The Structure of Mouse Cytomegalovirus m04 Protein Obtained from Sparse NMR Data Reveals a Conserved Fold of the m02-m06 Viral Immune Modulator Family

Nikolaos G. Sgourakis,1 Kannan Natarajan,2 Jinfia Ying,1 Beat Vogeli,1,3 Lisa F. Boyd,2 David H. Margulies,2,* and Ad Bax1,*
1Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
2Molecular Biology Section, Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA
3Present address: Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland
*Correspondence: dhm@nih.gov (D.H.M.), bax@nih.gov (A.B.)
http://dx.doi.org/10.1016/j.str.2014.05.018

SUMMARY

Immunoevasins are key proteins used by viruses to subvert host immune responses. Determining their high-resolution structures is key to understanding virus-host interactions toward the design of vaccines and other antiviral therapies. Mouse cytomegalovirus encodes a unique set of immunoevasins, the m02-m06 family, that modulates major histocompatibility complex class I (MHC-I) antigen presentation to CD8+ T cells and natural killer cells. Notwithstanding the large number of genetic and functional studies, the structural biology of immunoevasins remains incompletely understood, largely because of crystallization bottlenecks. Here we implement a technology using sparse nuclear magnetic resonance data and integrative Rosetta modeling to determine the structure of the m04/gp34 immunoevasin extracellular domain. The structure reveals a β fold that is representative of the m02-m06 family of viral proteins, several of which are known to bind MHC-I molecules and interfere with antigen presentation, suggesting its role as a diversified immune regulation module.

INTRODUCTION

Cytomegaloviruses (CMVs) are important models of pathogen-host interactions, widely recognized for their ability to interfere with host immune responses to accomplish the multifaceted task of inhibiting the recognition of infected cells by CD8+ T cells while avoiding destruction by natural killer (NK) cells according to the "missing-self" hypothesis (i.e., the lack of antigen-presenting major histocompatibility complex class I [MHC-I] molecules on the cell surface) (Lemmermann et al., 2012). To do this, the virus maintains a series of genes encoding immune evasion and regulatory proteins (Lilley and Ploegh, 2005). In particular, the m02-m16 family is a class of early-expressed genes of mouse CMV (MCMV) crucial for viral survival and infectivity. However, there is no identifiable homology to any determined structure in the Protein Data Bank (PDB), nor is there amino acid sequence similarity to any other protein family. A member of the m02-m06 class, the m04/gp34 protein, is unique in its ability to bind MHC-I molecules in the endoplasmic reticulum (ER) and accompany them to the cell surface (Kleijnen et al., 1997), while the closely related m06 protein binds MHC-I and directs it to the endosome (Tomas et al., 2010). Therefore, it has been proposed that by countering the MHC-retaining functions of other viral proteins (m06 and the MHC-I-like m152), m04 helps MCMV evade the NK cell response (Babić et al., 2010, 2011; Holtappels et al., 2006). The combined effects of m04, m06, and m152 on CD8+ T cells and NK responses reveal a complex and still poorly understood aspect of MCMV immune evasion (Pinto et al., 2006). Although the structural basis for the function of m152/gp40 immunoevasin has been previously revealed by X-ray crystallography (Wang et al., 2012), and structures of other MCMV MHC-I-like proteins have been similarly characterized (Adams et al., 2007; Berry et al., 2013; Mans et al., 2007; Natarajan et al., 2006), for other molecules, a detailed picture is still lacking because of difficulties in crystallization and the large sizes of their in vivo functional complexes, making conventional nuclear magnetic resonance (NMR) studies challenging. In particular, extensive crystallization trials in our laboratory using different constructs of m04 have repeatedly failed to yield well-differacting protein crystals.

Recent advances in computational methods using the program Rosetta (Leaver-Fay et al., 2011) allow accurate modeling of protein structures from sparse NMR data sets containing chemical shifts (CSs), residual dipolar couplings (RDCs), and a minimal subset of the proton-proton distances used by conventional methods (Shen et al., 2008). By combining the use of CS-Rosetta with extensive deuteration of side-chain protons to improve 13C relaxation and the use of transverse relaxation optimized spectroscopy (TROSY) at high magnetic fields to improve 15N and amide 1H relaxation (Pervushin et al., 1997), the structure of larger proteins and protein complexes can now be determined by solution NMR (Raman et al., 2010; Sgourakis et al., 2011). Methyl protons are reintroduced using site-specific labeling of Ilc, Leu, and Val (ILV) residues (Tugarinov et al., 2006), a scheme that greatly simplifies the NMR spectra while still providing a sparse set of long-range methyl-methyl nuclear...
data was found to outperform conventional protocols in terms of both precision (convergence) and accuracy relative to the target structure (Lange et al., 2012). This is due in part to the use of an empirically optimized all-atom energy function that defines the local hydrogen-bonding and side-chain-packing features once a sufficiently converged low-resolution model of the backbone (within 2–5 Å accuracy relative to the native structure) can be obtained using a sparse network of long-range experimental restraints. Thus, although experimental data are still required to define an overall backbone fold, Rosetta alleviates the need for extensive side-chain assignments and a high density of NOE restraints (in excess of ten restraints per residue) that is typically required by conventional protocols. For larger, more challenging systems in which the sparse ILV and amide-amide NOE data are insufficient to determine conclusively the target backbone structure, obtaining RDCs from spectra gathered in multiple alignment media is a powerful way to improve further structural convergence and validate the final Rosetta models, as shown in recent RASREC applications (Rao et al., 2014; Warner et al., 2011). Recent TROSY-based methods allow quantitative RDC measurements for larger proteins at high accuracy (Fitzkee and Bax, 2010).

Here we combine these advanced NMR technologies to determine the structure of the m04 extracellular domain (m04ED). The calculations converge to a well-defined structure showing an elaborate β sheet topology that is reminiscent of, but highly divergent from, the canonical immunoglobulin (Ig) fold (Williams and Barclay, 1988). Sequence alignments with other members of the viral m02-m06 family show that the structural features of the m04 core are broadly conserved among members of this family, suggesting that the interactions involved in MHC-I binding are likely to be similar as well.

RESULTS

The m04ED Binds MHC-I Molecules
m04 was originally identified as a 34 kDa glycoprotein that coimmunoprecipitates with MHC-I molecules upon MCMV infection (Kleijnen et al., 1997). Toward determining the molecular requirements of m04 recognition by MHC, previous studies have highlighted the requirement for proper MHC folding and association with its light-chain β2-microglobulin to accommodate m04 binding, suggesting that m04 interacts with properly conformed, peptide-loaded MHC-I molecules (Lu et al., 2006). This work further suggested that the transmembrane (TM) region of m04 is critical for its interaction with the MHC-I molecule H2-Kb. To test whether this is a general feature of m04–MHC interactions and to identify a minimal, functional m04 construct suitable for structural studies by NMR, we expressed the ED of the m04 protein (m04ED), excluding the signal peptide (SP), the TM region, and the intracellular C-terminal tail (Figure 1A). The protein was expressed in E. coli and therefore also lacked the posttranslational glycosylations present in the wild-type (WT) protein (WT m04 has five consensus glycosylation sites, of which three have been confirmed to be used in vivo [Kleijnen et al., 1997]). To evaluate the biological activity of recombinant m04 protein, we examined its interaction with the MHC-I molecule H2-Kb by surface plasmon resonance (SPR). Preliminary SPR experiments (Figure 1B) demonstrate a direct but weak ($K_D \geq 100 \mu M$) interactions...
Structure

m04 NMR Structure by Integrative Rosetta Modeling

interaction between m04ED and the MHC-I molecule H2-D\(^\beta\), even in the absence of the TM region and posttranslational modifications. By contrast, the MHC-I-like molecule, MULT-1, used here as a negative control, did not bind (Figure S1 available online).

Our results reveal a measurable interaction between MHC and the m04 ectodomain using a direct biophysical method. Previous studies demonstrated the requirement of the m04 TM region for efficient m04–MHC interaction in detergent cell lysates (Lu et al., 2006). Taken together, these results support a model of weak interaction between the luminal domains when strongly coupled via their TM domains.

NMR Backbone Relaxation Rates Identify a Well-Ordered Structural Core

The optimized m04ED construct shows well-dispersed 2D \(^1\text{H}-\text{\textsuperscript{15}N}\) heteronuclear single-quantum coherence (HSQC) spectra indicative of a folded, stable protein core (Figure 1C). We used backbone relaxation experiments to probe the extent of backbone conformational dynamics on the fast (picoseconds to nanoseconds) timescale (Kay et al., 1989). The resulting \(^{15}\text{N}\) \(R_1\) and \(R_2\) (Figures 2B and 2C) relaxation rates combined with the heteronuclear \(^{15}\text{N}-\text{\textsuperscript{1}H}\) NOE (Figure 2A) indicate the presence of a well-ordered structural core spanning residues 21 to 176, which includes a flexible loop at residues 95 to 102. The N-terminal region (residues 1–20) is highly mobile on the nanosecond timescale, as evidenced by decreased \(^{15}\text{N}-\text{\textsuperscript{1}H}\) NOE and increased \(R_1\) and decreased \(R_2\) relaxation rates. This highly charged segment (11 of 20 charged residues) includes 3.5 turns of a regular \(\alpha\) helix for residues 6 to 19, as indicated by the analysis of backbone CSs using the program TALOS-N (Shen and the N terminus, does not participate in strong packing interactions with the core structure. Taken together, these results reveal the presence of a well-ordered protein fold for residues 21 to 176, flanked by two terminal-capping sequences. We therefore focus on determining the structure of this m04ED core region to elucidate its fold and identify structural features that could serve as possible binding sites for MHC-I molecules.

**Figure 2.** \(^{15}\text{N}\) Amide Relaxation Rates and \(^{15}\text{N}-\text{\textsuperscript{1}H}\) NOE Ratios for the Full-Length m04ED Sequence, Recorded at 600 MHz \(\text{\textsuperscript{1}H}\) Frequency (A–C) \(R_2\) values were obtained from measured \(R_1\) rates after correction for off-resonance effects (Massi et al., 2004). Key residues that demarcate the structural core (21–176) and flexible loop region (95–102) are highlighted on the plots. The dotted line in (A) indicates the 0.5 \(^{15}\text{N}-\text{\textsuperscript{1}H}\) NOE cutoff value used to define rigid and dynamic structural elements within the m04ED molecule.

Bax, 2013), and the presence of short-range \(d_{nN(i,i+3)}\) and \(d_{nN(i,i+1)}\) NOE patterns (Wuthrich, 1986), confirming the \(\alpha\)-helical structure. The elevated dynamic characteristics and the absence of long-range \(\text{H}^\text{N}-\text{H}^\text{N}\) or \(\text{H}^\text{N}-\text{CH}_3\) NOEs to core residues suggest that the N-terminal helix interacts only transiently with the core. Similarly, the C-terminal residues 177 to 197 are highly mobile on the picosecond-to-nanosecond timescale, as indicated by the \(^{15}\text{N}\) relaxation data. This region is found primarily in a disordered loop/coil conformation that, like...
and CS data and confirm the power of the Rosetta structure determination protocol for this highly sparse NOE data set (1.5 long-range restraints per 10 residues). The remaining 26 C-terminal residues of the core sequence (151–176) were poorly converged in the first round of models, as no long-range H-H NOE NOEs could be identified for this part of the sequence. Analysis of the backbone CSs using TALOS-N (Shen and Bax, 2013) indicates that this segment contains three turns of an α-helix (residues 160–171).

To better define the conformation of the C-terminal helix, we measured RDCs in two alignment media (P1 phage and positively charged gel) that report on the orientation of N-H vectors with respect to an overall alignment frame (Bax and Grishaev, 2005). The two data sets are complementary in the sense that the two alignment tensors show a normalized scalar product (Sass et al., 2000) of only 0.51 (corresponding to a 60° angle in five-dimensional alignment tensor space), meaning that the two measurements are quite independent of each other. Additionally, the linear correlation between the raw RDC data sets is 0.36, further suggesting that the two data sets are for all practical purposes sufficiently independent (Figure S2C) (Tolman and Ruan, 2006). Inspection of the RDCs for core residues 160 to 171 (Figure S2A and S2B) reveals a kink in the C-terminal helix, centered at Ser167, in which the helical axis changes orientation (Figure S2 A and S2B). We find that the use of RDC restraints strongly improved the convergence of the C-terminal helix (Figures 4A and 4C), validating the previous structure. Careful analysis of the spectra revealed a total of 19 long-range H-H NOE NOEs that were included in the subsequent round of calculations (Table 1, No. 4). This set includes five restraints from Asp159, Val168, Leu169, and Leu172 that link the C-terminal helix to the main core of the structure (Figures 4B and 4D), while the remaining 14 restraints validate the β-sheet topology obtained in the previous models. The resulting models show improved convergence of the C terminus (backbone rmsd within 1.5 Å), and RDC cross-validation statistics (Q_{ref} = 0.58, R_p = 0.82). Finally, we identified 25 long-range CH₃-CH₃ NOEs in the corresponding methyl NOE spectra, 6 of which map to the C-terminal α-helix (Figures 4A and 4C), validating the previous round of models (Table 1, Nos. 5–7). All NOE cross-peaks with peak signal-to-noise ratios > 10, which also showed mirror peaks in the case of HNNH, H(NH), H(C)(CH), and H(C)CH 3D NOE spectra, were manually assigned and classified as intrasegmental, sequential or medium range (which further validated the backbone and methyl resonance assignments), or long range, corresponding to a sequence separation greater than five residues. Although the short- to medium-range NOEs were consistent with the first-round Rosetta models, the inclusion of such restraints in the calculations does not lead to any gains in convergence of the local backbone structure, which is already heavily constrained by the CS data. Including all the available long-range NOEs, backbone CSs and RDC data (Table 1, No. 7) led to a structural ensemble (Figure 3C) that is converged to within a backbone rmsd of 0.85 Å, excluding the loop region spanning residues 95 to 102 (Figure S3B). This loop is highly mobile on the picosecond-to-nanosecond timescale, as evidenced by the reduced 15N-{1H} NOE values (Figure S3A) and further supported by the near-zero RDC values (Figures S2A and S2B). We find that the use of RDC restraints strongly improves convergence for the backbone core spanning residues 21 to 176 (95%, with the remaining 5% of the sequence...
Structure

m04 NMR Structure by Integrative Rosetta Modeling

corresponding to the flexible loop at residues 95–102), including a short, ordered loop at residues 118 to 122 (Table 1, compare No. 5 with Nos. 6 and 7). The open, type-I α helix conformation of the loop (Figure 5B) is supported by analysis of the backbone CSs (Shen and Bax, 2012) and further validated by the absence of a long range R^H->H^N NOE between Leu49 and Asn122 that would be expected for an alternative closed conformation sampled in the RDC-free calculations (Figure 5A). The final full-atom models (Tables 1, No. 7, and 2; Figure 3C) have good packing and structural statistics (assessed separately using MOLPROBITY [Chen et al., 2010]) while satisfying all experimental NOE restraints and showing good fits to both RDC data sets (Figures S4A and S4B).

The m04 Core Structure Exemplifies a Highly Diverse Ig-Based Immuneevasin Fold

m04 does not show any homology to nonviral proteins by amino acid sequence comparisons or to any sequence of known structure in the PDB. At first sight, the m04 core resembles the fold of the Ig superfam, which can be further divided into four to nine main structural classes (Bateman et al., 1996; Halaby et al., 1999) depending on overall size, strand connectivity, and loop size. A Dali (Holm and Rosenström, 2010) search using the NMR structure as a target suggests a statistically significant structural similarity (Z score ≤ 5.0) to a variety of Ig-like molecules (Igs) such as various antibody chains (e.g., PDB accession number 1AXS-chain B) and human leukocyte antigen class-II Ig-like domains (e.g., PDB accession number 4HL1-chain H). Moreover, the disulfide bond at C142–C47 with Trp62 packing against it and the β strand arrangement are clearly indicative of an Ig fold. However, a closer look at the structural alignments reveals major differences from the canonical Ig superfam.

Although the Ig fold is typically characterized by seven strands on two sheets (Structural Classification of Proteins [Andreeva et al., 2008]), the m04 core structure has a total of nine strands, four of which (β2–β3–β6–β7) are short (4 or 5 residues each) and form one layer of the β sandwich, while five strands of 6 to 12 residues in length (β1–β5–β8–β9–β10) form an extended β sheet platform on the other side of the sandwich (Figure 6D). The five-stranded platform provides a nucleus beneath which the smaller four-stranded sheet packs at its upper end. The kinked C-terminal helix supports the lower end of the larger five-stranded platform. In addition, the strand connectivity is clearly different from the closest Ig fold, found in the antibody variable (V) domains (typically an A[β1], B[β2], E[β3], D[β4] sheet packed against C[β5], F[β6], G[β7]). Furthermore the spacing between the Cys residues forming the disulfide bond that links the two β layers (typically between the B and F strands) (Cys47, Cys142) is significantly larger than found in Igs (95 versus 55–75 residues) and connects β3 to β6. Finally, m04 has one additional disulfide bond connecting Cys26 (β1) and Cys147 (β4). These features, specifically the secondary structural elements forming the hydrophobic core of the fold and the second disulfide, are highly conserved in the m02-m06 family and among different MCMV strains (Corbett et al., 2007) (Figure 7). In particular, this family preserves the NAXWXEE/HW motif (in strand β4) throughout a large number of laboratory and wild-derived isolates of MCMV. This segment is a central element of the m04 fold, likely to play a key role in stabilizing the core structure, which provides a scaffold for the grafting of various loop sequences among different family members. Taken together, these results suggest that the m04 structure is likely a product of convergent evolution toward an Ig-like structure rather than a true structural homolog of Igs.

DISCUSSION

This description of the solution structure of the m04/gp34 glycoprotein ED reveals the power of a hybrid methodology that combines multidimensional NMR with sparse labeling and data collection techniques and integrative computational modeling (Ward et al., 2013). The value of progressive addition of experimental data sets in achieving a converged set of minimum energy structures is documented here and highlights the importance of including ILV NOEs and RDCs in the proper placement of structural elements. In particular, these distance and orientational restraints permitted the placement of the C-terminal α-helical segment to a defined position underlying a β sheet platform and precise modeling of a second loop conformation. The dynamic aspects of the m04ED structure revealed by NMR relaxation experiments, notably the highly mobile amino and carboxy-terminal segments as well as the central loop region, may relate to the function of the molecule in its ability to interact with a number of diverse MHC-I molecules as well as the difficulty in identifying suitable crystallization conditions.

The overall structure of m04ED reveals a complex β topology, punctuated with connecting regions containing both coil and helix, distantly related to the Ig fold. This core β structure permits variation in the connecting loops, as evidenced by the amino acid sequence diversity observed in other members of the
m02-m06 viral protein family, and should prove useful in designing crystallizable variants toward determining the structures of other family members. The closely related m06 also carries out an immunomodulatory function and binds MHC-I molecules (Reusch et al., 1999). Amino acid sequence analysis of the extended family that includes m02-m16 has suggested that all of these genes encode structurally related proteins, and the structure of m04ED provides a toehold in deriving their structures. Whether the structural similarities carry over to related functions or the m04 structure serves as a robust scaffold for loop variability remaining to be determined.

m04/gp34 is posttranslationally processed in the ER and contains at least three N-asparaginyl-linked carbohydrate moieties (Kleijnen et al., 1997). We have determined the NMR structure of m04 using a bacterially expressed construct, so it is important to consider the predicted location of the N-asparaginyl sites in the context of this structure. An analysis of the amino acid sequence of m04 using the NetNGlyc 1.0 Server (Blom et al., 2004) identifies seven potential N-X-S/T sites in m04ED, of which five, at positions 5, 32, 55, 116, and 131, are likely to be used. The core of m04ED contains four of these sites, and the locations of the Asn residues on the amino acid sequence alignment and on the ribbon diagram of the structure are shown in Figures 6 and 7. All of the glycosylation sites are located in accessible loop regions and are not expected to interfere with the proper folding of the core structure.

Immunoprecipitation experiments of MCMV-infected cells demonstrate that m04 is the major component associated with MHC-I molecules (Kleijnen et al., 1997). Biochemical experiments indicate the importance of a functional peptide-loading apparatus (including transporter associated with antigen processing and tapasin) as well as the TM region of m04 for the association (Lu et al., 2006). Our experiments, using a truncated m04 construct lacking the TM and cytoplasmic regions of the protein, indicate that the m04ED is sufficient for a transient interaction with MHC-I and that the m04ED may contribute to interaction with the lumenal domain of assembled MHC-I heterotrimers (MHC-I heavy chain, β2 m, and peptide). The apparent low affinity of the m04ED interaction with soluble MHC-I contrasts with the results of immunoprecipitation experiments of virus-infected or transfectant cell lysates, in which m04 is a predominant MHC-I-associated molecule (Kleijnen et al., 1997; Lu et al., 2006). These differences may also result from the clear differences in measurement of association of molecules constrained by 2D membrane surfaces as compared with 3D solution measurements. Alternatively, in addition to the demonstrated role of the TM regions, we cannot rule out the participation of additional cellular or viral components in this interaction. Inspection of the structure and consideration of other molecules that interact with MHC-I and MHC-I-like molecules (T cell receptors, CD8 coreceptor, NK receptors, and other viral immunoevasins including m152, US2, CV203, and E3-19K glycoprotein) prompts us to speculate on possible m04ED sites that might participate in MHC-I interactions. Examination of the surface electrostatic potential of m04ED (Figure 6B) suggests that the large basic groove along the
Structure
m04 NMR Structure by Integrative Rosetta Modeling

Figure 5. Possible Loop Conformations for Residues D115-1N125
(A and B) Closed (A) and open (B) loop conformations sampled in Rosetta structure calculations using only the sparse NOE distance restraints (A) or NOEs supplemented by amide RDCs from two alignment media (B). The rms of the residual $\chi^2$ in the RDC structure fits (PH/gel) is indicated for each conformation. A strong HN-HN NOE cross-peak between the well-resolved resonances of Leu49 and Asn122 (indicated with a yellow line in [A]) expected in the "closed" conformation was not observed in the HNHN and HNHH NOESY spectra, further supporting the "open" loop conformation. Although the NOE data alone are insufficient to define the loop, convergence to the "open" conformation is obtained when RDC restraints are included in addition to NOEs.

Table 2. NMR Restraints and Structural Statistics

<table>
<thead>
<tr>
<th>Distance restraints</th>
<th>Total NOE (long range)</th>
<th>H$^\alpha$-$\text{HN}$</th>
<th>H$^\alpha$-CH$_3$</th>
<th>CH$_3$-CH$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67</td>
<td>23</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>Total dihedral angle restraints$^a$</td>
<td>292</td>
<td>146</td>
<td>146</td>
<td></td>
</tr>
</tbody>
</table>

Structural Statistics$^{cd}$

- **RDC Q-factors**
  - Data set 1 (PH1) 0.36
  - Data set 2 (gel) 0.40
- Average rmsd (Å)
  - Backbone 0.85
  - All heavy atom 1.25

EXPERIMENTAL PROCEDURES

Sample Preparation for SPR and NMR
DNA encoding m04 was PCR amplified from virus containing cell culture supernatants and has the sequence of the K181 laboratory strain of MCMV (Uniprot ID A2Q6L0), with the additional mutations I31V, I44V. m04ED was expressed in E. coli as insoluble inclusion bodies, denatured in 6M guanidine-HCl, refolded in vitro, and purified by size exclusion and ion-exchange chromatography. The final sample conditions for NMR were 0.5 mM m04ED, 50 mM NaCl in 20 mM NaH$_2$PO$_4$ buffer (pH 6.5). The quality of the 2D $^1$H-$^15$N HSQC correlation spectra was further improved by the single C7S resonance region at residues 95 to 105.

Cell Press

Please cite this article in press as: Sgourakis et al., The Structure of Mouse Cytomegalovirus m04 Protein Obtained from Sparse NMR Data Reveals a Conserved Fold of the m02-m06 Viral Immune Modulator Family, Structure (2014), http://dx.doi.org/10.1016/j.str.2014.05.018

Structure 22, 1–11, September 2, 2014 ©2014 Elsevier Ltd All rights reserved 7
NMR Backbone, Side-Chain Assignments, Backbone Relaxation Rates, and ILV NOE Measurements

All experiments were recorded at a temperature of 12 °C using 600 MHz, 800 MHz, and 900 MHz cryoprobe-equipped Bruker spectrometers. We used an array of standard triple-resonance assignment experiments (HNCO, HN(CA)CO, HNCA, and HN(CAB)C) supplemented with H(NINH) and H(NH) 3D NOE spectroscopy (NOESY) data sets, recorded at 600 MHz using a mixing time of 250 ms. All resulting spectra were processed with NMRPipe (Delaglio et al., 1995) and analyzed with Sparky (http://www.cgl.ucsf.edu/home/sparky/). Orthogonal projections from each 3D data set were extracted on the basis of the centered HNCO peak positions and visualized using the strip-plotting interface in nmrDraw (scroll.tcl), obtaining a highly consistent network of final assignments, with a completeness > 95% (excluding Pro residues). The strip-plot visualization allowed identification of 23 strong long-range Hα-1Hβ NOEs, sufficient to define a protein fold for the m04ED core, as outlined in detail in the structure calculation section.

To assign selectively labeled ILV methyls, we used the SIM-HMCM(CGCBCA)C (Tugarinov et al., 2014) and HMCM(CGCBCA)C (Tugarinov and Kay, 2003) triple-resonance spectra recorded at 600 MHz that link the methyl resonances to the previously established backbone C′ and Cβ/Cγ resonance assignments, respectively. These experiments were performed using a deuterated sample that was specifically 13C labeled at the side chains of ILV residues such as to yield a linear spin system (Tugarinov et al., 2006). In this manner, we obtained complete assignments for all 66 labeled Cα, Cβ/Cγ, and Cβ/Cγ methyls in the m04 sequence (containing 2 Ile, 16 Leu, and 16 Val residues, respectively), which were identified in a high-resolution methyl heteronuclear multiple-quantum coherence spectrum obtained at 900 MHz using a separate ILV sample that was 13C labeled only at the methyl carbon atoms (Figure S5). Finally, using the same sample, we recorded two complementary 3D methyl-to-amide NOESY spectra at 900 MHz and two methyl-to-methyl NOESY data sets at 600 MHz using 250 and 200 ms mixing times, respectively. The processed methyl NOE data further validated the methyl assignments and provided at total of 44 long-range contacts (19 Hα-1Hβ and 25 CH2-CH2) that were manually picked and readily assigned using Sparky. Backbone amide 15N R1 relaxation rates and heteronuclear NOE rates were measured from a perdeuterated, amide 1H sample using TROSY-readout methods (Lakomek et al., 2012). R2 rates were obtained from rotating-frame R2" rates (Massai et al., 2004) measured under a spin-lock field strength of 2 kHz, after correction for the 15N off-resonance, tilted field. Uncertainties in the R1 and R2 rates were measured from the spectral noise levels using 21 Monte Carlo simulations, while uncertainties in the heteronuclear NOE rates were propagated directly from the noise levels in the reference raw data, relative to a range of RDC values estimated from a randomly distributed set of vectors assuming an alignment tensor of known Dα and R parameters:

\[
Q = \frac{\text{RMSD}_{\text{obs}} - \text{RMSD}_{\text{calc}}}{\sqrt{\text{DS}(4 + 3R^2)}} / S
\]

where Dα and R refer to the magnitude and rhombicity of the alignment tensor, and D\text{obs} and D\text{calc} are the calculated and observed dipolar couplings, respectively. Q\text{obs} is the Q factor computed for a linearly independent RDC data set, not used in the structure calculations. Pearson’s linear correlation coefficient R\text{P} is defined as

\[
R\text{P} = \frac{\text{cov}(D\text{calc}, D\text{obs})}{\text{SD}_{D\text{calc}} \times \text{SD}_{D\text{obs}}}
\]

where cov(D\text{calc}, D\text{obs}) is the covariance between the individual computed and calculated RDCs, and SD\text{calc} and SD\text{obs} are the SDs of the experimental and computed RDCs, respectively.

Iterative CS-Rosetta Structure Calculations

We used the recently developed RASREC protocol (Lange and Baker, 2012), guided by the backbone CSs, a total of 67 long-range NOE distance restraints (23 Hα-Hβ, 25 CH2-CH2, and 19 Hβ-CH2) and 231 RDCs from two alignment media as N-H vector orientation restraints. In summary, the approach is based on several cycles of Monte Carlo-based trials, which include the sampling of...
backbone fragments (3-mers and 9-mers) and β strand pairings. Structural features that consistently lead to optimization of a target function (defined as the sum of the Rosetta energy and the experimental NOE and RDC score terms) are recombined during six generations of iterative structure calculations. In a series of benchmark calculations with targets of known structures, the protocol showed improved sampling efficiency over both standard CS-Rosetta structure calculations (Shen et al., 2008) and conventional NMR structure determination programs (Herrmann et al., 2002) for β-rich proteins with complicated topologies (Lange et al., 2012).

While the backbone CSs are used to guide the selection of backbone fragments from high-resolution structures in the PDB (Vernon et al., 2013), long-range NOEs help define the overall protein fold and β strand connectivity. RDCs are used as bond vector orientation restraints relative to an overall alignment frame to better define the local backbone structure, particularly in loop regions that have few or no long-range NOEs. The calculations are heavily restrained by the experimental data during the low-resolution backbone search. However, the final placement of side-chain rotamers and fine-tuning of the backbone torsion angles is guided primarily by the Rosetta energy function (Leaver-Fay et al., 2011), with almost no bias from the experimental score terms.

NOEs were implemented as flat-bottom restraints with an upper limit of 4.0 Å and an exponential penalty function. In the absence of stereospecific methyl proton assignments, all pairwise combinations of protons within each pair of interacting methyl sites were averaged as \( r/C_0 \) to compute an effective NOE distance (Nilges, 1993). In all calculations, the connectivity of disulfide bonds was also used as an input restraint using an orientation-dependent potential (Raman et al., 2009). All calculations were carried out in 200 threads of an SGI UV2000 cluster using Intel Xeon E5-4640 processors at 2.40 GHz with 1,024 GB of memory, customized to accommodate the MPI interface required by RASREC. Typical calculation runtimes under these conditions were 10 to 12 hr.

ACCESSION NUMBERS

The final 10 lowest energy models were deposited in the PDB (accession number 2MIZ).

Figure 7. Multiple Sequence Alignment of m04ED with the Predicted EDs of Other m02-m06 Family Members

The approximate positions of conserved secondary structure elements are highlighted on the top of each alignment block (β1–β9: β strands; red ovals: α helices). The position of the two disulfide bonds on the m04ED structure is indicated with capital C letters and connecting dashed lines, while high-confidence predicted glycosylation sites on the m04ED sequence are indicated with stars. Protein sequences for m02 (UniProt ID YP_214010), m03 (YP_214011), m05 (YP_214013), and m06 (YP_214014) were all taken from the Smith MCMV genome (NC_004065). The sequence of m03.5 (ABM74010), not present in the Smith strain, was taken from strain G1F.

To identify the ER-lumenal domains of these proteins from the full-length protein sequence, we first identified the end of the SP using the SignalP (http://www.cbs.dtu.dk/services/SignalP/) and Phobius (http://phobius.sbc.su.se/) servers, then identified the start of TM domains using TMpred (http://www.ch.embnet.org/software/TMPRED_form.html). Sequence alignments were constructed using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and colored by sequence similarity using BoxShade (http://www.ch.embnet.org/software/BOX_form.html).
SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and can be found with this article online at http://dx.doi.org/10.1016/j.str.2014.05.018.

ACKNOWLEDGMENTS

We thank Dr. Oliver Lange for providing consultation with setting up the RASREC-Rosetta protocol and Dr. Vitali Tugarinov for assistance with the SIM-HMCM(CGCBCA) and HMCM(CG)CBCA triple-resonance experiments. We are grateful to Drs. Nathan May for preparing multiple sequence alignments of the m02 family and to Drs. Yang Shen and Lishan Yao for preliminary work on determining optimal m04 sample conditions, as well as to Dr. Manqing Hong for providing the MULT-1 protein sample. This work is supported by the National Institute of Diabetes and Digestive and Kidney Diseases and National Institute of Allergy and Infectious Diseases intramural research programs and the Intramural AIDS-Targeted Antiviral Program of the Office of the Director, NIH.

Received: March 6, 2014 
Revised: May 23, 2014 
Accepted: May 27, 2014 
Published: August 7, 2014

REFERENCES


Kleijn, M.F., Huppa, J.B., Lucin, P., Mukherjee, S., Farrell, H., Campbell, A.E., Koszinowski, U.H., Hill, A.B., and Ploegh, H.L. (1997). A mouse cytomegalovirus glycoprotein, gp34, forms a complex with folded class I MHC molecules in the ER which is not retained but is transported to the cell surface. EMBO J. 16, 685–694.


Structure

m04 NMR Structure by Integrative Rosetta Modeling


