Identification and Clustering Network of Virulent *Aeromonas Hydrophila* C16-13425 Hypothetical Proteins

Harun Pirim[1], Hasan C. Tekedar[2], Matt J. Griffin[2], Geoffrey C. Waldbieser[3], Larry A. Hanson[2]

[1]Industrial and Systems Engineering, Mississippi State University, [2]College of Veterinary Medicine, Mississippi State University, [3]Agriculture Research Service, U.S. Department of Agriculture

introduction

•Aeromonas hydrophila is a Gram negative mesophilic species ubiquitous in aquatic environments that causes infections in multiple host species, including fish.

•The U.S. channel catfish industry has been affected by virulent *A. hydrophila* (vAh) since 2009 and caused extensive mortalities and economic losses to the channel catfish industry in the United States.

•We sequenced the complete genome of an *A. hydrophila* strain C16-13425 that was isolated from an outbreak of Aeromonas septicemia in catfish from a commercial production pond in Mississippi.

•Many proteins (1082 out of 4879) from its genomes are not assigned a role.

•These unknown proteins are called hypothetical proteins and they remain to be elucidated so that their function and potential biological roles could be identified and assigned.

•We filtered the HPs through a pipeline similar to the one in (1).

•Pfam (2) and CATH databases (3) are used to retain hypothetical proteins at consensus.

•83 sequences were in common. These sequences are submitted to Blast (4), DEG (5), and PSORTdb (6) databases to obtain information about homolog sequences, essential genes, and subcellular localization.

•Results from these databases are summarized as a single table. Top 5 entries of the table are shown in Table 1.

network construction and analysis

•The table columns are features employed to construct a weighted similarity network of relationships between hypothetical proteins.

• Normalized Gower distance is computed using cluster package in R. The distance matrix is converted to a similarity matrix using the relationship $s_{ij} = 1 - d_{ij}$.

•Similarity scores are applied the threshold of 0.85 to construct a network with strong or the most similar relationships.

•A community structure finding algorithm is applied and three distinct clusters are formed.

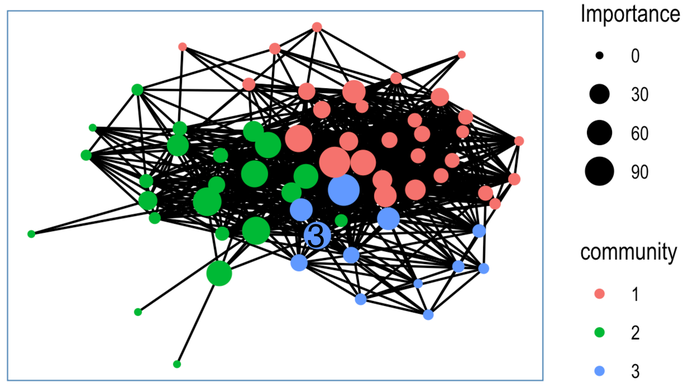

Figure 1: HP clusters with distinct colors

•These clusters of hypothetical proteins are to be investigated further.

•Figure 1 illustrates the clusters with different colors and the size of the nodes are proportional to the centrality (degree) of the nodes.

Table 1: Merged information from three databases

Sequence	Count_blast	AvgIdBlast	Unknown_psort	Cytoplasmic_psort	Periplasmic_psort	OuterMembrane_psort	CytoplasmicMembrane_psort
1	2.00	93.18	12.00	0.00	0.00	0.00	0.00
2	2.00	96.18	23.00	0.00	0.00	0.00	0.00
3	2.00	92.86	2.00	15.00	0.00	0.00	0.00
4	1.00	100.00	2.00	0.00	12.00	0.00	0.00
5	2.00	97.22	0.00	9.00	0.00	0.00	0.00
Sequence	**AvgIdentity_psort**	**DEGEukar_MaxID**	**DEGEukar_totalHits**	**DEGArchea_MaxID**	**DEGArchea_totalHits**	**DEGBacteria_MaxID**	**DEGBacteria_totalHits**
1	95.75	46.43	11.00	43.48	8.00	60.87	23.00
2	94.00	41.67	11.00	26.56	2.00	31.25	9.00
3	94.00	45.83	11.00	50.00	7.00	41.94	15.00
4	95.50	40.32	14.00	35.29	4.00	38.64	12.00
5	96.11	40.54	16.00	42.11	5.00	51.85	12.00

results and conclusion

•String database reports connection between hypothetical protein SeqID2 (from green cluster), SeqID3 (from blue cluster), SeqID4 (from red cluster) and the protein AHV34012.1 identified as RTX toxin from *Aeromonas hydrophila* YL17.

•Connections are circled in Figure 2.

•In conclusion, this research aims to reveal hypothetical proteins which interact closely making use of the available information in an efficient way.

•Network representation and analysis provides distinctive results about potential toxin proteins to investigate further.

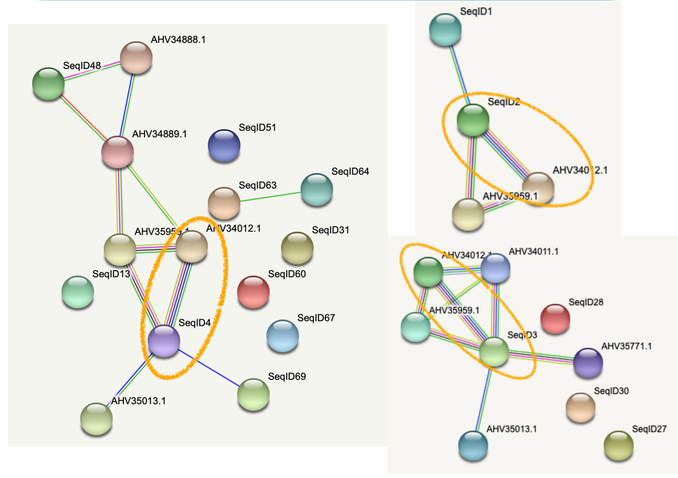

Figure 2: HP and RTX toxin relations

references

1. **da Costa WLO, Araujo CLA, Dias LM, Pereira LCS, Alves JTC, Araujo FA, Folador EL, Henriques I, Silva A, Folador ARC.** 2018. Functional annotation of hypothetical proteins from the Exiguobacterium antarcticum strain B7 reveals proteins involved in adaptation to extreme environments, including high arsenic resistance. PLoS One **13:**e0198965.

2. **Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A.** 2021. Pfam: The protein families database in 2021. Nucleic Acids Res **49:**D412-D419.

3. **Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA.** 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res **35:**D291-297.

4. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-410.

5. **Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, Zhang R.** 2021. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res **49:**D677-D686.

6. **Lau WYV, Hoad GR, Jin V, Winsor GL, Madyan A, Gray KL, Laird MR, Lo R, Brinkman FSL.** 2021. PSORTdb 4.0: expanded and redesigned bacterial and archaeal protein subcellular localization database incorporating new secondary localizations. Nucleic Acids Res **49:**D803-D808.