Molecular modelling of Androglobin

Ngaahule Jerry Junior Mukhathedzwa

A thesis submitted for the degree of Master of Science (by Dissertation) in Biological Sciences

Department of Biological sciences

University of Essex

Date of submission September 2020

Abstract

Experimental methods fail to express in vitro the full length of Androglobin (Adgb). Predominantly due to its size, the chimeric protein spans 1667 residues. Distinct from other members of this protein superfamily because of its circularly permutated heme domain, and N- terminal calpain domain, and an IQ binding motif. Here we will explore the molecular structure of Adgb.

This will be achieved by first exposing the sequence to disorder prediction algorithms with the aims of identifying intrinsically disordered regions (IDRs), that may not only add noise when subjecting the sequence to fold recognition techniques. These IDRs may also be in close proximity to key molecular recognition features (MoRFs).

Secondary structure prediction will not only be used to establish a common frame of reference between disorder prediction and fold recognition result, but it will also be integral to identifying the sites for potential MoRFs. Fold recognition techniques, when exposed to a query sequence, will produce tertiary structures homologous to the query sequence with known molecular functions. The structural information produced will be essential in building a complete computational model of Adgb, in hopes of guiding efforts to express *in vitro* the full length of the sequence.

Disorder prediction revealed 5 ordered regions, of which the pre-heme sequence harboured 3; of these 2 displaying homology to calpain domains IIa and III through profile-profile alignments using Clustal Omega (Sievers *et al.*, 2011) and a consensus formed by fold recognition techniques. Profile-profile alignments have also yielded 2 potential EF hand motifs, required for the calpain domain's proteolytic activity. The presence of an even number of EF hands suggests that Adgb may indeed function as a monomer. In the post-heme region of the sequence 2 ordered regions have been found.

Acknowledgements

Thank you, Prof. Reynolds, for being so understanding over the course of my MSD. Beyond just being extremely knowledgeable in your field, you have believed in me in the moments where I honestly didn't believe in myself. I honestly did not think I would finish this MSD. There are things that you have just done for my best interest, before I even knew it was in my best interest, and for that I cannot be thankful enough.

Thank you, Dr. Reeder. The time you have spent explaining some of the literature to me, has formed the basis for my understanding of the proteins concerned and their molecular functions.

Thank you, Emma Revill. If it was not because of you, honestly, I would have had a much harder time. From start to finish, you have facilitated so many things. For making sure I was getting the support I needed, so I could submit a body this work that stands as testament to my ability

Thank you, Mum, for being my personal ATM machine and being an emotional punching when I was stressed.

To all of these people and more, I cannot say thank you enough.

Chapter 1: Introduction

Androglobin

Androglobin (Adgb) is the 5th mammalian globin to be discovered, in addition to being the largest globin at 1667 residues in length. Adgb is roughly 10 times larger than the average globin, which typically has a length between 140 and 190 residues. While expressed in the heart, lungs, and brain, the site of primary expression for Adgb is in the testes, where it has been shown to play a vital role in post meiotic spermatogenesis, and remains integral to male reproduction (Hoogewijs *et al.*, 2012; Huang *et al.*, 2014).

The ability to facilitate gaseous exchange was once widely accepted to be the sole function of a globin. This was largely because common to all globins is a non-protein, gaseous ligand binding heme (iron protoporphyrin IX in Fig. 1.1d). Arguably this feature would have continued to define globins as a whole a stringent depiction of their true cellular functions.

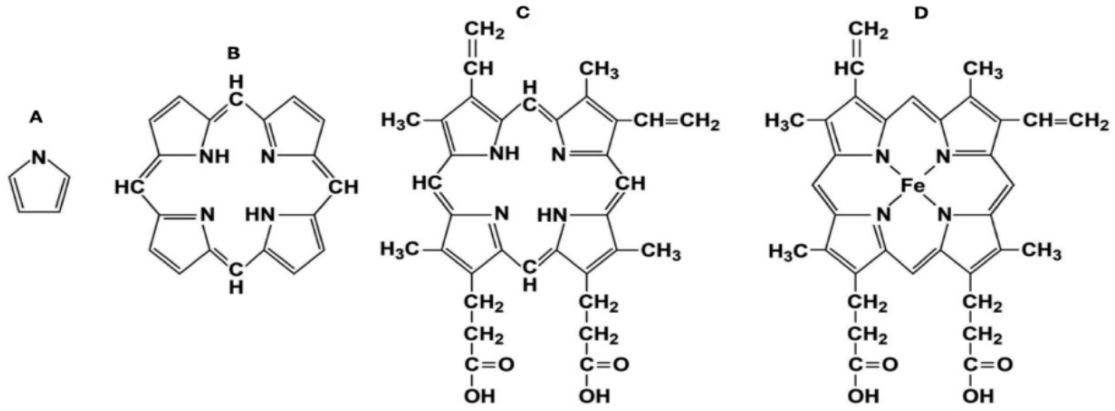

Fig. 1.1: Structure of Heme Porphyrin structures. (A) pyrrole ring, (B) porphyrin ring, (C) protoporphyrin IX, and (D) Heme, iron protoporphyrin IX (Source: Schmidt *et al.*, 2012).

The heme is a planar tetrapyrrolic protoporphyrin ring shown in Fig. 1.1d capable of chelating 4 out of 6 coordination sites on a single Fe^{2+} ion. Each one of the 4 dative bonds between the Fe^{2+} ion and the protoporphyrin ring, is formed by a nitrogen atom from a pyrrole ring in Fig.1.1a donating a lone pair of elections. The 5th binding site is datively bonded by an imidazole nitrogen from the 'proximal' (F8, i.e. the 8th residue on helix F) histidine side chain. This is one of the ways the globin protein incorporates the non-protein heme group into the protein. The 6th coordination site on Fe^{2+} in pentacoordinate globins, is able to reversibly bind preferably O_2, in addition to carbon monoxide (CO), nitric oxide (NO), and water. When oxidised to Fe^{3+} in the form of met-hemoglobin (met-Hb) it is instead able to bind cyanide, azide, NO, or water (Aronson *et al.*, 1994; Stryer, 1995; Nye and Lecomte, 2018).

When deoxygenated, hemoglobin (Hb) and myoglobin (Mb) exhibits a bent, domed geometry shown in Fig. 1.2a, where the Fe ion is shifted out of the plane were the protoporphyrin IX ring sits. Conversely, the oxygen bound conformation exhibits planar geometry, as shown in Fig. 1.2b. When the heme binds O_2 to the 6th coordination site of Fe, the heme is then stabilised and allows the Fe ion to sit on the same plane as the heme. This occurs at the non-binding oxygen atom on O_2, by a 'distal' (E7 i.e. 7th residue on helix E) histidine in the globin chain binding to the Fe-bound oxygen.

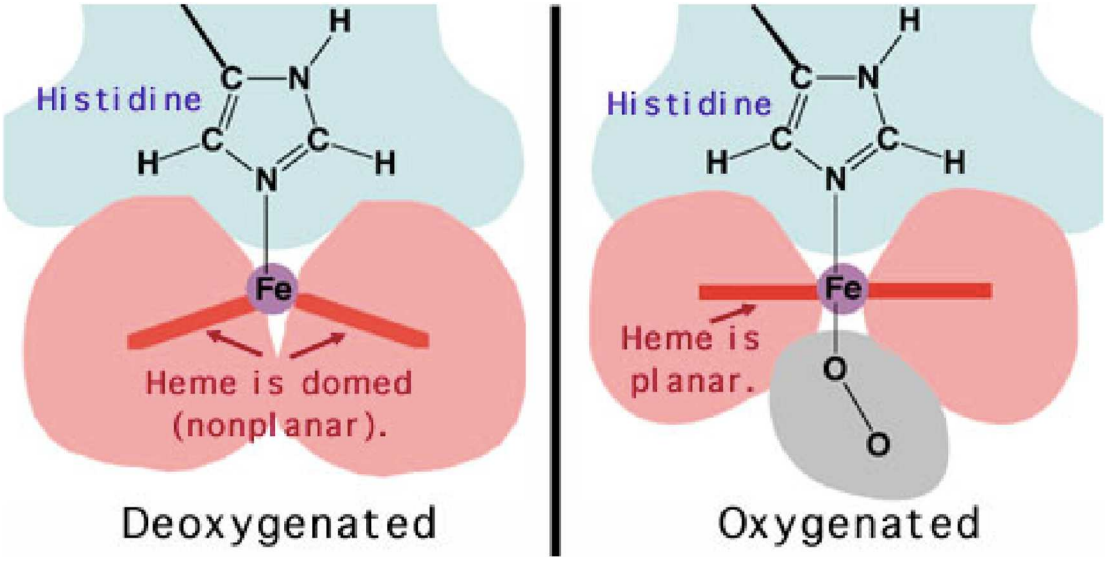

Fig. 1.2: Porphyrin structure geometry. (a) deoxygenated – nonplanar/domed (b) oxygenated – planar (Source: Traverso, 2004).

The presence of heme within the core of the protein that is capable of binding respiratory gases, means that the conclusion that globins are respiratory pigments is not completely unfounded (Dickerson and Geis, 1984). It remains true that the nature of the ligand, and the reason for their ligation, is currently the subject of an ever-evolving conversation.

The heme is encased within a 'globin-fold' which all globins share. The fold is comprised of a set of 7 helices termed A, B, C, D, E, F, G, and H, with the short D helix present in only some globins. In Adgb, the 'globin-fold' is located within the centre of the sequence and has been modelled as 8 alpha helices that aggregate to form the heme structure (Aronson *et al.*, 1994). The sequence structure in the revised model of Adgb only differs when compared to other globins, primarily in that the permutation of the heme domain results in helices D-H now being the first to be encoded at the N-terminus of the heme domain in the sequence for Adgb (F782 – D890). This is then followed by

helices A-C (D935 – Q980), with the interface between helices H and A accommodating an IQ calmodulin binding domain, a section of the protein 33 amino acids long (Fig. 1.3b). Adgb preserves its structural integrity, retaining its ability to bind heme in the hydrophobic heme pocket, despite the large number of structural changes that have occurred in the structural configuration of the protein. This is the first instance of circular permutation in a hemoglobin, that has been observed to be naturally occurring (Hoogewijs *et al.*, 2012).

In most globins, three very specific residues integral to stabilisation seem to be meticulously conserved. First is the 'proximal' histidine on the F alpha helix (F8), which functions to bind the protein to the non-protein heme bound Fe ion. This interaction effectively indirectly anchors the heme to the globin protein. The second conserved residue is the (E7) ligand-binding 'distal' glutamine in Adgb, but histidine in other Hbs, on the E alpha helix. This residue functions to stabilise heme bound oxygen with a hydrogen bond to the second oxygen atom on the heme bound O_2. Finally, the most conserved residue among all the globins is the interhelical CD1 Phenylalanine residue. This functions as an anchorage point to bind the heme group directly to the heme pocket of the protein, as opposed to indirectly through the Fe ion as is the case for residue F8 (Hargrove *et al.*, 1994; Roesner *et al.*, 2005). It is specifically mutations in this CD1 residue alone that seem to give rise to a number of proteinopathies. These often results in the formation of Heinz bodies, haemolytic anaemias, cyanosis, all in addition to decreasing oxygen affinity (Sonati *et al.*, 2006; Grimholt *et al.*, 2018). However, despite it being possible to generate a stable model of the heme domain of Adgb computationally in our laboratory, the current literature states it is not possible to express the protein by recombinant means *in vitro* (Bracke *et al*, 2018).

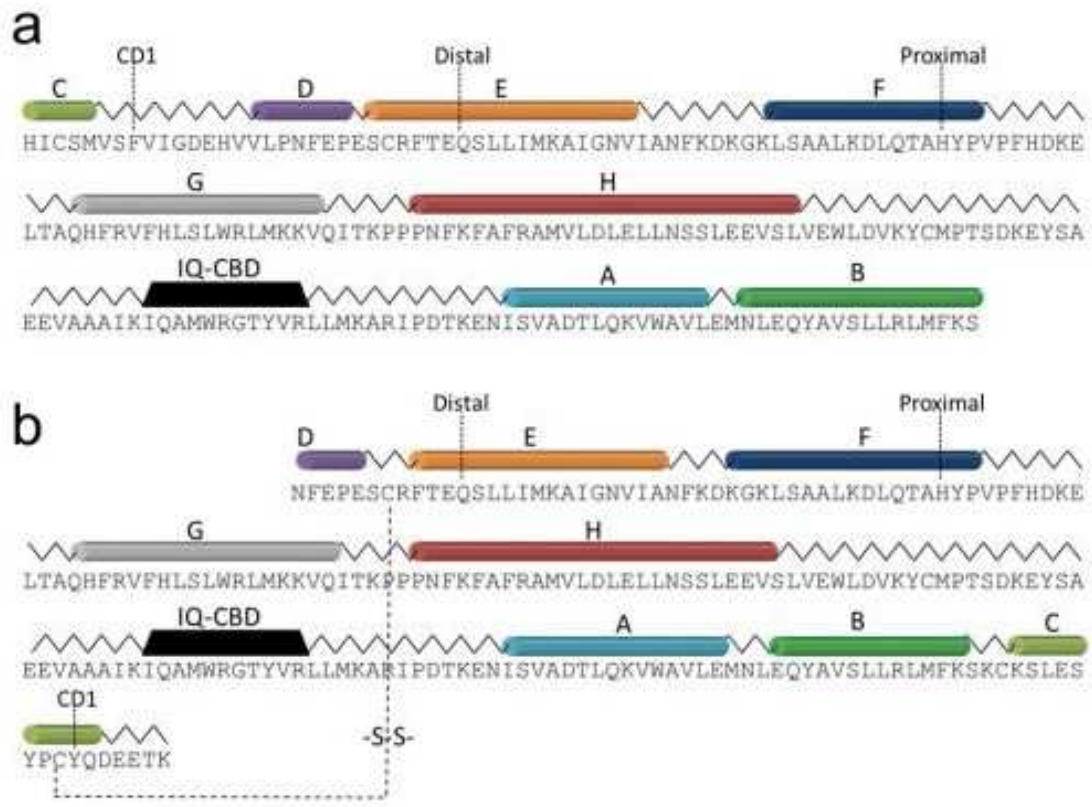

Fig. 1.3: Androglobin heme domain structural alignment. (a) Original structural alignment as shown in (Hoogewijs *et al.*, 2012) with domains C-H at N-terminus and A-B domain at C-terminus. IQ calmodulin binding domain (IQ-CBD) at residues 762-966. (b) Alternate alignment for heme domain remains practically identical save for revision of C alpha helix (to residues 780-984). Depicted within this Fig. the CD1 residue resides on the C-terminus, instead of the N-terminus. This provides stability to the proteins by means of a potential disulphide bond (-S-S-) in a similar fashion to that which is observed within the CD "loop" of human Neuroglobin (Ngb). This new revised stable sequence expressed the heme domain now known as Adgb-GD and has been expressed by recombinant means (Diagram kindly provided by B Reeder).

With this in mind, the substitution of Phe769 for Tyr975 in the revision of residue acting as the Adgb CD1 anchorage point (Reynolds, CA; personal communication), has

indicated that the new Reynolds lab alignment has seemingly provided newly found stability when expressing the heme domain of Adgb with this conserved substitution. This suggests that Adgb displays a preference of aromatic residue at this anchorage site in this globin, as the revised Adgb model retains an aromatic residue at the CD1 anchorage point.

The discrepancies between the two alignments in Fig. 1.3a and 1.3b can be attributed largely to the simple fact that the original Adgb alignment in Fig. 1.3a was created using only 34 know Adgb sequences to form a consensus sequence, using less sophisticated means of multiple sequence alignment. Since then, the number of known Adgb sequences has risen to nearly 600 sequences, which when coupled with a sophisticated profile alignment approach based on multiple pair-wise alignments (Lock *et al.*, 2014; Taddese *et al.*, 2014) has given a new more reliable multiple sequence alignment in Fig. 1.3b.

In short, the movement of alpha helix C from the N-terminus of the heme domain as shown in Fig. 1.1a to the C-terminus of the heme domain, results in helix D becoming the first helix from the heme domain to be encoded in the Adgb sequence at the N-terminus as shown in Fig. 1.3b. The shift yields a new CD1 anchorage point from that suggested by Hoogewijs *et al.*, 2012 placing the new CD1 anchorage point at Tyr975.

Once aggregated, and the heme group is secured within the globin fold, the globin may dimerise e.g. clam Hb forms and functions as a homodimer. Some are able to form even bigger tetrameric complexes e.g. Hemoglobin (Hb). Globins are also capable of functioning as a monomer as is the case for myoglobin. However, it remains unclear whether or not Adgb functions as a part of a complex or as a monomer (Aronson *et al.*, 1994).

The evidence seems to indicate that Adgb cannot function as a O_2 co-transport protein, despite all the similarities shared between Adgb and other globins. This inability to function as a co-transport protein is most apparent from the tendency for Adgb to rapidly undergo auto-oxidation from Fe^{2+} to Fe^{3+}. Upon oxidation to Fe^{3+} the ion is no longer able to reversibly bind O_2, which when coupled with the fact that O_2 levels in the testes are around 10-15 mmHg, makes it increasingly more unlikely that oxygen transportation is one the molecular functions of Adgb. This is mainly because myoglobin (Mb) would be half saturated (P_{50}) at ~2.8 mmHg, and thus would exhibit high affinity for O_2 at the same concentrations found in the testes. This in turn means that O_2 would more readily bind Mb in the same site where Adgb is primarily expressed (Endeward *et al.*, 2010).

Furthermore, Hb typically exhibits a P_{50} of ~27 mmHg. This means it would display a high affinity to O_2 in most tissues than Adgb, as venal concentrations of O_2 are typically ~45 mmHg (Rumi *et al.*, 2009). Ultimately, there also seems to be no evidence for the upregulation of Adgb mRNA transcription in a number of hypoxic tissue cell lines including human testicular germ cell tumour, which would have been an appropriate homeostatic response if Adgb primarily functioned as a respiratory pigment (Domino *et al.*, 1983; Reyes *et al.*, 2012).

The reality is, globins only really share a meticulously conserved heme domain. To conclude they all transport oxygen is just not only incorrect for Adgb, but a number of other globins. With co-transport discounted as one of the molecular functions Adgb may undertake, and the ability to catalyse reactions that convert NO and O_2 into nitrate (NO_3^-) being common to most, if not all, hemoprotein species then it stands to reason that the true physiological function Adgb may be rooted in NO homeostasis. In fact, the

ability to mediate NO metabolism seems to have been acquired by globins prior to the ability to co-transport O_2. This is an ability that is fully exploited by the nitrogen fixating leghaemoglobin and is shown in Eq. 1.1 (Kuhn *et al.*, 2017).

$$HbFe^{2+}O_2 + NO \rightarrow HbFe^{3+} + NO_3^-$$

Eq. (1.1) Dioxygenation of NO

Hemoglobin (Hb) and Myoglobin (Mb)

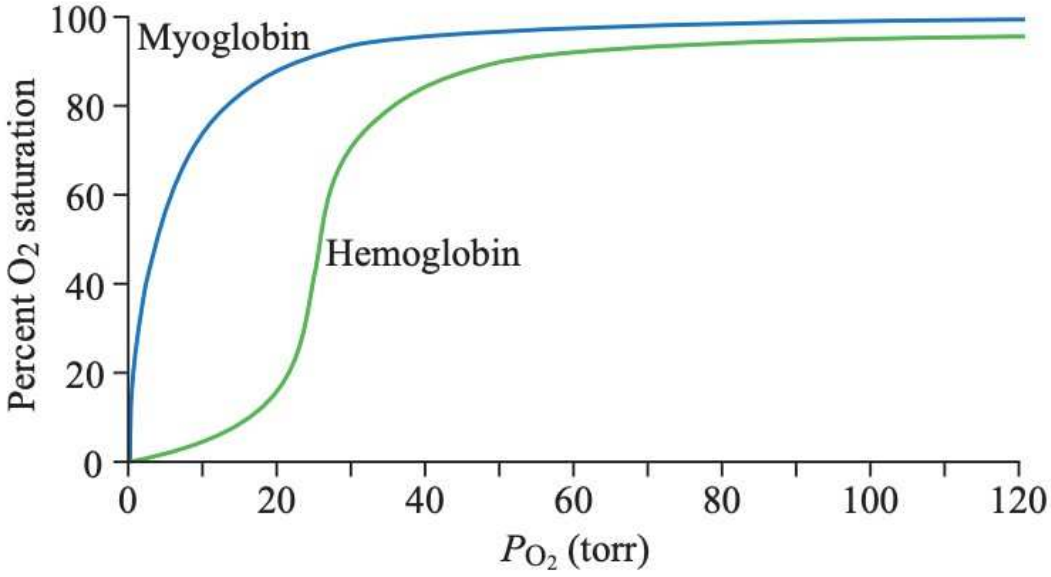

Fig. 1.4: Hemoglobin and Myoglobin O2 dissociation curve (Source: Ordway and Garry, 2004)

Hb functions as the tetrameric 64-kDa respiratory pigment. Comprised of a_2b_2 monomers all containing a single heme group, each tetramer is capable of cooperatively binding up to four O_2 molecules in total. The cooperative binding in Hb arises from its two-state model; a high O_2 affinity relaxed or R-state and a low O_2 affinity or T-state. Hb is deoxygenated in its T-state, resulting in the first molecule of O_2 to bind with the least affinity whilst in the lungs where O_2 concentrations are high. Despite the low affinity of the T-state, Hb will bind 2-3 molecules of O_2 in its T-state whilst the blood circulates the pulmonary circuit. When Hb is over half-saturated with O_2 it will undergo a conformational change into its high affinity R-state. This makes it easier for any additional O_2 to bind until Hb becomes fully saturated. Hb produces a sigmoidal cooperative binding curve for O_2 (green, Fig. 1.4). This cooperative binding model, given the abundance of Hb per erythrocyte in addition to the abundance of erythrocytes in circulation, results in the blood being able to absorb up to 50 times the amount of O_2 possible through its dissolution within the blood plasma alone (Helms and Kim-Shapiro, 2013).

Like Hb, Mb functions to transport O_2 as a gaseous exchange surface, however, unlike Hb it is only expressed in the cytoplasm of cardiac myocytes and striated oxidative skeletal muscle cells. Furthermore, Mb exists as a monomer, with a single O_2 binding coordination site, and produces a hyperbolic O_2 binding curve (blue, Fig 1.4). This curve shows O_2 affinity to have a positive causal relation to O_2 concentration. It primarily functions to facilitate the diffusion of O_2 from erythrocytes into the mitochondria, for use in oxidative phosphorylation, as well as functioning as an O_2 store for release during hypoxic/anaerobic conditions. This is especially the case in marine mammals and birds, which exhibit 10 – 30 times the amount of muscle Mb observed in predominantly

terrestrial animals. This is because marine mammals and birds are exposed to prolonged periods of oxygen deprivation and have evolved accordingly, in order to mitigate the effects of subaquatic hunting on muscle performance and fatigue.

The higher affinity for O_2 is displayed by myoglobin at a lower partial pressure of O_2 (pO_2) aids its ability to act as an intracellular oxygen store, by helping myoglobin stay almost fully saturated at P_{50} for Hb. The capacity for Mb to act as an oxygen store indirectly confers the capability for Mb to also act as a buffer for pO_2 during vigorous exercise, by resisting the sudden influx of O_2, thus ensuring cytosolic pO_2 remains constant. Desaturation of Mb when it is close to the plasma membrane results in increased uptake of O_2, which in turn diffuses into the mitochondria where the O_2 mediates oxidative phosphorylation. The facilitated diffusion by Mb complements and occurs concurrently with the simple diffusion of dissolved O_2 across the plasma membrane (Ordway and Garry, 2004).

Neuroglobin (Ngb) and Cytoglobin (Cygb)

The detection of neuroglobin (Ngb) and cytoglobin (Cygb) has radically redefined the nature and function of a globin protein. If globins were still considered purely respiratory pigments, the strong expression of Ngb and Cygb in the cytoplasm of the cells in the nervous system and hepatic tissues respectively, would have to be explained in terms of O_2 consumption. This is something which for Ngb is intuitive. This is as its

expression within in the brain, which accounts for only ~2% of human body weight but consumes ~20% of circulatory O_z even at rest. Ngb could feasibly facilitate diffusion of O_2 to neural tissues. The presence of Cygb within the cytoplasm of hepatic stellate cells (HSC) lacks explanation until the moment of their transdifferentiation, under conditions of oxidative stress or liver injury, into a type of myofibroblast (activated HSC). In this form they are highly fibrogenic and contractile, both of which are high energy consuming cellular activities. In these processes, energy is needed to form cross-links in the collagen molecules and contractions; O_2 drives the oxidative phosphorylation in order to produce that energy (Gandhi, 2012; Yoshizato *et al.*, 2016).

Since their discovery, these two new classes of globin have become synonymous with predominantly protecting sensitive tissues from chemical stresses, though Ngb has been linked to O_2 transport in invertebrates, as the viability for neurones in hypoxic conditions plummets in the absence of this neurologically associated globin. Though there is no definitive answer on their primary cellular roles, largely because the full range of their respective cellular functions are still not fully understood, they are known to be NADH reductases, and are capable of functioning as O_2 sensors (Pesce *et al.*, 2002; Gardner, 2005; Burmester and Hankeln, 2008).

Ngb is expressed predominantly within the tissues of the central nervous system and peripheral nervous system (CNS and PNS), in addition to a number of endocrine tissues. Concentrations Ngb expressed in a mouse brain are as low as <1 µmol l^{-1} and its structure is most analogous to Mb, in that it functions as a 16-kDa monomer approximately 151 residues long, in addition to exhibiting a typical globin fold. Much like Mb, it expressed within the cell cytoplasm, however, it has been found to also associate with the mitochondria at low concentration (~10%). Ngb can be expressed up to 100

times higher than the concentrations observed within the brain, much in the same way Mb is expressed in huge quantities within skeletal and cardiac muscle (Schmidt *et al.*, 2003).

The expression of Ngb within the CNS seems to be loosely confined to cells with a high metabolic rate, resulting in a high rate of O_2 consumption. Despite binding O_2 with affinity wavering between Mb and Hb, only a small proportion of human Ngb is saturated with O_2 at any given time under physiological condition. This is due largely to Ngb being a hexacoordinate globin, chelated by 4 pyrrole nitrogen atoms and 2 imidazole nitrogen atoms. In Ngb, the 'distal' E7 histidine binds the 6^{th} coordination site on Fe^{2+} directly in its deoxygenated state. This is the same E7 residue which in Hb and Mb only ever binds 6^{th} coordinate site on the Fe^{2+} when O_2 is already bound, therefore it binds the 6^{th} site indirectly.

Ngb exists in an equilibrium that favours its hexacoordinate conformation. As a consequence, any gaseous ligand i.e. O_2, CO, or NO, will have to outcompete the E7 histidine for the 6^{th} coordination site, resulting in it being very unlikely that Ngb functions as a respiratory protein. Competition for a coordination site on Fe^{2+} is only possible because a small portion of Ngb exists in this equilibrium as a pentacoordinate hemoprotein i.e. when the Fe^{2+} ion is chelated by 4 pyrrole nitrogen atoms and 1 imidazole nitrogen atom, leaving the 6^{th} coordination site able to bind gaseous ligands. (Burmester and Hankeln, 2009; Guidolin *et al.*, 2016)

In addition, Ngb is noticeably affected by fluctuations in pH, drastically altering the affinity with which O_2 will bind to Ngb. These changes in affinity may be described both in terms of a typical and a reversed Bohr shift, depending on temperature. Though its function is still far from being fully understood, we do know Ngb plays an active role

in providing protection to nerve cells against a number of pathologies that would damage the integrity of the cells, through scavenging reactive nitrogen species (RNS) and reactive oxygen species (ROS) to reduce the detrimental effects of toxic neural stimuli (Guidolin *et al.*, 2016).

Similarly to Ngb, Cygb's function is largely unknown, though it does seem to hold influence in O_2 homeostasis (Burmester *et al.*, 2002). Cygb shares a large number of similarities with other globins, exhibiting the typical 3-on-3 a helical globin fold. Though unlike Hb, Mb, and Ngb it is expressed in a wider array of cell types. The 190 long polypeptide is expressed within the cytoplasm of fibroblastic cells, oesophageal cells, smooth muscle cells, and melanocytes. Most notably Cygb is expressed in the hepatic stellate cells (HSCs), which are important regulators of repair and regeneration of a damaged liver. HSCs account for ~4% of the cell count within the liver and are vascularized by the hepatic portal vein, which carries 'hypoxic blood'. The O_2 deprived condition of this blood supply have been identified as a factor that contributes to HSCs activation, as HSCs are known to be sensitive to pO_2 fluctuations (Yoshizato *et al.*, 2016). The expression of Cygb by the stellate cells in the liver lends itself to the use as a biological marker of viability of stellate cells, and overall liver health within various pathological states (Motoyama *et al.*, 2014). There is a great deal of evidence in existence linking Cygb to the mitigation of the effects of oxidative stress. Evidence suggests Cygb issues protection from tissue damage in hypoxic conditions, such as those displayed during episodes of ischaemic blockage, which if left untreated may lead to localised and/or delocalised myocardial atrophy (Fago *et al.*, 2004).

With both Ngb and Cygb joining the ranks of nonerythroid globins, they both function with greater capacity regulating processes outside O_2 facilitated diffusion. This

may suggest that a globin may not even need to bind O_2 for an extended period of time, if at all, but merely be capable of remaining sensitive to physiological levels of saturation, thus functioning as a biological sensor. However, it is more intrinsic to globins to engage in addition to NO homeostasis, by managing it's transportation and metabolism (Ascenzi *et al.*, 2014).

Leghemoglobin (Lb), Non-symbiotic plant Hbs (nsHbs), and Truncated Hbs(trHbs)

Coined leghemoglobin (Lb), symbiotic Hbs work in synergism with nitrogen fixing bacteria on the root nodules of leguminous plant. When observed in healthy nodules, leghemoglobin would have an equilibrium ~20% oxyhemoglobin and ~80% deoxyhemoglobin. Despite this distinction, Lb can also be found in the root nodules of some non-leguminous plants. These Hbs are the most analogous, in structure, to Mb as they are expressed as a 16 kDa monomeric Hb. They primarily function in facilitating the continual transport of oxygen to the bacteria, up regulated in hypoxic conditions, they also share similar redox molecular activity as those displayed by both Hb and Mb. Similarly too, Lb ceases to be able to act as a respiratory Hb upon oxidisation of Fe^{2+} to Fe^{3+} by the slightly acidic conditions of the soil (Becana and Moran, 1994).

However, Lb's affinity to O_2, when it's not in its ferric state, is much higher than that displayed by its vertebral homologue Hb. In addition to this, it also functions to

buffer the O_2 concentrations; if not, it would threaten to disable nitrogenase enzymes required for nitrogen fixation (Appleby *et al.*, 1983; Moreau *et al.*, 1996; Kundu *et al.*, 2003).

Non-symbiotic plant Hbs (nsHbs) are further divided into classes with variable affinities to O_2; class 1 (nsHbs-1s) possess the highest affinity to O_2 within non-symbiotic Hbs, whilst class 2 (nsHbs-2s) have a comparatively lower affinity. Upregulated in conditions of flooding and colder temperature respectively, they are expressed in most, if not all, plant proteomes (Parent *et al.*, 2007). When observed in Arabidopsis, Hb1 is expressed at a low level during rest but is then quickly expressed strongly under hypoxia, indicative of its high affinity for O_2. The expression of Hb 2 in cold temperatures suggests a possible role in mitigating the effects of cold stress (Vigeolas *et al.*, 2011).

Plants produce nsHbs-1 in particular, in various situations including osmotic stresses, cold stress, hypoxia and nutrient deprivation. Both class 1 and class 2 nsHbs have been shown to be able to metabolise NO. The lower affinity to O_2 in nsHbs-2 may be indicative of the preference it has for the role of NO in metabolism, as opposed to O_2 co-transportation.

The expression of nsHbs-2 has also been linked to the reduction in blooming periods displayed by Arabidopsis, when over-expressing the Arabidopsis Hb2 (AHb2) gene. It is theorised that when AHb2 is under expressed, NO is not metabolised; it is these abnormally high levels of NO that have been linked to delayed flowering. Peroxidase-like activity and NO metabolism in nsHbs2's has eluded to the possibility of conferring protection from nitrosative stresses on the plant, as well as potentially interacting within some signal transduction pathways. Additional evidence also exists for nsHbs-2's inherent ability to up-regulate defence genes, mount a response through

largely ill-defined means, in efforts to fight off a potential pathogen invasion (Dordas, 2009; Reeder, 2010).

Truncated Hbs (trHbs) are so termed for the cleaved A helix in their heme domain. The globin fold in this globin differs from the typical 3-on-3 alpha helical sandwich, with the formation of a 2-on-2 sandwich. This structure forms antiparallel pairs from helical pairs B/E and G/H. Typically only spanning 20-40 amino acids, the N-terminal A-helix is virtually absent, whilst the CD/D helix has been trimmed with 2-3 residues, just shy of a complete alpha helical turn. In addition, the F helix is confined to a single helical turn that still contains the Fe^{2+} chelating proximal His F8 residue. In stark contrast, this Hb seems lack preference for a specific residue being present at the distal E7 residue, with at least 6 possible residue substitutions at this coordination site. The CD1 residue shows noticeably less variation possible when compared to the E7 residue, where Phe is the highly conserved CD1 residue in non-vertebrates. This residue in trHbs seems to be capable of accommodating Phe, Tyre, or His at this position. Both residues B9 and B10 are highly conserved as Phe and Tyr respectively in this Hb. However, ligand stabilisation is performed by only TyrB10, as opposed to the E7 residue in classical globins when O_2 binds. In the interest of compacting this particular globin homolog, very little is conserved (Ouellet *et al.*, 2002; Wittenberg *et al.*, 2002).

Strong links exists between trHbs and the bacillus like M. *tuberculosis*, as the evidence suggests that it uses trHbs, namely its ability to scavenge NO, to protect itself with the ability to detoxify NO. This is further supported by the fact that knocking out the glbN gene encoding for trHbN in M. *bovis* bacillus, decreases NO-scavenging in non-mitotic cells, an action that could be reendowed by reinsertion of the glbN gene in these same knock-out cells (Ouellet *et al.*, 2002).

Androglobin is still a relatively recently discovered protein when compared to other globins. Unlike hemoglobin, myoglobin, cytoglobin, and neuroglobin the molecular function of Adgb largely remains obscure when compared to other globins. Though considerably more is known about the molecular function of other globins, the full scope and nature of their functions is still uncertain.

Intrinsic disorder

Of all the proteins present within the human proteome, roughly 25% are thought to be fully flexible intrinsically disordered proteins (IDP). An additional 40% of all proteins are thought to contain intrinsically disordered regions (IDR) spanning at least 30 residues in length, yet the proteins in question will remain biologically viable (Dunker *et al.*, 2002). The fact that IDP and IDR still retain their molecular functions seems to stand in direct violation of the structural protein biology model of an induced fit "lock-key" model of protein-protein interaction, where the structure of a protein plays a definite and integral role in that protein's function. In this model, proteins bind their natural ligand with high affinity and high specificity, showing little to no promiscuity in the ligand they choose to bind (Uversky, 2019).

In the light of this it has now become imperative to consider, when modelling proteins in attempts to derive function, that the protein being modelled may harbour an IDR within an otherwise natively folded structure. If neglected, this may lead to a drastic

misappropriation in protein function, through the generating of false positives, as structurally different proteins may in fact perform similar cellular functions (Hensen *et al.*, 2012).

Despite IDPs being branded as fully flexible, it does not mean they may undergo any and all possible conformational changes. There is a predetermined number of conformers possible, heavily restricted by the sequence of the protein, which are characterised by a lack of hydrophobic interactions between individual residues. These hydrophobic interactions normally the induce entropy driven folding of a protein into a distinct 3-dimensional structure. Instead IDPs and IDRs are characterised by a large proportion of polar or charged residues, which tend not to aggregate by engaging in co-operative folding (Babu, 2016). This results in IDP/IDRs exhibiting low net hydrophobicity and high net charge through the disordered region or the full length of the disordered protein (Berg *et al.*, 2002; Hensen *et al.*, 2012; Deiana *et al.*, 2019).

Natively folded proteins are likely to contain much more rigid active binding sites, binding their natural ligand with high affinity, resulting in a highly substrate-specific induced fit model of complex formation. On the other hand, it seems that the lack of structure in IDPs and IDRs of certain proteins, remains to be the key factor in conferring their molecular function. IDPs, because of the fact they lack a structurally defined active site, are able to promiscuously bind with high specificity to a wide array of ligands. This is an ability exploited extensively within cell signalling pathways (Vacic *et al.*, 2007; Huang and Liu, 2013).

Binding to IDPs and IDRs usually results in a limited conformational change within the protein. The disorder-to-order transition is limited, as though this may endow an IDP/IDR the ability to partially fold, there are still segments within the protein that will

remain disordered. This conformation of IDPs is typically referred to as a fuzzy complex formation. The ability to make such exaggerated changes in conformation upon binding, and the time-scales in which IDPs are able to undergo said changes, is now thought to be as meticulously conserved as 3 dimensional structure in natively folded protein (Kosol *et al.*, 2013; Abyzov *et al.*, 2016; Babu, 2016). In cases where proteins lack a clearly defined, if any structure, considering the protein's dynamics upon binding seems to be a more pivotal and integral to elucidating IDPs and IDRs molecular function. This is in much in the same way that structure is related to function in a natively folded protein (Hensen *et al.*, 2012).

Despite Adgb having already had a defined structure attributed to the portions of its sequence structure that homologous to globin heme domains and a calpain-like protease domain, it is still highly likely that within the sequence there will IDR regions. These regions may themselves have a distinct function or merely function as linkers to ordered regions within the protein sequence. As aforementioned, attempting to deploy secondary structure predication and fold recognition techniques to portions of the sequence likely to contain disorder, may result in the generation of false positives, so it is imperative to predict regions likely to contain disorder. To achieve this, a series of methods will be used to determine the local propensity for disorder, at a residue level, within defined limits.

The first method used was available through the IUPred2A server, which used a combination of the IUPred2 method and the ANCHOR2 method to estimate disorder. Both use energy estimations to approximate disorder at the residue level. While the IUPred2 method is concerned with predicating general disorder of a given sequence, the ANCHOR2 method is more concerned with identifying potential disordered binding

regions. Both generate a score at each residue between 0 and 1 with 0.5 functioning as the threshold, which if a residue generates a score above this threshold it will be considered disordered (Mészáros *et al.*, 2018).

The recently updated Spot-disorder2 method utilised a combination of machine learning methods known as deep Squeeze-and-Excitation residual inception and long short-term memory. This will generate a score locally at the residue level, which Spot will assign a threshold of 0.46. Any residue generating a score above this value will be considered disordered, and vice-versa (Hanson *et al.*, 2019).

DISOPRED3 is made available through the same server as PSIPRED 4.0, and it originally trained to identify IDRs through the missing residues in structures gained through high resolution X-ray crystallography. However, in addition to DISOPREDs current iteration, an additional 2 independent predictors of disorder have been incorporated within DISOPRED3, making it a meta-predictor. One of these predictors will also seeks to identify and annotate protein-binding disordered regions. This method generates residue level scores, applying the threshold of 0.5, above which a residue will be considered disordered (Jones and Cozzetto, 2017).

RaptorX Property also utilizes machine learning, more specifically, both solvent accessibility and disorder predictions are based on deep convolutional neural feeds (DeepCNF). This method will generate a disorder score at a residue level, applying a threshold of 0.5, above which the residue generating that value will be labelled as disordered (Wang *et al.*, 2016).

Protein disorder prediction system (PrDOS) uses a position-specific score matrix (PSSM) or profile generated from subjecting the query amino acid sequence to two rounds of PSI-Blast searches. The query sequence PSSM is then run through 2

independent disorder prediction methods, the results of which are then combined. The first method relied of solved crystal structures to produce a learn data set for the supervised machine learning method known as support vector machine (SVM), where missing coordinates or residues in the structure were defined as disordered. The query sequence PSSM is then run through the same previously trained SVM, which gives an output at the residue level, for the propensity for that residue to be disordered. The second predictor uses a PSI-BLAST search for the alignment homologues detected, when using the PSSM formed from the query sequence. The probability that any given residue is disordered, is gained from the weighted ratio that other residues aligned within homologues are also disordered. The results are then combined, where the first method used is weighted 1.0 and the second method used is weighted 0.11. A probability between 0 and 1 will be generated at each residue, and a threshold of 0.5 applied (Ishida and Kinoshita, 2007).

GlobPlot2 rests on the assumption that P, the propensity for any single residue to be disorder, may be expressed as P = RC – SS. Originally in this equation RC was the propensity a residue would adopt a 'random-coil' and SS was the propensity any residue would adopt 'secondary structure' i.e. either helix or sheet. The definition for secondary structure has since been redefined in line with DSSP. Anything outside this DDSP derived distinction was defined as RC. The propensity is then used as an input for a sum function that generates a graph for the full length of the protein. The differential of graph function at each residue will produce a value between -1 and 1, which when applying a threshold of 0, above which a particular residue is indicated as disordered (Linding, Russell, *et al.*, 2003).

MFDp, much like PrDOS, relies on an SVM to render its prediction of disorder at a residue level. These SVMs are further divided into 3 classifiers, SVM_{ALL} (for all residues), SVM_{LONG} (for >30 residues), and SVM_{SHORT} (for <30 residues). Each SVM was assembled and trained independently using different parameters and different input features aggregated from IUPred, DISOclust, and DISOPRED3. For this reason, MFDp is considered a meta-predictor of disorder. The final probability for disorder is attained by combining the results of each individual SVM, then applying a threshold of 0.37, above which a particular residue will be considered disordered (Mizianty *et al.*, 2011).

MFDp2 combines the residue-level of disorder generated by MFDp with the sequence-level of disorder generated by DisCon to improve the quality of the prediction when both are used independently. The threshold value associated is 0.5 as opposed to 0.37, as was the case for MFDp (Mizianty *et al.*, 2013).

DISOclust derives its prediction of disorder from the per-residue error calculated using the average S-score present in multiple models produced by fold recognition methods for the query sequence. This is then followed by the analysis of the conservation of per-residue error across all models. The threshold of 0.5 is applied to the output values to determine disorder (Mcguffin, 2008).

PONDR VSL2 like MFDp, is also considered a meta-predictor. The results from PrDOS, DISpro, DISPROT, DISOPRED2, POODLE-S, IUPred, and DisEMBL are used as the input for a combination of neural network predictors for both short and long disordered regions. Short regions are defined as <30 residues and long are defined at >30 residues. The predictors of each type of disordered region are trained independently, with datasets of sequences of that specified sequence length. The final prediction is a result of the weighted averages between each predictor, determined by a second layer

predictor, where a threshold of 0.5 is applied to the generated disorder values (Xue *et al.*, 2010).

DisEMBL REM465 is based on an artificial neural network. It uses information from missing electron densities, within a proteins crystal structure to train the machine learning method. This is as these often denote portions of the sequence that contain intrinsic disorder. However, this particular definition of intrinsic disorder has a drawback associated with the way in which some protein structures are solved experimentally by truncating the protein, cleaving highly flexible or disordered regions. This results in the data sets being used to train the neural nets being skewed, as more proteins should contain missing electron densities than at present. The prediction will generate values between 0 and 1 at a residue level, where a threshold of 0.6 is applied to determine disorder (Linding *et al.*, 2003).

Molecular recognition features (MoRFs) divided into alpha-MoRFs, beta-MoRFs, and iota-MoRFs, and they are 10 – 70 residue long sequences that undergo disorder-to-order transition upon binding a natural ligand. They form alpha helical, beta sheet, or irregular random patterns respectively. This is why the results from disorder predictions made by especially ANCHOR2 and by the IUPred2A server, may shed light on possible portions of the sequence that could function as a binding site that can propagate the formation of structure (Gsponer and Madan Babu, 2009).

Secondary structure prediction

Traditional techniques have been increasingly subsumed under Computer aided drug discovery (CADD) as *in silico* techniques carry the added benefits of being time effective, inexpensive as well as carrying considerably less associated risks, as experimental work may lack efficacy (Agarwal, 2013). A known 3D structure is not only vital for understanding protein physiological function, as protein structure has been intrinsically linked to function, but it also forms the cornerstone for inferring a new possible drug-able site on the protein.

In the stead of experimental data, we will seek to build models of an unknown protein's structure solely from sequence alone. This is because it is widely accepted that all the information required to infer a 3D structure of a given protein, is present within the sequence alone. A chain of amino acids was thought to naturally and intuitively fold into the most thermodynamically stable conformation, without any external provocation. This is possible in theory, despite it not being possible in practise, as an already known protein template is required to infer structural information. This is because it is now known that chaperones and other small proteins help assist in protein folding, aiding the macromolecule to reach an energetically favourable conformation, through posttranslational modifications. This makes the task of predicting the 3D globular structure of the final macromolecule that much harder, despite the inherent ease in predicting whether a set of localised amino acids residues will adopt α-helical (H), β-strand(E) or irregular coiled conformations (C) (Singh, 2001; Cheng *et al.*, 2008). Experimental data always needs to be obtained to justify conclusions reached through molecular modelling techniques, by utilising such techniques as NMR or X-Ray crystallography.

Despite it being possible to infer secondary structures from sequence alone, this still leaves a gap in understanding the macromolecule in terms of its tertiary structure. Secondary structure prediction does, however, provide restraints for fold recognition that will go on to dictate the overall geometry of the protein. Ultimately, these restraints will develop understanding of how these microstructures will aggregate and form the 3D structure of a whole protein. Such predictions have been assessed to be up to 80% accurate when compared to experimental data (Pirovano and Heringa, 2010; Zhou and Troyanskaya, 2014; Drozdetskiy *et al.*, 2015).

In respect to secondary structure prediction, multiple sequence alignments (MSA) have proven invaluable, as they narrow the possibility of error arising from modelling from single homologous sequence. Mainly because MSAs contain more information on which residues are conserved between homologous proteins, either by sequence identities present or conserved substitution, in effect safeguarding against the occurrence of false-negatives (Soding, 2005).

Though it is integral that the template strands must be of a non-redundant nature. This mean that differing isoforms generated from the same gene that arise as a result of alternate splicing or posttranslational modification, will contaminate and distort MSAs, ultimately leading to the degree of homology being expressed by both the template strands to the unknown sequence strand being heavily skewed. The inaccuracy in the degree of homology displayed arises as a result of the MSA's inability to differentiate between mutually exclusive exons. Thus, in a situation where a domain exists pertaining to an exon present within only one alternatively spliced protein but not the other, the MSA will still seek to align whole of the protein ignorant of the fact that

certain sections of protein will not share homology with each other. Naturally this will lead to a false index of homology being generated (Nord, 2017).

BLOSUM matrices are used to asses this degree of homology present within template sequences prior to the sequences being subjected to an MSA. Both secondary structure prediction web-servers JPRED (Drozdetskiy *et al.*, 2015) and PSIPRED (Buchan and Jones, 2019) may utilise a user defined MSA to predict the presence of secondary structures within the query sequence structure. The use of more than one homologous sequence incurs a number of benefits. Mainly it allows for greater sequence coverage over the length of unknown protein, in addition to the possibility that one template may provide a better alignment for a single domain in the target protein than any other template, in this way the presence of one template may compliment the presence of another within the MSA. This in a target sequence with multiple domains, would help bolster accuracy, improve reliability, and confidence in the model generated. RaptorX Property does not always seek to use experimentally solved structures as templates in protein modelling, unlike JPRED and PSIPRED. Instead of protein structures, RaptorX Property will seek to utilise the target sequence's profile. Though unlike RaptorX, this will be done without assembling the tertiary or quaternary structure query sequence. Instead, RaptorX Property will seek to identify each individual secondary structure (Wang *et al.*, 2016).

Machine learning and neural networks

The prediction of secondary structure in an unknown query sequence may be achieved by utilising machine-learning algorithms. These will seek to loosely mimic biologically occurring neural networks and consist of at least 2 layers. The input layer refers the position of each residue within the sequence structure, whilst the output layer functions an activation threshold function, which the input is then fed into. However, before the input reaches the output function it must first be scaled, by multiplying with a weighting associated with any given input. The weighting of each input is something the algorithm must learn, by being repeatedly exposed to a test set of input data with a known target value. Once the input has been subsumed under both the corresponding weight and output function, i.e. a known sequence structure with known empirically verified secondary structures, then it is possible to determine the structure at a given residue of an unknown sequence.

In addition to these two indispensable layers, any other additional layers are considered hidden layers. These layers will function to break down and interconnect simpler units with similar topologies, thus adding complexity to the network and providing downstream functions with input produced from previous layer's output. When considering the conformation for any single residue in Fig 1.6 it is shown that the 8 residues flanking either side are also taken into consideration. For each of the 17 residues being considered when predicting secondary structure, each possible amino acid is used as a mutually exclusive input. In this example both the hidden layers and the output layer contain two scalar units the input must be subsumed under before the structure maybe determined (Singh, 2001).

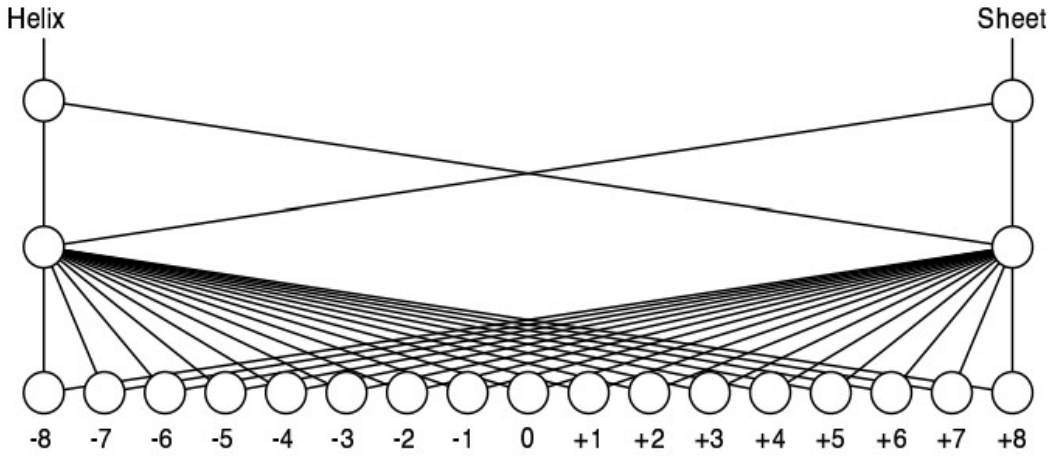

Fig. 1.5: Neural network topology for secondary structure prediction. (Source: Singh, 2001)

In addition to machine learning, the homology between a known protein and an unknown protein may be exploited and utilized to clarify the characteristics of unknown protein structures. This can be done in far more explicit terms than merely providing test data from which machine learning is made possible. This is only possible if pairwise sequence alignments determine that there is >30% sequence identities between the sequence structures, as sequence similarities above this value are likely to have comparable protein structure. Conversely sequence similarities below 25% have displayed less than 10% sequence identities between protein structures. Pairwise sequence similarity between 20% - 35% is widely known as the twilight zone for protein homology; the degree of homology between these ranges plummets drastically. In rare cases homology may still be present within proteins displaying exceptionally low sequence similarity <20% (Rost, 1999).

Fold recognition

Despite the staggering variation possible within proteins, there exists a finite number of folds (conformations) that a protein may adopt. This limit has been theorised to be no more than 10,000 possible unique protein folds, which arise from how secondary structures begin to aggregate. In the absence of experimental data, fold recognition proves vital to elucidating tertiary globular structure in computational biology, proving to be indispensable in aiding understanding of how proteins will aggregate. How for example, a set of beta sheets will form either a parallel or anti-parallel sheets before they form larger secondary structure such as a beta-sandwich or Greek key motifs (Wang, 1996; Koonin, Wolf and Karev, 2002; Saini *et al.*, 2016).

Firmly underpinned by the widely accepted fact that the structure of a given protein is conserved far more stringently than sequence, homology modelling, comparative modelling, and fold recognition are encompassed by and subsumed under the hypernym template-based modelling (TBM), which may also be referred to as threading. However, their accuracy is stunted by the mere fact that classical MSA techniques seek to only locally align the sequences queried never globally align. TBM have their reliability capped by the accuracy of the methods used to assess the homology between template sequences and the target sequence.

One way over overcoming this issue in accuracy is the use of multiple template fragments when modelling a structure of any given sequence. I-TASSER seeks to reassemble fragments aligned to a sequence structure, excised from a number of threading templates used. This is done with the aim of achieving a higher level of

homology locally, by matching the query sequence to a small fragment from a threading template, as opposed to attempting to achieve global alignment for the query sequence (Yang and Zhang, 2015).

The use of empirically solved protein structures holds the added benefit of containing energy functions derived from experimental data, which when coupled with physics-based approaches, can be used to simulated folding *in silico.* This is only possible through the use of statistical analysis techniques, that allow the evolutionary relationship between homologous sequences to be derived merely from the degree of similarity. It is these same evolutionary relationships that allow large databases of raw sequence data containing all known proteins to be efficiently searched. PHYRE2 allows MSAs to be utilised by first converting it into a hidden Markov model (HMM), which may then be rapidly compared against other HMMs with known structures, through HHsearch (Söding *et al.*, 2005) a robust homolog detection algorithm with which PHYRE2 will seek to build a rudimentary backbone structure for the protein containing none of the proteins monomeric side chains (Kelley *et al.*, 2015).

In the absence of sufficient homologs to produce a detailed sequence profile, statically learning methods may be employed in order to achieve protein modelling results similar to those produced by empirically based methods. RaptorX (Peng and Xu, 2011) achieves this by exploiting template structural elements in order to align the target sequence to multiple templates simultaneously. This generates a threading scoring function that relates the target sequence to the template structure. In this way the target sequence directly adopts to the template structure, as opposed to this same interaction between target sequence and template structure being mediated through a sequence alignment with the template and query sequence. This in turn removes any error that

may potentially arise from an inaccuracy in an MSA between any template sequences being used and the target sequence. However, in the presence of sufficient homologs able to produce an adequate pairwise alignment or MSA, RaptorX will use its inherent ability to make statistical inferences from structural alignments of the template. This ability to make statistical inferences helps to improve the sequence profiles generated in order to assist modelling.

HHPRED was the first alignment program to use a HMM for its pairwise comparison of sequence profiles. As with all the other servers, it was aimed at detecting the presence of remote homologues, by searching a large database of experimentally solved protein structures (Söding *et al.*, 2005).

Calpains

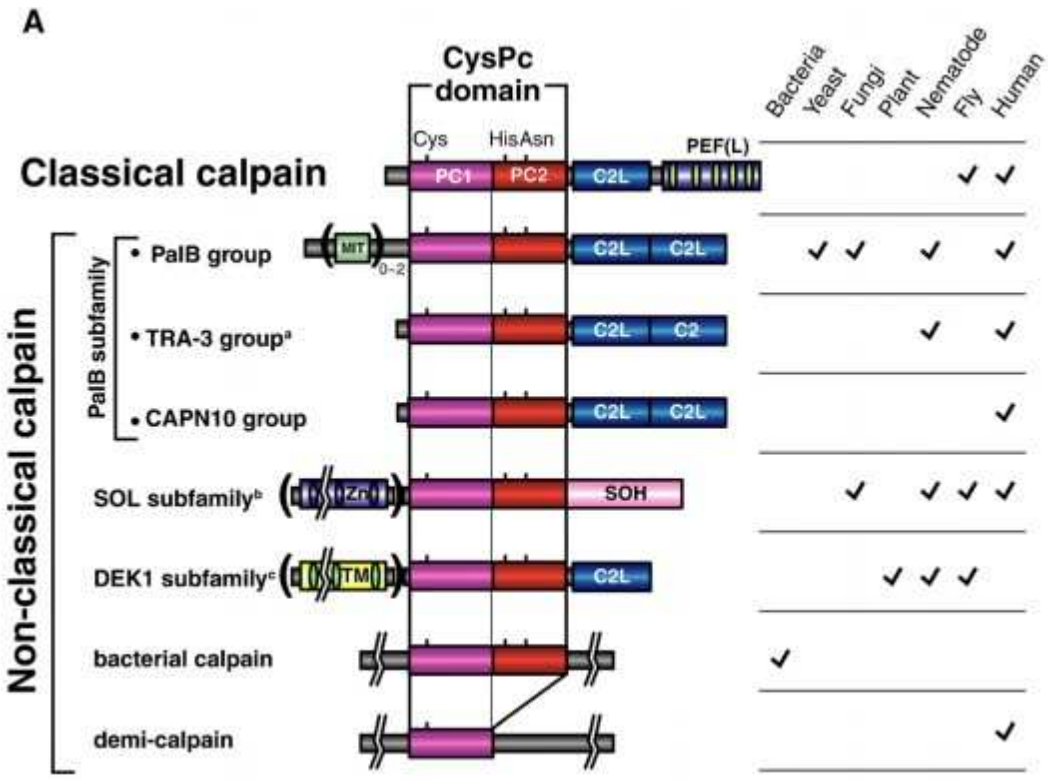

Fig. 1.6: Structural classification of calpain. PC1/PC2 represents the two distinctive parts of the catalytic triad that forms the calpain hydrolytic active site. C2L represents the Ca^{2+} dependent phospholipid binding C2-like domain comprised of an 8 stranded anti-parallel beta sandwich. PEF represent the 5 (penta-) EF hand Ca^{2+} binding motifs. SOH represents the small optic lobe (SOL)-homology domain, which has an unknown function; SOLs from calpain have a Zn^{2+}-finger binding motif at the N- terminus. MIT represent a microtubule inserting and transport motif. TM represents the transmembrane domain in DEK1 calpain (Ono and Sorimachi, 2012).

With the presence of a single N-terminal calpain domain already identified within the Adgb sequence (Hoogewijs *et al.*, 2012), it stands to reason that the remainder of

the sequence is likely to contain additional domains which share homology to other calpain domains. The super-family of proteases displays a great deal of structural variation, resulting in the differing isomers that may be expressed *in vivo*. So termed for their **cal**cium dependence, and their pa**pain** like characteristics, there are 15 different genes within the human genome that encode these cysteine proteases; 9 of those genes encoding for classical calpains. Commonly shared amongst both conventional and non-conventional calpains, is their papain-like protease domain (CysPc). This domain will function to modify substrate through controlled proteolysis, without destroying the target, is shown in its two composite parts, PC1 and PC2 (Fig. 1.6). Classical calpains will also contain a C2-like domain which is thought to engage in membrane localisation, mediated by calcium-dependent phospholipid binding. Composed of an 8 stranded antiparallel beta sandwich with 3 Ca^{2+} binding loops at end, this C2 domain is implicated in both signal transduction and membrane trafficking (Davletov and Sudhof, 1993; Bryan Sutton *et al.*, 1995; Farah and Sossin, 2012; Ono and Sorimachi, 2012). The 5 (Penta) EF-hand (PEF) domain acts in pairs to engage in Ca^{2+} binding, with the 5^{th} EF hand facilitating dimerization in m- and mu- calpains (Kolobynina *et al.*, 2016). The majority of non-classical calpains will usually lack a PEF domain from the generic classical calpain template, whilst retaining a C2-like domain, though this may not always be the case. Fig. 1.6 shows SOL sub-family, bacterial, and demi-calpains retaining merely the CysPc domain with no C2-like domain (Söding *et al.*, 2005; Ono and Sorimachi, 2012).

Classical mu- and m- calpain, so termed for the Ca^{2+} concentrations required *in vitro* to reach half V_{max} being in the µM and mM range respectively, are also known as calpain I and calpain II respectively. There is evidence to suggest that in mu-calpain the N-terminal localisation sequence (NLS) immediately before domain I, confers the ability

for mitochondrial import as its removal blocks the mitochondrial import of m-calpain. Similarly, addition of this same NLS to m-calpain, which natively lacks the ability for mitochondrial import, allows m-calpain to be imported into the mitochondria (Badugu *et al.*, 2008). This remains consistent with the *in vivo* localisation of mu-calpain within the intermembrane space of the mitochondria, where it functions to activate apoptosis inducing factor, and mitochondrial Bax (Garcia *et al.*, 2005). Furthermore, the inability for m-calpain to undergo mitochondrial import is supported by m-calpain being activated in the presence of epidermal growth factor (EGF), however, this only occurs when m-calpain is adjacent to and docked with the plasma membrane (Shao *et al.*, 2006).

M-calpain (Fig. 1.7) has been shown to catalyse the proteolysis of capase-12 in ER induced cell apoptosis, which seems to suggest possible tumour suppression capabilities, in addition to aiding gestation as apoptosis remains necessary (Zuo *et al.*, 2018). M- and mu-calpain have also been implicated in the processes of cytoskeletal remodelling integral to conferring cell-motility, cell-adhesion required for wound repair. The full-length classical calpain exists as a heterodimer with a “large” L-subunit -80kDa (domain I - domain IV).

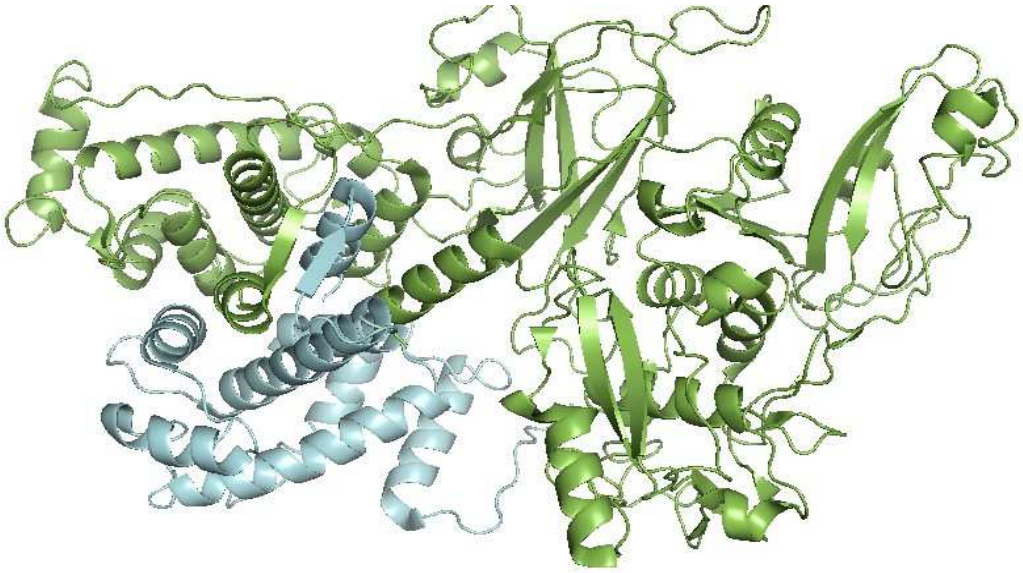

Fig. 1.7: Crystal structure of human m-calpain – 1KFU. The protease containing "L-subunit" shown in green, with the regulatory "S-subunit" shown in cyan.

As a cysteine protease the catalytic triad in human m-calpain, is highly conserved; it consists predominantly of residue Cys-105 located in sub-domain IIa (N-terminus domain II prior to crystallographic evidence). In addition, residues His-262, and Asn-286 both subdomain IIb (C-terminus domain II prior to crystallographic evidence) help to form the remainder of the proteolytic core. The cysteine functions as the site of catalysis in the proteolysis of a peptide bond in the substrate; histidine and asparagine function as proton donors/acceptors for the cysteine residue during this catalysis. The C2-like domain III has been suggested to interact electrostatically with the other domains in order to stabilise, and possibly regulate the protease potential of the L-subunit. The C2-like domain also holds a capacity to associate with the plasma membrane through its innate ability to bind phospholipid. Crystallographic structures have shown both domain IV and domain VI (S-subunit) to contain 5 EF-hand motifs each (Badugu *et al.*, 2008; Huang *et al.*, 2016). The standard EF-hand motif predominantly consists of a highly

conserved 12 amino acid Ca^{2+} binding loop, which is then immediately flanked either side by helices termed E and F respectively. Nonstandard EF-hand motifs come in 4 differing variations, with additions, insertion or substitution mutations, that further stabilise Ca^{2+} binding (Maki *et al.*, 2012).

However in both the case of domain IV + VI, only 4 of these motifs were capable of binding Ca^{2+}; the final EF-hand motif expressed at the C-terminus has been attributed with the heterodimerisation of the L- and S-subunits (Strobl *et al.*, 2000; Tompa *et al.*, 2001; Chou, 2010). Hetero-dimerization is something in calpains only exhibited by m- and mu-calpains. It is solely in these calpains where the S-subunit is expressed, and bonded non-covalently to the L-subunit, within the calpain family of proteases. The S-subunit's physiological activities are largely regulatory, as it predominantly functions as a "molecular-switch". It does so by harbouring the calmodulin-like domain VI, which is capable of binding Ca^{2+} to the 4 of the 5 EF-hand motifs present. The S-subunit also contains the Gly-rich hydrophobic domain V, later targeted for autolysation, though the methods by which autolysis is achieved still remain largely obscure (Kolobynina *et al.*, 2017).

Of the two, m-calpain has significantly more physiological activity *in vivo,* as evidenced by mu-calpain knock-out mice remaining embryonically viable. However, the mice appear deficient in the ability for phosphorylate tyrosine and the platelet aggregation, which in turn has adverse effects on the blood's ability to clot. M-calpain knock-out mice are not embryonically viable. This suggests an integral role in embryonic gestation, with the possibility of a more active role than mu-calpain in relation to cytoskeleton remodelling; a process on which embryonic development is wholly reliant upon. Furthermore, disruption of the small regulatory subunit encoded by the Capn4

gene in mice, results in the inactivation of both m- & mu-calpain, and consequently the embryo's lack of viability. This is due to the lack of development in the cardiovascular system, and the over accumulation in undeveloped erythroid progenitor cells. This pathology is consistent with the loss of the ability to regulate, and interact with cytoskeletal remodelling processes necessary for typical embryonic development (Dutt *et al.*, 2006).

EF hands

Adgb would not be able to function as a calcium-dependent protease without the ability to facilitate Ca^{2+}-binding, an ability conferred by EF hand motifs, though some EF hands are promiscuously bind Mg^{2+-} and Zn^{2+}. These are motifs are usually expressed in pairs, as the aforementioned proteins contain an odd number of these motifs they tend to dimerise. Proteins that contain a single EF hand are present within both eukaryotic and bacterial proteomes, though these seem to mediate protein-protein interaction as opposed to functioning as a Ca^{2+} binding protein. Characteristic of EF hands are their helix-loop-helix topology and a 12-residue long interhelical loop, which forms their Ca^{2+} binding site, as shown in Fig. 1.9. Upon binding calcium, these motifs will undergo a conformational change, and begin to display bipyramidal geometry (Zhou *et al.*, 2013; Kolobynina *et al.*, 2016).

Residues 1 (X), 3 (Y), 5 (Z), and 12 (-Z) in the loop form contact residues; they seek to coordinate the Ca^{2+} ion through oxygen atoms on their sidechain carboxyl or hydroxyl

groups. Contact residues 7 (-Y) and 9 (-X) chelate the ion through an oxygen atom the main chain carbonyl group and a bridge water respectively. Residue 12 (-Z) seems to be the most conserved residue, as it forms a bidentate coordinating ligand (Zhou *et al.*, 2009).

Calmodulin

The presence of a Calmodulin (CaM) binding domain, the IQ-motif expressed within ADGB's sequence, means the ability to bind Ca^{2+} may not be necessarily conferred by sequences homologous to domains IV+VI in both m- and mu-calpain. In their stead CaM, a highly conserved protein sharing as much as 90% similarities between human and plant isoforms, may fulfil this function. This is a suitable substitution to domain IV + VI as its possess the ability to be sensitive, and bind Ca^{2+} (Fig. 1.8) at concentration in the range of 10^{5} – 10^{-6} M, exhibiting a conformational change as it does so (Ono and Sorimachi, 2012).

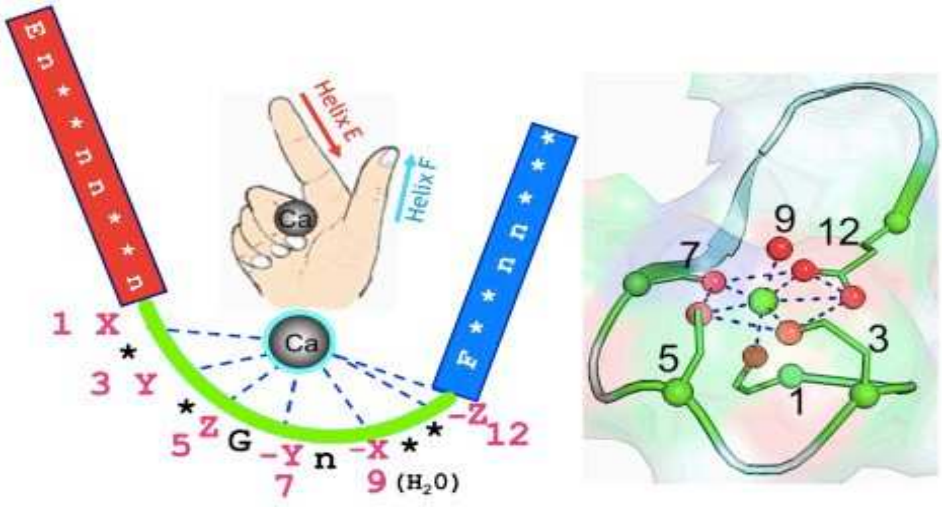

Fig. 1.8: CaM EF hand Ca2+ binding site. Calcium binding at the binding loop formed in the interface between helices E in red and F in blue. The loop coloured in green with red letters forms the 12-residue long Ca^{2+} binding loop. Highly conserved contact residues 1 (X), 3 (Y), 5 (Z), 7 (-Y), 9 (-X), and 12 (-Z) form the 12 residue long binding loop. (Zhou *et al.*, 2009)

CaM's association and constitutive activation of calcium-activated potassium channels, which display no inherent ability to directly bind calcium yet remains dependent on Ca^{2+} binding for their activation, adds further evidence of how suitable a substitution it is in the possible absence of m- and mu-calpains domain IV + VI. Further, CaM is implicated in the mediation of cellular processes also associated with calpains, including but not limited to, cell apoptosis, motility, and muscle contraction, providing further evidence for the conjecture that CaM may indeed mimic domain IV+ VI's regulatory role in Adgb (Ihara and MacDonald, 2007).

Furthermore, the ease with which CaM readily dimerises, and is in turn responsible for the regulation of an associated enzyme's activity, is further evidence for the functional similarities displayed by both CaM and domain IV+VI in m- and mu-

calpains. It does so by displaying an ability to potentially, though unlikely, bind Ca^{2+} at each of its 4 EF-hand domains. 2 of which are present on the N-lobe and C-lobe globular domains located at each terminal of the protein (Lafitte *et al.*, 1999; Chagot *et al.*, 2009; Li *et al.*, 2009; Tidow and Nissen, 2013).

Aims and Objectives

In chapter 2, we will seek to characterise the full length of Adgb using disorder prediction, this will be in the hopes of identifying any IDRs in the sequence, that may add noise when subjecting the sequence to fold recognition techniques. We hope that that after these predictions, we will have fully defined the areas in the protein sequence were will expect secondary structure and fold recognition techniques to produce a clear consensus of any structures present.

In addition to disorder predictions, we will also be using profile-profile alignments to identify possible EF hand motifs, as these would enable the protein to bind the calcium necessary, for the calcium-dependent processes carried out by the calpain domains known to be within the Adgb sequence. Finally, we will seek to plot the secondary structure, as predicted by a consensus between JPRED and PSIPRED, next to the disorder generated in hopes to identify key structural MoRFs that would enable protein folding in the areas where disorder prediction methods have predicted disorder.

In chapter 3 we hope to produce secondary structure and fold recognition predictions. Secondary structure prediction will seek to structurally characterise each residue as helical, sheet, or loop, in the regions we suspect to be well ordered. Whereas, fold recognition will provide both structural as well as functional information on the query sequences used, in aims of elucidating the molecular functions of Adgb as a whole.

Chapter 2. Disorder predictions in Androglobin

2.1. Introduction

With over half of the proteome either being classed as intrinsically disordered protein (IDP), or containing intrinsically disordered regions (IDR), it has become imperative to detect intrinsic disorder when attempting to computationally model a protein.

This helps lower the possibility of producing false positives, arising from attempting to model structure on to a target protein sequence, which otherwise would be likely to contain disorder, when utilising secondary structure prediction and fold recognition techniques to predict secondary and tertiary structures respectively.

In portions where the consensus formed between different disorder prediction methods of the region is largely disordered, unless they are helical recognition elements, we do not expect to see any secondary structure (i.e. helices and beta sheets) associated with these regions. Thus, in addition to the possibility of discovering these recognition elements, secondary structure prediction in conjunction with disorder prediction provides a check and balance on the reliability for the disorder predictions produced.

Profile alignments between Adgb's MSA and Calpain alignments have located regions in Adgb's sequence structure suspected to contain Calpain domains IIa and III. However, for these domains to retain their hydrolase molecular function, Ca^{2+} binding has to be a feature either intrinsic to Adgb or conveyed by other means. One way the ability to bind Ca^{2+} maybe attained, is through EF hand motifs possibly present within Adgb's sequence structure. Consequently, in addition to identifying secondary structure and disordered regions, we also seek to identify whether any regions of Adgb are

predicted to contain EF hands. Together these predictions will lay the basis for a focused fold recognition study of Adgb, at least as far as seems possible in the light of the disorder predictions.

2.2. Methods

Due to the exceptional length of the Adgb's sequence, it was integral to split it into three sections, as the 1667 residues are too long a sequence for many webservers. This was achieved using the multiple sequences alignment visualisation java tool Jalview (Waterhouse *et al.*, 2009). This first required the identification of the sequence limits for the pre-heme, heme domains, and post-heme regions. The residues before the heme domain formed the first fragment (M1 – N781), the remaining residues within the heme domain, including the IQ domain, (F782 – K985) formed the second fragment, and lastly the residues following the heme domain (I996 – K1667) formed the remaining fragment.

Once the sequence limits were identified within Jalview, the residues in the human Adgb sequence alone were highlighted, and the selection was then outputted into a textbox inside Jalview in FASTA format. The output sequence will begin with an angled bracket '>' on the first line, which is where the name of the sequence may typically be found, this is then followed by a hard-return where the protein sequence is will begin on the next line. Typical FASTA format dictates there to be 80 residues to each line. Sequence is copied and pasted into a text editor and saved into a file, while making

sure to replace the default file extension preferred by that text editor with the file extension '.FASTA' or '.fa' for a FASTA file type. This was repeated for the remains sequence fragments.

Once all three sequence fragments were used to create FASTA files, each in turn each file in turn was then submitted into each of the following 12 intrinsically disordered region prediction server:

- IUPred2A which will be referred to as IUPred
- ANCHOR2 which will be referred as ANCHOR (Mészáros *et al.*, 2018)
- SPOT-disorder 2 which will be referred to as SPOT (Hanson *et al.*, 2019)
- DISOPRED3 which will be Multi-layered Fusion Based Predictor (MFDp) (Mizianty *et al.*, 2011)
- referred to as DISOPRED (Jones and Cozzetto, 2017)
- Raptor Property (Wang *et al.*, 2016)
- Protein Disorder prediction Systems (PrDOS) (Ishida and Kinoshita, 2007)
- GlobPlot v 2.3 (Linding *et al.*, 2003)
- MFDp2 (Mizianty *et al.*, 2013)
- DISOclust (Mcguffin, 2008)
- PONDR VSL2 (Peng and Zhang, 2006)
- DisEMBL Intrinsic Protein Disorder Prediction v 1.5 which will be referred to as DisEMBL (Linding *et al.*, 2003)

Raw sets of data generated by each webserver provide a statistical inference of the disorder expected at each residue. The threshold dictated by each predictor respectively denotes a value above which, any particular residue will be considered disordered by that predictor. Graphs were produced by using the datasets, where in in Excel, the residue positions on the Adgb sequence were plotted as the independent variable on the X axis. This enabled us to establish as a common frame of reference, allowing for an effective comparison across the disorder predictors. The probability of that particular residue being disordered was treated as the dependent variable and was then plotted on the Y axis. A total of 12 graphs would be produced, in this way inside Excel for each of the 3 fragments, to which a second series of coordinates were plotted. This second series that was plotted along the Y axis, contained predictor specific disorder cut off value.

The graphs were imported into a graphical editing software, where they were resized to give them identical dimensions. It was important the graphs had identical dimension so we could maintain the common frame of reference that would allow for accurate comparison between webservers. A rectangle was placed at the corresponding residue positions, in each instance where the graph exceeded the threshold of disorder that has been dictated by each predictor used. This made it possible to replace each graph with solely the X axis and a series of rectangular blocks that denoted the residues which were predicted as disordered.

In order to further maintain a coherent frame of reference, when considering the more comprehensive predictions of secondary structures and fold recognition yet to come, JPRED (Drozdetskiy *et al.*, 2015) and PSIPRED (Buchan and Jones, 2019) secondary structure predictions were plotted alongside disorder predictions. To ensure the

reliability of the structures produced, on the secondary structure formed a consensus between webservers by being present in both JRED's and PSIPRED's alignment would be included. This was done in hopes of identifying key molecular recognition features (MoRFs) in close proximity to disorder, which may help mediate disorder-to-order transitions.

Due to calpains being calcium-dependent proteins, and two calpain domains believed to be present within Adgb's protein sequence (see chapter 3), it becomes important to elucidate and understand the mechanism by which calcium would bind, as it would be integral to the calpain function. Profile alignments using Clustal Omega (Sievers *et al.*, 2011) between the in-house Adgb MSA and two seeded alignments of an EF hand motifs obtained from the Pfam database (El-Gebali *et al.*, 2018) were obtained. Each profile alignment was performed using the Adgb MSA and 1 out a possible 2 seed files in turn: EF hand 1 (PF00036) and EF hand 8 (PF13833). These alignments produced FASTA format files viewable in Jalview. The first alignment between original Adgb MSA and a seed file represents the most strongly predicted putative EF hand region in Adgb. To find a second predicted putative EF hand region in Adgb, the section of the MSA that aligned to the seed file was cut out using Jalview. This process is then repeated with the now modified MSA and the EF hand MSA, as in the previous step. This process was repeated and ensures all possible EF hands within the MSA, may be identified using an EF hand MSA.

2.3. Results and Discussion

Disorder

2.3.1. Region M1 – N7

Generally, the sets of residues present within the interval M1 – N781, represented by Fig. 2.1, have been split into 3 main areas that are disordered. These three areas, which across all the predictors generally seem to contain disorder, are residues M1 – H80, L300 – V400, and F520 – L620. The ordered regions are therefore roughly H80 – L300, V400 – F520 and L620 – P780. It is in these structured areas that the majority of secondary structure predicted by JPRED and PSIPRED resides, in addition to containing Calpain domains IIa (E59-G327) and III (W632 – E775).

Of all the predictors utilised, IUPred's ANCHOR score has seemingly predicted the least amount of disorder. This disorder is within the same residue intervals that the general consensus across all the predictors seems to follow, despite the disorder being presented in a sporadic fashion and spanning a much narrower range of residues, typically under 40 consecutive residues. IUPred's score follows this general consensus; the server predicted residues within the same general consensus. Though the blocks of disorder seem to be in same residues M1 – H80, L300 – V400, and F520 – L620, they appear in a discontinuous and interrupted fashion in several places. There appears a number of new disconnected strips of disorder more than 5 consecutive disordered residues wide in each case, in the interval between T460 – K500. Only a single alpha helix (as predicted) overlaps with the disorder predicted by this particular method at the interval L300 – P320.

The SPOT disorder, DISOpred, PrDOS, and Raptor property predictions also follow the same general consensus in a similar fashion to IUPred. Additionally, SPOT, DISOpred, and PrDOS contain three blocks of disorder in the regions K60 – H80, K480 – K500 and L700 – G710. The disorder produced by DISOpred at K480 – K500 presents as two disconnected blocks no more than 10 residues in width. However, in the SPOT and PrDOS alignments this appears as a continuous block of disorder. While the localisation of all three of these blocks of disorder seems largely unchanged, the width of the blocks changed in width by less than 10 residues between each predictor, barring the PrDOS prediction at K480 – K500. PrDOS and Raptor property have predicted disorder at the end of this fragment roughly no more than 10 residues and 2 – 3 residues in width respectively.

Globplot has produced the most disconnected disorder prediction in this fragment, especially in the regions M1 – H80 and C765 – P780 and seems far more intermittent in its placement of disordered regions than any other predictor utilised. Blocks of disorder roughly 10 – 15 residues in width are present in the intervals K160 – P170, P470 – K480, and L700 – G710.

MFDp's disorder prediction is more similar to the PrDOS prediction than to that of MFDp2. However, it does not contain the small block of disorder present in the PrDOS prediction at K160 – P170, in addition to the disorder predicted at the M1 – H80 being a continuous block of disorder, as opposed to two distinct blocks. Further still, the presence of the solid block of disorder predicted at residues F520 – L620 is preceded by a block of disorder matching the one that can be seen in the PrDOS prediction at K480 – K500.

MFDp2 is the predictor that best matches the consensus gained across all predictors; this it is because the disorder bars have the same location along the sequence, yet the boundary for each block of disorder is narrower as shown at residues M1 – Q40, I310 – K380, and F520 – L620.

DISOclust and PONDR VSL2 have produced the two prediction most like each other; most of the solid blocks of disorder directly overlapping over each other. In DISOclust's case though, the location of disorder remains similar to that of PONDR, while the width of each solid block is larger by 5 – 20 residues in length. They both share a great degree of similarity with the general consensus predicted, which lacks the additional two blocks of disorder present in both DISOclust and PONDR at residues T420 – S440 and V640 – K660.

DisEMBL has predicted disorder that is encompassed fully by the general consensus of disorder predictors. The residues M1 – H80, L300 – V400, and F520 – L620 present with disorder, but it is not continuous between the ranges, as are 3 smaller blocks depicted in the first range, 3 in the second, and 2 in the third.

The fragments were formed to avoid regions in Adgb's MSA that presented a chaotic alignment; the interface at the end of one fragment and the beginning of another, represents a portion of the alignment where the MSA was considered disorganised. Therefore, each fragment would have displayed an organised alignment within the MSA. The space between fragment 1 and 2 coincide with the end of the first region of disorder, through residues M1 – H80, predicted as disordered in the general consensus though the first fragment is considered fully disordered by the same consensus.

Approximately 80 residues towards the end of fragment 3, is covered by disorder in the same consensus; the interface between fragment 3 and 4 falls in a portion of the alignment where the general consensus is that there is no predicted disorder within this region of Adgb's sequence. The end of fragment 4 overextends by approximately 50 residues, into an area that is considered disordered by the general consensus. Fragment 5 is situated in a portion of the sequence where the general consensus largely considers Adgb to be ordered, barring 2 separate blocks of disorder no more than 20 residues wide, which there exists no universal assent between servers thus their presence is somewhat negligible.

Allowing regions where the disorder server form a consensus of that sequence being largely disordered adds noise when deploying fold recognition techniques. To lower the chance of generating false positives, in addition to the regions E59 – G327 and W632 – E775 already believed to sharing similarities with Calpains domain IIa and III respectively the result seems to indicate that fold recognition techniques should be focused in on the region V400 – F520.

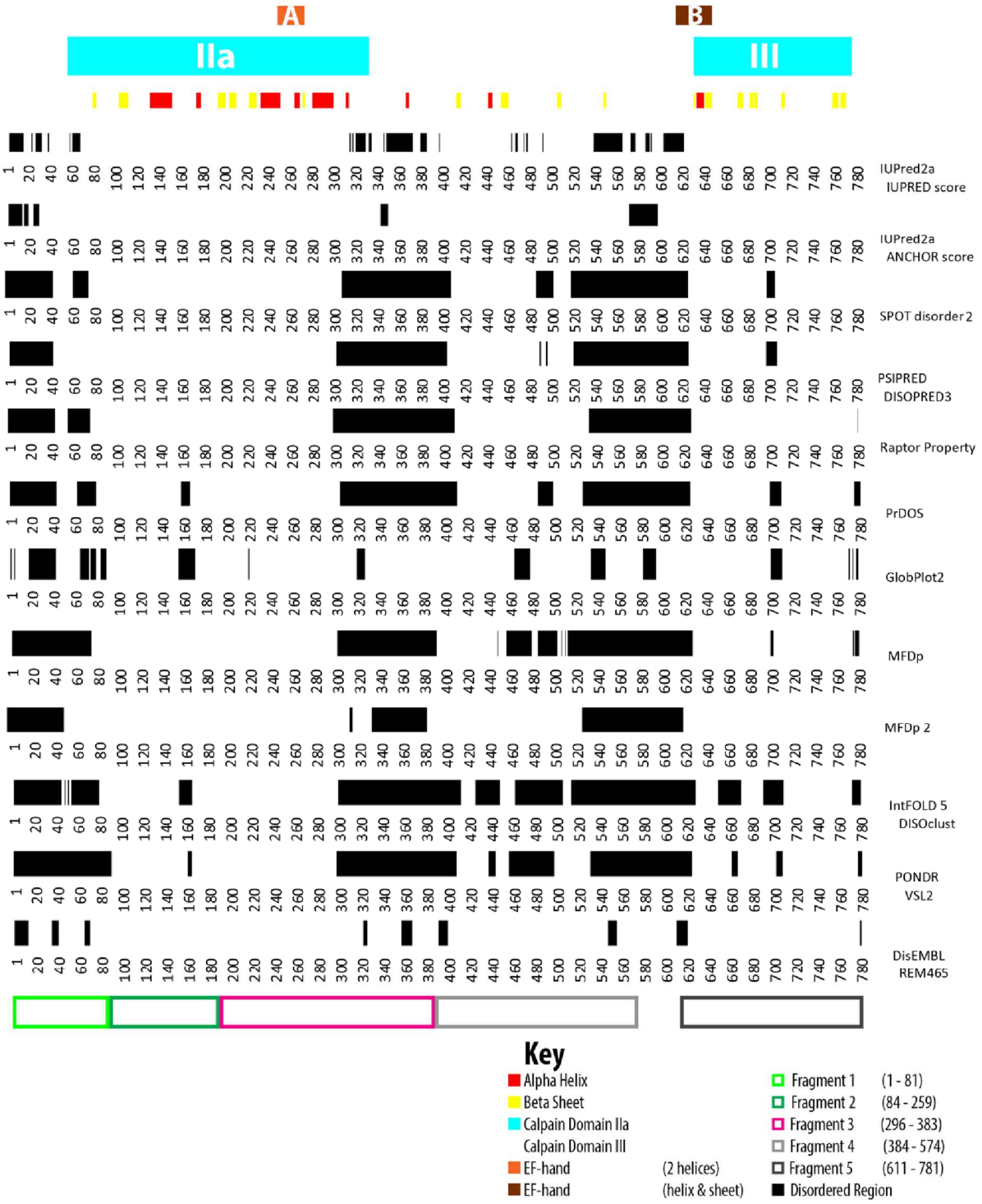

Fig. 2.1: Disorder prediction for the region before the heme domain (M1 – N781) by 12 different methods. Two EF hand labelled A and B produced by profile-profile alignments between Adgb MSA and EF hand MSA (PF00036) have been shown in orange and brown respectively. Orange EF hand (residues I250 – T270) relates to a region

that coincides with two alpha helical structures, making possible the EF hand to form a Ca2+ binding loop in the presence of a helix-loop-helix secondary structural configuration. The brown EF hand (residues E617 – I645) relates to a region that coincides with a single predicted helix within the bounds of that EF hand; the lack of a predicted helix-loop-helix conformation would be less likely. Two Calpain domains labelled IIa (residues E59 – G327) and III (residues W632 – E775) produced by profile-profile alignment between the Adgb MSA and the calpain MSA have been shown in cyan. Predicted secondary structure is shown below the calpain domains, displaying helices (red) and sheet (yellow) inferred from structure predicted by a joint consensus between JPRED and PSIPRED. Coloured rectangles with no fill have been used to denote the residue position where fragments where formed in the Adgb sequence. This was done in anticipation of disorder being present within section of the MSA that displayed chaotic alignments. Residue numbers next to fragments formed relate to Adgb sequence residue positions.

2.3.2. Region F782 – K985

Fig. 2.2 represent disorder predictions for the region F782 – K985; the format is similar to that of the disorder predicted in Fig. 2.1. The webserver predictions have generally mapped 3 areas where disorder is apparent: F782 – T790, L880 – K900, and C970 – K985. All the secondary structure predicted in this region generally lacks disorder, as the heme domain remains largely ordered. There is a reasonable overlap between the predicted alpha-helical secondary structure and the corresponding region inferred

from the permuted heme domain, but since the heme domain extends to residue L885, some of the disorder prediction methods have probably over-predicted in this region.

The IUPRED score, ANCHOR score, and SPOT disorder have near identical disorder prediction in this region, namely 2 blocks of disorder no more than 10 residues wide and a central block of disorder at residue positions L880 – K900. This central block of disorder interrupts helix H (shown in blue) at position P859 – S884 in Fig. 2.2.

DISOclust has the most similarity to the disorder predications made by IUPred and SPOT; the predicted disorder present at the end of Fig. 2.2 in DISOclust is at least twice the size of that predicted by IUPred and SPOT.

PrDOS, much like DISOclust, contains a significantly larger block of disorder towards the end of its prediction in Fig. 2.2. The major difference is that central block of disorder is significantly smaller in PrDOS than any of the other predictors discussed thus far, spanning the residues D890 – K900.

DISOpred has a similar distribution of disorder to the fore-mentioned predictors, barring the lack disorder that appears in the end of the fragment. However, its central block of disorder at position D880 – K900, appears in two parts at position approximately L885 – M895 separated by approximately 5 ordered residues, as opposed to one continuous block of disorder.

Raptor Property and DisEMBL both predicts completely ordered regions between residues T790 – Q980; only in the two termini of the heme region has disorder been predicted by the two predictors in near identical locations in the heme domain.

Globplot2 is similar to Raptor Property and DisEMBL in as much as the central portion of the sequence remains unmarred by the presence of any disorder. The two termini harbour predicted disorder in this region in the form of for two discontinuous

blocks: disorder at positions E785 – 790 and Q980 – K985, as opposed to a single continuous block of disorder at both termini produced by Raptor Property and DisEMBL.

MFDp contains two regions which it predicted as disordered at position D890 – K900, significantly smaller than that of IUPred, SPOT, and DISOclust. It also presents a second block of disorder at approximately position C970 – K985, whilst MFDp2 only depicts the single disordered region within a much narrower range of residues P780 – E785.

PONDR has a similar prediction to that generated by PrDOS and is the only one to contain a single block of disorder at approximately I930 – E935 that bisects that second stretch of ordered residues, between K900 – C970 in the general trend for disorder predicted, for 2 – 3 residues.

These results have little relevance to the fold recognition as this region corresponds to the permuted heme domain and since the insertion that corresponds to the IQ domain is disordered it is unlikely that fold recognition for this section would yield any significant results. The correspondence between the predicted alpha helices and the inferred heme alpha helices is good, but at the same time indicates that while the predictions can be reasonably reliable they are not 100% accurate and we should expect some discrepancies.

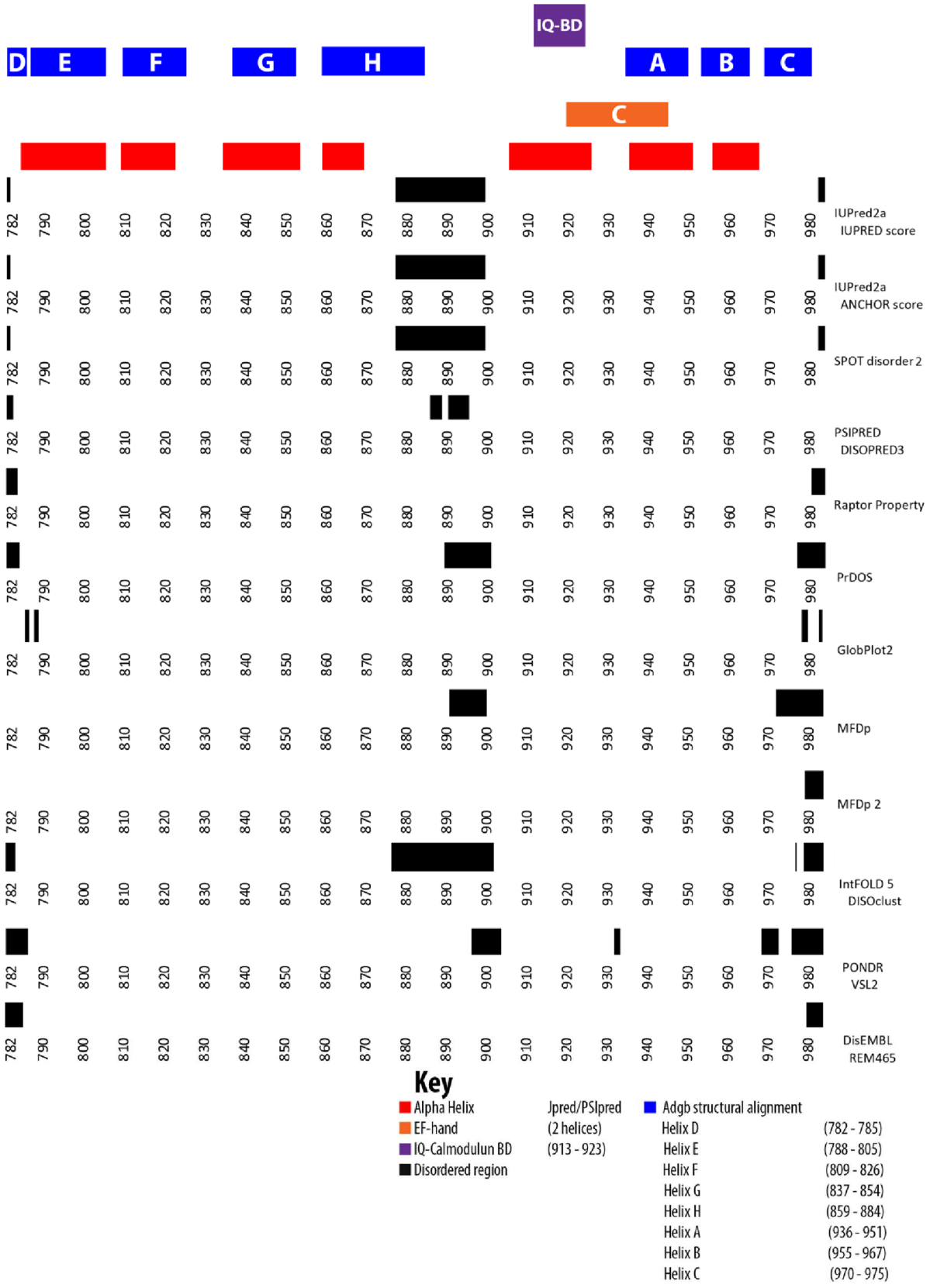

Fig. 2.2: Disorder prediction for the hemic region (F782 – K985), by 12 different methods. Disorder residues are noted by the presence of black blocks in the

corresponding residue positions; where the alignment contains no black, the structure can be assumed to be ordered. A single EF hand (residues Y921 – Q944) in this regions labelled C has been produce by profile-profile alignments between Adgb MSA and EF hand MSA (PF00036) has been shown in orange; this particular EF hand has been mapped to a region where 2 helices have been predicted, therefore it is plausible that there is he a Ca^{2+} binding loop located. The IQ-binding domain (E618 – I645) is shown in purple and alpha helices shown in blue. JPRED/PSIPRED secondary structure prediction is shown in red.

2.3.3. Region I996 – K1667

Largely, Fig. 2.3 shows the general consensus for the region I996 – K1667 across the predictors as being mostly ordered in the first half of this fragment through residues I986 – L1180. In this interval, secondary structure prediction has produced a beta-sheet rich alignment. Outside of the range I986 – L1180, there exists a great deal of variation in the amount of disorder predicted, in addition to all the alpha helices predicted in Fig. 2.3.

SPOT, DISOPRED, and Raptor Property have a great deal of similarity in their predictions; presenting 4 mainly solid blocks of disorder through approximately residues L1180 – T1240, D1280 – P1340, T1420 – S1520, and A1640 – K1660. SPOT and DISOPRED differ from this general trend by predicting a small amount of disorder that is less than 10 consecutive residues long, immediately after the block of disorder present at T1420 – S1520. DISOclust matches the general disorder trends shown by SPOT, DISOPRED, and Raptor Property at L1180 – T1240, D1280 – P1340, T1420 – S1520, and A1640 – K1660.

Additional disorder is present in DISOclust between the block V1260 – D1280 and Y1360 – T1420 in a discontinuous manner for a total of 50 residues through this region. The disorder present T1420 – S1520, overextends to residue Q1560, in addition to disorder being present at P1090 – D1110.

PrDOS has a shared similarity with the general trend, and presence disorder in L1180 – T1240, D1280 – P1340, and A1640 – K1660, but the block of disorder at T1420 – S1520 has been split into two distinct portions by a stretch of ordered resides between A1470 – G1480. Approximately 10 residues of disorder precede the disorder between T1420 – S1520 and residue 1400. The interval between S1520 – A1630 has 4 disconnected instances of predicted disorder, the largest being a block of disorder spanning 20 consecutive residues between Q1560 – V1580.

The IUPred score in this region produces far more disorder than ANCHOR. Only IUPred shares common disordered residues with the general consensus reached by SPOT, DISOPRED, and Raptor Property in the interval D1280 – P1340, as it overextends into Y1360. The region before this in both predictors is shown as being ordered, barring the disorder present within IUPred at L1230 – T1240. Between residues N1370 – D1600 IUPred predicts a number of intermittent blocks of disorder, separated by 9 separate stretches of ordered consecutive residues. The largest stretches of these ordered residues interrupting the predicted disorder is approximately 20 consecutive residues at position L1530 – D1550.

However, in the prediction produced by ANCHOR between G1290 – D1550, the disorder predicted is discontinuous in the same way as that produced by IUPred. The key difference being the ordered residues that interrupts the disorder predicted between

G1290 – D1550 are a much longer set of consecutive residues, the largest being approximately 40 residues in position H1305 – P1340.

MFDp and MFDp2 produce the largest and second largest amount of consecutive disordered residues of all the predictors used, through residues L1180 – K1660; MFDp 2 is interrupted by ordered residues in V1260 – S1270. MFDp differs from MFDp2 in the 30 residues preceding L1180 are also predicted as disordered. Three narrow discontinuous strips of disorder at P1130 – V1150 with a width of no-more than 5 consecutive residues in length each strip, precede the solid block of disorder present in MFDp at S1160 – K1660; this is in addition the disorder present at D1100 and at the start of the fragment in this prediction.

PONDR produces the third largest continuous strip of disorder at resides G1400 – K1660. The 10 residues predicted as disordered at D1380 are preceded by a block of disorder at L1180 – T1240 common to 4 other predictors, which in turn is preceded by further discontinuous disorder at D1160 – L1180.

DisEMBL and GlobPlot both produce the most intermittent disorder predictions, though not deviating outside the trend of disorder predicted at 1180 – 1240, 1280 – 1340, T1420 – S1520, and A1640 – K1660; only GlobPlot predicts a single set of residues outside these ranges as disordered, at S1190 – C1100.

Significantly, residues P1000 – L1180 are largely predicted to be ordered. These are prime candidates for fold recognition. The secondary structure prediction suggests that we are looking for a largely beta sheet structure. The remaining regions are largely predicted to be disordered. However, it may be possible to use fold recognition methods to investigate Y1360 – T1420 and S1520 – A1640.

Fragment 6 is present in sequence space that is largely considered ordered by the general consensus, save for the last 20 residues. Fragments 7, 9 11 and 12 would have been considered to have an organised region within the alignment but are fully disordered according to the general consensus. The interspace between fragment 7 and 8, in addition to the first 20 residues in fragment 8 are predicted as disordered.

Fragments 10 and 13 reside in portions of the sequence where no real consensus can be made of whether or not there is disorder generally, as half of the predictors predict disorder despite an organised alignment.

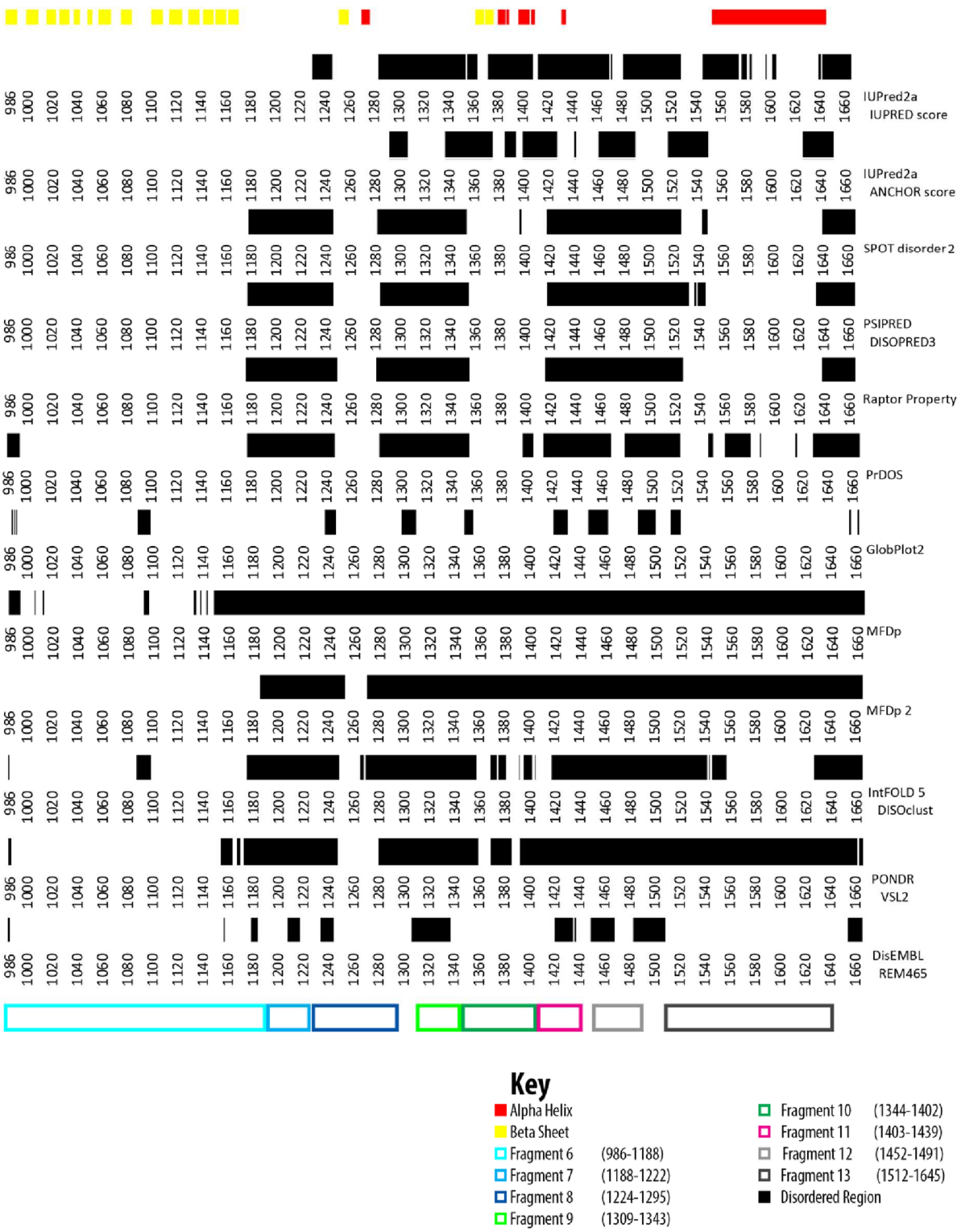

Fig. 2.3: Disorder prediction for the post heme region (I996 – K1667), by 12 different methods. Similarly to Fig. 2.1, the predicted secondary prediction is given at the top, helix (red) and sheets (yellow), inferred from a consensus from JPRED and

PSIPRED. Coloured rectangles with no fill are used to denote the Adgb sequence fragments used in fold recognition (see chapter 3).

EF hands

The key elements of a canonical EF-hand are 12 highly conserved residues – **DxDGDGxISxEEF** – with residues 1(X), 3(Y), 5(Z), 7(-X), 9(-Y), and 12(-Z) (where Z can be D or E, X can be X, Y can be X) functioning as binding residues in the chelation of Ca^{2+}. These residues help to form a Ca^{2+} binding loop between two alpha helices. The most conserved of these residues is present at position 12(-Z), as either glutamate (E) or aspartate (D), both of which function as a bidentate ligands, providing the two oxygens required for Ca^{2+} ligation; residues 1(X) and 3(Y) are also highly conserved as D/E. (Kretsinger, 1997; Zhou, Xue and Yang, 2013). A number of profile-profile alignments were used to map the EF hand MSA to a region of the Adgb MSA. The results of this exercise to identify EF-hands within the Adgb MSA are shown in Figs. 2.4 – 2.6.

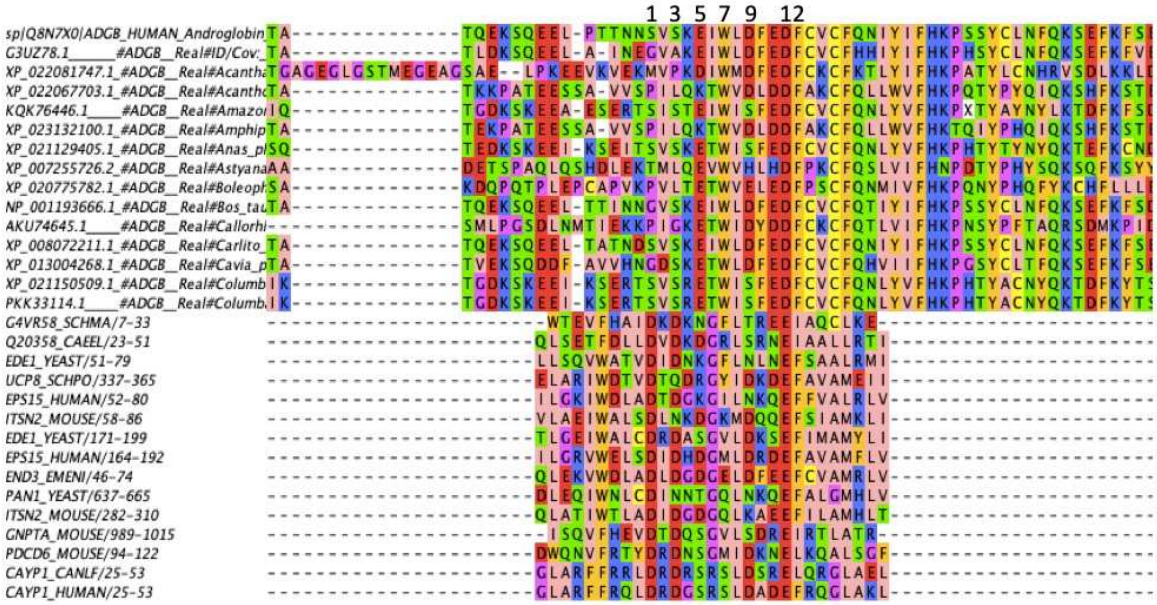

Fig. 2.4: The first Profile alignment between EF hand and Androglobin MSA.

This alignment was obtained by using the seed alignment EF hand 1 - PF00036 from the Pfam database; this aligned onto the Adgb MSA at residues E617 – I645.

2.3.4. EF hand 1 region E617 – I645

Fig. 2.4 shows the first EF hand to be mapped; the EF hand aligned with residue positions E617 – I645 of the Adgb sequence, by a profile-profile alignment between Adgb MSA and EF hand 1 (PF00036) MSA. The predicted EF hand immediately precedes the calpain domain III in Fig. 2.1, predicted to be running through residue positions W632 – E775. This EF hand aligns with an Adgb region that contains within its bounds a helix immediately flanked by two beta sheets either side, as predicted by a consensus formed between JPRED and PSIPRED, shown in Fig. 2.1 by the brown EF hand block labelled B. The first half of this EF hand is located in a portion of the Adgb sequence were 10 out of 12 of the predictors expect the protein to be disordered. However, as the border

between order and disorder is not precisely predicted, this may not be a serious problem.

Three glutamate residues in Fig. 2.4 within the top 3 sequences, can be seen aligning to 3 aspartate residues present within the bottom 3 sequences; this occurs in the 9th residue from the end of the EF hand sequences. This serves as evidence for conservative substitution of residues at that highly conserved 12th (-Z) residue.

The consensus formed for EF hand 1 (PF00036) in this region of the profile-profile alignment is when the 12th (-Z) residue is at the end of binding loop residues is **D**K**D**G**DG**GKIDF**E**E for which residues G7(-X) and I9(-Y) are not conserved when this consensus is compared to the canonical EF hand.

The glutamate (E) in the Adgb sequences and aspartate (D) in the EF hand sequences, functions as the end of the frame of the conserved binding residues, results in the Ca^{2+} binding loop being formed by the residues S626 – D637 of the Adgb sequence. The consensus sequence SVSK**E**TW**I**DF**ED** within the Adgb MSA forms the EF hand binding loop, potentially mediating Ca^{2+} binding, despite residues S1(X), S3(Y), W7(-X), and D9(-Y) deviating from the canonical EF hand motif, thus failing to be conserved within Adgb's sequence. Nevertheless, there is probably sufficient functionality to bind Ca^{2+}. Despite this however, the mere fact that this motif has been mapped onto a region of the sequence where the JPRED/PSIPRED consensus has not predicted the binding loop to reside between two helices, it seems unlikely that this stretch of amino acids will indeed bind Ca^{2+}.

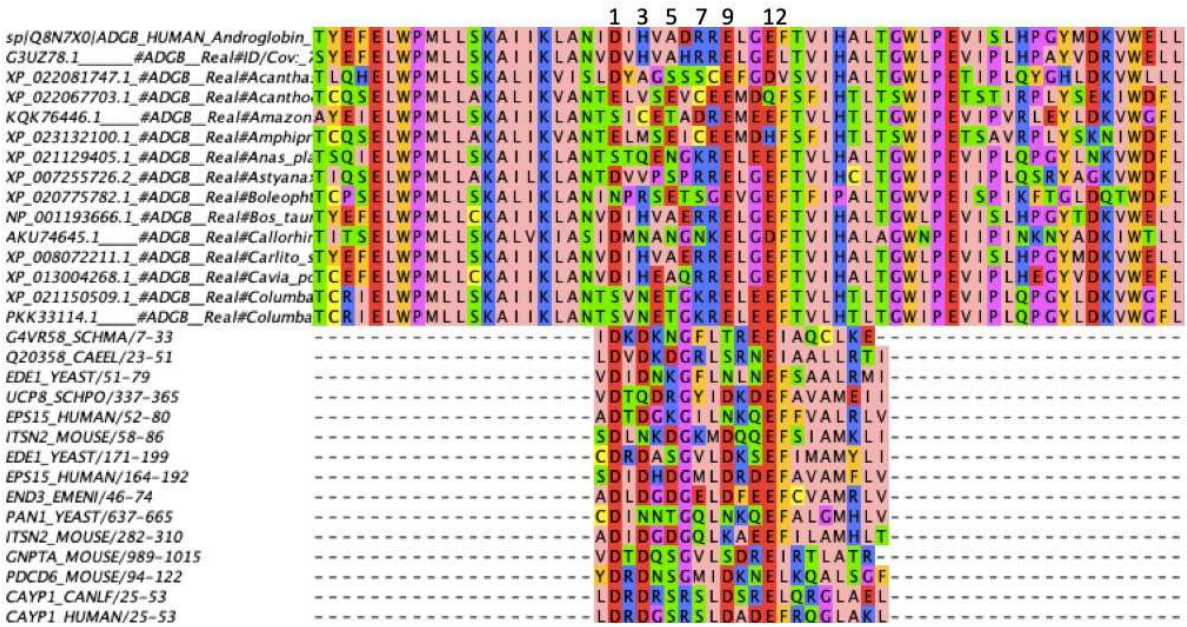

Fig. 2.5: The second Profile alignment between EF hand and Androglobin MSA.

This alignment was obtained by using the seed file EF hand 1 - PF00036 from the Pfam database. It aligned to the Adgb MSA at residues I250 – T270 on Adgb's sequence.

2.3.5. EF hand 2 region I250 – T270

Fig. 2.5 shows the resulting profile-profile alignment between EF hand 1 – PF00036 represented in Fig. 2.1 by the orange block labelled A, mapped though residue positions I250 – T270. It overlaps with the second half of calpain domain IIa, which has been residues E59 – G327. This EF hand encompasses the space between the final few residues of the preceding alpha helix, approximately 10 residue long, and a short alpha helix immediately after. There is a consensus across all the disorder predictors used in this region formed, that this particular EF hand occupies a sequence region expected to be ordered.

D1(X) is conserved across both Adgb and EF hand sequences in Fig. 2.5, this occurs in the 2nd residue from the start of the EF hand sequences; it stands to reason that this residue would function as the start of the frame of 12 conserved binding residues. Overall consensus is **D**IHVA**G**RRELG**E** in this region of the Adgb MSA. We note that residues H3(Y), A5(Z), R7(X), and E9(-Y) are not conserved within the general consensus of the present in the MSA form by the profile-profile alignment. Therefore, the putative Ca^{2+} binding loop is formed by the residues D251 – E262; Adgb's sequence through this loop is **D**IHVAHRRELG**E**, only conserving D1(X) and E12(-Z) from EF hand consensus sequence. However, despite the lack of conservation in residues, this EF hand maps onto region where JPRED and PSIPRED reach a consensus of there being two alpha helices where in the interspace between helices, this motif may therefore form a Ca^{2+} binding loop.

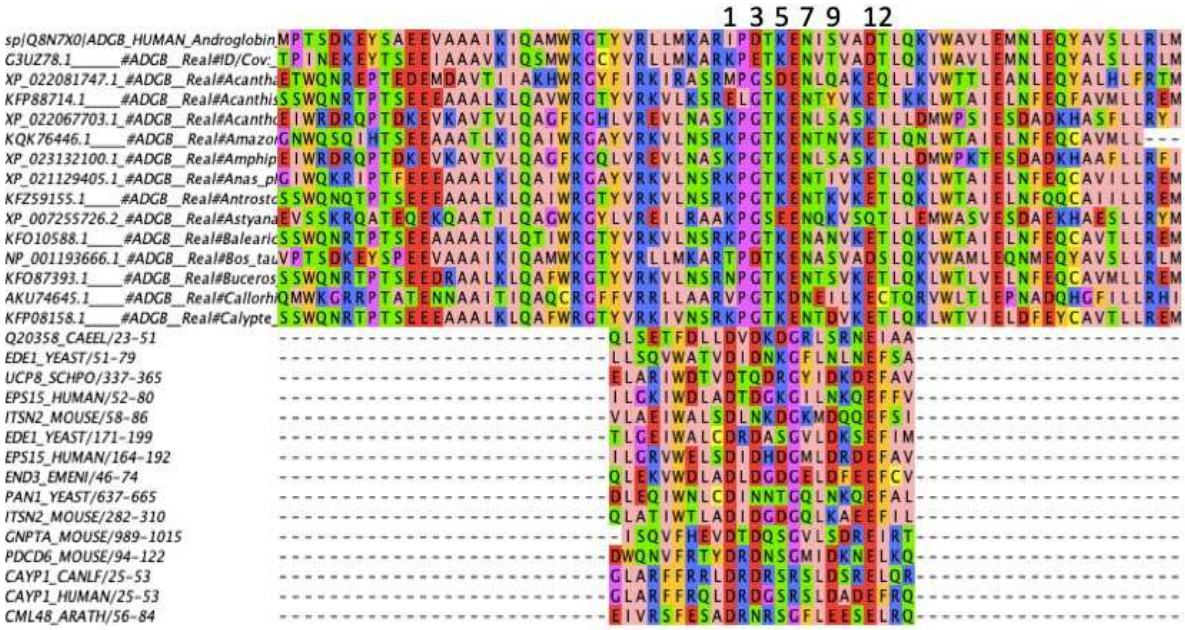

Fig. 2.6: The third Profile alignment between EF hand and Androglobin MSA.

This alignment was gained by using the seed file EF hand 1 - PF00036 from the Pfam database, and has aligned on to the Adgb MSA at residues Y921 – Q944 on the Adgb sequence

2.3.6. EF hand 3 region Y921 – Q944

The EF hand in Fig. 2.6 shares characteristics similar to the alignment in Fig. 2.4, in that the residue 12 (-Z) is conserved throughout the Adgb MSA, signalling the end of the frame of 12 conserved binding residues. An EF hand binding loop was mapped through residues I930 – D941 to in this region, formed by the sequence IP**D**TKEN**I**SVA**D**T within the Adgb sequence, resulting in D3 (Y) being the only other contact residue being conserved in Adgb sequence.

The alignment in Fig. 2.6 overlaps with the end of the 5th helix and the beginning of the 6th helix in the JPRED/PSIPRED secondary structure prediction. The region in which

this overlap is evident, is predominantly predicted as an ordered region close to the Adgb heme domain when considering the general consensus present in Fig. 2.2. In addition, this predicted EF hand from PF00036 overlaps the Calmodulin binding domain present within the same region, and the 5th helix from JPRED/PSIPRED, present in the Adgb heme structural alignment.

The fourth Profile alignment between EF hand and Androglobin MSA did not yield satisfactory results as the PF00036 part of the MSA became fragmented, suggesting that Adgb contains at most 3 EF hands.

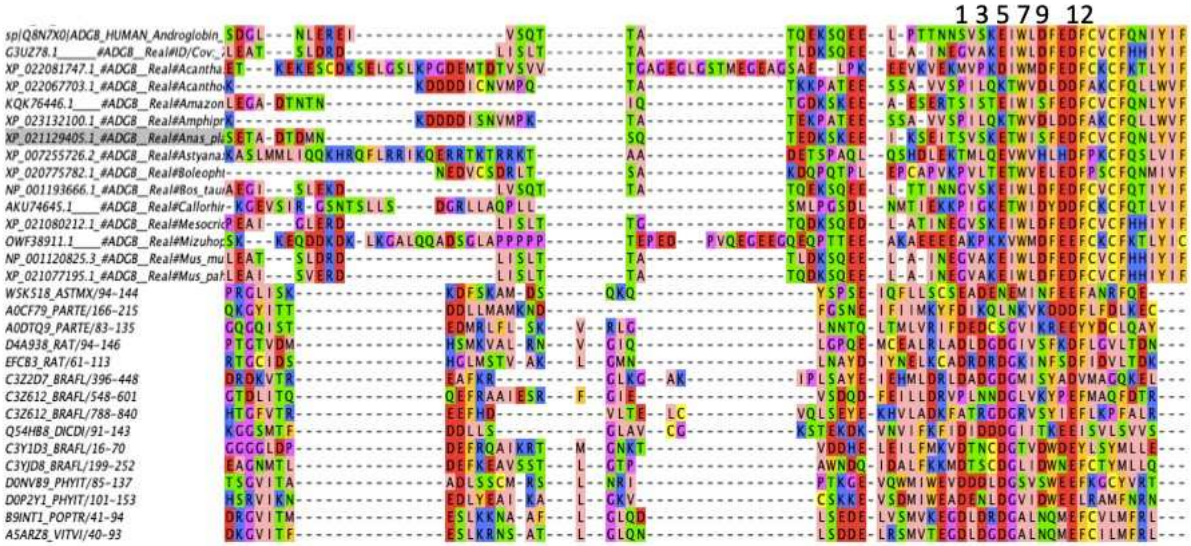

Fig. 2.7: The fourth Profile alignment between EF hand and Androglobin MSA.

This alignment was gained by using the seed file EF hand 8 (PF13833) from the Pfam database, and has aligned on to Adgb's MSA at residues D595 – I645 on Adgb's sequence

2.3.7. EF hand 1 region D595 – I645

Fig. 2.7 has produced an alignment in the same sequence space as that seen in Fig. 2.4, Identifying the same binding loop through residues S626 – D637, where residues E5 (Z) and D12 (-Z); S1(X), S3(Y), W7(-X), and D9(-Y) deviated front the canonical EF hand motif.

Whilst gaining an alignment in identical sequence spaces in both Figs.2.4 and 2.7 increases the reliability of the alignment, it is impossible to prove or disprove that these two alignments in addition to the alignments shown in Figs. 2.5 and 2.6, are genuine EF hands in the absence of any experimental data. This would entail expressing Adgb in the presence of Ca^{2+} in an effort to see if binding occurs. This would then be followed up by mutating those key residues 1(X), 3(Y), 5(Z), 7(-X), 9(-Y), and 12(-Z) to see if binding ceases when these residues are no longer expressed.

2.4. Conclusion

All the disorder predictions that have been generated for the heme domain function as a control; the structure is already known to a reasonable extent as the heme domain has already been expressed *in vitro*. Observations that may be inferred from Fig. 2.2 is that helices D and C are absent from the secondary structure prediction consensus, formed between JPRED/PSIPRED, suggesting these predictions are giving false negatives i.e. no structure where there should be structure. However, the C and D helices are short,

so this is not a major error. Furthermore, the consensus formed from the disorder predictors shows the disorder dividing the heme fragment into 2 continuous strips of ordered protein. This occurs at the interface spanning helices H and A. In this region, helix H is shown to extend into the disordered region, as predicted by a consensus from IUPred, ANCHOR, SPOT2, and DISOclust. The corresponding structure to helix H in the JPRED/PSIPRED consensus, is shown to end approximately 10 residues before the predicted disorder.

The CaM binding domain is shown in the same place as the 5th helix in the JPRED/PSIPRED consensus, immediately downstream of the disorder. This may not affect reliability as the CaM binding domain may be alpha helical in structure. A second protein-protein interaction motif, namely EF hand C, is present in a binding loop formed at the interface between helices 5 and 6 on the JPRED/PSIPRED consensus. Thus, the two elements necessary for a molecular recognition features (MoRFs) within the protein are present; a structural feature that functions as the site for protein-protein interaction in close proximity of disorder (Vacic *et al.*, 2007).

The disorder predicted in Fig. 2.1 shows the consensus partitioning the sequences into three major areas, K60 – P320, S390 – F520, and L620 – N781, where we expect the protein to be natively folded. Profile-profile alignments have mapped calpain domains IIa and III through the residue E59 – G327 and T632 – E775 respectively. This leaves the region S390 – F520 as the only region in the pre-heme sequence we expect to yield relevant and related structures when using fold recognition techniques.

Two potential EF hand motifs were also mapped to this region in the sequence, as shown in Fig. 2.1. If both motifs are found to be able to form 2 separate Ca^{2+} binding loops, it will suggest that they would be able regulate and control the proteolytic activity

of the calpain domains present. In addition, they are both mapped near the regions P320 – V400 and F520 – L620, where the consensus is that the sequence is disordered. Only a single EF hand motif may be able to form a binding loop, due to EF hand B in Fig. 2.4 and 2.7 lacking the typical helix-loop-helix topology characteristic of canonical binding loops. Whether both EF hands are found to be functional or whether only EF hand A in Fig. 2.5 is functional, the binding would happen in such close proximity to predicted disorder that it would be characteristic of a MoRF. This means ligand binding or protein-protein interactions, depending on whether or not the EF hands will function as a pair, may propagate a disorder-to-order transition in disordered regions.

The results in Fig. 2.3 suggest that the only region where a clear consensus exists for the sequence being ordered is P1000 – L1180. Whilst the rest of the sequence is predicted as largely disordered, there remains pockets where the sequence is likely to be ordered through the residues Y1360 – T1420 and especially S1520 – A1640. The long >30 residue alpha helix in S1520 – A1640 makes this region particularly interesting in respect to fold recognition, as this is the longest continuous alpha helix that has been predicted anywhere in the sequence structure. The structure present in the regions through the ordered residues T1240 – V1280 and Y1360 – T1420, and the disorder flanking either side of these ordered regions, makes them possible sites where MoRFs may be present. However, this would depend on some sort of molecular interaction to propagate a disorder-to-order transition.

Chapter 3: Fold recognition in androglobin

3.1. Introduction

Despite comparative genomics studies having elucidated the presence of a 5th member in the globin superfamily of protein, there is still remains a gap in knowledge in terms of a comprehensive, and complete structure of Androglobin (Adgb).

Outside harbouring an N-Terminal calpain-like domain, a heme domain, and an IQ calmodulin-binding domain (Hoogewijs *et al.*, 2012), not much is known about the 1667 residue long chimeric protein. However, the site of primary expression within the testes, and its rapid auto-oxidation, suggested a primary cellular function other than oxygen transport. When Adgb's cellular function was determined to primarily be as a mediator of the nitric acid cycle, the notion that globins generally function as a gaseous exchange surface was drawn into contention.

As result of this gap in knowledge, and the inability to express the protein in its totality outside the heme domain, it has become imperative to characterise the topology displayed by Adgb by fold recognition. Moreover, modelling the protein by computational means could help guide current efforts to express and purify the protein *in vitro.*

3.2. Methods

A multiple sequence alignment (MSA) containing Human Adgb, and a number orthologs, was fragmented into 16 separate FASTA files within the MSA editing program Jalview (Waterhouse *et al.*, 2009). The fragments only contained a portion of sequence which were determined to contain well-aligned regions of sequence, including conserved motifs, across all the sequences in the MSA. In other words, the segments of the MSA that were deemed ordered, as shown in Fig. 3.1a, would form the query sequences that would then create the FASTA files. The messy segments, shown in Fig. 3.1b, would form the interface between query sequence fragments. In this way, the chaotic portions of the Adgb MSA would be excluded when secondary structure prediction and fold recognition techniques were deployed, in order to elucidate the nature of the topology present in the Adgb sequence.

The FASTA files were created from the sequence limits present in Table 3.1. this was achieved in an identical fashion to the sequence fragments generated for the disorder analysis in Chapter 2. The residues in the human Adgb sequence alone were highlighted, and the selection was then outputted into a textbox inside Jalview in FASTA format. Sequence was copied, pasted and saved into a file inside a text editor. Replacement of the file extension from the default extension used to '.FASTA' or '.fa' is required to allow the web-servers to detect the file. This was repeated for the remains sequence fragments.

Once all 16 sequence fragments were used to create individual FASTA files. Each of these files in turn was then submitted into each of the following secondary structure prediction servers:

- JPRED (Drozdetskiy *et al.*, 2015)
- PSIPRED (Buchan and Jones, 2019)
- RaptorX Property (Wang *et al.*, 2016)

The prediction gained from each of these servers was produced using the default parameters available in each server. The secondary structure prediction methods used have produced graphical alignments in both JPRED and PSIPRED. Each server will map secondary structure on to the query sequence at a residue level, in addition to generating a confidence level between 0 – 9 associated with the predicted structure at a residue level. Aligning the structure predictions produced in each of the servers to one another, required the resizing of each individual output. This was achieved in a graphical editing software. RaptorX Property yielded the secondary structural alignment as a '.txt' file. Therefore, it made it necessary to colour the text, which represented the Adgb sequence in a given fragment, as a means of indicating the nature of structure that had been mapped onto each specific residue using Excel.

Using of the 'Courier' font ensured consistency in not only producing uniform spacing between characters, but in also producing uniform width in each of the characters, which allowed for accurate alignment. Once coloured, it was possible to export the alignment as an image and align it with the graphical output from JPRED and PSIPRED already aligned in the graphical editing software.

Primarily, fold recognition techniques will serve to provide suggestions of the sequence fragments molecular function, based on the tertiary structure present, in the

absence of the experimental data concerning that particular query sequences. Secondly, fold recognition is also able to guide understanding on how these secondary structures are liable to aggregate and form super secondary structures such as, though not limited to, beta-barrels, Greek-key motifs, or Rossmann folds. Finally, some of the fold recognition techniques used, namely PHYRE2 and I-TASSER, will also generate secondary structure predictions. The secondary structure prediction produced by PHYRE2 and I-TASSER were used in combination with results from JPRED, PSIPRED, and RaptorX Property. The fold recognition matches produced against known protein folds, was achieved using the default parameters present in the following servers:

- PHYRE2 (Kelley *et al.*, 2015)
- RaptorX (Källberg *et al.*, 2014)
- I-TASSER (Yang and Zhang, 2015)
- HHPRED(Zimmermann *et al.*, 2018)

Each of these methods will give a measure of confidence in the reliability of the structure produced. This measure will be given globally, across the length of the sequence fragment that the threading template has mapped to, as opposed to locally at a residue level. Each of the fold recognition methods being used will then rank each threading templates being used according the confidence produced by each model. HHPRED is the only fold recognition method that will solely produce a graphical alignment between the template sequence and query sequence, similar to the secondary structure prediction alignments. PHYRE2, RaptorX, and I-TASSER will all produce a '.zip' file contain all the alignment information and a number of PDB files.

Once it is retrieved from each server, it will be possible to open top ranking threading template from each server in the same PyMoL session. It is possible to superimpose the models on top of each other, by aligning all the models onto a single chosen model, using the 'align' function located under the 'actions' menu.

Though HHPRED does not produce PDB files for the modelled structure, it is possible to introduce the threading template used in HHPRED into this same PyMoL session where the structures from PHYRE2, RaptorX, and I-TASSER have already been aligned. Using the 'fetch' command within PyMoL, followed by the PDB code of the threading template present in HHPRED's graphical alignment, the whole PDB can be retrieved. As the query sequence only aligned to a portion of the threading template, the portion of the template that provided a match will need to be excised out the whole structure. This is achieved by identifying the sequence limits on the template where the query sequence produced a match, in the sequence window of PyMoL. Once the length of the match is highlighted the selection can be extracted using from the PDB using the 'extract object' function under the 'actions' menu. This will create a separate object from the original PDB file, containing only the portion of the PDB that was used as threading template by HHPRED, provided the appropriate sequence limits were identified within the threading templates PDB.

3.3. Results and Discussion

Due to the fact that fold recognition techniques would struggle to identify any potential homologous protein folds within an intrinsically disordered region (IDR), as it would inherently lack any discernible structure, the Adgb sequence was fragmented to remove any potential disorder.

Firstly, this was achieved in Jalview by considering the portions of the MSA that appeared to have a disorganised and chaotic alignment, when human Adgb was aligned with other homologous proteins. As structure is conserved more stringently than sequence across homologues, chaotic alignments is an indication of the lack of conservation. This is in addition to being a sign of the possible absence of secondary structure in that given region, because of the direct relationship between structure and function. Therefore Adgb's protein sequence was fragmented in such a way that organised portions of the MSA shown in Fig. 3.3b, which are likely to contain structure integral to the proteins function, remained intact whilst excising the portions of Adgb's sequence that produced disorganised alignments shown in Fig. 3.3a within the MSA in order to produce the fragments present in Table 3.1 in such a way that the sequence between fragments presented a chaotic alignment.

A second set of fragments were then formed according to the consensus positions of the 12 disorder prediction servers. Adgb's sequence is likely to contain disordered residues and so it seemed appropriate to remove these before deploying fold recognition techniques as the disorder would most likely add noise to the predictions. The resultant fragments are given in Table 3.2.

Table 3.1: MSA fragments. Residue positions used to form fragments, as inferred from the multiple sequence alignment

Fragment	**1**	**2**	**3**	**4**	**5**	**6**
Fig. no.	2	3	4	5	6	7
Adgb res.	**M1_F81**	**D84_L295**	**P296_F383**	**K384_S574**	**T612_N781**	**I986_I1188**

7	8	9	10	11	12	13
8a	8b	9a	9b	10a	10b	11
I1188_I1123	**Q1224_E1295**	**E1309_E1343**	**P1344P_A1402**	**I1403_E1440**	**K1452_E1491**	**I1512_A1644**

Table 3.2: Disorder fragments. Residue positions used to form fragment, as inferred from the disorder predictions (see Chapter 2)

Fragment	**14**	**15**	**16**	**17**	**18**
Fig. no.	12	13	14	15	16
Adgb res.	**H80_L300**	**S390_F520**	**F782_D890**	**K900_Q980**	**L1540_E1650**

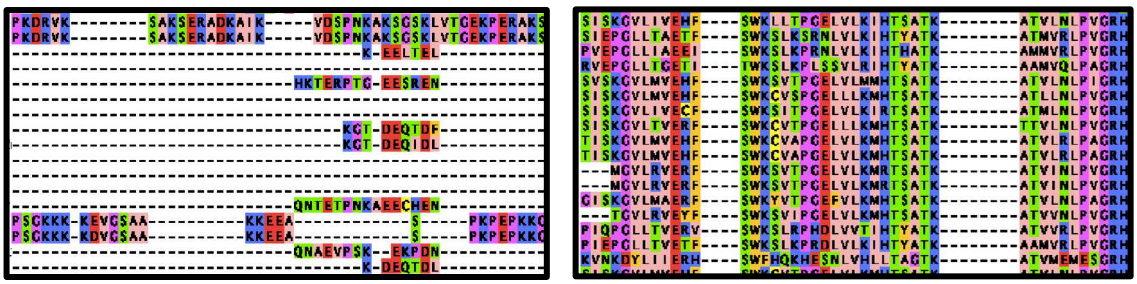

Fig. 3.1: disordered and ordered regions in MSA. (a) Disordered, chaotic alignment vs (b) Ordered, highly conserved

3.3.1. Multiple sequence fragments

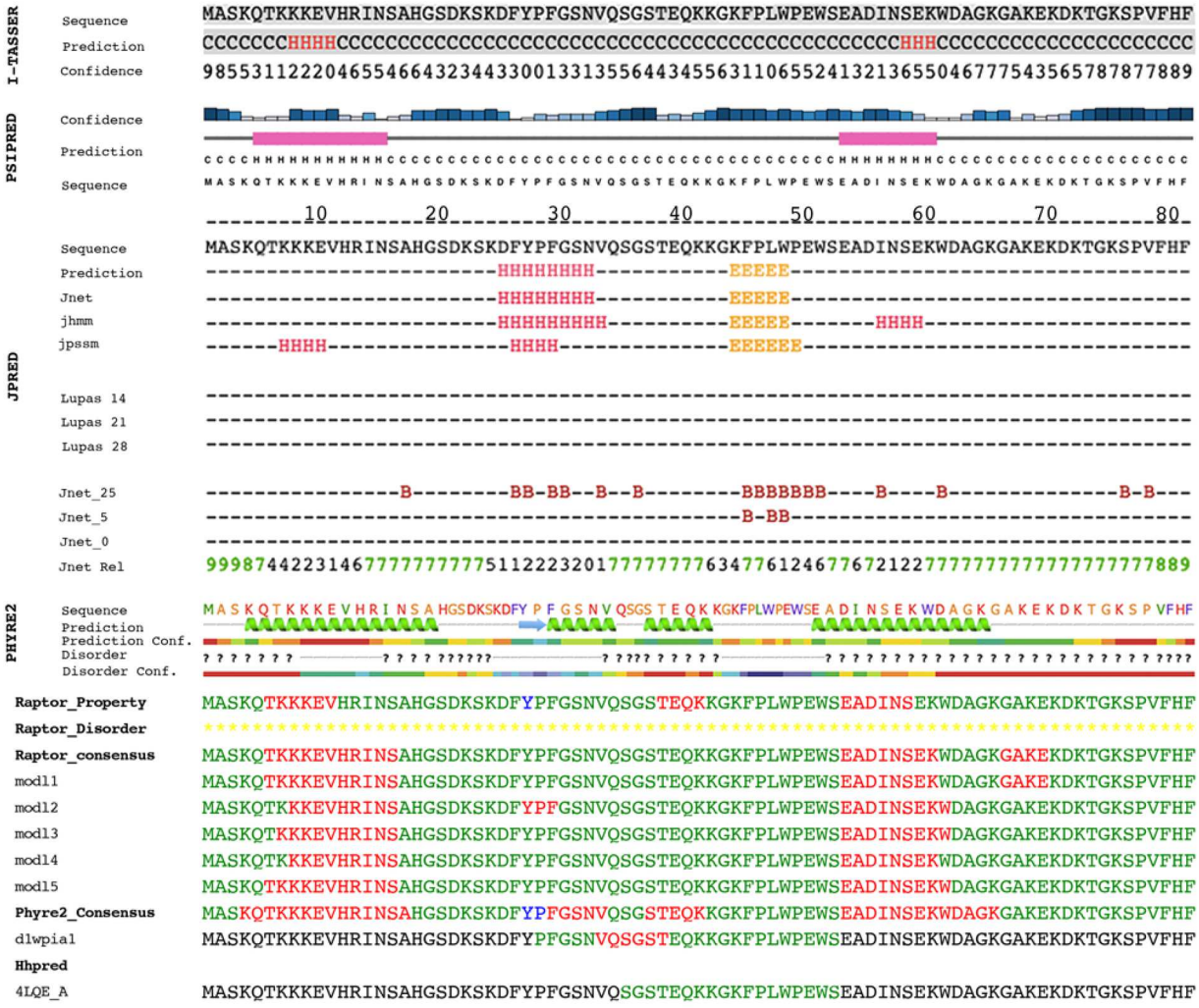

Fig. 3.2: Secondary structures, and results of fold recognition for fragment M1 – F81. I-TASSER, Raptor X (Raptor_consensus), and PHYRE2 all produce PDB files, which will also be analysed as part of their fold recognition analysis. JPRED, PSIPRED, RaptorX_Property, HHPRED all function to simply detect the presence secondary structure within the fragments without attempting to assemble any 3D structures. JPRED colouring red=helix, yellow=sheet, and PSIPRED colouring pink=helix, yellow sheet, both also give a measure of confidence at every residue. Alignments presented below the PHYRE2 graphic were aligned within Excel. PSIPRED generates blue bars with their colour

and size as a scale, with a large blue bar equating to high confidence, and a pale, small bar equating to low confidence. JPRED expresses confidence as an integer, with 0 (low confidence) through to 9 (high confidence).

3.3.1.1.MSA Fragment 1: M1 – F81

PSIPPRED, and I-TASSER both predicted helices at the two opposing ends of the fragment. JPRED's simple prediction (top line) displays a helix, and sheet in a section of the sequence located between both the helices predicted by PSIPRED, and I-TASSER. PHYRE2 has largely coiled structure with 4 helices, and a single sheet preceding the second helix predicted. There is also a large amount of disorder present within the PHYRE2 prediction, with a lot of the predicted structures also expected to fall into regions already predicted as disordered. Barring the coil between helix 3 and 4, only the central mass of the first helix and the second helix immediately preceded by the beta-sheet, have been predicted as ordered within this first fragment formed. RaptorX_Property predicted 3 helices, and sheet in very similar positions to those displayed by PHYRE2 but differing in length. The measure of disorder within the RaptorX_Property alignment has depicted the full length of this fragment as disordered. The models produced by RaptorX largely correlate, in terms of secondary structure, with the results collected from PSIPRED and I-TASSER. They exhibited two helices at opposing ends of the fragment. Modl1 however, contains a 4 amino acid helix after the second helix which is carried over into the consensus model.

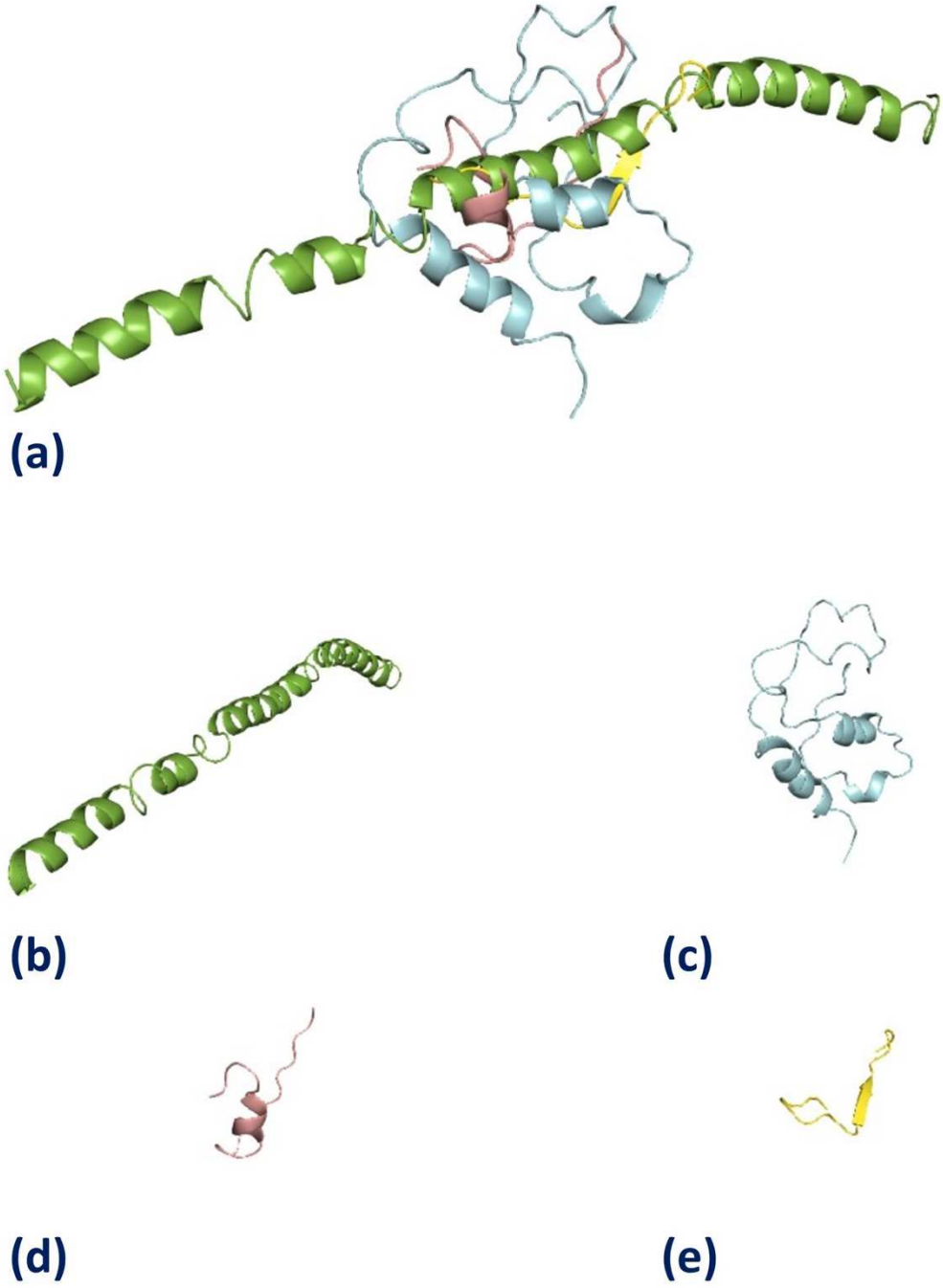

Fig. 3.3: Models produced in the region M1 – F81 of the Androglobin sequence.

(a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 3J9A. (c) final model produced by RaptorX using threading template Modl1. (d) threading template 1WPI used by PHYRE2. (e) threading template 4QLE used by HHPRED.

Fig. 3.3b contains the structure 3J9A (smallest capsid protein for the herpes simplex virus). This was the top ranked threading template in I-TASSER, producing a Z score 1.45. Typically, Z-scores greater than 1 correlate with a good threading alignment between the template and the query sequence. In addition, the sequence coverage offered by this template of 96% increases the significance of the threading template through the region.

The final model covered the whole length of the query sequence with a C score of -3.11. Generally, C scores greater than -1.5 are considered good predictions, making the C score generated for the final model in this case was particularly low, meaning I-TASSER lacks confidence in the alignment produced. This model produced a TM Score of 0.36±0.12. TM score is measure of the accuracy, of a modelled protein when compared to the known native structure of the threading template. A model that exhibits values greater than 0.5 are generally considered to have good topology, with values less than 0.17 being considered to have randomly generated similarity when compared to the threading template. This means that the final model produced, marginally fell short of the threshold for correct topology.

Fig. 3.3c contains the structure Modl1, which was used as a threading template used mapping to the full length of the sequence fragment, providing 100% sequence coverage in this region. Modl1 has a P value of 0. P values <0.05, which are usually associated with moderate degree of confidence, as 1/20 randomly generated and incorrect model would generate scores higher than the modelled protein. P values <0.01, for the same reason, would correlate to high confidence. This is as an incorrect model has a 1/100 probability of producing similar scores. P values <0.1 correlate with low confidence in the models generated, as 1/10 models would generate similar scores. The

P value observed for the alignment between the query sequence and the template Modl1, suggests that it is not possible for a randomly generated and incorrect model to produce a higher score. When the P value is coupled with the lack of PDB code, they suggest that this model was folded by RaptorX *ab initio* in the absence of experientially solved crystal structures.

When Modl1 was used as a threading template by RaptorX, the alignment generated a score of 4. The best alignment score for a model in this region produced by RaptorX would equate to full length of the domain. A moderate alignment would be approximately half the length of the domain, whilst a bad alignment score would equate to 0. This means that modl1 has a particularly bad alignment score considering this region is 81 residues in length. The unnormalized Global Distance Test (uGDT) must be considered in conjunction with the alignment score to get a true idea of the quality of the model being generated. A model that produces a uGDT greater than 50, for a sequence greater than 100 residues in length, is considered a high-quality model by RaptorX. However, for sequences less than 100 residues in length as is the case for this sequence fragment, the normalised GDT value is a more appropriate indicator of the quality of the model. Applying the same threshold of values greater than 50 to indicate a good quality model. As aforementioned, the domain is 81 residues in length and the GDT values predicated for this threading template of 65 this model can be considered a high-quality model, despite the P value generated.

Fig. 3.3d contains the structure 1WPI (alpha helical protein fold belongs to thioredoxin family of proteins). This template provided PHYRE2 with a threading template for the region P28 – W51, equating to 29% sequence coverage. The alignment gained a poor confidence score of 6.4%, however, in order to be confident that the

modelled protein has adopted an appropriate fold, it needs to have produced a confidence in excess of 90%. Below this value the only inferences that can be made is of the sequences secondary structure alone. Despite the low sequence coverage and low confidence associated with the alignment, the number of identical residues in identical positions was 29%. The lack of sequence coverage and the fact this model displays the least amount secondary structure depicted by fold recognition techniques, fuels the conclusion that this model lacks biological significance in the context of this sequence fragment.

Fig. 3.3e contains only a small portion of the structure 4LQE (MepB – DNA binding protein), mapping to the sequence through the residues S35 – S52. The alignment provided 21% sequence coverage of the sequence fragment, when 4QLE was used as a threading template by HHPRED. Though the structure in question is a single beta-sheet strand, HHPRED only mapped a loop conformation to the query sequence. Despite the template having a moderately high sequence similarity, with 22% of the residues in the query sequence either being a conserved substitution or sharing identical residues at the same position in the template sequence, the model generated a poor probability of 6.6%. The probability was judged as poor as homology may only be truly inferred if the model displays >95% probability, however, a model may still be significant if it has generated probabilities of 50% - 35%.

Of the 4 methods, only the RaptorX model in Fig. 3.3c gives a plausible result with some degree of confidence, this is the only fold recognition technique that was able to provide full sequence coverage over the full length of this sequence fragment. We conclude that this fragment is a mixture of loop and helix, with a structure that may possibly resemble that given in Fig. 3.3

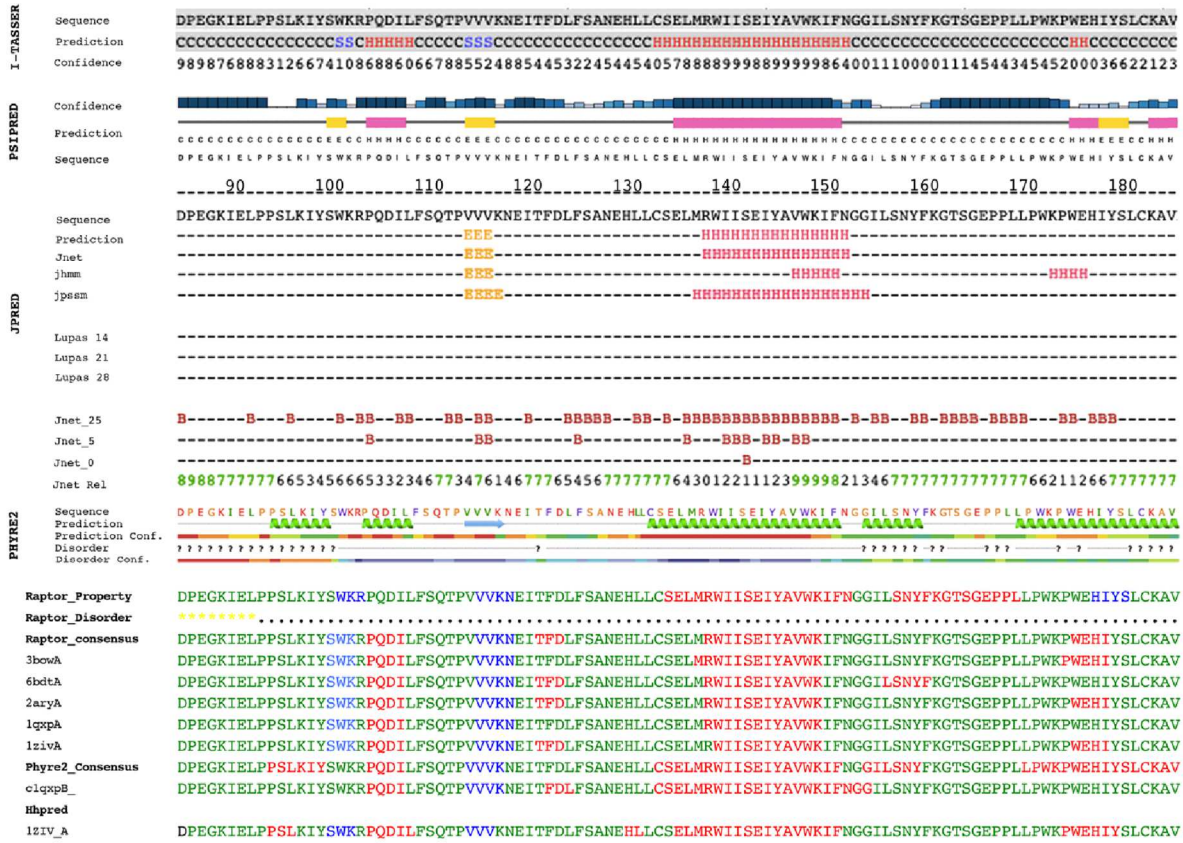

Fig. 3.4.1: Secondary structures, and results of fold recognition for fragments D84 – L295.

3.3.1.2. MSA Fragment 2: D85 – L295

Unlike the previous fragment in Fig. 3.2 D84 – L295, this portion of the sequence in Figs. 3.4 presents less disorder predicted by both PHYRE2 and RaptorX Property. Fig. 3.4.1 also shows clearer structures across the servers, though there are still regions where there is disagreement in the structures predicted. I-TASSER and PSIPRED predicted the same secondary structures barring the helix and loop at the C-terminus of the PSIPRED prediction, where PSIPRED predicts an additional sheet and helix present. Only in the central helix, and sheet is JPRED shown to have predicted the same structures

as the previous two servers. The prediction in PHYRE shows second and third helix, and only sheet predicted are shown to be in agreement with the PSIPRED and I-TASSER predictions in terms of positioning of these structures.

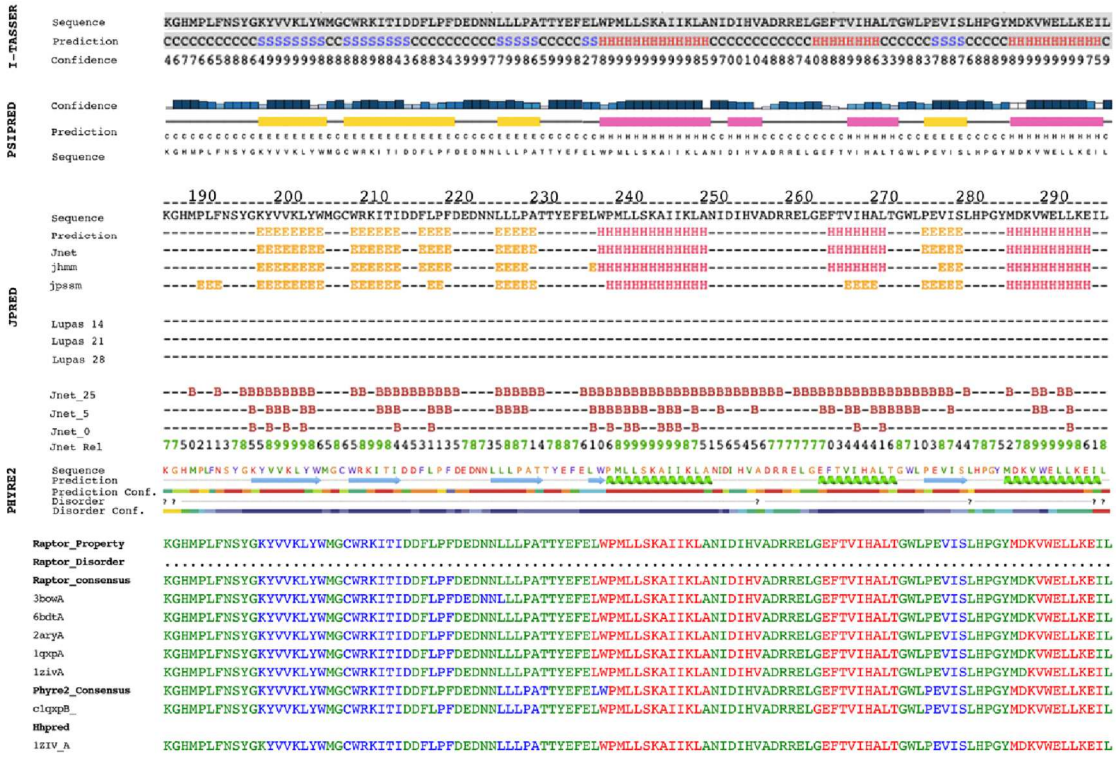

Fig. 3.4.2: Secondary structures, and results of fold recognition for fragments D84 – L295

The second half of the alignment for fragments D84 – L295 in Fig. 3.4.2 presents even less disorder compared to the first half of fragment D84 – L295, as predicted by

PHYRE2. RaptorX Property predicted absolutely no disorder. There is a clearer consensus in the structures present in Fig. 3.4.1 across the independent servers.

The first half of the region in Fig. 3.4.1 displays a sheet-rich region; I-TASSER, PSIPRED, and PHYRE2 placed the first 3 predicted structures in roughly the same region, differing only on the length of structures predicted. The JPRED alignment shows the second, and third sheet predicted spanning the same residue positions as the second sheet predicted by PSIPRED.

The second half of the region in Fig. 3.4.2 shows very clear structures across all of the servers, with each one predicting a minimum of 3 helices, and an additional sheet between the second and third helices. The lack of disorder probably makes this a good region for expression *in vitro*. JPRED, and I-TASSER have independently predicted a small amount of sheet preceding the first helix. RaptorX has predicted a second helix after the first throughout all of its models, and PSIPRED also produced a helix in roughly the same place.

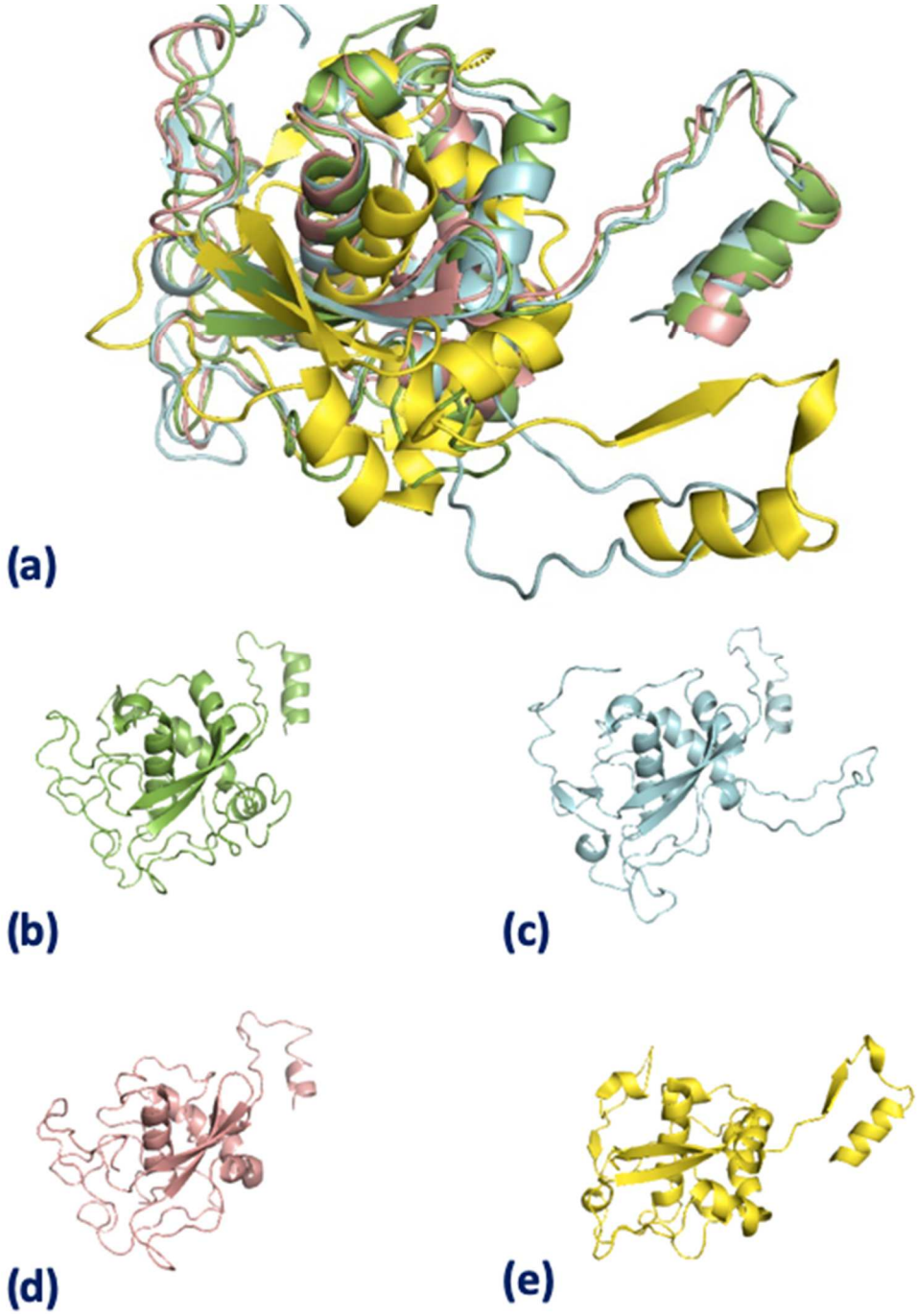

Fig. 3.5: Models produced in the region D84 – L295 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 1KFU. (c) final model produced by RaptorX using threading template 3BOW – a homologue of 1KFU. (d) threading template 1QXP, a calpain homologous to

both 1KFU and 3BOW, was used by PHYRE2. (e) threading template 1ZIV, the catalytic domain of human calpain 9, was used by HHPRED.

Fig. 3.5b shows the structure 1KFU (human m-calpain), which was used as the top-ranking template when modelling the final I-TASSER model, producing a C-score of -0.42. This stands as a clear indication of a good prediction as it has generated confidence that falls above the threshold for C score (-1.5). The associated TM score of 0.66±0.13, which even at the lower bounds of this range of values (>0.5) would suggest that the model has good topology and has been modelled in such a way that it is similar to the natively folded template. The threading template 1KFU produced a Z score of 1.68, which is sufficiently above the threshold (1) indicating good alignment through the 82% sequence coverage.

Fig. 3.5c contains the structure 3BOW (m-calpain) the threading template used in RaptorX, had a P value of 2.7×10^{-8}. This P value falls significantly below 1×10^{-2} which corresponds to high confidence, in fact the P value generated falls below the 1×10^{-3} threshold for certainty, that the model generated has adopted an appropriate fold. The alignment produced a total of 128 matches out of a possible 212 residues in this domain, suggested the alignment of the sequence fragment to the template sequence is moderate-high as it surpasses what would be considered a median value for an alignment score in this domain. When alignment score is considered in conjunction with the uGDT score of 105, which is above the threshold for what is considered a good model for a protein this size (>50) , it further increases confidence in the quality of the model produced using the threading template 3BOW.

This model has 2 more beta-sheets predicted in its structure than Fig 3.5c, however, both models are largely the same, as both Fig. 3.5c and 3.5d represent homologous calpain domains.

Fig. 3.5d shows PHYRE2 using 1QXP (mu-like calpain) as the top model for threading. It generated a very high confidence level of 100%, covering the full length of the query sequence submitted into PHYRE for the region D84 – L295. Despite the full sequence coverage offered by this PDB, there were a lot of gaps interspersed within the alignment, resulting in only 174 matches out of the 212 total residues present in this sequence fragment actually being modelled directly from the threading template.

The alignment produced an alignment with 22% identities, however, the significance of identical residues diminishes with the increasing number of gaps, as with enough gaps inserted into an alignment, any protein can be made homologous to a protein with which it shares no evolutionary relation.

Fig 3.5e shows the structure 1ZIV (human catalytic domain of calpain 9), being used as a template for threading for the sequence fragment by HHPRED, through the residues D84 – L295. The alignment generated covered the full-length sequence barring a single reside at the beginning. This template generated a probability of 99.91%, suggesting we can be certain that this region in the sequence is homologous to calpain. This model generated a moderate sequence similarity of 32.2%, which is still in the 'twilight zone' i.e. 25% - 35% similarity, and identity of 22%.

Both halves of the sequence fragment D84 – L295 in Fig. 3.4.1 and 3.4.2 have had a number of calpains display a huge amount of homology when utilising different fold recognition methods. In fact, each server has used a different calpain as a threading template in this sequence fragment, and such a consensus across all these servers

increases the reliability, and confidence in the conclusion that fragment D84 – L295 as a whole is homologous to calpain. Though to state this unequivocally in the absence of experimental data would be dubious. It is imperative to express this region *in vitro* to truly ascertain the structural nature of this sequence fragment.

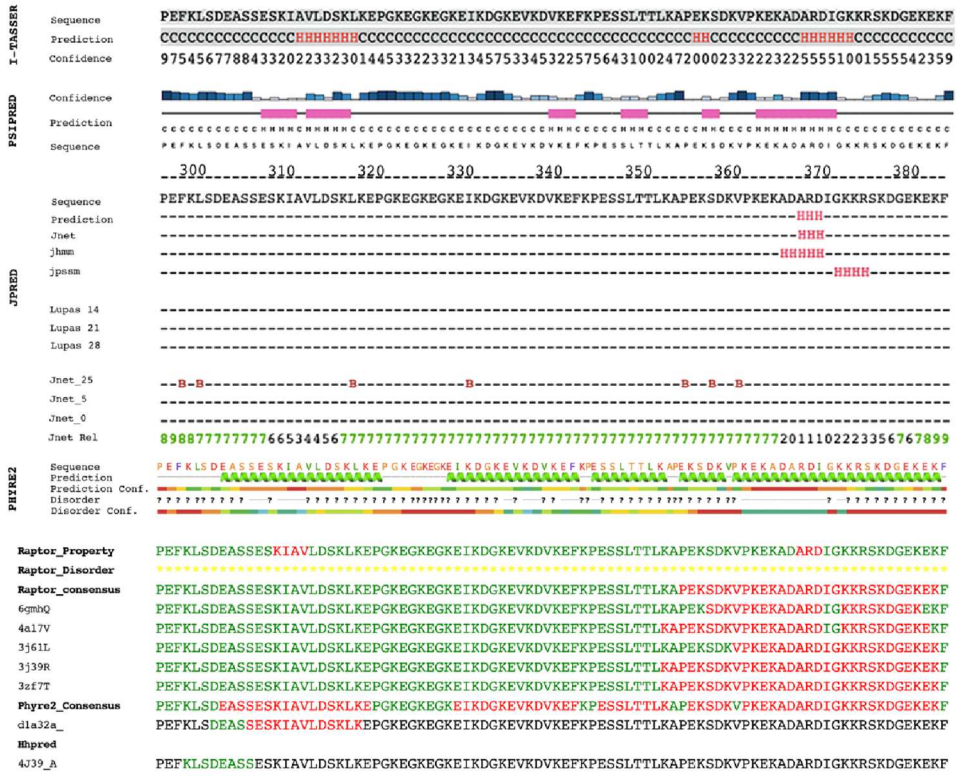

Fig. 3.6: Secondary structures, and results of fold recognition for fragments P296 – F383.

3.3.1.3. MSA Fragment 3: P296 – F383

This region in Fig. 3.6 is predicted as largely disordered by PHYRE2. RaptorX Property predicts this fragment to be fully disordered, as opposed to merely the majority of the fragment being disordered. I-TASSER shows remarkable lack of confidence, with values of equal to or less than 5 at each residue for the first two helices predicted. PSIPRED displaying the same lack of confidence in the first four helices predicted; many of the residue positions have pale grey rectangles, with just a few light/dark blue bars on the second helix.

The is no clear consensus across the servers, as all predicted alpha helices have variable lengths. The only real consensus is in the 3 residues predicted as helices by JPRED. JPRED only predicts a helix in the positions where the confidence in both ITASSER, and PSIPRED may be considered moderate to good; these residues coincide with a small region in which PHYRE2 has predicted the structure to be ordered. However, this is the same region within which confidence generated by JPRED alignments confidence falls significantly towards zero.

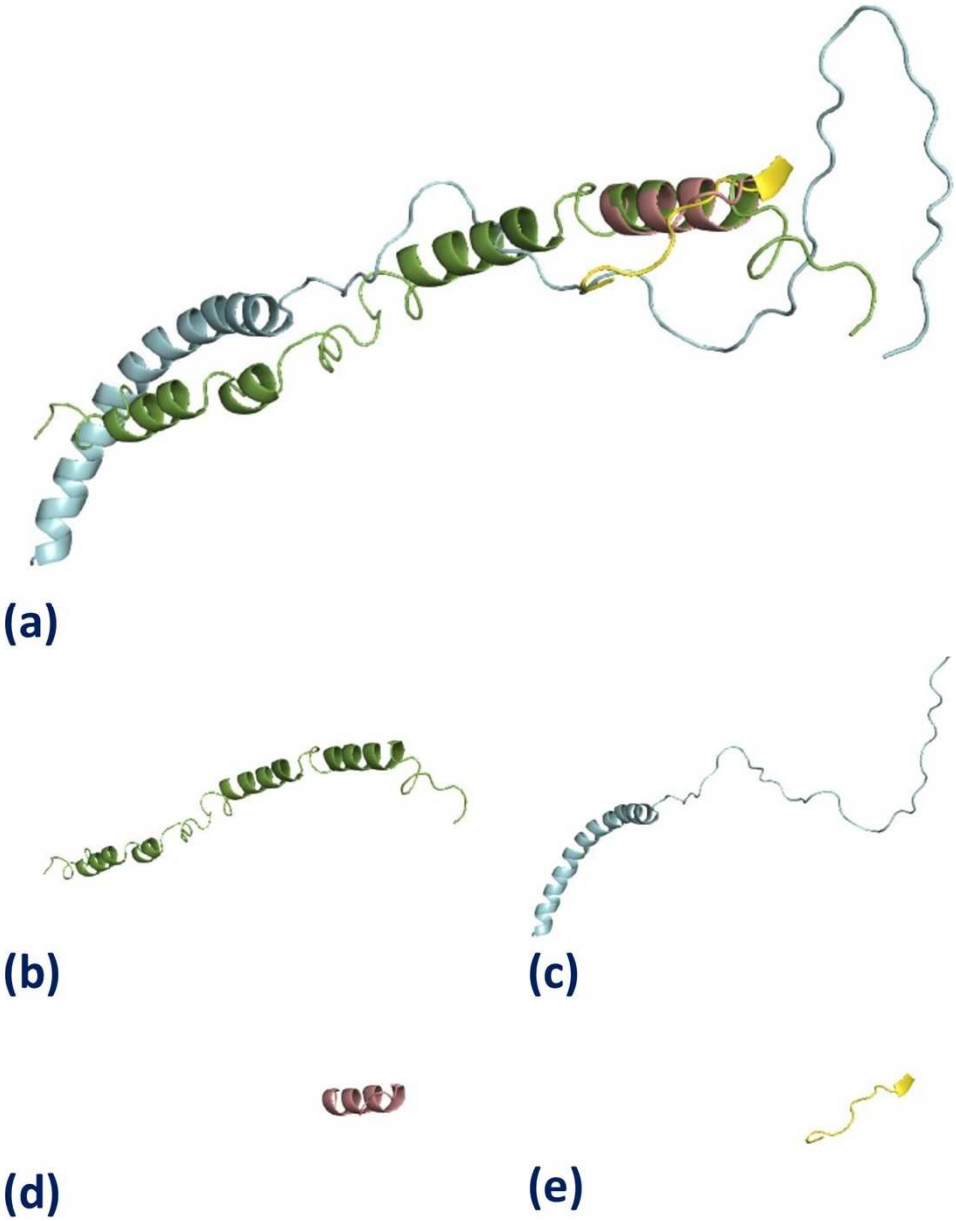

Fig. 3.7: Models produced in the region P296 – F383 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 6EXN. (c) final model produced by RaptorX using threading template 6GMH. (d) threading template 1A32 was used by PHYRE2. (e) threading template 4J39 was used by HHPRED.

Fig. 3.7b contains the structure 6EXN (spliceosome P complex), which was the top-ranking threading template used for the final, largely helical, model produced by I-TASSER. This model generated a C score of -2.04 in this sequence fragment, falling below the value -1.5 which is considered good. The TM score associated with the final model produced was 0.47±0.15. As it was above the threshold associated with TM score (>0.5), this indicated that when the final model was compared to the native structure of the threading template, it was found to be an accurate model and evaluated as having adopted correct topology. The threading template generated a Z-score of 1.44, which is above the threshold (>1) that corresponds to a good threading alignment across 98% of the query sequence.

Fig 3.7c contains the structure 6GMH (activated transcription factor complex), which provided a threading template to RaptorX. The region was modelled predominantly with loop residue that accounted for over half the residues in this sequence fragment. The model generated a P value of 1.4×10^{-2} alongside an alignment score of 6 out of 88 residues in this region. The P value generated for this model was moderate, as it failed to fall below the 1×10^{-2} for high confidence. the alignment score is bad when considering the length of the fragment, as a moderate alignment score be ~40.

The GDT score for this model is a more appropriate measure of accuracy, as the sequence fragment is in excess of 100 residues. The alignment between the threading template and the sequence fragment produced a GDT score of 20, falling short of the threshold for an accurate model (>50).

Fig. 3.7d contains the structure 1A32 (ribosomal protein s15 from bacillus stearothermophilus). The model contains a single helix flanked by a number of loop residues either side of the helix. The lack of reliability in this model stems largely from

the fact the threading template only aligned a portion of the query sequence submitted between positions 301D – 318K, amounting to only 18% coverage in this sequence fragment, for which it produced a poor confidence score of 9.73%. Despite this model exhibiting 35% identities, which is the upper limits of the 'twilight zone' for homology modelling, PHYRE2 produced the least amount of sequence coverage in this interval of all the fold recognition techniques used.

Fig. 3.7e contains the structure 4J39 is an RNA-binding protein complexed with double helical RNA. This PDB was used as a model for residues F298 – S306, generating a poor probability of 12.41%. This is despite its high sequence similarity of 98.4% with 67% positive identities. These values were likely caused by the short region over which this threading template was mapped to the sequence, which was only 8 residues. This region accounts for less than a 10^{th} of the residues present in the PDB. We need to question this model as we do not expect androglobin to be complexed with RNA.

Of the 4 fold recognition techniques used, I-TASSER in Fig. 3.4b produced the best confidence and alignment score in this region with the PDB 6EXN. The P value for 6GMH in Fig. 3.4c fails to fall below the ranges 1×10^{-2} where the probability would indicate that RaptorX had a high degree of confidence that the model generated had adopted an appropriate fold. This is as both PHYRE2 in Fig. 3.4d and HHPRED in Fig. 3.4e provide low sequence coverage, in addition to incredibly low confidence in the model predicted. Here we can conclude little more than the idea that the fragment is likely to contain loop and helix.

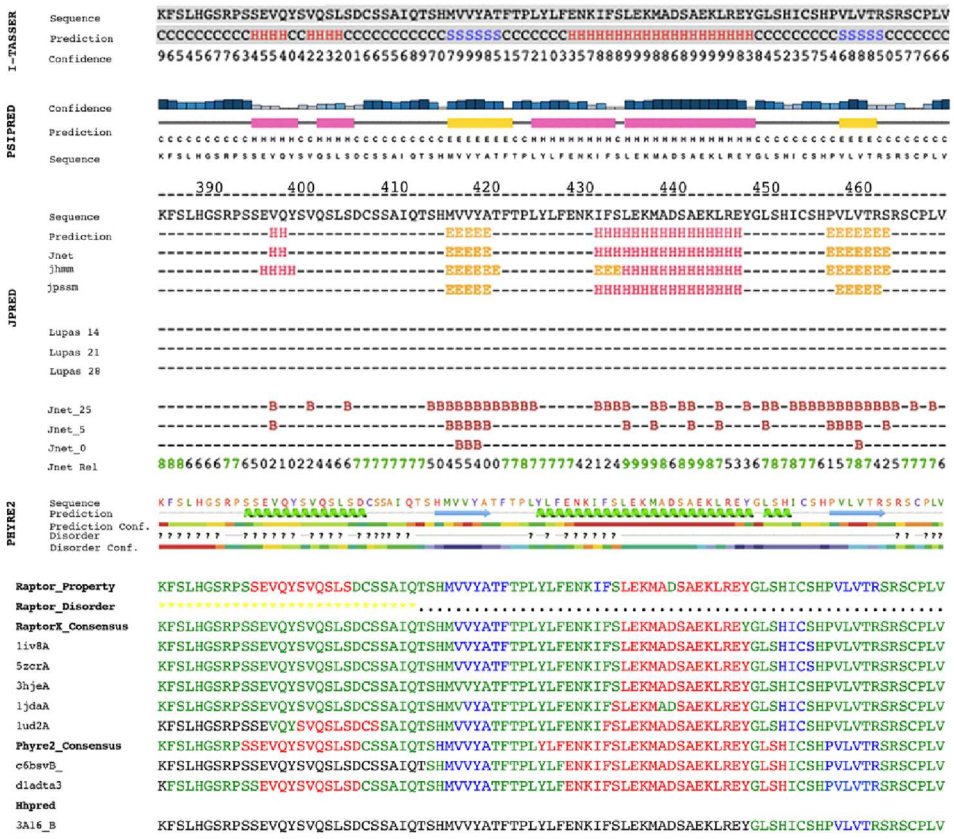

Fig. 3.8.1: Secondary structures, and results of fold recognition for fragments K384 – S574

3.3.1.4. MSA Fragment 4: K384 – S574

Fragment 4 lies in the middle of an ordered region of Fig. 2.1. We understand that the first and third ordered regions correspond to calpain; this was confirmed for the first ordered region in section 3.3.1.2 (D84 – L295). Fragment 4 therefore holds the key to understanding the structure of the whole pre-heme region of androglobin – provided that meaningful fold recognition results can be obtained.

The first half of this region in Fig. 3.8.1 presents with less disorder, and more structure across the servers when compared to the previous region in Fig. 3.6 More

disorder was predicted by PHYRE2 than RaptorX Property. Order is only present in the PHYRE2 prediction in the region surrounding the residues in the first predicted beta-sheet, and the last two thirds of the second helix PHYRE2 has predicted, all the way up to the end of the last beta-sheet predicted. We largely observe clear structures across all servers, but there is disagreement in the placement of last sheet by RaptorX. JPRED, and ITASSER show a consensus on the location of structures predicted within the query fragment sequence, except for the second helix in the I-TASSER alignment. The PSIPRED prediction is a closer to the I-TASSER prediction than JPRED, with the only structural difference predicted between the two servers is the length, and conformation of its third helix. In PSIPRED, the third helix predicted by I-TASSER is bisected by a coil conformation for a single residue, for which the confidence is low, and the N-terminal end of the first helix has been extended by several residues.

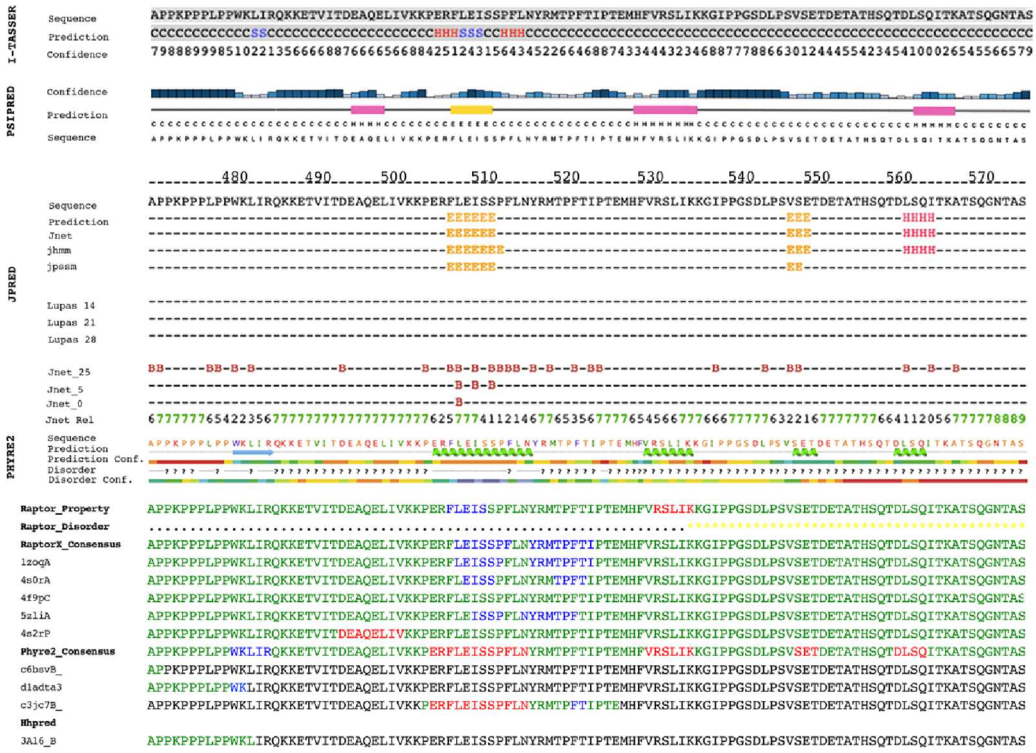

Fig. 3.8.2: Secondary structures, and results of fold recognition for fragments K384 – S574

The second half of this region K384 – S574 in Fig. 3.8.2 has more disorder and less structure when compared to the first half of this region. There seems to a reasonable consensus between servers in the first beta-sheet predicted in JPRED. PHYRE2 is the only server at this position to have a different structure as it predicts a helix at this position. The only other consensus across servers that lies within this portion of the fragment is in the second and last helix predictions produced by PSIPRED, and PHYRE2. The first of these two helices predicted have been supported by the RaptorX Property prediction.

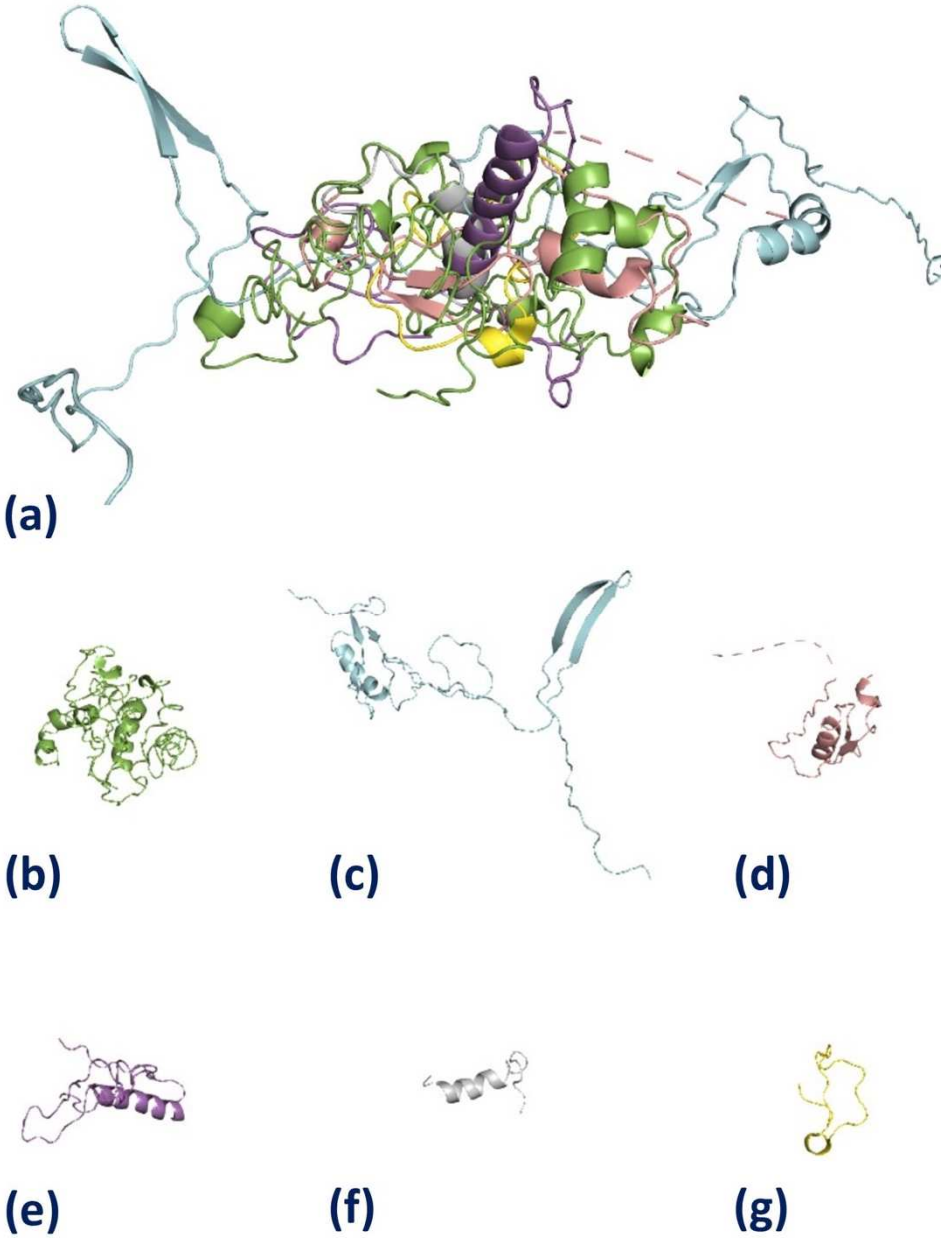

Fig. 3.9: Models produced in the region K384 – S574 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon, magenta, and grey), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 2PFF. (c) final model produced by RaptorX using the threading templates 1IV8 and 4S0R. (d) 6BSV, (e) 1ADT, and (f) 3JC7 were all used by PHYRE2 as threading templates. (g) threading template 3A16 was used by HHPRED.

Fig. 3.9b contains the structure 2PFF (Yeast fatty acid synthase), which was used as a threading template in 3 out of 10 instances I-TASSER sought an alignment, achieving a Z score of 2.32 and a sequence coverage of 91%. The Z score generated suggests this alignment between the sequence and the threading template is a good alignment, as the Z score generated is significantly higher than the threshold for a good alignment (>1). This is supported by the fact that the alignment between the two sequences largely lacks gaps.

Despite the exceptionally high Z score corresponding to the threading template used, the final model produced through this region has a poor C score of 4.20 and a TM score of 0.47±0.15. Though the C score indicated that the I-TASSER severely lacked confidence in the model produced, as it fell significantly short of the threshold for what is considered s good alignment (-1.5). However, the TM score has a mean value that falls marginally below the threshold (>0.5), indicating that the model was produced with a good amount accuracy when taking into account the range of values possible. The alignment displays a lot of loop regions, which account for over half of the model, in addition to a number of incomplete helical turns, likely contributing to the low C score.

Fig. 3.9c contains the model produced in by two separate threading templates. The structure 1IV8 (Maltooligosyl trehalose synthase), was the first of the threading templates used. It was the top-ranking threading template in DOM1 when RaptorX was used to model this region. This model produced a P-value of 2.9×10^{-02} which equates to a moderate amount confidence, as it fails to fall below the 1×10^{-02} for high confidence in this region. An alignment score of 26 out of 80 residues was generated in this domain, suggesting this is be a moderately poor alignment. The GDT score, in the domain <100 residues, is a more appropriate measure of quality than GDT. This domain generates a

GDT of 47, which is marginally below the threshold for what is considered a high-quality model in a domain of this size (>50).

The PDB 1ZOQ (Interferon regulatory factor 3 complex) was the top-ranking threading template used to model DOM2 of the model in Fig. 3.9c, generating low confidence as denoted by a P value of 1×10^{1}. When the threading template was aligned to the query sequence, it produced an alignment of 11 residues for a domain 111 residues in length, indicating Raptor X considered this an even worse alignment than the alignment produced for DOM1. The uGDT score provides a more appropriate measure of accuracy in this region, as this domain > 100 residues, meaning the score generated for this domain of 17 suggests this model is not a high-quality model.

Out of the two domains modelled by RaptorX in this region, 1IV8 produced the highest alignment score in relation to DOM1 as the sequence length was 80 residues, in addition to a better P value. The confidence associated with DOM1 was moderate, whereas the confidence associated with DOM2 was low. The threading template 1ZOQ produced an alignment score that was less than half of that displayed by 1IV8 in addition to an incredibly poor P value.

Fig 3.9d shows the structure 6BSV (Xyloglucan Xylosyltransferase), is one of 3 models that will be referenced to as a threading template used by PHYRE2 to model the query sequence in this in this region. The only portion of the model used runs through residue positions S413 – C465, amounting to 29% sequence coverage in this region. This model produced a moderate-low confidence score of 28.4% and 22% identities. Low confidence produced by this model is not enough to discount structural similarities that may exist between query and template sequence, though it is enough to discount any functional relation between the two sequences.

Fig 3.9e shows the structure of 1ADT (Zinc binding domain), which was the second of the 3 models chosen from the possible threading templates used in PHYRE2. Though it was used to model the same region as the first model PHYRE2 produced in Fig. 3.5e, it does so with a greater sequence coverage through the region F385 – K480, which amounts to 49% sequence coverage through this sequence fragment, despite the lower confidence level of 24.9%. This is a shortfall in the confidence generated is exasperated by the 14% identities present in the alignment between the threading template and the sequence fragment that generated this model. It is significantly less than that produced by Fig 3.9d.

Fig 3.9f contains the helix modelled when 3JC7 (Eukaryotic CMG helicase hydrolase protein) was used as a threading template by PHYRE2. The templated aligned to the sequence fragment through the residue positions P502 – E525, amounting to 12% sequence coverage of the query sequence provided. This alignment is too short to inform on functionality of the sequence fragment. Adding further to the unreliability of the model, it generated the confidence score of 10% in the region and 21% identities were produced by the alignment between query and template.

Fig. 3.9g contains the structure of 3A16 (Aldoximine dehydrate), which provided HHPRED with a threading template through a portion of the query sequence provided. The model produced through residues G448 – L481L, is depicted as mostly loop residues, with the only visible structure being limited to a single helical turn. The model generated through this region produced 28% identities, and a probability of 25.6%. The sequence similarity between threading template and sequence fragment is high enough that both the template sequence and the query sequence are structurally similar. The confidence level expressed in both DOM1 of the Raptor X predication in Fig. 3.9c and PHYRE2 in Fig.

3.9d provide the sequence models displaying the greatest confidence, but considering the PDB's contain dissimilar structures, there is no real consensus between the two folds.

We conclude therefore, in the absence of a consensus amongst the fold recognition results that clear insights into the structural nature of this fragment remain elusive.

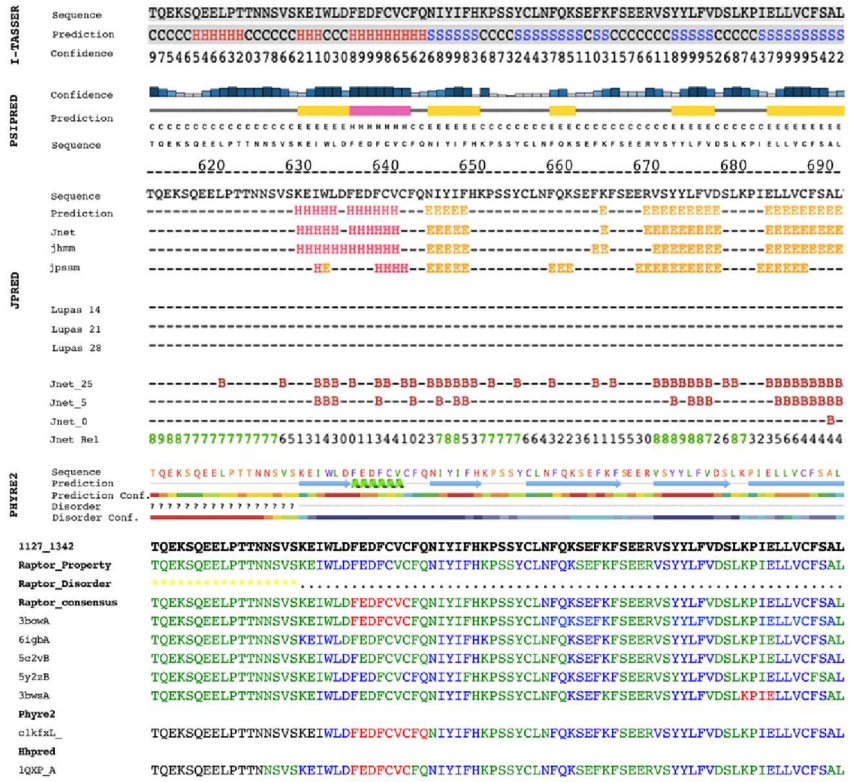

Fig. 3.10.1: Secondary structures, and results of fold recognition for the first half of fragment T612 – N781

3.3.1.5. MSA Fragment 5: T612 – N781

This is the fragment that we believe corresponds to calpain domain III (Hanna *et al.*, 2008). Here we seek to verify (or otherwise) this result. Disorder predicted by RaptorX Property is present in a quarter of this fragment in Fig. 3.10.1 and it overlaps with none of the predicted secondary structures. Clearer structures, with the latter half of the fragment predicted as a beta-sheet rich region, where 3 of the 4 beta sheets predicted in PSIPRED after the first helix in this alignment are supported by all of the servers in terms of localisation in the sequence. This region has a high confidence score associated with the first and third beta-sheet from PHYRE2, PSIPRED, and ITASSER. The

only other consensus reached across servers, is the presence of the helix immediately preceding the beta-sheet rich region of this fragment, baring PSIPRED which produces the its first beta-sheet prediction in this region. The helix has an above average level of confidence generated, with moderate to high confidence score in the case of PHYRE2 throughout the length of the helix and is also predicted by RaptorX and HHPRED.

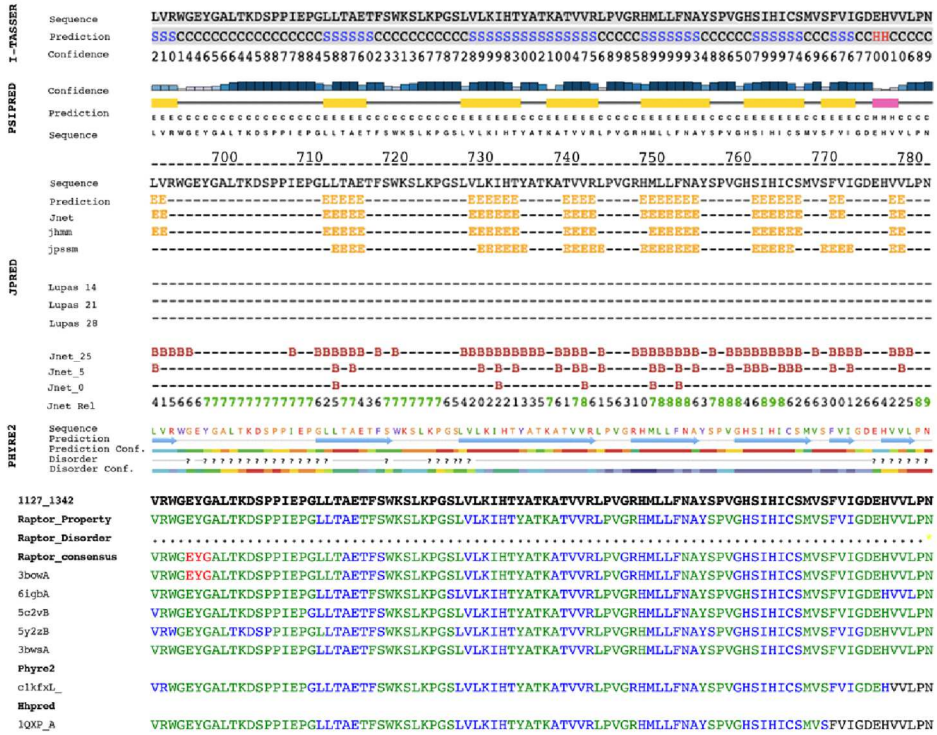

Fig. 3.10.2: Secondary structures, and results of fold recognition for the second half of fragment T612 – N781

Disorder from PHYRE2 mostly appears in regions predicted as having predominantly loop residues in this portion of the sequence. There are clear structures predicted across the servers, similar to the previous fragment in Fig. 3.10.1, with most of the structure predicted overlapping either partially or completely; clear consensus of nearly twice the beta-sheets across the length of the whole fragment in Fig. 3.10.2 when compared to Fig. 3.10.1. This consensus is present across all servers including Raptor and HHPRED, barring the third beta-sheet in the alignment produced by both I-TASSER and PHYRE2 where sheet 3 and 4 are continuous. Confidence for this helix in both of the aforementioned servers declines halfway through the beta-sheet in both these servers. The last structure in this fragment breaks the consensus in this portion of the sequence, as no real consensus is reached between the presence of sheet or helix, this is in addition to PHYRE2 and HHPRED failing to map a template structure to this portion of the sequence.

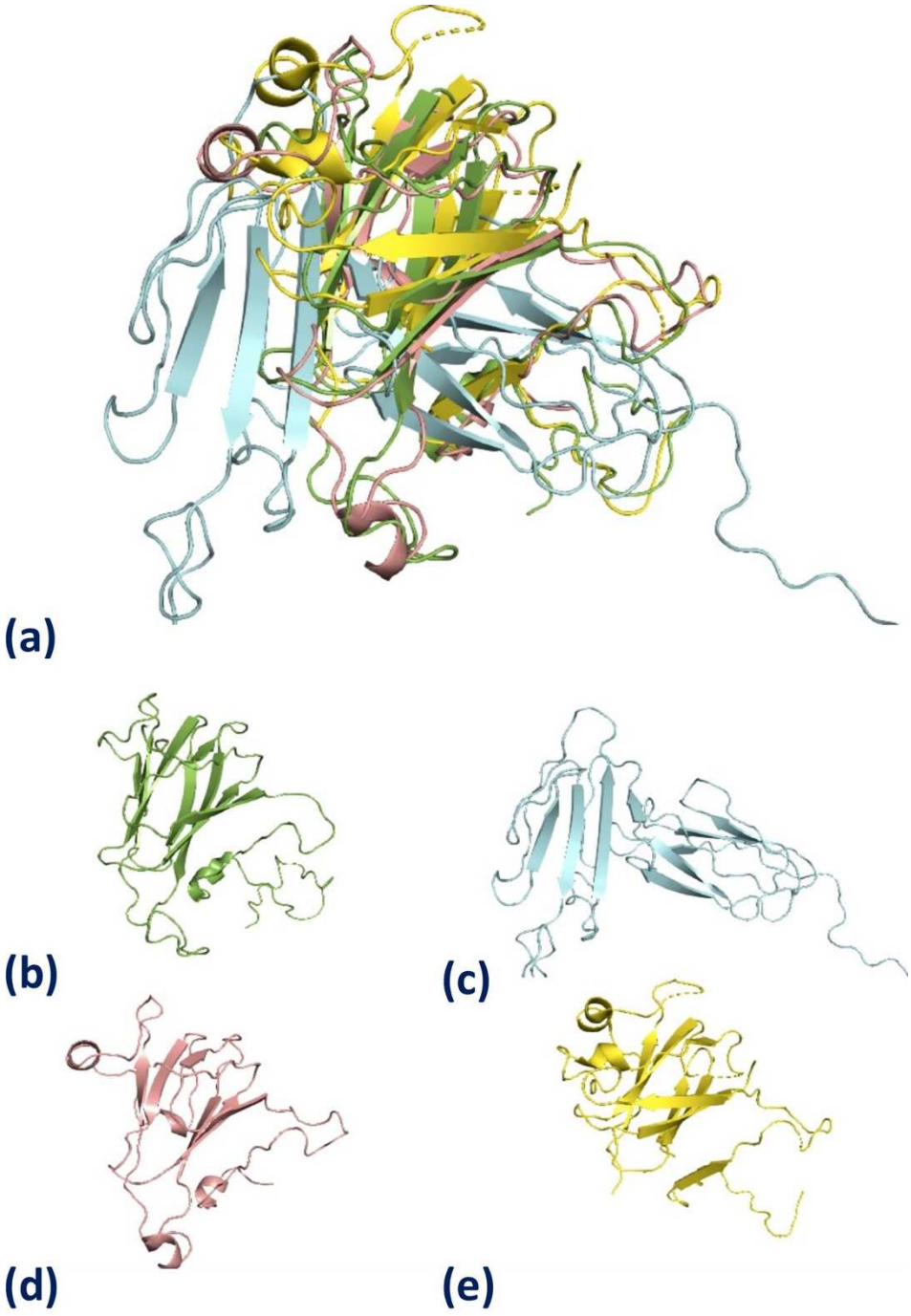

Fig. 3.11: Models produced in the region T612 – N781 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 1KFU. (c) final model produced by RaptorX using the threading templates 3BOW. (d) threading template 1KFX was used by PHYRE2. (e) threading template 1QXP was used by HHPRED.

Fig 3.11b contains the structure 1KFU (Human m-calpain), which was used as the main template for threading the query sequence by I-TASSER in this region. This threading template was used by I-TASSER in 5 out of 10 instances, generating a Z-score of 1.46 and a sequence coverage of 97%, it is the highest-ranking threading template. The Z score produced was above the threshold (>1), an Indication that this was a good alignment between the template sequence and the query sequence. In addition, the final model produced by I-TASSER generated a TM score of 0.62±0.14, and a C score of -0.71. Both these parameter were above their independent threshold. A TM score above 0.5 signifies that the model was evaluated as having the correct fold in relation to the natively folded threading template, and it displayed a good level confidence as the C score exceeded the threshold value of -1.5.

Fig. 3.11c contains the structure 3BOW (Rat m-calpain complexed with Calpastatin), which was used as the top threading template used by RaptorX when modelling this region. It produced a probability score of 2.1×10^{-3} and an alignment of 49 residues out of the 81 residues in this region. The P value corresponds with a high degree of confidence, as it falls below the threshold of 1×10^{-3} where RaptorX would be certain the sequence has adopted an appropriate fold. In addition, the alignment score surpasses the median threshold in this domain. GDT produced a score of 40 for this domain, falling marginally below the threshold for what is considered an accurate and good quality model in this region (>50). However, this domain also produced a uGDT score of 69, which exceeded the threshold (>50), suggesting that only a portion of this model is good.

Fig. 3.11d contains the structure 1KFX (Human m-calpain), which was used as the threading template by PHYRE2, producing a confidence score of 93.2% covering residues

W632 – H776 of the sequence providing 84% sequence coverage. Both the confidence and sequence coverage make us certain that this is an appropriate fold adopted by the sequence fragment when it was modelled through this region. Only 144 of the residues within this sequence fragment have been aligned, meaning that 26 residues from the sequence fragment failed to align to the threading template, producing 14% identities.

Fig. 3.11e contains the structure 1QXP (Rat m-calpain), which was used in HHPRED as a threading template, producing a probability of 91.8%. This resulted in high degree of confidence amounting to certainty, that the predicted for this region through residues N625 – S769 has produced an appropriate fold. The 85% sequence coverage offered by this threading template, despite the low sequence similarity of 11.7% and identities of 13% generated by the sequence alignment, further supports the high confidence generated by the template.

However, since all of the 4 fold recognition servers give the consensus of this region being homologous to calpain, with high confidence for each server, the individual descriptors that attempt to asses sequence similarities, identities and the relative quality of the models being produced become less significant in light of the consensus. Even if a conclusion on function may not be possible, a clear beta-barrel structure can be seen in all structures produced by the fold recognition results. Thus, the fold recognition confirms the initial hypothesis, that this region is related to calpain. The identification of this region being related to calpain has not yet been published

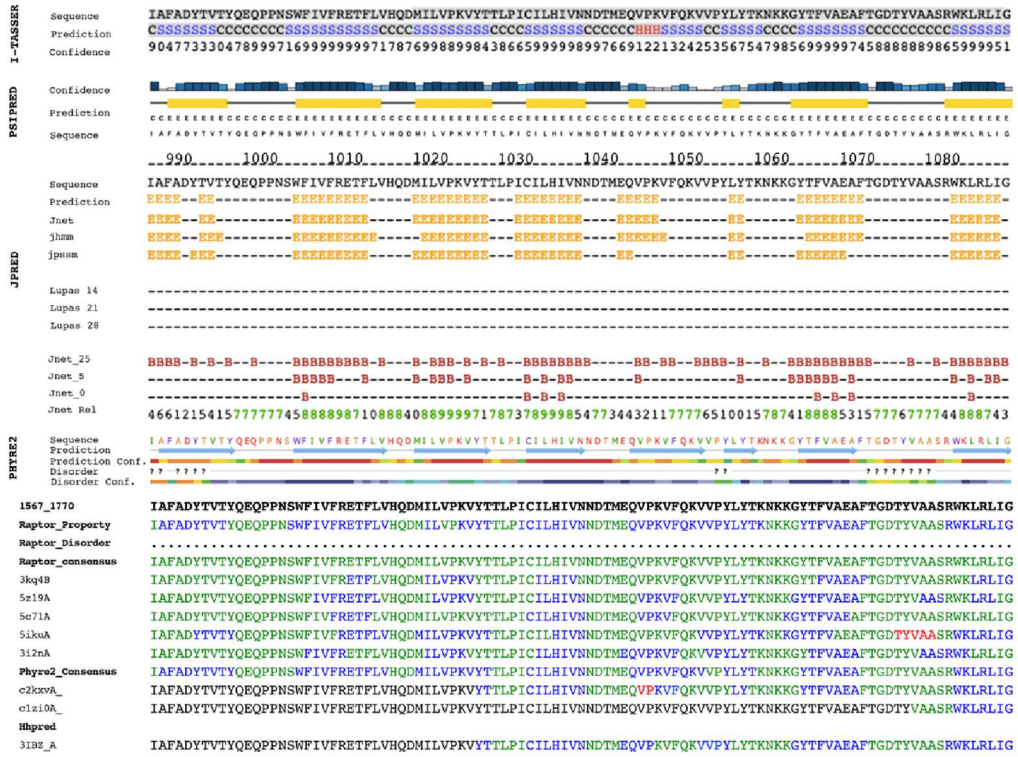

Fig. 3.12.1: Secondary structures, and results of fold recognition for fragments I986 – I1188

3.3.1.6. MSA Fragment 6: I986 – I1188

I986 – I1188 corresponds to the largest region of ordered residues in the post heme region of the androglobin sequence (Fig. 2.3). There are no prior indications as to the structure of this region. The region is characterised by well-defined secondary structure in similar fashion to the previous fragment in Fig. 3.10.2. The clear consensus across all servers is that this region is rich in beta-sheet, with the first half of this region having a total of 8 supported beta-sheet structures (Fig. 3.10.1). The structures are observed differing only in the size of each sheet predicted, whilst the consensus of their localisation within the sequence remains intact. Furthermore, approximately half of its

residues are considered disordered by PHYRE2. The confidence levels across all the servers appears exceptional for sheets 2, 3, 4, 7, and 8; declining briefly to moderate to low confidence in the fifth and sixth beta-sheet predicted in this region. RaptorX model 5IKU predicts helix in the loop conformation predicted between sheets 7 and 8. Additionally, a few residues predicted in the I-TASSER alignment preceding the fifth beta-sheet have been predicted as helix, for which a single PHYRE2 model has similarly predicted these same residues as helix despite the consensus of this region being beta-sheet.

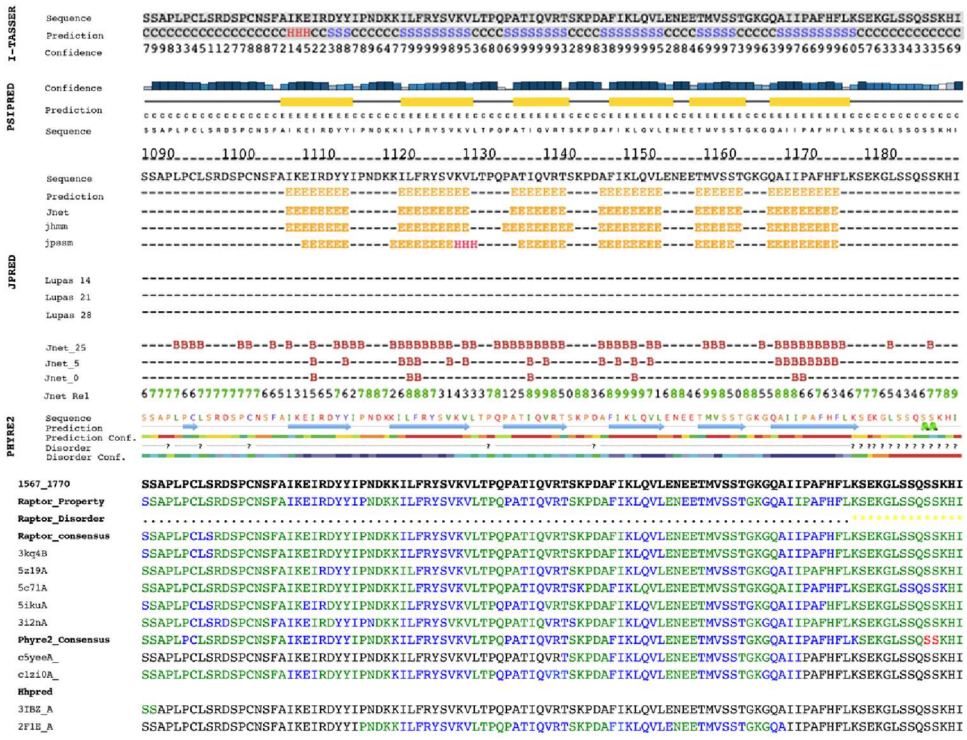

Fig. 3.12.2: Secondary structures, and results of fold recognition for fragments I986 – I1188

This regions shares a near identical consensus to first half of this fragment in Fig. 3.12.1 as they both depict a beta-sheet rich region, totalling 6 beta-sheet structures, in the consensus for this half of the sequence fragment. This region of the protein contains very few isolated pockets of disorder. There are a few residues of beta-sheet predicted in the alignments generated by PHYRE2, RaptorX, and Raptor Property that precede the first clearly defined beta-sheet, for which a consensus reached across all the servers. RaptorX does not support this consensus for the first beta-sheet. Confidence scores are between moderate and high for secondary structures predicted from the second beta-sheet through to the sixth beta-sheet, across all servers, whilst there are below average confidence values for the first beta-sheet. I-TASSER predicts helix in the same portion of the first beta-sheet predicted in the general consensus. PHYRE2 produces the only helix predicted at the end of this region.

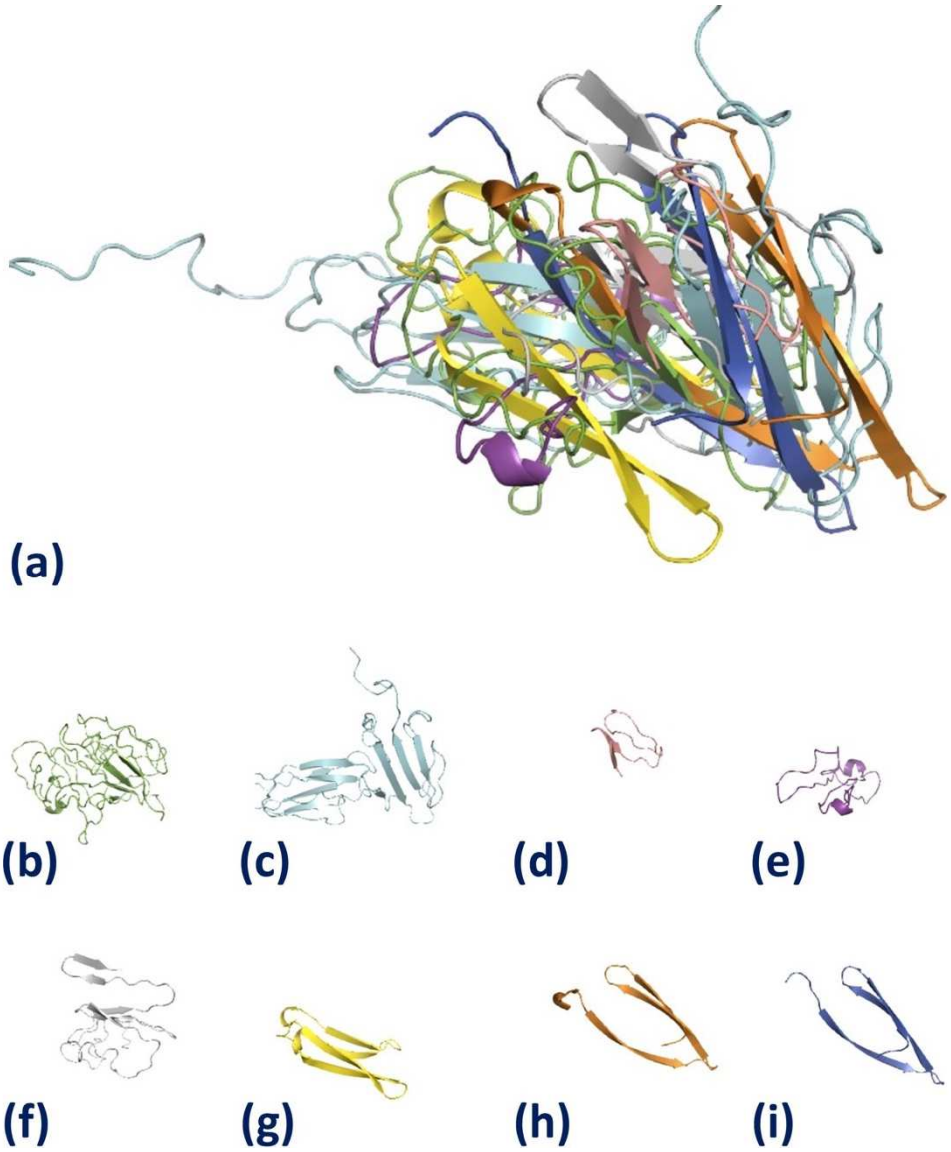

Fig. 3.13: Models produced in the region I986 – I1188 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon, magenta, and grey), and HHPRED (yellow, orange, and blue). (b) final model produced by I-TASSER using threading template 5Z6D. (c) final model produced by RaptorX using the threading templates 3KQ4. (d) 5YEE, (e) 2KXV, and (f) 1ZIO were all used by PHYRE2 as threading templates. (g) 3IBZ, (h) 2F1E, and (i) 1TZA were all used by HHPRED as threading templates.

Fig. 3.13b contains the structure 5Z6D (Abundant Perithecial Protein from Neurospora crassa), which was used as a threading template. The template produced a moderate Z score of 1.10, when considering the threshold associated with Z score is >1. The threading template mapped through the region I986 – I1188 with 75% sequence coverage. This was due to the large number of gaps within the alignment, when the sequence fragment was mapped on to the threading template. Despite the low, though acceptable, Z score produced by the template the final model I-TASSER produced had an exceptionally low C score of -4.53. The final model was found to fall short of the -1.5 threshold for good amount of confidence in the structure generated. Furthermore, the TM score associated with the final model generated by I-TASSER is 0.24±0.07, which is well below the threshold of 0.5. The lower bounds of the TM score would suggest that the structure adopted by the final model is too dissimilar from the natively folded form of the threading template to have the correct topology.

A single and incomplete helical turn is present within the model in a region that otherwise exhibits 2 pairs of antiparallel beta-sheet; in essence these structures resemble half a beta-sandwich through this region. Over half of the residues in the model have taken a loop conformation.

Fig. 3.13c contains the structure 3KQ4 (Intrinsic Factor-Cobalamin complexed Cubilin), which provided RaptorX with its threading template across the full length of sequence fragment 6. The model produced a P value of 2.8×10^{-3}, indicating a high degree of certainty in the structure produced, failing to produce a P value less than the 1×10^{-3} required for certainty. In addition, the alignment score of 53 residues out of a possible 203, was moderately low as it failed to reach median values for a protein domain this size. However, the uGDT of 49 generated in this region for this model, is only marginally

below the threshold of 50 that is necessary for a model for this size to be considered a high-quality model.

This model seems to have the same 2 pairs of antiparallel beta-strands in relatively similar portions of the sequence as those present in Fig. 3.13.b. However, in addition to the 2 pairs of antiparallel beta-sheets it contains 5 beta-sheet strands in a conformations that resembles a beta-sandwich and a single short beta-sheet strand. This model also displays far less loop region than that present in Fig. 3.13b.

Fig. 3.13d contains the structure 5YEE (lokiprofilin1/rabbit actin complex), which was the highest-ranking threading template used by PHYRE2 in this region. The template aligned to residues T1139 – I1168 in the sequence fragment, providing 14% sequence coverage. This model also produced a high confidence score of 73.9% and a total of 40% identities. Both of these descriptors indicated that the model is structurally similar to the sequence fragment, as the identities produced are above the 'twilight zone' for homology modelling. Given the low sequence coverage, the match may only be able to inform on protein structure through the set of residues that have aligned onto the query sequence.

Fig. 3.13e contains the structure 2KXV (calcium-binding properties of the tellurite2 resistance protein terd), which was the second ranking threading template used in this region by PHYRE2. This threading template produced a confidence score of 65% and 23% identities over the 53 residues that have been aligned between the two sequences. The greater sequence coverage offered by this model, in addition to the model's ability to bind calcium and potentially modulate the activity of calcium-dependent calpain within Adgb, makes it a more appealing template for homology modelling, despite the decrease in identities when compared to Fig. 3.13d. In addition,

this model gave greater sequence coverage when compared to Fig. 3.13d, as it gave an alignment through the residues Y1024 – G1086 and provided this sequence fragment 30% coverage from through this alignment.

Fig. 3.13g contains the structure 3IBZ (putative tellurium resistant like protein), which was used as a threading template by HHPRED through the residues Y1024 – S1088. This equated to 32% sequence coverage. This threading template also produced a probability of 63.7% and 13% identities. The probability produced seems to suggest this is a reasonable model, especially when considering the alignment produced 32% similar residues in this region, the low sequence coverage may explain reasonably high probability generated. The protein model has Ca^{2+} binding properties, so may have some functional similarity in addition to the structural similarity, since Adgb is expected to bind calcium.

Fig. 3.13h contains the structure 2F1E (ApaG from Xanthomonas axonopodis), which was used as the second ranked PDB used by HHPRED through the region I986 – I1188. This provided this fragment with an additional threading template further downstream of Fig.3.7i. This model produced a probability of 62% when used as a threading template through the residues P1114 – A1166 in the Adgb sequence. Both threading template and the query sequence were found to be 37% similar to one another, producing 15% positive identities in the aligned region. The similarities generated are above the 'twilight zone'. This suggests that this model would be structurally similar enough to the query sequence to be used as a template in homology modelling. ApaG holds an unknown function, despite being found in a wide range of bacterial genomes. Even if the function of ApaG was known, it is unlikely that the

function may be inferred due the small amount of sequence coverage offered through this alignment.

Fig.3.13i contains the structure 1TZA (ApaG), which has been used by HHPRED through the same set of residues, P1114 – A1166, as was seen in Fig. 3.13h. This model produced an identical probability of 62%, though this model produced 19% identities and 41% similarity between the threading sequence and query sequence. Though the reliability in model is bolstered by HHPRED mapping the same protein from two different PDBs through the same region of the query sequence, it is impossible draw a conclusion in the absence of experimental data, other than that this region is predominantly beta-sheet.

All structures yielded by fold recognition servers in this fragment with variable degrees of confidence in the model produce. However, they form a consensus of Adgb, in this portion of the sequence, being predominantly beta-sheet; with limited alpha helix, as can be seen in the structures presented in Figs. 3.13b, 3.13e, 3.13h, and 3.13i. Unfortunately, despite the lack of disorder in this region (Fig. 2.3), there is no consensus from the fold recognition as to what the fold of this region may resemble.

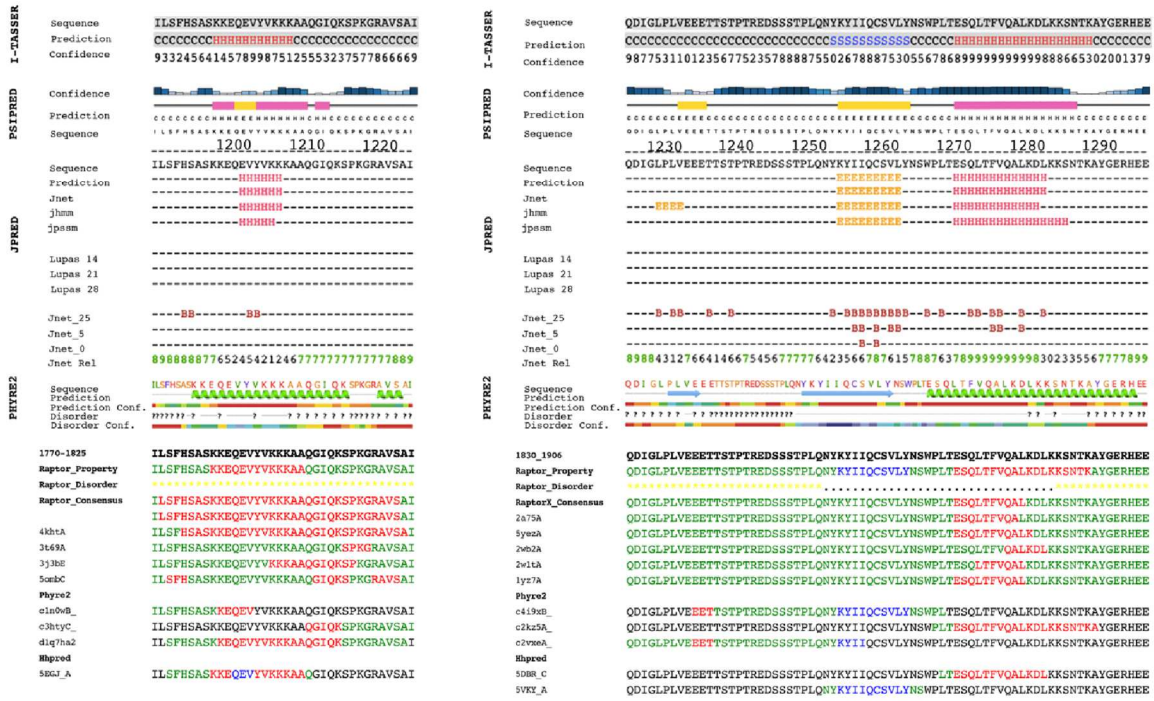

Fig. 3.14: Secondary structures, and results of fold recognition for fragments (14.1) I1188 – I1223 and (14.2) Q1224 – E1295.

3.3.1.7. MSA Fragment 7 & 8: I1188 – I1223 & Q1224 – E1295

Fragments 7 and 8 cover a region with a high degree of disorder (Fig. 2.3) and so it is unlikely that any strong hits will be retrieved from the fold recognition. Nevertheless, we will describe the fold recognition results for this region.

The Raptor Property disorder predication covers some, if not all, of the secondary structure predicted in this region. Fig. 3.14.1 shows a consensus, servers predicting the presence of helix with overlapping therefore similar localisation but spanning a differing number of residues. High level of confidence was produced towards the centre of this helix by I-TASSER and PHYRE2, whilst PSIPRED's confidence score peaks towards the end

of this same helix. JPRED's confidence stays within moderate bounds throughout the course of the helix. In both PSIPRED and HHPRED this helix is bisected by the presence of beta-sheet, which is present for 3 residues in the exact same position on both alignments. The alignment in PHYRE2 and the RaptorX consensus sequence, in addition to a number of the PDBs used to model this, have predicted a significantly longer helix when compared JPRED, Raptor Property, or I-TASSER.

Fig. 3.14.2 presents the same trend in terms of the disorder generated by both methods, in that it covers small portions of the secondary structures predicted. Clear consensus across all servers, excluding RaptorX and property, for the second beta-sheet predicted in this fragment with high confidence scores in I-TASSER, PSIPRED, and towards of the end of the sheet by JPRED. This helix has a strong confidence score, which only seems to be below average at the beginning of the predicted sheet.

There is also a very strong consensus for the presence of an alpha helix across all servers, though there exists a great deal of variation in terms of the length of the helix, especially in the RaptorX alignment; PHYRE2 produces the longest helix at this position. The confidence score for this helix declines as it gets longer, hitting very low confidence limits in the exit loop of this helix.

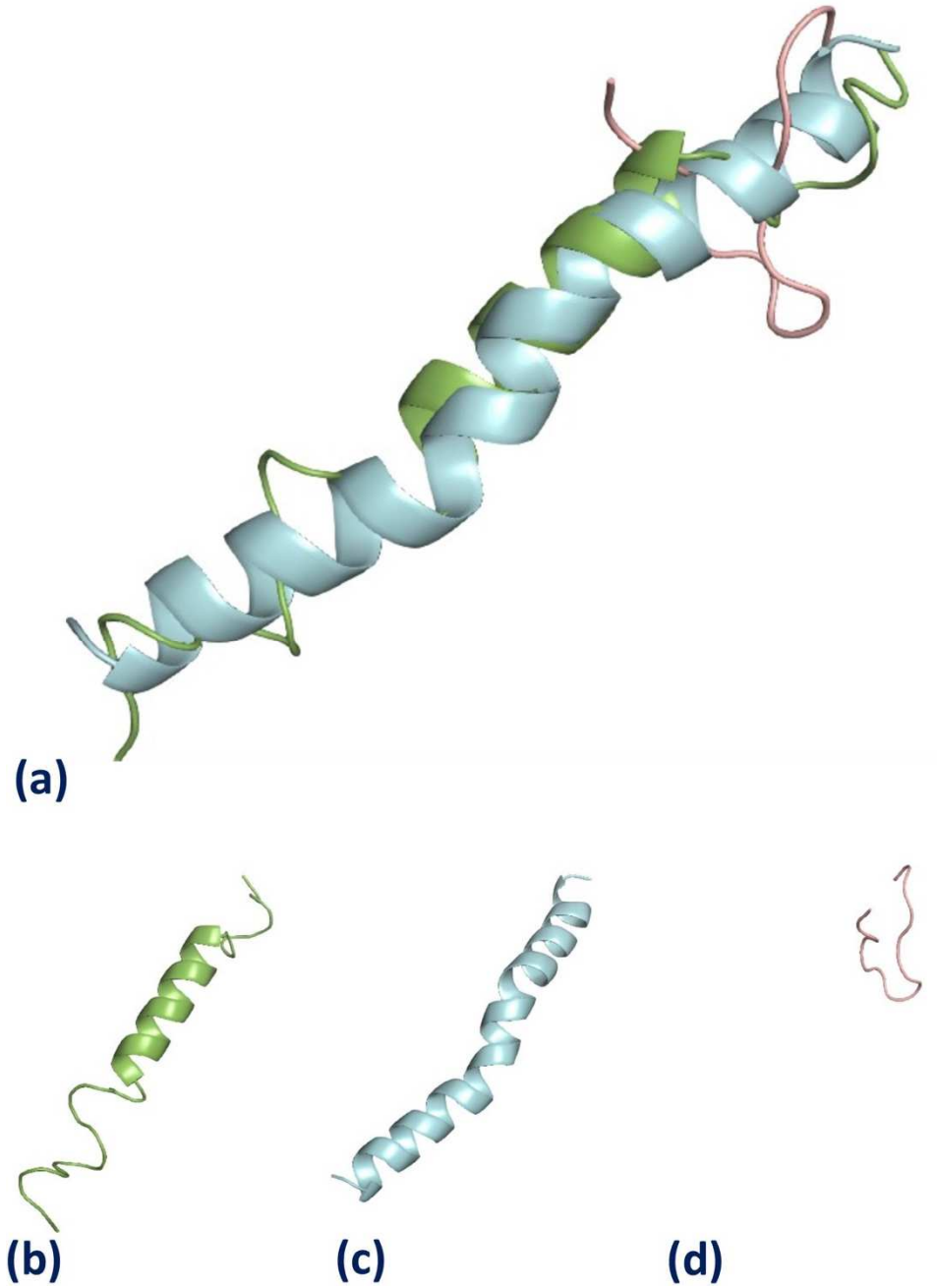

Fig. 3.15: Models produced in the region I1188 – I1223 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), and PHYRE2 (salmon). (b) final model produced by I-TASSER using threading template 2WW9. (c) final model produced by RaptorX using the threading templates 4JNP. (d) threading template 1N0W was used by PHYRE2.

Fig. 3.15b contains the structure 2WW9 (Yeast Ssh1 complex bound to the yeast 80S ribosome), which when used as a threading template generated a Z score of 1.49. This was above the threshold for a good alignment (>1), when providing I-TASSER with a threading template through residues L1189 – I1223. The alignment covered 97% of the sequence fragment, for which the final model produces a C score of -1.21 and a TM score of 0.56±0.15.

Both C and Z score produced indicate this is a good alignment between the sequence fragment and the threading template, and a moderate amount of confidence indicated by the C score generated being greater than the threshold for what is considered a good level of confidence (-1.5). The TM score also characterises the model as having adopted an appropriate fold when compared to the natively fold threading template. This is as the mean for the range of values generated, is greater than the threshold hold assigned to TM score (>0.5). This model shows a single alpha helix flanked either side by a number of loop residues with only a single residue, I1188, failing to map onto the templates structure within this region. Due to the small size of the fragment, there is likely to be no functional correlation between the template function and function of the query sequence in Adgb.

Fig. 3.15c contains the obsolete PDB 4JNP, superseded by 4R7W, (5-methylcytosine deaminase). This structure was used as a threading template by RaptorX. This threading template generated moderate amount of confidence, denoted by a P value of 1.6×10^{-02} through the full length of the 36-residue long sequence fragment, as it failed to generate a P value less than the 1×10^{-02} threshold associated with a high degree of confidence. The alignment score of 4 out of a total 36 residues indicates that

despite the moderate confidence in this the structure produced, RaptorX considered this model to be a bad alignment between the query sequence and the threading template. This model produced a GDT score of 48, which marginally fell short of the threshold for a protein of this size being considered a good quality model.

Fig. 3.15d contains the structure 1N0W (RAD51-BRCA2 BRC repeat complex), which was used by PHYRE2 as the top threading template through this region, generating a very low confidence score of 37% through the residues I1188 – V1201 and a sequence coverage of 36% in this sequence fragment. Despite the confidence generated being good enough to model structure from, and the number of identities being 43%, sufficiently this could sufficiently greater than the 'twilight zone' where structural homology may be inferred, this could all be due to the limited sequence coverage provided by the PDB in particular. This is in addition to the fact that the loop structure present within the model is an incredibly common feature amongst proteins, thus the match produced by this threading template onto this portion of the sequence fragment is insignificant.

The remaining models produced by PHYRE2 in this region have lower confidence levels below 10% indicating that even structural inferences possibly made would most likely be wrong.

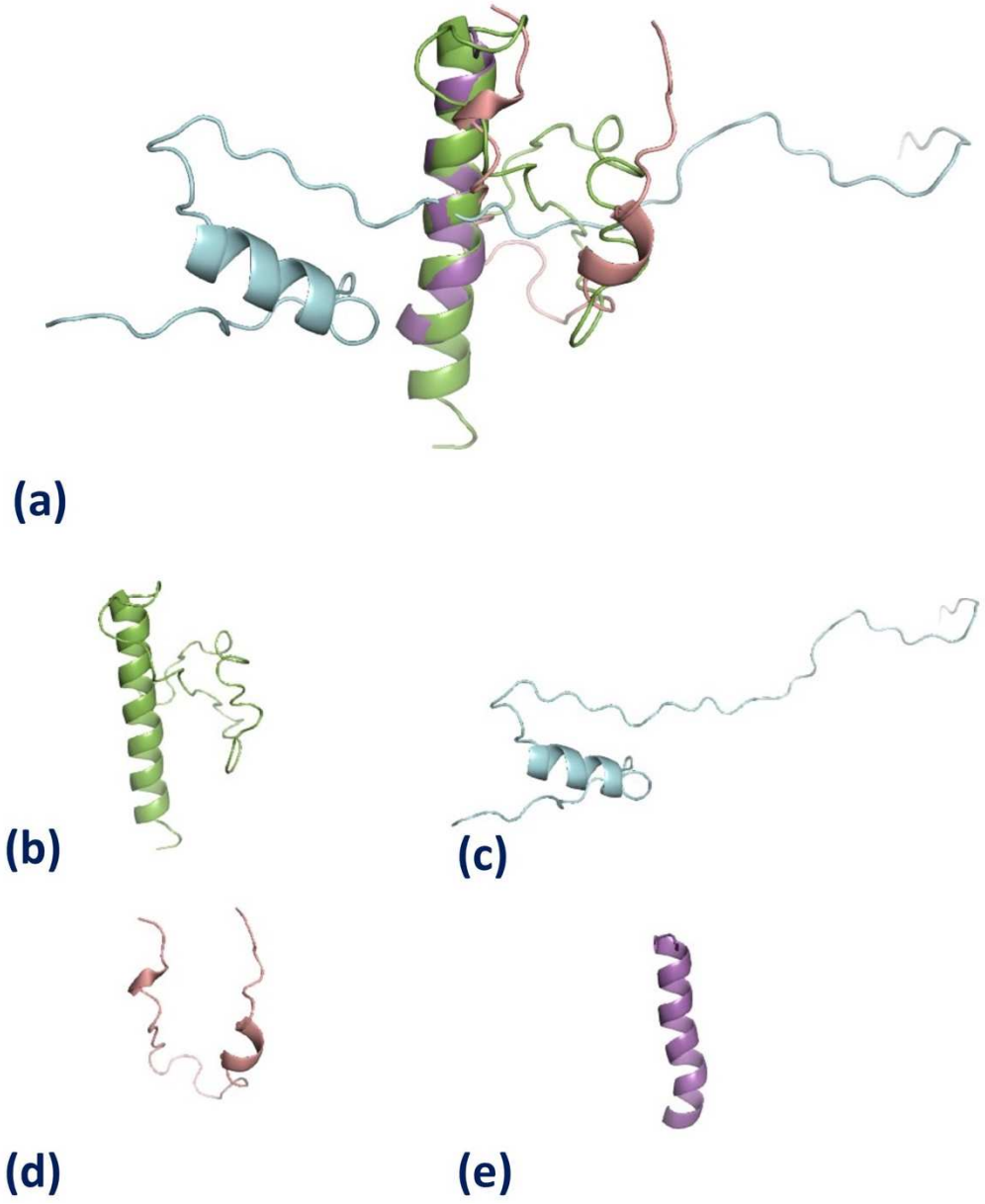

Fig. 3.16: Models produced in the region Q1224 – E1295 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), and PHYRE2 (salmon and magenta). (b) final model produced by I-TASSER using threading template 2RNQ. (c) final model produced by RaptorX using the threading templates 2A75. (d) 4I9X, and (e) 2KZ5 were all used by PHYRE2 as threading templates.

Fig. 3.16b contains the structure 2RNQ (C-terminal acidic domain of TFIIE alpha), which was the top-ranking template used by I-TASSER for residues E1232 – E1295. This structure covered 78% of the query sequence, for which it produced a Z score of 0.72. The final model produced a C score of -2.91 and a TM score of 0.38±0.13. Both Z and C score indicated that I-TASSER lacks confidence in the model produced in this region as the both fell short of their thresholds of 1 and -1.5 respectively. This indicated that it was both a bad alignment between the threading template and the query sequence, in addition to I-TASSER lacking confidence in the final model that was produced. Conversely, when considering the upper bounds of the values generated for the TM score, it indicates that the model has adopted the correct topology, as it would surpass the designated threshold of 0.5. This protein has a lot of loop residues preceding the alpha helix, accounting for more than half of the residues in this region.

Fig. 3.16c contains the structure 2A75 (Trypanosoma rangeli Sialidase In Complex), which was used as a threading template by RaptorX through the full length of this fragment. The alignment produced residues with loop structure and a limited alpha helical structure. The model generated a P value of 3.4×10^{-02} indicting that RaptorX has a moderate amount of confidence in the model it produced within the sequence limits I1188 – I1223, as it failed to fall under the threshold 1×10^{-02} that is associated with a high degree of confidence. Despite the moderate P value, only 22 residues out a possible 72 residues aligned, which fails to reach median values in this region. Given the size of this model, GDT was used to assess the overall quality of the model, generating a value of 36 in this region, suggesting this is a poor model as it fails to reach the threshold for a good quality model (>50). This region is predicted as predominantly loop by RaptorX, sharing

similar structures with Fig. 3.16b. The consensus between different methods helps increase reliability of results.

Fig. 3.16d contains the structure 4I9X (human cytomegalovirus glycoprotein UL141 targeting the death receptor TRAIL-R2), which provided PHYRE2 with a threading model through the residues E1233 – L1267. The template covered 47% of the sequence fragment, while producing a confidence score of 28.5% for this region and 29% identities. Though the confidence produced by PHYRE2 was marginally below 30%, thus this template may still bare structural similarities to the query sequence fragment, it has mapped onto the query sequence with less than 50% coverage.

Fig. 3.16e contains the structure 2KZ5 (Transcription factor NF-E2 subunit's DNA binding domain from Homo sapiens), which was used as the second ranking threading template by PHYRE2 in this region. This template produced a confidence score of 25.3% and 43% identities, when mapping to the query sequence fragment, through residues P1265 – A1288. This alignment provided 30% sequence coverage downstream of 4I9X in Fig.3.16d. the low sequence coverage overshadows the fact that it has produced enough identities for homology modelling.

The top models used as a threading template by HHPRED were 5EQJ (two-subunit tRNA m1A58 methyltransferase) and 5DBR (Ca^{2+} CaM), for fragments 7 and 8 and produced of 27.8% and 61.5% probabilities with the associated structures. Though the PDBs mapped the Adgb sequence through residues S1190 – Q1209 and L1267 – L1281, when evaluating their respective sequences, it was found that there was no structure through the residues that provided a match to the Adgb sequence in the template PDB. It is highly likely that when crystallising the respective proteins, these regions in the

proteins had fully flexible coil conformations which X-ray crystallography inherently had difficulty mapping.

Given the variety of different templates proposed, and the low confidence levels of these fold recognition hits, it is not possible to offer any conclusions as to the structural nature of this region, other than that it is largely a mixture of coil and alpha helical secondary structure.

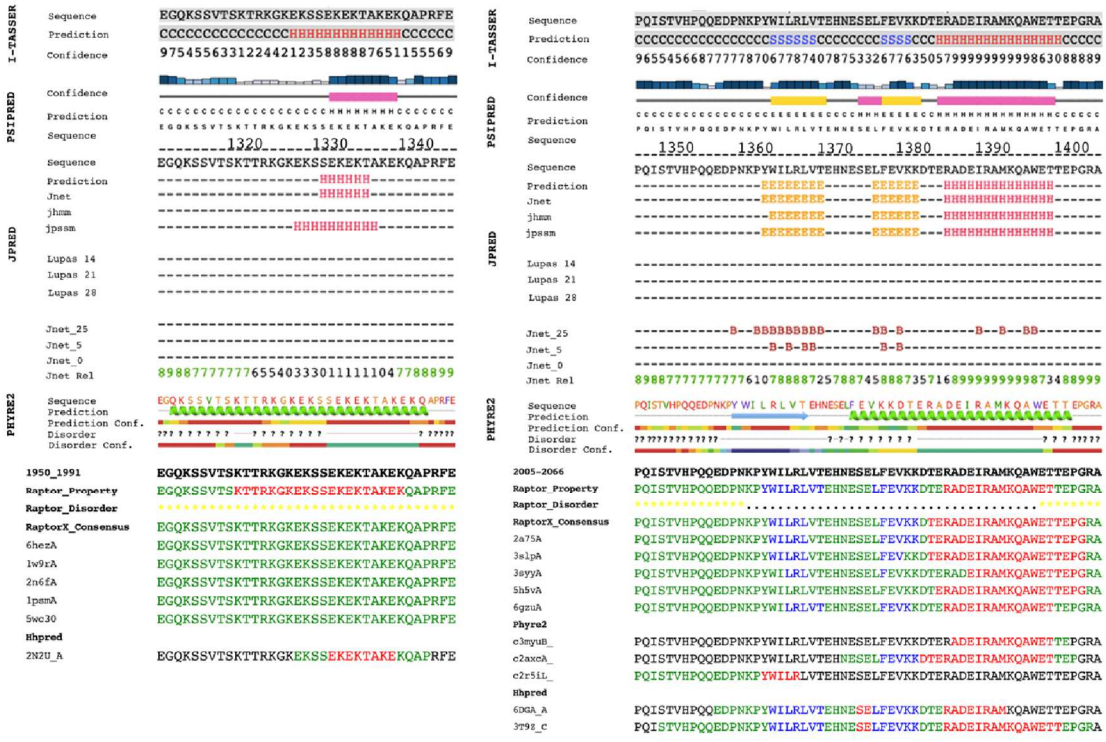

Fig. 3.17: Secondary structures, and results of fold recognition for fragments (17.1) E1309 – E1343 and (17.2) P1344 – A1402

3.3.1.8. MSA Fragment 9 & 10: E1309 – E1343 & P1344 – A1402

Again, fragments 9 and 10 cover a region with a high degree of disorder (Fig. 2.3) and so it is unlikely that any strong hits will be retrieved from the fold recognition. Nevertheless, we will again describe the fold recognition results for this region.

Disorder covers a great deal of the fragment in Fig. 3.17.1 when considering the Raptor Property alignment. Only helix, not beta sheet, is predicted in the region across all the servers, but they differ in the length of the helix being predicted. PHYRE2 predicts the helix spanning nearly the full length of the fragment. There is overlap in the helices from I-TASSER, JPRED, and PSIPRED in the portion of sequence where the PHYRE2 disorder prediction shows the presence of ordered residues. There are high confidence

scores produces by PSIPRED throughout the course of the helices predicted in this region and in the second half of the helix produced by I-TASSER; JPRED produces an exceptionally low confidence score for the helix predicted, despite sharing the same localisation as the helices produced by PSIPRED and I-TASSER.

Fig. 3.17.2 contains far less disorder predicted, when compared with the disorder prediction from the first fragment in Fig. 3.17.1, by both methods. The regions that lack disorder in PHYRE2 disorder prediction shows consensus across the servers for a single beta-sheet and alpha helix of varying length with excellent confidence scores in all the servers presented. These two well defined structures are separated by beta-sheet predicted in all the servers at that position, barring the PHYRE 2. The helix mapped by the PHYRE 2 server overlaps with the beta-sheet predicted by all other servers. The beta sheet predicted in this region by PSIPRED is preceded by a short helix, for which the confidence score dramatically decreases.

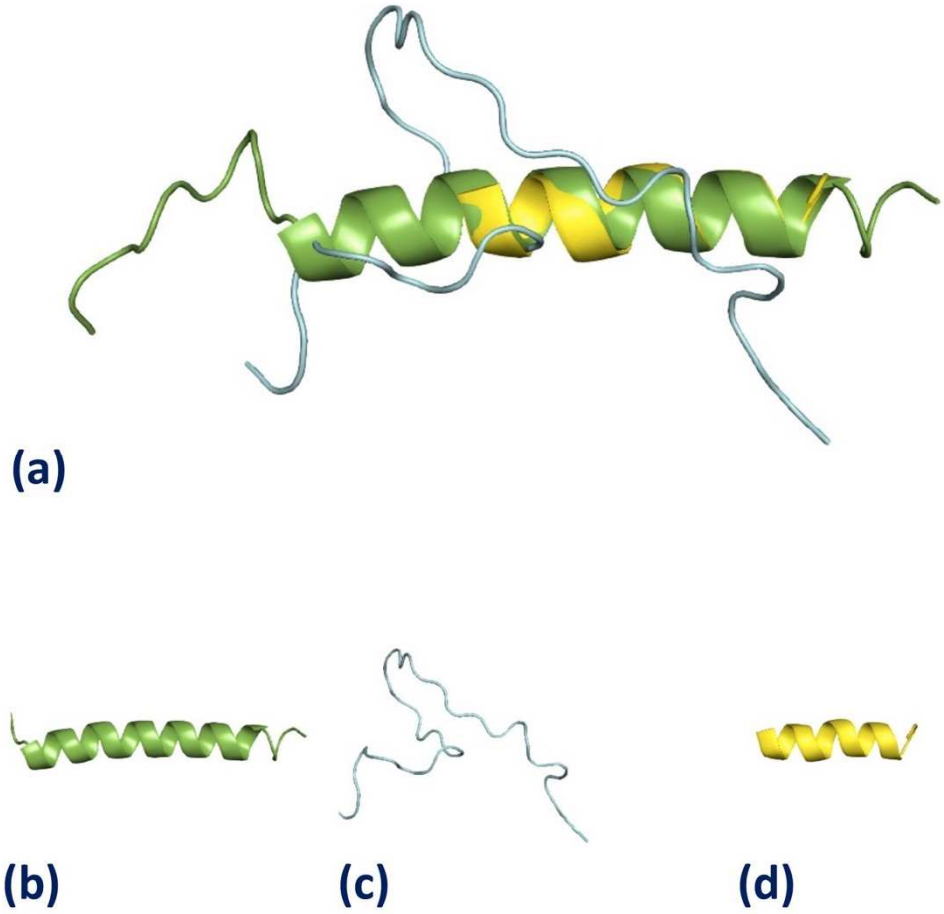

Fig. 3.17: Models produced in the region E1309 – E1343 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 1VX7. (c) final model produced by RaptorX using the threading templates 6EHZ. (d) threading template 2N2U was used by HHPRED.

Fig. 3.17b contains the obsolete structure 1VX7, which was superseded by 3J79 (Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine). When I-TASSER modelled this sequence fragment, 1VX7 was used as the top ranked threading template for the final model, producing 100% sequence coverage of the sequence fragment, for which it generated a Z score of 1.34. This means that I-TASSER

considered the alignment between threading template and query sequence as a good alignment, since the Z score generated was greater than the threshold associated with this descriptor (1). The final model produced a C score of -1.34 and a TM score of 0.55±0.15, these values reflect the confidence and quality of the model, indicating that the alignment for the sequence to the threading template and the confidence in the final model produced were all agreeable. This is as they both generated values greater than their threshold of -1.5 and 05 respectively.

However, due to the limited size of the fragment, and the fact that this is simply a single helix flanked by two loops, it is impossible to infer anything beyond structure throughout this sequence fragment. Many proteins may exhibit single helices for a similar number of turns flanked by coiled coils.

Fig. 3.17c contains the structure 6EHZ (Murine CXCL12 gamma isoform), which was used by RaptorX in modelling the full length of the sequence fragment through residues E1309 – E1343, for which it produced a P value of 4.6×10^{-03} associated with the alignment between the sequence fragment and the threading template. The P value generated for this model suggests Raptor X modelled this structure with a high degree of confidence, as it falls below the 1×10^{-02} threshold. RaptorX produced an alignment score of 5 residues out of a total 35 residues and a GDT score of 47. Though the alignment score seemingly indicates that this alignment is considered poor by the server, as it does not reach median values for a sequence this size despite the P value equating to moderate-high confidence in the model being produced. The GDT score falls marginally below the threshold for a high-quality model for a sequence fragment this size (>50). The confidence level for this model become ever more insignificant in light of the coiled

coil conformation that this model adopts, as there is nothing inherently unique about the fold, all proteins will have coiled regions in their structure at any one point in time.

Fragment 9, bound by the residues P1344 – A1402, failed to provide a threading template in PHYRE2 that could be evaluated in this region.

Fig. 3.17d contains the structure 2N2U (Ferredoxin Fold PROTEIN sfr3), which was used by HHPRED as a threading template for this sequence fragment. This template produced a poor alignment, denoted by the probability of 2.9% and producing 44% identities when mapped to the sequence fragment by HHPRED. In addition, the limited amount of structure and the low sequence coverage offered through residues E1325 – P1340, overshadows the high number of identities produced. Furthermore, the sequence coverage only accounts for half of the residues within an already small sequence fragment and the probability generated by the alignment, suggests it is unlikely that even the structure modelled has any biological significance.

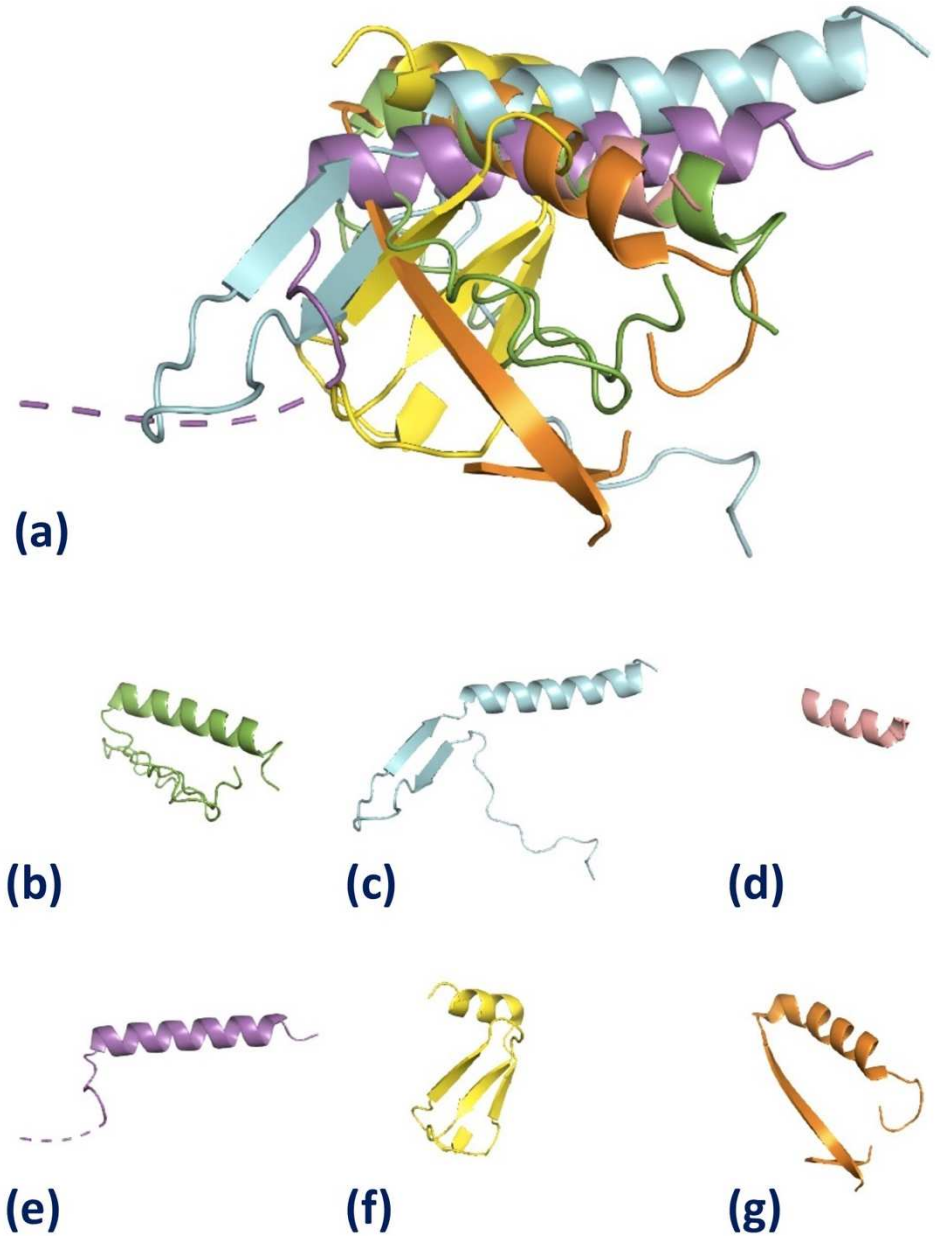

Fig. 3.18: Models produced in the region P1344 – A1402 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon and magenta), and HHPRED (yellow and orange). (b) final model produced by I-TASSER using threading template 6O91. (c) final model produced by RaptorX using the threading templates 2A75. (d) 3MUY and (e) 2AXC were all used by PHYRE2 as threading templates. (f) 6DGA and (g) 3T9Z were all used by HHPRED as threading templates.

Fig. 3.18b contains the structure 6O91 (Horse liver L57F alcohol dehydrogenase complexed with NAD and pentafluorobenzyl alcohol), which functioned as a threading template for I-TASSER, providing 97% sequence coverage for sequence fragment 10. This model generated a Z score 0.73 which is sufficiently below the threshold of 1, to conclude that the alignment between threading template and structure was poor. The corresponding final model built by I-TASSER had a C score of -2.09 and a TM score of 0.47±0.15. The C score and TM score for the alignment between this threading template and sequence fragment, fell short of their thresholds of -1.5 and 0.5 respectively. This means that I-TASSER lacks confidence in the final model produced., however, the TM score generated for this model suggests that the model has adopted an appropriate fold and does not exhibit random similarity when compared to the native folded structure of the threading template.

Similarly to Fig. 3.17b, the model only shows a helix flanked by two coiled coil, though the coiled in this model was significantly longer than the produced in Fig. 3.17b. however, the loop-helix-loop topology observed in both these models is a feature of a lot of proteins, thus it is impossible to draw a conclusion beyond basic conclusions that may be made about structure in this region. However, most notable about this prediction is the fact that it has a long-coiled region proceeding the helix, is in a region of the sequence where secondary structure prediction had 2 beta-sheet strands.

Fig.3.18c contains the structure 2A75 (Trypanosoma rangeli Sialidase In Complex), which was the top threading template used by RaptorX through the full length of the sequence fragment P1344 – A1402. This alignment generated a P value of 5.9×10^{-03} for the final model structure produced by the alignment, which correlated to high confidence in the structure produced as it falls below the 1×10^{-02} threshold for high

confidence. Out of 59 residues this threading template produced an alignment score of 32, indicating that the server considered it a moderate alignment in terms of quality, which when coupled with the high confidence denoted by the P value makes this a good alignment. Despite this, the GDT value fell short of the threshold for a sequence fragment this size (>50), with a value of 42 generated for this protein, meaning RaptorX considered it is a poor-quality model because the alignment marginally missed the threshold the high confidence generated by mitigate the poor quality of the model,

This fragment begins with a loop region covering approximately a quarter of this sequence. There is a single helix toward the C terminus of this fragment, and a 2 anti-parallel beta-sheet strands preceding this helix. The two anti-parallel stands could easily be accommodated with the long region with consecutive residues in a loop conformation at the beginning of the helix in Fig3.9.2b, making the two structure fairly analogous.

Fig. 3.18d contains the structure 3MUY (Mycoplasma genitalium MG289), which was 1 of 3 threading templates chosen from a number of template models mapped to this sequence fragment by PHYRE2.This template generated the highest confidence level of the 3 models chosen with 60.7% through the residues A1384 – E1398. This model is sufficiently above the 'twilight zone' where homology may be inferred as it produced 47% identities.

The low sequence coverage of 23% offered by this template mode, overshadows the moderate-high confidence produced and the high number of identities across the alignment. It remains impossible however, to infer anything beyond helical structure with some loop residues from this model in the absence of experimental data.

Fig. 3.18e contains the structure 2AXC (ColE7 translocation domain), which was the second ranked threading model used by PHYRE2 in this sequence fragment, mapping

though N1370 – P1399. This model gave greater sequence coverage of 49% than 3MUY in Fig. 3.18d. This was at the cost of the model producing a lower confidence score of 31.7% and lower number of identities, producing only 20% conserved residues at specific positions on the alignment.

Fig 3.18f contains the structure 6DGA (Cronobacter turicensis RpfR quorum-sensing receptor RpfF interaction domain), which was used as a threading template by HHPRED and mapped through the residues E1354 – M1390 in the sequence fragment. This alignment produced a probability of 49.5% and 27% identities between the sequence fragment and threading template. In addition, the number of similar residues between query sequence structure and templates structure, approximately 47.2%. The probability and sequence similarities generated for this model indicates we are able to infer structural similarities.

Fig. 3.18g contains the structure 3T9Z (A. fulgidus GlnK3, ligand-free), which was mapped through the same region as Fig. 3.18f, though doing so with greater amount of sequence coverage through the residues S1356 – R1401. This alignment generated a probability of 28.1% and 15% identities through the mapped region. The model has missing co-ordinates between the two antiparallel beta-sheet, most likely due to the interface between the two structures being fully flexible and thus inherently harder to crystallise. This makes the modelled structures in corresponding region of the sequence fragment unreliable.

Despite this, Figs. 3.18c, 3.18f, and 3.18g share similar structures, with Figs. 3.18c and 3.18g being the most similar to one another. However, all these

structures display antiparallel strand followed by an alpha helix, only difference being that in Fig. 3.18f there 3 antiparallel strands present not 2. The only way to know

if this consensus formed between models produced by HHPRED is true, is to express the fragment and to determine its structure experimentally.

The first fragment is predominantly helical and preceded by extensive loop regions, whereas the second fragment is generally an antiparallel beta-sheet followed by a helix. However, it is not possible to draw a conclusion outside this, as none of the threading templates have produce high level of confidence to prove themselves as significant, in the absence of a consensus on fold formed. Experimental data would be required to truly make a definitive conclusion in these region.

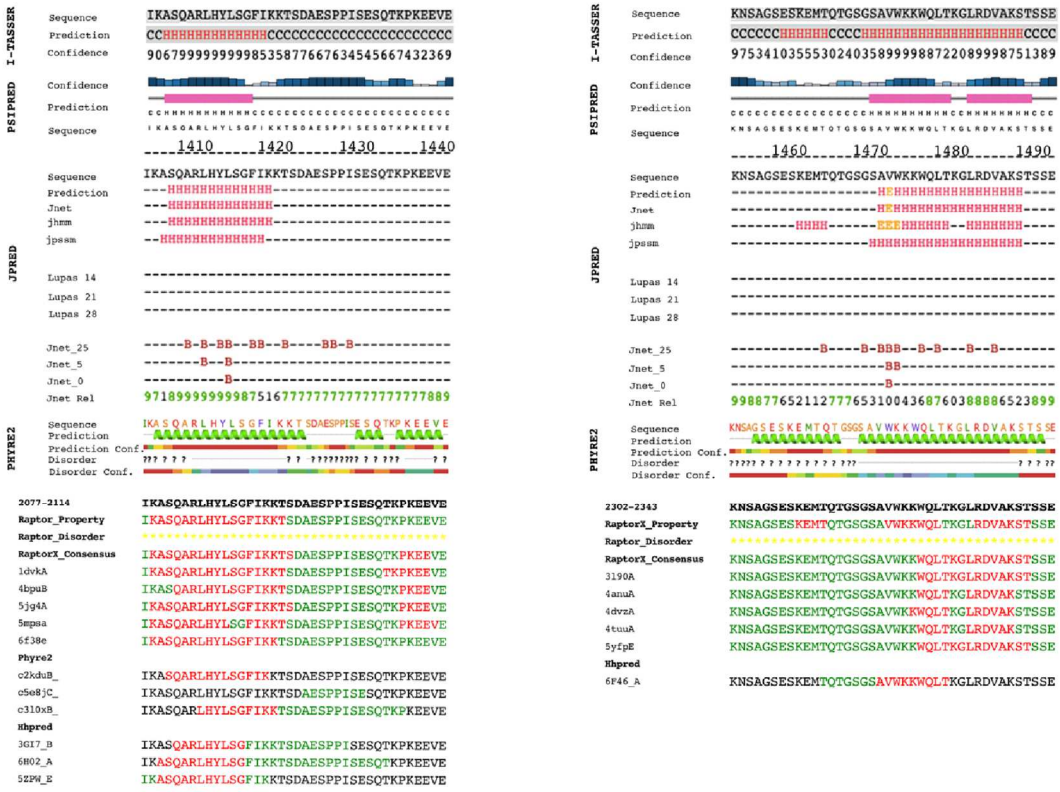

Fig. 3.19: Secondary structures, and results of fold recognition for fragments (19.1) A1402 – E1440 and (19.2) K1452 – E1491.

3.3.1.9. MSA Fragment 11 & 12: A1402 – E1440 & K1452 – E1491

Generally, the distribution of disorder is most like that depicted in Fig. 3.18.1. This is as RaptorX Property predicts the length of both fragments in Fig. 3.19 as fully disordered. PHYRE2 predicted the majority of the fragment as disordered, barring the regions where there is a consensus across most of the servers. It is important to note that the disorder confidence is extremely low through the regions PHYRE2 has predicted as ordered, when compared to the confidence for the predicted secondary structures.

Fig. 3.19.1 has two isolated regions where disorder in PHYRE2 ceases, to be replaced by predicted alpha helical structure. The first of the two alpha helices that reside in an ordered portion of the fragment according to PHYRE2.This first helix produces the most robust consensus across every single server within this sequence fragment; this as opposed to the second alpha helix for Fig. 3.19.1, where a consensus may only be reached between PHYRE2 and RaptorX. The consensus gained from the localisation for this first helix is further supported by the extremely high confidence scores produced by I-TASSER, PSIPRED, JPRED, PHYRE2.

Fig. 3.19.2 shows two alpha helical structures present, the first of which is not supported by PSIPRED or RaptorX. The confidence scores produced is higher towards the end of this same helix in PHYRE2, where it overlaps with the helix produced in ITASSER and JPRED; I-TASSER has moderate confidence in this region where consensus is gained. The second helix in this fragment gains a greater consensus than the first, barring PSIPRED, a number of RaptorX models, and Raptor Property generating two separate helices separated by two, two, and four loop regions respectively. Though ITASSER generates a continuous helix, the confidence scores in same loop residues is extremely low. The two termini of this helix are also marred by low confidence scored generate by I-TASSER, PSIPRED, JPRED, and a PHYRE2 across a number of residues.

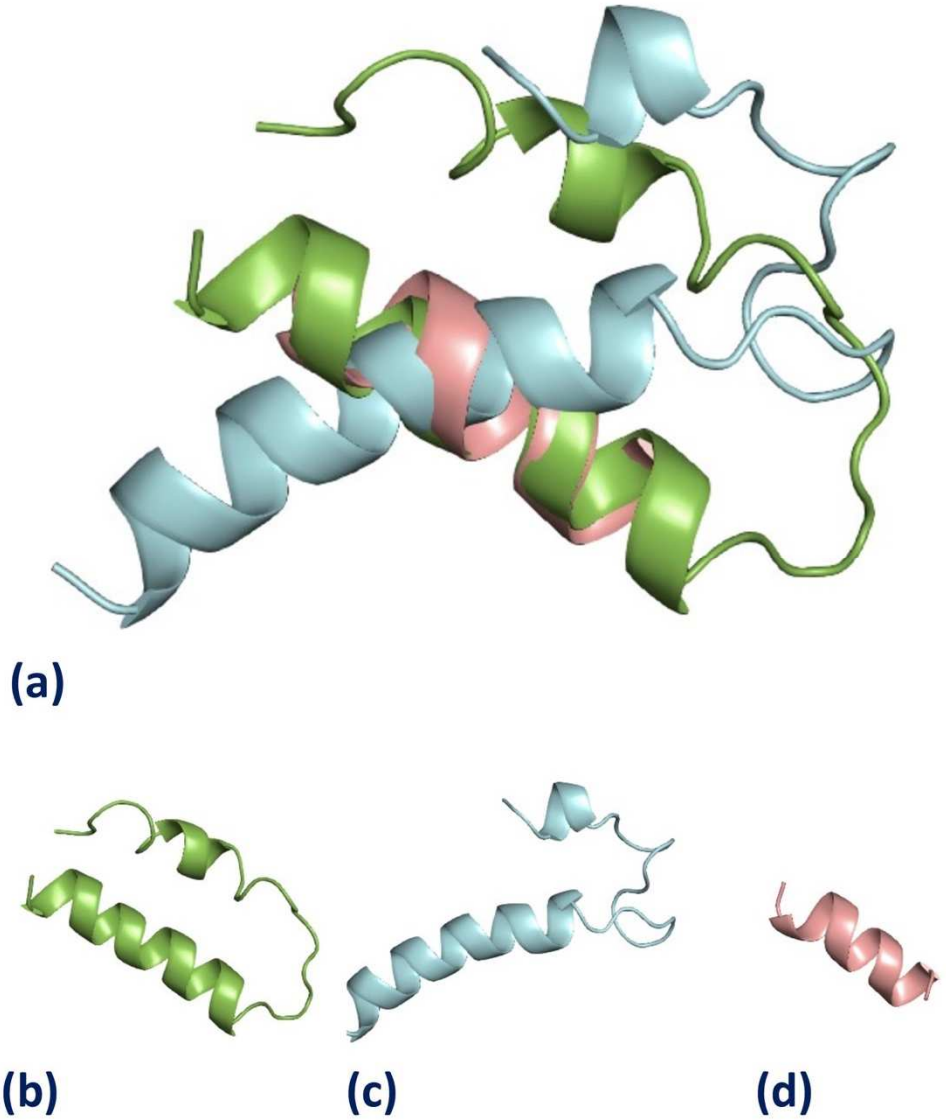

Fig. 3.20: Models produced in the region A1402 – E1440 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), and PHYRE2 (salmon). (b) final model produced by I-TASSER using threading template 4JIO. (c) final model produced by RaptorX using the threading templates 5JG4. (d) threading template 2KDU was used by PHYRE2

Fig. 3.20b contains the structure 4JIO (Bro1 V domain and ubiquitin), which was used by I-TASSER as a threading template when modelling the full length of this sequence fragment through the residues A1402 – E1440. It produced an incredibly low

Z score of 0.66, falling short of the its threshold value of 1. The final model in turn produced a C score of -1.61 and an associated TM score of 0.52±0.15 when using this threading template, indicating that while the confidence produced for the final model was marginally below by 0.11 from what would have been considered a good amount of confidence. However, the TM score generated by this model indicates that I-TASSER considered this model to have an appropriate fold when compared to the natively folded threading template.

Fig. 3.20c contains the structure 5JG4 (Effector protein LpiR1), which was the top-ranking model used by RaptorX for this sequence fragment, aligning to the full length of the 38-residue fragment where it produced a P value 1.1×10^{-1}. This indicated a degree of uncertainty, as it failed to generate a value below the threshold value of 1×10^{-1}, suggesting that 1/10 randomly generated structures will adopt a similar fold. However, in spite of the poor P value generated for this domain RaptorX produced an alignment score of 16 out of 38 and a GDT of 54, indicating that this alignment was considered to be moderate through this region but that the model was considered high quality, as the alignment score was near median values in this region and GDT exceeded the threshold assigned (>50).

This fits conformation of model generated in Fig. 3.19.1 where there are two helices in this region at the opposing termini; the first helix is significantly longer than the second.

Fig. 3.20d contains the structure 2KDU (A novel 1-26 calmodulin binding motif with a bipartite binding mode), which was the top ranked threading used by PHYRE2 with a confidence score of 12.9% mapping to the Adgb sequence through the residues S1405 – K1418, providing 31% coverage for the sequence fragment.

Despite the high sequence identities of 38% produced by the alignment, the low sequence coverage in addition to the low confidence generated by the top ranked threading template in this region suggests this portion of the sequence is a hard target to model. Furthermore, the remaining threading templates that have been produced in this region are unlikely to provide any new functional information or any structural information that is not already present within the secondary structure analysis in this portion of the sequence, as the confidence levels associated with those structures decrease further with each model. The only conclusion that would be evident would be the characterisation of secondary structure in this region, structure that has already been fully explored in Fig. 3.19.1.

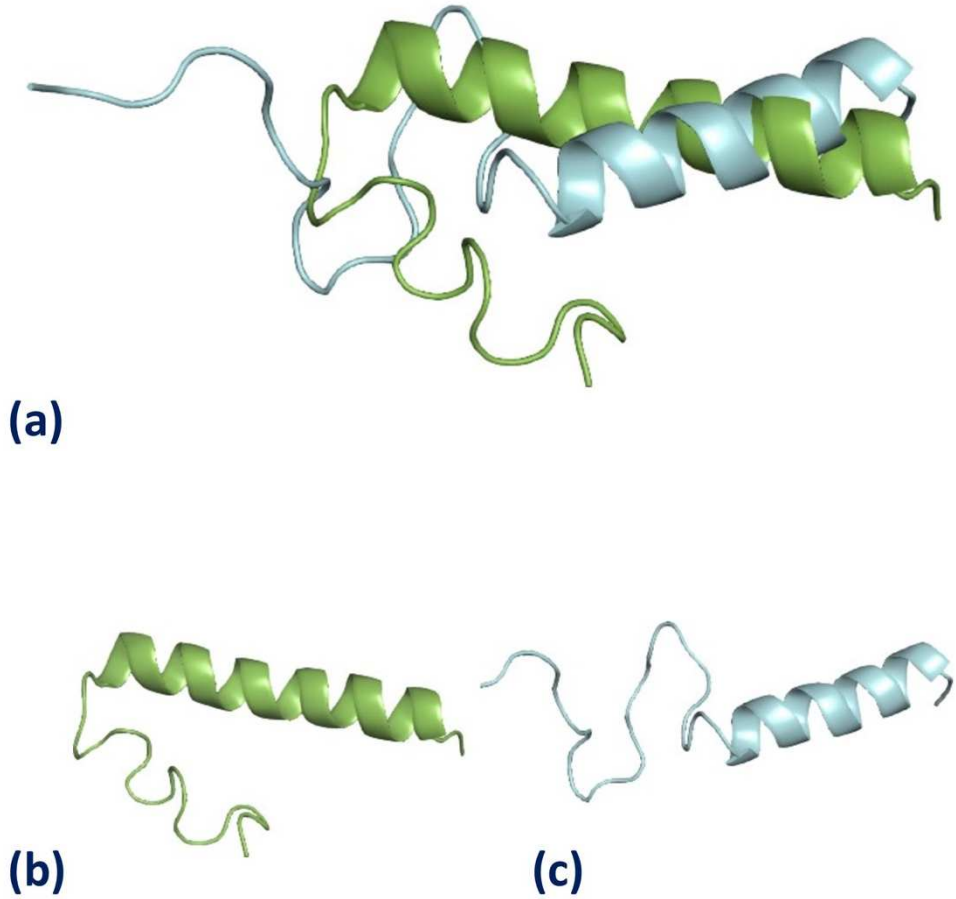

Fig. 3.21: Models produced in the region K1452 – E1491 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green) and RaptorX (cyan). (b) final model produced by I-TASSER using threading template 1IHQ. (c) final model produced by RaptorX using the threading templates 5YFP.

Fig. 3.21b contains the structure 1IHQ (N-terminus of a rat short alpha tropomyosin with the n-terminus encoded by exon 1b), which was the top threading template used by I-TASSER through the region K1452 – E1491. This template only failed to map to the first two residues in this region. It provided a sequence coverage of 95% and a Z score for 1.13 for the final model, meaning this threading template was considered to have a good alignment in relation to the sequence fragment, as it

generated a value above the Z score threshold of 1. The final model produced a C score of -1.85, which is marginally lower than the -1.5 threshold for what is considered a good confidence in the modelled structure having adopted a correct fold. This structure also produced a TM score of 0.49±0.15, 0.01 short of its 0.5 threshold. This is an acceptable short fall as the standard deviation of the mean value generated exceeds this threshold. The secondary structure shown in this model matches secondary structure prediction produced in Fig. 3.19.2, as both are predominantly helical with a number of coiled residues preceding this helix. Because the fragment is so short, the query is unlikely to share any functional relation to threading template.

Fig. 3.21c contains the structure 5YFP (Exocyst Complex), which was used by RaptorX as a threading template through the residues K1452 – E1491, providing 100% sequence coverage over the residues present in this fragment. The alignment associated with this PDB produced a P value of 4.9×10^{-2} correlating with a moderate amount of confidence, as it falls short of threshold 1×10^{-2} for high level of confidence. The model produced generated an alignment score of 12 out of 40 residues and a GDT score of 43. The alignment score also reflects the relative lack of confidence in the model produced, as it does not reach median values for this domain. Despite this though the value generated for GDT indicates that the quality of the model marginally misses the threshold for what is considered a high-quality model (>50).

The modelled structure fits the conformation of residues shown in Fig. 3.19.2, where there is a single helix mapped in this region, preceded by a number of loop residues.

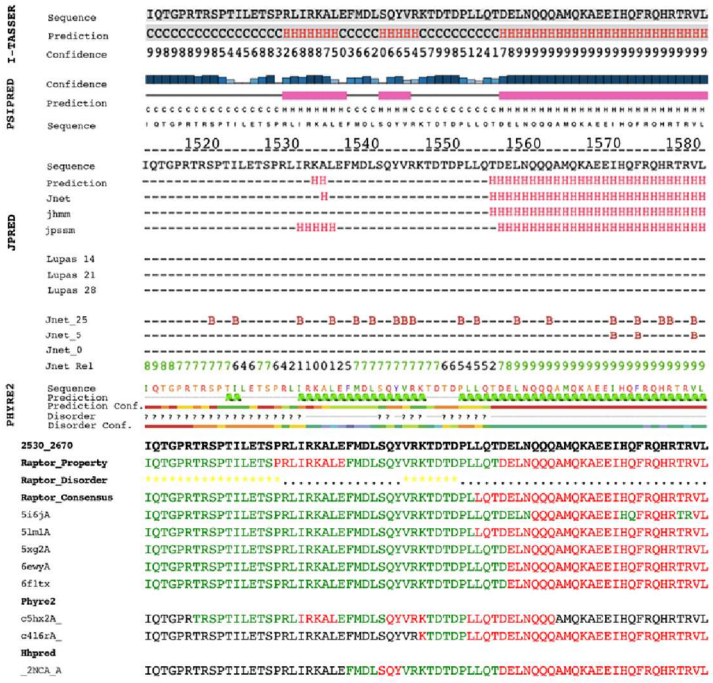

Fig. 3.22.1: Secondary structures, and results of fold recognition for fragments I1512 – A1644

3.3.1.10. MSA Fragment 13: I1512 – A1644

There is consensus across all the servers for the localisation of the first and third helices in this region, with a variation in the length of the helices generated (Fig. 3.22.1). Most of the variation is apparent within JPRED's prediction, where the helix can be seen spanning 2 residues. Further agreement is gained between PSIPRED, ITASSER, the PHYRE2 model 5HX2, and the HHPRED model 2NCA over the positioning of the second helix produced in this region. The first two of the helices produce moderate to high confidence levels in both PSIPRED and I-TASSER, whilst the third produced extremely high confidence. The first and third helix produced a low and high confidence in the JPRED alignment respectively. The general consensus drawn from disorder in this region

is that both PHYRE2 and Raptor Property map disorder onto the region preceding the first helix and loop conformation between the first and third helices in the general structural consensus.

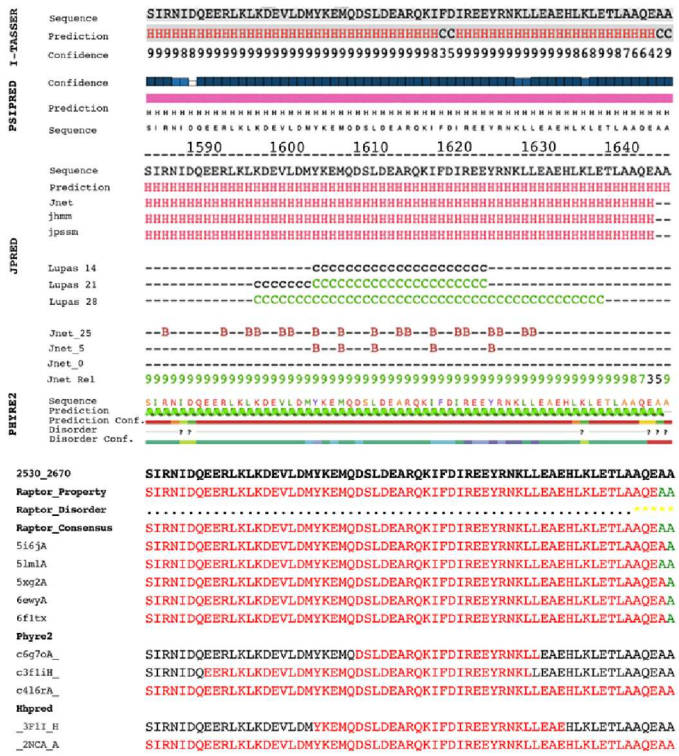

Fig. 3.22.2: Secondary structures, and results of fold recognition for fragments I1512 – A1644

Clear consensus across servers with nearly the whole length of this fragment, save the last two residues, presenting as the helical continuation of Fig. 3.22.1 The secondary structure predicted generates high confidence scores in I-TASSER, PSIPRED,

JPRED, and PHYRE2. This helix seems to solely be interrupted in the alignment generated in I-TASSER for two residues in the middle of this sequence, for which the confidence drops in the same alignment. Disorder prediction generated in PHYRE2 only has high confidence towards the end of this fragment, where there is a consensus with Raptor Property. Before this region in the sequence, the disorder confidence score is depicted as being low to moderate.

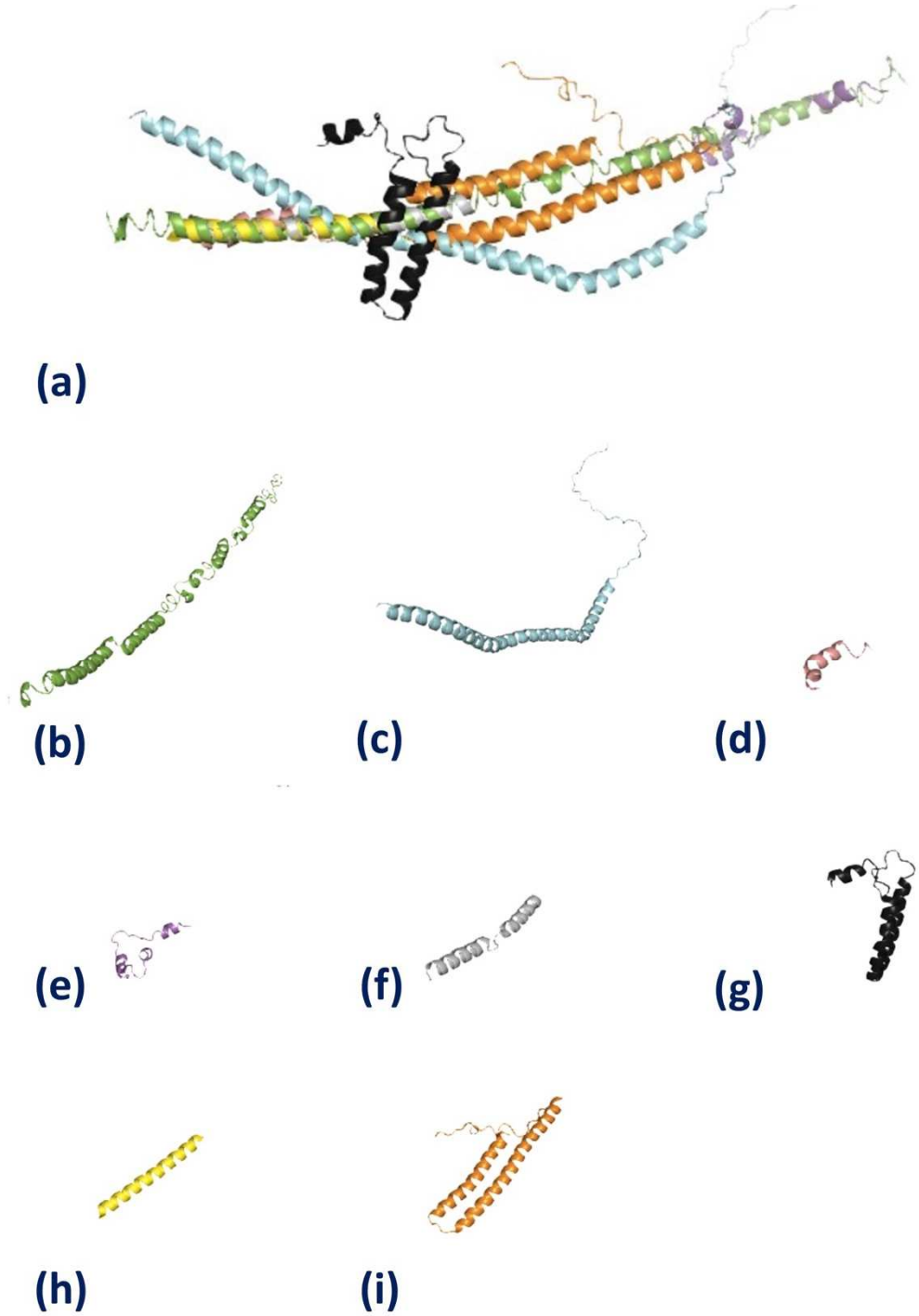

Fig. 3.23: Models produced in the region I1512 – A1644 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon, magenta, grey, and black), and HHPRED (yellow and orange). (b) final model produced by I-TASSER using threading template 5HMO. (c) final model produced by RaptorX using the threading templates 5LM1. (d) 6Q7O, (e) 5HX2, (f) 3F1I, and (g)

4L6R were all used by PHYRE2 as threading templates. (h) 3F1I and (i) 2NCA were all used by HHPRED as threading templates.

Fig. 3.23b contains the structure 5HMO (myosin X motor activity), which was the highest-ranking threading template used by I-TASSER. The alignment had the template aligning to the sequence fragment through the residues I1512 – Q1641, it provided 97% sequence coverage for this fragment, only failing to align with the last 3 residues in this sequence fragment. This model generated a Z score of 1.21, which was higher than the threshold of 1, for a good alignment between the threading template and the query sequence. The C score generated by the threading template for the final model I-TASSER produced, fell short of the threshold (-1.5), with a score of -2.02. This indicates that I-TASSER considers the confidence that the final model has adopted the appropriate fold, below what is considered good. However, the TM score generated of 0.47±0.15 suggests some structural similarity between the natively folded template's structure and the query sequence, as the mean value generated was 0.03 below the threshold for TM score (0.5).

Fig. 3.23c contains the structure 5LM1 (HD-PTP phosphatase in complex with UBAP1), which was the top-ranking threading template used by RaptorX in this region. This threading template produced a moderate alignment score of 75 residues out of a possible 133 residues in this region, as such, it exceeded median values for a sequence fragment of this length. The model generated a P value of 1.5×10^{-2} for the region, indicating a moderate amount of confidence associated with the probability that a randomly generated model could not exhibit the same structure. This was as the P value produced did not fall under the threshold of 1×10^{-2} that is associated with a high degree

of confidence. The same threading template also produced a uGDT of 54 in this region, exceeding the threshold for a good quality alignment for a sequence this size (>50). This indicated that RaptorX considered this a high-quality model for a domain this size. The structure shown can be seen matching the consensus gained from structural alignments of this being an unusually long helical region preceded by a large number of loop residues. It is unlikely this model will yield any functional information, as the probability associated with this match is too low, but it will remain useful from a structural perspective especially as it supports the consensus formed in the secondary structure predictions.

Fig. 3.23d contains the structure 6Q7O (human alkaline ceramidase 3), which was the top threading template used for this sequence fragment by PHYRE2, providing limited sequence coverage through the residues D1607 – L1628. This only amounted to 15% sequence coverage for which a confidence score of 52.9% was generated. The low sequence coverage almost overshadows the moderate confidence and sequence identities of 27% being generated for this model. This is in addition to the limited amount of structure present in the portion of the PDB used as a threading template, being a generic feature proteins as a whole, means this alignment fails to hold any biological significance.

Fig. 3.23e contains the structure 5HX2 (Baseplate wedge protein gp7), which was the second highest ranking threading template used by PHYRE2. This template mapped to the region preceding fig.3.23d through residues T1518 – Q1561, producing a sequence coverage of 32% and a confidence level of 43.5% for the modelled structure. The larger sequence coverage offered by this model increases the reliability for the model produced, however, it is still unlikely there is a functional relationship between the

sequence fragment and the threading template in light of the 17% identities generated being particularly low. The model produced is still just a small portion of the sequence when compared to the full length of the sequence fragment.

Fig. 3.23f contains the structure 3F1I (Hepatocyte growth factor-regulated tyrosine kinase substrate), which was the third ranking PDB used as a threading template by PHYRE2 in this sequence fragment. The template was mapped through residues E1589 – L1627, generating a confidence level of 40.1% and 41% identities. Though this model produced a moderate amount of confidence and identities. The only real conclusion that may be drawn from this structure is the presence of alpha helices, which stands to reason, as secondary structure prediction methods have already predicted this region as largely helical.

Fig. 3.23g contains the structure 4L6R (Soluble cytochrome b562 and Glucagon receptor chimera), which was the 5th ranking threading model used by PHYRE2 in this region. This template was chosen for the fact that is displays the largest sequence coverage of any threading model in this sequence fragment. The model mapped through the residues K1545 – A1640, providing 70% sequence coverage, with a confidence level of 36.5% and 17% identities. Both the confidence and identities generated are low. This model seems to follow the clear consensus being formed by the products of fold recognition techniques in this sequence fragment being predominantly helical in nature, with interspersed residues with loop conformations. This is despite the consensus from secondary structure predication characterising the region as one long continuous alpha helix.

Fig. 3.23h contains the structure 3F1I (Hepatocyte growth factor-regulated tyrosine kinase substrate), which was the top threading model used by HHPRED in this

region, aligning through residues M1601 – E1631 in this fragment, for which it produced a probability of 73.9%. This model also produced 40% identities and 59.9% similarities between the two sequences when aligned. Despite being the same PDB that was used as a threading template produced by the PHYRE2 prediction in fig. 3.23f, the probability that a randomly folded protein will adopt the same fold as fig. 3.23h is almost twice that of fig. 3.23f, despite the model in PHYRE2 providing a larger sequence coverage through this sequence fragment. The higher sequence similarities, identities, and confidence may arise from the fact that fig. 3.23f covers almost twice the number of residues as fig. 3.23h

Fig. 3.23i contains the structure 2NCA (Hsp90 co-chaperone Cdc37), which was the second threading template produced by HHPRED in this region of the sequence. This template provided greater sequence coverage than the model displayed in fig. 3.23g, through the residues F1537 – E1642 in addition to producing a probability of 40.1% in this region. The model also generated 18% identities and 24.6% residues similar between the query and template sequence, all whilst fitting the general consensus of the fold recognition servers in this region.

Structurally this seems to be the most credible model produced by HHPRED, as almost the full length of the threading template was utilised in the modelling of this region. The fold recognition for this region did not a produce clear consensus amongst the different methods, nor did its present models with a high confidence. Two factors make fold recognition for this fragment difficult, namely the short length of the fragments, meaning that the fold recognition results do little more than support the secondary structure prediction methods and that the prevalence of disorder in this region makes fold recognition difficult.

3.3.2. Disordered prediction fragments

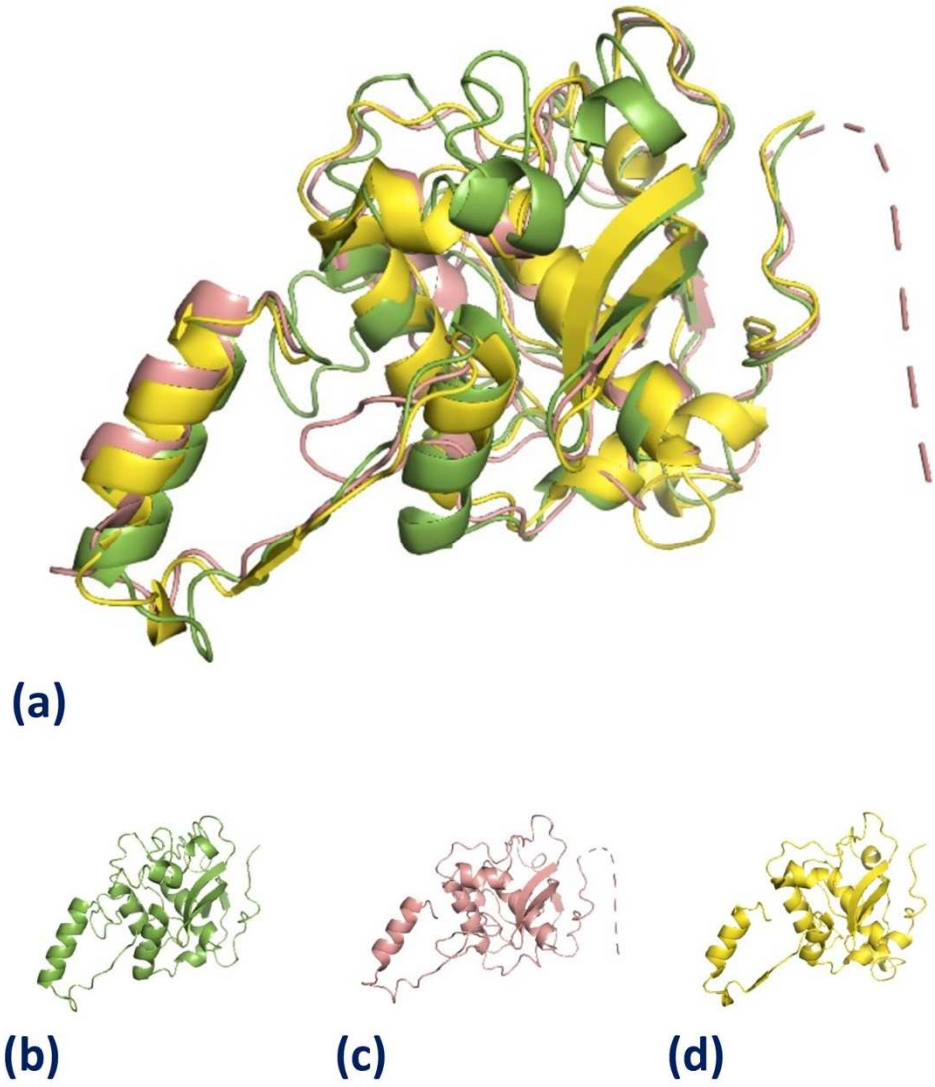

Fig. 3.24: Models produced in the region A1402 – E1440 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 1KFU. (c) threading template 1QXP was used by PHYRE2. (d) threading template 1QXP was used by HHPRED.

3.3.2.1. Disorder Fragment 14: H80 – L300

Fig. 3.24b contains the structure 1KFU (Human m-Calpain Form II), which is the same threading template used in Fig. 3.5b, was once again used by I-TASSER as a threading template in this region, providing a C score of -0.31 and TM score of 0.67±0.13. Both the values produced were only marginally better than the previous alignment produced in I-TASSER, as such these values still correlate with a high degree of confidence, as they post surpassed their respective thresholds of -1.5 and 0.5. This indicated that the structure that has been modelled adopted an appropriate fold and was structurally similar to the natively folded threading template. This threading template also produced a Z score of 1.71 associated with the alignment gained between the threading template and the query sequence fragment, as it generated a value greater than the Z score threshold of 1. Furthermore, this structure was mapped with over the 83% sequence coverage of the query sequence. This structure however contains 3 more beta-sheet strands than in Fig. 3.5b, where only single pair antiparallel beta-sheet strands are present. This addition in secondary structure maybe the reason for the slight change in values when compared to Fig. 3.5b.

Fig. 3.24c contains the structure 1QXP (human calpain), which has been used as the top-ranking threading template used by PHYRE2 in this region, in a similar fashion to fig. 3.5e, it mapped to the sequence fragment with 100% confidence. Though the full length of the sequence fragment was used in the alignment between sequence fragment and threading template, the alignment produced a number of gaps despite the 23% identities, which meant this threading in reality gave approximately 99% sequence coverage. This alignment aligned 183 residues out of 221, which is the same number of

residues aligned, relative to the length of the sequence fragments produced, as were aligned in fig. 3.5e.

Fig. 3.24d contains the structure 1QXP (human calpain), which was used as a threading template by HHPRED, producing probability 99.95% and 22% identities of the full length of the sequence fragment. There were a number of gaps present within the sequence alignment, resulting in 181 residues being aligned through this region.

Once again in this sequence fragment, much like figs. 3.5, the fold recognition methods used have all used calpain threading templates to model the query sequences throughout this region. In addition to this, the mere fact that PHYRE2 and HHPRED, for this sequence fragment, have aligned the same threading template to this sequence fragment further increases reliability in the confidence generated by each fold recognition technique alone.

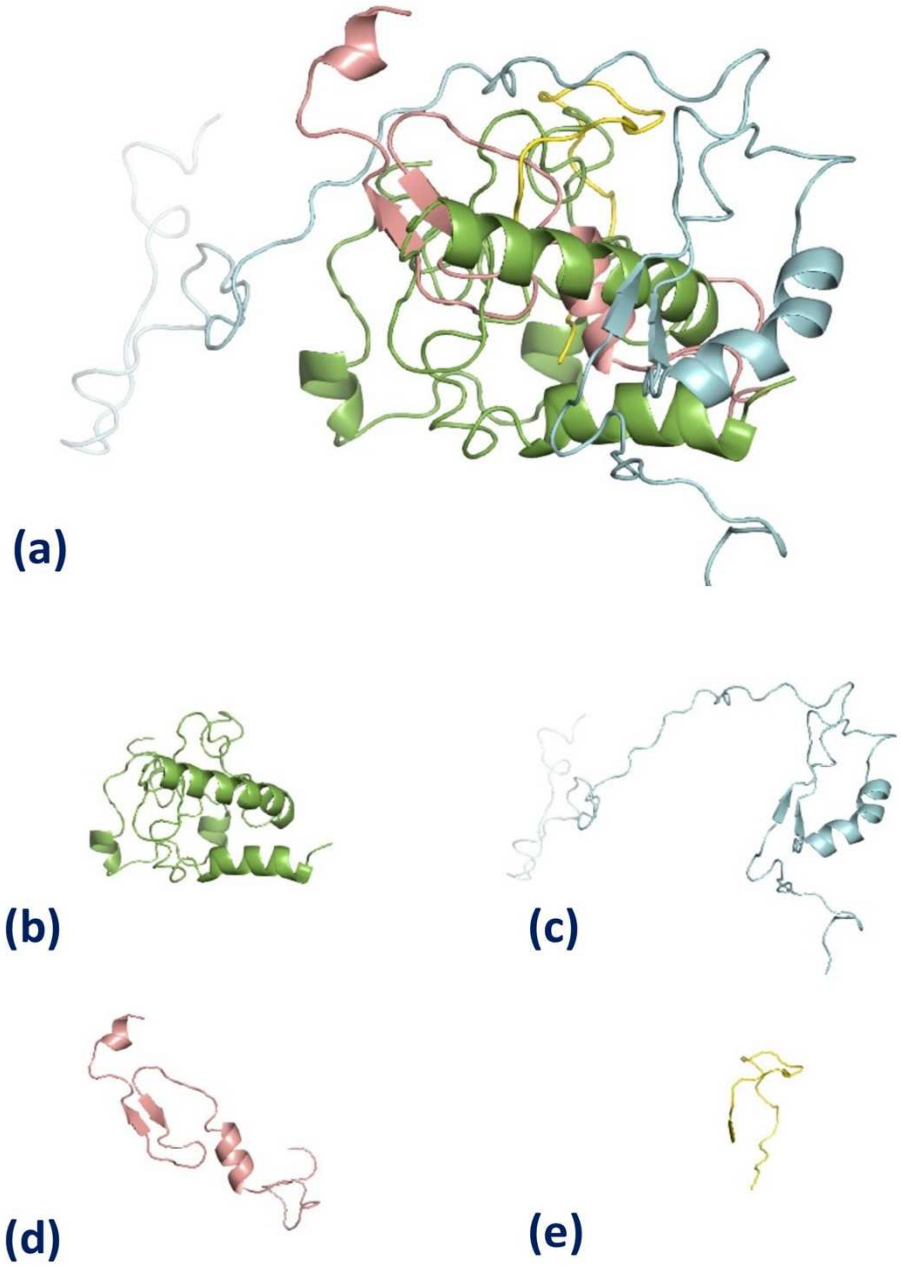

Fig. 3.25: Models produced in the region S390 – F520 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 6N2B. (c) final model produced by RaptorX using the threading templates 1IV8 and 6GQ3 (DOM1 and DOM2 respectively). (d) threading template 6BSV was used by PHYRE2. (e) threading template 3A16 was used by HHPRED.

3.3.2.2. Disorder Fragment 15: S390 – F520

This region corresponds to the second ordered region prior to the heme. There are no published accounts of the plausible structure of this region. This region roughly corresponds to fragment 4, which gave no clear consensus, but the shorter fragment considered here may help to give clearer results.

Fig. 3.25b contains the structure 6N2B (Caldicellulosiruptor kristjanssonii Tapirin C-terminal domain), which was the top-ranking threading template used by I-TASSER. This template produced a Z score 1.01 and 97% sequence coverage. This suggest that the alignment between the threading template and the sequence fragment was considered a good alignment, as it marginally exceeds the threshold associated with Z score of 1. However, the C score and TM score generated by the final model were -3.92 and 0.29±0.09 respectively. Both scores fell short of the threshold values for good scores of -1.5 and 0.5 respectively. This indicated that the confidence generated by this model was poor, and even though the TM score does not fall below the 0.17 threshold that would suggest that this final model produced displayed random similarity to the threading template, the lower bound of the range of values of the TM score is only 0.03 from this value.

It is unlikely that the model produced in this region has any structural relevance to the sequence fragment, let alone functional relevance. As such the model created for this region by I-TASSER in fig. 3.9b was probably more reliable.

Fig. 3.25c contains the final model produced by 2 threading templates. The first of the two templates utilised, 1IV8 (Maltooligosyl trehalose synthase), was the top-ranking threading template produced by RaptorX in DOM1, generated a P value of

2.9×10^{-2} associated with a moderate amount of confidence. This is as the P value generate failed to exceed the threshold of 1×10^{-2} associated with high confidence. This structure was mapped trough the region S390 – R463, for which this template provided 100% sequence coverage. The model produced a GDT score 49, which is marginally below the threshold for a good quality model produced in a domain this size(>50), a better score than the GDT score produced through the region modelled in fig. 3.9c by the exact same threading template. This alignment produced an alignment of 25 residues out of 74, making the alignment low-moderate. However, the size of the domains differs by 6 residues and may possibly explain the discrepancy displayed in the GDT scores.

The second of the 2 templates utilised, 6GQ3 (Asparagine synthetase glutamine-hydrolyzing), was the top-ranking threading template produced by RaptorX for DOM2, generated a P value of 4.7×10^{-2}. This marginally exceeded the 5×10^{-2} threshold associated with a moderate amount of confidence in the structure modelled. The threading template was used to model through the region S464 – F520, providing 100% sequence coverage in this domain. The GDT score of 32 fell significantly below the threshold for a domain this size (>50), suggesting that this is a poor-quality model. This is a conclusion supported by the alignment of 13 residues out of a total 56 residues in this domain, which does not even reach median values for domains this size, suggesting that the alignment between the query sequence and the threading template was particularly poor.

Fig. 3.25d contains the structure 6BSV (Xyloglucan Xylosyltransferase), which was the top threading template utilised by PHYRE2 in this region. However, remarkably in this instance, it has produced 25% identities and a confidence level of 79.5%. This

amounts to over double confidence that that was generated in the corresponding fig. 3.9d which used the same template. This model was only mapped onto a portion of this sequence fragment between the confines of 418Y – 470P, providing the same amount of sequence coverage of 29% as that evident in fig. 3.9d.

The rise in confidence produced by this template suggests this model is far more significant than first concluded in fig. 3.9d, as the confidence suggests this model may inform on more than just structure through this region. In addition, the revision of the residues that bind this particular sequence fragment lead to the spike in confidence for this structure produced. This revision was only made in the light of the disorder predictions; this stand as testament to how integral it is to remove residues prone to disorder, in aims of removing noise and focus fold recognition techniques.

Fig. 3.25e contains the structure 3A16 (Aldoximine dehydratase) is an identical hit to the hit in fig. 3.9g, which again presents as mostly loop conformation with the beginnings of a beta-sheet. This was the only threading template used by HHPRED through this region producing a probability of 24.7% and 35% identities through the region (459V – 481L). Though probability is lower than the alignment in fig. 3.9g, the model in fig. 3.25e produces more identities. Despite this, the model offers too little sequence coverage to provide information on the functionality of this sequence region. There is a limited amount of structure that is present in this model aside from loop residues, so it is unlikely that this model may be able to inform on any structural elements through the region either.

None of the hits in this section give clear structures: the majority of hits contain a significant number of loop residues. This suggests that the disorder predictions given

in Fig. 2.1 that predict more disorder (e.g. IntFOLD5) are probably more correct than those that predict less disorder (e.g. DISOPRED3).

3.3.2.3. Disorder Fragment 16: T612 – N781

This fragment corresponds exactly to Fragment 5 discussed in section 3.3.1 and so the work is not repeated here.

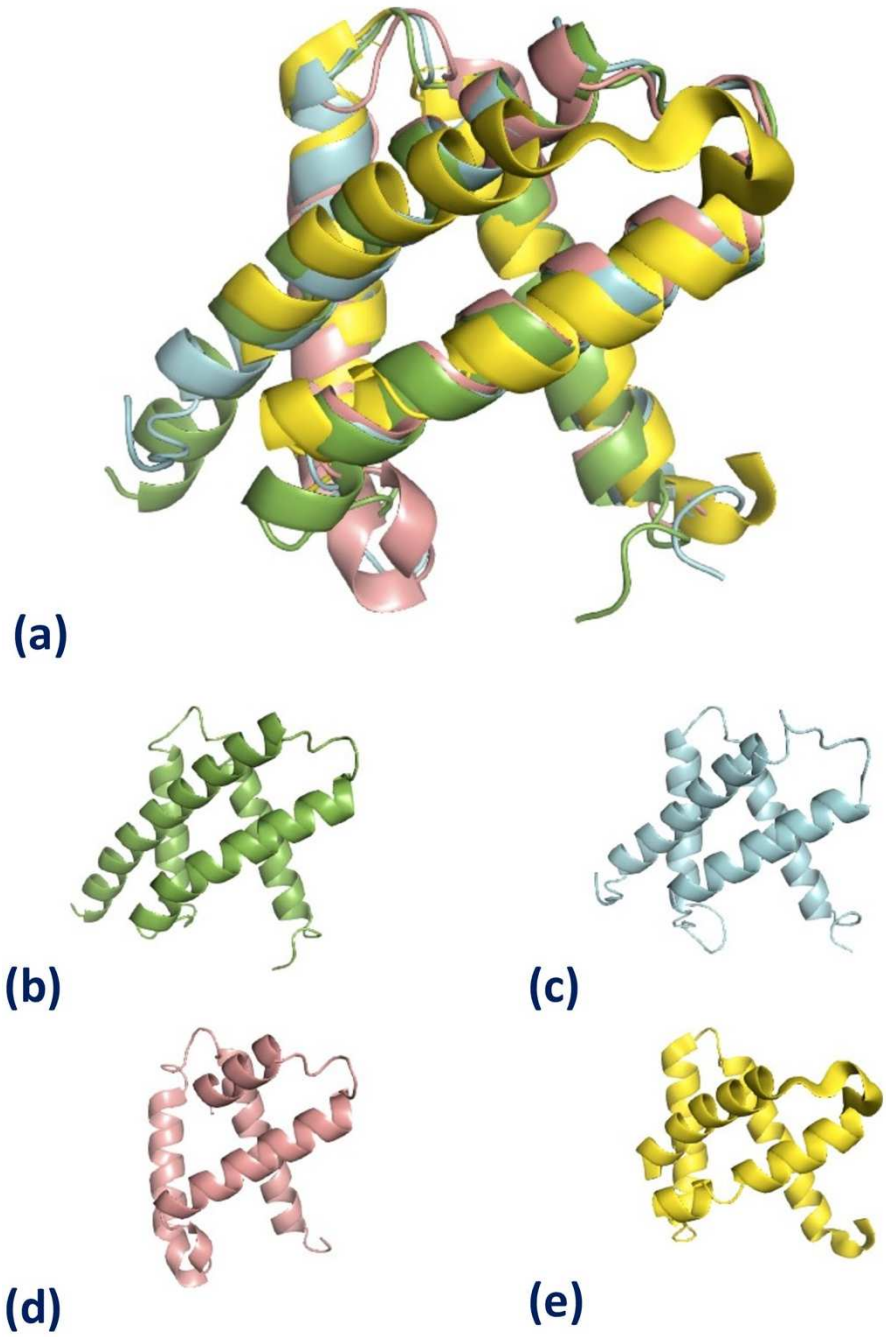

Fig. 3.26: Models produced in the region F782 – D890 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), RaptorX (cyan), PHYRE2 (salmon), and HHPRED (yellow). (b) final model produced by I-TASSER using threading template 1JF3. (c) final model produced by RaptorX using the threading template 1HBG. (d) threading template 1Q1F was used by PHYRE2. (e) threading template 4HRR was used by HHPRED.

3.3.2.4. Disorder Fragment 16: F782 – D890

These residues correspond to the first domain in the cyclically permuted heme fragment, that contains the calmodulin-binding domain and so function as a control.

Fig. 3.26b contains the structure 1JF3 (Component III Glycera Dibranchiata Monomeric Hemoglobin), which was used as the top-ranking threading template in this region by I-TASSER. This threading template produced a Z score of 1.25 and sequence coverage of 92%. The Z score alone suggests this is a good threading alignment between the query sequence and the threading template, as it exceeds the threshold associated with Z score of 1. The final model I-TASSER produced a C score of -0.42 and a TM score of 0.66±0.13, both of which support the reliability of the model. The C score correlates to a good confidence in the model, and it is likely that this model has biological relevance, as it exceeds the threshold associated with C score of -1.5. The TM score produced suggests that the model produced by this threading alignment is a high-quality alignment, exhibiting a similar fold to the threading template being used, as the mean value generated exceeds the TM score threshold of 0.5. The cyclically permuted androglobin should have 8 helices (A-H), of which two are short. The first domain contains 1 short helix and 4 long helices. I-TASSER has therefore correctly modelled the first androglobin domain (helices E-H). This is the pattern that we see emerging from the other fold recognition servers, as shown below.

Fig. 3.26c contains 1HBG (Glycera dibranchiata haemoglobin), which was the structure used as a threading template by RaptorX in this sequence fragment. The threading template provided 100% sequence coverage and a P value of 2.3×10^{-3}. This

descriptor correlates with a high degree of confidence in the structure being modelled as it falls below the threshold of 1×10^{-2} associated high confidence.

This model also produced an alignment score of 101 out of a possible 109 residues in this sequence fragment, suggesting the alignment between threading template and query sequence was high quality. This is something supported by the uGDT generated in this model of 75 indicating this to be a high-quality model, as it generated a score that exceeds the uGDT threshold for a sequence this size (>50).

Fig. 3.26d contains the structure 1Q1F (murine neuroglobin), which was utilised by PHYRE2 as a threading templates through the region P784 – V870, for which it produced 98.9% confidence and 21% identities. This model presents and 4 helices and half a single helical between the second and third helices mapped to this structure.

Fig. 3.26e contains the structure 4HRR (Scapharca tetrameric haemoglobin), which was used to provide HHPRED with a threading template in this region, aligning to P784 **– N**877 within the sequence fragment. This PDB produced a probability of 99.3% for this region and 13% identities. The model produced shows 5 alpha helical structures through this portion of the sequence, aligning a total 89 residues out of 109 residues possible in this query sequence.

Figs. 3.26 as a whole functioned as controls to assess the accuracy of fold recognition techniques being used, as this fragments formed from residues when the helices D – H of the heme domain in Adgb are known to reside. Only 2 of the fold recognition techniques used, PHYRE2 and HHPRED, successfully produced 5 separate helices for this sequence fragment, which further increases the reliability of all the results produced by these two servers.

In addition, none of the fold recognition methods used failed to positively identify the sequence fragment as homologous to haemoglobin, which increases the reliability in the structures being generated with the consensus formed.

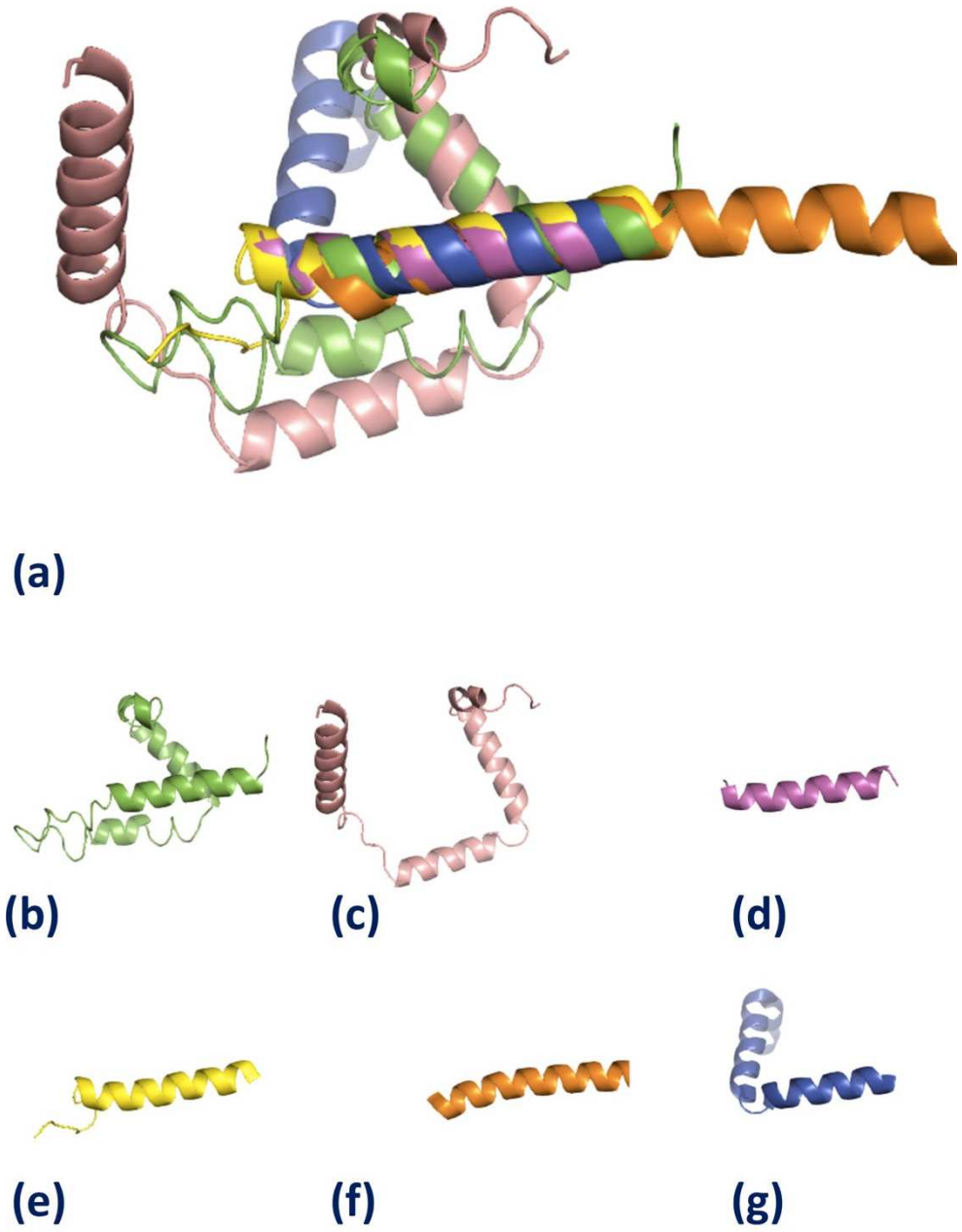

Fig. 3.27: Models produced in the region K900 – Q980 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), PHYRE2 (salmon and magenta), and HHPRED (yellow, orange, and blue). (b) final model produced by I-TASSER using threading template 4R8G. (c) 1O1N and (d) 4LZXwere both used by PHYRE2 as threading templates. (e) 4LZX, (f) 2NCA, and (g) 3BOM were all used by HHPRED as threading templates.

3.3.2.5. Disorder Fragment 17: K900 – Q980

This fragment corresponds to the second part of the hem domain, namely helices A – C, in which helix C is short. The shorter fragment may present a more difficult fold recognition challenge than that is section 3.3.2.4, particularly as it contains the IQ (calmodulin-binding) domain.

Fig. 3.27b contain the structure 4R8G (Myosin-1c tail in complex with Calmodulin), which was used to provide I-TASSER with a threading model through this region which produced a Z score of 1.45 and 94% sequence coverage. The Z score generated exceeds the associated threshold of 1, suggesting I-TASSER considered the alignment between the template and the query sequence as a good alignment. The alignment that produced the model failed to align the first 5 residues of this sequence fragment to the threading template. However, the final model produced a C score of -2.03, indicating that while the confidence generated was not sufficient enough to make this a good model, as it did not exceed the assigned threshold of -1.5. Despite this, the C score was not poor enough that the model holds no biological significance in terms of structure. The model also produced a TM score of 0.43±0.14, which at the top of the range of this values generated means that I-TASSER considered the model to have a similar fold to the natively folded template.

The threading template is complexed with calmodulin, in the same region that was used to model this sequence fragment, which is reasonable in light of the fact that this sequence fragment contains a calmodulin binding domain. Despite this, the fact that

I-TASSER failed to identify this region as haemoglobin homologous protein, decreases reliability in this method.

Figs. 3.27c and 3.27d contain the structures 1O1N (Deoxy haemoglobin) and 4LZX (IQ-domain-containing G protein). These structures have been mapped onto this region of the protein by PHYRE2, producing confidence scores of 96.3% and 95.5% respectively. These two models aligned to 913I – 980Q, providing 82% sequence coverage and 904A – 924L, providing 24% sequence coverage respectively. Both these structures have mapped this region as mostly helical, with loop conformations functioning as the link between each helical structure. The green model has 4 helices, with the fourth being continuous with the IQ binding domain helix coloured fig. 3.27d. This region is known to have the heme domain helices A – C, and as such the modelled structures confirm the presence of helical structure expected in this region.

The IQ-domain present in Fig. 3.27d coloured in magenta, is supported by the threading template used in this HHPRED prediction by the same PDB in fig. 3.27e. This structure produced a probability of 97.8% between residue positions 904A – 932D in the sequence fragment, producing 34% identities through the alignment. Despite the low sequence coverage, both the probability generated, and the consensus formed between PHYRE2 and HHPRED for this model, not only increases the reliability in this alignment but also the likelihood that this model hold functional relevance to this region of the protein.

Fig. 3.27f contains the structure 5BIW (Calcium binding EF-hand domain), which was used as the threading template by HHPRED in the region, mapping onto the sequence fragment through residues 907V – 950L. The helical binding domain was predicted to be present with a probability of 96.75% and 9% identities. This model

produced significantly less identities than Fig. 3.27e. However, the probability suggests this model remains biologically significant with reference to Androglobin.

Fig. 3.27g contains the structure 3BOM (trout haemoglobin), which was used by HHPRED as a threading template, through the residue positions 936N – 977P in this sequence fragment. This fragment mapped 2 – 3 alpha helices with a probability of 92.6%.

Ultimately, fig. 3.27 functioned as control. In detecting homology present between the sequence fragments provided and other heme proteins increases reliability in the models generated by these fold recognition techniques. In addition, the presence of an IQ-binding domain in Fig. 3.27d and 3.27e and an EF hand calcium binding domain in Fig. 3.27f further bolsters the reliability of fold recognition techniques utilized. Once again, PHYRE2 and HHPRED have been shown to be the most reliable of the methods utilised, as they formed a consensus on not only the 3 globin helices present, but also the IQ binding domain represented by 4LZX. These results not only function as a control, they also managed to produce supporting evidence for the third EF hand previously mapped within the structure of the heme domain (see Chapter 2).

3.3.2.6. Disorder Fragment 18: I986 – ~L1180

This fragment is the same as in section 3.3.1.6. so, the results are not repeated here.

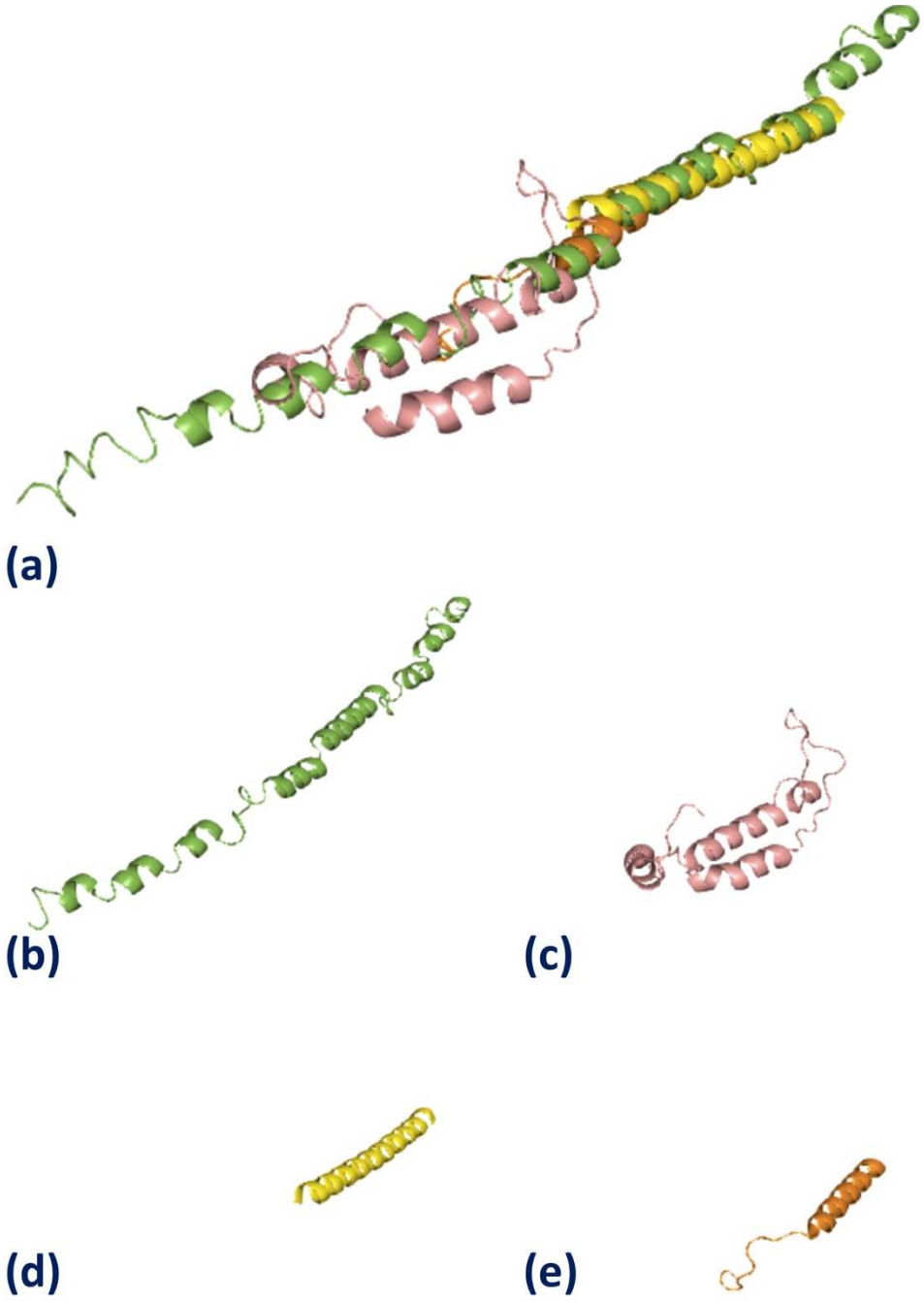

Fig. 3.28: Models produced in the region L1540 – E1650 of the Androglobin sequence. (a) superimposed structures produced in I-TASSER (green), PHYRE2 (salmon and magenta), and HHPRED (yellow and orange). (b) final model produced by I-TASSER using threading template 6U0U. (c) 4JKV was used by PHYRE2 as threading templates. (d) 3F1I and (e) 1DCE were both used by HHPRED as threading templates.

3.3.2.7. Disorder Fragment 18: L1540 – E1650

Fig. 3.28b contains the structure 6U0U (Protofilament Ribbon Flagellar Proteins Rib43a-L), which provided I-TASSER with a threading template, generating a Z score of 1.35. This alignment covered the residues through D1555 – T1649, where it provided 85% sequence coverage for this fragment. This Z score indicated that the alignment between the threading template and the query sequence was considered a good alignment, as it exceeded the threshold associated with Z score of 1. The final model generated a C score of -1.45 and a TM score of 0.54±0.15. Both these values indicate that there is a good amount of confidence associated with the model produced in I-TASSER and that this model has adopted a fold that similar to the threading template indicating that this model holds some biological significance, as both C score and TM score exceeded both their threshold of -1.5 and 0.5 respectively.

Fig. 3.28c contains the structure 4JKV (human smoothened membrane receptor), which provided the threading structure for PHYRE2. The template mapped between the residues 1546K – 1628L, equating to relatively high sequence coverage of 73%. It also produced a moderate confidence level of 65.4% and 24% identities. However, as 4JKV is a membrane spanning G protein coupled receptor, this association is somewhat random.

Fig. 3.28d contains the structure 3F1I (Hepatocyte growth factor-regulated tyrosine kinase substrate), which was also presented in figs. 3.23g and 3.23i, was used by HHPRED as a threading template for the helical structure in this sequence fragment through the residues 1599L – 1642E. This alignment produced a moderate-high probability of 74.9% and identities of 39% in the region modelled.

Fig. 3.28e contains the structure 1DCE (Rab geranylgeranyltransferase) was used to model this sequence fragment in the residues preceding Fig. 3.16c, through the region 1555T – 1585N. This model produced a confidence level of 23.3% and identities of 10%.

Fig. 3.28c and 3.28d display some degree of confidence in the structures predicted. In addition, both structures are modelled after folds found in cell membrane receptors, but androglobin is not predicted to have a membrane-spanning helix. The only consensus that may be had is the fact that this region seems to be largely helical. There exists a great deal of discrepancy between servers as to how long the and how many individual helices are present, but these results seem to refute the consensus formed by secondary structure prediction methods, of this region being one singular continuous helix.

3.4 Conclusion

3.4.1 Pre-heme MSA fragments

The calpain domains IIa and III have been shown, through profile-profile alignments, to be present in 2 of the 3 well-defined and ordered regions. These 3 ordered regions of the sequence preceding the heme domain, were characterised by disorder predictions in chapter 2. The results of both the disorder prediction and profile-profile alignments can be seen in Fig. 2.1. The profile-profile alignments mapped domains IIa and III through the residues E59 – G327 and W632 – E775 respectively. These two calpain domains mapped in chapter 2 correspond to 3.3.1.2 MSA fragment 2 and 3.3.1.5 MSA Fragment 5, which were produced using the sequence limits defined by residues D85 – L295 and T612 – N781 respectively. Fold recognition results from fragments 2 and 5 show the sequences as being homologous to calpain proteins, indicated by the use of calpain threading templates in each of the 4 independent servers. The clear consensus formed with a high degree of confidence produced by each server, in conjunction with the profile-profile alignments obtained in chapter 2, provides strong evidence in the absence of experimental data for these two sequence fragments being homologous to calpain domains. In terms of secondary structure, both 3.3.1.2 and 3.3.1.5 produced clearer consensus of predicted structure in the second half if both sequence fragments shown in Fig. 3.3b and 3.6b.

The third well-ordered region found in the pre-heme region of the Adgb sequence, as predicted by the consensus of disorder predictions shown in Fig. 2.1 and coincides with the sequence limit of 3.3.1.3. MSA fragment 3. This fragment resides

between the two calpain domains, in the other 2 ordered regions thought to be present in the pre-heme region of Adgb, through the residues K384 – S574. In contrast to fragments 2 and 5, this fragment produced clearer consensus of structures predicted in the first half of the fragment shown in Fig. 3.5a. However, the second half of this fragment produced far less structure. This was in addition to failing to produce a clear consensus of structure though to be present, with the second structure present in the PSIPRED alignment of Fig. 3.5b being the only structure present in more than 3 servers. Furthermore, fold recognition results for this fragment also failed to produce a consensus for this fragment despite the disorder prediction produced in the Fig. 2.1 characterising this sequence fragment as largely ordered. Therefore it is not possible to draw conclusion from the structure produced in this region, beyond the mere fact that the majority of the models produced are large amount of loop and helical residues. This is with the exception I-TASSER, which produced a pair of both parallel and antiparallel beta-sheets in the models shown in Fig. 3.5d.

The remainder of the pre-heme sequence is covered by 3.3.1.1 MSA Fragment 1 through residues M1 – F81 and 3.3.1.3 MSA Fragment 3 through the residues P296 – F383. The disorder prediction produced in chapter 2 have characterised these two regions in the sequence largely disordered. Due to disorder predicted in these regions, we expected that these sequence fragments would fail in producing a consensus in both the secondary structure prediction and the fold recognition. The MSA fragment 1 did indeed fail to produce a meaningful threading template, despite it producing two helices in the PSIPRED structural alignment in Fig. 3.2a. The structure predicted in PSIPRED was supported by the other, though there was no consensus gained for the length of the structures produced. The MSA fragment 3 also failed to produce a clear consensus for

structural alignments and fold recognition. The only structure present in Fig. 3.4a for which a consensus has been formed between servers, is the helix produced by the JPRED alignment. Though no consensus if formed from fold recognition, the conclusion that this fragment contains helix and loop is consistent with the results of the structural alignment gained in this region.

3.4.2 Post-heme MSA fragments

The region covered through the residues I986 – I1188, by 3.3.1.6 MSA fragment 6 was characterised by disorder prediction, as the largest post-heme continuous stretch of residues to be ordered in Fig. 2.3. As a result, it was expected that this fragment would produce the most meaningful alignment from secondary structure prediction and fold recognition results. This expectation was due to the lack of disorder expected in this fragment. This fragment produced an incredibly clear consensus, across all servers, on the structures expected to be present in both halves of the sequence fragment created in this region shown in Fig. 3.7a and 3.7b. The consensus was of this region being beta-sheet rich with a negligible number of helices predicated. The fold recognition results for this fragment failed to form a consensus of the protein present in this region of the sequence, however, the results seem to indicate that the beta-sheets tend to aggregate to form antiparallel strands. Most notably Fig. 3.7d shows a beta sandwich flanked to the left by a pair of parallel and antiparallel beta-sheets, though the variation in the structures observed in this region means it is not possible to definitely state that this is the structure through this region.

The last 100 residues in the C- terminus of Adgb have been shown by secondary structure prediction to contain an alpha helix spanning >30 residues, which starts in the second half of Fig. 3.11a and going on to cover the entire length of the Fig. 3.11b. The sequence in this region is covered by 3.3.1.10 MSA fragment 13 through the residues I1512 – A1644. The presence of the helix in this region of the sequence, and was predicted to be a coiled-coil by the Lupas methods (Lupas, Van Dyke and Stock, 1991). However, despite this consensus reached through secondary structure prediction, fold recognition methods were not able to produce a consensus when modelling this sequence fragment. This resulted in the use of dissimilar threading templates being produced, that though were largely had loop and helical residues, failed to support the consensus that was produce by secondary structure prediction of this region being a long continuous alpha helix preceded by residues with loop conformations. Barring RaptorX in Fig. 3.11d, none of the fold recognition techniques resembled the alignment gained by secondary structure prediction methods.

The interface between fragments 6 and 13 was predicted as largely, disordered by disorder prediction and such we did not expect fold recognition to produce any meaningful results in this region. Out of all the structure predicted in secondary structure results from 3.3.1.7. MSA fragments 7 and 8, through residues I1188 – I1223 and Q1224 – E1295 respectively, fragment 8 produced the clearest structures, with a consensus formed for the second beta-sheet and first helix in this fragment. The discrepancy in the length of predicted structures and their localisation in this fragment

has the least amount of variation between the two fragment. When considering 3.3.1.9. MSA fragments 11 and 12, through the residues A1402 – E1440 and K1452 – E1491 respectively, the clearest structures are produced in fragment 11. This is as the

consensus in fragment 12 fails to establish whether or not the second helix in the ITASSER prediction is continuous, or if it is divided in a similar fashion to the PSIPRED and Raptor Property prediction. Though fragments 7, 8, 11, and 12 fail to form a consensus of the protein present in their respective regions, they have produced models that have residues that are predominantly loop and helical in nature, largely lacing alpha helices. In addition, the lack of structures produced by HHPRED in fragment 12 and the missing coordinates in the structures produced by HHPRED for fragment 7 and 8, further supports the conclusion that was reached (see Chapter 2) of this region being intrinsically disordered; this would inherently make the sequence hard to crystallise in this region.

Furthermore, use of 5DBR as a threading template for fragment 8 by HHPRED is notable, as Adgb contains a CaM IQ binding motif in its heme domain, which may seek to interact with this portion of the protein sequence. If this interaction occurs, the CaM mapped through residues L1267 – L1281, is in close proximity to disorder and may be able to act as a MoRF (Vacic *et al.*, 2007) in this region shown in Fig. 2.3. This becomes more likely when considering the clear consensus of secondary structure that was predicted in fragment 8. This means that if CaM is truly present in this region and was able bind the IQ binding motif in the heme domain, that it may propagate an order-to-disorder transition both downstream of fragment 8 and upstream of the IQ binding motif in the heme domain.

Despite the consensus in this region characterising Adgb as disordered, the secondary structure prediction in 3.3.1.8 MSA fragments 9 and 10 produced a clear consensus of structure especially in fragment 10. The mixture of helices, sheet and loop residues in this region is evident from the structures produced by fold recognition, methods failing to form a consensus. This left us unable to draw a conclusion beyond

fragment 9 being predominantly helical, with loop residues. However, fragment 10 shows antiparallel beta-sheet to be present in addition to alpha helix.

3.4.3 Disorder fragments

Fragments were also formed to encompass the regions considered as largely ordered; these regions were deduced by a consensus of different disorder prediction methods. This was done in hopes of removing noise from the sequence fragments and increasing the quality of the models generated. The fold recognition methods used formed the same consensus in 3.3.2.1 disorder fragment 14 as was formed for 3.1.2 MSA fragment 2. This consensus characterised this region in the sequence, through the residues H80 – L300, as homologous to calpain.

With redefined sequence limits of L1540 – E1650, disorder fragment 18 failed to produce a consensus that would clarify the nature of the protein through this region. Despite all the structures in this region being predicted as predominantly helical, there still seems to be no consistency in regard to the length of the helices that are produced in this region. Therefore, much like fragment 13, it is not possible to conclude whether or not this region in the sequence contains a single continuous alpha helix, or a number of disjointed smaller alpha helices.

Redefining the sequence limits in 3.3.2.2 disorder fragment 15 to S390 – F520 has failed to produce a consensus, as was the face for MSA fragment 4. Though despite the lack consensus, the use of 6BSV as pictured in Fig. 3.13c as a threading template by PHYRE2 in this region is worth noting, as this alignment has produced a confidence level over double that seen in MSA fragment 4 shown in Fig. 3.5e. Despite the increase in the

confidence produced by this fragment alone, it is not enough to comment beyond the structure of this region being helical with a pair of antiparallel beta-sheet strands. Though the increase in confidence does support the use of disorder prediction, in an effort to reduce noise from disordered residue and produce a threading template with greater statistical significance.

As was the case for disorder prediction, 3.3.2.4. disorder fragment 16 and 3.3.2.5. disorder fragment 17 was formed from the heme domain sequence of Adgb and functioned as a control to assess the reliability of the methods used . Fragment 16 is known to contain 5 helices D – H in this region, however, the fold recognition consensus has produced 4 clear helices in each case. PHYRE2 and HHPRED in this region have produced an extra helix, after the second helix, which in both cases is a little over a single helical turn in length. The limited number of turns in this third helix contradicts the known structure, as such a small helix would be expected to be helix D in the N-terminus of this sequence fragment. Despite this, the consensus of fold recognition results in this sequence fragment has independently produced 4 globin threading templates.

Fragment 17 is expected to contain 3 heme helices and an additional IQ binding domain. The model produced in I-TASSER has no relation to globin and has produced a structure from myosin in complex with CaM, which may be explained by the presence of the IQ binding motif in this sequence fragment, as the myosin tail may contain the same IQ binding motif that promotes it to form a complex with CaM in the same way Adgb does. PHYRE2 and HHPRRED both produced 3 helices from globin proteins as expected. In both the PHYRE2 and the HHPRED prediction, the first helix produced in this alignment is in close proximity with IQ binding domain shown in Fig. 3.15b, 3.15c, and 3.15e is

something that is consistent with the already known structure of the heme domain of Adgb.

3.4.4 Heme domain EF hand

The presence of an EF binding domain in Fig. 3.27f further supports the results gained from the profile-profile alignments (see Chapter 2), produced between the Adgb MSA and the EF hand seeded alignments. To further verify the presence of this EF hand in light of the HHPRED threading template used in Fig. 3.27f, profile-profile alignment was produced using the human Adgb sequence through the residues N781 – D990. The choice to use only the human sequence will ensure an alignment achieved between the EF hand seeded alignment and the Adgb sequence, will be due to conserved EF hand contact residues in human Adgb alone as opposed to conservation of contact residues across homologues.

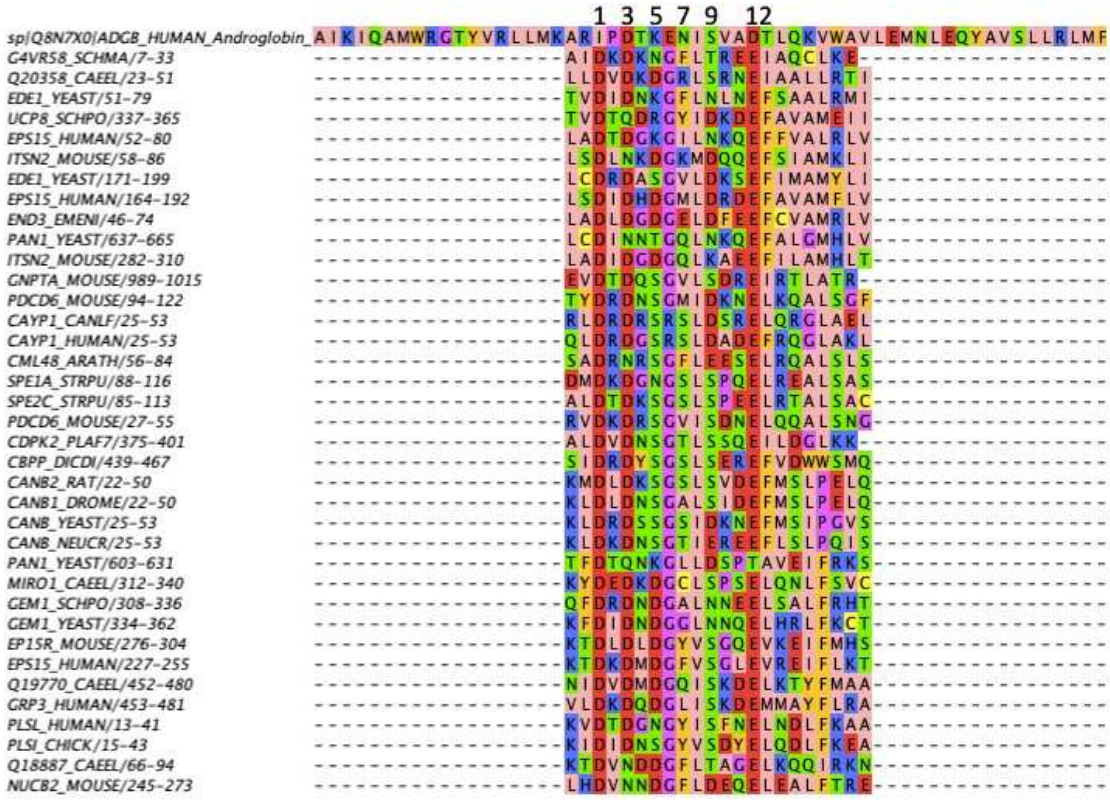

Fig. 3.29: The third Profile alignment between EF hand and Androglobin heme sequence. This alignment was gained by using the seed file EF hand 1 - PF00036 from the Pfam database, and has aligned on to the Adgb heme sequence at residues A928 – V949 on the Adgb sequence

This alignment has mapped the EF hand from PF00036 to the heme domain in the same region of the Adgb sequence shown in Fig. 2.7. The EF hand mapped by HHPRED in Fig. 3.27f lies between residues 907V – 950L in the Adgb sequence. It stands to reason that the alignment produced in HHPRED would result in the contact residues IP**D**TKEN**I**SVA**D**T shown in Fig. 3.29.

Given more time, it would have been interesting to evaluate results from RaptorX models. This server was unusually slow in generating the structures pertaining to the disorder fragments, and as a result, it failed to produce some of the structures. This was due to the fact that when the sequences were submitted, the structure prediction methods were being updated to version 2. It would be interesting to see the difference in the structures being produced, in light of the new prediction method. Furthermore, subjecting these structures to molecular dynamic simulations would help explore the conformational space in which some of these structures exist. It is important to note, that while these proteins are static when crystallised, they remain flexible and mutable when expressed *in vivo.*

Statement of significance

The fusion of portions of protein homologous to calpain with the O_2 sensing heme domain, may essentially confer the ability to mediate cell apoptosis to this specific globin. This has been evidenced by the down regulation of Signal transducers and activators of transcription 3 (STAT3), seen when Adgb is knocked down in glioma cell lines, in addition, to the ability for m-calpain regulate apoptosis in gestation. This means Adgb may prove to be a potentially druggable target in some cancers some specific cancers. Globins as a whole have been linked to breast cancer, hepatocellular carcinoma, and cervical cancer, as a means of imbuing the ability to resist different forms of therapies including chemo- and radiotherapy. This functions as strong evidence that an aberrant form of Adgb has the potential to function as an oncogene leading to the propagation of cancers. (Huang *et al.*, 2014; Zuo *et al.*, 2018).

The expression of Adgb in higher level that is found in fertile male spermatozoa when compared to infertile male spermatozoa suggests as integral role in spermatogenesis. This Is intuitive as both m- and mu-caplain have been implicated in cytoskeletal remodelling which remains integral to meiosis and spermatogenesis. As such the Adgb may potentially be a druggable target for therapies in the treatment for infertility in men (Hoogewijs *et al.*, 2012).

Bibliography

Abyzov, A. et al. (2016) ‘Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation’, Journal of the American Chemical Society. American Chemical Society, 138(19), pp. 6240–6251. doi: 10.1021/jacs.6b02424.

Agarwal, N. (2013) 'In-Silico Drug Design: A revolutionary approach to change the concept of current Drug Discovery Process', INDIAN JOURNAL OF PHARMACEUTICAL & BIOLOGICAL RESEARCH (IJPBR). Available at: www.ijpbr.in (Accessed: 26 December 2018).

Appleby, C. A., Tjepkema, J. D. and Trinick, M. J. (1983) ‘Hemoglobin in a Nonleguminous Plant, Parasponia: Possible Genetic Origin and Function in Nitrogen Fixation’, Science, 220(4600), pp. 951–953. doi: 10.1126/science.220.4600.951.

Aronson, H.E., Royer, W.E.Jr., and Hendrickson, W.A. 1994. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci. 3: 1706– 1711.

Ascenzi, P., Gustincich, S. and Marino, M. (2014) ‘Vinogradov SN, Moens L. Diversity of globin function: enzymatic, transport, storage, and sensing’, Acta Physiol, 66, pp. 501–514. doi: 10.1002/iub.1267.

Babu, M. M. (2016) ‘The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease’, Biochemical Society Transactions, 44(5), pp. 1185–1200. doi: 10.1042/BST20160172.

Badugu, R. K. et al. (2008) ‘N terminus of calpain 1 is a mitochondrial targeting sequence’, Journal of Biological Chemistry, 283(6), pp. 3409–3417. doi: 10.1074/jbc.M706851200.

Becana, M. and Moran, J. F. (1994) 'Structure and function of leghemoglobins', An. Estac. Exp. Aula Dei, 21(3), pp. 203–208.

Bracke, A., Hoogewijs, D. and Dewilde, S. (2018) 'Exploring three different expression systems for recombinant expression of globins: Escherichia coli, Pichia pastoris and Spodoptera frugiperda'. Analytical Biochemistry 2018, 543, pp. 62-70. doi: 10.1016/j.ab.2017.11.027.

Buchan, D. W. A. and Jones, D. T. (2019) 'The PSIPRED Protein Analysis Workbench: 20 years on', Nucleic Acids Research. doi: 10.1093/nar/gkz297.

Burmester, T. et al. (2002) 'Cytoglobin: A Novel Globin Type Ubiquitously Expressed inVertebrate Tissues', Molecular Biology and Evolution, 19(4), pp. 416–421. doi: 10.1093/oxfordjournals.molbev.a004096.

Burmester, T. and Hankeln, T. (2008) 'Neuroglobin and Other Nerve Haemoglobins', Dioxygen Binding and Sensing Proteins. Springer Milan, pp. 211–222. doi: 10.1007/978-88-470-0807-6_18.

Burmester, T. and Hankeln, T. (2009) 'What is the function of neuroglobin?', Journal of Experimental Biology, 212(10), pp. 1423 LP – 1428. doi: 10.1242/jeb.000729.

Chagot, B. et al. (2009) 'Solution NMR Structure of the C-terminal EF-hand Domain of Human Cardiac Sodium Channel Na V 1.5 * □ S Downloaded from', THE JOURNAL OF BIOLOGICAL CHEMISTRY, 284(10), pp. 6436–6445. doi: 10.1074/jbc.M807747200.

Cheng, J. et al. (2008) 'Machine Learning Methods for Protein Structure Prediction', IEEE REVIEWS IN BIOMEDICAL ENGINEERING, 1. doi: 10.1109/RBME.2008.2008239.

Chou, J. (2010) 'CALPAIN 2 ACTIVATION, AUTOLYSIS, AND SUBUNIT DISSOCIATION', Master of Science Thesis. Queen's University, Kingston, Ontario, Canada

Davletov, B. A. and Sudhof, T. C. (1993) 'A single C2domain from synaptotagmin I is sufficient for high affinity Ca2+/phospholipid binding', Journal of Biological Chemistry, 268(35), pp. 26386–26390.

Deiana, A. et al. (2019) 'Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell', PLoS ONE, 14(8). doi: 10.1371/journal.pone.0217889.

Domino, K. et al. (1983) 'Influence of mixed venous oxygen tension ($P\bar{V}(O2)$) on blood flow to atelectatic lung', Anesthesiology. Anesthesiology, 59(5), pp. 428–434. doi: 10.1097/00000542-198311000-00012.

Dordas, C. (2009) 'Nonsymbiotic hemoglobins and stress tolerance in plants', Plant Science, 176(4), pp. 433–440. doi: 10.1016/j.plantsci.2009.01.003.

Drozdetskiy, A. et al. (2015) 'JPred4: a protein secondary structure prediction server', Nucleic Acids Research, 43(1), pp. 389–394. doi: 10.1093/nar/gkv332.

Dunker, A. K. et al. (2002) 'Intrinsic disorder and protein fluctuations', Biochemistry. American Chemical Society, 41(21), pp. 6573–6582. doi: 10.1021/bi012159.

Dutt, P. et al. (2006) 'm-Calpain is required for preimplantation embryonic development in mice', BMC Developmental Biology. BioMed Central, 6(1), p. 3. doi: 10.1186/1471-213X-6-3.

Endeward, V., Gros, G. and Jürgens, K. D. (2010) 'Significance of myoglobin as an oxygen store and oxygen transporter in the intermittently perfused human heart: A model study', Cardiovascular Research, 87(1), pp. 22–29. doi: 10.1093/cvr/cvq036.

Fago, A. et al. (2004) 'Critical Review Functional Properties of Neuroglobin and Cytoglobin. Insights into the Ancestral Physiological Roles of Globins'. doi: 10.1080/15216540500037299.

Farah, C. A. and Sossin, W. S. (2012) 'The role of C2 domains in PKC signaling', Advances in

Experimental Medicine and Biology. Springer, Dordrecht, 740, pp. 663–683. doi: 10.1007/978-94-007-2888-2_29.

Gandhi, C. R. (2012) 'Oxidative Stress and Hepatic Stellate Cells: A PARADOXICAL RELATIONSHIP.', Trends in cell & molecular biology. NIH Public Access, 7, pp. 1–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27721591 (Accessed: 19 April 2020).

Garcia, M., Bondada, V. and Geddes, J. W. (2005) 'Mitochondrial localization of μ-calpain', Biochemical and Biophysical Research Communications, 338(2), pp. 1241–1247. doi: 10.1016/j.bbrc.2005.10.081.

Gardner, P. R. (2005) 'Nitric oxide dioxygenase function and mechanism of flavohemoglobin, hemoglobin, myoglobin and their associated reductases', Journal of Inorganic Biochemistry. Elsevier Inc., 99(1), pp. 247–266. doi: 10.1016/j.jinorgbio.2004.10.003.

Grimholt, R. M. et al. (2018) 'Hb Oslo [β42(CD1)Phe→Ile; HBB: c.127T>A]: A Novel Unstable Hemoglobin Variant Found in a Norwegian Patient', Hemoglobin, 42(2), pp. 78–83. doi: 10.1080/03630269.2018.1468773.

Gsponer, J. and Madan Babu, M. (2009) 'The rules of disorder or why disorder rules', Progress in Biophysics and Molecular Biology, 99(2–3), pp. 94–103. doi: 10.1016/j.pbiomolbio.2009.03.001.

Guidolin, D. et al. (2016) ‘Neuroglobin, a Factor Playing for Nerve Cell Survival’, International journal of molecular sciences, 17(11). doi: 10.3390/ijms17111817.

Hanson, J. et al. (2019) 'Enhancing protein intrinsic disorder prediction by utilizing deep squeeze and excitation residual inception and long short-term memory networks'. Genomics, Proteomics & Bioinformatics, 17: 645-656 (2019).

Hargrove, M. S. et al. (1994) ‘Stability of Myoglobin: A Model for the Folding of Heme Proteins’, Biochemistry. Biochemistry, 33(39), pp. 11767–11775. doi: 10.1021/bi00205a012.

Helms, C. and Kim-Shapiro, D. B. (2013) ‘Hemoglobin-mediated nitric oxide signaling’. doi: 10.1016/j.freeradbiomed.2013.04.028.

Hensen, U. et al. (2012) ‘Exploring protein dynamics space: the dynasome as the missing link between protein structure and function.’, PloS one. Public Library of Science, 7(5), p. e33931. doi: 10.1371/journal.pone.0033931.

Hoogewijs, D. et al. (2012) ‘Androglobin: A Chimeric Globin in Metazoans That Is Preferentially Expressed in Mammalian Testes’, Molecular Biology and Evolution, 29(4), pp. 1105–1114. doi: 10.1093/molbev/msr2

46.

Huang, B. et al. (2014) ‘Androglobin knockdown inhibits growth of glioma cell lines.’, International journal of clinical and experimental pathology. e-Century Publishing Corporation, 7(5), pp. 2179–84.

Huang, J., Zhu, Xiaoping and Zhu, X (2016) ‘The Molecular Mechanisms of Calpains Action on Skeletal Muscle Atrophy’, Physiol. Res, 65, pp. 547–560.

Huang, Y. and Liu, Z. (2013) 'Do intrinsically disordered proteins possess high specificity in protein-protein interactions?', Chemistry - A European Journal. Chemistry, 19(14), pp. 4462–4467. doi: 10.1002/chem.201203100.

Ihara, E. and MacDonald, J. A. (2007) 'The regulation of smooth muscle contractility by zipper-interacting protein kinase', Canadian Journal of Physiology and Pharmacology. Can J Physiol Pharmacol, pp. 79–87. doi: 10.1139/Y06-103.

Ishida, T. and Kinoshita, K. (2007) 'PrDOS: prediction of disordered protein regions from amino acid sequence', Nucleic Acids Research, 35. doi: 10.1093/nar/gkm363.

Jones, D. T. and Cozzetto, D. (2017) 'DISOPRED3: Precise Disordered Region Predictions With Annotated Protein-Binding Activity', Oxford. Bioinformatics, 31(6), pp. 1401–1403. doi: 10.1093/BIOINFORMATICS.

Källberg, M. et al. (2014) 'RaptorX server: A Resource for Template-Based Protein Structure Modeling', Methods in molecular biology, Clifton, N.J. , pp. 17–27. doi: 10.1007/978-1-4939-0366-5_2.

Kelley, L. A. et al. (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', Nature Protocols. Nature Publishing Group, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.

Kolobynina, K. G. et al. (2016) 'Emerging roles of the single EF-hand Ca2+ sensor tescalcin in the regulation of gene expression, cell growth and differentiation', Journal of Cell Science. Company of Biologists Ltd, pp. 3533–3540. doi: 10.1242/jcs.191486.

Kolobynina, K. G. et al. (2017) 'Correction: Emerging roles of the single EF-hand Ca 2+ sensor tescalcin in the regulation of gene expression, cell growth and differentiation'. doi: 10.1242/jcs.200378.

Koonin, E. V., Wolf, Y. I. and Karev, G. P. (2002) 'The structure of the protein universe and genome evolution', Nature, 420(6912), pp. 218–223. doi: 10.1038/nature01256.

Kosol, S. et al. (2013) 'molecules Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy', Molecules, 18, pp. 10802–10828. doi: 10.3390/molecules180910802.

Kretsinger, R. H. (1997) 'EF-hands embrace', Nature Structural Biology, pp. 514–516. doi: 10.1038/nsb0797-514.

Kuhn, V. et al. (2017) 'Red Blood Cell Function and Dysfunction: Redox Regulation, Nitric Oxide Metabolism, Anemia', Antioxidants and Redox Signaling. Mary Ann Liebert Inc., pp. 718–742. doi: 10.1089/ars.2016.6954.

Kundu, S., Trent, J. T. and Hargrove, M. S. (2003) 'Plants, humans and hemoglobins.', Trends in plant science, 8(8), pp. 387–93. doi: 10.1016/S1360-1385(03)00163-8.

Lafitte, D. et al. (1999) Evidence of noncovalent dimerization of calmodulin, European Journal of Biochemistry. doi: 10.1046/j.1432-1327.1999.00284.x.

Li, W. et al. (2009) 'EF hands at the N-lobe of calmodulin are required for both SK channel gating and stable SK-calmodulin interaction', Journal of General Physiology, 134(4), pp. 281–293. doi: 10.1085/jgp.200910295.

Linding, R. et al. (2003) 'GlobPlot: Exploring protein sequences for globularity and disorder', Nucleic Acids Research, 31(13), pp. 3701–3708. doi: 10.1093/nar/gkg519.

Linding, R. et al. (2003) 'Protein disorder prediction: Implications for structural proteomics', Structure. Cell Press, 11(11), pp. 1453–1459. doi: 10.1016/j.str.2003.10.002.

Littlechild, J. A. (2013) 'Protein structure and function', Introduction to Biological and Small Molecule Drug Research and Development: Theory and Case Studies. W H Freeman, pp. 57–79. doi: 10.1016/B978-0-12-397176-0.00002-9.

Lock, A. et al. (2014) 'One motif to bind them: A small-XXX-small motif affects transmembrane domain 1 oligomerization, function, localization, and cross-talk between two yeast GPCRs', Biochimica et Biophysica Acta - Biomembranes. Elsevier B.V., 1838(12), pp. 3036–3051. doi: 10.1016/j.bbamem.2014.08.019.

Maki, M. et al. (2012) 'Evolutionary and physical linkage between calpains and penta-EF-hand Ca2+-binding proteins', FEBS Journal, 279(8), pp. 1414–1421. doi: 10.1111/j.1742-4658.2012.08560.x.

Mcguffin, L. J. (2008) 'Intrinsic disorder prediction from the analysis of multiple protein fold recognition models', Bioinformatics, 24(16), pp. 1798–1804. doi: 10.1093/bioinformatics/btn326.

Mészáros, B., Erdös, G. and Dosztányi, Z. (2018) 'IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding', Nucleic Acids Research. doi: 10.1093/nar/gky384.

Mizianty, M. J. et al. (2011) 'Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources', Bioinformatics, 27(13), pp. i489–i496. doi: 10.1093/bioinformatics/btq373.

Mizianty, M. J., Peng, Z. and Kurgan, L. (2013) 'MFDp2', Intrinsically Disordered Proteins. Informa UK Limited, 1(1), p. e24428. doi: 10.4161/idp.24428.

Moreau, S. et al. (1996) 'Leghemoglobin-derived radicals. Evidence for multiple protein-derived radicals and the initiation of peribacteroid membrane damage.', The

Journal of biological chemistry. American Society for Biochemistry and Molecular Biology, 271(51), pp. 32557–62.doi: 10.1074/JBC.271.51.32557.

Motoyama, H. et al. (2014) 'Cytoglobin is expressed in hepatic stellate cells, but not in myofibroblasts, in normal and fibrotic human liver', Laboratory Investigation. Nature Publishing Group, 94(2), pp. 192–207. doi: 10.1038/labinvest.2013.135.

Nord, A. (2017) 'Mirage: A Novel Multiple Protein Sequence Alignment Tool', Graduate Student Theses, Dissertations, & Professional Papers. Available at: https://scholarworks.umt.edu/etd/11110 (Accessed: 26 December 2018).

Nye, D. B. and Lecomte, J. T. J. (2018) 'Replacement of the Distal Histidine Reveals a Noncanonical Heme Binding Site in a 2-on-2 Hemoglobin', Biochemistry. American Chemical Society, 57(40), pp. 5785–5796. doi: 10.1021/acs.biochem.8b00752.

Ono, Y. and Sorimachi, H. (2012) 'Calpains-An elaborate proteolytic system ☆', BBA - Proteins and Proteomics, 1824, pp. 224–236. doi: 10.1016/j.bbapap.2011.08.005.

Ordway, G. A. and Garry, D. J. (2004) 'Myoglobin: An essential hemoprotein in striated muscle', Journal of Experimental Biology. The Company of Biologists Ltd, pp. 3441–3446. doi: 10.1242/jeb.01172.

Ouellet, H. et al. (2002) 'Truncated hemoglobin HbN protects Mycobacterium bovis from nitric oxide', Proceedings of the National Academy of Sciences, 99(9), pp. 5902–5907. doi: 10.1073/pnas.092017799.

Parent, C. et al. (2007) 'A novel nonsymbiotic hemoglobin from oak: cellular and tissue specificity of gene expression', New Phytologist. John Wiley & Sons, Ltd, 0(0), pp. 071106233614004-??? doi: 10.1111/j.1469-8137.2007.02250.x.

Peng, J. and Xu, J. (2011) 'RaptorX: exploiting structure information for protein alignment by statistical inference.', Proteins. NIH Public Access, 79 Suppl 1(Suppl 10), pp. 161–71. doi: 10.1002/prot.23175.

Peng, Q. Y. and Zhang, Q. F. (2006) 'Precise positions of Phoebe determined with CCD image-overlapping calibration', Monthly Notices of the Royal Astronomical Society. BioMed Central, 366(1), pp. 208–212. doi: 10.1186/1471-2105-7-208.

Pesce, A. et al. (2002) 'Neuroglobin and Cytoglobin. Fresh Blood for the Vertebrate Globin Family', EMBO reports. EMBO Rep, 3(12). doi: 10.1093/EMBO-REPORTS/KVF248.

Pirovano, W. and Heringa, J. (2010) 'Protein secondary structure prediction.', in Methods in molecular biology (Clifton, N.J.). Humana Press, pp. 327–348. doi: 10.1007/978-1-60327-241-4_19.

Reeder, B. J. (2010) 'The redox activity of hemoglobins: From physiologic functions to pathologic mechanisms.' Antioxidants and Redox Signaling, 13 (7). 1087 - 1123. ISSN 1523-0864

Reyes, J. G. et al. (2012) 'The Hypoxic Testicle: Physiology and Pathophysiology', Oxidative Medicine and Cellular Longevity. Hindawi Publishing Corporation, 2012, p. 15. doi: 10.1155/2012/929285.

Roesner, A. et al. (2005) 'A globin gene of ancient evolutionary origin in lower vertebrates: Evidence for two distinct globin families in animals', Molecular Biology and Evolution. doi: 10.1093/molbev/msh258.

Rost, B. (1999) 'Twilight zone of protein sequence alignments', Protein Engineering, Design and Selection. Oxford Academic, 12(2), pp. 85–94. doi: 10.1093/protein/12.2.85.

Rumi, E. et al. (2009) 'Blood p50 evaluation enhances diagnostic definition of isolated erythrocytosis', Journal of Internal Medicine. John Wiley & Sons, Ltd, 265(2), pp. 266–274. doi: 10.1111/j.1365-2796.2008.02014.x.

Saini, H. et al. (2016) 'Protein Fold Recognition Using Genetic Algorithm Optimized Voting Scheme and Profile Bigram'. doi: 10.17706/jsw.11.8.756-767.

Schmidt, M. et al. (2003) 'How does the eye breathe? Evidence for neuroglobin-mediated oxygen supply in the mammalian retina', Journal of Biological Chemistry. American Society for Biochemistry and Molecular Biology, 278(3), pp. 1932–1935. doi: 10.1074/jbc.M209909200.

Schmidt, W. N., Mathahs, M. M. and Zhu, Z. (2012) 'Heme and HO-1 inhibition of HCV, HBV, and HIV', Frontiers in Pharmacology, 3 OCT. doi: 10.3389/fphar.2012.00129.

Shao, H. et al. (2006) 'Spatial Localization of m-Calpain to the Plasma Membrane by Phosphoinositide Biphosphate Binding during Epidermal Growth Factor Receptor-Mediated Activation', Molecular and Cellular Biology. American Society for Microbiology, 26(14), pp. 5481–5496. doi: 10.1128/mcb.02243-05.

Sievers, F. et al. (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', Molecular Systems Biology. European Molecular Biology Organization, 7, p. 539. doi: 10.1038/msb.2011.75.

Singh, M. (2001) Predicting Protein Secondary and Supersecondary Structure. Available at: https://www.cs.princeton.edu/~mona/Chapter29.pdf (Accessed: 28 December 2018).

Soding, J. (2005) 'Protein homology detection by HMM-HMM comparison', Bioinformatics. Oxford University Press, 21(7), pp. 951–960. doi: 10.1093/bioinformatics/bti125.

Söding, J., Biegert, A. and Lupas, A. N. (2005) 'The HHpred interactive server for protein homology detection and structure prediction', Nucleic Acids Research, 33(2). doi: 10.1093/nar/gki408.

Sonati, M. F. et al. (2006) 'Hemoglobin Hammersmith [β42 (CD1) Phe → Ser] in a Brazilian girl with congenital Heinz body hemolytic anemia', Pediatric Blood & Cancer. John Wiley & Sons, Ltd, 47(6), pp. 855–856. doi: 10.1002/pbc.20851.

Strobl, S. et al. (2000) 'The crystal structure of calcium-free human m-calpain suggests an electrostatic switch mechanism for activation by calcium', Proceedings of the National Academy of Sciences of the United States of America, 97(2), pp. 588–592. doi: 10.1073/pnas.97.2.588.

Stryer, L. (1995) In: Biochemistry 4th edition. W.H. Freeman.

Sutton R.B., et al. (1995) 'Structure of the first C2 domain of synaptotagmin I: a novel Ca2+/phospholipid-binding fold', Cell 1995, 80 (6), pp. 929-938

Taddese, B. et al. (2014) 'Do Plants Contain G Protein-Coupled Receptors?', Plant Physiology Ò, 164, pp. 287–307. doi: 10.1104/pp.113.228874.

Tidow, H. and Nissen, P. (2013) 'Structural diversity of calmodulin binding to its target sites', FEBS Journal, 280(21), pp. 5551–5565. doi: 10.1111/febs.12296.

Tompa, P. et al. (2001) 'Domain III of calpain is a ca2+-regulated phospholipid-binding domain.', Biochemical and biophysical research communications, 280(5), pp. 1333–9. doi: 10.1006/bbrc.2001.4279.

Traverso, M. (2004) Conformational changes upon Binding of Oxygen, Washington University in St. Louis. Available at: http://www.chemistry.wustl.edu/~edudev/LabTutorials/CourseTutorials/Tutorials/Hemoglobin/conformation.htm (Accessed: 10 April 2020).

Uversky, V. N. (2019) 'Intrinsically disordered proteins and their "Mysterious" (meta)physics', Frontiers in Physics. Frontiers Media S.A., p. 10. doi: 10.3389/fphy.2019.00010.

Vacic, V. et al. (2007) 'Characterization of molecular recognition features, MoRFs, and their binding partners', Journal of Proteome Research. J Proteome Res, 6(6), pp. 2351–2366. doi: 10.1021/pr0701411.

Vigeolas, H., Hühn, D. H. and Geigenberger, P. (2011) 'Nonsymbiotic hemoglobin-2 leads to an elevated energy state and to a combined increase in polyunsaturated fatty acids and total oil content when overexpressed in developing seeds of transgenic arabidopsis plants', Plant Physiology, 155(3), pp. 1435–1444. doi: 10.1104/pp.110.166462.

Wang, S. et al. (2016) 'RaptorX-Property: a web server for protein structure property prediction', Nucleic acids research, 44(W1), pp. W430–W435. doi: 10.1093/nar/gkw306.

Wang, Z.-X. (1996) 'How many fold types of protein are there in nature?', Proteins: Structure, Function, and Genetics, 26(2), pp. 186–191. doi: 10.1002/(SICI)1097-0134(199610)26:2<186::AID-PROT8>3.0.CO;2-E.

Waterhouse, A. M. et al. (2009) 'Jalview Version 2--a multiple sequence alignment editor and analysis workbench', Bioinformatics, 25(9), pp. 1189–1191. doi: 10.1093/bioinformatics/btp033.

Wittenberg, J. B. et al. (2002) 'Truncated hemoglobins: a new family of hemoglobins widely distributed in bacteria, unicellular eukaryotes, and plants.', The Journal of biological chemistry. American Society for Biochemistry and Molecular Biology, 277(2), pp. 871–4. doi: 10.1074/jbc.R100058200.

Xue, B. et al. (2010) 'PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids', Biochim Biophys Acta, 1804(4), pp. 996–1010. doi: 10.1016/j.bbapap.2010.01.011.

Yang, J. and Zhang, Y. (2015) 'I-TASSER server: new development for protein structure and function predictions', Web Server issue Published online, 43. doi: 10.1093/nar/gkv342.

Yoshizato, K. et al. (2016) 'Review Discovery of cytoglobin and its roles in physiology and pathology of hepatic stellate cells'. doi: 10.2183/pjab.92.77.

Zhou, J. and Troyanskaya, O. G. (2014) Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction. Available at: http://proceedings.mlr.press/v32/zhou14.pdf (Accessed: 26 December 2018).

Zhou, Y., Frey, T. K. and Yang, J. J. (2009) 'Viral calciomics: Interplays between Ca2+ and virus', Cell Calcium. Elsevier Ltd, pp. 1–17. doi: 10.1016/j.ceca.2009.05.005.

Zhou, Y., Xue, S. and Yang, J. J. (2013) 'Calciomics: Integrative studies of Ca2+-binding proteins and their interactomes in biological systems', Metallomics, 5(1), pp. 29–42. doi: 10.1039/c2mt20009k.

Zimmermann, L. et al. (2018) 'A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core', Journal of Molecular Biology. Academic Press, 430(15), pp. 2237–2243. doi: 10.1016/J.JMB.2017.12.007.

Zuo, S. et al. (2018) 'CRTH2 promotes endoplasmic reticulum stress-induced cardiomyocyte apoptosis through m-calpain', EMBO Molecular Medicine. EMBO, 10(3). doi: 10.15252/emmm.201708237.