Global and Local Protein Sequence Alignments of Selected Pathogenic Substrate Binding Proteins

Ipsita Krishnamurthy

May 14 2020

1 Introduction

During infection, *Staphylococcus aureus* (S. aureus) and other pathogens must acquire a broad array of nutrients, namely transition metal ions, such as manganese, zinc, and iron. These metal ions are necessary for basic cellular function, such as the regulation of gene expression, as well as structural and catalytic function of various metalloproteins (Radin et al., 2018). To combat invading pathogens, vertebrates leverage the essentiality of transition metals by restricting their availability in a process known as nutritional immunity (Kehl-Fie et al., 2013). Bacterial ATP-binding cassette (ABC) transporters play a critical role in nutrient acquisition under host-limiting conditions and serve as potential therapeutic targets to attenuate virulence. The *S. aureus* MntABC transporter is the primary transporter responsible for Mn(II) sequestration in metal-limiting conditions and is necessary for enzymatic neutralization of reactive oxygen species during the host immune response (Kehl-Fie et al., 2013; Handke et al., 2018).

The MntABC transporter can be broken down into its three constituent parts: MntA, an ATP-ase, MntB, a transmembrane domain, and MntC, an extracytosolic soluble binding

Key:
: single, fully conserved residue
. conservation between groups with

Figure 1: Protein sequence alignment of MntC and known soluble binding proteins from other pathogens. *S.* aureus MntC protein sequence is aligned to TroA, PsaA, and ZnuA sequences, sourced from Uniprot. ZnuA is a Zn(II) binding SBP in *E. coli*; TroA is a proposed Zn(II) transporter found in *Treponema pallidum*; PsaA is an Mn(II) and Zn(II) binding SBP from *Streptococcus pneumoniae*. Figure was generated using multiple sequence alignment program, Clustal Omega.

each MntC homolog and MntC. I hypothesize that the local alignments will highlight shared motifs between the SBPs that are relevant to function.

2 Computational Approach

The PAM250 matrix is frequently used to score aligned peptide sequences to determine the similarity of those sequences. The numbers given above were derived from comparing aligned sequences of proteins with known homologous proteins and determining the \accepted point mutations" (PAM) observed. Each column and row in the PAM250 matrix represents one of the 20 standard amino acids. A function to read the PAM250 matrix text file was graciously provided by Anna!

move east, j increases and i stays the same. In this case, the value at global_alignment[i][j] is equivalent to the value at global_alignment[i][j-1]. If we move south on our protein alignment, the value at global_alignment[i][j] is equivalent to the value at global_alignment[i-1][j]. If we move in a southeast direction across the protein sequence alignment, the amino acids in both sequences can match or mismatch. The penalty of this mismatch is determined by the PAM250 scoring matrix. When we move diagonally across a table, the value at the new node can be determined by a recurrence function. This value corresponds to the maximum score obtained by movement south, east, or diagonally. If we move diagonally across the table, the value at global_alignment[i][j] equals the value at global_alignment[i-1][j-1] + the penalty from the PAM 250 scoring matrix.

In global alignments, the largest score is the final score calculated at the \destination" or the bottom right of the global alignment table generated. As we build our global alignment, we also build a backtracking table that allows us to visualize the \path" that the algorithm takes to generate our alignment. We assign each movement as south (`S'), east (`E'), or diagonal (`D'). This backtracking table is helpful to visually check whether the program functions for smaller test global alignments and determine the direction of the path taken during the global alignment.

We will also write a local alignment function that will also require an input of two protein sequences, a penalty for insertion or deletion, and the PAM250 matrix in order to output the local alignments and final score. The algorithm is very similar to the algorithm for global alignment, with one notable difference. In a local alignment, we can imagine a \free taxi ride" that takes us from the source (0,0) in the local alignment table to the beginning of our alignment. The local alignment function adds zero-weight edges from the source to every other node, thereby allowing us to take this \free taxi ride" to any node in the graph. In the function, we define the free taxi ride denoted by the symbol `*'.

When the value of local_alignment[i][j] is less than 0, we take advantage of a \free taxi ride" from the source (0, 0) to skip that node. For this reason, the largest score from a local alignment is not the score at the end of the path, but rather is the largest score in the table. We can write a for loop that iterates through all the rows (i) and columns (j) of our local_alignment table to find the maximum value in the table. We can similarly use a backtracking matrix, as we did for the global alignment, to determine the direction of the path taken during the local alignment.

The repl.it with the above code is linked: https://repl.it/@IpsitaKrishnam1/Final-Project.

3 Results and Discussion

Results from the global and local sequence alignment are consistent with the previously derived multiple sequence alignment comparing MntC, PsaA, TroaA, and ZnuA. MntC and PsaA have the most similar sequences, followed by MntC and TroA, and last MntC and ZnuA. Global sequence alignment of MntC with the SBP PsaA yields a relatively high score of 757. The local sequence alignment of MntC with the SBP PsaA returns a high score of 834, supporting the hypothesis that a local sequence alignment would better predict sequential similarity across SBPs.

Global sequence alignment of MntC with the SBP TroA yields a score of 358. Local sequence alignment of MntC with the SBP TroA returns a score of 438, again supporting that a local sequence alignment potentially reveals greater sequential similarity between MntC and TroA than the mulitple sequence alignment or global sequence alignment.

Global sequence alignment of MntC with SBP ZnuA yields a score of 193. Local sequence alignment of MntC with ZnuA returns a score of 254.

Taken together, the results suggest that a local sequence alignment provides a better assessment of the similarity between the protein sequences of MntC with PsaA, TroA, and ZnuA respectively. As expected from the multiple sequence alignment initially run, PsaA and MntC have the highest conserved sequence identity of the SBPs tested (Figure 1).

I wanted to see whether the increase in score corresponded to an increase in the percentage of matching amino acids between each pair. As local sequence alignments compare the best matched substrings of the greater protein sequence, I hypothesized that the local sequence alignment would show a higher percentage of conserved residues in comparison to the global sequence alignment. Aafter computing the percentage of conserved residues, I found that a higher local alignment score did not necessarily correspond to a greater percentage conserved sequence identity. When we compare the global and local sequence alignments of PsaA and MntC, for example, the global alignment yields a 48% conserved sequence identity and the local alignment yields a 47$ conserved sequence identity, although the local alignment score is greater.

Similar comparisons between global and local sequence alignments between MntC and TroA reveal that global sequence alignment yields a 28% conserved sequence identity and a local sequence alignment returns a 30% conserved sequence identity.

The conserved sequence identity is even lower between MntC and ZnuA. The global sequence alignment results yield a 19% conserved sequence identity and local sequence alignment results yield a 20% conserved sequence identity between the SBPs.

In the future, I could run another algorithm to determine the longest common subsequence between pairs of proteins may be more helpful to visualize common motifs within the proteins. It seems unusual to me that structural homologs of MntC would be so different sequentially from each other.

Figure 2: Overlay of crystal structures of Zn(II)-bound MntC and PsaA (Ahuja et al., 2015; Lawrence et al., 1998). Crystal structures of MntC (shown in pink) and the PsaA (shown in orange) display highly conserved structural identity. A. The residues implicated of PsaA in Zn(II)-binding are shown in yellow. B. The residues implicated in metal ion coordination in both MntC (pink) and PsaA (shown in yellow) are highly conserved. In MntC, the residues involved in zinc ion coordination are D281, E206, H67, and H140. In PsaA, the residues involved in zinc ion coordination are D280, E205, H67, and H139. Labels in figure correspond to Zn(II)-bound MntC crystal structure sequence. Access codes for the Pdb files of the PsaA (1PSZ) and MntC (4NNO) crystal structures are available on https://www.rcsb.org/.

PsaA, for example, is highly structurally similar to MntC (Figure 2). PsaA and MntC have very conserved residues implicated in metal ion coordination (Figure 2B). The metal-coordinating residues for both proteins include an aspartic acid, a glutamic acid, and two histidines. PsaA also has been shown to bind irreversibly to Zn(II) (Couñago et al., 2014). As with Zn(II)-bound MntC, Zn(II) cannot be chelated from PsaA by the addition of EDTA

(Couñago et al., 2014; Ahuja et al., 2015). Couñago *et al.* show that the inability of PsaA to satisfy preferred octahedral coordination chemistry of Mn(II) allows for reversible Mn(II) binding (Couñago et al., 2014). Tetrahedral coordination of Zn(II), however, locks PsaA into a closed conformation upon binding (Couñago et al., 2014). The similarity in both the binding behavior and metal-binding sites of PsaA and MntC suggests that there *should* be common motifs between the MntC and PsaA sequences.

Increasing the indel penalty to -10 only reduces the overall score observed, but doesn't modulate the length of the substrings compared during the local sequence alignment. This suggests that comparing the longest common subsequences, which should return shorter shared motifs between pairs of SBPs (given the observed lack of conserved residues), may elucidate shared sequence motifs in the protein.

This project sought to determine whether local sequence alignments between MntC and its structural homologs would reveal any conserved motifs hidden during multiple sequence alignment. I found that the local and global sequence alignments of MntC and other SBPs yield very similar results. The local sequence alignments I ran did not reveal any motifs obscured by the multiple sequence alignment.