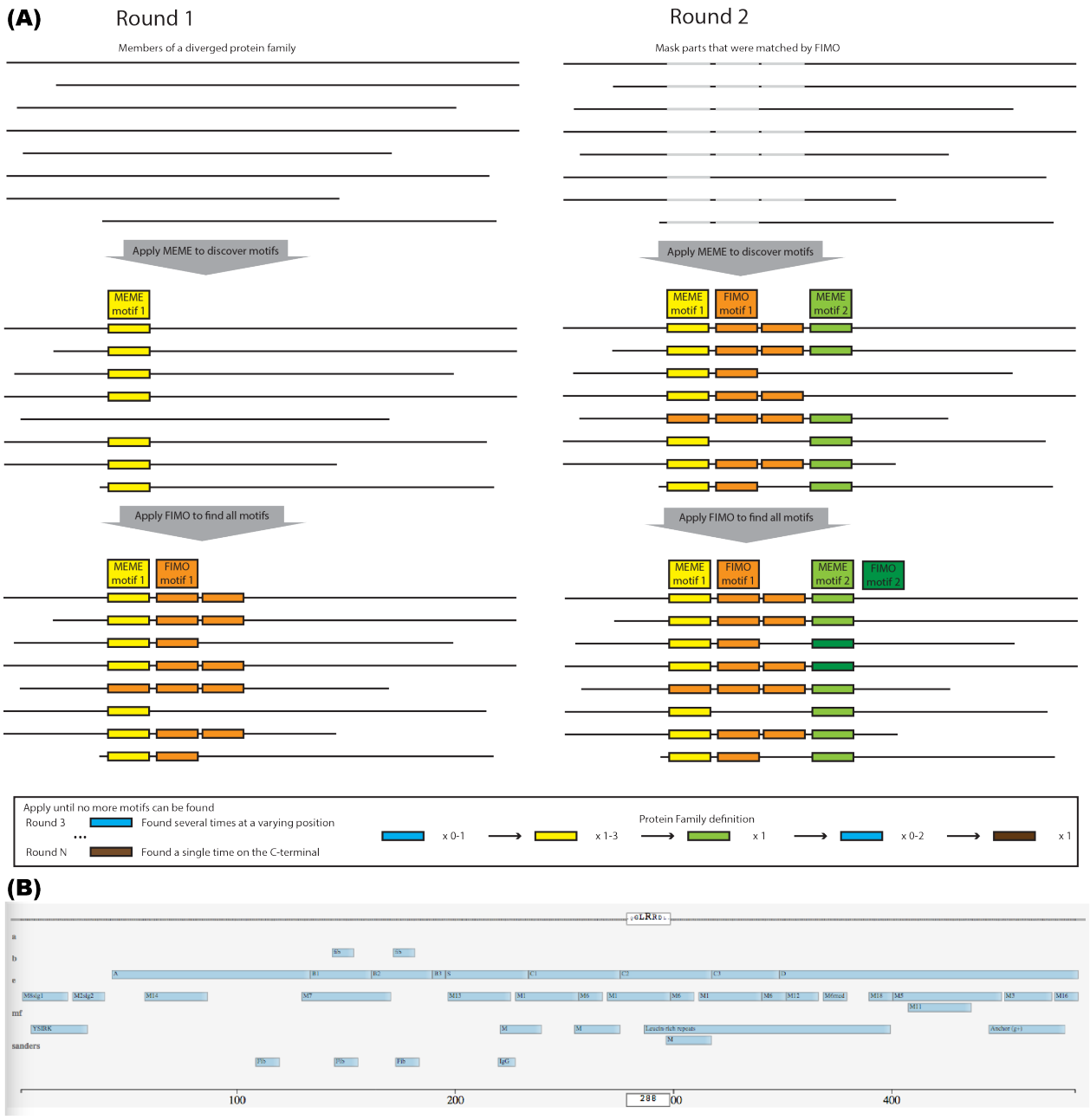

Figure 1: **A**- Two rounds (first to the left, second to the right) of the algorithm is displayed. It starts with a collection of sequences and then discovers motifs in this collection using MEME. It uses FIMO to find additional occurrences of the motif within the sequence collection. In the second round, the motifs are masked (gray bars) before MEME is applied once more. The algorithm iterates until no more motifs are found, or the sequence collection is fully annotated. **B**- The auto-generated result of our approach on M1 protein.

Results

We parsed an M protein sequence collection of 123 sequences, and the algorithm ended after 18 rounds, resulting in 18 motifs. We used the SF370 M1 protein as the reference which contains motifs M01-M03, M05-M08, M11-M14, M16-M17 but not M04, M09-10, M15, and M18-20. We found a total of 123 architectures and of these, 85% (104) are associated with a single serotype. Architecture m01:3, m02:1, m03:1, m05:1, m06:4, m07:3, m08:1, m11:1, m12:1, m14:1, m16:1, m17:1 is the architecture that exists in most serotypes (emm52, emm23, emm16, emm83, emm10). For the 18 emm1 proteins, we identified seven architectures, see Table 1 for emm1 architectures. For emm1.1, we see that M08 and M02 are the first and second part of the YSIRK signal peptide. M03 largely overlap with the anchor region. M1 and M6 correspond to the C repeats, M07 overlaps with the B repeats although we fail to identify the second and third B-domain. M13 finds the S region and M14 overlap partly with the A domain. The D domain is largely split into several motifs - M12, M18 followed by M05. These results are also shown in Figure 1-B.

Table 1: Architectures identified for emm1 protein.

m07:3	m01:1	m06:2	m05:1	1
m07:3	m01:2	m06:4	m05:1	1
m07:3	m01:3	m06:4	m05:0	1
m07:3	m01:4	m06:5	m05:1	1
m07:3	m01:2	m06:3	m05:1	2
m07:2	m01:3	m06:4	m05:1	2
m07:3	m01:3	m06:4	m05:1	10

Conclusions

In this paper, we demonstrate a proof-of-principle approach to parsing large sequence families into motifs using a greedy approach. This simple approach can easily handle situations where parts of proteins are repeated or re-arranged, and this can be time-consuming using other approaches. We observe that we over-parse some domains, but also observe that many of these large domains are only partly conserved over the sequence collection. Our results recapitulate the [5] paper, but in a completely automated fashion allowing us to scale to an arbitrary number of protein families. Given the speed and flexibility of our approach, we believe it will be useful in breaking analyzing surface protein of pathogens as these proteins are under high selective pressure and therefore cannot be analyzed using more traditional approaches such as multiple-sequence alignments (MSAs). Our attempts to use various MSA algorithms failed due to high sequence variability in regions between motifs and the varying number of motifs. Also, motif searching approaches failed and only identified a small subset of the motifs our approach discovered. Preliminary data indicates that many of the newly discovered motifs are not always present together with adjacent motifs, indicating that they might have different and independent functions. Interestingly, many of our newly discovered motifs are not found in any of the emm1 strains, and some of these might be responsible for binding other ligands, such as plasminogen [11].

References

1. O'Neill J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Rev Antimicrob Resist. 2014. http://amr-review.org/Publications

2. Malmstrom J et al. Streptococcus pyogenes in Human Plasma: adaptive mechanism analyzed by mass spectrometry-based protomics. J. Biol. Chem. 2012. 287(2): 1415-1425. DOI:10.1074/jbc.M111.267674

3. Bhardwaj G et al. Accurate de novo design of hyperstable constrained peptides. Nature. 2016. 538:329?335. DOI:10.1038/nature19791

4. Macheboeuf P et al. Streptococcal M1 protein constructs a pathological host fibrinogen network. Nature. 2011. 472:64-68. DOI:10.1038/nature09967

5. Akesson P, Schmidt KH, Cooney J, Bjorck L. M1 protein and protein H: IgGFc- and albumin-binding streptococcal surface proteins encoded by adjacent genes. Biochemical Journal. 1994. 300(3):877-886. DOI: 10.1042/bj3000877

6. Sjoholm K. Targeted Proteomics and Absolute Protein Quantification for the Construction of a Stoichiometric Host-Pathogen Surface Density Model. Mol Cell Proteomics. 2017. 16(4):29-41. DOI:10.1074/mcp.M116.063966

7. Fu L et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012. 28 (23):3150-3152. DOI: https://doi.org/10.1093/bioinformatics/bts565

8. L. Bailey T et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009. 37:202-208. DOI: https://doi.org/10.1093/nar/gkp335

9. E. Grant C et al. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011. 27(7):1017-1018. DOI: https://doi.org/10.1093/bioinformatics/btr064

10. Mukhyala K et al. Visualization of protein sequence features using JavaScript and SVG with pViz.js. Bioinformatics. 2014. 30(23):3408-3409. DOI: 10.1093/bioinformatics/btu567

11. Ringdahl U et al. Analysis of Plasminogen-Binding M Proteins of Streptococcus pyogenes. Methods. 2000. 21(2):143-150. DOI: https://doi.org/10.1006/meth.2000.0985