Modeling the Tertiary Structure of a Multi-domain Protein

Fahad Almsned
Krasnow Institute for Advanced Study, School of Systems Biology, George Mason University

Gideon Gogovi
Center for Simulation and Modeling, Dept of Computational and Data Sciences,
George Mason University

Nicole Bracci
National Center for Biodefense and Infectious Diseases,
School of Systems Biology,
George Mason University

Kylene Kehn-Hall*
National Center for Biodefense and Infectious Diseases,
School of Systems Biology,
George Mason University

Estela Blaisten-Barojas†
Center for Simulation and Modeling, Department of Computational and Data Sciences,
George Mason University

Amarda Shehu‡
Dept of Computer Science,
Dept of Bioengineering,
Center for Simulation and Modeling,
School of Systems Biology,
George Mason University

ABSTRACT

Due to the central role that tertiary structure plays in determining protein function, resolving protein tertiary structures is an integral research thrust in both wet and dry laboratories. Dry laboratories have primarily focused on small- to medium-size proteins. However, proteins central to human biology and human health are often quite complex, containing multiple domains and consisting of thou- sands of amino acids. Such proteins are challenging for various reasons, including the inability to crystallize. We present a case study of structure determination for the Rift Valley fever virus L-protein, a a large, multi-domain protein with currently no available tertiary structure. We employ this case study as an emerging paradigm and demonstrate how to leverage the rich and diverse landscape of bioinformatics tools for building tertiary structure models for multi-domain proteins with thousands of amino acids.

CCS CONCEPTS

• **Applied computing → Molecular structural biology**;

KEYWORDS

multi-domain protein, tertiary structure, structure determination

ACM Reference Format:
Fahad Almsned, Gideon Gogovi, Nicole Bracci, Kylene Kehn-Hall, Estela Blaisten-Barojas, and Amarda Shehu. 2018. Modeling the Tertiary Structure of a Multi-domain Protein. In *ACM-BCB’18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3233547.3233702

1 INTRODUCTION

Tertiary structure plays a central role in determining the biological activity of a protein, as in the cell proteins adapt their structure to complement the structure of their molecular partners [4]. Due to the premise that tertiary structure is crucial for unraveling the biological activity of a protein, resolving protein tertiary structures is a compelling research thrust in both dry and wet laboratories [27]. Traditionally, in the absence of close sequence homologs with known tertiary structures in the Protein Data Bank (PDB) [2], dry laboratories have either focused on template-free structure prediction limited to small- to medium-size proteins [17, 21, 22, 35], threading-based structure prediction [25], or a combination of the two [38]. Template-based modeling has become more popular than template-free modeling. This is primarily due to the fact that, in principle, template-based modeling allows building of protein tertiary structures possessing sequences with more than 300 amino acids. However, the quality of this approach relies on identifying remote homologs with known well-resolved structures [10]. Proteins with thousands of amino acids are currently problematic test cases for tertiary structure determination in dry laboratories. Such proteins also challenge wet laboratories, as their size often makes it difficult to achieve crystallization.

Currently, there are no methods or protocols for holistically determining tertiary structures of multi-domain proteins with long sequences of amino acids. Efforts to build tertiary structures of multi-domain proteins are limited to assuming that the structure of the composing domains are known. As a consequence, the discovery of tertiary structures focuses on exploring how the different domain structures can be arranged in 3D in search of low-energy stable arrangements that are potential tertiary structures of the multi-domain protein [33, 34]. However, some of the most interesting cases in viral biology, as showcased here via the RVFV L-protein, involve multi-domain proteins with little or no structural information at the domain level. To make matters worse, the delineation of domain boundaries may also be unknown or uncertain.

In this paper we present a case study of tertiary structure determination for the Rift Valley fever virus (RVFV) L-protein containing 2092 amino acids and use this case study as a driving example of the complexity and the various considerations involved. As we demonstrate, this protein is a good example that illustrates how the

[*] Corresponding Author: kkehnhal@gmu.edu
[†] Corresponding Author: blaisten@gmu.edu
[‡] Corresponding Authors: ashehu@gmu.edu

ACM-BCB’18, August 29-September 1, 2018, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233702

research community can leverage the currently rich and diverse landscape of bioinformatics tools to build structural models for multi-domain proteins with thousands of amino acids.

The paper is organized as follows. Section 2 describes the relevant chemical components of the RVFV L-protein, enumerates the different servers used for determining the domain sizes, their boundaries and structure, and details the protocol for assembling the protein domains. Section 3 contains our results from the process of selecting the more reliable structure models for the adopted domains and their evaluation. This section describes the mechanism of assembling the domains together and provides three predicted RVFV L-protein tertiary structures. The paper is concluded in section 4.

2 METHODS SURVEY

Homology or comparative modeling is a reliable approach for predicting tertiary structure based on given amino acid sequences. However, search of the PDB [2] with the NCBI Basic Local Alignment Search Tool (BLAST) [1] for proteins homologous to the full 2092 amino acids sequence of the RVFV L-protein displaying 20% or better identity [9] was negative. Search with NCBI SMART BLAST [26] for exploration of the landmark database of sequences reveals that the RVFV L-protein sequence is homologous to those of the Ambe, Joa, and Salobo viruses with identities in the range 60 − 61%. Unfortunately, the tertiary structure of these viruses polymerases is also unknown.

RVFV is a tri-segmented negative-sense RNA virus consisting of an S, M, and L segment [6]. The L segment encodes the RNA-dependent RNA polymerase (RdRp) and is known as the L-protein. The L-protein is 2092 amino acids long containing an RNA polymerase domain and an endonuclease domain. Few studies have been performed specifically on RVFV L-protein. Therefore much of the knowledge about the protein is gained from related viruses. The primary function of the L-protein is to perform viral mRNA transcription and genomic replication of viral RNA. The L-protein forms L-L dimers through interaction between the N-terminus (a.a. 1-222) and the C-terminus (a.a. 1219-2092) of the protein [37]. Within the polymerase domain, there are SDD sequence motifs that are highly conserved among negative stranded RNA viruses [20, 31].

In the absence of a homologous sequence with known structure, we propose to create a structural model by identifying dominant domains of the entire RVFV L-protein sequence that have known structure and assembling the partial structures into a single model. To implement the approach one may use fully-automated servers, but we devise a strategy for controlling the domain identification leading to template structures that incorporate biological insight of the system under study.

Our first alternative is RAPTORX [23]. This server recurs to remote homology recognition and fully-automates the process of structure prediction of proteins with long sequences by determining if the targeted sequence consists of multiple domains or not. Next, the PDB is searched automatically for structures remotely homologous to the domains that are used as templates. High-quality structural models for numerous targets with only remote homology templates have been predicted by this process [14]. Advantages of RAPTORX include a high quality target-templates alignment algorithm, a novel nonlinear scoring function, and a probabilistic-consistency algorithm. Currently, RAPTORX processes a sequence of 200 amino acids in about 3-5 minutes. Obtaining the tertiary structure of he RVFV L-protein sequence took 3 days.

The second alternative consists of identifying the domains incorporating biological insight without recurring to structure homology. Multiple options exist for domain identification. For example, conserved regions can be identified via NCBI DELTA-BLAST [5] in knowledge-based databases, such as the Pfam database [13]. Another possibility is the use of domain identification servers such as ThreaDomEX [32], Dobo [12], DomCut [29], and DomPred [15]. However, the outcome from these two approaches might differ in the number of domains predicted as well as in their boundaries. Therefore, biological insight becomes crucial for finalizing the domains identification. In our example of the RVFV L-protein, identification of three domains is information suggested from the polymerases of the arenavirus. We use this insight to reconcile findings on the number of domains and the domain boundaries, as explained in section 3.

Once the domains are identified, a tertiary structure should be modeled for each domain. A route often used consists of assembling *ab initio* protein models such as those delivered by ROSETTA [28] and QUARK [35]. However, these approaches are not viable when the domain exceeds 200 amino acids. Instead, threading-based methods, such as I-TASSER [25], become useful by automatically identifying structural templates based on sequence-environment alignments. The resulting structural models are ranked on a statistical-based *goodness* score. This scoring method is not sufficient for asserting which is the best structure of the domain under investigation. More options need to be considered. Biological insight can leverage the process by proposing specific structural templates that are not obviously obtained by I-TASSER. In addition, a Molecular Dynamics protocol can be considered to produce dynamically the structure of a domain as summarized in the next paragraph. Moreover, each of the structural templates obtained via these different approaches can be further energetically refined/minimized.

For the identified domains containing the protein N- and C-termini, we conduct all-atom molecular dynamics (MD) simulations in implicit solvent [30] at constant pH = 7.4 [18] using AMBER and the ff12SB force field [7]. The domains considered are simulated at a constant temperature of 310 K via the Langevin thermostat with a collision frequency of 5 ps^{-1} and a time step of 1 fs. The equilibration stage implements a round-robin [16] approach of segmenting the domain in smaller portions, equilibrating each of them, and then rebuilding the complete domain and continuing the run for 10ns at a time. This acceleration strategy is very effective, allowing to reach the production stage faster. The production stage is composed of a collection of 20 trajectories of 5ns each, for a total run time of 100 ns.

The structural models built with the above mentioned approaches can be refined. Although several protein refinement methods exist in the literature [17], we use 3D^{refine} [3]. Another refinement approach relies on energy minimization using the AMBER ff12SB force field and implicit solvent as in the MD simulations. This refinement is conducted only on structures of domains with less than 500 amino acids due to the computational costs. The steepest decent method for minimization is used with a convergence criterion for gradients of 0.0001 kcal/mol/Å.

The approaches described in previous paragraphs, I-TASSER with automated template identification, I-TASSER with biologically-based and MD-based templates, yield several structural models for each domain. We narrow down to 2-3 structural models for each domain by ranking them with the Critical Assessment of Protein Structure Prediction (CASP) metrics [38]. Even constraining to this small number of ranked possible models for the domains, the total number of structural models for the full assembled L-protein is substantial. The full L-protein models can be assembled from the domains in several ways. We consider AIDA [34] and Chimera [24]. AIDA computes the best spatial arrangement of domains by an optimization process. Meanwhile, Chimera connects single-domain models via the *join* function that assigns fixed bond lengths and angles to the relevant bonds and angles in proteins.

The resulting full-length models can be ranked via various metrics to reveal a few top structural models of the tertiary structure of the protein under study. We evaluate end-termini domains based on the AMBER ff12SB minimized potential energy. Structural models of the entire multi-domain protein (obtained by assembling structural models of the constitutive domains) are evaluated via metrics utilized by the CASP community. Specifically, we consider the MolProbity scores Clash-score, Rot-out, Ram-out, Ram-fv, and MP-score [11]. Clash-score is the number of severe atomic clashes (with overlap < 0.4Å) per 1000 atoms. Rot-out is the rotamer outlier score that measures the percentage of side-chain configurations classified as rotamer outliers. Ram-fv is the Ramachandran favored score measuring the percentage of backbone Ramachandran configurations classified as outliers. The MP-score aggregates all these scores per the formula: $0.426 \cdot \ln(1 + \mathrm{Clash} - \mathrm{Score}) + 0.33 \cdot \ln(1 + \max(0, \mathrm{Rot} - \mathrm{out} - 1)) + 0.25 \cdot \ln(1 + \max(0, (100 - \mathrm{Ram} - \mathrm{fv})) - 2)) + 0.5$.

3 RESULTS

Our first model of the RVFV L-protein tertiary structure is obtained from RAPTORX. This server predicts a model with three domains, D1, D2, D3, that span $1 - 67$, $68 - 1622$, and $1623 - 2092$ amino acids, respectively. Domain D1 contaains the N-terminus and is PDB id 4q84A with sequence identity 9%, while D2 corresponds to PDB id 5amrA with sequence identity 12%, and D3 is based on PBD id 4qhpA and 2hpoA with sequence identity 6%. In addition, RAPTORX predicts 27 disordered positions at amino acids $1 - 3$ in the of D1, $1992 - 1996$ and $2074 - 2092$ in D3.

Several possible ways of fragmenting the L-protein into domains were identified computationally via NCBI DELTA-BLAST [5]. From those, we kept two alternatives with important biological information. In the first, the domain length and boundaries are both reconciled with functional information from the Arenavirus L-protein domains [8]. Figure 1 shows the phylogenetic tree of aligned core polymerase domains from members of the Arenaviridae, Orthomyxoviridae, and Bunyaviridae families [8]. Reconciling the computationally-predicted domains with those shown in Fig. 1 led us to adopting three domains for the RVFV L-protein: L1 spanning $1 - 200$, L2 spanning $201 - 1500$, and L3 spanning $1501 - 2092$ amino acids.

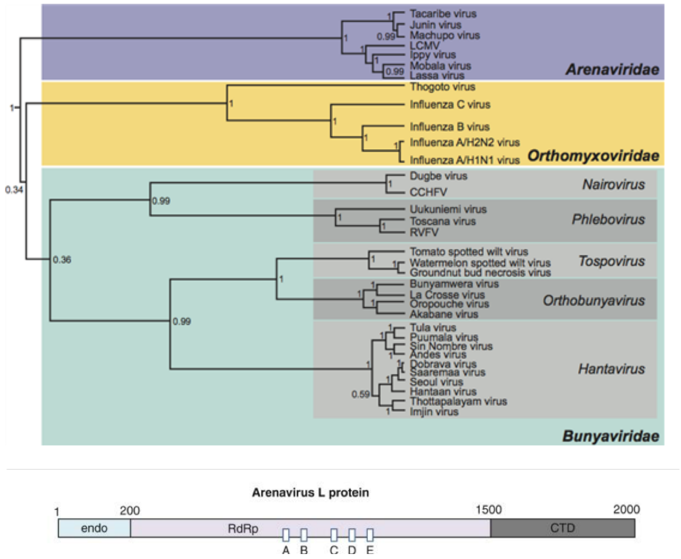


Figure 1: Top panel: aligned core polymerase domains from members of the Arenaviridae, Orthomyxoviridae, and Bunyaviridae families. Bottom panel: functional domains of the Arenaviridae L-proteins from [8].

Our second alternative for domain length and boundaries is again based on biological input and has three domains: $\tilde{L}1$ spanning $1 - 222$, $\tilde{L}2$ spanning $223 - 1500$, and L3 spanning $1501 - 2092$ amino acids. Both $\tilde{L}1$ and $\tilde{L}2$ have been shown to be stable in cell culture models of infection, with segment $1 - 222$ forming stable oligomers [37].

3.1 Structure of the RVFV L-protein Domains

Summarizing, we have adopted three domains for the RVFV L-protein, either L1-L2-L3 or $\tilde{L}1$-$\tilde{L}2$-L3. The next task is to establish the best structure for each domain and proceed to their assemblage. We employed I-TASSER [25] for the structure determination based on templates, and fed our pre-selected, biologically-inspired, templates to this server.

Table 1 shows the C-score of five models for L1 and seven models for L2 or $\tilde{L}2$ obtained with different inputs. For model L1-nt and L2-nt, I-TASSER picks the template. However, we feed our selected template structure with biological insight for each of the models L1-4miw, L1-5ize, L1-5hsb, and L1-5j1n listed in table 1. The L1 domain is an endonuclease domain. All endonucleases contain the motif P-D-x(13,45)-E-x(19,22)-K [19]. This motif is present in the RVFV L-protein, as well as in the L-proteins of the Hantan, Lassa, and Andes viruses. The endonucleases of the three latter viruses have structures in the PDB with ids 4miw and 5j1n (Lassa virus), 5ize (Hantan virus), 5hsb (Andes virus). Table 1 lists the I-TASSER solved structure with highest C-score per fed template. From these five models for the L1 domain, we select the two marked in bold for further evaluation.

Similarly, table 1 shows the C-score of six models of the L2 domain, where a template is fed to the server (PDB ids 5amq, 5amr, 1yuy, 4xhi, 4ucy, 5fje). Once again, these templates are based on

Table 1: Structure models of the L1 and L2 domains built with I-TASSER [25].

PDB id	Organism	Best C-score
L1-nt	-	-4.32
L1-4miw	Lassa virus	-1.55
L1-5ize	Hantaan virus	**-0.97**
L1-5hsb	Andes virus	**-0.74**
L1-5j1n	Lassa virus	-1.45
L2-nt	-	-0.09
L2-5amq	La Crosse	**0.07**
L2-5amr	La crosse	-0.09
L2-1yuy	Hepatitis C	-0.05
L2-4xhi	Thosea Asigna	**0.04**
L2-4ucy	Metapneuumovirus	**0.17**
L2-5fj6	Phi6	-0.04

the biological insight for the L2 domain, which is a polymerase domain. As illustrated in Fig. 1, all polymerases possess the palm $A \rightarrow E$, which contains the motifs A (D-x(3)-W-x(5)), B (QG-x(3)-Y-x-SSLL), C (SDD), D (Y-x(3)-K), E (x-F-x(2)-E) [8]. The C motif is crucial to the function of all polymerases. Indeed, the RVFV L-protein contains the C motif. Therefore, based on this insight, the following protocol is implemented: (i) the PDB is searched for all RdRPs of which the L2 domain of the RVFV L-protein polymerase is a member, (ii) only those RdRp polymerases containing the C (SDD) motif or its GDD and KDD variants are retained. This process resulted in PDB ids 5amq, 5amr, 1yuy, 4xhi, 4ucy, and 5fje. Each of these structures is fed to I-TASSER as a template for the L2 domain, and only the model with highest C-score produced by I-TASSER in each setting is retained and listed in Table 1. We select three viable models, their C-score are reported in bold in the table.

For the $\tilde{L}1$ and $\tilde{L}2$ domains, the same highest-score structural models of L1 and L2 (see table 1) are considered, plus a third model for $\tilde{L}1$ obtained from MD and minimized as described in section 2. We refer to this model as $\tilde{L}1$-MD and its evaluation is in table 2.

Table 2: Property and energetics evaluation of $\tilde{L}1$-MD and L3 domain: Potential energiy PE, number of atoms in structure N, radius of gyration R$_g$, end-to-end distance R$_{e-e}$, and maximum radius R$_{max}$.

Name	PE (kcal/mol)	PE/N (kcal/mol)	R$_{e-e}$ (nm)	R$_g$ (nm)	R$_{max}$ (nm)
$\tilde{L}1$-MD	-6992.8	-2.0	8.305	2.455	5.053
L3$_{1-360}$	-10603.	-1.81	7.231	2.454	5.466
L3$_{361-592}$	-7467.4	-2.08	3.890	2.972	5.109

The L3 domain contains the L-protein C-terminus and is 591 amino acids long. Table 3 shows the C-score of two models obtained with different settings. L3-nt is the best model obtained allowing I-TASSER to automatically identify a template. An additional setting is considered, where I-TASSER is used to improve the L3-nt model by feeding back this model as a template with its last 231 amino acids replaced by the structure of a well-equilibrated and minimized MD version listed in the third row of table 2. This model is referred as L3-MD in table 3 and has a lower C-score than the original L3-nt model. We produced two extra models for the L3 domain. In both of them we include two segments for L3: (i) the first 360 aminoacids of the L3-MD model are minimized within the ff12SB force field in implicit solvent environment, listed in the 2nd row of table 2, and (ii) the last segment of 231 amino acids already MD-optimized and listed in the 3rd row of table 2. These two segments are assembled using AIDA [34] and Chimera [24]. The resulting models are referred to as L3-AIDA and L3-Chimera.

Table 3: Structure models of the L3 domain built with I-TASSER [25].

Setting	Description	Best C-score
L3-nt	C-terminus L3 I-TASSER model	-1.68
L3-MD	C-terminus L3 I-TASSER model improved by MD 231 a.a. segment	-1.95

3.2 Assembled Full-length Structural Models

Table 4 summarizes all the structural models obtained for the full-length L-protein, including the RAPTORX model and the models computed via assembling different structural models for the L1, L2, L3 domains and the alternative $\tilde{L}1$), $\tilde{L}2$), L3 domains. The full-length models are evaluated through the Molprobity scores listed in Section 2. These 18 models can be refined, as we describe in the next paragraph. Therefore, table 4 reports the Molprobity scores before and after refinement. The two models with the best MP-score for the set of models of the assembled L1-L2-L3 domains and for the set of assembled $\tilde{L}1$-$\tilde{L}2$-L3 domains are highlighted in bold.

Furthermore, the 18 full-length models listed in Table 4 are refined, as described in Section 2, and re-evaluated after refinement. The new scores are reported in the table under the column labeled "after." The four refined models with highest MP-score (bold font) are subjected to an intensive new refinement protocol, as follows. The three domain structures in the full-length model are refined with 3D^{refine}. The resulting single-domain structures are re-assembled with AIDA yielding an optimal multi-domain arrangement. Finally, the AIDA-computed full-length models are refined again with 3D^{refine}. Evaluation of the three best and final models is provided in Table 5 and depicted in Fig. 2.

The first structure in table 5 has the best scores. This model structure is shown at the top in Figure 2. For comparison, this structure is used as reference and the other two in table 5 are aligned to it with FATCAT [36]. This algorithm optimizes the alignment and minimizes the number of rigid-body movements (twists) around pivot points (hinges) introduced in the reference structure. According to FATCAT, the L1-5hsb+L2-4xhi+L3-nt and L1-5ize+L2-4xhi+L3-MD structures are significantly similar, with a p-value of 5.66×10^{-15} (raw score of 3645.61). The structure alignment identifies 1824/2092 equivalent positions with an RMSD of 6.86 Å, with 4 twists. The L1-5hsb+L2-4xhi+L3-nt and L1-MD+L2-4xhi+L3-MD structures are also significantly similar, with a p-value of 1.60×10^{-124} (raw score of 3336.60). The structure alignment identifies 1717/2092 equivalent positions with an RMSD of 7.66 Å, with 4 twists as well. Finally, the L1-5ize+L2-4xhi+L3-MD and L1-MD+L2-4xhi+L3-MD structures

Table 4: Molprobity evaluation before and after refinement of the full-length L-protein structures.

Model	MP-score		Clash-score		Rot-out%		Ram-out%		Ram-fv%	
	before	after	before	after	before	after	before	after	before	after
RAPTORX	3.74	3.75	167.85	81.92	4.09	10.97	3.55	3.83	88.30	88.71
L1-5ize+L2-5amq+L3-nt	3.75	3.67	127.81	66.39	3.07	7.69	11.67	8.52	72.87	82.44
L1-5ize+L2-4xhi+L3-nt	**3.54**	**3.46**	124.15	61.79	2.41	5.54	7.03	4.88	83.0	86.27
L1-5ize+L2-4ucy+L3-nt	3.64	3.50	124.12	63.27	3.12	6.02	6.70	5.55	82.15	86.17
L1-5hsb+L2-5amq+L3-nt	3.75	3.68	127.77	69.56	3.07	7.75	11.53	7.94	72.68	82.92
L1-5hsb+L2-4xhi+L3-nt	**3.58**	**3.45**	126.58	61.75	2.64	5.38	6.65	5.17	82.73	86.60
L1-5hsb+L2-4ucy+L3-nt	3.62	3.49	122.47	62.41	3.01	6.02	6.41	5.31	82.06	86.65
$\tilde{L}$1-5ize+$\tilde{L}$2-4xhi+$L$3-nt	3.95	3.75	147.78	74.1	4.30	7.91	11.64	8.42	70.83	80.53
$\tilde{L}$1-5ize+$\tilde{L}$2-4xhi+$L$3-MD	**3.83**	**3.70**	151.30	73.34	3.34	7.93	10.49	7.85	75.23	82.54
$\tilde{L}$1-5ize+$\tilde{L}$2-4xhi+$L$3-AIDA	3.92	3.79	143.40	73.36	4.30	8.82	11.86	9.76	71.95	80.05
$\tilde{L}$1-5ize+$\tilde{L}$2-4xhi+$L$3-Chimera	3.99	3.91	171.69	88.81	4.25	9.84	11.49	10.07	72.43	79.58
$\tilde{L}$1-5hsb+$\tilde{L}$2-4xhi+$L$3-nt	3.92	3.76	147.84	77.25	4.03	8.02	11.18	8.18	71.59	81.15
$\tilde{L}$1-5hsb+$\tilde{L}$2-4xhi+$L$3-MD	**3.83**	3.73	152.62	74.62	3.39	8.23	10.21	7.70	75.66	83.16
$\tilde{L}$1-5hsb+$\tilde{L}$2-4xhi+$L$3-AIDA	3.93	3.78	148.82	77.23	4.25	8.23	11.76	9.33	72.42	80.19
$\tilde{L}$1-5hsb+$\tilde{L}$2-4xhi+$L$3-Chimera	4.01	3.90	175.88	89.62	4.46	9.36	11.17	9.73	73.07	79.29
$\tilde{L}$1-MD+$\tilde{L}$2-4xhi+$L$3-nt	3.91	3.73	144.21	72.28	4.09	8.02	11.07	7.94	71.48	81.53
$\tilde{L}$1-MD+$\tilde{L}$2-4xhi+$L$3-MD	3.84	**3.70**	152.15	76.02	3.50	7.42	9.87	7.42	75.95	83.43
$\tilde{L}$1-MD+$\tilde{L}$2-4xhi+$L$3-AIDA	3.90	3.77	144.23	72.81	4.09	8.66	11.21	9.04	72.66	80.57
$\tilde{L}$1-MD+$\tilde{L}$2-4xhi+$L$3-Chimera	3.97	3.90	172.83	90.22	4.03	9.63	10.97	10.16	72.87	80.30

Table 5: Molprobity evaluation of the best full-length structural models of the RVFV L-protein.

Model	MP-score	Clash-score	Rot-out%	Ram-out%	Ram-fv%
L1-5hsb+L2-4xhi+L3-nt	**3.40**	61.23	4.95	5.12	87.42
$\tilde{L}$1-5ize+$\tilde{L}$2-4xhi+L3-MD	3.51	65.95	5.16	6.79	84.02
$\tilde{L}$1-MD+$\tilde{L}$2-4xhi+L3-MD	3.49	64.36	5.38	6.36	85.31

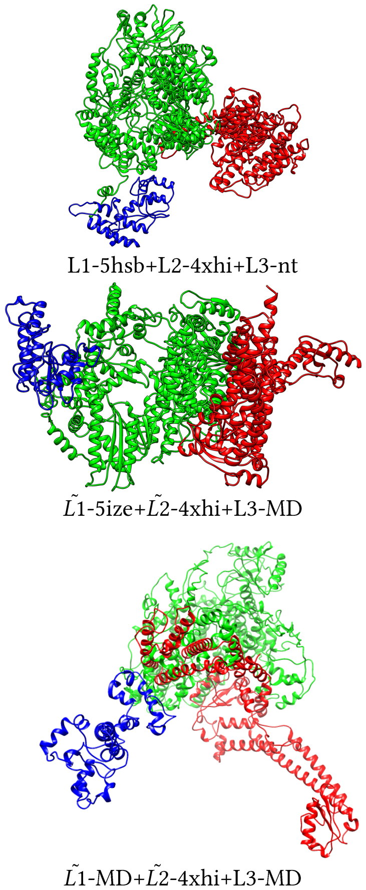
are also significantly similar, with a p-value of 0.00 (raw score of 5619.35). The structure alignment identifies 2014/2092 equivalent positions with an RMSD of 1.34 Å, with 5 twists.

4 CONCLUSION

This paper shows the complexity entailed in obtaining physically realistic structural models of multi-domain proteins containing thousands of amino acids. As the results demonstrate, the inclusion of strong biological insight guides very efficiently the existing computational approaches for the identification of domains, the building of single-domain structural models, the assembly of these models into structural models of the full protein, and the refinement of such structures. As we demonstrate here for the RVFV L-protein, the biological insight leveraged the otherwise badly-defined options for defining domain boundaries and templates used in the process of building structural models of single domains.

An important consideration is the extent to which the built structural models represent the biologically-active state of the protein under study. In the absence of feedback from wet-laboratory data, this evaluation is nontrivial. Indeed, the problem of decoy selection in computational structural biology is known. While research is active in this area, currently, it is limited to decoys obtained for short, single-domain proteins not exceeding 300 amino acids. In ongoing research on the RVFV L-protein, we are exploring several avenues for investigating the credibility of the computed models presented in this paper. For example, expression constructs exist already for the 1-222 amino acids region (the $\tilde{L}$1 domain). Further expression constructs will allow us to relate modeling data with biological function directly investigated in our wet-lab. It is worth noting, that we have strong evidence from the ongoing MD simulations that the structural models presented here only capture restricted views of the RVFV L-protein biologically-active state. In fact, the structural models may only provide snapshots of the possible flexibility of the L- protein, particularly concerning its termini. The results obtained via RAPTORX point to disorder predicted for amino acids in the N-terminus (L1 domain) and C-terminus (L3 domain). These findings are corroborated by our MD simulations that predict that the two termini of the RVFV L- protein, at the thermodynamic conditions studied, are quite flexible and may adapt their structure upon the environment in which they are immersed under physiological conditions.

Given the sparsity of information on the RVFV L-protein and lack of crystal structure, the findings presented here on structural models of the RVFV L- protein will facilitate protein-protein interaction studies and drug discovery efforts. In addition, given the high sequence identity of the RVFV L-protein and other RdRps, L-protein structural models may serve as valuable templates to further expanding the structural characterization of other polymerases.



L1-5hsb+L2-4xhi+L3-nt

$\tilde{L}$1-5ize+$\tilde{L}$2-4xhi+L3-MD

$\tilde{L}$1-MD+$\tilde{L}$2-4xhi+L3-MD

Figure 2: The three final structures for the RVFV L-protein. The L1, L2, L3 domains are in blue, green, red, respectively.

ACKNOWLEDGMENTS

This research was partially supported by a 2017-Multidisciplinary Research Award from the Provost Office. All MD simulations were done in the ARGO cluster of the Office of Research Computing at George Mason University.

REFERENCES

[1] S. Altschul, F. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* 215 (1990), 403–410.

[2] H. M. Berman, K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. *Nat. Struct. Bio.* 10 (2003), 980–980.

[3] D. Bhattacharya and J. Cheng. 2013. 3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins: Struct, Funct, and Bioinf* 81 (2013), 119–131.

[4] D. D. Bohr and P. E. Wright. 2008. How do proteins interact? *Science* 320, 5882 (2008), 1429–1430.

[5] G.M. Boratyn, A. A. Schäffer, R. Agarwala, S. F. Altschul, and et al. 2012. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* 7:12 (2012), 1–14.

[6] M. Bouloy and F. Weber. 2010. Molecular biology of rift valley Fever virus. *Open Virol J* 4 (2010), 8–14.

[7] D. A. Case, R. M. Betz, D. S. Cerutti, T. E. III Cheatham, and et al. 2016. AMBER 2016. (2016). University of California, San Francisco.

[8] K. H. Choi. 2012. Viral Polymerases. *Adv. Exp. Med. Biol.* 726 (2012), 267–304.

[9] C. Chothia and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* 5, 4 (1986), 823–826.

[10] P. R. Daga, Patel. R. Y., and R. J. Doerksen. 2018. Template-based Protein Modeling: Recent Methodological Advances. *Curr. Top. Med. Chem.* 10 (2018), 84–94.

[11] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, and et al. 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* 35(Web Server issue) (2007), W375–83.

[12] J. Eickholt, X. Deng, and J. Cheng. 2011. DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinf.* 12 (2011), 43.

[13] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, and et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucl. Acids Res.* 44 (2016), D279–D285.

[14] M. Källberg, H. Wang, S. Wang, J. Peng, and et al. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7 (2012), 1511–22.

[15] R. L. Marsden, L. J. McGuffin, and D. T. Jones. 2009. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.* 11 (2009), 2814–2824.

[16] G. Miao, J. Zander, K. W. Sung, and B. Slimane. 2016. *Fundamentals of Mobile Data Networks.* Cambridge University Press.

[17] K. M.S. Misura and D. Baker. 2005. Progress and challenges in high-resolution refinement of protein structure models. *Proteins: Struct., Func. and Bio.* 59, 1 (2005), 15–29. DOI:https://doi.org/10.1002/prot.20376

[18] J. Mongan, D. A. Case, and J. A. McCammon. 2004. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* 25 (2004), 2038–2048.

[19] B. Morin, B. Coutard, M. Lelke, and et al. 2010. The N-terminal domain of the arenavirus L protein is an RNA endonuclease essential in mRNA transcription. *PLoS Pathog* 6 (2010), e1001038.

[20] R. Müller, O. Poch, M. Delarue, D. H. Bishop, and M. Bouloy. 1994. Rift-Valley Fever Virus L-Segment - Correction of the Sequence and Possible Functional-Role of Newly Identified Regions Conserved in RNA-Dependent Polymerases. *J. Gen. Virol., Pt 6* 75 (1994), 1345–1352.

[21] B. Olson, K. A. De Jong, and A. Shehu. 2013. *Off-Lattice Protein Structure Prediction with Homologous Crossover.* ACM, New York, NY. 287–294 pages.

[22] B. Olson and A. Shehu. 2013. *Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface.* ACM ,Washington, D.C. 430–439 pages.

[23] J. Peng and J. Xu. 2011. Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins* 79 (2011), 161–171.

[24] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. UCSF Chimera: A visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 13 (2004), 1605–1612.

[25] A. Roy, A. Kucukural, and Y. Zhang. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5 (2010), 725–738.

[26] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *PNAS* 95 (1998), 5857–5864.

[27] A. Shehu. 2010. *Protein Structure Prediction: Method and Algorithms, Eds. Rangwala, H. and Karypis, G.* Wiley Book Series on Bioinformatics, Hoboken, NJ.

[28] D. Shortle, K. T. Simons, and D. Baker. 1998. Clustering of low-energy conformations near the native structures of small proteins. *PNAS* 95 (1998), 11158–11162.

[29] M. Suyama and O. Ohara. 2003. DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 19 (2003), 673–4.

[30] V. Tsui and D. A. Case. 2001. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* 56 (2001), 275–291.

[31] S. Venkataraman, B. V. L. S. Prasad, and R. Selvarajan. 2018. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses* 10, 2 (2018), E76.

[32] Y. Wang, J. Wang, R. Li, Q. Shi, Z. Xue, and Y. Zhang. 2017. ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. *Nucl. Acids Res.* 45 (2017), W400–W407.

[33] A. M. Wollacott, A. Zanghellini, P. Murphy, and D. Baker. 2007. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci* 16 (2007), 165–175.

[34] D. Xu, L. Jaroszewski, Z. Li, and A. Godzik. 2015. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. *Bioinformatics* 31 (2015), 2098–2105.

[35] D. Xu and Y. Zhang. 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Struct., Func. and Bio.* 80, 7 (2012), 1715–1735.

[36] Y. Ye and A. Godzik. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19, Suppl 2 (2003), ii246–ii255.

[37] A. Zamoto-Niikura, K. Terasaki, T. Ikegami, C. J. Peters, and S. Makino. 2009. Rift Valley Fever Virus L Protein Forms a Biologically Active Oligomer. *J Virol* 83, 24 (2009), 12779–12789.

[38] Y. Zhang. 2014. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins: Suppl. 2* 82 (2014), 175–187.