Original Article

In silico molecular dynamics of human glycophorin A (GPA) extracellular structure

Serena Ekman[1,2]^, Robert Flower[1,2]^, Stephen Mahler[1]^, Alison Gould[3], Ross T. Barnard[1,4]^, Catherine Hyland[2]^, Martina Jones[1]^, Alpeshkumar K. Malde[1,5,6]^, Xuan T. Bui[1,2]^

[1]ARC Training Centre for Biopharmaceutical Innovation, Australian Institute of Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD, Australia; [2]Australian Red Cross Lifeblood (Formerly Australian Red Cross Blood Service), Research and Development, Kelvin Grove, QLD, Australia; [3]Australian Red Cross Lifeblood (formerly Australian Red Cross Blood Service), Research and Development, Alexandria, NSW, Australia; [4]School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Australia; [5]Institute for Glycomics, Griffith University, Gold Coast Campus, Queensland, Australia; [6]MaldE Scientific, Brisbane, QLD, Australia

Contributions: (I) Concept and design: R Flower, AK Malde, S Ekman, XT Bui; (II) Administrative support: A Gould, S Mahler, M Jones, R Flower; (III) Provision of study materials or patients: S Ekman, AK Malde, XT Bui; (IV) Collection and assembly of data: S Ekman, AK Malde; (V) Data analysis and interpretation: S Ekman, AK Malde, XT Bui, RT Barnard, A Gould; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Xuan T. Bui. ARC Training Centre for Biopharmaceutical Innovation, Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Building 75, Cnr College Rd & Cooper Rd, St Lucia QLD 4072, Australia. Email: x.bui@uq.edu.au; Alpeshkumar K. Malde. Institute for Glycomics, Griffith University, Gold Coast Campus, Queensland, 4222, Australia. Email: a.malde@griffith.edu.au.

Background: Glycophorin A (GPA) is one of two glycoproteins found on the surface of human red blood cells (RBCs) that constitute the MNS blood group system. The structure of GPA's extracellular domain is unknown despite previous attempts using X-ray crystallography and NMR spectroscopy. As a result, there is a knowledge gap regarding antigen presentation. This hinders the production of monoclonal antibodies (mAbs) against MNS antigens.

Methods: *In silico* modelling techniques including homology modelling and ab initio predictions were implemented to develop a proposed structure for the monomeric form of the GPA extracellular domain. Developed structures were then subjected to molecular dynamics (MD) simulations.

Results: The results obtained indicated that the monomeric extracellular domain of GPA is most likely intrinsically disordered, with the exception of a β-hairpin-like structure spanning the exon 3–4 junction. Further analysis showed this β-hairpin-like structure was not observed when starting from an extended or cyclical peptide structure within the time-scale used in the MD simulation study, suggesting that linear or cyclical peptide fragments of this region are unsuitable representations for the purposes of antigen presentation. Lastly, of the MNS antigens produced by single amino acid variations found in the exon 3–4 junction, only the ERIK antigen (p.Gly78Arg; MNS37) was found to alter the β-hairpin-like structure.

Conclusions: The monomer of the extracellular domain of GPA has a high level of disorder, with the exception of the antigenic exon 3–4 junction, which adopts a β-hairpin-like structure. Our work suggests that linear peptides and expression of the monomeric form of GPA might be of limited use for immunisation or screening processes used in antibody identification. Further understanding of the antigenic determinants of GPA will require a more sophisticated combination of laboratory and computational approaches, as well as consideration of possible structural changes as a result of dimerisation.

Keywords: Glycophorin A (GPA); *in silico* modelling; intrinsically disordered protein; β-hairpin

^ ORCID: Serena Ekman, 0000-0001-8015-5309; Alpeshkumar K. Malde, 0000-0002-8181-1619; Robert Flower, 0000-0002-7257-1844; Stephen Mahler, 0000-0003-2403-1437; Ross T. Barnard, 0000-0002-5685-4828; Martina Jones, 0000-0002-5154-6017; Catherine Hyland, 0000-0002-4124-6168; Xuan T. Bui, 0000-0003-0881-7691.

© Annals of Blood. All rights reserved.

Received: 16 July 2020; Accepted: 30 December 2020; Published: 30 June 2021.
doi: 10.21037/aob-20-51
View this article at: http://dx.doi.org/10.21037/aob-20-51

Introduction

Glycophorin A (GPA) is a single-pass transmembrane sialoglycoprotein found in human red blood cells (RBCs), and is one of two glycoproteins that constitute the MNS blood group system. The MNS blood group system contains multiple clinically significant antigens, and has been associated with haemolytic disease of the foetus and newborn (1).

Unfortunately, there is currently a lack of monoclonal antibodies (mAbs) available for a number of antigens in the MNS blood group system (2). Previous attempts in our laboratory at producing mAbs against various glycophorin peptides in both GPA KO mice, and biopanning experiments resulted in antibodies with strong affinity towards the peptides, but little or no affinity to an RBC. It is possible that there are structural features associated with antigenicity of GPA that are not present in the peptide sequence alone. Therefore, a structural model of the extracellular domain of GPA (GPA-ECD) may act as a stepping-stone in understanding antigen presentation, and subsequently improve the way mAbs are identified.

Typically, protein structures are solved through the process of X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, the GPA-ECD is heavily glycosylated, and contains 16 O-linked and 1 N-glycosylation sites (3,4). This high level of glycosylation hinders the formation of a stable crystal for X-ray crystallography, and the high degree of heterogeneity associated with glycosylation adds further complexity when analysing any produced structures (5,6). It is possible to remove glycosylation through the use of neuraminidase cleavage or to express a soluble version of the protein in a bacterial system (7), however, there are still problems in retaining native structure through the preceding methods. NMR has been previously used to solve the transmembrane domain structure of GPA (8), however, there is limited information about the secondary structure of GPA-ECD. What is available from NMR and circular dichroism (CD) studies, which includes studies with neuraminidase cleavage of glycosylation, indicates that there is a disordered portion of GPA-ECD structure (9,10). Apart from consensus that GPA is a single-pass transmembrane protein, no detailed structure of GPA-ECD or its antigens has emerged from these methods to date.

Due to the limited availability of experimental data, combined with the difficulty of obtaining structural data for the GPA-ECD, three structure prediction methods of increasing complexity were used: (I) homology/comparative modelling, (II) threading and (III) *ab initio* simulation. This was performed to better understand the molecular basis of GPA antigenicity, as well as whether these *in silico* modelling methods can provide a prediction of GPA secondary structure and subsequent antigen conformation. Although glycosylation is a factor in antigenicity, there is limited evidence to suggest it has a significant role in protein backbone structure. Due to the increased level of complexity as well as processing power required to simulate glycosylation, this study produced an initial model of GPA-ECD structure based on an amino acid backbone. Protein glycosylation will be added once a structure has been determined in order to investigate its role in antigenicity.

Methods

Generation of 3D structures of GPA-ECD

GPA-ECD

The amino acid sequence of GPA was obtained from Uniprot (Universal Protein Resource, RRID:SCR_002380) (11) accession number P02724 (GLPA_HUMAN). The amino acid sequence of the extracellular domain of glycophorin A (GPA-ECD, residues 19–89; 71 residues) were submitted to Robetta (Robetta RRID:SCR_018805) (12), iTasser (iTasser RRID:SCR_018803) (13) and FALCON (FALCON RRID:SCR_018804) (14). The top 5–10 models generated by all three programs were used for analysis. Five *ab initio* models generated by Robetta (12) were further analysed and validated using atomistic molecular dynamics (MD) simulations (see "MD simulations" below). In addition, a previously predicted GPA model with a β-barrel structure (15) generated using the program FALCON (14), for the extracellular domain GPA-ECD (residues 21–91; 71 residues) was also used for MD simulations. All amino acid sequences for GYPA

© Annals of Blood. All rights reserved.

models can be seen in *Table 1*.

GPA-ECD exon 3–4 junction hairpin peptide

A representative structure was extracted from clustering data produced on the GPA-ECD exon 3–4 junction from the combined trajectories of the 5 Rosetta GPA-ECD models. Two structures (residues 64–85; 22 residues) were selected, in the form of a beta-hairpin with N-terminal acetyl and C-terminal amide caps added using PyMOL (PyMOL, RRID:SCR_000305) version #2.2.3 (16).

GPA-ECD exon 3–4 junction linear peptide

An extended peptide model of the GPA-ECD exon 3–4 junction (residues 64–85; 22 residues) was produced using the builder tool of PyMOL (16) with N-terminal acetyl and C-terminal amide caps added.

GPA-ECD exon 3–4 junction circular peptide

Two circular peptide models for GPA-ECD exon 3–4 junction (residues 64–85; 22 residues) were produced using Rosetta (12). The topology files were generated using a combination of GROMOS (17) and GROMACS (GROMACS, RRID:SCR_014565).

GPA-ECD exon 3–4 junction single amino acid mutations

FoldX software (FoldX, RRID:SCR_008522) (18) was used to create single amino acid changes in the GPA-ECD exon 3–4 junction hairpin peptide (produced using method above) according to known MNS antigens caused by single amino acid mutations in Table S1. The modelled structures retained the N-terminal acetyl and C-terminal amide caps.

MD simulations

All MD simulations were performed using the GPU version Gromacs 2019.1 (19) on the Wiener HPC cluster at the University of Queensland. The Gromos 54A7 force field was used to model protein structures (20,21). Each protein was placed in a cubic periodic box with the minimum distance of 1.2 nm between protein surface and wall. Each system was solvated with SPC (22) water model. The protonation states of titratable groups were chosen appropriate to pH 7.0, where N-terminal, Arg and Lys residues were protonated while C-terminal, Asp and Glu residues were deprotonated. The neutral ε-tautomer for His residues was used. No counter ions were added.

Each system (protein + water) was energy minimized and equilibrated for 200 ps with the heavy atoms of the protein positionally restrained before commencing a series of unrestrained MD simulations. All simulations were performed at constant temperature (298 K) and pressure (1 atm) using a Berendsen thermostat (coupling time of 0.1 ps) and barostat [coupling time of 1.0 ps and isothermal compressibility of 4.575×10^{-4} $(kJ/mol/nm^3)^{-1}$] (23,24). A single cut-off of 1.4 nm was used for all non-bonded interactions. The neighbour list was updated every 0.010 ps (every 5 steps). To correct for the truncation of the electrostatic interactions beyond the 1.4 nm long-range cut-off a reaction-field correction was applied using a dielectric permittivity of 78. The equations of motion were integrated using the leapfrog scheme and a step-size of 0.002 ps. Initial velocities at a given temperature were derived from a Maxwell-Boltzmann distribution. All bonds were constrained using the LINCS algorithm with a lincs_order of 4 (25). The SPC water molecules were constrained using the SETTLE algorithm (26). MD simulations were performed for 200 ns for each system (with the exception of GPA-ECD exon 3–4 junction single amino acid mutation models which were run for 100 ns); all coordinates, velocities, forces and energies were saved every 10,000 steps (20 ps) for analysis.

Analysis using RMSD, Clustering and Visual Analysis

The stability of protein structures was analysed using root mean square deviation (RMSD) as obtained by fitting the backbone atoms and calculating the RMSD for backbone atoms. Additional secondary-structure analysis was performed to check the stability and inter-change of secondary structure elements including visual analysis of the trajectory in VMD program (27) (Visual Molecular Dynamics, RRID:SCR_001820) version 1.9.3. The clustering of the relevant combined MD simulation trajectories (5 Robetta models for GPA-ECD, GPA-ECD exon 3–4 junction region as a representative hairpin, extended and circular model, and 7 FoldX GPA-ECD exon 3–4 junction mutation models), was performed using the inbuilt analysis tool in GROMACS. The samples were run for 200 ns trajectory contained 10,000 frames, with the exception of the exon 3–4 junction mutation models which were run for 100 ns for a total of 5,000 frames. The clustering method of Daura *et al.* [1999] as implemented in GROMACS under the name 'GROMOS method' was used (28). For clusters of structures in an MD simulation trajectory, the RMSD of atom positions between all pairs

© Annals of Blood. All rights reserved.

Table 1 GYPA extracellular sequence (M antigen) accession number P02724 (GLPA_HUMAN) sequence identification numbers for modelled structures 1–6, exon 3–4 peptide and International Society of Blood Transfusion (ISBT) amino acid numbering system

Single letter amino acid code	ISBT	Structure 1–5	Structure 6	Exon 3–4
M	1	–	–	–
Y	2	–	–	–
G	3	–	–	–
K	4	–	–	–
I	5	–	–	–
I	6	–	–	–
F	7	–	–	–
V	8	–	–	–
L	9	–	–	–
L	10	–	–	–
L	11	–	–	–
S	12	–	–	–
E	13	–	–	–
I	14	–	–	–
V	15	–	–	–
S	16	–	–	–
I	17	–	–	–
S	18	–	–	–
A	19	1	–	–
S	20	2	–	–
S	21	3	1	–
T	22	4	2	–
T	23	5	3	–
G	24	6	4	–
V	25	7	5	–
A	26	8	6	–
M	27	9	7	–
H	28	10	8	–
T	29	11	9	–
S	30	12	10	–
T	31	13	11	–
S	32	14	12	–
S	33	15	13	–
S	34	16	14	–
V	35	17	15	–
T	36	18	16	–
K	37	19	17	–
S	38	20	18	–
Y	39	21	19	–
I	40	22	20	–
S	41	23	21	–
S	42	24	22	–
Q	43	25	23	–
T	44	26	24	–
N	45	27	25	–
D	46	28	26	–
T	47	29	27	–
H	48	30	28	–
K	49	31	29	–
R	50	32	30	–
D	51	33	31	–
T	52	34	32	–
Y	53	35	33	–
A	54	36	34	–
A	55	37	35	–
T	56	38	36	–
P	57	39	37	–
R	58	40	38	–
A	59	41	39	–
H	60	42	40	–
E	61	43	41	–
V	62	44	42	–
S	63	45	43	–
E	64	46	44	1
I	65	47	45	2
S	66	48	46	3
V	67	49	47	4

Table 1 (*continued*)

© Annals of Blood. All rights reserved.

Table 1 (*continued*)

Single letter amino acid code	ISBT	Structure 1–5	Structure 6	Exon 3–4
R	68	50	48	5
T	69	51	49	6
V	70	52	50	7
Y	71	53	51	8
P	72	54	52	9
P	73	55	53	10
E	74	56	54	11
E	75	57	55	12
E	76	58	56	13
T	77	59	57	14
G	78	60	58	15
E	79	61	59	16
R	80	62	60	17
V	81	63	61	18
Q	82	64	62	19
L	83	65	63	20
A	84	66	64	21
H	85	67	65	22
H	86	68	66	–
F	87	69	67	–
S	88	70	68	–
E	89	71	69	–
P	90	–	70	–
E	91	–	71	–

of structures were determined. For each structure, the number of alternate iterations either similar or dissimilar, based on backbone atom RMSD values less than or equal to a specified value determined by RMS distribution analysis, was calculated. The structure with the highest number of neighbours was taken as the centre of a cluster, and formed the complete cluster together with all its neighbours. The structures of this cluster were thereafter eliminated from the pool of structures, and the process was repeated until the pool of structures was empty. An RMS distribution plot was generated to select appropriate cut-off RMSD values in the cases where more than one unique cluster was found.

Secondary structural features were automatically identified via the internal algorithms of the Visual Molecular Dynamics program (VMD). The program feature 'Timeline' was used to produce a graphical representation of secondary structure associated with each amino acid across the timeline of the full MD simulation. This tool was used to determine the specific amino acids associated with secondary structures, as well as the duration of the presence of the secondary structural feature throughout the simulation.

Circular dichroism (CD) spectropolarimetry of GPA-ECD

A peptide containing the GPA extracellular domain sequence [SSTTGVAMHTSTSSSVTKSYISSQTNDTH KRDTYAATPRAHEVSEISVRTVYPPEEETGERVQLA HHFSEPE] was synthesised by Thermo Fisher. CD spectra of GPA-ECD peptide (100 µM in 20 mM $KHPO_4$, pH 6.0) were acquired under constant N_2 flush using a Jasco J-810 spectropolarimeter. Measurements were taken at 0.2-nm wavelength increments from 195 to 250 nm at 100 nm/min using a cell with a path length of 1 mm, bandwidth of 2 nm, response time of 1 s and five accumulations, and corrected for buffer baseline contribution.

Results

Development and analysis of initial structures

Homology modelling

Homology/comparative modelling was attempted, although due to overall very low sequence identity and low homology of the GPA-ECD to known 3D structures of proteins (both solved by crystallography as well as NMR as reported in the Protein Data Bank (PDB) [Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), RRID:SCR_012820)] this method was not suitable. Partial matches obtained by performing BLAST (29) (NCBI BLAST, RRID:SCR_004870) analysis against the PDB database revealed fragments of 20-30 residues in length from GPYA-EC exhibiting a sequence identity of 38–50% with known structures. However, homology modelling techniques require a minimum of approx. 150 residues with homology of at least 30–50% (30,31). As all the returned sequences were so short, it is unlikely that they reflect the appropriate secondary/tertiary structure as they are missing numerous interactions with other surrounding structures. Most of these segments were helical in nature (Table S2), but were too small to build a reliable homology model.

© Annals of Blood. All rights reserved.

Threading—iTasser and FALCON

As homology modelling was an inappropriate method for determining a suitable starting structure, the threading method (iTasser) was implemented for the structure prediction for GPA-ECD. The fold prediction is made by "threading" each amino acid in the target sequence to a position in the top 10 template structures generated from the LOMETS threading programs. For each target, iTasser uses the SPICKER program to cluster large ensembles of structural conformations based on pair-wise structure similarity and selects the top 5 models, where the confidence of each is measured by a C-score (32-40). However, this method relies on the fact that the total number of protein folds in nature is much smaller than the total number of known sequences (41,42).

None of the 5 models of GPA-ECD produced by iTasser had positive or relatively high C-scores, indicating a low level of confidence. Visual analysis using PyMOL and VMD revealed these 5 models mostly exhibited a random-coil structure, in contrast to comparative modelling of short GPA sequence segments (approx. 25 amino acids long), where all results returned as α-helical structures.

Subsequently, an alternative threading method, FALCON was used. FALCON uses homologous templates to calculate common structural frameworks using the TBM (Template Based Modelling, including homology and threading) module. If the TBM module fails to identify protein homologues, the *ab initio* module is activated. Initial attempts for GPA-ECD led to generation of 10 structural models, but none had confidence levels that justified continuation of study.

A previous model had been created using the FALCON program (15), so that program was also adopted for the current work. The previous model was created using residues 21–91 (71 residues), as compared to 19–89 (71 residues) used in the current study. The previously created structure consisted of 5 highly coiled, antiparallel β sheets in 1-2-3-5-4-1 topology, referred to as an OB-fold (oligonucleotide/oligosaccharide-binding fold). In our attempts to replicate the previous model using FALCON, the β-barrel structure was obtained only as 10^{th} best-ranked model. Additionally, this model was only possible to obtain using the amino acid residues 21–91, where no β-barrel structures appeared using residues 19–89. Notwithstanding this, our produced replicate model was renamed 'Structure 6' and included as one of the selected models for further analysis. However, the sensitivity of the threading method towards a small addition/deletion of residues on the end-regions was surprising, and more investigations may be needed (beyond the scope of the current work) to support this observation.

Ab Initio modelling—Robetta

Due to the divergent structural observations obtained from the homology modelling and threading methods, prediction of the 3D structure of GPA-ECD was attempted using the program Robetta that implements *ab initio* methods. The *ab initio* Relax application consists of two main steps. The first step is a coarse-grained fragment-based search through conformational space using a knowledge-based "centroid" score function that favours protein-like features. The second optional step is all-atom refinement using the Robetta full-atom force field (43). Robetta was selected due to its success in the bi-annual "Critical Assessment of Protein Structure Prediction" (CASP) (44) experiments, which test the latest methods for protein structure prediction. The top 5 models obtained from Robetta were renamed as Structures 1–5 respectively and chosen for further investigation. When selecting models, weight was given not only to the ranking of the model by the Robetta program, but also to the expectation that residues involved in glycosylation were exposed to the surface.

Analysis of GPA-ECD structures

All six initial predicted structures (5 from Robetta and 1 from FALCON) exhibited distinct secondary structural features in the form of 3 to 6 β-sheets (listed in Table S3). There was a consistent feature of two β-sheets across all models, where the first β-sheet was located within the region 64–71 [EISVRTVY] and the second was located within the region 80–87 [RVQLAHHF] (Table S3). Although each β-sheet differed slightly in length and location, these β-sheets were also retained across all models throughout the simulation using the GROMOS force field. The combined region 64–85 is highlighted in red in *Figure 1* as a comparison between the initial structure predictions. This β-hairpin is also the site of a number of GPA antigens (Table S1). Another region of consensus was a β-sheet found to span the region 25–30 [VAMHTS] (Table S3). However, the β-sheet identified at 25–30 disappeared within 20 ns of MD simulation and was thus not considered a stable β-sheet region. Using secondary structure prediction program PsiPred (45) short β-sheets were predicted to be located in corresponding location 64-71 [EISVRTVY] (Figure S1). Despite the consensus of

© Annals of Blood. All rights reserved.

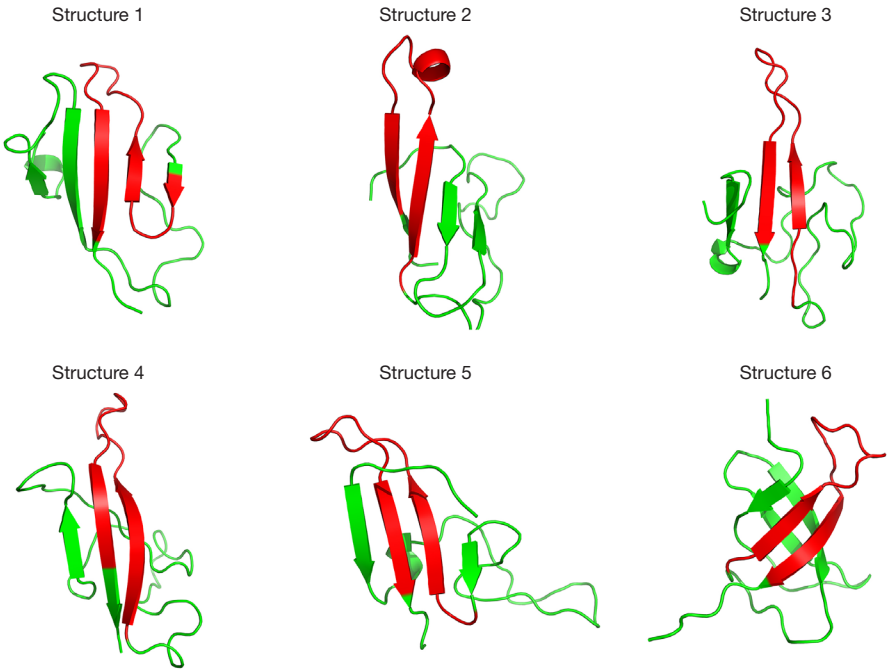

Figure 1 Initial predicted structures of the extracellular domain of glycophorin A. The extracted stable exon 3–4 region (extracted from the results of 1 μs molecular dynamics simulation) is highlighted in red (EISVRTVYPPEEETGERVQLAH) for comparison between the identified stable β-hairpin against initial structure predictions. Images visualised using PyMOL (15).

the three aforementioned β-sheets, the structure for the entire GPA-ECD of the initial predictions varies greatly, and is likely to be associated with the inherent uncertainties in predicting secondary structure from primary sequence alone (46,47).

MD simulations were performed in explicit water on the six initial structures to validate the stability of these secondary structures. Furthermore, as the initial structures differed from one another, MD was used to determine if two or more would converge into a single stable representative form. After 200 ns of MD in explicit solvent, visual analysis indicated that all models were unable to retain their secondary structure, failed to conform into a single common structure, and all tended towards complete unfolding of the proteins captured at 0, 100 and 200 ns (*Figure 2*).

For Structures 1, 2 and 3, a notable feature was the lengthening or shortening of β-sheets as well as significant twisting of the β-sheets seen by visual analysis (*Figure 2*). This distortion of shape correlates with the fluctuating RMSD values shown in the Y-axis of Figure S2 for Structures 1, 2 and 3. A stable conformation was not found for Structure 4 as it completely unfolds to form a random-coil structure after 200 ns of MD simulation (*Figure 2*), also illustrated in Figure S2. The RMSD value of Structure 4 continuously increases from 0.3 to 0.6 nm over the entire 200 ns simulation, indicating continuous protein backbone alteration. Structure 5 had more widely fluctuating RMSD values (Figure S2, where the changes in secondary structure are visible across *Figure 2*). Structure 6 presented with the highest level of secondary structure stability. Although the RMSD values varied within the first 50 ns between 0.3 and 0.6 nm (Figure S2), indicating an unstable starting structure [most likely the alteration of the β-sheet at 27–30 (MHTS)], then at 50 ns there was a sharp spike in RMSD to 0.65 nm indicating a transition in the secondary structure, which remains stable at 0.6–0.65 nm for the rest of the 200 ns simulation. This revealed that a reasonably stable structure has been obtained, which is corroborated by visual inspection of the structure at each time point shown in *Figure 2*. The RMSF was also calculated for each of the protein structures (Figure S3) where Structures 1, 2, 4 and 5 show high levels of fluctuations throughout the protein. Structures 3 and 5 appear to have lower levels of fluctuation, with the exception of two peaks both between residues 40 to 50, and 70 to 80, as well as high fluctuation on the termini

© Annals of Blood. All rights reserved.

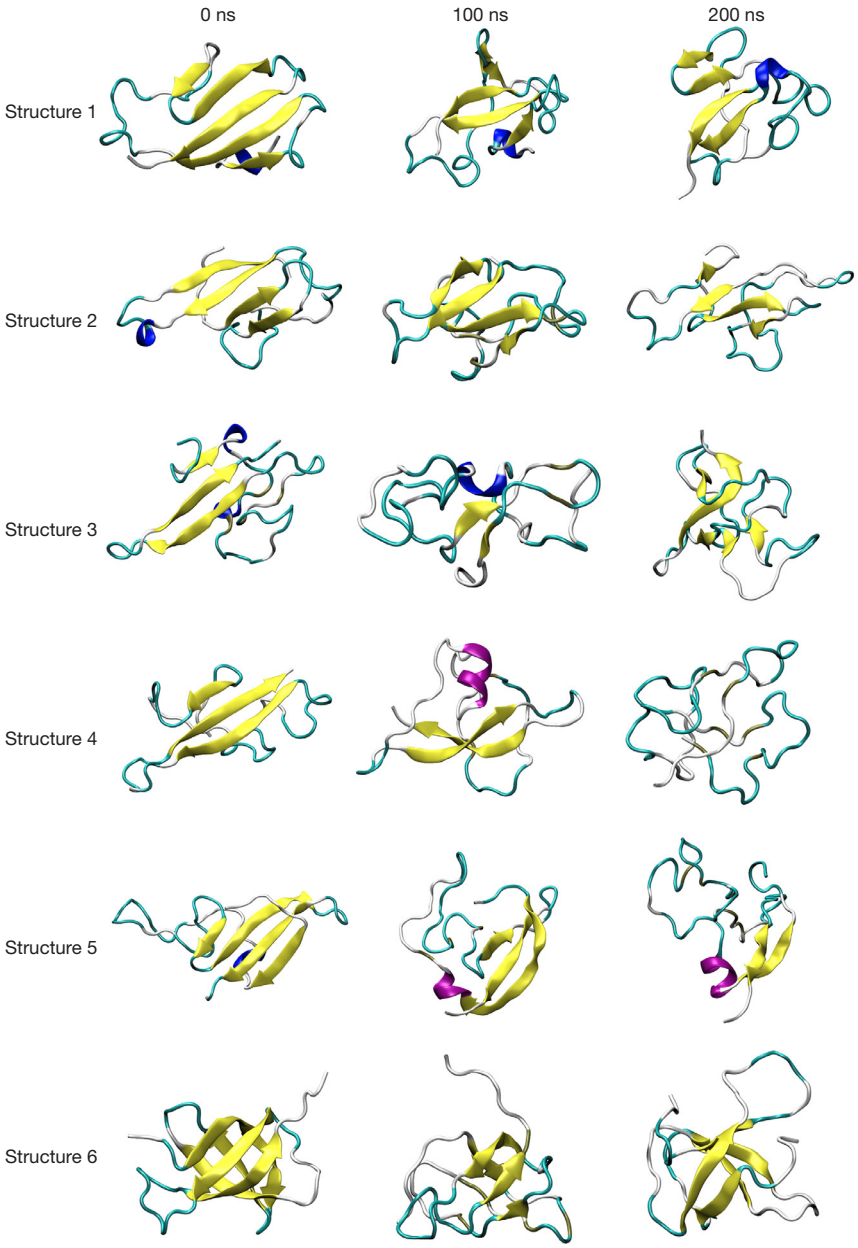

Figure 2 Snapshots of all 6 initial extracellular predicted glycophorin A structures after molecular dynamics at 0 ns, 100 ns and 200 ns. Structure represented in cartoon format, with colouring based on secondary structure; α-helices are coloured maroon, 3–10 α-helices are coloured dark blue, β-sheets are coloured yellow, and loops are coloured white and cyan. Images derived from VMD software (26).

of the protein. However, under visual analysis (*Figure 2*) it can be seen that the secondary structural features shift from β-sheets to random coils, as well as changes in the tertiary structure.

Across all six structures, the twisting of β-sheets and presence of transient α-helical segments (*Figure 2*) suggests a propensity to shift into an α-helical structure. However, the majority of observed α-helical segments appeared intermittently at different locations across the six structures as well as at different time-points during the simulation (Table S4). The majority of the helices produced were also 3–10 helices which have been proposed

© Annals of Blood. All rights reserved.

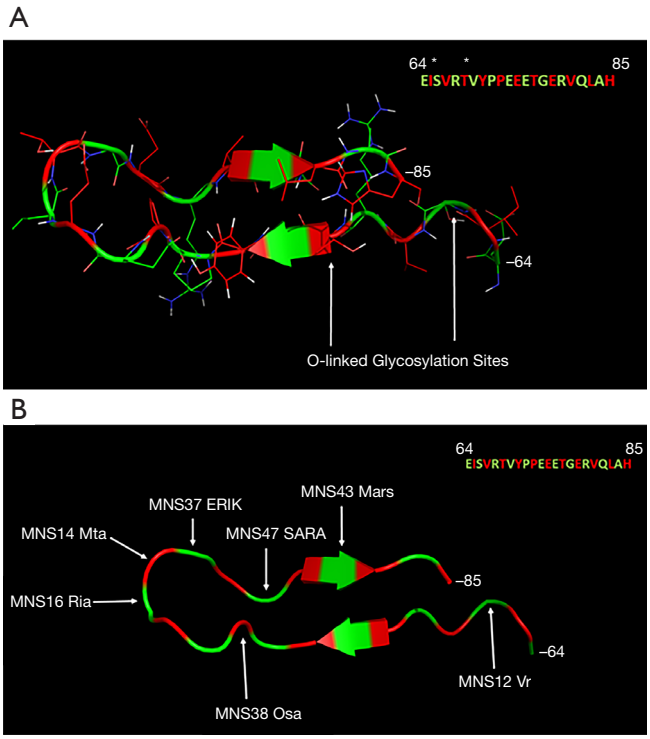

Figure 3 A representative model of the glycophorin A (GPA) exon 3–4 junction hairpin structure visualised using PyMOL (15). (A) Depiction of GPA exon 3–4 junction hairpin structure with cartoon backbone and line side chains, with labelled O-linked glycosylation sites. (B) Depiction of GPA exon 3–4 junction hairpin structure showing cartoon backbone only, with labelled single amino acid variants.

to be an intermediate conformation, although four (of twelve) observed helices were not 3–10 helices. Lastly, the α-helical segments were on average 3 amino acids long and thus were not considered to be a significant structural feature even if they were retained for the majority of the 200 ns simulation.

Although the overall secondary structure failed to find a stable form, there were two stable β-sheets. The β-sheet at 81–87 [VQLAHHF] was retained throughout the 200 ns MD simulation for Structures 1,2,3,5 and 6, as well as for more than 100 ns in Structure 4. This was also observed for the β-sheet at 64–67 [EISV] (with the exception of Structure 3). These two β-sheets are located in the exon 3–4 junction region. These sheets are present in the same location as in the initial Robetta and FALCON predicted structures (*Figure 1* in red), indicating that these β-sheets may be a stable component of the GPA-ECD. A representative structure of the exon 3–4 junction region was obtained from the combined trajectories of 1 μs of MD simulation, this structure exhibited a β-hairpin-like structure with an internal protein backbone, with side chains and glycosylation sites exposed externally (*Figure 3A*). This structure is also the location for seven GPA antigens which have been identified on *Figure 3B* and Table S1.

In relation to the addition of counter-ions, there is currently no strong evidence that neutralisation of the protein enhances stabilisation, and is highly dependent on the model variant and procedure used (48). However, a test for MD simulation of Structure 1 was performed in a 100 mM NaCl environment and run for 100 ns in explicit water, where no additional stability was seen.

GPA-ECD as an intrinsically disordered-like protein

Considering the size of the GPA protein, four independent approaches were used (*ab initio*, threading, homology modelling and MD simulations) to derive a consensus about the secondary structure of its extracellular domain. The results of the MD simulations across the 6 structures lead to the observation that monomeric form of GPA-ECD most likely exists in an intrinsically disordered form with a small predicted region of local stability of the β-hairpin-like structure across the exon 3–4 junction.

In addition, experiments were performed to test if GPA-ECD is disordered, using circular dichroism (CD) spectropolarimetry. A non-glycosylated peptide of the GPA-ECD was analysed and the CD spectrum indicated that the peptide does not have significant α-helical or β-sheet secondary structure elements and it is most likely in a random coil conformation in solution (*Figure 4*).

Another method of predicting IDPs is through sequence analysis (49). This follows the principle that IDPs have lower frequencies of order-promoting residues (W, C, I, F, D and L) and hydrophobic/aromatic residues, as well as higher frequencies of disorder-promoting residues (R, P, Q, G, E, S, A and K) and charged/polar residues (50,51). This pattern can be seen in the GPA-ECD (as shown in *Table 2*), where over half of the residues within the extracellular domain are disorder promoting, and only approximately one fifth are order promoting. Lastly, secondary structure prediction based on the entire GPA protein sequence using PsiPred and DISOPRED (45,52) predicted disordered regions predominantly in exon 2 and 4 (Figure S1).

© Annals of Blood. All rights reserved.

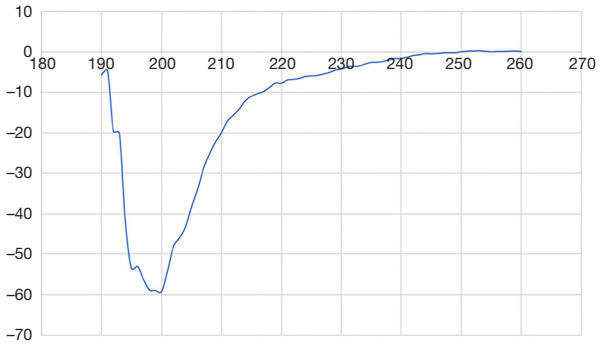

Figure 4 Circular dichroism (CD) analysis of the glycophorin A extracellular domain (GPA-ECD) peptide.

Table 2 Table of order promoting and disorder promoting residues in the modelled glycophorin A extracellular domain (GPA-ECD) (amino acids 19-92—inclusive of 1 amino acid on either side of the extracellular domain)

Amino acid	# in GPA-ECD	Amino acid	# in GPA-ECD	Total
Order promoting residues				
W†	0	F†	1	16/73
C¶	0	V‡	6	
Y†	3	D§	2	
I‡	3	L‡	1	
Disorder promoting residues				
R§	4	E§	8	40/73
P‡	4	S¶	12	
Q¶	2	A‡	6	
G‡	2	K§	2	

†, aromatic; ‡, hydrophobic; §, charged; ¶, polar.

Antigen presentation in GPA-ECD

The M/N (MNS1/2), MNTD (MNS46) and Nya (MNS18) antigens were located in regions on the GPA-ECD that were highly unstable and had no consensus for secondary structure across the six models after MD simulation. As a result, no conclusive data could be determined for secondary structure prediction of these antigenic sites.

The exon 3–4 junction was the only region of stability across the six models, and is the location for seven MNS antigens encoded by single amino acid changes (Table S1 and *Figure 3B*). To determine whether these alterations would reduce the stability of the β-hairpin-like structure, or allow the formation of new secondary structures, seven new structures were produced. Each new β-hairpin-like structure contained a single amino acid change corresponding to one of the seven MNS antigens (Table S1). It was found that after 100 ns of MD simulation, only one single amino acid variation caused disruption to the β-hairpin structure; p.Gly78Arg (MNS:37 ERIK+) (*Figure 5*). Although slight alterations were noted in p.Ser66Tyr (MNS:12 Vr+), p.Pro73Ser [MNS:38 Os(a+)] and p.Gly82Lys (MNS:–42,43 MARS+), by the end of the 100 ns simulation the β-hairpin-like structure had re-formed. For p.Glu76Lys [MNS:16 Ri(a+)], p.Thr77Ile [MNS:14 Mt(a+)] and p.Arg80Ser (MNS:47 SARA+), the core β-hairpin-like structure was maintained. Additionally, no single amino acid changes were able to produce a stable alternative structure to the β-hairpin. This indicates that only the amino acid change associated with the ERIK antigen causes a structural change in GPA-ECD, and this would be expected to have an impact on the presentation of the antigen for antibody production. It should also be noted that p.Ser66Tyr (MNS:12 Vr+) mutation is also an O-linked glycosylation site, and the alteration of the Serine to a Tyrosine would result in the elimination of this glycosylation site, which would be expected to change the surface presentation of GPA-ECD.

Development of suitable peptides mimicking the β-hairpin-like structure of the exon 3–4 junction region

The coordinates of the representative hairpin structure obtained from the 1 μs MD simulation of GPA-ECD were used to extract a shorter (22 amino acid) peptide sequence that was used for further computational analysis. The hairpin peptide ran for 200 ns MD simulation and was able to retain its secondary structure despite slight extension and shortening of the β-sheet (shown in *Figure 6*). As peptide synthesis is most commonly carried out linearly, the extended peptide structure (with no pre-determined structure) was also tested to determine if a linear peptide would be able to fold into the β-hairpin-like structure.

MD simulations starting with the linear peptide, in contrast to the simulations carried out from the hairpin structure, were unable to fold into the β-hairpin-like structure or conform into a single stable structure after 200 ns of MD simulations (shown in *Figure 7*).

As the linear peptide was deemed to be an unsuitable candidate for laboratory experimental purposes, it was hypothesised that a circular peptide might replicate the modelled β-hairpin-like structure without the need for a

© Annals of Blood. All rights reserved.

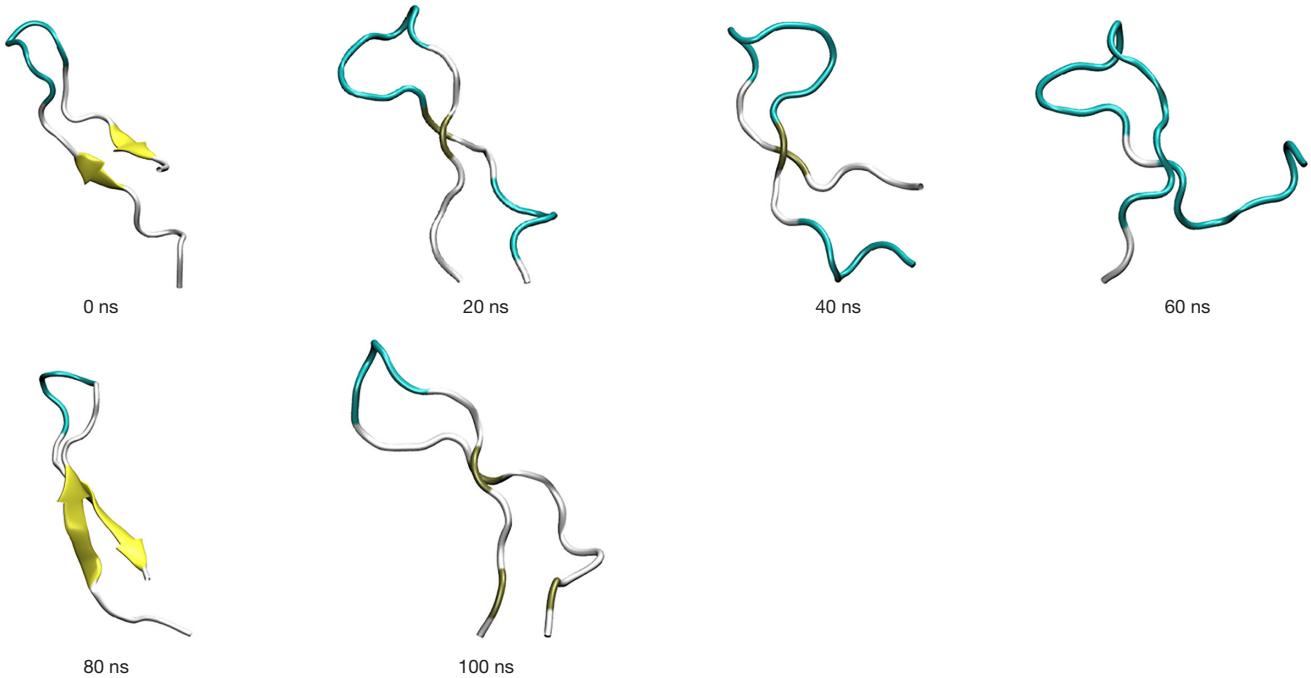

Figure 5 Representation of glycophorin A exon 3–4 structure MNS37 ERIK mutation (p.Gly78Arg) at 20 ns time points starting from a hairpin conformation. Peptide represented in new cartoon format, with colouring based on secondary structure; white and blue are loops, yellow is β-sheets. Images derived from VMD software (26).

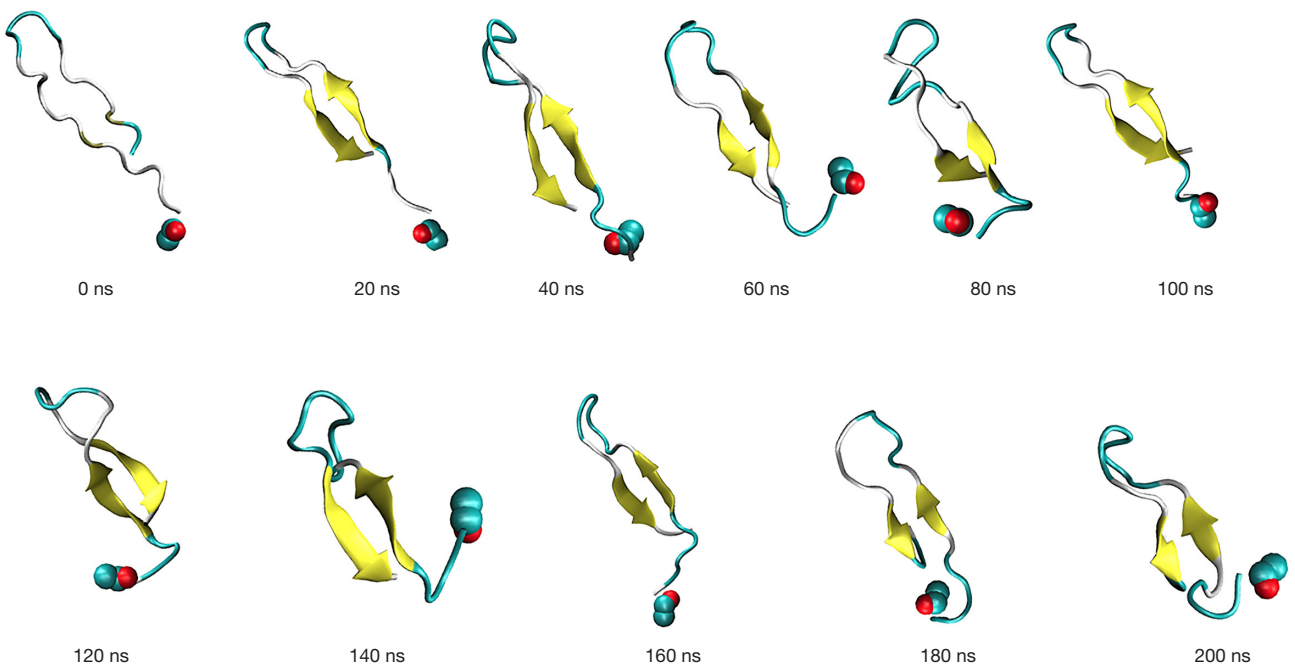

Figure 6 Representation of glycophorin A exon 3–4 structure at 20 ns time points starting from a hairpin conformation. N-terminus represented by Van Der Walls (VDW) ball structure, and peptide represented in new cartoon format, with colouring based on secondary structure; white and blue are loops, yellow is β-sheets. Images derived from VMD software (26).

© Annals of Blood. All rights reserved.

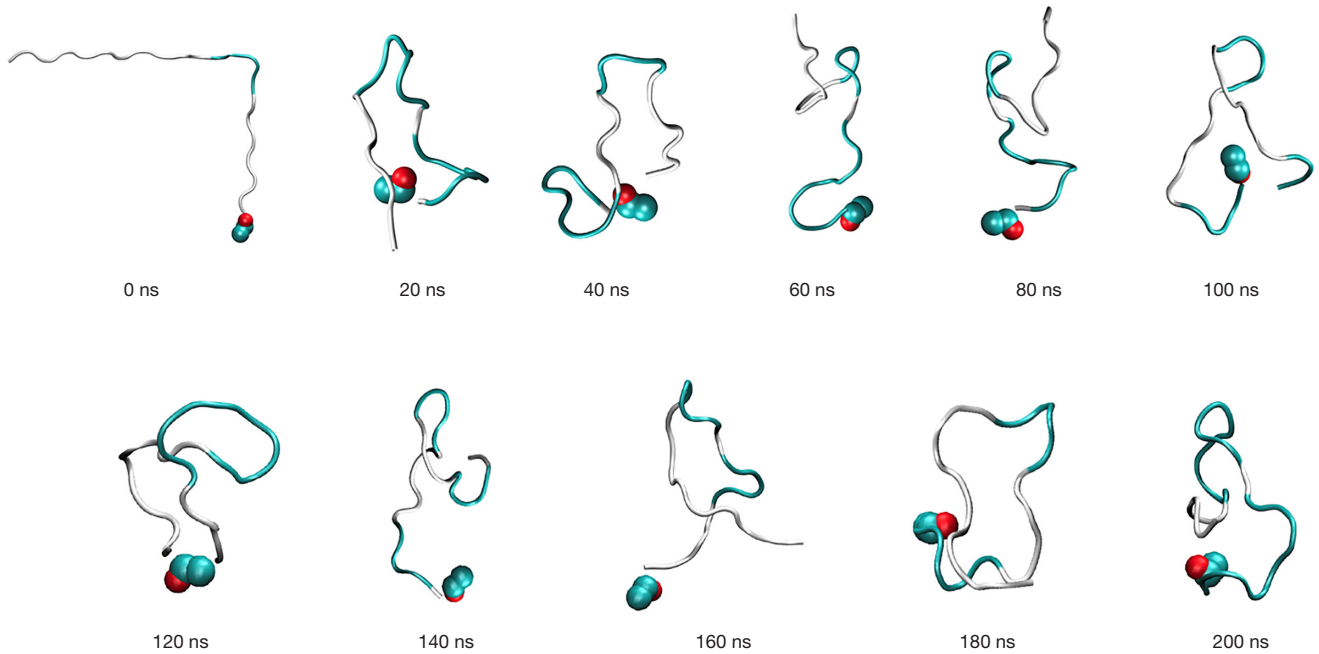

Figure 7 Representation of glycophorin A exon 3–4 structure at 20 ns time points starting from an extended conformation. N-terminus represented by Van Der Walls (VDW) ball structure, and peptide represented in new cartoon format, with colouring based on secondary structure; white and blue are loops. Images derived from VMD software (26).

protein scaffold. The modelled circular peptide had limited ability to retain the β-hairpin-like structure, as can be seen in *Figure 8*. Although the circular peptide was able to retain a hairpin-like structure for the first 80 ns, initial twisting can be seen at 120 ns, as well as further folding and collapsing of the circular peptide from 160 ns onward. With a shortened circular peptide length this twisting and folding may reduce.

When starting from a β-hairpin structure, or within the confines of the entire GPA-ECD, the β-hairpin peptide was stable. However, peptides that did not start as a β-hairpin (starting from a circular or linear peptide) were unable to re-form the β-hairpin-like structure after 100 ns MD.

Discussion

From the initial searches for homologous protein structures, it was assumed that our predicted structure would similarly contain α-helices. However this was not the case across any of our models. This is likely due to the predicted behaviour of short peptide fragments (20–30 amino acids) not necessarily being replicated in a large molecular structure. In particular, that the formation of β-sheets relies heavily upon tertiary structure and cannot be replicated in smaller fragments (53). As secondary structure is determined by a multiplicity of molecular interactions outside the short domain subjected to sequence comparison, it is not surprising that the short sequences of homology did not align with the produced models.

A key setback in our initial models is the lack of glycosylation. There is evidence that glycosylation might not alter the structure and folding of proteins, but rather acts to enhance protein stability via destabilizing the unfolded protein state (54,55). Alternatively, glycosylation removal allows for recognition of misfolded glycoproteins for proteasome degradation (56). Additionally, a previous study by Ekman *et al.* [2019] using a similar structure to Structure 6 indicated that the addition of glycosylation (substituting full glycosylation for α-N-Acetyl-D-galactosamine molecules) had a slight stabilization effect but did not play a major role in the determination of GPA-ECD structure (15). On this basis, it is unlikely that the lack of glycosylation is the cause of the disordered nature of the simulated structures. As a result, there was sufficient confidence in the ability to produce accurate secondary structures using the protein backbone and *in silico*

© Annals of Blood. All rights reserved.

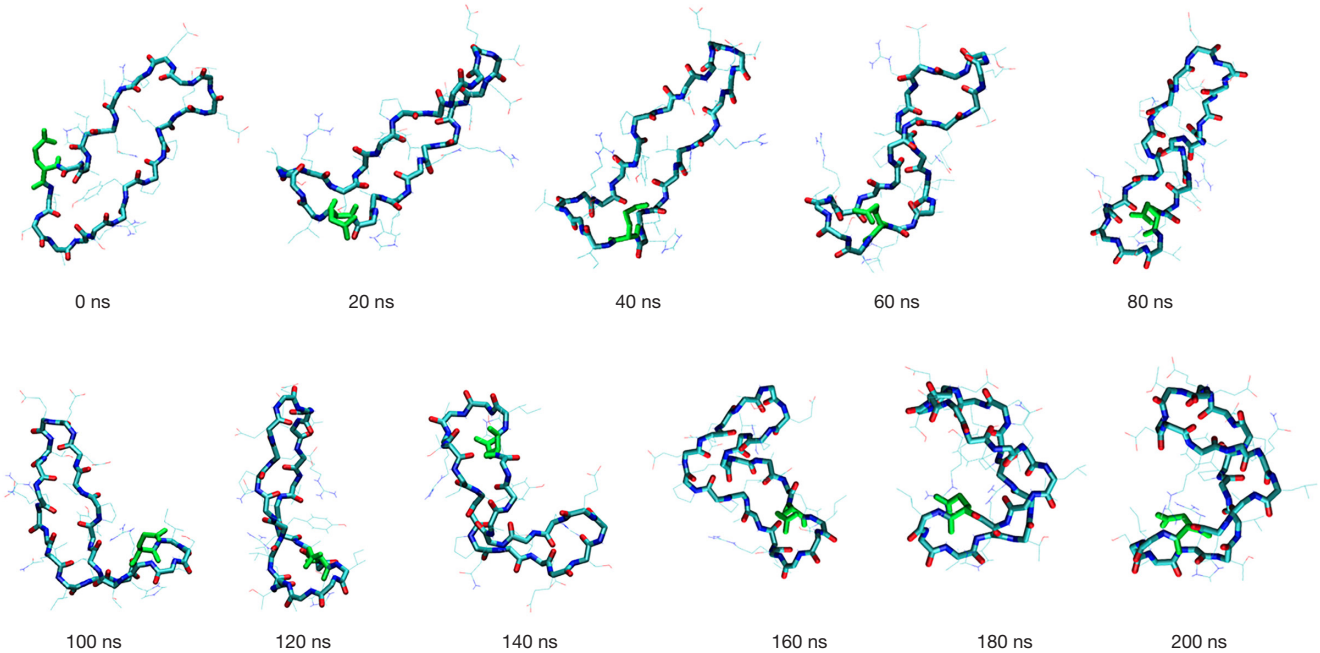

Figure 8 Representation of glycophorin A exon 3–4 structure at 20 ns time points starting from a cyclic conformation. N-terminus coloured in green, and peptide backbone represented in licorice format, and side chains in bond format, with colouring based on secondary structure. Images derived from VMD software (26).

methods without glycosylation, with the intention to add glycosylation once the base structure had been determined.

Although the programs used in this study were deemed suitable in the production of a protein backbone, both Robetta and FALCON have been developed and tested against ordered structures (12,14,57). Thus, challenges arise when attempting to model structures with higher levels of disorder. To overcome these challenges, MD simulations are used in adjunct to such predictive programs and allow for the "settling" of the predicted structures into energetically favourable and hence more probable states. The necessity of the use of MD can be seen when comparing the extracted Exon 3–4 region to its counterpart from the initial protein predictions (highlighted in red in *Figure 1*), as the β-strands are slightly altered. Furthermore, investigation needs to be undertaken to understand why the Exon 3–4 β-strands were retained whereas the other predicted β-strands disappeared over the 200 ns of MD simulation.

The CD spectropolarimetry results support the hypothesis that the monomeric non-glycosylated GPA-ECD has a high level of disorder, yet seem to contradict the presence of a stable β-hairpin within the GPA monomer. However, it should be noted that although CD spectropolarimetry, works well for proteins that are completely composed of α-helices and to a lesser extent completely composed of β-strands, it has been suggested that CD spectra is not effective at distinguishing mixed α-helical or β-sheet elements with particular difficulty identifying non-canonical β-strands (58). It is highly likely that the stable β-hairpin observed in our computer models was obscured in the CD experiments.

The combination of CD spectropolarimetry results, sequence based prediction, and the visual results of the MD simulations supports the theory that the monomeric extracellular region of GPA has a high level of disorder. As IDPs are able to conform into different 3D-structures, allowing them to interact with a variety of different proteins (51), the intrinsically disordered-like character of monomeric GPA-ECD could explain GPA's ability to interact with numerous RBC surface proteins. An important feature of GPA is its ability to form stable homodimers (59), but this is not inconsistent with the possibility that the monomeric GPA-ECD is IDP-like. IDP monomers are known to form stable homodimers through specific interactions (60,61). As GPA dimerisation is facilitated via transmembrane interactions (8), it is possible that membrane driven dimerisation stabilises the extracellular domain of the monomer.

© Annals of Blood. All rights reserved.

When attempting to identify antigenic sites, or develop antibodies against the GPA-ECD, the monomeric form of this molecule may not be a suitable representation of antigens in the context of an RBC. It is likely that antigen presentation in GPA-ECD is comprised of a combination of features including specific amino acid sequences, orientation, and presence of glycosylation. Consideration should also be made to the fact that the process of dimerisation may affect ECD antigen presentation. Additionally, GPA has direct interaction with, and exists in a complex alongside other RBC membrane proteins (62,63). These interactions may also play a role in the tertiary structural presentation of the ECD, as well as antigen presentation such as the case of the Wrb antigen (64). Nonetheless, the stable β-hairpin-like structure that was evident in the MD simulations may be used to improve antigen presentation of the exon 3–4 junction region for future laboratory-based experiments.

The models of a peptide structure of the β-hairpin indicates that the hairpin structure is stable even when removed from the influence of the neighbouring amino acids. This further implies that if the hairpin structure is synthetically produced as a peptide in the β-hairpin-like form it should retain its structure and thus be a useful embodiment of the native structure.

The inability of the circular or extended peptide to fold into a β-hair pin might have occurred due to an activation energy barrier that restricts the "foldability" of this structure. This energy barrier may not be overcome within the time-scale of MD simulations attempted in the present study. Running the peptide simulation for a longer time-scale or altering the parameters of the simulation to overcome the activation energy barrier may allow for the folding of the peptide.

Another possibility is that the formation of a hairpin structure requires the stabilisation energy of other interactions within the extracellular domain, and hence the isolated linear and circular peptides were unable to fold into the β-hairpin-like state. This inability to form the correct secondary structure from a linear or circular peptide has implications in experiments using synthetic peptides as replicas for antigens such as in epitope mapping and biopanning.

Despite the inability of the modelled circular and linear peptides to form the β-hairpin structure, circular or linear peptides are deemed to be preferred starting structures. This is as neither requires a protein scaffold which would increase the ease of laboratory-based experiments for this antigenic region. Further testing may be performed utilising differing sequence lengths (both lengthening and shortening), to determine an optimum sequence for peptides that can better replicate the β-hairpin-like structure of the exon 3–4 junction region. In addition, the use of other peptide presentation systems needs to be investigated to re-create the putative native secondary structure.

The ultimate aim of this work is to understand the antigenicity associated with structural elements of GPA and in the future glycophorin B (GPB) and hybrid glycophorins. Although glycosylation might not play a significant role in protein structure, its presence is crucial in certain antigen presentation. Thus although some information can be gleaned regarding antigenicity from structural models alone, additional detail still needs to be added to these models to fully understand antigen presentation and antibody recognition. As GPA has such high levels of glycosylation the addition of glycosylation requires computing capabilities beyond what was available for this project. The specific effects of glycosylation on antibody recognition and antigen presentation, will only be determinable once an appropriate protein structure for GPA-ECD has been determined and specific glycosylation has been added to the proposed structure based on experimental findings. Future studies adding glycosylation to GPA would also need to account for the complexity associated with differing styles of glycosylation across different cell lines (65), particularly if GPA is expressed on alternative (non-erythrocyte) cell lines for experimental purposes.

Additional studies, both computation and experimental will be required to validate the predicted extracellular structure for GPA, as well as the production of GPA homodimers. Once it can be determined that this modelling process is able to produce suitably accurate extracellular representations *in silico*, further models can be produced. GPA could be used as a template for homology modelling of GPB and other hybrid glycophorins due to their high level of homology (4). The model produced could also act as a template for the development of a GPA-ECD homodimer model.

Conclusions

In conclusion, this study predicts that the monomeric extracellular domain of non-glycosylated GPA has a high level of disorder; with the exception of the antigenic region that spans the exon 3–4 junction (residues 64–85), and presents as two β-sheets flanking a loop in a hairpin-like structure. Although this structure is locally stable within a

© Annals of Blood. All rights reserved.

simulation of the entire GPA extracellular domain, when starting from a linear or circular peptide, it failed to fold into the β-hairpin structure during the time-scale used in the current MD simulation study. This suggests that stabilisation of the β-sheet-hairpin secondary structure depends upon interactions with other regions of the extracellular domain that are brought into proximity during folding of the tertiary structure. As such, laboratory based experiments to replicate this region will require careful consideration of both secondary and tertiary structure. Lastly, it was shown that after 100 ns MD simulation on the mutated hairpin peptide, the only single amino acid variation, p.Gly78Arg (the ERIK antigen of the MNS system) led to the loss of the β-hairpin secondary structure.

Acknowledgments

Computational resources were provided by the Research Computing Centre (RCC) at the University of Queensland (UQ). All simulations were performed on the Wiener (GPU) cluster. The authors would like to thank Assoc. Prof. Mehdi Mobli and Centre for Advanced Imaging (CAI) at UQ for their assistance with performing the Circular Dichroism experiment.

Funding: Funding for the project was provided from the ARC Training Centre for Biopharmaceutical Innovation (CBI) from the Australian Research Council (ARC), and the Australian Red Cross Lifeblood. Australian Governments fund Australian Red Cross Lifeblood for the provision of blood, blood products and services to the Australian community.

Footnote

Data Sharing Statement: Available at http://dx.doi.org/10.21037/aob-20-51

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/aob-20-51). Prof. RF serves as an unpaid editorial board member of *Annals of Blood*. The other authors have no conflicts of interest to declare.

Ethical Statement: the authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Poole J, Daniels G. Blood Group Antibodies and Their Significance in Transfusion Medicine. Transfus Med Rev 2007;21:58-71.
2. Heathcote DJ, Carroll TE, Flower RL. Sixty Years of Antibodies to MNS System Hybrid Glycophorins: What Have We Learned? Transfus Med Rev 2011;25:111-24.
3. Pisano A, Redmond JW, Williams KL, et al. Glycosylation sites identified by solid-phase Edman degradation: O-linked glycosylation motifs on human glycophorin A. Glycobiology 1993;3:429-35.
4. Daniels G. MNS Blood Group System. In: Daniels G. editor. Human Blood Groups. Victoria: Blackwell Science Asia Pty; 2007.
5. Chang VT, Crispin M, Aricescu AR, et al. Glycoprotein structural genomics: solving the glycosylation problem. Structure 2007;15:267-73.
6. Lee JE, Fusco ML, Saphire EO. An efficient platform for screening expression and crystallization of glycoproteins produced in human cells. Nat Protoc 2009;4:592.
7. Stura EA, Nemerow GR, Wilson IA. Strategies in the crystallization of glycoproteins and protein complexes. J Cryst Growth 1992;122:273-85.
8. MacKenzie KR, Prestegard JH, Engelman DM. A Transmembrane Helix Dimer: Structure and Implications. Science 1997;276:131-3.
9. Schulte TH, Marchesi VT. Conformation of Human Erythrocyte Glycophorin A and Its Constituent Peptides. Biochemistry 1979;18:275-80.
10. Dill K, Hu S, Berman E, et al. One- and two-dimensional NMR studies of the N-terminal portion of glycophorin A at 11.7 Tesla. J Protein Chem 1990;9:129-36.
11. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506-15.
12. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32:W526-31.

© Annals of Blood. All rights reserved.

13. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40.
14. Wang C, Zhang H, Zheng WM, et al. FALCON@home: a high-throughput protein structure prediction server based on remote homologue recognition. Bioinformatics 2016;32:462-4.
15. Ekman S, Flower R, Hyland C, et al. In silico model for Glycophorin A (GPA) structure. Pathology 2019;51:S123.
16. DeLano WL. Pymol: An open-source moleuclar graphics tool. CCP4 Newsletter On Protein Crystallography 2002;40:82-92.
17. Christen M, Hünenberger PH, Bakowies D, et al. The GROMOS software for biomolecular simulation: GROMOS05. J Comput Chem 2005;26:1719-51.
18. Schymkowitz J, Borg J, Stricher F, et al. The FoldX web server: an online force field. Nucleic Acids Res 2005;33:W382-8.
19. Abraham MJ, Murtola T, Schulz R, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1-2:19-25.
20. Oostenbrink C, Villa A, Mark AE, et al. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 2004;25:1656-76.
21. Schmid N, Eichenberger AP, Choutko A, et al. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. Eur Biophys J 2011;40:843.
22. Berendsen HJC, Postma JPM, van Gunsteren WF, et al. Interaction Models for Water in Relation to Protein Hydration. In: Pullman B, editor. Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981. Dordrecht: Springer Netherlands; 1981:331-42.
23. Berendsen HJC, Postma JPM, Gunsteren WFv, et al. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684-90.
24. van Gunsteren WF, Billeter SR, Eising AA, et al. Biomolecular Simulation: The GROMOS96 Manual and User Guide. Biomos; Zürich; 1996.
25. Hess B, Bekker H, Berendsen HJC, et al. LINCS: A linear constraint solver for molecular simulations. J Comput Chem 1997;18:1463-72.
26. Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comput Chem 1992;13:952-62.
27. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33-8, 27-8.
28. Daura X, van Gunsteren WF, Mark AE. Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. Proteins 1999;34:269-80.
29. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215:403-10.
30. Haddad Y, Adam V, Heger Z. Ten quick tips for homology modeling of high-resolution protein 3D structures. PLoS Comput Biol 2020;16:e1007449.
31. Vyas VK, Ukawala RD, Ghate M, et al. Homology modeling a fast tool for drug discovery: current perspectives. Indian J Pharm Sci 2012;74:1-17.
32. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35:3375-82.
33. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 2004;101:7594-9.
34. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846-56.
35. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443-53.
36. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195-7.
37. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192-201.
38. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302-9.
39. Feig M, Rotkiewicz P, Kolinski A, et al. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86-97.
40. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001-14.
41. Grant A, Lee D, Orengo C. Progress towards mapping the universe of protein folds. Genome Biol 2004;5:107.
42. Schaeffer RD, Daggett V. Protein folds and protein folding. Protein Eng Des Sel 2011;24:11-9.
43. Bradley P, Misura KMS, Baker D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. Science 2005;309:1868-71.
44. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system.

© Annals of Blood. All rights reserved.

Proteins 2016;84:15-9.
45. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195-202.
46. Bywater RP. Protein folding: a problem with multiple solutions. J Biomol Struct Dyn 2013;31:351-62.
47. Bywater RP. Prediction of protein structural features from sequence data based on Shannon entropy and Kolmogorov complexity. PLoS One 2015;10:e0119306.
48. Drabik P, Liwo A, Czaplewski C, et al. The investigation of the effects of counterions in protein dynamics simulations. Protein Eng 2001;14:747-52.
49. Oldfield CJ, Dunker AK. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annu Rev Biochem 2014;83:553-84.
50. Uversky VN. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Front Phys 2019. doi: 10.3389/fphy.2019.00010.
51. Dyson HJ. Making Sense of Intrinsically Disordered Proteins. Biophysical journal 2016;110:1013-6.
52. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 2015;31:857-63.
53. Schneider JP, Kelly JW. Templates That Induce .alpha.-Helical, .beta.-Sheet, and Loop Conformations. Chem Rev 1995;95:2169-87.
54. Shental-Bechor D, Levy Y. Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. Proc Natl Acad Sci U S A 2008;105:8256-61.
55. Mitra N, Sinha S, Ramya TNC, et al. N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. Trends Biochem Sci 2006;31:156-63.
56. Parodi AJ. Role of N-oligosaccharide endoplasmic reticulum processing reactions in glycoprotein folding and degradation. Biochem J 2000;348 Pt 1:1-13.
57. Protein Structure Prediction Center. The Critical Assessment of protein Structure Prediction (CASP). University of California. 2020. Available online: https://predictioncenter.org/index.cgi
58. Khrapunov S. Circular dichroism spectroscopy has intrinsic limitations for protein secondary structure analysis. Anal Biochem 2009;389:174-6.
59. Gerber D, Shai Y. In Vivo Detection of Hetero-association of Glycophorin-A and Its Mutants within the Membrane. J Biol Chem 2001;276:31229-32.
60. Sigalov AB. Unusual biophysics of immune signaling-related intrinsically disordered proteins. Self Nonself 2010;1:271-81.
61. Danielsson J, Liljedahl L, Bárány-Wallje E, et al. The intrinsically disordered RNR inhibitor Sml1 is a dynamic dimer. Biochemistry 2008;47:13428-37.
62. Williamson RC, Toye AM. Glycophorin A: Band 3 aid. Blood Cells Mol Dis 2008;41:35-43.
63. Mankelow TJ, Satchwell TJ, Burton NM. Refined views of multi-protein complexes in the erythrocyte membrane. Blood Cells Mol Dis 2012;49:1-10.
64. Poole J. Red cell antigens on band 3 and glycophorin A. Blood Rev 2000;14:31-43.
65. Croset A, Delafosse L, Gaudry JP, et al. Differences in the glycosylation of recombinant proteins expressed in HEK and CHO cells. J Biotechnol 2012;161:336-48.

doi: 10.21037/aob-20-51
Cite this article as: Ekman S, Flower R, Mahler S, Gould A, Barnard RT, Hyland C, Jones M, Malde AK, Bui XT. *In silico* molecular dynamics of human glycophorin A (GPA) extracellular structure. Ann Blood 2021;6:11.

© Annals of Blood. All rights reserved.

Supplementary

Table S1 Single amino acid changes in glycophorin A (GPA) exon 3-4 region, and associated side chain alterations

Amino acid change	MNS antigen	Side chain change
p.Ser66Tyr	MNS:12 or Vr+	Polar to hydrophobic
p.Pro73Ser	MNS:38 or Os(a+)	Hydrophobic to polar
p.Glu76Lys	MNS:16 or Ri(a+)	Negative charged to positive charged
p.Thr77Ile	MNS:14 or Mt(a+)	Polar to hydrophobic
p.Gly78Arg	MNS:37 or ERIK+	Small aliphatic side chain to large positively charged side chain
p.Arg80Ser	MNS:47 or SARA+	Large positive charge to small polar
p.Gln82Lys	MNS:-42,43 or MARS+	Polar uncharged to positive charged

© Annals of Blood. All rights reserved.
http://dx.doi.org/10.21037/aob-20-51

Table S2 Homologous sequences to glycophorin A (GPA), with associated PDB structures. % identity indicates level of homology to GPA sequence. Secondary structure of homologous sequence identified from the PDB structure

			19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
GPA			A	S	S	T	T	G	V	A	M	H	T	S	T	S	S	S	V	T	K	S	Y	I	S	S	Q	T	N	D
PDB-10B1	45% identity		A	–	–	–	–	–	V	–	M	–	–	+	–	+	S	S	+	+	–	S	Y	+						
	Loose helix	13	A	F	L	G	E	R	V	T	M	T	C	T	A	T	S	S	L	S	S	S	Y	L	33					
PDB-4NOA	45.5% identity								+	A	–	–	T	S	T	+	–	+	+	T	–	+	–	I	–	S	Q	T	–	D
	Helical						82		I	A	T	P	T	S	T	T	Y	T	L	T	A	T	P	I	N	S	Q	T	R	D

			40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
GPA			I	S	S	Q	T	N	D	T	H	K	R	D	T	Y	A	A	T	P	R	A	H	E	V	S	E	I	S	V	R	T	V	Y	P	P	E	E
PDB-3RCN	38.5% identity		I	–	S	+	–	–	D	+	–	+	R	–	–	+	A	A	–	P	R	–	–	–	+	S	E	+										
	Helical	453	I	W	S	E	H	L	D	S	P	R	R	V	Q	F	A	A	F	P	R	L	S	A	I	S	E	V	478								
PDB-5X1E chain F (Icm0)	38.5% identity												+	D	–	Y	A	–	T	–	–	–	–	–	+	–	+	–	–	+	–	T	–	Y	P	P	E	E
	Helical										722		K	D	K	Y	A	G	T	V	A	N	E	L	I	K	D	F	Q	I	A	T	S	Y	P	P	E	E	747

			58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87
GPA			R	A	H	E	V	S	E	I	S	V	R	T	V	Y	P	P	E	E	T	G	E	R	V	Q	L	A	H	H	F	S
PDB-1H5Q	38% identity		R	–	H	+	–	S	–	I	–	+	–	–	–	–	–	P	E	E	–	T	G	+	–	+	–	L				
	Helical	215	R	D	H	Q	A	S	N	I	P	L	N	R	F	A	Q	P	E	E	M	T	G	Q	A	I	L	L	240			
PDB-5F1P	50% identity																P	E	–	E	–	–	–	R	V	–	L	–	H	H	–	+
	Helical (198-199)													198		P	E	R	E	A	A	Y	R	V	M	L	P	H	H	L	T	213

(–) indicates no corresponding amino acid, (+) indicates amino acid with similar identity, identical amino acids are shown with their one letter identifier. Amino acid numbering (based on the PDB structures) flanks the homologous sequences.

© Annals of Blood. All rights reserved.

http://dx.doi.org/10.21037/aob-20-51

Table S3 Locations of β-sheets and α-helices for initial structures 1-6 before MD simulation, using ISBT amino acid numbering.

Structure #	β-sheet	β-sheet	β-sheet	β-sheet	β-sheet	β-sheet	α-helix	α-helix
Structure 1	26-27	37-41		63-65	68-71	80-85	21-23	
Structure 2	25-29	36-38			65-70	83-87		73-75
Structure 3	25-27				69-71	81-87	36-39	75-77
Structure 4	26-28				66-71	81-87		
Structure 5	26-28			61-63	67-71	82-85	32-35	
Structure 6	28-30	39-43	49-53		63-69	82-85		

© Annals of Blood. All rights reserved.
http://dx.doi.org/10.21037/aob-20-51

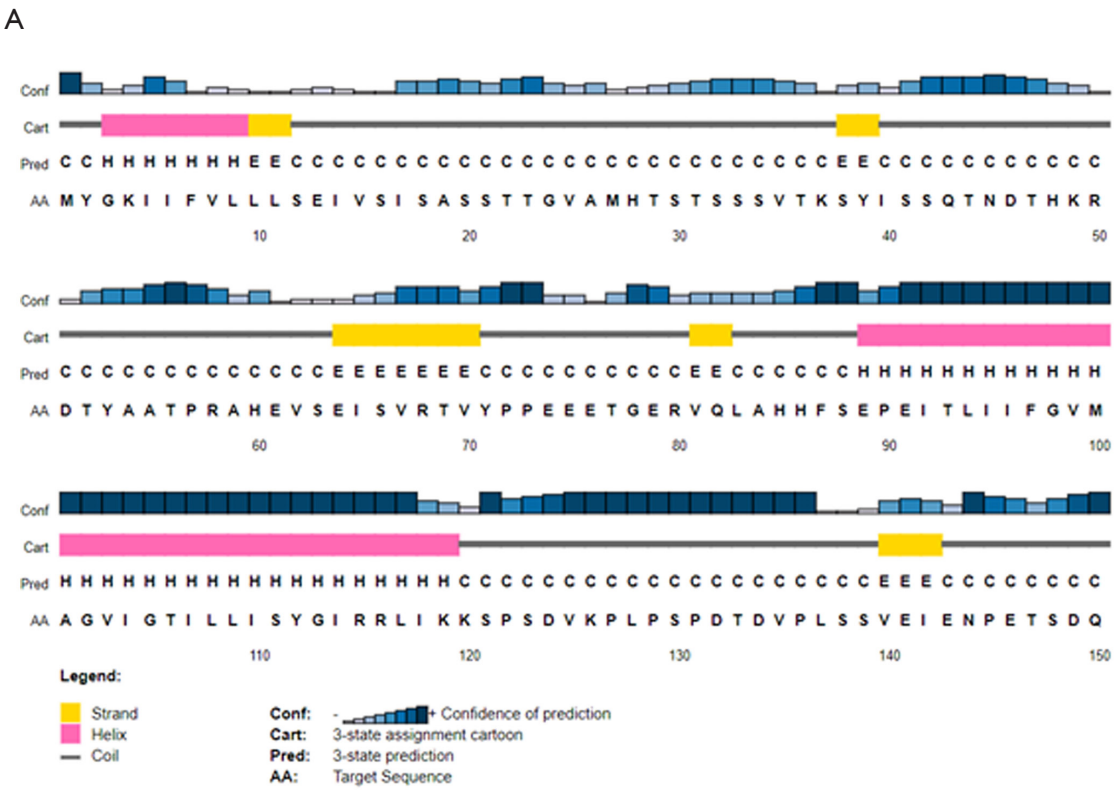

Figure S1 (A) Secondary structure prediction of GPA protein and (B) disorder of GPA protein as determined by PsiPred and DISOPRED (45,52).

© Annals of Blood. All rights reserved.
http://dx.doi.org/10.21037/aob-20-51

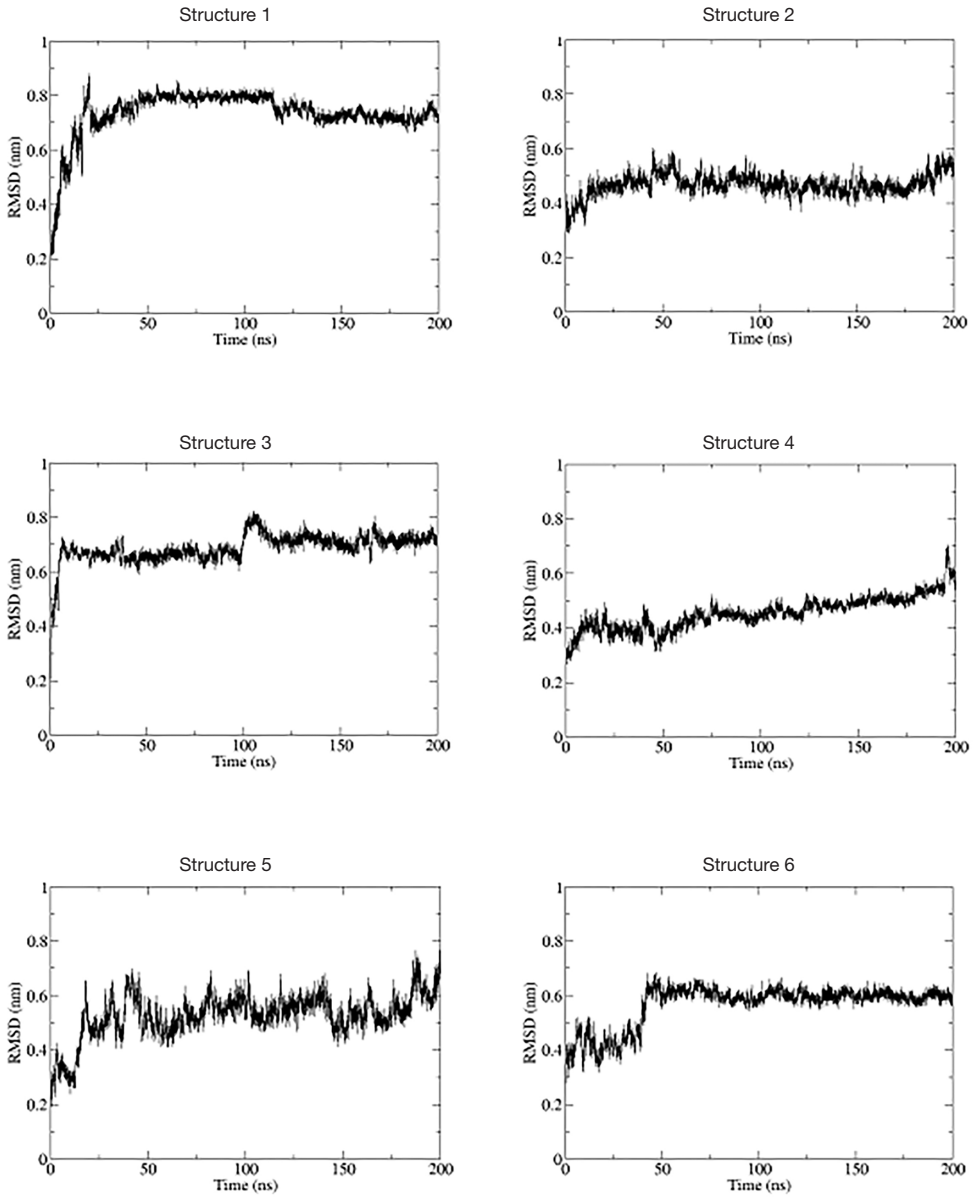

Figure S2 Root Mean Square Deviation (RMSD) analysis of protein backbone, based on protein backbone for 200 ns MD.

© Annals of Blood. All rights reserved.
http://dx.doi.org/10.21037/aob-20-51

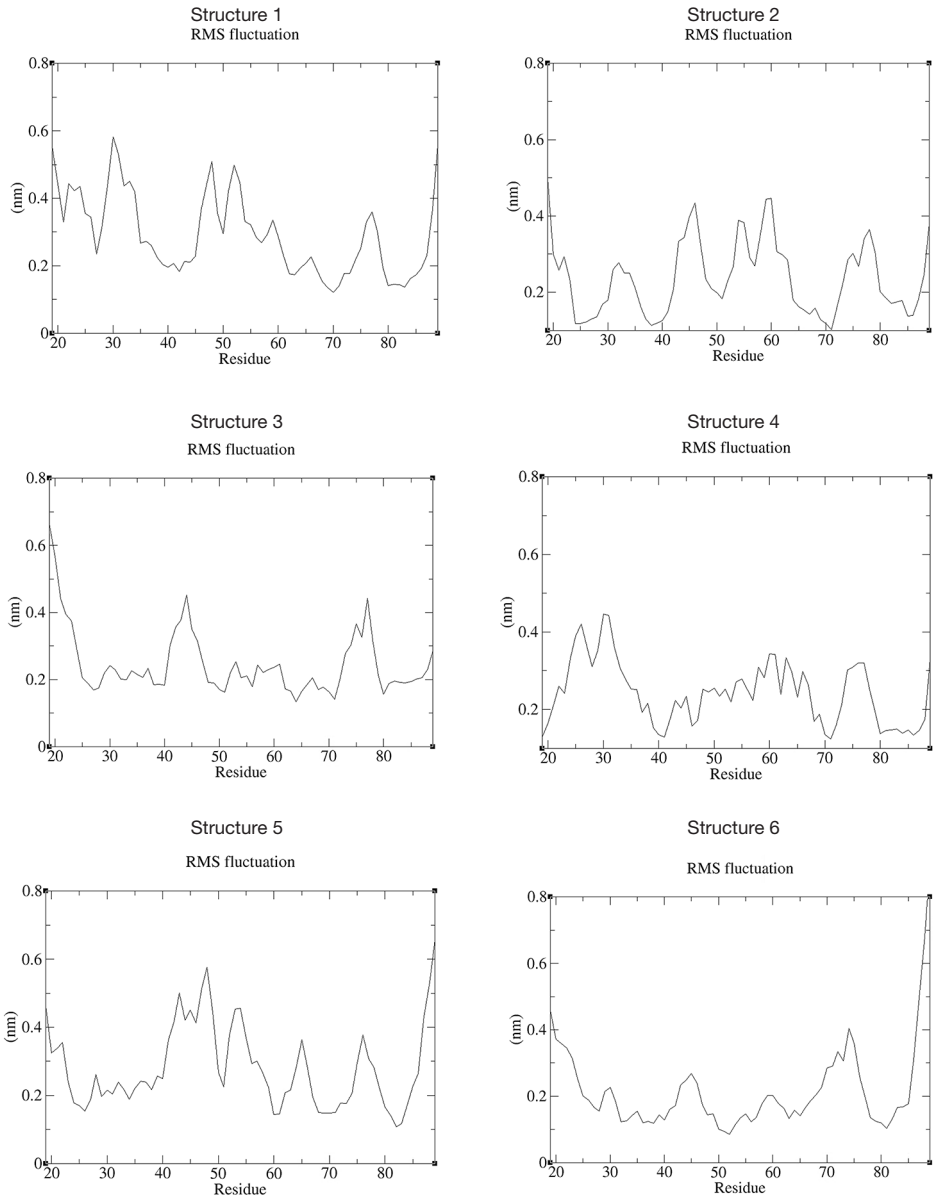

Figure S3 Root Mean Square Fluctuation (RMSF) analysis of protein backbone, based on protein α-carbons for 200 ns MD.

© Annals of Blood. All rights reserved.
http://dx.doi.org/10.21037/aob-20-51

Table S4 Table of α-helix formation across 6 glycophorin A (GPA) structures, including time of helix formation over 200 ns MD simulation in explicit solvent, using ISBT amino acid numbering

Structure #	Amino acids involved in α-helix formation	Time of formation
Structure 1	α-helix: 22-25	130 ns onward
	3-10 helix: 33-37	24-48 ns
Structure 2	3-10 helix (intermittent): 59-61	16-30 ns
	3-10 helix (intermittent): 73-76	120-180 ns
Structure 3	α-helix/3-10 helix: 36-39	0-120 ns
Structure 4	3-10 helix (intermittent): 30-32	90-180 ns
	3-10 helix (intermittent): 47-49	70-190 ns
	3-10 helix (intermittent): 59-61	0-200 ns
Structure 5	α-helix: 32-37	0-200 ns
	3-10 helix (intermittent): 47-49	0-200 ns
	α-helix: 47-49	120-145 ns
Structure 6	3-10 helix (intermittent): 75-78	0-70 ns

© Annals of Blood. All rights reserved.
http://dx.doi.org/10.21037/aob-20-51