

# Identification of Putative Virulence Factors by Genome-Scale Analysis of Large Double-Stranded DNA Viruses

Stefan Ponko<sup>1</sup>, Steven Wiley<sup>2</sup>, Ajamete Kaykas<sup>1</sup>



<sup>1</sup>VLST Corp, 307 Westlake Avenue North, Suite 300, Seattle, WA 98109, sponko@vlstcorp.com, akaykas@vlstcorp.com

<sup>2</sup>Imdaptive, 3010 NW 56th St. Seattle, WA 98107

## Introduction

We have developed a database of large double-stranded DNA virus genomes to identify functional protein families and potential virulence factors. The database, and associated toolset is used to identify and prioritize potential virulence factors based on protein motif(s), topology, homology to human proteins, host species, deletion in lab strains, non-requirement for replication *in vitro*, and distribution across viral species. Prioritized virulence factors are tagged, expressed and screened using immunoprecipitation follow by mass spectrometry (IP-MS) to identify binding partners. The cellular binding partners are then evaluated for their potential as therapeutic intervention points for treating autoimmune diseases and inflammation.

## Selection Process

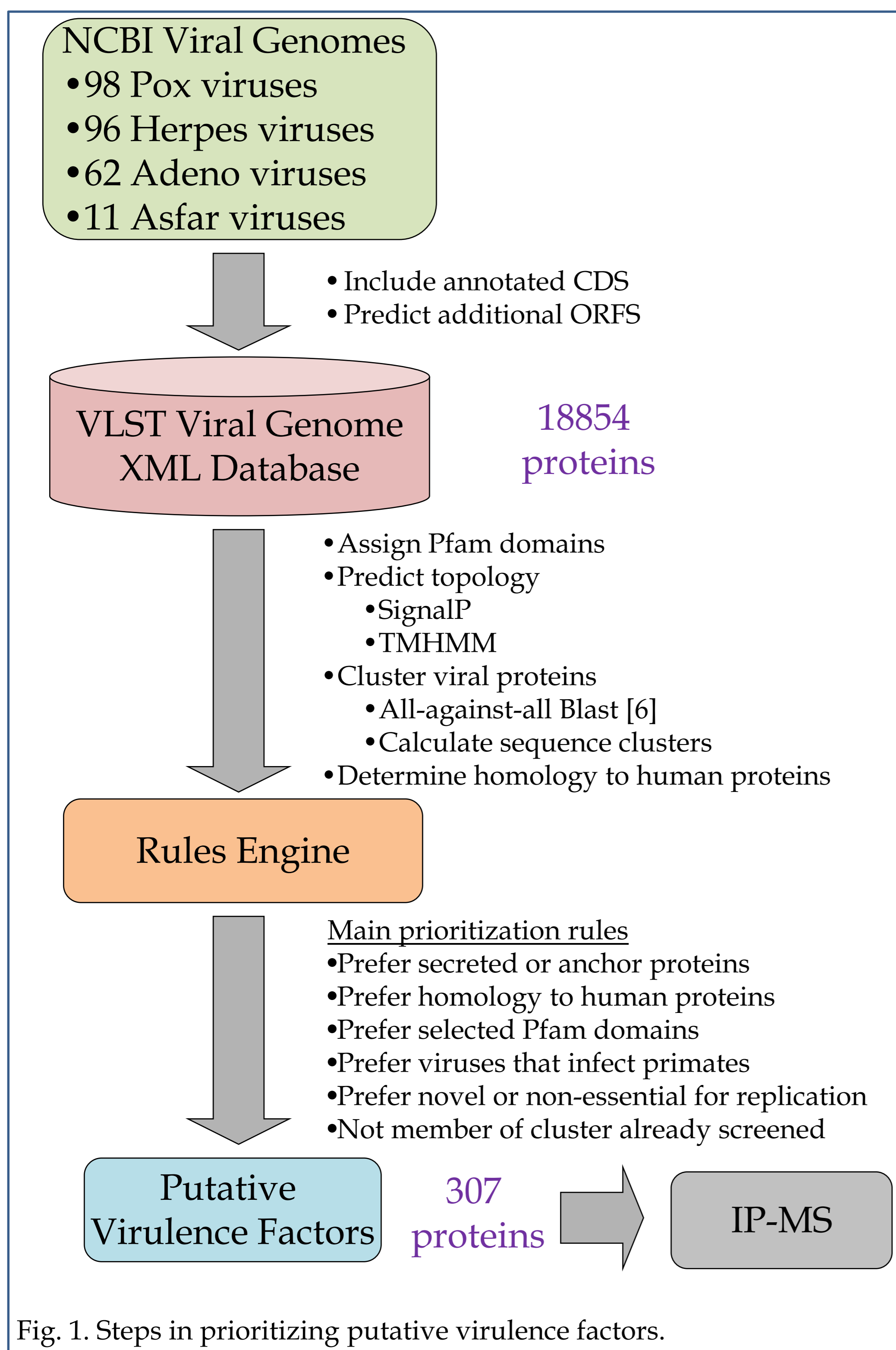


Fig. 1. Steps in prioritizing putative virulence factors.

## Methods

The viral database consists of 98 Pox, 96 Herpes, 11 African Swine Fever, and 62 Adenoviral genomes, and translations obtained from Genbank [1]. Translations are combined with our predicted ORFS to create a unique set of viral proteins. SignalP [2] and TMHMM [3] are used to identify potential secreted and transmembrane proteins. Each protein is analyzed by Pfam [4] to identify functional families, and all matching Pfam families are tracked with each protein record. In addition, we apply quality threshold (QT) clustering [5] across the entire dataset of 18854 viral proteins to group proteins by sequence homology. Combining the two strategies allows us to determine distributions of functional families across viral genomes (Table 1). Additionally, when new viral proteins are identified for binding partner screening, clusters containing previously screened proteins may be excluded from further consideration, allowing maximum initial screening diversity.

Cluster	Members	Pfam Families
1	115	ASFV_360 (some ANK)
2	106	Serpin (some GIY-YIG, VAR1)
3	100	BTB, BACK, Kelch_1 repeats (and some VAR1)
4	94	DNA_pol_B_exo, DNA_pol_B
5	92	Adeno_E1A
6	87	DUF249 (some ANK)
7	82	ANK repeats
8	82	v110
9	78	DNA_pack_C, DNA_pack_N, Terminase_6, Terminase_1
10	72	Adeno_terminal (some DUF2382)

Table 1. Pfam families for top 10 largest protein clusters.

A rules engine from the JBoss organization [7] scores viral proteins by applying user-defined rules (listed at left) based on various protein attributes, set and cluster membership. Each rule is weighted to reflect its relative importance. Proteins with higher relative scores are preferentially selected for screening. The scores and rules are saved back to the database for future reference.

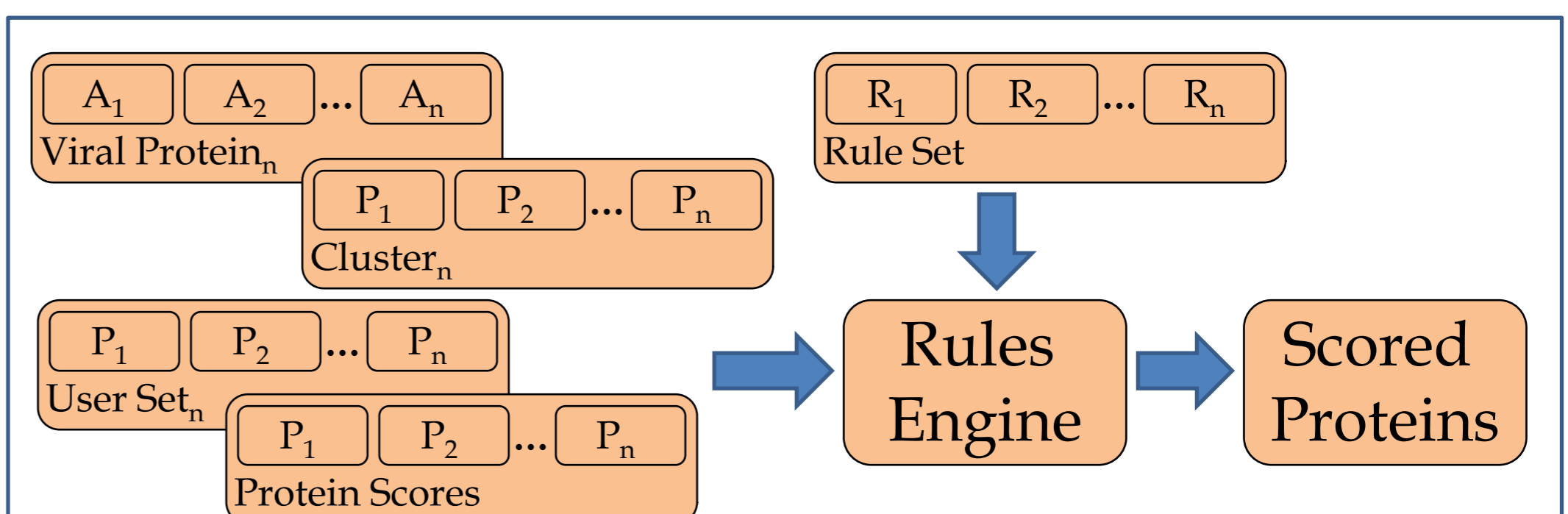


Fig. 2. User-defined rules are applied to score proteins based on protein attributes, as well as cluster and user-defined set membership. A=protein attribute, e.g. topology, P=viral protein, R=user-defined rule

A web application provides a user interface for viewing the viral genomes, proteins, homologs to human proteins in the IPI database [8], Pfam domains, and QT cluster membership and for maintenance operations such as downloading, and processing the viral proteins.

## Summary

- We have developed an extensive database of dsDNA viral proteins.
- A rules engine prioritizes putative virulence factors for screening for cellular binding partners.
- To date, we have identified 307 potential virulence factors for which binding partners are currently being identified by IP-MS.

## References

- [1] Benson, D.A., Karsch-Mizrachi, L., Lipman, D.J., Ostell, J. and Wheeler, D.L. 2003. Genbank, *Nucleic Acids Research*, 31(1):23-27.
- [2] Bendtsen, J. D., Nielsen, H., von Heijne, G. and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal Of Molecular. Biology*, 340:783-795.
- [3] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Mol. Biology*, 305:567-580.
- [4] Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L., Studholme D.J., Yeats C. and Eddy S.R. 2004. The Pfam protein families database. *Nucleic Acids Research*, 32:D138-D141.
- [5] Heyer L.J, Kruglyak S. and Yooseph S. 1999. Exploring expression data: identification and analysis of coexpressed genes. *Genome Research*, 9(11):1106-15. Review.
- [6] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389-3402.
- [7] JBoss DROOLS, v4.0.7, <http://www.jboss.org/drools>
- [8] Kersey P. J., Duarte J., Williams A., Karavidopoulou Y., Birney E. and Apweiler R., 2004. The International Protein Index: An integrated database for proteomics experiments. *Proteomics*, 4(7):1985-1988.