CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Agrobiology, Food and Natural Resources

Department of Chemistry

Structural and interaction study of MEG family proteins and their role in liver fibrosis onset

.....

Dissertation thesis

Author: **Ing. Štěpánka Nedvědová**

Supervisor: **Ing. Matyáš Orsák, Ph.D.**

Consultants: **Dr. Maggy Hologne, Ph.D., HDR**

Prof. Adriana Erica Miele, Ph.D.

Prague 2023

N° National de Thèse

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sien de

L'Université Claude Bernard Lyon 1

École Doctorale 206

École Doctorale de Chimie

Spécialité de doctorat : Biochimie des protéines

Discipline : Chimie

Présentée et soutenue publiquement le 22/09/2023

par Štěpánka Nedvědová

Structural and interaction study of MEG family proteins and their role in liver fibrosis onset
Étude de la structure et de l'interaction des protéines de la famille MEG et de leur rôle dans l'apparition de la fibrose hépatique

Devant le jury composé de :

Pavel TLUSTOŠ	Professeur CZU, Prague, République tchèque	Président du jury
Christina SIZUN	Chargée de Recherche CNRS ICSN-CNRS, Gif-sur-Yvette, France	Rapporteur
Sonia LONGHI	Directeur de Recherche CNRS AFBM, Marseille, France	Rapporteur
Frederic VELLIEUX	Senior researcher BIOCEV, Vestec, République tchèque	Examinateur
Maggy HOLOGNE	Maître des Conférences UCBL1, Lyon, France	Directeur de thèse
Saida MEBAREK	Professeur UCBL1, Lyon, France	Examinateur
Jaroslav HAVLÍK	Senior researcher CZU, Prague, République tchèque	Examinateur
Petr MARŠÍK	Senior researcher CZU, Prague, République tchèque	Examinateur

Declaration

I hereby declare that I have written my dissertation "Structural and interaction study of MEG family proteins and their role in liver fibrosis onset" independently and that I have properly cited and referenced all sources and literature used. I further declare that no third-party copyrights have been infringed in connection with the writing of my doctoral dissertation.

In Prague, date ..... ……..... Author's signature

REMERCIEMENTS & PODĚKOVÁNÍ

Tout d'abord, je me dois de remercier mes deux superviseurs de l'UCBL1, sans lesquels ce travail n'aurait pas pu être réalisé. Vous êtes toutes deux des femmes uniques et fortes dans le domaine des sciences, que j'admirerai et respecterai toujours énormément. Merci d'être des modèles pour les étudiants et les jeunes scientifiques !
Merci à Adriana de m'avoir soutenue et aidée à mettre en place l'ensemble du programme de Barrande Fellowship. Merci pour tous les projets et articles sur lesquels nous avons travaillé ensemble, pour les conseils professionnels et personnels, pour m'avoir aidée à rédiger ma thèse, pour le soutien et le temps que tu m'as accordé. Et surtout, merci de m'avoir montré comment communiquer et comment faire face à des situations difficiles avec grâce.
Maggy, merci beaucoup de m'avoir acceptée comme étudiante à la dernière minute et de m'avoir sauvée. Merci pour tous les conseils en biophysique et en RMN, pour l'aide considérable apportée aux analyses et à leur interprétation, pour tout ton temps, pour toute l'aide administrative, pour tous les articles et pour l'aide apportée à la préparation de la thèse. Merci pour ton approche amicale et chaleureuse à mon égard et pour m'avoir permise de me sentir moins étrangère et seule en France !
Merci à Olivier Walker pour m'avoir accueillie dans son laboratoire, pour son approche, pour m'avoir aidée avec le calcul de la structure et la dynamique moléculaire et pour toutes nos discussions. Merci à tout le laboratoire BIOSYS de l'ISA pour l'environnement de travail merveilleux. Merci à Cyrielle pour tous les conseils, les avis/idées, mais surtout pour tous les déjeuners et les discussions que nous avons eus ensemble.
Et enfin, un grand merci à Florence, qui est et une personne merveilleuse que je garderai toujours dans mon cœur. Merci pour ton aide, qui a été considérable, pour toutes tes histoires, tes conseils, ton soutien et pour ta gentillesse, qui est si rare dans le milieu académique.
Merci également à Fred pour m'avoir convaincue de suivre la voie d'une collaboration avec la France dans le cadre de la cotutelle et pour m'avoir mise en contact avec Adriana.

Děkuji minulému i současnému vedení FAPPZ za jeho podporu a za to, že mi umožnilo nestandardní studium v rámci Barrande Fellowshipu. Jmenovitě děkuji profesorce Ivě Langrové za její důvěru, podporu a pomoc. Děkuji rovněž proděkanovi Haklovi za jeho vysoce profesionální a lidský přístup a jeho pomoc. Děkuji paní doktorce Makovcové za všechnu (nejen administrativní) pomoc, za její čas a obětavost. Děkuji doktoru Dvořákovi a jeho týmu za návrh tématu, na kterém jsem pracovala a za jejich pomoc s klonováním. Můj největší dík patří mému staronovému vedoucímu práce doktoru Matyášovi Orsákovi za to, jak mě celé studium podporoval a zachraňoval. Je mi ctí, že nakonec mohu obhájit pod záštitou katedry chemie, která je jedinečně skvělým pracovním prostředím, za což děkuji jejím členům a opět jejímu vedoucímu, který ji svým laskavým a nevídaným způsobem řídí a udržuje v chodu.
Závěrečný dík patří mým třem nejbližším, bez jejichž nepřetržité podpory a pochopení bych nemohla tuto práci dokončit

ABSTRACT

MEG proteins are unique to parasitic trematodes, particularly *Schistosoma mansoni*. These proteins are encoded by micro-exon genes, which have a unique structure consisting of short micro-exons alternated by long introns. Through genome sequencing of *S. mansoni*, 35 genes encoding 87 verified MEG proteins were identified. These genes show no similarity to other annotated genes except those found in *Schistosoma spp*. Furthermore, they lack identifiable motifs or functional domains. MEGs undergo alternative splicing, resulting in the generation of multiple isoforms from a single mRNA.

Most MEG families have several members, each with several spliced isoforms, enabling *S. mansoni* to produce a diverse range of proteins based on the internal environment of the human host. Transcriptomic studies have suggested that MEG proteins are involved in host-parasite interactions, particularly in the penetration of eggs through the intestinal wall. These proteins exhibit a unique structure with a signal peptide at the N-terminus, indicating their secretion. Additionally, MEG proteins often contain a significant amount of cysteine residues and are predicted to have a disordered protein structure (IDP).

Despite the intriguing nature of MEG proteins, there have been limited studies investigating their biophysical properties. Only a few publications have focused on transcriptomic and localization analyses, with even fewer dedicated to their structural and functional characterization. Therefore, the current dissertation aims to address this knowledge gap by focusing on a trio of MEG proteins—MEG 2.1, MEG 3.2, and MEG 6—that have been identified in *S. mansoni* eggs.

To study the MEG proteins, various expression systems, and conditions were tested to obtain recombinant proteins for structural analyses. Unfortunately, despite extensive efforts, none of the three MEG proteins were obtained in sufficient quantities for NMR analysis. MEG 6 protein was not expressed at all, while MEG 2.1 isoform 1 showed instability in the selected expression system. MEG 3.2 isoform 1 was successfully expressed using a specific protocol for toxic protein expression in a bacterial system, but the required concentration of both unlabeled and labeled protein for NMR analysis could not be achieved. Additionally, MEG 3.2 isoform 1 exhibited stability issues.

Given the challenges in recombinant protein expression, a "divide and reconstruct" strategy was employed. Short peptides derived from isoforms 1, 2, and 3 of MEG 2.1 were chemically synthesized. However, these peptides posed solubility issues in biologically relevant buffers and organic solvents. To overcome this, 100% deuterated dimethyl sulphoxide (DMSO-d6) was used for NMR measurements, while acetonitrile (also mixed with trifluoroethanol) (TFE) were tested as suitable solvents for circular dichroism (CD) analyses.

NMR experiments were performed on the synthesized peptides using homonuclear and heteronuclear techniques. The peptides were measured at natural abundance, resulting in lengthy analysis times. Subsequently, structural calculations using CYANA software were conducted. The reconstructed and minimized isoform 1 of MEG 2.1 exhibited the formation of short and long helices during molecular dynamics simulations. Additionally, phylogenetic

analysis, classification based on primary structure, and *ab initio* structural predictions were performed.

Through a combination of biochemical, biophysical, and bioinformatic analyses, this dissertation aimed to characterize the poorly studied egg-secreted MEG proteins. Notably, MEG 2.1 remains the only MEG family for which all splice variants have been structurally described. Overall, this study contributes to the understanding of MEG protein diversity and provides a foundation for future investigations into their functional significance in host-parasite interactions.

RÉSUMÉ

Les protéines MEG sont propres aux trématodes parasites, en particulier *Schistosoma mansoni*. Ces protéines sont codées par des gènes à micro-exons, qui ont une structure unique composée de micro-exons courts alternants avec de longs introns. Le séquençage du génome de *S. mansoni* a permis d'identifier 35 gènes codant 87 protéines MEG vérifiées. Ces gènes ne présentent aucune similitude avec d'autres gènes annotés, à l'exception de ceux que l'on trouve chez *Schistosoma spp*. Ils ne présentent pas de motifs ou de domaines fonctionnels identifiables. Les MEG subissent un épissage alternatif, ce qui entraîne la production de plusieurs isoformes à partir d'un seul ARNm.

La plupart des familles de MEG comptent plusieurs membres, chacun ayant plusieurs isoformes épissées, ce qui permet à *S. mansoni* de produire une variété de protéines en fonction de l'environnement interne de l'hôte. Des études transcriptomiques ont suggéré que les protéines MEG sont impliquées dans les interactions hôte-parasite, en particulier dans la pénétration des œufs à travers la paroi intestinale. Ces protéines présentent une structure unique avec un peptide signal à l'extrémité N-terminale, ce qui indique qu'elles sont sécrétées. Les MEGs contiennent souvent un nombre important de cystéine et sont modélisées avec une structure protéique désordonnée (IDP).

Malgré la nature intrigante des protéines MEG, peu d'études ont été menées sur leurs propriétés biophysiques. La thèse vise donc à combler cette lacune en se concentrant sur un trio de protéines MEG - MEG 2.1, MEG 3.2 et MEG 6 - qui ont été identifiées dans les œufs de S*. mansoni*.

Pour étudier les MEGs, différents systèmes d'expression ont été testés afin d'obtenir des protéines recombinantes pour des analyses structurales. Malgré des efforts considérables, aucune des trois MEGs n'a pu être obtenue en quantité suffisante pour être analysée par RMN. La protéine MEG 6 n'a pas été exprimée du tout. L'isoforme 1 de la MEG 2.1 s'est révélée instable dans le système d'expression S2. L'isoforme 1 de MEG 3.2 a été exprimée avec succès en utilisant un protocole pour l'expression de protéines toxiques dans un système bactérien, mais la concentration requise de protéines marquées et non marquées pour l'analyse RMN n'a pas pu être atteinte. En outre, l'isoforme 1 de la MEG 3.2 a présenté des problèmes de stabilité.

Compte tenu des difficultés liées à l'expression des protéines recombinantes, une stratégie de "division et reconstruction" a été employée. Des peptides courts issus des isoformes 1, 2 et 3 de la MEG 2.1 ont été synthétisés chimiquement. Ces peptides présentent des problèmes de solubilité dans les tampons et les solvants biologiquement pertinents. Pour résoudre ce problème, le DMSO-d6 à 100 % a été utilisé pour les mesures RMN, tandis que l'acétonitrile (également mélangé avec TFE) a été utilisé comme solvant pour les analyses de CD.

Des expériences RMN ont été réalisées sur les peptides synthétisés à l'aide de techniques homonucléaires et hétéronucléaires. Les spectres RMN ont été enregistrés sur des peptides en abondance naturelle, ce qui a entraîné de longs temps d'analyse. Des calculs structuraux ont ensuite été effectués à l'aide du logiciel CYANA. L'isoforme 1 reconstruite et minimisée de

MEG 2.1 a montré la formation d'hélices courtes et longues pendant les simulations de dynamique moléculaire. En outre, une analyse phylogénétique, une classification basée sur la structure primaire et des prédictions structurales ab initio ont été réalisées.

Grâce à une combinaison d'analyses biochimiques, biophysiques et bioinformatiques, cette thèse visait à caractériser les protéines MEG sécrétées dans l'œuf. MEG 2.1 reste la seule famille de MEG pour laquelle toutes les variantes d'épissage ont été décrites structuralement. Cette thèse contribue à la compréhension de la diversité des protéines MEG et fournit une base pour de futures recherches sur leur importance fonctionnelle dans les interactions hôte-parasite.

LIST OF ABBREVIATIONS

MEG - Micro-Exon Genes
EV - Extracellular Vesicle
TNF-α – Tumor Necrosis Factor Alpha
IPSE - Interleukin-4-Inducing Principle
NPC2 - Niemann-Pick C2
ESP15 - Egg-Secreted Proteins 15
WHO - World Health Organization
KO - Knock-Out
E/S - Excretory/Secretory
VAL - Venom Allergen-Like
SCP/TAPS - Sperm-Coating Protein/ Tpx-1/Ag5/PR-1/Sc7
dN/dS - a ratio of non-synonymous to synonymous substitutions
IPTG - Isopropyl-β -Thio-Galactopyranoside
OD_{600} - optical density of a sample measured at a wavelength of 600 nm
LB - Luria-Bertani Broth
Amp - Ampicillin
FPLC - Fast Protein Liquid Chromatography
BME - 2-Mercaptoethanol
SEC - Size-Exclusion Chromatography
IEX - Ion Exchange Chromatography
DMSO - Dimethyl sulfoxide
CD - Circular Dichroism
DLS - Dynamic Light Scattering
NMR - Nuclear Magnetic Resonance
TOCSY - Total Correlation Spectroscopy
NOESY - Nuclear Overhauser Effect Spectroscopy
HSQC - Heteronuclear Single Quantum Coherence Spectroscopy
DMD - Discrete Molecular Dynamics
cryo-EM - Cryogenic Electron Microscopy
MX - Macromolecular Crystallography
AF2 - Alpha Fold 2
plDDT - Predicted Local Distance Difference Test
TFE - trifluoroethanol
TCEP - tris(2-carboxyethyl)phosphine
NOE - Nuclear Overhauser Effect
IDP - Intrinsically Disordered Protein
tPSA - Topological Polar Surface Area
MD - Molecular Dynamics

TABLE OF CONTENTS

1 INTRODUCTION

Schistosomiasis is a vector-borne parasitic disease affecting over 250 million of the world's poorest people in sub-Saharan Africa, Brazil, and South-East Asia (Fig. 1.1) and is one of the most devastating parasitic diseases in the world, with up to 779 million people at risk (McManus et al. 2018). Beyond distribution in (sub)tropical regions, the introduction of human schistosomiasis to Southern Europe was reported in Corsica (Berry et al. 2016; Oleaga et al. 2019) and Almeria (Salas-Coronas et al. 2021).

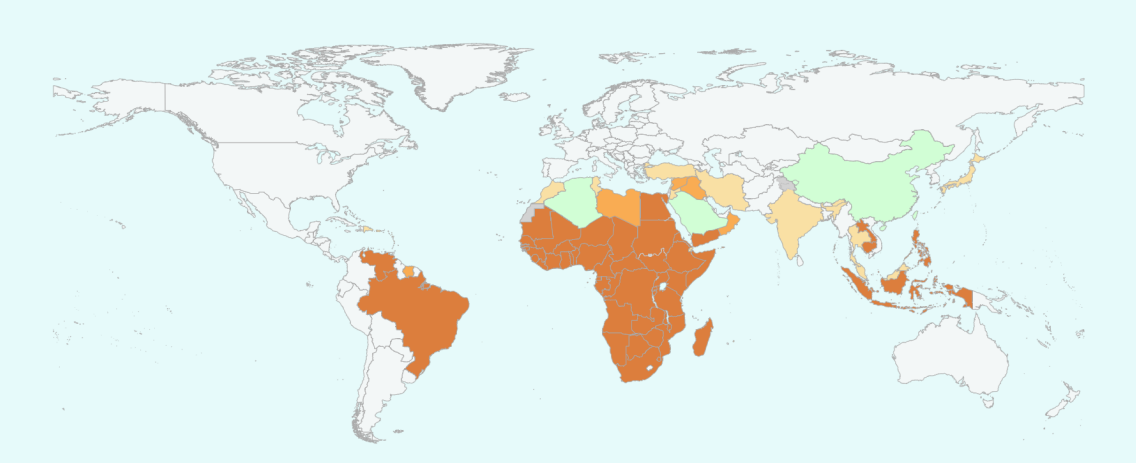

Figure 1.1 - Status of schistosomiasis in endemic countries in 2020 by WHO. Green color for countries with no preventive chemotherapy required, light orange for countries with the status "interruption of transmission to be confirmed", orange color for countries with the status of transmission to be determined, dark orange color for countries requiring preventive chemotherapy and grey color for countries that have the status "not applicable". Adapted from Global Health Observatory (GHO) interactive graph of the World Health Organization. Neglected diseases - Schistosomiasis - Status of schistosomiasis endemic countries 2020 -. https://apps.who.int/neglected_diseases/ntddata/sch/sch.html (Dyson and Wright).

Disease-causing pathogens are dioecious trematodes of the genus *Schistosoma,* living in the blood vessels of either the bladder (*Schistosoma haematobium*) or the intestine (*Schistosoma mansoni*, *Schistosoma japonicum,* and related species) (Lockyer et al. 2003; Howe et al. 2017). People become infected when larval forms of the parasite (cercariae) are released by freshwater snails and penetrate the skin during contact with infested water. Infected freshwater snails in irrigation canals, streams, and lakes transmit the flatworm parasite as villagers gather to collect water to support their household needs or perform agricultural and fishing activities. Precariously, there is just one partially effective drug, called praziquantel, that is authorized by the WHO for disease treatment and mass drug administration (MDA); therefore, the threat of resistance is ever-present (WHO 2022). Praziquantel is effective against all major forms of schistosomiasis, although it is more active against mature adult parasites than juvenile ones. Due to its efficacy, safety, price, and lack of alternatives, praziquantel has been the drug of choice for over 40 years, despite its limitations and concerns about resistance that have been confirmed not only in the laboratory but also in the field (Caffrey 2015; Vale et al. 2017). The mechanism of action of praziquantel is the influence

of calcium influx throughout the body of the parasite, muscle contraction, and surface modification. The key to all these mechanisms of action is the interference with the parasite's calcium metabolism, leading very rapidly to spastic paralysis of the musculature and then to morphological changes in the tissue (Cioli, Pica-Mattoccia, and Archer 1995; Cioli et al. 2014; Vale et al. 2017). The effectiveness of this drug depends on a number of factors, whether pharmacokinetic, its metabolization, but also on the species of *Schistosoma*, its strain (for *S. mansoni*: Belo Horizonte, Puerto Rican, Mwanza, Egyptian), stage, and also depends on the type of infection (single-sex female/single-sex male/bisexual) (Saoud 1966; Soliman et al. 1986; Pica-Mattoccia and Cioli 2004; Vale et al. 2017).

According to WHO, at least 236.6 million people required preventive treatment in 2019. The death estimates due to schistosomiasis need to be re-assessed, as it varies between 24,072 (1) and 200,000 (2) globally per year (WHO 2023). At the dose delivered, praziquantel is curative, but it does not prevent reinfections, and the reliance on just one drug means that resistance is a constant threat. Praziquantel also has many pharmacological and pharmaceutical weaknesses relating to its spectrum of efficacy, metabolism, dose, formulation, and taste (McManus et al. 2018; Caffrey 2015, 2007; Coulibaly et al. 2018; Olliaro, Delgado-Romero, and Keiser 2014; Zwang and Olliaro 2017). As a result, new medications and drug targets must be discovered. A pediatric praziquantel formulation is currently being developed that could be used in these large-scale treatment campaigns for preschool children. At the same time, there is still no cure or vaccine that could be used as a preventative measure. The five species of blood fluke cause two main forms of human schistosomiasis - intestinal and urogenital (Table 1.1).

Table 1.1 - Schistosoma species causing two major types of human schistosomiasis and its geographical distribution. Adapted from Global Health Observatory (GHO) interactive graph of World Health Organization. Neglected diseases - Schistosomiasis - Status of schistosomiasis endemic countries 2021, https://apps.who.int/neglected_diseases/ntddata/sch/sch.html (2021)

	species	***geographical distribution***
intestinal schistosomiasis	*Schistosoma mansoni*	Africa, the Middle East, the Caribbean, Brazil, Venezuela, and Suriname
	Schistosoma japonicum	China, Indonesia, the Philippines
	Schistosoma mekongi	Several districts of Cambodia and the Lao People's Democratic Republic
	Schistosoma guineensis and related *S. intercalatum*	Rain forest areas of central Africa
urogenital schistosomiasis	*Schistosoma haematobium*	Africa, the Middle East, Corsica (France)

1.1 The life cycle of Schistosomes

Schistosomes have two hosts: freshwater snails and mammals. They reproduce sexually in mammalian hosts, including humans, cattle, dogs, cats, rodents, pigs, horses, goats, and wild primates, and asexually in snails. The life cycle of *Schistosoma* starts with the eggs (n. 5 in Fig. 1.2) shed from an infected mammalian host (with feces or urine, depending on species). The eggs then hatch and release the free-living miracidia (n. 5 in Fig. 1.2), which penetrate a specific snail intermediate host (n. 6 in Fig. 1.3), which is *Biomphalaria for S. mansoni*, *Oncomelania for S. japonicum*, *Bulinus* for *S. haematobium, S. intercalatum, S. guineensis*. The only known intermediate host for *S. mekongi* is *Neotricula aperta* (CDC 2019). In the snail, asexual reproduction produces the subsequent formation of two stages of sporocysts. Sporocysts reproduce and develop into free-living cercariae, which are released in freshwater by the snails (n. 7 in Fig. 1.2). The infective cercariae swim, penetrate the skin of the mammalian host, and shed their forked tails, transforming into schistosomulae (n. 2 in Fig. 1.2). The schistosomulae migrate to the lungs *via* the venous circulation (n. 3 in Fig. 1.2), then to the heart, and then to the liver, where they mature and escape *via* the portal vein system (n. 3 in Fig. 1.2). Male and female adult worms copulate and reside in the mesenteric venules, the location of which varies by species (with some exceptions). *S. japonicum*, for example, is more commonly found in the superior mesenteric veins that drain the small intestine), while *S. mansoni* is more commonly found in the inferior mesenteric veins that drain the large intestine. Both species, however, may live in either place and can move between them (n. 4 in Fig. 1.2). *S. intercalatum* and *S. guineensis*, like *S. mansoni*, live in the inferior mesenteric plexus but lower in the gut. *S. haematobium* lives in the bladder's vesicular and pelvic venous plexus, although it can also be detected in the rectal venules. Female worms lay eggs in the portal and peri-vesical systems' small venules; the eggs can range in size from 7 to 28 mm, depending on the species. They are progressively transported toward the lumens of the intestine (*S. mansoni, S. japonicum, S. mekongi, S. intercalatum*/*guineensis*) or the bladder and ureters (*S. haematobium*), where they are removed with feces or urine (CDC 2019).

Each adult pair of *S. mansoni* worms produces approximately 300 eggs daily, which induce chronic inflammatory responses in the visceral organs, liver, or bladder, called a granuloma. The eggs endeavor to pass through the mesenteric vessels and across the intestinal wall into the intestinal lumen (Mooee and Sandgeound 1956). This leads to life-threatening organ damage and a lifetime of chronic pain and stunted growth. Schistosomes do not elicit immune protection; therefore, re-infection is the major threat in endemic countries. Untreated infection can lead to extensive granulomas formation, which gradually occupies a large proportion of the liver tissue, eventually inducing a blockage of the blood flow back to the heart through the portal system, creating portal hypertension, pulmonary hypertension, and esophageal varices resulting in death, if untreated (Colley and Secor 2014).

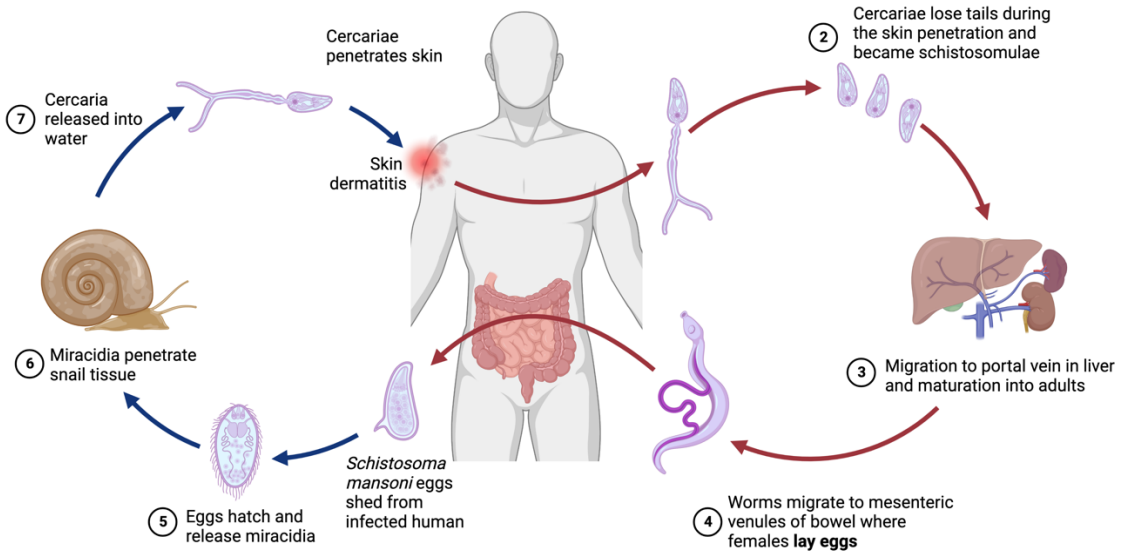

Figure 1.2* - Life cycle of *Schistosoma mansoni. Made with the use of BioRender.

1.2 Symptoms and clinical presentation of schistosomiasis

Symptoms of schistosomiasis are generated by the body's reaction to the eggs rather than to the worms themselves. A large number of infections are asymptomatic; thus, the affected individual is not treated in the early stages of schistosomiasis. Following cercariae penetration, a local cutaneous hypersensitivity reaction might occur, manifesting as tiny, itchy maculopapular sores. Acute schistosomiasis, called Katayama fever, is a systemic hypersensitivity reaction caused by *S. mansoni* and *S. japonicum* that can occur weeks after the first infection. Fever, cough, abdominal pain, diarrhea, hepato-splenomegaly, and eosinophilia are some of the systemic symptoms and indicators. In some cases, *Schistosoma* infections can cause lesions in the central nervous system. *S. japonicum* eggs in the brain can cause cerebral granulomatous illness, and granulomatous lesions around ectopic eggs in the spinal cord can also arise in *S. mansoni* and *S. haematobium* infections. Continued infection can lead to granulomatous responses and fibrosis in the affected organs such as the liver and spleen, as well as other signs and symptoms, such as embolic egg granulomas in the brain and spinal cord. Haematuria, scarring, calcification, squamous cell carcinoma, and embolic egg granulomas in the brain and spinal cord are all consequences of *S. haematobium* schistosomiasis (McManus et al. 2018; CDC 2019). In advanced untreated cases, fibrosis of the bladder and ureter, as well as kidney dysfunction, may be diagnosed, as well as bladder cancer. Genital sores, vaginal bleeding, pain during sexual intercourse, and nodules in the vulva are all symptoms of urogenital schistosomiasis in women. Urogenital schistosomiasis can cause seminal vesicles, prostate, and other organ pathology in men. Infertility is one of the long-term, permanent outcomes of urinary schistosomiasis.

In all cases, schistosomiasis causes anemia, stunting, and a loss of learning ability in children, but these effects are usually reversible with treatment. Chronic schistosomiasis

can make it difficult for people to work; it can eventually lead to death if untreated or very late treated. Because of hidden pathologies such as liver and renal failure, bladder cancer, and ectopic pregnancies caused by female genital schistosomiasis, it's difficult to assess the number of deaths directly caused by schistosomiasis (WHO 2023). In addition, some symptoms and typical manifestations of schistosomiasis caused by eggs can often be mistaken for symptoms of other diseases (Gobbi et al. 2015).

1.3 Egg migration and pathology

Local inflammations and the above-described serious pathologies are primarily caused by eggs that progressively accumulate in the host body because living adult worms found in blood vessels typically do not directly produce symptoms or trigger any localized inflammation (McManus et al. 2018). Because schistosome eggs lack evident motility capabilities, their evacuation (Fig. 1.3 and 1.4) is believed to be substantially reliant on host-driven processes. Schistosomes exert a range of immunomodulatory effects in the intestinal tissues to promote the formation of granulomas around transiting eggs, which is an important step in egg excretion(Doenhoff et al. 1981; Mathew and Boros 1986) (Byram and von Lichtenberg 1977). However, successful egg passage is not an easy task on which the completion of the life cycle and survival of the parasite depends. Only 22 % of all eggs produced are passed in the feces, 18 % remain in the wall of the large intestine, 32 % in the small intestine, 26 % in the liver, and 2 % in the mesenteries and its associated lymph nodes and pancreas. The number of eggs discovered in the spleen and lungs is negligible (McManus et al. 2018).

Adult worms typically survive for 5-10 years in the veins of the host while they release anti-inflammatory, vasoregulatory, and anticoagulation factors (Shariati et al. 2011). The deposition of eggs exerts an inflammatory process about 4-6 weeks after infection. The eggs' products cause progressive Th2-driven tissue fibrosis on one hand and attenuation of a tissue-damaging Th1 response *via* direct stimulation of a Th2 response on the other. The creation of granulomas (Fig. 5) surrounding tissue-entrapped eggs is a dynamic process involving adaptive immune responses that protect hepatocytes against cytotoxic egg products (Pirovich, Da'dara, and Skelly 2019). Egg secretions are known to be a mixture of highly potent immunomodulatory proteins, nucleic acids, lipids, and glycoproteins, such as IPSE/alpha-1 (interleukin-4-inducing principle of *S. mansoni* eggs) or omega-1 (Mathieson and Wilson 2010; Cass et al. 2007), which stimulate robust Th2 cell responses (Jankovic et al. 1999; Lundy and Lukacs 2013). The granuloma begins to form when the eggs reach maturity (after 6 days of development) within the intestinal wall or when imprisoned in the liver. Immature eggs are immunologically inactive until they reach that stage (Takaki et al. 2021). T cells, B cells, M2 macrophages, neutrophils, eosinophils, and mast cells are concentrated in granulomas, which are highly organized, well-defined multicellular clusters (Costain, MacDonald, and Smits 2018; Hams, Aviello, and Fallon 2013). Intestinal granulomatous inflammation facilitates egg penetration into

the gastrointestinal lumen. However, the development of fibrotic lesions in the place of the liver granuloma is the main cause of the pathology (Hams, Aviello, and Fallon 2013). The development of granuloma around the egg is divided into four major stages (Amaral et al. 2017; Lenzi et al. 1998) that differ in the organization of the layers, the number of inflammatory cells, and the presence of collagen fibers. The size, cellular composition, and extracellular matrix deposition of intestinal and hepatic granulomas differ. These changes reflect the infected liver's greater amount of dead eggs and differences in egg secretions between the two organs (Linder 2017; Weinstock and Boros 1983). Severe fibrosis, blood flow blockages, portal hypertension, and portocaval anastomosis result from the formation of hepatic granulomas, increasing the risk of life-threatening bleeding. Intestinal infection in humans can induce pseudo-polyps formation, ulceration, and stricture formation, despite the fact that pathology in the intestine is often less severe than in the liver (Barsoum, Esmat, and El-Baz 2013). Egg extravasation is also facilitated by angiogenesis, endothelial activation, and interactions with blood clotting components inside the vasculature (Shariati et al. 2011; Mebius et al. 2013). Another ingenious technique of *S. mansoni* eggs is the use of extracellular vesicles (EVs) for intercellular, inter-tissue communication without serious damage to host tissues (Bischofsberger et al. 2020). The use of EVs as a tool for balancing host immunity is a widespread tool for parasites, especially helminths (Khosravi et al. 2020).

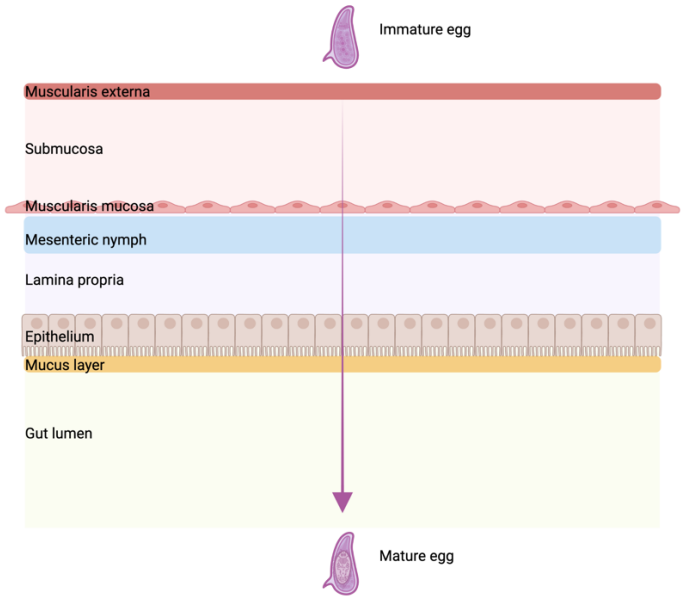

Figure 1.3 **- Migration of *S. mansoni* eggs from the mesenteric veins through the intestinal wall to the intestinal lumen.**

So far, there is only very limited knowledge about excretory-secretory (E/S) products of *S. mansoni* eggs and their ability to stimulate egg excretion or intestinal granuloma formation. Ultrastructural studies reveal pore openings in the shell of schistosome eggs (Jones et al. 2008; Cao, Wang, and Long 1982). When eggs mature, they cause host inflammatory cells to migrate from the vasculature endothelia and be discharged by

host excretions. This results in pathological processes in the surrounding intestinal tissues, which are essential for effective egg release into the gut environment. The process of egg release can be divided into four stages: a) egg release into the bloodstream and attachment to the endothelium; b) immune-dependent granuloma formation; c) transition between endothelium and epithelium; and d) release into the intestinal lumen (Schwartz and Fallon 2018). With a few exceptions, there is a considerable knowledge gap about the entire spectrum of secreted proteins that potentially interact with host tissues to facilitate the transition across the gut epithelium.

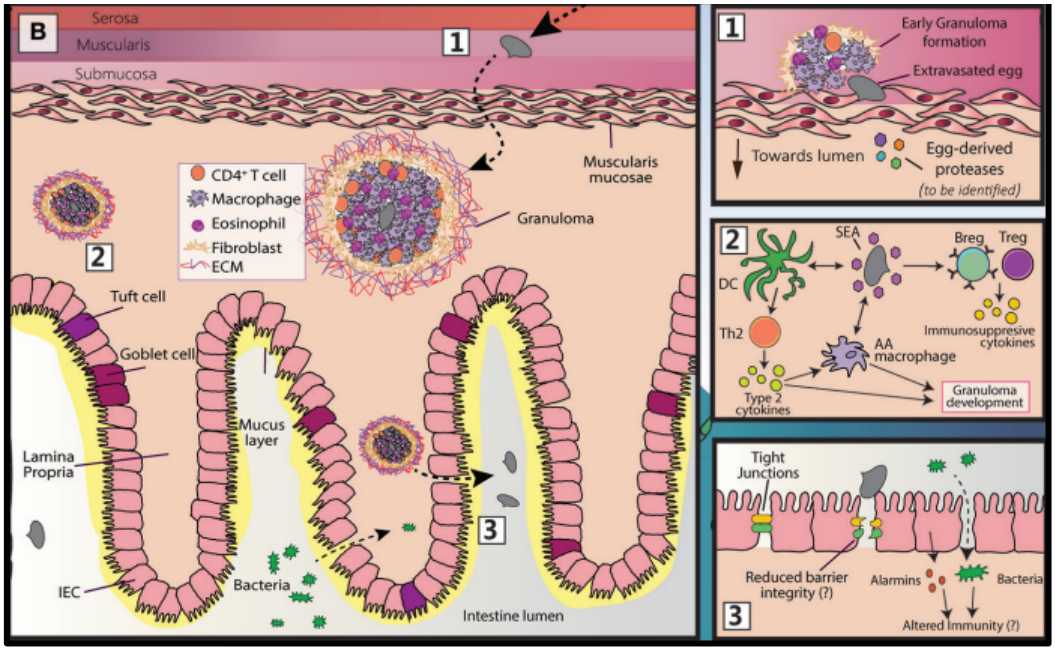

***Figure 1.4 - S. mansoni* egg migration adapted from "Schistosome Egg Migration:** Mechanisms, Pathogenesis and Host Immune Responses." by Costain, A. H. et al. 2018. *Frontiers in immunology*, 9, 3042.

1.4 Host-parasite molecular interactions

Parasites, because of their lifestyle, have been forced to develop a variety of different countermeasures, thanks to which they can remain in the host at least until they can start generating offspring. These mechanisms are very sophisticated, and not all are fully understood. This regulation of host immunity has both intended and unforeseen outcomes, both beneficial and detrimental (Maizels and McSorley 2016; Riffkin et al. 1996). Successful helminth parasites combine specific molecular strategies to deal with the threat of immediate immune attack with quantitative and dynamic properties that allow long-term establishment within an individual host (Maizels et al. 1993). Parasites can use immuno-evasion or immune-modulation strategies, or a combination of both (Riffkin et al. 1996). Immuno-evasion is a passive way of avoiding the effects of the immune response and can be represented, for example, by the formation of cysts, granulomas (i.e., encapsulation), or the secretion of proteins that protect the surface of the parasite (and its stages) from the effects of interactions with immunity. Most of the

time, it is the secretion of peptidoglycan or glycoprotein. Some parasites express and secrete proteases which then cleave the immunoglobulin G (IgG) (Kong et al. 1994) or proteases secreted by the schistosomula *S. mansoni*, which cleave IgG molecules bound to their surface, thus again modulating the immune response of the organism (Auriault et al. 1981). Some parasites secrete zinc metalloproteases that help degrade immune components (Hill et al. 1993) or secrete serine proteases that inhibit interleukin-2 proliferation (Nicolas-Gaulard, Moiré, and Boulard 1995). The second strategy is active immunemodulation. This strategy includes suppression of T- and B-cells, down-regulation of interleukins, production of TNF-α inhibitor, modulation of some cytokines, or even secretion of proteins that mimic their host proteins (Riffkin et al. 1996).

Achieving a better understanding of the molecular architecture of human biology, pathogenesis, and host-parasite interactions is possible by combining transcriptomic, genomic, and proteomic data (Han et al. 2009). Interactions between host and parasite include several life stages in the life cycle of *S. mansoni:* penetrating cercariae, migrating schistosomula, multiplying adults and eggs trapped in host tissues or migrating through the intestinal wall (McManus et al. 2018; Pearce et al. 1998). Mature schistosomes have also evolved highly powerful strategies for evading and even exploiting the cellular and humoral immune responses that they elicit (Maizels et al. 1993; McKerrow 1997) (Pearce and MacDonald 2002; Pearce and Sher 1987).

Other players in the field of parasite-host interactions are the aforementioned extracellular vesicles (EVs), whose secretion has been observed in a number of helminths. For the maintenance of biological processes (migration, maturation, reproduction) and host immunity evasion, every life stage of the schistosome life cycle has evolved its own inventive abilities to interact with the host via EVs. EVs are composed of a combination of proteins, lipids, nucleic acids, and the exact composition of individual EVs depends on the individual life stages of the schistosome. Proteomic studies have shown that *S. mansoni* exosomes contain a variety of proteases (serine, cysteine, metalloproteases), glycolytic enzymes, proteins that affect blood clotting. At the same time, Schistosomal EVs have been shown to contain different types of RNA molecules (mRNA, miRNA, small non-coding RNA - sncRNA). The authors agree that the proteins found in EVs may be potential therapeutic targets. (Tritten and Geary 2018; Bischofsberger et al. 2020; Nowacki et al. 2015; Sotillo et al. 2016; Samoil et al. 2018; Zhu et al. 2016)

Only a few *S. mansoni* egg molecules have been experimentally examined, and their function elucidated (Costain, MacDonald, and Smits 2018) (Macedonia and Mosimann 1994), despite the fact that several have been isolated and identified (Mathieson and Wilson 2010). Proteomic analyses of the *S. mansoni* egg secretome revealed the content of a wide range of proteins secreted from the eggs (Cass et al. 2007). It is well known that several abundant egg proteins are highly glycosylated (Dunne, Jones, and Doenhoff 1991; Schramm et al. 2006; Fitzsimmons et al. 2005). Glycomics analysis confirmed Omega-1 and IPSE as the most abundant glycoproteins in the secretome. It also showed

a high proportion of N- and O-glycan structures in the egg secretome. The authors suggest that these glycan epitopes could be another means for *S. mansoni* to modify host immunity by interacting with Th2 cells and interleukins (Jang-Lee et al. 2007).
The proteins IPSE/alpha-1, omega-1, Sm-p40, HSP70, Niemann-Pick C2 (NPC2), and peptidylglycine alpha hydroxylating monooxygenase were found among the most highly abundant ones found in egg secretions (Cass et al. 2007; Mair et al. 2004). A number of proteins with different functional categories of effects have been as well discovered: antioxidant proteins (peroxiredoxin 1, thioredoxin, …), heat shock, chaperone, and protein folding proteins (HSP70, polyubiquitin C, …), calcium-binding proteins (translationally controlled tumor protein, calpain), inflammation-inducing proteins (IPSE/alpha-1, venom allergen-like proteins), glycolytic or glycolytic feeder pathway enzymes (fructose bisphosphate aldolase, enolase,…), and scavenging pathway proteins (NPC2 - cholesterol transporter) (Cass et al. 2007).
Of the 188 proteins identified in the egg secretome, 118 contain traditional signal sequences or are secreted by a classical pathway. The rest of the analyzed egg-secreted proteins were either too short for predictive analysis or were proteins that were truncated but conserved ORFs. In addition, the predictive neural network SecretomeP v. 2.0 is trained on mammalian proteins, so it does not have suitable comparison datasets for pure helminth proteins (Cass et al. 2007). The majority of all 188 secreted egg proteins analyzed are involved in molecular functions (ion, nucleic acid, and protein binding, catalytic activity - hydrolase, oxidoreductase, transferase activity), biological processes (cellular death, homeostasis, metabolism, cell communication) or are part of cellular components (cytoplasmic organelles, protein complexes), as revealed by gene ontology analysis (Cass et al. 2007). The results of the small number of proteomic studies published so far vary depending on the egg extraction and egg-secreted proteins extraction protocol (Cass et al. 2007; Mathieson and Wilson 2010).
Mathieson and Wilson (2010) describe a relatively lower number of proteins and their isoforms/variants than Cass et al. (2007). Mathieson and Wilson (2010) criticize the egg and egg-secreted protein extraction protocol and claim that the proteomic analysis of Cass et al. (2007) is an analysis of the cytosolic proteins of dead and dying eggs and claim that there are far fewer secreted proteins than reported by Cass et al. (2007). However, their results for the most abundant proteins are consistent - IPSE, Omega-1, thioredoxin peroxidase - were confirmed in the study (Mathieson and Wilson 2010; Cass et al. 2007). The ESP15-family/ESP-like family proteins (“egg secreted proteins” - described in the Result 4.1 bioinformatic analysis of MEG superfamily chapter as members of the MEG 2 family) are added to these jointly identified most abundant proteins (Mathieson and Wilson 2010). The authors also claim that another of the described glycoproteins abundantly represented in the egg secretome, kappa 5, is not part of the secreted proteins because it is too large (97 kDa) to penetrate the eggshell; they claim that this protein is part of the hatching fluid. (Mathieson and Wilson 2010). Also, proteins that are already one of the targets for WHO vaccines were discussed in this comparative

proteomic study. These are paramyosin (motor protein), glutathione S-transferasetriose (defense), phosphate isomerase (energy metabolism), and Sm14 (fatty acid binding protein), which were also relatively abundant in the egg secretome (Bergquist et al. 2002; Mathieson and Wilson 2010).

Egg-specific glycoprotein IPSE-1/alpha-1 is a common E/S product that has been shown to stimulate basophils to produce anti-inflammatory IL-10 (interleukin-10) and IL-4, which then induces alternative macrophage activation and shift the immune response to Th2 polarization (Knuhr et al. 2018) (Schramm et al. 2006; Hewitson, Grainger, and Maizels 2009; Kaur et al. 2011; Haisch et al. 2001). Omega-1 (glycosylated T2 RNase) is another E/S glycoprotein that stimulates dendritic cells to produce fewer Th1 pro-inflammatory molecules, promoting the Th2 anti-inflammatory response (Hewitson, Grainger, and Maizels 2009; Everts et al. 2009). Omega-1 was recently investigated utilizing CRISPR/Cas9 technology, and it was discovered that KO eggs do not polarize Th2 responses; therefore, the granulomas surrounding these eggs are substantially less than those surrounding non-genetically modified eggs. IPSE/alpha-1 has been described as a sticky protein due to its high glycosylation and a positive charge at physiological pH. For this reason, it is assumed that IPSE could have potential use to keep more proteins near the egg (Mathieson and Wilson 2010).

In vitro, monoclonal antibodies that depleted omega-1 and IPSE/alpha-1 also reduced hepatotoxicity (Ittiprasert et al. 2019). Important studies on the proteomic identification of egg E/S products in *S. mansoni* have been reported; in particular, Williams and coworkers (Cass et al. 2007) identified IPSE-1, omega-1, and schistosome specific proteins with a unique multi-isoformic organization – called micro-exon gene (MEG) family (Fig. 6) - were identified as one group of secreted proteins (Mathieson and Wilson 2010). Mathieson and Wilson (2010) pointed to the presence of MEG 3 and ESP15 (MEG 2) family (and their isoforms) antigens in the egg secretome. Berriman et al. (2009) detected the presence of MEG 2, MEG 3, and MEG 6 in *S. mansoni* eggs (Fig. 5 panel b) (Berriman et al. 2009). MEG proteins are described in detail in the chapter MEG family proteins below.

Another of the interesting protein families described in the *S. mansoni* egg secretome are the Venom Allergen-Like (VAL/SmVAL) proteins. More than 200 VAL proteins were identified by transcriptomic analyses, all taxonomically belonging to the Trematoda, Cestoda, Monogenea, and Turbellaria (Chalmers and Hoffmann 2012). SmVAL family of proteins was first described after the analysis of *S. mansoni* transcriptome, and the name is derived from their similarity with wasp venom allergen (Verjovski-Almeida et al. 2003). SmVAL proteins contain SCP/TAPS protein domain; thus, they are members of the Sperm-coating protein/Tpx-1/Ag5/PR-1/Sc7 family, which varies in length between 120 and 170 amino acids. Tertiary structural studies of SCP/TAPS domain have demonstrated that this domain adopts a highly conserved α-β-α sandwich conformation (Fernández et al. 1997; Henriksen et al. 2001; Groves et al. 2004; Asojo, Loukas, et al. 2005; Shikamoto et al. 2005; Asojo, Goud, et al. 2005). Chalmers et al. (2008) study

divides SmVAL proteins into two groups based on their phylogenetic analysis and sequence similarities (Chalmers et al. 2008). Group 1 of SmVAL contains a signal peptide and a motif of six conserved cysteines that are able to form disulfide musts; group 2 differs in that it does not contain a signal peptide and does not have these conserved cysteines (Chalmers et al. 2008). Like MEG proteins, SmVAL proteins are also found among the developmental stages of *S. mansoni*, and their presence has been confirmed in eggs: SmVAL2, SmVAL3, SmVAL5, SmVAL9, SmVAL10, SmVAL11, and SmVAL13 in eggs (Chalmers et al. 2008; Cass et al. 2007). Very interesting is SmVAL6, which is structurally strikingly close to MEG proteins. For the SmVAL6 protein, 35 different isoforms have been identified, which are generated (as for the Schistosomal MEG protein) by alternative splicing. It contains an SCP/TAPS domain, which makes it a member of the SmVAL protein family, but the rest of its sequence is composed mainly of short symmetric exons, a structural pattern that has been described for Schistosomal MEG proteins (Chalmers et al. 2008). Verjovski-Almeida and DeMarco (2011), supported by Chalmers (2012), hypothesized that SmVAL6 was formed by recombination between the ancestral SmVAL6 that contained only the first four exons with a MEG, creating this SmVAL/MEG combined structure (Chalmers and Hoffmann 2012; Chalmers et al. 2008). SmVAL proteins exhibit better protein homology; therefore, a number of homology modeling of their structure has already been performed (for SmVAL1, SmVAL4, SmVAL13). In addition, the structure of the SCP/TAPS domain has already been resolved, as described above. Furthermore, an X-ray structure (with resolution 2.16 Å) of the *S. mansoni* VAL4 protein (Q1X6L4_SCHMA) has already been solved (Kelleher et al. 2014). Genes encoding both MEG and VAL proteins show high levels of nonsynonymous/synonymous) dN/dS substitutions. The two antigens that are vaccine candidates Sm29 and TSP-2 show similar dN/dS values to those found for MEG and VAL proteins. All of these proteins are exposed to the host immune system, so it is assumed that the genes encoding them have undergone a higher and accelerated evolutionary pressure to bypass host immunity. This accelerated co-evolutionary pressure on potential candidate molecules for vaccine production could greatly complicate the development of this vaccine (Philippsen, Wilson, and DeMarco 2015). Some of the MEGs have been used in the form of short synthetic peptides for immunization tests in mice. The most reactive ones were selected to create a protective vaccine, unfortunately, with very low efficacy (Farias et al. 2021).

1.5 Schistosomal MEG family proteins

SmMEG family proteins are an enigmatic group of schistosome-specific proteins previously described in the human parasite *Schistosoma mansoni*. The name MEG is an abbreviation of the words “micro-exon genes”, which also explains the nature of the proteins of this group (Fig. 1.5). SmMEGs are relatively small proteins - the longest of them is composed of 188 amino acids (UniProt 2022). The genes (*meg*) that encode SmMEG proteins are unique in that they contain up to 75 % of the sequence encoded by short, mostly symmetric exons (Berriman et al. 2009). A total of 35 *meg* genes of *Schistosoma mansoni* have been annotated, which are characterized by an alternation of long introns (0.1 - 5 kbp) and short symmetric exons, whose length ranges between 6 and 81 bp; the most abundant exons are 15 bp long (Berriman et al. 2009; Howe et al. 2017). The intron structure is also interesting - longer introns (up to 4 kbp) are located in the middle of the sequence, while their length shortens towards the 5' and 3' ends (length varies around 100 - 500 bp) (Howe et al. 2017). The WormBase Parasite database contains a total of 40 MEG-encoding genes, which, in addition to the 35 genes encoding *S. mansoni* proteins, also encode MEG proteins of *S. rodhaini*, *S. bovis*, *S. japonicum* and S*. haematobium* (Howe et al. 2017). The UniProt database contains 108 MEG proteins belonging mostly to *S. mansoni* but also to *S. haematobium,* and *S. japonicum* (UniProt 2022). Genes encoding MEG proteins show no similarity to any annotated genes, except *Schistosoma* spp.; at the same time, they do not contain any identifiable motifs or functional domains (Berriman et al. 2009). The proteins encoded by these genes also show no significant homology to non-*Schistosomal* MEG proteins. Most MEG proteins contain a signal peptide at the N-terminus, indicating that they are secreted proteins (Berriman et al. 2009). Additionally, most of them contain a relatively significant amount of cysteine, and a large part of the MEG members is predicated with a large part of the disordered structure (Felizatti et al. 2020). The primary protein structure makes MEGs challenging for *in vitro* studies.

The nomenclature and numbering of individual MEG proteins (and their isoforms) are confusing because some of the MEG proteins are named MEG-2 ESP15 (egg-secreted proteins), some of them are named as MEG-3 (Grail) family and antigen 10.3. At the same time, their numbering maintains continuity from MEG 1 to MEG 14, and then there is a gap in numbering, and the next one is MEG 26, which continues to MEG 32 (UniProt 2022).

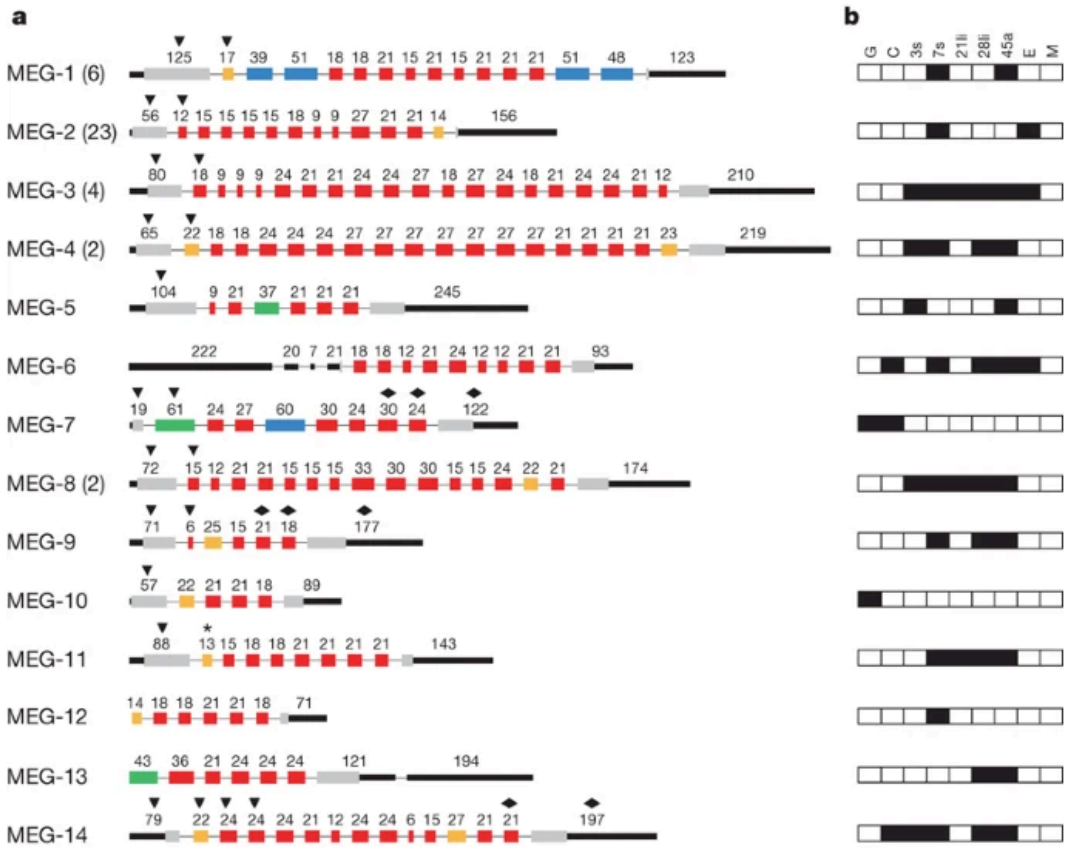

Figure 1.5 - Schematic representation of gene structure from MEG family members. a) Boxes represent exons and the numbers above their size in nucleotides. Black triangles indicate exons encoding *predicted signal peptides and transmembrane helices. Other characteristics associated with exons are indicated by color and grouped as follow: micro-exons having lengths of either multiples of 3 bp (red) or indivisible by 3 bp (orange); exons longer than 36 bp and having lengths of either multiples of 3 bp (blue) or indivisible by 3 bp (green); putative initiation and termination exons (grey); untranslated region. b) Diagram showing the presence of individual MEG families at different life stages: C - cercaria;* ***E - egg****; G - germball; M - miracidium; 3s and 7s, 3- and 7-day schistosomula; 21li and 28li, 21- and 28-day liver worms; 45a, 45-day adult worm pairs. Adapted from "The genome of the blood fluke Schistosoma mansoni" by Berriman, M. et al., 2009, Nature, 460, 352–358.*

So far, only MEG 14, MEG 24, and MEG 27 have been structurally characterized by circular dichroism. Of these three proteins, only MEG 14 and Sm10.3 protein (one isoform of the MEG 4 family) were recombinantly expressed, while MEG 24 and MEG 27 were chemically synthesized for the purpose of these biophysical studies (Felizatti et al. 2020; Lopes et al. 2013; Martins et al. 2014). Felizatti et al. (2020) found that both MEG 24 and MEG 27 contain a structure predominantly composed of α-helix, and the overall character of these molecules is amphipathic. This structure is similar to that of membrane-active peptides, whose actions disrupt membrane integrity (e.g., antimicrobial peptides) (Avci, Sariyar Akbulut, and Ozkirimli 2018). Both of these MEGs interact with mimetic membranes to reduce membrane fluidity and promote membrane leakage. Whole-mount in situ hybridization (WISH) experiments performed in the same study proved that MEG 24 is transcribed in subtegumental cells and MEG 27 in the esophagus of adult worms (Felizatti et al. 2020).

MEG 14 has been determined by synchrotron radiation circular dichroism (SRCD) as a member of intrinsically disordered proteins (IDPs). Additionally, it has been shown that

by changing the external conditions, partial folding can be achieved (Lopes et al. 2013). Conformational changes/structural morphing of MEG 14 protein was demonstrated in interaction with human S100 protein, which is involved in the inflammatory response. Unfolded MEG 14 has been shown to fold upon binding to negatively charged membranes or calgranulins (mediators of inflammatory processes) (Orcia et al. 2017).

Sm10.3 (MEG 4 family member) was expressed for the purpose of immunization and hemagglutination assays. At the same time, *Sm10.3* was confirmed to be expressed in esophageal epithelia, esophageal lumen, and intestinal epithelia, which could suggest a possible role of the Sm10.3 protein (MEG 4 family member) in the host blood feeding process (Martins et al. 2014).

It has been shown that several genes of the MEG family are developmentally regulated. However, all the MEGs are preferentially expressed in the intra-mammalian phase of the life cycle. As already mentioned, MEG proteins are among the most abundant components of the egg secretome (Mathieson and Wilson 2010; Berriman et al. 2009). The bioinformatically predicted secretion of the MEG 3 family and MEG 2 family, based on the presence of a signal peptide sequence, was confirmed by protein identification of *S. mansoni* egg-secreted MEG proteins. Moreover, many different isoforms of MEG proteins were identified, which proves that they undergo alternative splicing (DeMarco et al. 2010).

It is difficult to infer a function from the sequence, but some hypotheses were proposed based on the localization of transcripts and proteins within some *S. mansoni* life stages. In adult schistosomes, a variety of MEGs were detected in the esophageal glands (Wilson et al. 2015); therefore, it was suggested that the secreted MEGs could interact with leukocytes that are part of the worm blood meal. Berriman et al. (2009) pointed to the presence of MEG 2, MEG 3, and MEG 6 in the eggs (Fig. 5). It has been hypothesized that MEG 2 and MEG 3 family proteins might be involved in the interaction of eggs with host tissues - specifically the gut (Mathieson and Wilson 2010; DeMarco et al. 2010; Castro-Borges and Wilson 2022). This hypothesis is also supported by the unpublished transcriptomic data from Dvorak's lab (Fig. 6), where significantly higher expression of all three studied MEG-family proteins was observed in mature eggs compared to immature ones (Fig. 6). The MEG 2 and MEG 3 family had already been linked to the migration of the schistosomulum and the subshell envelope of the mature egg (DeMarco et al. 2010; Wilson 2012). The authors proposed that the MEG proteins produced by the eggs and larvae interact with and change vascular endothelium function (Wilson 2012). MEG 2 and MEG 3 family proteins were also found in head glands (adults) and sub-shell of eggs, MEG 3.4 was found in the anterior and posterior esophageal glands (adults), and MEG 6 was found in the tegument and cell bodies of adults (Castro-Borges and Wilson 2022).

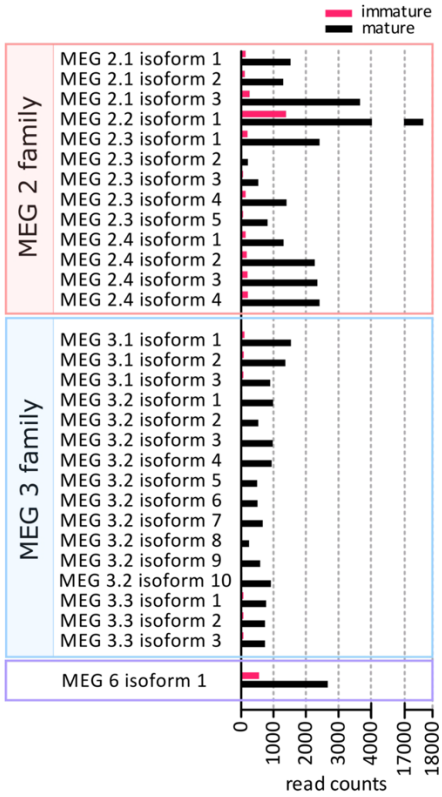

***Figure 1.6* - Comparison of the expression of MEG proteins in *S. mansoni* mature/immature eggs.** Expression values were counted from Illumina short reads originating from RNA isolated from *S. mansoni* eggs. Unpublished preliminary data, Dvorak's lab 2023.

Each family of MEGs has its members, and each member has one or multiple isoforms. In view of the data obtained and confirmed by us regarding the expression of MEGs in *S. mansoni* eggs, 3 members of the MEG family of proteins were selected for this thesis. On the basis of unpublished transcriptomic data from Dr. Dvorak's lab (Fig. 1.6), MEG 2.1 from the MEG 2 family; MEG 3.2 from the MEG 3 family and MEG 6, which is the only member of the MEG 6 family, were selected. For all studied proteins, the longest isoforms were selected (most often referred to as isoform 1), which cover the whole cassette of exons, within which alternative splicing and subsequent formation of different proteins occur. It is plausible that *Schistosomes* use MEGs as a comprehensive tool for their adaptation to the environment and/or immunemodulation in interactions with the host. With the help of alternative splicing, *S. mansoni* can create up to 18 different proteins from one unspliced isoform, as it is the case of MEG 1 for example.
MEG 3.2 protein has 10 isoforms (Fig. 1.7), MEG 2.1 has three and MEG 6 has only one.

Figure 1.7 – Alternative splice variants of MEG 3.2 protein. The deduced structure for sequenced transcripts of the MEG-3.2 gene is displayed. Boxes represent the coding region for exons; narrow lines represent the introns (not to scale). Exons generated using an alternative splicing site are shown in grey. Coding exons that are being read in a frame different from the most abundant isoform (1) are shown as a rainbow box. Exons are shown to scale, but for illustrative purposes, intron length is not proportional to size. Numbers in the boxes (and their color) indicate exon size. Colors for individual lengths are pink - 18 bp, light blue - 9 bp, lavender - 24 bp, light yellow - 24 bp, green - 27 bp, gray - 19/62/23/71/20 bp. Figure adapted from "Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts" by DeMarco, R. et al., 2010, *Genome Research*, 20, p. 1112-1121. Copyright 2010 by Cold Spring Harbor Laboratory Press.

The structure of micro exons is not new to researchers; it has been described in many genes across animal and plant phyla (Volfovsky, Haas, and Salzberg 2003). However, it is the unique character of the Schistosomal MEGs that makes the gene and MEG proteins unique. The work on solving the structure of these enigmatic proteins is important not only in terms of finding new potential drug targets, which are needed in the case of the global schistosomiasis threat. Solving the structure of the MEG proteins will allow homologous modeling of other MEG proteins, finding potential binding/interaction partners, and thus may reveal much about the human immune system, about which there is still much to learn. MEG proteins are considered to be one of the many very powerful tools of *S. mansoni* to regulate the host immune response and/or to avoid the attention of the immune system. Their structural and biological functions are yet unclear, although partial structural elements and some of their *in vivo* localizations have already been described.

2 SCIENTIFIC HYPOTHESIS AND OBJECTIVES

H1: Alternative splicing of MEG family proteins is a powerful tool of *Schistosoma mansoni* to achieve maximum variability of expressed proteins for best immunomodulation.

H2: MEG family proteins are important players among the secreted proteins as they interact with the host and assist in egg passage through the intestinal wall.

O1: In order to elucidate the structure and variability of isoforms within individual members of the MEG protein families, recombinant proteins will be expressed, purified, and then biophysically analyzed. To complete this objective, I proposed to express and purify MEG 2.1 (isoform 1), MEG 3.2 (isoform 1), and MEG 6 proteins in various expression systems (bacterial, yeast, insect, cell-free) to obtain soluble properly folded proteins for further investigation. The final step will be a thorough biophysical and structural characterization of the complexes using nuclear magnetic resonance (NMR) spectroscopy (liquid-state), dynamic light scattering (DLS), and circular dichroism (CD).

O2: Given the lack of research on interaction partners for the whole family of MEG proteins, the structure of the investigated proteins will have to be solved first. Then it will be possible to proceed with interaction studies. Screening of potential interaction partners will first be carried out in the framework of *in silico* modeling, such as prediction of putative interaction surface(s) and molecular docking.

REFERENCES

Amaral, Kátia B, Thiago P Silva, Felipe F Dias, Kássia K Malta, Florence M Rosa, Sócrates F Costa-Neto, Rosana Gentile, and Rossana CN Systém. 2017. 'Histological assessment of granulomas in natural and experimental Schistosoma mansoni infections using whole slide imaging', *PLoS One*, 12: e0184696.

Asojo, Oluwatoyin A, Gaddam Goud, Kajari Dhar, Alex Loukas, Bin Zhan, Vehid Deumic, Sen Liu, Gloria EO Borgstahl, and Peter J Hotez. 2005. 'X-ray structure of Na-ASP-2, a pathogenesis-related-1 protein from the nematode parasite, Necator americanus, and a vaccine antigen for human hookworm infection', *Journal of molecular biology*, 346: 801-14.

Asojo, Oluwatoyin A, Alex Loukas, Mehmet Inan, Rick Barent, Jicai Huang, Brad Plantz, Amber Swanson, Mark Gouthro, Michael M Meagher, and Peter J Hotez. 2005. 'Crystallization and preliminary X-ray analysis of Na-ASP-1, a multi-domain pathogenesis-related-1 protein from the human hookworm parasite Necator americanus', *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 61: 391-94.

Auriault, C, MA Ouaissi, G Torpier, H Eisen, and A Capron. 1981. 'Proteolytic cleavage of IgG bound to the Fc receptor of Schistosoma mansoni schistosomula', *Parasite immunology*, 3: 33-44.

Avci, Fatma Gizem, Berna Sariyar Akbulut, and Elif Ozkirimli. 2018. 'Membrane active peptides and their biophysical characterization', *Biomolecules*, 8: 77.

Barsoum, Rashad S, Gamal Esmat, and Tamer El-Baz. 2013. 'Human schistosomiasis: clinical perspective', *Journal of advanced research*, 4: 433-44.

Bergquist, Robert, Maged Al-Sherbiny, Rashida Barakat, and Richard Olds. 2002. 'Blueprint for schistosomiasis vaccine development', *Acta Tropica*, 82: 183-92.

Berriman, Matthew, Brian J Haas, Philip T LoVerde, R Alan Wilson, Gary P Dillon, Gustavo C Cerqueira, Susan T Mashiyama, Bissan Al-Lazikani, Luiza F Andrade, and Peter D Ashton. 2009. 'The genome of the blood fluke Schistosoma mansoni', *Nature*, 460: 352-58.

Bischofsberger, Miriam, Franziska Winkelmann, Anne Rabes, Emil C Reisinger, and Martina Sombetzki. 2020. 'Pathogen-host interaction mediated by vesicle-based secretion in schistosomes', *Protoplasma*, 257: 1277-87.

Byram, JE, and Franz von Lichtenberg. 1977. 'Altered schistosome granuloma formation in nude mice', *The American Journal of Tropical Medicine and Hygiene*, 26: 944-56.

Caffrey, Conor R. 2007. 'Chemotherapy of schistosomiasis: present and future', *Current opinion in chemical biology*, 11: 433-39.

———. 2015. 'Schistosomiasis and its treatment', *Future medicinal chemistry*, 7: 675-76.

Cao, HM, YF Wang, and S Long. 1982. 'A study of ultrastructure of egg shell of Schistosoma japonicum. I. Transmission electron microscopic observation of S. japonicum egg', *Annales de parasitologie humaine et comparee*, 57: 345-52.

Cass, Cynthia L, Jeffrey R Johnson, Lindsay L Califf, Tao Xu, Hector J Hernandez, Miguel J Stadecker, John R Yates III, and David L Williams. 2007. 'Proteomic analysis of Schistosoma mansoni egg secretions', *Molecular and biochemical parasitology*, 155: 84-93.

Castro-Borges, William, and R Alan Wilson. 2022. 'Schistosome proteomics: updates and clinical implications', *Expert Review of Proteomics*: 1-15.

CDC. 2019. 'Schistosomiasis'. https://www.cdc.gov/dpdx/schistosomiasis/index.html.

Chalmers, Iain W, and Karl F Hoffmann. 2012. 'Platyhelminth Venom Allergen-Like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum', *Parasitology*, 139: 1231-45.

Chalmers, Iain W, Andrew J McArdle, Richard MR Coulson, Marissa A Wagner, Ralf Schmid, Hirohisa Hirai, and Karl F Hoffmann. 2008. 'Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the Schistosoma mansoni venom allergen-like (SmVAL) gene family', *BMC genomics*, 9: 1-20.

Cioli, Donato, Livia Pica-Mattoccia, and Sydney Archer. 1995. 'Antischistosomal drugs: past, present… and future?', *Pharmacology & therapeutics*, 68: 35-85.

Cioli, Donato, Livia Pica-Mattoccia, Annalisa Basso, and Alessandra Guidi. 2014. 'Schistosomiasis control: praziquantel forever?', *Molecular and biochemical parasitology*, 195: 23-29.

Colley, DG, and WE Secor. 2014. 'Immunology of human schistosomiasis', *Parasite immunology*, 36: 347-57.

Costain, Alice H, Andrew S MacDonald, and Hermelijn H Smits. 2018. 'Schistosome egg migration: mechanisms, pathogenesis and host immune responses', *Frontiers in immunology*, 9: 3042.

Coulibaly, Jean T, Mamadou Ouattara, Beatrice Barda, Jürg Utzinger, Eliézer K N'Goran, and Jennifer Keiser. 2018. 'A rapid appraisal of factors influencing praziquantel treatment compliance in two communities endemic for schistosomiasis in Côte d'Ivoire', *Tropical medicine and infectious disease*, 3: 69.

DeMarco, Ricardo, William Mathieson, Sophia J Manuel, Gary P Dillon, Rachel S Curwen, Peter D Ashton, Alasdair C Ivens, Matthew Berriman, Sergio Verjovski-Almeida, and R Alan Wilson. 2010. 'Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts', *Genome research*, 20: 1112-21.

Doenhoff, MJ, S Pearson, DW Dunne, Q Bickle, S Lucas, J Bain, R Musallam, and O Hassounah. 1981. 'Immunological control of hepatotoxicity and parasite egg excretion in Schistosoma mansoni infections: stage specificity of the reactivity of immune serum in T-cell deprived mice', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 75: 41-53.

Dunne, DW, FM Jones, and MJ Doenhoff. 1991. 'The purification, characterization, serological activity and hepatotoxic properties of two cationic glycoproteins (α1 and ω1) from Schistosoma mansoni eggs', *Parasitology*, 103: 225-36.

Dyson, H Jane, and Peter E Wright. 2021. 'NMR illuminates intrinsic disorder', *Current Opinion in Structural Biology*, 70: 44-52.

Everts, Bart, Georgia Perona-Wright, Hermelijn H Smits, Cornelis H Hokke, Alwin J van der Ham, Colin M Fitzsimmons, Michael J Doenhoff, Jürgen van der Bosch, Katja Mohrs, and Helmut Haas. 2009. 'Omega-1, a glycoprotein secreted by Schistosoma mansoni eggs, drives Th2 responses', *Journal of Experimental Medicine*, 206: 1673-80.

Farias, Leonardo P, Gillian M Vance, Patricia S Coulson, Juliana Vitoriano-Souza, Almiro Pires da Silva Neto, Arporn Wangwiwatsin, Leandro Xavier Neves, William

Castro-Borges, Stuart McNicholas, and Keith S Wilson. 2021. 'Epitope mapping of exposed tegument and alimentary tract proteins identifies putative antigenic targets of the attenuated schistosome vaccine', *Frontiers in immunology*, 11: 624613.

Felizatti, Ana P, Ana E Zeraik, Luis GM Basso, Patricia S Kumagai, Jose LS Lopes, Bonnie A Wallace, Ana PU Araujo, and Ricardo DeMarco. 2020. 'Interactions of amphipathic α-helical MEG proteins from Schistosoma mansoni with membranes', *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1862: 183173.

Fernández, César, Thomas Szyperski, Thierry Bruyere, Paul Ramage, Egon Mösinger, and Kurt Wüthrich. 1997. 'NMR solution structure of the pathogenesis-related protein P14a', *Journal of molecular biology*, 266: 576-93.

Fitzsimmons, Colin M, Gabriele Schramm, Frances M Jones, Iain W Chalmers, Karl F Hoffmann, Christoph G Grevelding, Manfred Wuhrer, Cornelis H Hokke, Helmut Haas, and Michael J Doenhoff. 2005. 'Molecular characterization of omega-1: A hepatotoxic ribonuclease from Schistosoma mansonieggs', *Molecular & Biochemical Parasitology*, 144: 123-27.

Gobbi, Federico, Giulia Martelli, Luciano Attard, Dora Buonfrate, Andrea Angheben, Valentina Marchese, Laura Bortesi, Maria Gobbo, Elisa Vanino, and Pierluigi Viale. 2015. "Schistosoma mansoni eggs in spleen and lungs, mimicking other diseases." In.: Public Library of Science San Francisco, CA USA.

Groves, Matthew R, Audrey Kuhn, Astrid Hendricks, Susanne Radke, Ramon L Serrano, J Bernd Helms, and Irmgard Sinning. 2004. 'Crystallization of a Golgi-associated PR-1-related protein (GAPR-1) that localizes to lipid-enriched microdomains', *Acta Crystallographica Section D: Biological Crystallography*, 60: 730-32.

Haisch, Karin, Gabriele Schramm, Franco H Falcone, Christian Alexander, Max Schlaak, and Helmut Haas. 2001. 'A glycoprotein from Schistosoma mansoni eggs binds non-antigen-specific immunoglobulin E and releases interleukin-4 from human basophils', *Parasite immunology*, 23: 427-34.

Hams, Emily, Gabriella Aviello, and Padraic G Fallon. 2013. 'The schistosoma granuloma: friend or foe?', *Frontiers in immunology*, 4: 89.

Han, Ze-Guang, Paul J Brindley, Sheng-Yue Wang, and Zhu Chen. 2009. 'Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction', *Annual review of genomics and human genetics*, 10: 211-40.

Henriksen, Anette, Te P King, Osman Mirza, Rafaél I Monsalve, Kåre Meno, Henrik Ipsen, Jørgen N Larsen, Michael Gajhede, and Michael D Spangfort. 2001. 'Major venom allergen of yellow jackets, Ves v 5: structural characterization of a pathogenesis-related protein superfamily', *Proteins: Structure, Function, and Bioinformatics*, 45: 438-48.

Hewitson, James P, John R Grainger, and Rick M Maizels. 2009. 'Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity', *Molecular and biochemical parasitology*, 167: 1-11.

Hill, DOLORES E, HR Gamble, MARCIA L Rhoads, RAYMOND H Fetterer, and JF Urban. 1993. 'Trichuris suis: a zinc metalloprotease from culture fluids of adult parasites', *Experimental parasitology*, 77: 170-78.

Howe, Kevin L, Bruce J Bolt, Myriam Shafie, Paul Kersey, and Matthew Berriman. 2017. 'WormBase ParaSite− a comprehensive resource for helminth genomics', *Molecular and biochemical parasitology*, 215: 2-10.

Ittiprasert, Wannaporn, Victoria H Mann, Shannon E Karinshak, Avril Coghlan, Gabriel Rinaldi, Geetha Sankaranarayanan, Apisit Chaidee, Toshihiko Tanno, Chutima Kumkhaek, and Pannathee Prangtaworn. 2019. 'Programmed genome editing of the omega-1 ribonuclease of the blood fluke, Schistosoma mansoni', *Elife*, 8: e41337.

Jang-Lee, Jihye, Rachel S Curwen, Peter D Ashton, Berangere Tissot, William Mathieson, Maria Panico, Anne Dell, R Alan Wilson, and Stuart M Haslam. 2007. 'Glycomics analysis of Schistosoma mansoni egg and cercarial secretions', *Molecular & Cellular Proteomics*, 6: 1485-99.

Jankovic, Dragana, Thomas A Wynn, Marika C Kullberg, Sara Hieny, Patricia Caspar, Stephanie James, Allen W Cheever, and Alan Sher. 1999. 'Optimal vaccination against Schistosoma mansoni requires the induction of both B cell-and IFN-γ-dependent effector mechanisms', *The Journal of Immunology*, 162: 345-51.

Jones, Malcolm K, Sze How Bong, Kathryn M Green, Philadelphia Holmes, Mary Duke, Alex Loukas, and Donald P McManus. 2008. 'Correlative and dynamic imaging of the hatching biology of Schistosoma japonicum from eggs prepared by high pressure freezing', *PLoS Neglected Tropical Diseases*, 2: e334.

Kaur, Ishwinder, Gabriele Schramm, Bart Everts, Thomas Scholzen, Karin B Kindle, Christian Beetz, Cristina Montiel-Duarte, Silke Blindow, Arwyn T Jones, and Helmut Haas. 2011. 'Interleukin-4-inducing principle from Schistosoma mansoni eggs contains a functional C-terminal nuclear localization signal necessary for nuclear translocation in mammalian cells but not for its uptake', *Infection and Immunity*, 79: 1779-88.

Kelleher, Alan, Rabih Darwiche, Wanderson C Rezende, Leonardo P Farias, Luciana CC Leite, Roger Schneiter, and Oluwatoyin A Asojo. 2014. 'Schistosoma mansoni venom allergen-like protein 4 (SmVAL4) is a novel lipid-binding SCP/TAPS protein that lacks the prototypical CAP motifs', *Acta Crystallographica Section D: Biological Crystallography*, 70: 2186-96.

Khosravi, Mojdeh, Elnaz Sadat Mirsamadi, Hamed Mirjalali, and Mohammad Reza Zali. 2020. 'Isolation and functions of extracellular vesicles derived from parasites: the promise of a new era in immunotherapy, vaccination, and diagnosis', *International journal of nanomedicine*: 2957-69.

Knuhr, Katrin, Kristina Langhans, Sandra Nyenhuis, Kerstin Viertmann, Anna M Overgaard Kildemoes, Michael J Doenhoff, Helmut Haas, and Gabriele Schramm. 2018. 'Schistosoma mansoni egg-released IPSE/alpha-1 dampens inflammatory cytokine responses via basophil interleukin (IL)-4 and IL-13', *Frontiers in immunology*, 9: 2293.

Kong, Y, Y-B Chung, S-Y Cho, and S-Y Kang. 1994. 'Cleavage of immunoglobulin G by excretory–secretory cathepsin S-like protease of Spirometra mansoni plerocercoid', *Parasitology*, 109: 611-21.

Lenzi, Henrique L, Eitan Kimmel, Helio Schechtman, Marcelo Pelajo-Machado, Waldemiro S Romanha, Ronaldo G Pacheco, Mario Mariano, and Jane A Lenzi. 1998. 'Histoarchitecture of schistosomal granuloma development and involution: morphogenetic and biomechanical approaches', *Memórias do Instituto Oswaldo Cruz*, 93: 141-51.

Linder, Ewert. 2017. 'The schistosome egg in transit', *Ann Clin Pathol*, 5: 1110.

Lockyer, AE, PD Olson, P Østergaard, D Rollinson, DA Johnston, SW Attwood, VR Southgate, P Horak, SD Snyder, and TH Le. 2003. 'The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of Schistosoma Weinland, 1858', *Parasitology*, 126: 203-24.

Lopes, Jose Luiz S, Debora Orcia, Ana Paula U Araujo, Ricardo DeMarco, and Bonnie A Wallace. 2013. 'Folding factors and partners for the intrinsically disordered protein micro-exon gene 14 (MEG-14)', *Biophysical journal*, 104: 2512-20.

Lundy, Steven K, and Nicholas W Lukacs. 2013. 'Chronic schistosome infection leads to modulation of granuloma formation and systemic immune suppression', *Frontiers in immunology*, 4: 39.

Macedonia, JG, and JE Mosimann. 1994. 'Kinetics of egg production and egg excretion by Schistosoma mansoni and S. japonicum in mice infected with a single pair of worms', *American Journal of Tropical Medicine and Hygiene*, 50: 281.

Mair, Gunnar R, Mark J Niciu, Michael T Stewart, Gerry Brennan, Hanan Omar, David W Halton, Richard Mains, Betty A Eipper, Aaron G Maule, and Tim A Day. 2004. 'A functionally atypical amidating enzyme from the human parasite Schistosoma mansoni', *The FASEB journal*, 18: 114-21.

Maizels, Rick M, Don AP Bundy, Murray E Selkirk, Deborah F Smith, and Roy M Anderson. 1993. 'Immunological modulation and evasion by helminth parasites in human populations', *Nature*, 365: 797-805.

Maizels, Rick M, and Henry J McSorley. 2016. 'Regulation of the host immune systém by helminth parasites', *Journal of Allergy and Clinical Immunology*, 138: 666-75.

Martins, Vicente P, Suellen B Morais, Carina S Pinheiro, Natan RG Assis, Barbara CP Figueiredo, Natasha D Ricci, Juliana Alves-Silva, Marcelo V Caliari, and Sergio C Oliveira. 2014. 'Sm 10.3, a Member of the Micro-Exon Gene 4 (MEG-4) Family, Induces Erythrocyte Agglutination In Vitro and Partially Protects Vaccinated Mice against Schistosoma mansoni Infection', *PLoS Neglected Tropical Diseases*, 8: e2750.

Mathew, RANJIT C, and DL Boros. 1986. 'Anti-L3T4 antibody treatment suppresses hepatic granuloma formation and abrogates antigen-induced interleukin-2 production in Schistosoma mansoni infection', *Infection and Immunity*, 54: 820-26.

Mathieson, William, and R Alan Wilson. 2010. 'A comparative proteomic study of the undeveloped and developed Schistosoma mansoni egg and its contents: the miracidium, hatch fluid and secretions', *International journal for parasitology*, 40: 617-28.

McKerrow, JH. 1997. 'Cytokine induction and exploitation in schistosome infections', *Parasitology*, 115: 107-12.

McManus, Donald P, David W Dunne, Moussa Sacko, Jürg Utzinger, Birgitte J Vennervald, and Xiao-Nong Zhou. 2018. 'Schistosomiasis (primer)', *Nature Reviews: Disease Primers*, 4: 13.

Mebius, Mirjam M, Perry JJ van Genderen, Rolf T Urbanus, Aloysius GM Tielens, Philip G de Groot, and Jaap J van Hellemond. 2013. 'Interference with the host haemostatic systém by schistosomes', *PLoS pathogens*, 9: e1003781.

Mooee, DV, and JH Sandgeound. 1956. 'The Relative Egg producing Capacity of Schistosoma mansoni and S chistosoma japonicum', *American Journal of Tropical Medicine and Hygiene*, 5: 831-40.

Nicolas-Gaulard, I, Nathalie Moiré, and C Boulard. 1995. 'Effect of the parasite enzyme, hypodermin A, on bovine lymphocyte proliferation and interleukin-2 production via the prostaglandin pathway', *Immunology*, 85: 160.

Nowacki, Fanny C, Martin T Swain, Oleg I Klychnikov, Umar Niazi, Alasdair Ivens, Juan F Quintana, Paul J Hensbergen, Cornelis H Hokke, Amy H Buck, and Karl F Hoffmann. 2015. 'Protein and small non-coding RNA-enriched extracellular vesicles are released by the pathogenic blood fluke Schistosoma mansoni', *Journal of extracellular vesicles*, 4: 28665.

Olliaro, Piero, Petra Delgado-Romero, and Jennifer Keiser. 2014. 'The little we know about the pharmacokinetics and pharmacodynamics of praziquantel (racemate and R-enantiomer)', *Journal of Antimicrobial Chemotherapy*, 69: 863-70.

Orcia, Debora, Ana Eliza Zeraik, Jose LS Lopes, Joci NA Macedo, Clarissa Romano Dos Santos, Katia C Oliveira, Leticia Anderson, Bonnie A Wallace, Sergio Verjovski-Almeida, and Ana PU Araujo. 2017. 'Interaction of an esophageal MEG protein from schistosomes with a human S100 protein involved in inflammatory response', *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861: 3490-97.

Pearce, Edward J, Anne La Flamme, Elizabeth Sabin, and Laura Rosa Brunet. 1998. 'The initiation and function of Th2 responses during infection with Schistosoma mansoni', *Mechanisms of Lymphocyte Activation and Immune Regulation VII: Molecular Determinants of Microbial Immunity*: 67-73.

Pearce, Edward J, and Andrew S MacDonald. 2002. 'The immunobiology of schistosomiasis', *Nature Reviews Immunology*, 2: 499-511.

Pearce, EJ, and A Sher. 1987. 'Mechanisms of immune evasion in schistosomiasis', *Contributions to microbiology and immunology*, 8: 219-32.

Philippsen, Gisele S, R Alan Wilson, and Ricardo DeMarco. 2015. 'Accelerated evolution of schistosome genes coding for proteins located at the host–parasite interface', *Genome Biology and Evolution*, 7: 431-43.

Pica-Mattoccia, Livia, and Donato Cioli. 2004. 'Sex-and stage-related sensitivity of Schistosoma mansoni to in vivo and in vitro praziquantel treatment', *International journal for parasitology*, 34: 527-33.

Pirovich, David, Akram A Da'dara, and Patrick J Skelly. 2019. 'Why do intravascular schistosomes coat themselves in glycolytic enzymes?', *BioEssays*, 41: 1900103.

Riffkin, Michael, HENG-FONG SEOW, David Jackson, Lorena Brown, and Paul Wood. 1996. 'Defence against the immune barrage: helminth survival strategies', *Immunology and cell biology*, 74: 564-74.

Samoil, Vitalie, Maude Dagenais, Vinupriya Ganapathy, Jerry Aldridge, Anastasia Glebov, Armando Jardim, and Paula Ribeiro. 2018. 'Vesicle-based secretion in schistosomes: analysis of protein and microRNA (miRNA) content of exosome-like vesicles derived from Schistosoma mansoni', *Scientific reports*, 8: 3286.

Saoud, MFA. 1966. 'The infectivity and pathogenicity of geographical strains of Schistosoma mansoni', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 60: 585-600.

Schramm, G, A Gronow, J Knobloch, V Wippersteg, CG Grevelding, J Galle, H Fuller, RG Stanley, PL Chiodini, and H Haas. 2006. 'IPSE/alpha-1: a major immunogenic component secreted from Schistosoma mansoni eggs', *Molecular and biochemical parasitology*, 147: 9-19.

Schwartz, Christian, and Padraic G Fallon. 2018. 'Schistosoma "eggs-iting" the host: granuloma formation and egg excretion', *Frontiers in immunology*, 9: 2492.

Shariati, F, JL Pérez-Arellano, C Carranza, J Lopez-Aban, B Vicente, M Arefi, and A Muro. 2011. 'Evaluation of the role of angiogenic factors in the pathogenesis of schistosomiasis', *Experimental parasitology*, 128: 44-49.

Shikamoto, Yasuo, Kyoko Suto, Yasuo Yamazaki, Takashi Morita, and Hiroshi Mizuno. 2005. 'Crystal structure of a CRISP family Ca2+-channel blocker derived from snake venom', *Journal of molecular biology*, 350: 735-43.

Soliman, Gamil N, Faiza M El Assal, Noshy S Mansour, and Kohar Garo. 1986. 'Comparison of two Egyptian strains of Schistosoma mansoni in hamsters', *Zeitschrift für Parasitenkunde*, 72: 353-63.

Sotillo, Javier, Mark Pearson, Jeremy Potriquet, Luke Becker, Darren Pickering, Jason Mulvenna, and Alex Loukas. 2016. 'Extracellular vesicles secreted by Schistosoma mansoni contain protein vaccine candidates', *International journal for parasitology*, 46: 1-5.

Takaki, Kevin K, Gabriel Rinaldi, Matthew Berriman, Antonio J Pagán, and Lalita Ramakrishnan. 2021. 'Schistosoma mansoni eggs modulate the timing of granuloma formation to promote transmission', *Cell Host & Microbe*, 29: 58-67. e5.

Tritten, Lucienne, and Timothy G Geary. 2018. 'Helminth extracellular vesicles in host–parasite interactions', *Current opinion in microbiology*, 46: 73-79.

UniProt, Consortium The. 2022. 'UniProt: the Universal Protein Knowledgebase in 2023', *Nucleic Acids Research*, 51: D523-D31.

Vale, Nuno, Maria João Gouveia, Gabriel Rinaldi, Paul J Brindley, Fátima Gärtner, and José M Correia da Costa. 2017. 'Praziquantel for schistosomiasis: single-drug metabolism revisited, mode of action, and resistance', *Antimicrobial agents and chemotherapy*, 61: e02582-16.

Verjovski-Almeida, Sergio, Ricardo DeMarco, Elizabeth AL Martins, Pedro EM Guimarães, Elida PB Ojopi, Apuã CM Paquola, João P Piazza, Milton Y Nishiyama Jr, João P Kitajima, and Rachel E Adamson. 2003. 'Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni', *Nature genetics*, 35: 148-57.

Volfovsky, Natalia, Brian J Haas, and Steven L Salzberg. 2003. 'Computational discovery of internal micro-exons', *Genome research*, 13: 1216-21.

Weinstock, JV, and DL Boros. 1983. 'Organ-dependent differences in composition and function observed in hepatic and intestinal granulomas isolated from mice with Schistosomiasis mansoni', *Journal of immunology (Baltimore, Md.: 1950)*, 130: 418-22.

WHO. 2022. 'WHO guideline on control and elimination of human schistosomiasis', World Health Organization.

———. 2023. 'Schistosomiasis Fact Sheet', Accessed 11/04. https://www.who.int/news-room/fact-sheets/detail/schistosomiasis.

Wilson, R Alan. 2012. 'Virulence factors of schistosomes', *Microbes and infection*, 14: 1442-50.

Wilson, R Alan, Xiao Hong Li, Sandy MacDonald, Leandro Xavier Neves, Juliana Vitoriano-Souza, Luciana CC Leite, Leonardo P Farias, Sally James, Peter D Ashton, and Ricardo DeMarco. 2015. 'The schistosome esophagus is a 'hotspot'for microexon

and lysosomal hydrolase gene expression: implications for blood processing', *PLoS Neglected Tropical Diseases*, 9: e0004272.

Zhu, Shanli, Sai Wang, Yu Lin, Pengyue Jiang, Xiaobin Cui, Xinye Wang, Yuanbin Zhang, and Weiqing Pan. 2016. 'Release of extracellular vesicles containing small RNAs from the eggs of Schistosoma japonicum', *Parasites & vectors*, 9: 1-9.

Zwang, Julien, and Piero Olliaro. 2017. 'Efficacy and safety of praziquantel 40 mg/kg in preschool-aged and school-aged children: a meta-analysis', *Parasites & vectors*, 10: 1-16.

List of figures

List of tables

3 METHODOLOGY

3.1 Bioinformatics

Bioinformatic analyses were performed to determine the characteristics of MEG proteins under study, such as hydrophobicity, charge, prediction of secondary and tertiary structure, and as a complement to structural studies, for the refinement of the determined structure. Combined with the biochemical and biophysical part of the project, it allowed to gain a more comprehensive view of the MEG family proteins.

3.1.1 Phylogenetics and primary sequence analysis

A thorough bioinformatic analysis of the three studied MEG family proteins, and their isoforms was performed. Protein sequences of MEG family proteins were obtained from the UniProt database (UniProt 2022), last accessed on April 2023). A phylogenetic tree of all previously described and submitted MEG protein sequences has been constructed. There are 103 *Schistosoma mansoni* MEG proteins and 107 *Schistosoma* genus proteins (including *Schistosoma japonicum*, *Schistosoma haematobium*) annotated in the Uniprot database. These sequences were then manually trimmed to remove duplicate entries, which displayed the same sequence with two different accession numbers, one from NCBI and one from UniProt. For each discovered duplication, only the UniProt sequence was strictly preserved to maintain the cross-annotation of the WormBase ParaSite database (Howe et al. 2017).

Phylogenetic trees were built with both Simple phylogeny and PRANK (Loytynoja and Goldman 2008) on the EBI server (Madeira et al. 2022). The one from PRANK was more consistent with the gene clustering and also with the type of retrotransposon sequences which had been found at the boundaries of the *megs*; therefore it was retained and visualized on iTOL (interactive Tree of Life) server (Letunic and Bork 2007). Emboss on the EBI server (Madeira et al. 2022) was used to put in evidence conserved linear motifs, which were displayed with Weblogo (Crooks et al. 2004). Jalview software (Waterhouse et al. 2009) was also used to visualize the results. Primary sequence analysis was performed by ProtParam tool (Gasteiger et al. 2005) on the Expasy website (Gasteiger et al. 2003); the results on calculated molecular weights, isoelectric point, aliphatic index and GRAVY index are presented in Table A. N-glycosylation predictions were made using NetNGlyc - 1.0 (Gupta and Brunak 2001), O-glycosylation predictions were made using NetOGlyc - 4.0 (Steentoft et al. 2013). Graphical overview of aligned sequences from the blastp was performed in use of the Constraint-based Multiple Alignment Tool (COBALT) (Papadopoulos and Agarwala 2007).

3.1.2 Ab Initio Protein Structure Prediction

A series of predictive modelling of MEG protein structures was also performed. Deep learning-based modelling software – AlphaFold2 (Jumper et al. 2021), Robetta (Song et al. 2013) (Baek et al. 2021), and the ESMFold (Lin et al. 2023) – were used for *ab initio* protein structure prediction. The Swiss-Model (Waterhouse et al. 2018) server was used to determine as many homologous proteins as possible. For AlphaFold 2 predicted local distance difference test (pLDDT) score (0-100) is a per-residue confidence score, with values greater than 90 indicating high confidence, and values below 50 indicating low confidence. This measure estimates whether the predicted residue has similar distances to neigbhoring C-alpha atoms (within 15 Angstroms) in agreement with distances in the true structure.

NCBI Blast (Standard Protein BLAST – blastp (Mahram and Herbordt 2015)) was also used to identify protein homology. The resulting models (from AlphaFold, Robetta and Swiss-Model) were analyzed and used for structure comparison and subsequent modelling in MODELLER (Webb and Sali 2016), which was used as a plugin in the UCSF Chimera (Pettersen et al. 2004). UCSF Chimera X (Pettersen et al. 2021) and Pymol (Schrödinger and DeLano 2020) software were also used for the visualization results of the coordinate files.

3.2 Recombinant protein expression

The first stage of expression was characterized by cloning design of the selected MEG proteins to be expressed in a heterologous host. We decided to start from MEG 3.2 (isoform 1), MEG 2.1 (isoform 1) and MEG 6 - Fig. 3.1.

```
                          10          20          30          40          50          60
MEG_2.1_isoform_1   MKLSGANCLVVFSLLQLLVAFSHCDINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLG  65
MEG_3.2_isoform_1   MLFVALILIISLHSFDCVFTARETQQECVRHCGGHNEYVTRYCGGLCSGSTGPQTFYCYLGCSHN  65
MEG_6               MVQNPKNTKKINRTIRRSTKTVIVITDRVQNIVLGHRLLHHRIPTIKRSKSHGINKNETVSNLFP  65

                     70          80          90          100         110         120
MEG_2.1_isoform_1   NYQRHQRMKNYLEEVCENFIYTP                                          88
MEG_3.2_isoform_1   ASNQNDFDKCLPKCNGSPQLTESSCQNDCGRVTTHPELCGIVCGGNVGDSFPLCLYNCDQGNGSG 130
MEG_6

                          140
MEG_2.1_isoform_1
MEG_3.2_isoform_1   NFDECKTKCYEMAGR                                                   145
MEG_6
```

Figure 3.1 - Sequences of MEG 2.1 isoform 1 (88 aa), MEG 3.2 isoform 1 (145 aa) and MEG 6 (65 aa) proteins.

Detailed cloning design, plasmid sequence, restriction enzymes used, primer design are given in the following tables: Table 3.1 for MEG 2.1 isoform 1, Table 3.2 for MEG 3.2 isoform 1, and Table 3.3 for MEG 6.

Table 3.1 - Overview of the cloning design for MEG 2.1 isoform 1 that have been used to express this protein.

name & expression system	DNA sequence	protein sequence	plasmid	restriction enzymes	primers
MEG 2.1 isoform 1 "long" (with SP) *E. coli*	atgaagttatccggcgcaaactgtttagtcgtattcagtcttcttcagttacttgttgcttcacactgcgacataaacgacataacatgcaacaaaacggtatgctgcgcttccgaggatggaaaaaaggttctttatgttgcgagaaggatggctgtccaatcccctcaacaccagacctttttgctgggcaattatcaaaggcaccagagaatgaagaattacctggaagaggtctgtgaaaactttatctacacaccctga	MAMKLSGANCLVVFSLLQLLVAFSHCDINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEVCENFIYTPENLYFQAFEHHHHHH	pET SUMO Champion	TOPO cloning with linearized vector supplied in the kit	Forward: 5'-ATGAAGTTATCCGGCGCAAA-3' Reverse: 5'-TTGAAATAGATGTGTGGGACT-3'
MEG 2.1 isoform 1 "long" (with SP) *E. coli*	cagccggcgatggccatgaagttatccggcgcaaactgtttagtcgtattcagtcttcttcagttacttgttgcttttttcacactgcgacataaacgacataacatgcaacaaaacggtatgctgcgcttccgaggatggaaaaaagggttctttatgttgcgagaaggatggctgtccaatcccctcaacaccagaccttttgctgggcaattatcaaaggcaccagagaatgaagaattacctggaagaggtctgtgaaaactttatctacacacccgaaaacctgtacttccaagcgttcgagcaccaccaccaccaccactga	MAMKLSGANCLVVFSLLQLLVAFSHCDINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEVCENFIYTPENLYFQAFEHHHHHH	pET 22(b)+	Nco I Xho I	Forward: 5'-CAGCCGGCGATGGCCATGAAGTTATCCGGCGCAAACTGTTTA-3' Reverse: 5'-CACTTTTGGACATGAAGGTTCGCAAGCTCGTGGTGGTGG-3'
MEG 2.1 isoform 1 "long" (with SP) *Komagataella phaffii*	gagaaagagaggctgaagctgcaaagttatccggcgcaaactgtttagtcgtattcagtcttcttcagttacttgttgcttttttcacactgcgacataaacgacataacatgcaacaaaacggtatgctgcgcttccgaggatggaaaaaaggttctttatgttgcgagaaggatggctgtccaatcccctcaacaccagaccttttgctgggcaattatcaaaggcaccagagaatgaagaattacctggaagaggtctgtgaaaactttatctacacacccggtcatcatcatcatcatcattaaggaattcacgtggcccag	MKLSGANCLVVFSLLQLLVAFSHCDINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEVCENFIYTP	pPICZα B	Pst I	Forward: 5'-GAGAAAGAGAGGCTGAAGCTGCA-3' Reverse: 5'-GTAATTCCTTAAGTGCACCGGGTC-3'
MEG 2.1 isoform 1 "long" (with SP) cell-free	gcggccgcgagaatctttatttttcagggcatgaaactatcgggagcaaactgtttggtagtcttcagcctactacaacttcttgtggcattttcacactgtgatattaatgacataacatgcaacaagacagttgttgcgcatcagaagacggtaaaaaaggttccctatgttgtgagaaagatggttgtccaattccaagcactccagatcttttgcttggaaattaccagcgccatcaacgaatgaaaaattatttagaggaagtgtgcgaaaatttcatatacacgccataataaggatcc	MKLSGANCLVVFSLLQLLVAFSHCDINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEVCENFIYTP	pIVEX2.4d	Not I Bam HI	N/A - synthetized
MEG 2.1 isoform 1 "short" (deletion of 24 aa from N-terminus) cell-free	gcggccgcgagaatctttatttttcagggcgctcgtgaaactcaacaagaatgtgtacgacattgtggtggacacaatgaatatgtgactcgatactgtggtggtctgtgttctggcagcacaggaccacaaacattctattgttatctcggatgcagtcataacgccagtaaccaaaacgatttcgacaaatgtttaccaaagtgtaatggtagtccccagcttactgagtcatcgtgtcagaatgactgtggtcgtgttaccacacaccctgaattgtgtggtatcgtttgtggtggaaatgttggagactcatttccactgtgtttgtatactgcgatcagggaaatggttcgggaaactttgacgaatgtaaaacaaagtgctacgaaatggcggggacggtgataaggatcc	DINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEVCENFIYTP	pIVEX2.4d	Not I Bam HI	N/A - synthetized

MEG 2.1 isoform 1 "short" (deletion of 24 aa from N-terminus) S2 *Drosophila* cells	gacataaacgacataacatgcaacaaaacggtatgctgcgcttccgaggatggaaaaagggttctttatgttgcgagaaggatggctgtccaatcccctcaacaccagacctttgctgggcaattatcaaaggcaccagagaatgaagaattacctggaagaggtctgtgaaaactttatctacacaccctaa	DINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEVCENFIYTP	pMT_BiP_SLIN	Kpn I Xho I	Forward: 5'-GGTACCAGACATAAACGACATAACATGCAACAAA-3' Reverse: 5'-ACACTTTTGAAATAGATGTGTGGGATTGAGCTC-3'

Table 3.2 - Overview of the cloning design for MEG 3.2 isoform 1 that have been used to express this protein.

Name & expression system	DNA sequence	Protein sequence	Plasmid	Restriction enzymes	Primers
MEG 3.2 isoform 1 - "long" (with SP) *E. coli*	atgcttttcgttgctttgattcttatcatttcacttcattccttcgattgtgtattcaccgcaagagaaaccccaacaggaatgtgtcagacattgcggtggtcataatgaatacgtcacacgttactgtggaggacttgttcaggctcaacgggaccacagacattctactgttacttaggttgctctcataatgcttccaaccaaaatgactttgacaagtgccttccaaaatgcaacggctcccctcagcttacggaatcctcatgccaaaacgattgtggaagggtcactacccatccagaactttgcggcatcgtgtgtggtggtaatgtggtgactccttcccattatgcttgtataattgtgaccagggaaacggatcaggcaacttcgatgagtgtaaaacgaaatgctatgaaatggccggaaggtga	MLFVALILIISLHSFDCVFTARETQQECVRHCGGHNEYVTRYCGGLCSGSTGPQTFYCYLGCSHNASNQNDFDKCLPKCNGSPQLTESSCQNDCGRVTTHPELCGIVCGGNVGDSFPLCLYNCDQGNGSGNFDECKTKCYEMAGR	pET SUMO Champion	TOPO cloning with linearized vector supplied in the kit	Forward: 5'-ATGCTTTTCGTTGCTTTGATT-3' Reverse: 3'-ATACTTTACCGGCCTTCCACT-5'
MEG 3.2 isoform 1 - "long" (with SP) *E. coli*	cagccggcgatggccatgcttttcgttgctttgattcttatcatttcacttcattccttcgattgtgtattcaccgcaagagaaacccaacaggaatgtgtcagacattgcggtggtcataatgaatacgtcacacgttactgtggaggactttgttcaggctcaacggaccacagacattctactgttacttaggttgctcctcataatgcttccaaccaaaatgactttgacaagtgccttccaaaatgcaacggctcccctcagcttacggaatcctcatgccaaaacgattgtggaaggtcactacccatccagaactttgcggcatcgtgtggtggtaatgtgggtgactccttcccattatgcttgtataattgtgaccagggaaacggatcaggcaacttcgatgagtgtaaaacgaaatgctatgaaatggccggaagggaaaacctgtacttccaagcgttcgagcaccaccaccaccaccactga	MAMLFVALILIISLHSFDCVFTARETQQECVRHCGGHNEYVTRYCGGLCSGSTGPQTFYCYLGCSHNASNQNDFDKCLPKCNGSPQLTESSCQNDCGRVTTHPELCGIVCGGNVGDSFPLCLYNCDQGNGSGNFDECKTKCYEMAGRENLYFQAFEHHHHHH	pET-22(b)+	Xho I Nco I	Forward: 5'-CAGCCGGCGATGGCCATGCTTTTCGTTGCTTTGATTCTTATCATTTCA-3' Reverse: 5'-CTTTTGGACATGAAGGTTCGCCAGCTCGTGGTGGTGG-3'
MEG 3.2 isoform 1 - "short" (deletion of 20 aa from N-terminus) *E. coli*	gcaagagaaacccaacaggaatgtgtcagacattgcggtggtcataatgaatacgtcacacgttactgtggaggactttgttcaggctcaacgggaccacagacattctactgttacttaggttgctctcataatgcttccaaccaaaatgactttgacaagtgccttccaaaatgcaacggctcccctcagcttacggaatcctcatgccaaaacgattgtggaagggtcactacccatccagaactttgcggcatcgtgtggtggtaatgtgggtgactccttcccattatgcttgtataattgtgaccagggaaacggatcaggcaacttcgatgagtgtaaaacgaaatgctat gaaatggccggaaggtga	ARETQQECVRHCGGHNEYVTRYCGGLCSGSTGPQTFYCYLGCSHNASNQNDFDKCLPKCNGSPQLTESSCQNDCGRVTTHPELCGIVCGGNVGDSFPLCLYNCDQGNGSGNFDECKTKCYEMAGR	pET-22(b)+	Xho I Nco I	Forward: 5'-CAGCCGGCGATGGCCATGGCAAGAGAAACCCAACAGGAATGTGTCAGAC-3' Reverse: 5'-CTTTTGGACATGAAGGTTCGCCAGCTCGTGGTGGTGG-3'

MEG 3.2 isoform 1 "long" (with SP) *Komagataella phaffii*	gagaaagagaggctgaagctgcacttttcgttg ctttgattcttatcatttcacttcattccttcgatt gtgtattcaccgcaagagaaacccaacaggaa tgtgtcagacattgcggtggtcataatgaatac gtcacacgttactgtggaggactttgttcaggct caacgggaccacagacattctactgttacttag gttgctctcataatgcttccaaccaaaatgactt tgacaagtgccttccaaaatgcaacggctcccc tcagcttacggaatcctcatgccaaaacgattg tggaagggtcactacccatccagaactttgcgg catcgtgtgtggtggtaatgtgggtgactccttc ccattatgcttgtataatttgtgaccagggaaac ggatcaggcaacttcgatgagtgtaaaacgaa atgctatgaaatggccggaaggggtcatcatc atcatcatcattaaggaattcacgtggcccag	MLFVALILIISLHSFDCV FTARETQQECVRHCG GHNEYVTRYCGGLCSG STGPQTFYCYLGCSHN ASNQNDFDKCLPKCN GSPQLTESSCQNDCGR VTTHPELCGIVCGGNV GDSFPLCLYNCDQGN GSGNFDECKTKCYEM AGR	pPICZα B	Pst I	Forward: 5'-GAGAAAGAGA GGCTGAAGCTGCA-3' Reverse: 5'-GTAATTCCTTAAGT GCACCGGGTC-3'
MEG 3.2 isoform 1 "long" (with SP) cell-free	gcggccgcgagaatctttattttcagggcatgct gttcgttgcattgattctgatcatctctctccact cattcgactgtgtattcacagctcgtgaaactca acaagaatgtgtacgacattgtggtggacaca atgaatatgtgactcgatactgtggtggtctgtg ttctggcagcacaggaccacaaacattctattg ttatctcggatgcagtcataacgccagtaacca aaacgatttcgacaaatgtttaccaaagtgtaa tggtagtccccagcttactgagtcatcgtgtcag aatgactgtggtcgtgttaccacacaccctgaa ttgtgtggtatcgtttgtggtggaaatgttggag actcatttccactgtgtttgtataactgcgatca gggaaatggttcgggaaactttgacgaatgta aaacaaagtgctacgaaatggcgggacggtg ataaggatcc	MLFVALILIISLHSFDCV FTARETQQECVRHCG GHNEYVTRYCGGLCSG STGPQTFYCYLGCSHN ASNQNDFDKCLPKCN GSPQLTESSCQNDCGR VTTHPELCGIVCGGNV GDSFPLCLYNCDQGN GSGNFDECKTKCYEM AGR	pIVEX2.4d	Not I Bam HI	N/A - synthetized
MEG 3.2 isoform 1 "short" (deletion of 20 aa from N-terminus) cell-free	gcggccgcgagaatctttattttcagggcgctc gtgaaactcaacaagaatgtgtacgacattgtg gtggacacaatgaatatgtgactcgatactgtg gtggtctgtgttctggcagcacaggaccacaaa cattctattgttatctcggatgcagtcataacgc cagtaaccaaaacgatttcgacaaatgtttacc aaagtgtaatggtagtccccagcttactgagtc atcgtgtcagaatgactgtggtcgtgttaccac acaccctgaattgtgtggtatcgtttgtggtgga aatgttggagactcatttccactgtgtttgtata actgcgatcagggaaatggttcgggaaactttg acgaatgtaaaacaaagtgctacgaaatggcg gacggtgataaggatcc	ARETQQECVRHCGGH NEYVTRYCGGLCSGST GPQTFYCYLGCSHNAS NQNDFDKCLPKCNGS PQLTESSCQNDCGRVT THPELCGIVCGGNVGD SFPLCLYNCDQGNGSG NFDECKTKCYEMAGR	pIVEX2.4d	Not I Bam HI	N/A - synthetized
MEG 3.2 isoform 1 "short V1" (deletion of 20 aa from N-terminus) S2 *Drosophila* cells	gcaagagaaacccaacaggaatgtgtcagac attgcggtggtcataatgaatacgtcacacgtt actgtggaggactttgttcaggctcaacgggac cacagacattctactgttacttaggttgctctcat aatgcttccaaccaaaatgactttgacaagtgc cttccaaaatgcaacggctcccctcagcttacg gaatcctcatgccaaaacgattgtggaagggtc actacccatccagaactttgcggcatcgtgtgtg gtggtaatgtgggtgactccttcccattatgctt gtataattgtgaccagggaaacggatcaggca acttcgatgagtgtaaaacgaaatgctatgaa atggccggaaggtga	ARETQQECVRHCGGH NEYVTRYCGGLCSGST GPQTFYCYLGCSHNAS NQNDFDKCLPKCNGS PQLTESSCQNDCGRVT THPELCGIVCGGNVGD SFPLCLYNCDQGNGSG NFDECKTKCYEMAGR	pMT_BiP_SL IN	Kpn I Xho I	Forward: 5'-GGTACCAGCAAG AGAAACCCAACAGGA AT-3' Reverse: 5'-ATACTTTACCGGCC TTCCACTGAGCTC-3'

MEG 3.2 isoform 1 "short V2" (deletion of 16 aa from N-terminus) S2 *Drosophila* cells	tgtgtattcaccgcaagagaaacccaacagga atgtgtcagacattgcggtggtcataatgaata cgtcacacgttactgtggaggactttgttcaggc tcaacgggaccacagacattctactgttactta ggttgctctcataatgcttccaaccaaaatgact ttgacaagtgccttccaaaatgcaacggctccc ctcagcttacggaatcctcatgccaaaacgatt gtggaagggtcactacccatccagaactttgcg gcatcgtgtgtggtggtaatgtgggtgactcctt cccattatgcttgtataattgtgaccagggaaa cggatcaggcaacttcgatgagtgtaaaacga aatgctatgaaatggccggaaggtga	CVFTARETQQECVRHC GGHNEYVTRYCGGLCS GSTGPQTFYCYLGCSH NASNQNDFDKCLPKC NGSPQLTESSCQNDC GRVTTHPELCGIVCGG NVGDSFPLCLYNCDQ GNGSGNFDECKTKCYE MAGR	pMT_BiP_SL IN	Kpn I Xho I	Forward: 5'-GGTACCATGTGTA TTCACCGCAAGAGAA AC-3' Reverse: 5'-ATACTTTACCGGCC TTCCACTGAGCTC-3'

Table 3.3 - Overview of the cloning design for MEG 6 that have been used to express this protein.

Name & expression system	DNA sequence	Protein sequence	Plasmid	Restriction enzymes	Primers
MEG 6 *E. coli*	atggttcaaaatccgaaaaacaccaaaaaaat caatcgcactatccgccgtagcactaaaacagt gattgtcatcacagaccgcgtccaaaacatcgt gctgggccaccgcttgttacaccatcgtattccc acgattaaacgctccaaaagccacggcatcaa taaaaacgaaaccgtgtctaacttatttccata g	MVQNPKNTKKINRTIR RSTKTVIVITDRVQNIV LGHRLLHHRIPTIKRSK SHGINKNETV SNLFP	pET-22(b)+	Xho I Nco I	Forward: 5'-CAGCCGGCGATG GCCATGGTTCAAAAT CCGAAAAACACCA-3' Reverse: 5'-TTTGGCACAGATT GAATAAAGGTATCGA GCTCGTGGTGGTGGT -3'
MEG 6 *Komagataella phaffii*	atggttcaaaatccgaaaaacaccaaaaaaat caatcgcactatccgccgtagcactaaaacagt gattgtcatcacagaccgcgtccaaaacatcgt gctgggccaccgcttgttacaccatcgtattccc acgattaaacgctccaaaagccacggcatcaa taaaaacgaaaccgtgtctaacttatttccaga aaacctgtacttccaagcgttcgagcaccacca ccaccaccactga	MVQNPKNTKKINRTIR RSTKTVIVITDRVQNIV LGHRLLHHRIPTIKRSK SHGINKNETVSNLFPE NLYFQAFEHHHHHH-	pPICZα B	Pst I	Forward: 5'AAGAGAGGCTGAA GCTATGGTTCAAAAT CCGAAAAACACCA-3' Reverse: 5'-GAATAAAGGTCTT TTGGACATGAAGGT TAGCCCAAGATCTCT TGTTTTTGAGTAG-3'

3.2.1 Expression in Yeast

This was done by In Fusion cloning system using synthetic DNA (produced and codon-optimized for *Komagataella phaffii* – sequences in Tables 3.1, 3.2, and 3.3), specific primers (Tables 3.1, 3.2, and 3.3), restriction enzyme PstI (NEB) and *Pichia pastoris* pPICZαB cloning vector for secretory methanol induced expression (EasySelect™ Pichia Expression Kit For Expression of Recombinant Proteins Using pPICZ and pPICZα (ThermoFisher) in *Komagataella phaffii*/*Pichia pastoris*).

3.2.2 Expression in bacteria

At the same time, protein cloning into the bacterial *Escherichia coli* expression system was performed. Therefore, the same synthetic DNA fragments for MEG 2.1, MEG 3.2, and MEG 6 (Tables 3.1, 3.2, 3.3) were used for cloning into the bacterial expression system. In the framework of *E. coli* expression, two vectors were used: pET 22b(+) (ThermoFisher), which allows periplasmic expression and polyhistidine-tag (6xHisTag) for subsequent protein

purification *via* affinity chromatography; and Champion™ pET SUMO Expression vector (Thermo Fisher Scientific), which helps to increase protein solubility thanks to fusion with SUMO protein and includes a 6xHisTag for subsequent protein purification *via* affinity chromatography. All three synthetic DNA of MEG 2.1, MEG 3.2 and MEG 6 proteins were cloned into the pET 22b(+) expression vector (Tables 3.1 - 3.3).

Only MEG 2.1 and MEG 3.2 coding sequences were cloned into Champion pET SUMO. Expression of the above-described proteins (three different proteins, two different vectors) was first performed in the commonly used *E. coli* expression strain BL21-Gold(DE3) Competent Cells (Novagen®).

Subsequently, another *E. coli* strain was also tested, Rosetta(DE3) Competent Cells (Novagen®). This bacterial strain was chosen because of the presence of rare codons for *E. coli* in both sequences of MEG 2.1 and MEG 3.2. In fact, all of ordered synthetic DNA had been optimized for the yeast expression system.

The next tested bacterial expression strain was One Shot™ BL21(DE3)pLysS Chemically Competent *E. coli* (Thermo Fisher Scientific). This system is suitable for T7 promoter expression and contains the pLysS plasmid, which expresses low levels of T7 lysozyme, supposed to help to block the basal leaking expression of recombinant proteins by inhibiting the T7 RNA polymerase. This strain is recommended in the literature for the expression of potentially toxic proteins.

The last bacterial strain that was tested was a strain that combines the two previously mentioned ones and which is suitable for expression of toxic proteins: Rosetta(DE3)pLysS Competent Cells (Novagen®). The Rosetta(DE3)pLysS bacteria used were not the commercial stock, but homemade chemically prepared, in order to reduce bacteriophage contamination. First, the cloned and sequenced constructs (~100 - 200 ng of plasmid) were transformed into the selected bacterial strain, and the mixture was left to incubate on ice for 30 minutes. Heat shock was then performed at 42 °C for 45 seconds. Next, the plasmid and cell mixture were again left on ice for 5 minutes. At the end of incubation, 200 - 250 µL of SOC (Super Optimal broth with Catabolite repression - 2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose) or LB (Luria-Bertani Broth - 5 g/L NaCl, 10 g/L Tryptone, 5 g/L Yeast Extract) media was added and left at 37 °C in a shaking incubator for 1.5 hours at 150 rpm, after which the bacteria were plated on LB agar, containing the antibiotic for the selection of positive clones. In our case ampicillin (Waterhouse et al.) at a final concentration of 100 µg/mL was used. Plates were allowed to incubate overnight at 37 °C. Subsequently, one colony was selected and used to create a pre-culture. The volume of the pre-culture was determined depending on the resulting volume of the expression medium; the dilution coefficient was 40 (i.e., for 1 L of the medium, a pre-culture of 25 mL was prepared). This pre-culture was prepared by overnight growth in the shaking incubator of liquid LB/Amp medium at 37 °C, spinning at 150 rpm. In the morning, the pre-culture was poured into tempered LB/Amp/glucose media (final concentration of glucose 20 mM) of the desired volume and let grow at 37 °C, 150 rpm shaking. Optical density (OD_{600}) of the bacterial growth was monitored and measured at regular intervals. After reaching an OD_{600} of 0.6 - 0.8,

the bacteria were allowed to settle at 3000 g for 10 minutes to remove glucose from the medium. Subsequently, the bacteria were gently resuspended again in a warmed LB/Amp medium, to avoid temperature shock, and protein production was induced by addition of 0.1 - 1 mM IPTG (isopropyl-β-thio-galactopyranoside). The culture was then left at 37 °C for either 3 hours and harvested or allowed to grow overnight and harvested in the morning.
Each large-scale expression was always preceded by a pilot expression. The first expression assay was according to "standard protocol" (induction of 0.1 - 1 mM IPTG at OD_{600} 0.6/0.8, followed by harvesting of bacteria 3 hours after induction or harvested after overnight culture), followed by optimization. Depending on the results obtained, the optimization of the expression process followed. All tested expression conditions are in Tables 3.4, 3.5, 3.6, and 3.7, which include different IPTG concentrations, different OD_{600} of the culture before IPTG induction, different induction times, and different growth temperatures.

Table 3.4 - Tested conditions of expression for MEG 2.1 isoform 1.

Protein name	Prot Param	Plasmid	cloning cell type	expression cell type	induction concentration	temperature of induction	expression levels	purification/stability	notes
MEG 2.1 isoform 1 - long version	MW = 11.96 kDa pI = 6.01 Ɛ(280) = 6460 (M-1cm-1) Ɛ(280) = 0.540 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	BL21 (DE3)	0.1 mM IPTG, 0.5 mM IPTG; 1 mM IPTG	37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 2.1 isoform 1 - long version	MW = 11.96 kDa pI = 6.01 Ɛ(280) = 6460 (M-1cm-1) Ɛ(280) = 0.540 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	Rosetta(DE3)	0.5 mM IPTG; 1 mM IPTG	37 °C O/N; 30 °C O/N, 18 °C O/N	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 2.1 isoform 1 - long version	MW = 11.96 kDa pI = 6.01 Ɛ(280) = 6460 (M-1cm-1) Ɛ(280) = 0.540 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	BL21(DE3)pLysS	0.1 mM IPTG, 0.5 mM IPTG; 1 mM IPTG	37 °C O/N; 30 °C O/N, 18 °C O/N	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 2.1 isoform 1 - long version	MW = 17.94 kDa pI = 5.81 Ɛ(280) = 110430 (M-1cm-1) Ɛ(280) = 0.637 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	Rosett(DE3)pLysS	no expression under the special rifampicin protocol (methodology)	no expression under the special rifampicin protocol (methodology)	X	X	
MEG 2.1 isoform 1 - long version	MW = 21.036 kDa pI = 5.08 Ɛ(280) = 6460 (M-1cm-1) Ɛ(280) = 0.307 (L g-1cm-1)	pET SUMO Champion	Stellar Competent cells	BL21(DE3)	0.5 mM IPTG; 1 mM IPTG	37 °C O/N; 30 °C O/N, 18 °C O/N	no expression in the soluble fraction / no expression in the inclusions	X	hard to grow
MEG 2.1 isoform 1 - long version	MW = 21.036 kDa pI = 5.08 Ɛ(280) = 6460 (M-1cm-1) Ɛ(280) = 0.307 (L g-1cm-1)	pET SUMO Champion	Stellar Competent cells	Rosetta(DE3)	0.5 mM IPTG; 1 mM IPTG	37 °C O/N; 30 °C O/N, 18 °C O/N	no expression in the soluble fraction / no expression in the inclusions	X	almost impossible to grow
MEG 2.1 isoform 1 - long version	MW = 21.036 kDa pI = 5.08 Ɛ(280) = 6460 (M-1cm-1) Ɛ(280) = 0.307 (L g-1cm-1)	pET SUMO Champion	Stellar Competent cells	BL21(DE3)pLysS	0.5 mM IPTG; 1 mM IPTG	37 °C O/N; 30 °C O/N, 18 °C O/N	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 2.1 isoform 1 - long version	MW =9.79 kDa pI = 5.53 Ɛ(280) = 4970 (M-1cm-1) Ɛ(280) = 0.456 (L g-1cm-1)	pPIC zALPHA B; PstI	Stellar Competent cells	*Pichia pastoris*	MetOH induced expression (according to the protocol)	30°C according to the protocol	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 2.1 isoform 1 - long version	MW = 9.79 kDa pI = 5.53 Ɛ(280) = 4970 (M-1cm-1) Ɛ(280) = 0.508 (L g-1cm-1)	pIVEX2.4d	BL21(DE3)	cell-free	/	/	no transformation into BL21 (for midiprep isolation - for Grenoble)	X	"long" with SP; synthesis into pIVEX2.4d; not send to Grenoble - no transformation
MEG 2.1 isoform 1 - short version	MW = 7.2 kDa pI = 4.9 Ɛ(280) = 4845 (M-1cm-1) Ɛ(280) = 0.671 (L g-1cm-1)	pIVEX2.4d	BL21(DE3)	cell-free	/	/	no expression	X	"short" without SP; synthesis into pIVEX2.4d; done in Grenoble
MEG 2.1 isoform 1 - short version	MW = 7.2 kDa pI = 4.9 Ɛ(280) = 4845 (M-1cm-1) Ɛ(280) = 0.671 (L g-1cm-1)	pMT_BiP_SLIN	Stellar Competent cells	S2 Drosophila cells	see Methodology	see Methodology	expression	StrepTactin XT	

Table 3.5 - Tested conditions of expression for MEG 3.2 isoform 1 - long version.

Protein Name	Prot Param	Plasmid	cloning cell type	expression cell type	induction concentration	temperature of induction	expression levels	purification/stability	notes
MEG 3.2 isoform 1 - long version	MW = 17.94 kDa pI = 5.81 Ɛ(280) = 110430 (M-1cm-1) Ɛ(280) = 0.637 (L g-1cm-1)	pET 22(b)+	Stellar Competent cells	BL21 (DE3)	0.1 mM; 0.5 mM IPTG; 1 mM IPTG	37°C , 37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 3.2 isoform 1 - long version	MW = 17.94 kDa pI = 5.81 Ɛ(280) = 110430 (M-1cm-1) Ɛ(280) = 0.637 (L g-1cm-1)	pET 22(b)+	Stellar Competent cells	Rosetta(DE3)	0.5 mM IPTG; 1 mM IPTG	37°C , 37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction / no expression in the inclusions	X	
MEG 3.2 isoform 1 - long version	MW = 17.94 kDa pI = 5.81 Ɛ(280) = 110430 (M-1cm-1) Ɛ(280) = 0.637 (L g-1cm-1)	pET 22(b)+	Stellar Competent cells	BL21(DE3)pLysS	0.1 mM; 0.5 mM IPTG; 1 mM IPTG	37°C , 37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction	X	
MEG 3.2 isoform 1 - long version	MW = 17.94 kDa pI = 5.81 Ɛ(280) = 110430 (M-1cm-1) Ɛ(280) = 0.637 (L g-1cm-1)	pET 22(b)+	Stellar Competent cells	Rosetta(DE3)pLysS	expression under the special rifampicin protocol (methodology)	expression under the special rifampicin protocol (methodology)	low expression	stable only with at least 0.5 M NaCl	
MEG 3.2 isoform 1 - long version	MW = 27.01 kDa pI = 5.10 Ɛ(280) = 11430 (M-1cm-1) Ɛ(280) = 0.423 (L g-1cm-1)	pET SUMO Champion	Stellar Competent cells	BL21 (DE3)	0.5 mM IPTG; 1 mM IPTG	37°C , 37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction	X	hard to grow
MEG 3.2 isoform 1 - long version	MW = 27.01 kDa pI = 5.10 Ɛ(280) = 11430 (M-1cm-1) Ɛ(280) = 0.423 (L g-1cm-1)	pET SUMO Champion	Stellar Competent cells	Rosetta(DE3)	0.5 mM IPTG; 1 mM IPTG	37°C , 37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction	X	almost impossible to grow
MEG 3.2 isoform 1 - long version	MW = 27.01 kDa pI = 5.10 Ɛ(280) = 11430 (M-1cm-1) Ɛ(280) = 0.423 (L g-1cm-1)	pET SUMO Champion	Stellar Competent cells	BL21(DE3)pLysS	0.5 mM IPTG; 1 mM IPTG	37°C , 37°C O/N; 30°C O/N, 18°C O/N	no expression in the soluble fraction	X	
MEG 3.2 isoform 1 - long version	MW = 15.77 kDa pI = 5.39 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.630 (L g-1cm-1)	pPICZ alpha B	Stellar Competent cells	*Pichia pastoris*	MetOH induced expression (according to the protocol)	30°C according to the protocol	no expression	X	codon optimized for *Pichia; protocol EasySelect Pichia Expression Kit*
MEG 3.2 isoform 1 - long version	MW = 15.77 kDa pI = 5.39 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.630 (L g-1cm-1)	pIVEX2.4d	BL21(3D)	cell-free	x	x	no expression	X	done in Grenoble

Table 3.6 - Tested conditions of expression for MEG 3.2 isoform 1 - short version.

Protein Name	Prot Param	Plasmid	cloning cell type	expression cell type	induction concentration	temperature of induction	expression levels	purification/stability	notes
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	BL21 (DE3)	x	no transformation at 37°C	x	X	no transformation in BL21 (in comparison with MEG6 and MEG 3.2 L)
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	BL21(DE3)pLysS	x	no transformation at 37°C	x	X	no transformation in BL21 (in comparison with MEG6 and MEG 3.2 L)
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	Rosetta(DE3)pLysS	x	no transformation at 37°C	x	X	no transformation in BL21 (in comparison with MEG6) - check for transformation; + DNA degradation (-20°C storage)
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pIVEX2.4d	BL21(3D)	cell-free	/	/	no expression	X	without SP, done in Grenoble
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pMT_BiP_SLIN	Stellar Competent cells	S2 Drosophila cells	see Methodology	see Methodology	no expression	X	MEG3.2_short_V1 (- 16 AA)
MEG 3.2 isoform 1 - short version	MW = 13.95 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.712 (L g-1cm-1)	pMT_BiP_SLIN	Stellar Competent cells	S2 Drosophila cells	see Methodology	see Methodology	no expression	X	MEG3.2_short_V2 (-20 AA)

Table 3.7 - Tested conditions of expression for MEG 6.

Protein Name	Prot Param	Plasmid	cloning cell type	expression cell type	induction concentration	temperature of induction	expression levels	purification/stability	notes
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	BL21 (DE3)	x	no transformation at 37°C	x	X	no transformation in BL21 (in comparison with MEG6 and MEG 3.2 L)
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	BL21(DE3)pLysS	x	no transformation at 37°C	x	X	no transformation in BL21 (in comparison with MEG6 and MEG 3.2 L)
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pET 22b(+)	Stellar Competent cells	Rosetta(DE3)pLysS	x	no transformation at 37°C	x	X	no transformation in BL21 (in comparison with MEG6) - check for transformation; + DNA degradation (-20°C storage)
MEG 3.2 isoform 1 - short version	MW = 13.5 kDa pI = 5.44 Ɛ(280) = 9940 (M-1cm-1) Ɛ(280) = 0.736 (L g-1cm-1)	pIVEX2.4d	BL21(3D)	cell-free	/	/	no expression	X	without SP, done in Grenoble

Only the strain Rosetta(DE3)pLysS followed a special protocol. This protocol was specific in that recovery in LB/Amp (100 µg/mL) after transformation of 100 ng of plasmid pET 22(b)+ with MEG 3.2 “long” version (with SP) into bacteria was carried out at 25 °C for two hours. The bacteria were then plated on LB/agar/Amp (100 µg/mL) plates, and the recovery was left at 25 °C for two days. Bacterial pre-culture was inoculated from multiple colonies (bacteria scraping) into 25 mL of LB/Amp (100 µg/mL). This pre-culture was incubated at 37 °C for 4 hours and then transferred to 1 L of LB/Amp (100 µg/mL)/20 mM glucose, which was tempered to 37 °C prior to inoculation. After reaching an OD_{600} of 0.6, the culture was induced with 100 mM IPTG (final concentration) for one hour. Rifampicin (100 µg/ml) was then added to the culture, and after two hours the bacteria were harvested. The pellet was frozen overnight at -80 °C.

This was the only functional protocol for the expression of MEG 3.2 isoform 1. In fact the addition of rifampicin, which is the most effective antibiotic inhibiting the transcription of bacterial RNA polymerase (Lama and Carrasco 1992; Du et al. 2021) will block the translation of all the bacteria proteins except the protein of interest. Addition of rifampicin after IPTG induction also simplifies protein labeling for subsequent NMR structural studies. The addition of rifampicin results in selective labeling of only the target heterologous protein (Almeida et al. 2001). This protocol was adapted from Dr. Francesca Fiorini's protocol at IBCP (Lyon) under her supervision. For this protocol it is essential to use bacteria that have already been multiplied once from a commercial stock, in order to reduce the potential occurrence of bacterial phages.

3.2.3 Isotopic labeling of MEG 3.2 protein in BL21(DE3) bacteria

In the morning, several transformed colonies were transferred into 10 mL of LB media with 10 ml of the ampicillin stock (100 mg/mL). For the rest of the day this culture was left in the incubator at 37 °C under shaking at 150 rpm. In the evening, 2 mL of this culture was added to a pre-culture of minimal medium for isotopic labeling. The pre-culture of 200 mL was prepared with ddH_2O, 1.5 g Na_2HPO_4, 0.6 g KH_2PO_4, 0.1 g NaCl and 0.2 g $^{15}NH_4Cl$ which had been previously sterilized by autoclaving.

At the same time, the solution of the trace metals was prepared, following the recipe: 5 g EDTA dissolved in 700 mL ddH_2O. Separately 0.5 g $FeCl_3 \cdot 6H_20$, 0.005 g ZnO, 0.001 g $CuCl_2 \cdot 2H_2O$, 0.001 g $Co(NO_3)_2 \cdot 6H_2O$ and 0.001 g $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$ was dissolved in the smallest possible volume of 5 M HCl. All the prepared solutions were then added to the 700 mL EDTA solution and topped up with ddH_2O to a volume of 1 L; pH of this solution was adjusted to 7. This stock solution is kept at 4 °C in a dark bottle.
After cooling of the minimal medium after autoclaving, sterile filtered solutions were added: 200 μL of the ampicillin stock (100 mg/mL), 200 μL of $MgSO_4$ stock (1 M), 200 μL of $CaCl_2$ stock (1 M), 200 μL of the thiamine stock (1 M), 2 g of glucose, 2 mL of prepared trace metals solution. To this prepared medium, 2 mL of culture in LB/Amp was added after an all-day growth. This culture was left in the incubator at 37 °C, shaking at 150 rpm overnight. Along with the preparation of 200 mL of minimal medium for the pre-culture, 2 L of minimal medium were prepared for the labelling itself. This medium contained 15 g Na_2HPO_4, 6 g KH_2PO_4, 1 g NaCl and 2 g $^{15}NH_4Cl$ in water and was sterilized by autoclaving. Prior to use, this minimal medium was supplemented with 2 mL of the ampicillin stock (100 mg/mL), 2 mL of $MgSO_4$ stock (1 M), 2 mL of $CaCl_2$ stock (1 M), 2 mL of the thiamine stock (1 M), 20 g of glucose, 20 mL of prepared trace metals solution. The amount of overnight preculture in the minimal media (200 mL) was added to these 2 L of complete minimal media so that the resulting measured OD600 was 0.1. This culture was left in the incubator at 37 °C, shaking at 150 rpm, until the OD600 reached 0.6. The bacteria were then induced by adding 2 mL of sterile IPTG stock solution (1 M). This culture was left in the incubator at 37 °C, shaking at 150 rpm for one hour. Rifampicin (100 μg/ml) was then added to the culture, and after two hours, the bacteria were harvested by centrifugation at 3,000x g for 15 min at room temperature and finally the pellet was frozen overnight at -80 °C.

3.2.4 Expression in cell-free system

Another expression system that was tested for MEG 3.2 (with signal peptide) and MEG 2.1 (without/with signal peptide) was cell-free protein expression. Pilot expressions were performed at the Institut de Biologie Structurale in Grenoble (CNRS). For this expression, cloning of MEG 3.2 and MEG 2.1 (both in variants without and with signal peptide) into the cell-free vector pIVEX2.4d (plasmid and inserts design specified in Tables 3.1 and 3.2) was performed.

3.2.5 Expression in insect cells

The next tested expression system was an insect system – the Schneider-2 Drosophila cell line. The expression vector used for S2 expression was pMT/BiP/SLIN (Dr. Barinka's lab, BIOCEV, Vestec) (Tables 3.1 and 3.2) with two StrepII tag sequences for further purification using affinity chromatography, BiP signal peptide, SLIN tag (StrepII - FLAG - TEV site - StrepII - TEV site). This system was chosen because the plasmid contains the BiP signal peptide, which secretes proteins from the S2 host cell into the medium and is cleaved off during secretion; secretion of proteins from cells into the media is another possible strategy for the expression of toxic proteins. Only "short versions", without signal peptide, of MEG 2.1 and MEG 3.2

proteins were cloned into the expression vector. For MEG 3.2, two versions (with 16 amino acid deletion and with 20 amino acid deletion) were created based on signal peptide prediction (sequence and cloning design in Table 3.2). Stable S2 transfectants were transferred to a sterile 500 mL Erlenmeyer flask into 150 mL of serum-free SF-900 II (Sigma) and the growth continued at 27 °C with shaking 120-130 rpm to the final density, approximately 10 x 10^6/mL. Cells were then split into two 2L sterile Erlenmeyer flasks (each with 0.7 L of serum-free SF-900 II media) and grown to the final concentration of 1 x 10^6/mL and afterwards incubated overnight with 0.7 mM $CuSO_4$ which triggered their over-expression. Cell growth was monitored daily (for 5 - 7 days). Medium was harvested upon reaching plateau of the cell growth (typically 30 x 10^6/mL).

3.3 Protein purification

3.3.1 Fast protein liquid chromatography (FPLC) - bacterial expression

Harvesting of bacteria after three hours or overnight after IPTG induction was performed by centrifugation at 3,000x g for 15 min at room temperature. The pellet was then resuspended in lysis buffer, using a ratio of 10 ml buffer per 1 g of cell pellet. Lysis buffer is composed by 50 mM Tris/HCl pH 8; 10 mM imidazole; 1 M NaCl; 5 mM β-Mercaptoethanol (BME); 10% glycerol; cOmplete™ Protease Inhibitor Cocktail, EDTA-free (Roche). After dissolving the bacterial pellet in the lysis buffer, sonication on ice in a metal beaker was performed, using 2x1 minute (with an amplitude of 70) by measuring the temperature of the sample and allowing it to cool sufficiently. After sonication, centrifugation was performed at 15,000x g for 30 minutes at 4 °C. The supernatant was used to analyze the soluble fraction, and the pellet was further used to analyze the insoluble fraction on a denaturing gel electrophoresis (SDS-PAGE).

FPLC purification system (ÄKTA go protein purification system from Cytiva) was used to achieve the best possible separation of proteins in the soluble fraction.

The supernatant was loaded onto a 1 ml or 5 ml (depending on the volume of the culture) Cytiva HisTrap™ FF Column equilibrated with loading/wash buffer (50 mM Tris/HCl pH 8, 10 mM imidazole; 1 M NaCl; 5 mM β-Mercaptoethanol; 10% glycerol). The supernatant was circulated twice on the column, before applying a stepped gradient (5% B, 10% B, 20% B, 50% B, 100% B) for elution, while collecting 1 ml fractions. The elution/ buffer B contained 50 mM Tris/HCl pH 8; 500 mM imidazole; 500 mM NaCl; 10 % glycerol; 5 mM β-Mercaptoethanol. The column loading and fractionation flow rate was 1 mL/min, and the wash flow rate before and after separation was 2 mL/min.

3.3.2 Size-exclusion chromatography (SEC) - bacterial expression

After that the individual fractions following affinity purification were analyzed by SDS-PAGE, all those containing the MEG protein of interest were pooled to be further purified using a size exclusion chromatography column on the Akta Pure system. Two columns were used (depending on the amount of purified protein) – HiLoad® 16/600 Superdex® 75 pg Cytiva

column 28-9893-33, L × ID 60 cm × 16 mm, 34 µm avg. part. size (120 mL) and Superdex 75 Increase 5/150 GL5 column 5x 153-158 mm, 9 µm avg. part. size (3 mL). The buffer used was identical to the elution buffer from FPLC – 50 mM Tris/HCl pH 8; 500 mM imidazole; 500 mM NaCl; 10 % glycerol; 5 mM β-Mercaptoethanol. If necessary, concentrations were performed on Amicon Ultra centrifugal filter units Ultra-15, MWCO 10 kDa, before injecting the desired volume via a loop into the system (500 µL loop for 3 mL column, 5 mL loop for 120 mL column). The separation was carried out within the pressure and flow rate recommended by the manufacturer.

3.3.3 Purification of the S2 expressions - gravity-flow affinity chromatography and size exclusion chromatography

The medium was harvested by centrifugation at 500× g followed by centrifugation of the supernatant at 10,000× g. Thaw conditioned media was filtered through 0.22 µm filter. The supernatant was completed with protease inhibitors (Roche) and concentrated from 700 mL to 45 mL using concentration/dialysis by tangential flow filtration TFF (Millipore Mosheim France). The concentrate was then dialyzed again with TFF into the equilibration buffer, 100 mM Tris-HCl pH 8.0, 300 mM NaCl. MEG 2.1 protein was purified by affinity chromatography using a StrepTactin XT resin (Coulibaly et al.). The column was equilibrated with 5 column volumes with loading buffer (100 mM Tris-HCl pH 8.0, 150 mM NaCl). Filtered and dialyzed media were loaded to the column (1 mL/min) and the flow-through fraction was collected. Column was washed with 10 column volumes of the loading buffer (100 mM Tris-HCl pH 8.0, 150 mM NaCl) and the flow-through fraction was collected again. Protein was eluted with 10 column volumes with elution buffer (100 mM Tris-HCl pH 8.0, 150 mM NaCl, 5 mM d-Desthiobiotin - Sigma cat. no. D 1411). Pooled elution fractions were subjected to size exclusion chromatography using a Superdex Hiload16/600 75 pg Cytiva column. Fractions containing MEG 2.1 isoform 1 protein were pooled, concentrated, flash frozen in liquid nitrogen and stored at −80 °C.

3.4 Biophysical analysis

The samples used for biophysical analyses were of two types: recombinant MEG 2.1 isoform 1 expressed in insect cells and MEG 3.2 isoform 1 expressed in E. *coli*. All protein samples were at very low concentrations for subsequent NMR determination of the tertiary structure (the highest concentration achieved was 109 µM for ^{15}N labeled MEG 3.2 isoform 1). Proteins were dissolved in buffers suitable for protein stabilization, most frequently 20 mM Tris/HCl at pH 8 for MEG 3.2 isoform 1 or 50 mM sodium phosphate buffer at pH 8 for MEG 2.1 isoform 1.

The second type of sample were lyophilized synthetic peptides. These peptides were dissolved to a final concentration of 2 mM in deuterated dimethyl-sulfoxide (DMSO-d6) for NMR analysis in natural abundance of ^{13}C and ^{15}N isotopes. At the same time, these peptides were partially dissolved in 100% acetonitrile, and 50 % acetonitrile plus 50 % trifluoroethanol, at concentrations of about 10 µM for secondary structure analyses using CD.

Biophysical analyses were performed to determine the secondary and tertiary structures of peptides and proteins.

3.4.1 Circular Dichroism

Circular dichroism (CD) analysis was used to determine the secondary structure. Samples (purified proteins and synthetic peptides) were measured in various buffers: Tris/HCl up to 10 mM pH 8, MES up to 10 mM pH 6, PIPES up to 10 mM pH 6 for two recombinant proteins (MEG 3.2 and MEG 2.1 isoform 1 in both cases); 100 % acetonitrile and 50 % acetonitrile/50 % TFE for the synthetic peptides. As the peptides exhibited very low solubility in biological buffers (Tris/HCl, phosphate or MES buffer) or organic solvents (chloroform, acetone, methanol, isopropanol), acetonitrile was used, in which the peptides were partially soluble. Acetonitrile proved to be a suitable solvent for CD measurements as it does not absorb between 180 and 250 nm, where the features essential for protein/peptide secondary structure determination lie. Conversely, DMSO is not suitable for CD analyses due to its high absorbance in the above-mentioned ranges of wavelengths. The lyophilized peptides were resuspended to a nominal concentration of 10 µM, then the samples were centrifuged, and the recovered supernatant was measured. Samples were measured in a HELLMA Macro-Cuvette 100-QS, 1mm Quarz Glass 100-1-40 in the range from 180 nm to 280 nm with a step size of 0.5 nm, a bandwidth of 1 nm and in five repetitions for the entire spectrum. The Chirascan VX instrument from AppliedPhotophysics was used for all the measurements. The spectra presented in results are the average of the 5 measures after subtraction of the average of 5 baselines.

3.4.2 Dynamic Light Scattering

Dynamic light scattering (DLS) was used to determine the size distribution of proteins in the solution. Purified protein MEG 3.2 was measured in a 50 mM Tris/HCl pH 8, 200 mM NaCl, 10 % glycerol and 5 mM BME buffer. The measurement was carried out at a temperature of 25 °C and was performed in three repetitions. The analyses were performed using the Zetasizer Nano ZS instrument (MalvernPanlytical), and the data were evaluated with Malvern Instruments software Zetasizer Ver. 6.34.

3.4.3 Nuclear Magnetic Resonance Spectroscopy

The synthetic peptides of MEG 2.1 (isoform 1, 1a, 1b, 1c, 1d, 1f, 1g, 2a, 2b and isoform 3) were dissolved in a volume of 450 µL of deuterated dimethyl sulfoxide (DMSO-d6) (Eurisotop) to a concentration of 2 mM. The prepared sample was then transferred to a 5 mm NMR sample tube and centrifuged. The temperature measurement was 27 °C. All of the peptides were measured at the Varian Inova spectrometer operated at a ^{1}H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C; isoforms 1 and 3 were measured also at the Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) and equipped with a triple HCN cryoprobe (Infranalytics Lille).

For proton resonance assignments we used homonuclear experiments as zero quantum TOCSY (zTOCSY) experiment with 16 scans and mixing time of 80 ms and 100 ms; NOESY experiment (Thrippleton and Keeler 2003) with 28 scans and a mixing time of 400 ms. For ^{15}N and ^{13}C resonance assignments we used heteronuclear experiments in natural abundance: ^{1}H-^{15}N HSQC experiment (Kay, Keifer, and Saarinen 1992) with 600 scans and ^{1}H-^{13}C HSQC experiment with 256 scans, with and without ^{1}H-^{13}C multiplicity editing. To complete ^{1}H and ^{13}C assignments, we also acquired one ^{1}H-^{13}C HSQC-TOCSY experiment (adapted from (Becker et al. 2019) (Kövér, Hruby, and Uhrín 1997)) with 400 scans and a mixing time of 80 ms. Except for the ^{1}H-^{15}N HSQC experiment, we have used a double pre-saturation for each 2D experiments at 2.48 ppm and 3.33 ppm to suppress signals of DMSO and H_2O respectively. To monitor the peptide stability, between each 2D experiment, we have inserted a standard 1D ^{1}H experiment with the same double pre-saturation for DMSO and H_2O signals. All these experiments were recorded for the different peptides with the Varian spectrometer.

The longest peptide of isoform 1 (without signal peptide) and isoform 3, were dissolved in 200 µL DMSO-d6 (Eurisotop) to a concentration of 2 mM. The experiments were recorded with the 1.2 GHz Bruker spectrometer in Lille (Infranalytics). The following experiments were recorded: ^{1}H-^{1}H TOCSY (Cavanagh and Rance 1990) (mixing time = 60 ms, 16 scans), ^{1}H-^{1}H NOESY (mixing time = 120 ms, 48 scans), ^{1}H-^{13}C edited HSQC (Boyer, Johnson, and Krishnamurthy 2003) (8 scans) and sofast ^{1}H-^{15}N HSQC (Schanda and Brutscher 2005) (256 scans).

2D data were processed using nmrpipe and nmrdraw processing software (Delaglio et al. 1995). Peak assignment and visualization were performed using Poky (Lee et al. 2021) and Sparky software (Lee, Tonelli, and Markley 2015).

3.4.4 Structure refinement and molecular docking

CYANA software (Güntert and Buchner 2015) version 2.1 was used for automatic structure calculations based on assigned spectra (^{1}H-^{1}H NOESY, ^{1}H-^{13}C HSQC and ^{1}H-^{15}N HSQC). The algorithm used on CYANA program is based on a consistent probabilistic treatment of the NOE assignment process. The goal of the procedure is to reduce ambiguities of NOE assignments that could lead to erroneous distance restraints. The structure calculation is an iterative process over 7 cycles. The same input data are used for all the cycles and comprise the assigned chemical shift list, the amino acid sequence of the peptides and a list containing the positions and the volume of the NOE cross peaks measured in 2D ^{1}H-^{1}H NOESY spectrum. Finally, the unambiguous distance restraints are included as input of the structure calculation with simulated annealing by the fast CYANA algorithm developed by Güntert et al. (Güntert, Mumenthaler, and Wüthrich 1997). The accuracy of the calculated structures improves with each subsequent cycle.

The peptide structures (1a, 1b, 1f, 1g, 2a and 2b) obtained after CYANA treatment were built together using UCSF Chimera software (Pettersen et al. 2004) (with a default setting Join Models section for C-N peptide bond). In order to reconstruct the complete structure of MEG 2.1 isoform 1 and isoform 2 (with the signal peptide), the AlphaFold2 (Jumper et al. 2021)

prediction of MEG 2.1 isoform 3 (signal peptide) was used, which was also linked in UCSF Chimera Build Structure with the already linked measured peptides. For the MEG 2.1 isoform 1 protein, it was necessary to truncate the predicted AlphaFold2 (Jumper et al. 2021) model of isoform 3 and insert a 4-amino acid peptide (-FSHC-) between it and the 1a peptide. For MEG 2.1 isoform 2 this modification was not necessary, here only the AlphaFold2 model of the signal peptide (MEG 2.1 isoform 3) was combined with the assembled structure (2a and 2b peptide).

Such reconstructed models were subjected to energy minimization using the Chiron protein structure refinement server (Ramachandran et al. 2011) that minimizes steric clashes in proteins using short discrete molecular dynamics (DMD) simulations until it attains an 'acceptable clash score' (the resulting protein structure has normalized clash score that is comparable to high-resolution protein structures (<2.5 Å)).

The minimized structure of the complete MEG 2.1 isoform 1 protein was screened using DeepSite neural-network software (Jiménez et al. 2017) *via* the PlayMolecule web application (Skalic et al. 2019) (Martínez-Rosell, Giorgino, and De Fabritiis 2017) in order to find possible binding pockets. Consistent with the NMR measured and analyzed data, a relevant binding pocket at the C-terminus of MEG 2.1 isoform 1 was identified. This pocket was further used for blind screen molecular docking.

Molecular docking was performed using the AutoDock Vina software (Trott and Olson 2010) (Eberhardt et al. 2021). The Open Babel translation program (O'Boyle et al. 2011) was used to convert all formats necessary for the docking process *via* AutodockVina. MGLTools and AutoDock tools (Morris et al. 2009) were also used to prepare the binding pocket and ligands. ZINC20 database (Irwin et al. 2020) of commercially available compounds was used for the virtual screening. For the purpose of the blind virtual screening of the potential ligands, the Tranches database layout was used for subsets download. Thousands of molecules were screened in this distribution with consideration of the nature of the predicted binding pocket. The 7942 selected molecules were from the subset of: Representation - 3D molecules, Highest reactivity - Clean, Minimum Purchasability - In Stock, Representation pH(s) - Reference and all allowed charges.

Molecular dynamics simulations were carried out using the ACEMD software (Harvey, Giupponi, and Fabritiis 2009) with the Amberff14SB force field (Maier et al. 2015) and TIP3P water model (Jorgensen et al. 1983). The system was minimized and equilibrated under constant pressure (1 atm) and temperature (300 °K) conditions using a time step of 4 fs, a non-bonded cutoff of 9 Å, and particle-mesh Ewald long-range electrostatics with a grid of 52 × 81 × 103 with spacing of 1 Å. The system was first equilibrated during 10 ns, then a production run of 4 μs was performed. The processing of the trajectory and the secondary structure timeline analysis were carried out by using VMD (Humphrey, Dalke, and Schulten 1996) and the GROMACS dssp utilities (Kabsch and Sander 1983) (Touw et al. 2014).

3.4.5 Toxicity test of extracellular MEG on bacterial cells

All three tests were performed in *E. coli* BL21(DE3) expression bacteria in 10 mL of LB medium without antibiotics, the growth was performed at 37 °C, shaking at 150 rpm. Peptide samples were added to the media in three different forms at an OD_{600} of 0.6 and then the OD_{600} was measured at time intervals up to values approaching 2.0. For all three toxicity tests a BL21(DE3) bacterial culture grown in LB was chosen as a negative control. All samples were first allowed to grow together in a total volume of LB media so that subsequent peptide addition was performed at an OD_{600} of 0.6 identical for all samples.

The first test was performed by adding multiple peptides dissolved in DMSO-d6 at a concentration of 2 mM, which were previously used for NMR analyses. The peptide sample thus prepared was added to the medium at OD_{600} 0.6 to the final concentration in the medium of 10 nmol. For the second control of this toxicity test, 12.5 μL of DMSO-d6 were added to LB medium, corresponding to the volume of peptide dissolved in DMSO-d6. The following peptides of MEG 2.1 were tested: isoform 3, isoform 2a, isoform 2b, isoform 3 + 2a + 2b together in one sample (to mimic a pseudo full-length MEG 2.1 isoform 2).

The second test was performed by adding 100 μL of peptide in DMSO-d6 to 10 mL of media to avoid exceeding the critical limit of DMSO in the bacterial culture that could cause bacterial death. The negative control was again BL21(DE3) bacteria grown in LB without antibiotics; the second control was BL21(DE3) bacteria grown in LB without antibiotics with the addition of 100 μL of DMSO to 10 mL of media at OD_{600} 0.6. The following peptides of MEG 2.1 were tested: isoform 1; isoform 1 + 3 together (to mimic a pseudo full-length MEG 2.1 isoform 1); isoform 1a; 1b; 1c; 1f; 1g; isoform 3; isoform 2a; isoform 2b; isoform 3 + 2a + 2b together in one sample.

The third test was performed by directly adding the lyophilized peptides to the medium at OD_{600} 0.6. The peptide addition was approximately 0.1 g. The negative control was again a growth of bacteria in LB medium. The following peptides of MEG 2.1 were tested: isoform 1; isoform 1 + 3 together; isoform 1a; 1b; 1c; 1f; 1g; isoform 3; isoform 2a; isoform 2b; isoform 3 + 2a + 2b together in one sample.

REFERENCES

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, and R Dustin Schaeffer. 2021. 'Accurate prediction of protein structures and interactions using a three-track neural network', *Science*, 373: 871-76.

Becker, Johanna, Martin RM Koos, David Schulze-Sünninghausen, and Burkhard Luy. 2019. 'ASAP-HSQC-TOCSY for fast spin system identification and extraction of long-range couplings', *Journal of Magnetic Resonance*, 300: 76-83.

Boyer, Robert D, Ross Johnson, and Krish Krishnamurthy. 2003. 'Compensation of refocusing inefficiency with synchronized inversion sweep (CRISIS) in multiplicity-edited HSQC', *Journal of Magnetic Resonance*, 165: 253-59.

Cavanagh, John, and Mark Rance. 1990. 'Sensitivity improvement in isotropic mixing (TOCSY) experiments', *Journal of Magnetic Resonance (1969)*, 88: 72-85.

Crooks, Gavin E, Gary Hon, John-Marc Chandonia, and Steven E Brenner. 2004. 'WebLogo: a sequence logo generator', *Genome research*, 14: 1188-90.

Delaglio, Frank, Stephan Grzesiek, Geerten W Vuister, Guang Zhu, John Pfeifer, and AD Bax. 1995. 'NMRPipe: a multidimensional spectral processing system based on UNIX pipes', *Journal of biomolecular NMR*, 6: 277-93.

Eberhardt, Jerome, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. 2021. 'AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings', *Journal of chemical information and modeling*, 61: 3891-98.

Gasteiger, Elisabeth, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D Appel, and Amos Bairoch. 2003. 'ExPASy: the proteomics server for in-depth protein knowledge and analysis', *Nucleic Acids Research*, 31: 3784-88.

Gasteiger, Elisabeth, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. 2005. *Protein identification and analysis tools on the ExPASy server* (Springer).

Güntert, Peter, and Lena Buchner. 2015. 'Combined automated NOE assignment and structure calculation with CYANA', *Journal of biomolecular NMR*, 62: 453-71.

Güntert, Peter, Christian Mumenthaler, and Kurt Wüthrich. 1997. 'Torsion angle dynamics for NMR structure calculation with the new program DYANA', *Journal of molecular biology*, 273: 283-98.

Gupta, Ramneek, and Søren Brunak. 2001. 'Prediction of glycosylation across the human proteome and the correlation to protein function'.

Harvey, Matt J, Giovanni Giupponi, and G De Fabritiis. 2009. 'ACEMD: accelerating biomolecular dynamics in the microsecond time scale', *Journal of chemical theory and computation*, 5: 1632-39.

Howe, Kevin L, Bruce J Bolt, Myriam Shafie, Paul Kersey, and Matthew Berriman. 2017. 'WormBase ParaSite− a comprehensive resource for helminth genomics', *Molecular and biochemical parasitology*, 215: 2-10.

Humphrey, William, Andrew Dalke, and Klaus Schulten. 1996. 'VMD: visual molecular dynamics', *Journal of molecular graphics*, 14: 33-38.

Irwin, John J, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. 2020.

'ZINC20—a free ultralarge-scale chemical database for ligand discovery', *Journal of chemical information and modeling*, 60: 6065-73.

Jiménez, José, Stefan Doerr, Gerard Martínez-Rosell, Alexander S Rose, and Gianni De Fabritiis. 2017. 'DeepSite: protein-binding site predictor using 3D-convolutional neural networks', *Bioinformatics*, 33: 3036-42.

Jorgensen, William L, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. 1983. 'Comparison of simple potential functions for simulating liquid water', *The Journal of chemical physics*, 79: 926-35.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and Anna Potapenko. 2021. 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596: 583-89.

Kabsch, Wolfgang, and Christian Sander. 1983. 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers: Original Research on Biomolecules*, 22: 2577-637.

Kay, Lewis, Paul Keifer, and Tim Saarinen. 1992. 'Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity', *Journal of the American Chemical Society*, 114: 10663-65.

Kövér, Katalin E, Victor J Hruby, and Dušan Uhrín. 1997. 'Sensitivity-and gradient-enhanced heteronuclear coupled/decoupled HSQC–TOCSY experiments for measuring long-range heteronuclear coupling constants', *Journal of Magnetic Resonance*, 129: 125-29.

Lee, Woonghee, Mehdi Rahimi, Yeongjoon Lee, and Abigail Chiu. 2021. 'POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules', *Bioinformatics*, 37: 3041-42.

Lee, Woonghee, Marco Tonelli, and John L Markley. 2015. 'NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy', *Bioinformatics*, 31: 1325-27.

Letunic, Ivica, and Peer Bork. 2007. 'Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation', *Bioinformatics*, 23: 127-28.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, and Yaniv Shmueli. 2023. 'Evolutionary-scale prediction of atomic-level protein structure with a language model', *Science*, 379: 1123-30.

Loytynoja, Ari, and Nick Goldman. 2008. 'Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis', *Science*, 320: 1632-35.

Madeira, Fábio, Matt Pearce, Adrian RN Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. 2022. 'Search and sequence analysis tools services from EMBL-EBI in 2022', *Nucleic Acids Research*, 50: W276-W79.

Mahram, Atabak, and Martin C. Herbordt. 2015. 'NCBI BLASTP on High-Performance Reconfigurable Computing Systems', *ACM Trans. Reconfigurable Technol. Syst.*, 7: Article 33.

Maier, James A, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. 2015. 'ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB', *Journal of chemical theory and computation*, 11: 3696-713.

Martínez-Rosell, Gerard, Toni Giorgino, and Gianni De Fabritiis. 2017. 'PlayMolecule ProteinPrepare: a web application for protein preparation for molecular dynamics simulations', *Journal of chemical information and modeling*, 57: 1511-16.

Morris, Garrett M, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. 2009. 'AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility', *Journal of computational chemistry*, 30: 2785-91.

O'Boyle, Noel M, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. 2011. 'Open Babel: An open chemical toolbox', *Journal of cheminformatics*, 3: 1-14.

Papadopoulos, Jason S, and Richa Agarwala. 2007. 'COBALT: constraint-based alignment tool for multiple protein sequences', *Bioinformatics*, 23: 1073-79.

Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. 2004. 'UCSF Chimera—a visualization system for exploratory research and analysis', *Journal of computational chemistry*, 25: 1605-12.

Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Elaine C Meng, Gregory S Couch, Tristan I Croll, John H Morris, and Thomas E Ferrin. 2021. 'UCSF ChimeraX: Structure visualization for researchers, educators, and developers', *Protein Science*, 30: 70-82.

Ramachandran, Srinivas, Pradeep Kota, Feng Ding, and Nikolay V Dokholyan. 2011. 'Automated minimization of steric clashes in protein structures', *Proteins: Structure, Function, and Bioinformatics*, 79: 261-70.

Schanda, Paul, and Bernhard Brutscher. 2005. 'Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds', *Journal of the American Chemical Society*, 127: 8014-15.

Schrödinger, L, and Warren DeLano. 2020. "PyMOL." In.

Skalic, Miha, Gerard Martínez-Rosell, José Jiménez, and Gianni De Fabritiis. 2019. 'PlayMolecule BindScope: large scale CNN-based virtual screening on the web', *Bioinformatics*, 35: 1237-38.

Song, Yifan, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, and David Baker. 2013. 'High-resolution comparative modeling with RosettaCM', *Structure*, 21: 1735-42.

Steentoft, Catharina, Sergey Y Vakhrushev, Hiren J Joshi, Yun Kong, Malene B Vester-Christensen, Katrine T-BG Schjoldager, Kirstine Lavrsen, Sally Dabelsteen, Nis B Pedersen, and Lara Marcos-Silva. 2013. 'Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology', *The EMBO journal*, 32: 1478-88.

Thrippleton, Michael J, and James Keeler. 2003. 'Elimination of zero-quantum interference in two-dimensional NMR spectra', *Angewandte Chemie International Edition*, 42: 3938-41.

Touw, Wouter G., Coos Baakman, Jon Black, Tim A. H. te Beek, E. Krieger, Robbie P. Joosten, and Gert Vriend. 2014. 'A series of PDB-related databanks for everyday needs', *Nucleic Acids Research*, 43: D364-D68.

Trott, Oleg, and Arthur J Olson. 2010. 'AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading', *Journal of computational chemistry*, 31: 455-61.

UniProt, Consortium The. 2022. 'UniProt: the Universal Protein Knowledgebase in 2023', *Nucleic Acids Research*, 51: D523-D31.

Waterhouse, Andrew, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, and Lorenza Bordoli. 2018. 'SWISS-MODEL: homology modelling of protein structures and complexes', *Nucleic Acids Research*, 46: W296-W303.

Waterhouse, Andrew M, James B Procter, David MA Martin, Michèle Clamp, and Geoffrey J Barton. 2009. 'Jalview Version 2—a multiple sequence alignment editor and analysis workbench', *Bioinformatics*, 25: 1189-91.

Webb, Benjamin, and Andrej Sali. 2016. 'Comparative protein structure modeling using MODELLER', *Current protocols in bioinformatics*, 54: 5.6. 1-5.6. 37.

List of figures

List of tables

4 RESULTS

4.1 Bioinformatic analysis of MEG superfamily

With the rapid growth of genomic and proteomic data, bioinformatics has become an essential component in the study of proteins, their structures, and their functions. In particular, bioinformatics plays a critical role in the identification of gene and protein sequences, the prediction of protein structures and the analysis of protein interactions. By combining computational methods with experimental techniques, bioinformatics has enabled us to gain new insights into the molecular mechanisms underlying protein function and is facilitating the design of new drugs and therapeutics. Therefore, bioinformatics has emerged as a key discipline in protein structure research, with important implications for medicine, biotechnology, and other areas of science.

Despite the vast number of bioinformatic tools that can be used nowadays, it is still necessary to interpret their results in a critical way and to support them with experimental analyses. In the case of MEG proteins, the biggest obstacle to bioinformatic structural studies is that these proteins have no homologous partners and therefore any *in silico* homology modelling proves to be very difficult. At the same time, these are mostly short unstructured proteins, therefore their structure is very difficult to predict even in the case of homology. Moreover, the primary structure of MEG proteins is often rich in cysteine and thus the correct formation of native disulfide bonds can become very challenging for predicting the structure, both for homologous and *ab initio* structural predictions software.

Before I discuss MEG 2.1, MEG 3.2, and MEG 6 proteins in more detail, I will introduce MEG proteins as a group. As I have already mentioned (Methodology, 3.1.1 Phylogenetics and primary sequence analysis), there are 103 *Schistosoma mansoni* MEG proteins and 108 *Schistosoma* genus proteins (also including *S. japonicum*, *S. haematobium*) annotated in the UniProt database (last access date April 2023). There are also 8 proteins annotated as ESP15/ESP15-family/egg secreted protein ESP15-like proteins, within the MEG 2 family, 4 proteins annotated as MEG Grail family (within the MEG 3 family) and 51 proteins annotated as “micro-exon genes” in the UniProt database. Within the WormBase ParaSite database (Howe et al. 2017) there are 40 genes under the entry "MEG Schistosoma", 3 genes for "GRAIL", 2 genes for "ESP15" and one gene for "antigen 10.3". All these names designate MEG proteins and/or *meg* genes of the genus *Schistosoma*. In addition to the fact that one protein is often tagged as both MEG and ESP15/Grail, one protein/gene is also submitted under different names and sequence numbers, despite being the same protein or nucleotide sequence. For example, MEG 2.2 is in the UniProt database twice: both entries refer to 83 aa sequence that differs in two amino acids.

Table 4.1 - List of duplicity and multiplicity of some S. mansoni MEG proteins sharing the same name in UniProt DB but displaying different lengths and amino acid sequences (last access date April 2023). Within the parenthesis there is the unique UniProt identifier.

Protein name	**Amino acid length of sequences with the same acronym (UniProt identifier)**				
MEG 1	186 (A0A5K4F8B3)	179 (A0A5K4F8U8)			
MEG-1 family	155 (A0A3Q0KKC4)	154 (A0A5K4EKN1)			
MEG 2.2	83 (A0A5K4FFX0)	83 (D7PD77)			
MEG-10	56 (A0A3Q0KQ39)	55 (G4LYD0)			
MEG-13	130 (A0A3Q0KLA7)	125 (A0A5K4EL02)			
MEG-14	86 (M1GUG5)	156 (Q8ITD5)	98 (Q8ITE1)		
MEG-14 isoform 3	150 (A0A5K4EK08)	141 (A0A1C9A1H6)			
MEG-2 (ESP15 family)	69 (A0A3Q0KR24)	81 (C4QG05)	73 (C4QPR6)	48 (C4QPR9)	44 (C4QPS0)
MEG 27	55 (A0A0U5KIV9	55 (A0A5K4F014)			
MEG-3 (Grail) family	151 (A0A3Q0KMS0)	151 (A0A3Q0KMU6)			
MEG-7	116 (G4V7W5)	145 (A0A5K4EUJ7)			
MEG-8 family	188 (G4VCW5)	140 (G4VLP3)			

Since all proteins in Table 1 are annotated only as “inferred from nucleotide sequences” in the UniProt database, we can assume that these duplicated/multiplied entries were generated due to challenging sequencing of the alternatively spliced *meg* genes. As already said, these genes have a unique sequential organization, which contains from 10 to 20 short exons that are interspersed with long introns, whose length is several times longer (from 100 to 5000 bp). Micro-exons, on the other hand, are most often multiples of three base pairs (bp) and range in size from 6 bp to 81 bp; the most common exon length is 15 bp (Berriman et al. 2009; DeMarco et al. 2010; Howe et al. 2017). Therefore, these multiplicity annotations and naming can lead to confusion, incorrect sequence analyses and other misunderstandings.

4.1.1 Phylogenetics and primary sequence analysis

For the purpose of bioinformatic analyses in this work and terminology clarity, trimming and elimination of duplicate inputs that had the same sequence and different names were performed. Whenever there was any doubt in the sequences, only those verified as proteins with a UniProt ID were considered as the decisive ones, for consistency with cross-annotation with the WormBase ParaSite DB. The final version of the "cleaned" sequences contains 35 meg genes encoding 87 confirmed MEG proteins in *S. mansoni* (all the sequences in Annex B). The difference between the number of coding genes and the number of proteins is caused by the number of isoforms generated from one pre-mRNA by alternative splicing. MEG proteins are annotated and divided into 26 families and further into family members and then into individual isoforms: for example, MEG 2 family – family member MEG 2.1 – isoform 1). These families are grouped from MEG 1

to MEG 16 families, then there is a gap in the numbering and the next family starts at MEG 26 and goes to MEG 32.

To simplify and clarify the confusing nomenclature, we have proposed a modification of their naming based on sequence similarity, conserved motifs, and their organization within their placement on the genes. In fact, the 35 *meg* genes are located on 7 autosomes (from 1 - 7, Fig. 4.1) and one sexual chromosome (ZW, Fig. 4.1). All chromosomes contain at least one *meg* gene, as it is the case of chromosomes 2, 4 and the sex chromosome (Zwang and Olliaro 2017) (*meg-8* in the second chromosome, *meg-15* in the fourth, and *meg-6* in the ZW chromosome). MEG 6 on the sexual chromosome is the only one of the entire MEG-family that does not contain cysteines and does not have a predicted signal peptide, two traits that make it different from the rest of the family.

At the other end of the spectrum, the most *meg*-occupied chromosome is the third one, which contains 13 different ones. The third chromosome is occupied by representatives of MEG 1, MEG 2, and MEG 3 families. Proteins encoded by those genes are abundantly represented in egg secretions (Anderson et al. 2015; DeMarco et al. 2010; Lu et al. 2021). In addition, MEG 3 family and MEG 1 family are the families with the highest number of confirmed isoforms produced by alternative splicing. The second most populated chromosome is the first one, which contains 8 *meg* genes, one of which is referred to as *antigen 10.3* (Fig. 4.1).

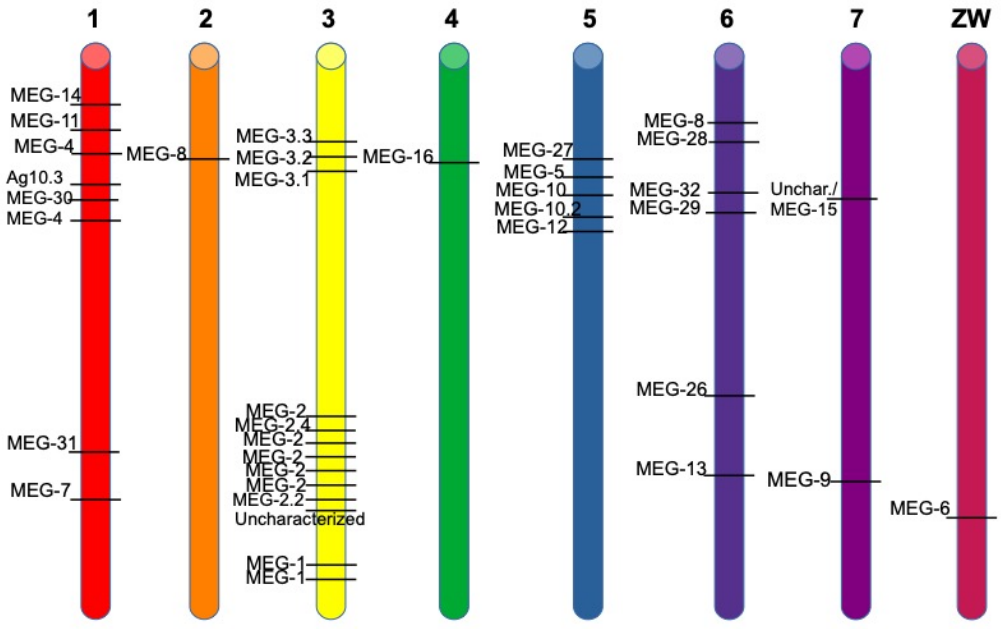

Figure 4.1 - Schematic representation of S. mansoni haplotype. The approximate position of each MEG gene on each chromosome (colored cylinder) is indicated by a black bar and its name on the WormBase ParaSite is noted on the left. The chromosome number is on top of each cylinder.

Together with the gene distribution on the individual eight chromosomes, a phylogenetic tree was constructed (Fig. 4.2), which also contributed to a better understanding of the character of the entire MEG superfamily. From this phylogenetic tree it is distinguishable at first glance that these are not 26 completely distinct families. From all verified sequences it is possible to define two principal subfamilies, which we have arbitrarily decided to show here as red and blue clades (Fig. 4.2). The red clade contains proteins of the MEG-29, MEG-28, MEG-31 family, one member of the MEG-2 family (named MEG 2), MEG-9, MEG 3 "Grail", MEG-1, and an uncharacterized protein Smp 326790.2. It is also noticeable that the above-mentioned single member of the MEG-2 family (named as MEG 2, with the UniProt ID C4QPS0, encoded by Smp_180340.1) was separated together with MEG-29 from the rest of the red clade, therefore they are also separated by color shade (darker red).

The second family, designated as the blue clade, is slightly more diverse and contains 5 sub-clades, the largest of which contains several members of the MEG 2 (ESP15) family, along with MEG 11, MEG 15, MEG 6, and MEG 8 (encoded by Smp 172180.1) members (royal blue). The second sub-clade of the blue clade is the lighter blue one containing MEG 10.3, MEG 12, MEG 30, MEG 8 (encoded by Smp 171190.1), MEG 27 and MEG 26. From this division also came another small sub-clade containing three MEG-10 proteins. Next to this small sub-clade, there is a fourth blue one (light cyan) containing MEG 14 and MEG 13 proteins, which was separated from the previous ones. The last early separated (parallel to the whole blue clade) sub-clade (sky blue) contains MEG 16, MEG 32, and MEG 7 proteins.

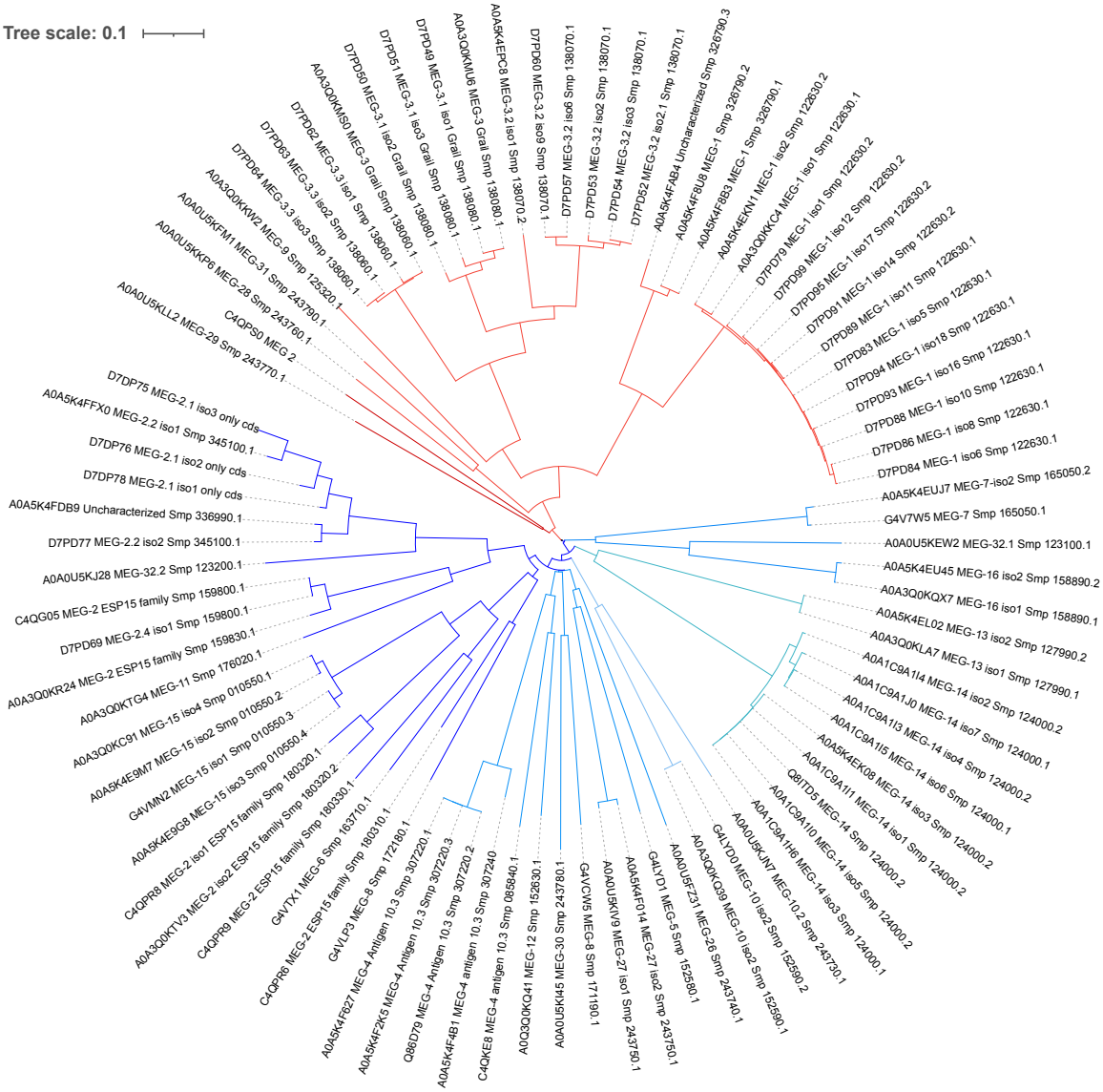

Figure 4.2 - Phylogenetic tree colored by clustering of the clades by sequence similarity. *The clades are colored in red and blue.* In the red clade an early event has separated MEG 29 and MEG 2 (ESP15, coded by Smp_183040.1) from the rest and it was colored in dark red. Similarly, on the blue clade, MEG 7, MEG 32, and MEG 16 departed early from the clade and are highlighted in light blue.

It is worth noticing that the strategies to achieve the largest diversity between the two clades are different: in fact, the blue clade has achieved it by sub-branching, i.e., gene duplication and sequence divergence; while the red clade has largely made use of alternative splicing to create the largest possible protein variability. Both of these strategies are one of the many successful tools of the parasites, which seek to maximize protein variability in order to escape the attention of the host immune system (DeMarco et al. 2010; Philippsen, Wilson, and DeMarco 2015; Fneich et al. 2016; Hull and Dlamini 2014).

Following multiple sequence alignment of the 87 confirmed sequences of the MEG super-family of proteins, we can state that there are very few consensus motifs conserved, except the N-terminal signal peptide, which is needed for secretion (Fig. 4.3). Apolar residues such as Pro, Phe, and Leu are regularly spaced and highly represented. Charged residues, mostly basic Lys (K), are more conserved in the C-terminal part, while glutamic acid (E) is interspersed at more or less regular intervals of 15-20 aa (without taking into account the gaps). Soon after the signal peptide, all the MEGs present a stretch of hydrophobic and aromatic residues ended by a basic one: FLffϕFX_6W*p*(K/H/R). Where X is any residue, p is a polar amino acid and ϕ is a hydrophobic one.

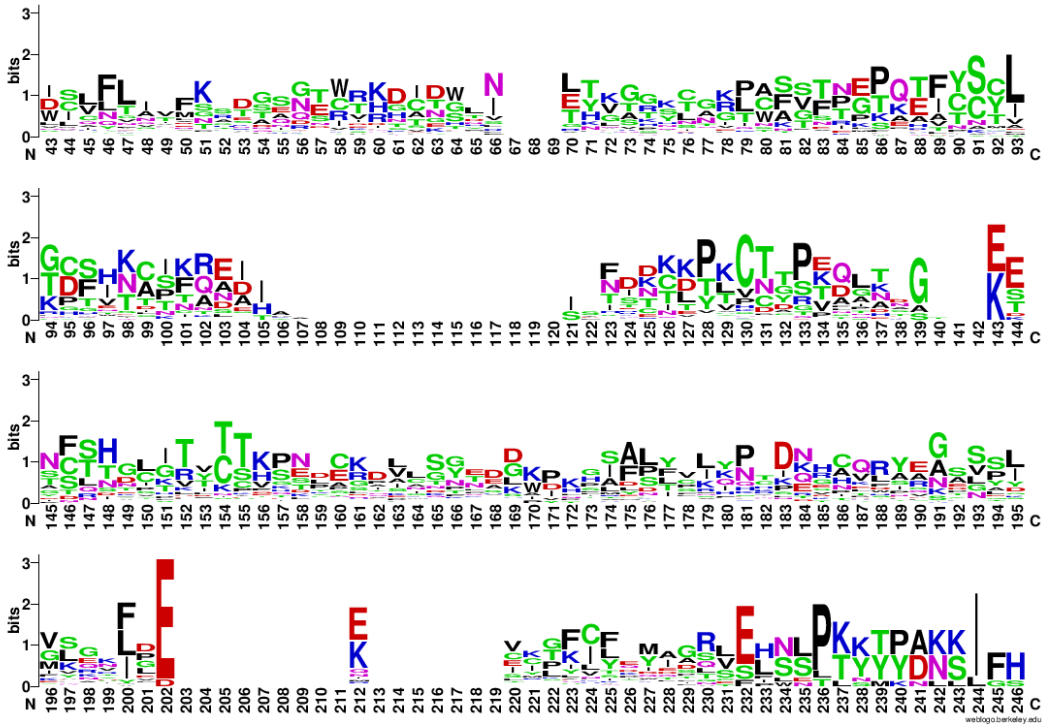

Figure 4.3 - Weblogo representation of the alignment of all the 87 MEG protein sequences. The sequences of the signal peptide have been omitted for clarity. Even if the longest protein is 189 residues long, the number of gaps lengthens the aligned sequences to 246 residues.

Independently of the clade, it is apparent that MEG proteins possess sticky sequences, which we have verified by calculating the aliphatic index and the GRAVY index with ProtParam (Duvaud et al. 2021), data that we have included in the Annex A. If we split the two clades and align the sequences separately (35 proteins for the red clade and 52 for the blue clade), we can appreciate that the contribution of conserved Cys and Phe to the overall alignment comes from the red clade (Fig. 4.4). On the other hand, the basic residues at the C-termini are contributed by the isoforms of the blue clade. Proline residues are conserved in both clades.

A hydrophobic motif at the N-terminus, soon after the signal peptide, is also present in the red clade (Fig. 4.4), but with a slightly different sequence [FxxLFL(I/R)(V/D/E)Fxx(D/E)]. Moreover, we can appreciate that this first linear motif is followed by other four conserved motifs: CGGL*pp*G; (D/E)F(D/I/E)KC$\phi\phi$(R/K); $Cx_{5/7/9}Hx_{3/5/7}C$; and CLY*pp*DX_3L(Y/F/D)V. In total, five short linear motifs characterize the

red clade from the N- to the C-terminus, the first one being in common with the blue clade. It would be interesting to experimentally check whether these peptides are conserved because they are antigenic or because they confer some structural features to the IDPs.

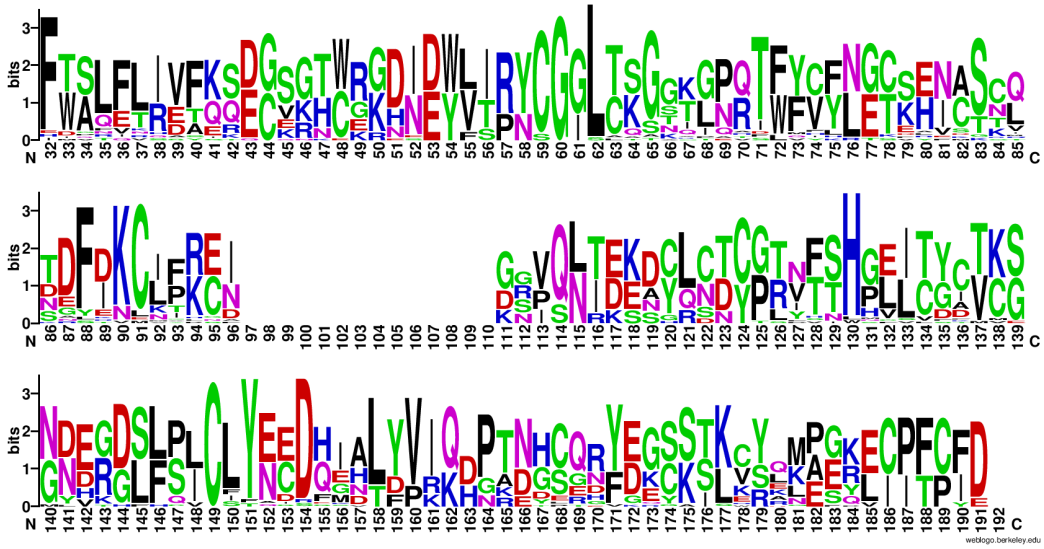

Figure 4.4 - Weblogo representation of the alignment of the red clade composed of 35 MEG proteins. The clade is composed of MEG-1, MEG-3, MEG-9, MEG-28, MEG-29, MEG-31 and C4QPS0 of MEG-2 family. The sequence of the signal peptide has been omitted for clarity.

4.1.2 Sequence analysis of MEG 3.2, MEG2.1 and MEG 6

In this thesis I have focused my attention to MEG 3.2 isoform 1, MEG 2.1 isoform 1 and MEG 6 proteins (sequences in Methodology, 3.2 Recombinant protein expression, Fig. 3.1) because in Dr. Jan Dvorak's laboratory it was previously found that their mRNAs were particularly abundant in eggs excreted/secreted transcriptome analysis (Introduction Fig. 1.6).

MEG 3.2 belongs to the red clade; it is one of the *meg* genes with the highest number of splice variants (10 isoforms) and it is located on the highest occupied chromosome no. 3. Their alignment (Fig. 4.5) points to an alternative splicing process involving the central exons, the N-terminal signal peptide being conserved in all isoforms. In the course of cleaning and verification of relevant *S. mansoni* MEG proteins/isoforms for bioinformatic analyses, six protein sequences were selected from all possible proteins annotated as "MEG 3.2" in the UniProt database. Only the sequences confirmed both as genes and complete proteins were retained; therefore, we eliminated isoforms 5, 7, 8 and 10 because of their partial sequences. At the same time isoform 4, which has a complete nucleotide sequence, but it is not annotated in the WormBase ParaSite database, was also trimmed. This trimming resulted in 5 isoforms of the MEG 3.2 family and two "isoforms 1": one with UniProt ID D7PD52 (encoded by Smp_138070.1) and the other with UniProt ID A0A5K4EPC8 (encoded by Smp_138070.2), which we aligned in Fig. 4.6. Their sequences differ in the N-terminus and C-terminus; however, the central part is identical. Both these isoforms 1 were annotated based on transcriptomic data, so we can ask whether they are really two different isoforms or just partial transcripts.

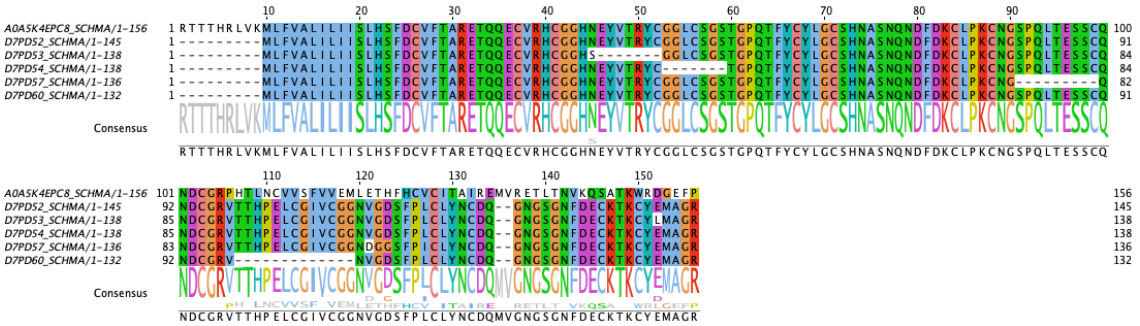

Figure 4.5 - MuscleWS alignment of retrieved sequences of trimmed and verified MEG 3.2 isoforms from the Uniprot database. Alignment was performed with default MuscleWS settings and was coloured with the Clustal X color scheme: hydrophobic residues are colored blue, positively charged are red, negatively charged magenta, polar are green, cysteines coral, glycines orange, prolines dark yellow, aromatic cyan and unconserved white. Consensus was calculated by Jalview 2.11.2.6 program and displayed as a weblogo.

The alignment of all trimmed isoforms (Fig. 4.5) highlights the spliced exons across the MEG 3.2 family. The remaining four isoforms of the MEG 3.2 protein were visibly derived from isoform 1 (D7PD52, encoded by Smp_138070.1). In contrast to the second isoform 1 (A0A5K4EPC8, encoded by Smp_138070.2), all of them have conserved N-terminus signal peptides and MAGR motif at the C-terminus. With the exception of isoform 2 (D7PD53), all other D7PD52-derived isoforms have a conserved 20 amino acids long C-terminus.

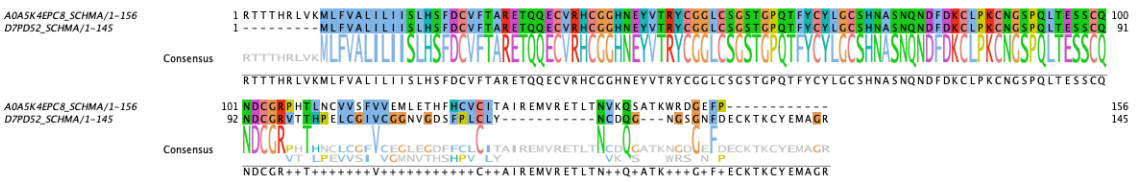

Figure 4.6 - MuscleWS alignment two MEG 3.2 isoforms 1 from the UniProt database. Alignment was performed with default MuscleWS settings and was coloured following the Clustal X color scheme. Consensus was calculated by Jalview 2.11.2.6 program and displayed as a weblogo.

MEG 3.2 isoform 1 contains three potential N-glycosylation motifs, but none of them were predicted to be positive. In contrast, S98 was identified as a very likely O-glycosylation site (Fig. 4.7).

A

```
Name:  Sequence          Length:  145
MLFVALILIISLHSFDCVFTARETQQECVRHCGGHNEYVTRYCGGLCSGSTGPQTFYCYLGCSHNASNQNDFDKCLPKCN      80
GSPQLTESSCQNDCGRVTTHPELCGIVCGGNVGDSFPLCLYNCDQGNGSGNFDECKTKCYEMAGR
................................................................................      80
................................................................      160

(Threshold=0.5)

No sites predicted in this sequence.
```

```
##gff-version 2
##source-version NetOGlyc 4.0.0.13
##date 23-5-28
##Type Protein
#seqname        source  feature start   end     score   strand  frame   comment
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        11      11      0.0303602       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        14      14      0.0310662       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        20      20      0.0517673       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        24      24      0.0534801       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        40      40      0.202861        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        48      48      0.0744034       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        50      50      0.152281        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        51      51      0.0788678       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        55      55      0.0319494       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        63      63      0.0831898       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        67      67      0.0652243       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        82      82      0.28052 .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        86      86      0.418518        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        88      88      0.327339        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        89      89      0.247069        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        98      98      0.723246        .       .       #POSITIVE
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        99      99      0.111447        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        115     115     0.026183        .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        129     129     0.0366456       .       .
TR_D7PD52_D7PD52_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        137     137     0.0303775       .       .
```

B

Figure 4.7 - Prediction of N-glycosylation (A) and O-glycosylation (B) for MEG 3.2 isoform 1 protein. Asn-Xaa-Ser/Thr motif in the sequence output below are highlighted in blue.

MEG 2.1 (Fig. 4.8) belongs to the blue clade, it has only three isoforms, and it is also located on the highest occupied chromosome 3. Their alignment indicates that alternative splicing has occurred in the middle of the sequence, leaving the YPT motif at the C-terminus intact in all three sequences; the N-terminal signal peptide is also conserved. The third isoform is practically composed by a signal peptide and a conserved C-terminal motif of three amino acids. MEG 2.1 isoform 1 is amphipatic (GRAVY -0.05), while isoform 2 is definitely hydrophobic (GRAVY 0.51) and isoform 3 is strongly hydrophobic (GRAVY 1.14), which is consistent with the fact that it is a signal peptide sequence with the YTP conserved motif at the C-terminus. The aliphatic index indicates increasing thermal stability with shortening sequence - i.e., isoform 3 is predicted to be extremely thermostable (AI 146.15), while isoform 1 is predicted to be slightly less thermally stable (AI 86.36) (Kristjansson and Kinsella 1991). Overall, all the sequences of the MEG 2.1 family are considerably hydrophobic (see Annex A and below 4.3.1 Chemical synthesis of MEG 2.1 isoforms 1, 2, and 3 experimentally).

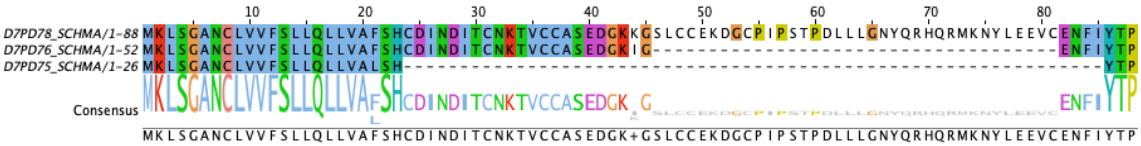

Figure 4.8 - MuscleWS alignment of retrieved sequences of all three MEG 2.1 isoforms from the Uniprot database. Alignment was performed with default MuscleWS settings and was coloured following the Clustal X color scheme. Consensus was calculated by Jalview 2.11.2.6 program and displayed as a weblogo.

MEG 2.1 isoform 1 contains two predicted glycosylation sites. N-glycosylation is predicted at the N32 position, in the NXT motif (Fig. 4.9A), and O-glycosylation is predicted for S58 (Fig. 4.9B).

A

```
Name:  Sequence          Length:  88
MKLSGANCLVVFSLLQLLVAFSHCDINDITCNKTVCCASEDGKKGSLCCEKDGCPIPSTPDLLLGNYQRHQRMKNYLEEV      80
CENFIYTP
...............................N.................................................      80
........                                                                             160

(Threshold=0.5)
--------------------------------------------------------------------------------
SeqName       Position  Potential    Jury    N-Glyc
                                    agreement result
--------------------------------------------------------------------------------
Sequence       32 NKTV    0.6778     (8/9)    +
--------------------------------------------------------------------------------
```

```
##gff-version 2
##source-version NetOGlyc 4.0.0.13
##date 23-5-28
##Type Protein
#seqname        source  feature start   end     score   strand  frame   comment
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        4       4       0.0211332       .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        13      13      0.0226399       .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        22      22      0.0194509       .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        30      30      0.0645757       .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        34      34      0.0154552       .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        39      39      0.228991        .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        46      46      0.0859516       .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        58      58      0.512257        .       .       #POSITIVE
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        59      59      0.185864        .       .
TR_D7PD78_D7PD78_SCHMA  netOGlyc-4.0.0.13       CARBOHYD        87      87      0.00654129      .       .
```

B

Figure 4.9 - Prediction of N-glycosylation (A) and O-glycosylation (B) for MEG 2.1 isoform 1 protein. Asn-Xaa-Ser/Thr motif in the sequence output below are highlighted in blue. Asparagines predicted to be N-glycosylated are highlighted in red.

MEG 6 (Fig. 4.10) is a representative of the blue clade (the second subclade, together with the MEG-2 family) and it is the only one located on the sexual ZW chromosome. This protein is not spliced, it has not a predicted signal peptide and does not contain any cysteines. All these features visibly distinguish it from other MEG proteins; however, MEG 6 protein sequence is more similar to MEG-2 (ESP15) and MEG-8 proteins. This protein is also very interesting for its predicted isoelectric point of 12.14, which is an extreme value indicating that this protein will be always positively charged in physiological conditions. This very basic pI value is indeed peculiar and found only in nuclear proteins, such as histones (Schwartz, Ting, and King 2001). MEG 6 was originally thought to be easier to express than MEG 2.1 and MEG 3.2 because it has only one isoform and because of the absence of cysteine and the predicted signal peptide, however its expression has never been achieved (see below 4.2.1.1.3 MEG 6 and 4.2.4 Yeast expression), possibly the above extreme pI value may be one of the explanations.

Figure 4.10 - Sequence of MEG 6 protein coloured with the Clustal X color scheme.

Despite the fact that MEG 6 has no predicted signal peptide, the prediction of glycosylation pointed to two possible sites for N-glycosylation: two asparagines (N12 and N57), again in the NXT motif (Fig. 4.11A). No potential O-glycosylation was predicted (Fig. 4.11B).

A

```
Name:  Sequence          Length:  65
MVQNPKNTKKINRTIRRSTKTVIVITDRVQNIVLGHRLLHHRIPTIKRSKSHGINKNETVSNLFP
...........N.............................................N........                    80

(Threshold=0.5)
--------------------------------------------------------------------------------
SeqName       Position  Potential    Jury    N-Glyc
                                     agreement result
--------------------------------------------------------------------------------
Sequence        12 NRTI   0.7162      (9/9)   ++
Sequence        57 NETV   0.5713      (7/9)   +
--------------------------------------------------------------------------------

##gff-version 2
##source-version NetOGlyc 4.0.0.13
##date 23-5-28
##Type Protein
#seqname        source  feature start   end     score   strand  frame   comment
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        8       8       0.0818101       .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        14      14      0.0868976       .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        18      18      0.10095 .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        19      19      0.0543545       .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        21      21      0.00894204      .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        26      26      0.00970337      .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        45      45      0.164537        .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        49      49      0.220042        .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        51      51      0.278169        .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        59      59      0.0314097       .       .
TR_G4VTX1_G4VTX1_SCHMA  netOGlyc-4.0.0.13        CARBOHYD        61      61      0.0352909       .       .
```

B

Figure 4.11 - Prediction of N-glycosylation (A) and O-glycosylation (B) for MEG 6 protein. Asn-Xaa-Ser/Thr motifs in the sequence output below are highlighted in blue. Asparagines predicted to be N-glycosylated are highlighted in red.

4.1.3 Comparison of MEG 2.1 isoform 1, MEG 3.2 isoform 1, and MEG 6 with other Schistosomal MEGs

According to blastp, MEG 2.1 isoform 1 is homologous only to *S. mansoni, S. rodhaini, S. margrebowiei, and S. japonicum* hypothetical proteins or other MEG-family (most frequently ESP15-family) proteins (Table 4.2).

Table 4.2 - 22 sequences producing significant alignments with MEG 2.1 isoform 1 from blastp analysis.

Protein Annotation [species]	Max Score	Total Score	Query Cover	E value	% identity	Acc. Len
MEG 2.1 isoform 1 [*S. mansoni*]	182	182	100 %	5E-58	100.00	88
unnamed protein product [*S. rodhaini*]	166	166	100 %	9E-51	89.77	116
egg-secreted protein ESP15 [*S. mansoni*]	163	163	93 %	3E-50	95.12	84
MEG 2.2 *[S. mansoni*]	146	146	96 %	2E-43	84.71	83
unnamed protein product *[S. margrebowiei*]	102	102	96 %	4E-26	67.02	94
unnamed protein product [*S. margrebowiei*]	102	102	96 %	5E-26	66.32	95
MEG-2 (ESP15) family [*S. mansoni*]	93.6	93.6	97 %	3E-22	58.14	112
MEG 2.1 isoform 2 *[S. mansoni*]	89.4	89.4	51 %	2E-21	97.78	52
MEG-2 (ESP15) family [*S. mansoni*]	82.0	82.0	73 %	4E-18	60.61	73
hypotheticial protein [*S. japonicum*]	56.2	56.2	88 %	7E-08	43.04	85
hypotheticial protein [*S. japonicum*]	56.2	56.2	88 %	7E-08	43.04	85
hypothetical protein EWB00_009851 [*S. japonicum*]	53.5	53.5	95 %	1E-06	36.47	88
hypotheticial protein [*S. japonicum*]	53.5	53.5	88 %	1E-06	39.24	104
hypothetical protein EWB00_009851 [*S. japonicum*]	53.1	53.1	90 %	1E-06	38.27	79

hypothetical protein EWB00_009849 [*S. japonicum*]	52.8	52.8	85 %	2E-06	36.71	88
hypotheticial protein [*S. japonicum*]	52.0	52.0	85 %	3E-06	36.71	87
hypotheticial protein [*S. japonicum*]	49.3	49.3	85 %	4E-05	36.71	87
unnamed protein product [*S. rodhaini*]	49.3	49.3	89 %	4E-05	38.55	91
MEG 2.1 isoform 3 [*S. mansoni*]	45.1	45.1	26 %	6E-04	95.65	26
hypothetical protein EWB00_009849 [*S. japonicum*]	45.1	45.1	85 %	0.002	34.67	83
unnamed protein product [*S. margrebowiei*]	44.7	44.7	90 %	0.002	39.29	87
hypotheticial protein [*S. japonicum*]	41.6	41.6	85 %	0.030	31.58	79

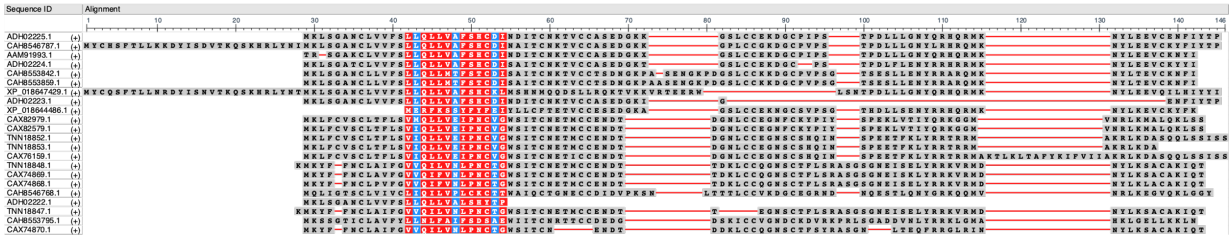

Figure 4.12 - Graphical overview from Constraint-based Multiple Alignment Tool (COBALT) for 22 sequences producing significant alignments with MEG 2.1 isoform 1. Positions where the majority of sequences match the MEG 2.1 isoform 1 sequence are colored in grey, while positions that contain a large proportion of mismatches are represented as red lines. Red-filled boxes indicate highly conserved positions. Red-framed amino acids indicate highly conserved positions and blue ones indicate lower conservation.

In light of the Blastp results (Table 4.3 and Fig. 4.12), I have also attempted to produce 3D models of the MEGs under study by using SWISS-MODEL and AlphaFold2, in parallel to produce the proteins in yields sufficient for experimental 3D structure determination. SWISS-MODEL found for MEG 2.1 isoform 1 a total of 34 templates which were filtered down to three (Table 4.3 below): 19.44 % identity with zinc finger (CCCH-type domain-containing protein 7A; 2d9m.1.A) and 21.43 % identity with methyl-CpG-binding domain protein 2 (2l2l.1.B). AlphaFold2 predicted MEG 2.2 protein (86.75 % identity). None of these templates has an experimentally resolved 3D structure.

Table 4.3 - Filtered 3 top templates matching MEG 2.1 isoform protein sequence from the SWISS-MODEL.

Template	Seq Identity	Oligo-state	QSQE	Found by	Method	Resolution	Seq Similarity	Coverage	Description
A0A5K4FFX0.1.A	86.75	monomer	-	AFDB search	AlphaFold v2	NA	0.59	0.94	MEG 2.2
2d9m.1.A	19.44	monomer	-	HHblits	NMR	NA	0.29	0.41	Zinc finger CCCH-type domain containing protein 7A
2l2l.1.B	21.43	monomer	-	HHblits	NMR	NA	0.29	0.32	Methyl-CpG-binding domain protein 2

Based on homology modelling analyses of selected MEG proteins, it was found that the MEG 3.2 protein isoform 1 exhibits partial homology within the MEG-3 family and simultaneously with all isoforms of the MEG 3.2 family as could be presumed, since the variability of this family is based on alternative splicing (Table 4.4 and Fig. 4.13). It also

shows fewer homologies within other schistosomes and other avian blood flukes from genus *Trichobilharzia* (*Trichobilharzia regenti* and *T. szidati*).

Table 4.4 - 89 sequences producing significant alignments with MEG 3.2 isoform 1 protein from blastp analysis.

Protein Annotation [species]	Max Score	Total Score	Query Cover	E value	% identity	Acc. Len
MEG 3.2 isoform 1 [*S. mansoni*]	297	297	100 %	2E-101	100.00	145
MEG 3.2 isoform 3 [*S. mansoni*]	278	278	100 %	5E-94	95.17	138
MEG 3.2 isoform 2 [*S. mansoni*]	271	271	100 %	3E-91	93.79	138
MEG 3.2 isoform 6 [*S. mansoni*]	266	266	100 %	4E-89	91.72	136
MEG 3.2 isoform 9 [*S. mansoni*]	262	262	100 %	9E-88	91.03	132
MEG 3.2 isoform 8 [*S. mansoni*]	233	233	77 %	6E-76	99.11	169
MEG 3.1 isoform 1 [*S. mansoni*]	222	222	100 %	1E-71	74.17	151
MEG 3.1 isoform 3 [*S. mansoni*]	212	212	100 %	1E-67	72.97	147
MEG-3 (Grail) family [*S. mansoni*]	207	207	93 %	9E-66	74.65	146
MEG 3.2 isoform 5 [*S. mansoni*]	207	207	71 %	2E-65	94.17	168
MEG 3.2 isoform 10 [*S. mansoni*]	200	200	68 %	4E-63	96.97	155
MEG 3.1 isoform 2 [*S. mansoni*]	198	198	83 %	2E-62	78.23	131
MEG 3.3 isoform 1 [*S. mansoni*]	195	195	100 %	4E-61	64.29	151
unnamed protein product [*S. haematobium*]	189	189	100 %	1E-58	64.86	148
unnamed protein product [*S. bovis*]	189	189	100 %	1E-58	64.86	148
MEG 3.2 isoform 4 [*S. mansoni*]	184	184	61 %	1E-57	100.00	89
unnamed protein product [*S. haematobium*]	181	181	100 %	2E-55	64.19	142
MEG-3 (Grail) family [*S. mansoni*]	177	177	100 %	6E-54	60.39	143
MEG 3.3 isoform 2 [*S. mansoni*]	171	171	100 %	2E-51	58.44	147
MEG 3.3 isoform 3 [*S. mansoni*]	170	170	100 %	2E-51	58.44	143
unnamed protein product [*S. bovis*]	170	170	100 %	4E-51	60.81	141
hypotheticial protein [*S. japonicum*]	169	169	100 %	1E-50	52.70	152
hypotheticial protein [*S. japonicum*]	162	162	100 %	5E-48	52.41	148
hypotheticial protein [*S. japonicum*]	150	150	100 %	2E-43	47.97	152
hypothetical protein MS3_00000020 [*S. haematobium*]	146	146	83 %	6E-42	61.16	132
unnamed protein product [*S. margrebowiei*]	140	140	100 %	9E-40	54.73	134
MEG 3.2 isoform 7 [*S. mansoni*]	139	139	46 %	8E-39	94.12	152
unnamed protein product [*S. mattheei*]	134	194	100 %	3E-35	59.82	302
unnamed protein product [*S. bovis*]	105	105	60 %	4E-26	62.22	101
unnamed protein product [*T. regenti*]	94.0	94.0	100 %	6E-21	32.52	161
unnamed protein product [*T. regenti*]	84.3	84.3	86 %	4E-17	37.01	167
unnamed protein product [*T. szidati*]	83.2	83.2	84 %	4E-15	35.71	549
unnamed protein product [*T. szidati*]	81.3	81.3	86 %	6E-16	34.92	168
unnamed protein product [*T. szidati*]	80.5	80.5	84 %	1E-15	34.13	170
unnamed protein product [*T. szidati*]	79.3	79.3	99 %	9E-14	34.01	542
unnamed protein product [*S. intercalatum*]	78.2	78.2	47 %	8E-16	57.97	74
unnamed protein product [*T. regenti*]	77.0	77.0	64 %	5E-15	39.36	106
unnamed protein product [*T. szidati*]	72.8	72.8	88 %	4E-13	34.38	125

unnamed protein product [*S. spindale*]	72.0	72.0	97 %	2E-12	29.75	156
unnamed protein product [*T. regenti*]	70.5	70.5	86 %	7E-12	31.20	162
unnamed protein product [*S. guineensis*]	67.4	67.4	95 %	9E-11	31.41	156
unnamed protein product [*S. intercalatum*]	66.6	66.6	95 %	2E-10	31.41	156
unnamed protein product [*S. curassoni*]	65.9	65.9	44 %	6E-11	49.23	81
unnamed protein product [*S. bovis*]	64.7	64.7	97 %	1E-09	30.38	156
unnamed protein product [*S. curassoni*]	62.4	62.4	95 %	8E-09	30.13	156
hypotheticial protein [*S. japonicum*]	61.6	61.6	97 %	3E-08	28.22	211
MEG 3.4 isoform 1 [*S. mansoni*]	60.8	60.8	97 %	3E-08	26.58	156
unnamed protein product [*S. margrebowiei*]	60.8	60.8	95 %	3E-08	29.49	156
unnamed protein product [*T. regenti*]	60.8	60.8	84 %	7E-08	27.81	197
unnamed protein product [*T. szidati*]	60.5	60.5	97 %	5E-08	31.51	156
hypotheticial protein [*S. japonicum*]	60.5	60.5	82 %	1E-07	29.27	230
hypotheticial protein [*S. japonicum*]	60.5	60.5	82 %	1E-07	29.27	241
hypotheticial protein [*S. japonicum*]	60.5	60.5	82 %	2E-07	29.27	241
hypotheticial protein [*S. japonicum*]	60.1	60.1	82 %	2E-07	29.27	241
hypotheticial protein [*S.* japonicum]	60.1	60.1	82 %	2E-07	29.27	241
hypotheticial protein [*S. japonicum*]	60.1	60.1	82 %	2E-07	29.27	241
hypotheticial protein [*S. japonicum*]	60.1	60.1	82 %	2E-07	29.27	241
hypotheticial protein [*S. japonicum*]	60.1	60.1	82 %	2E-07	29.27	241
unnamed protein product [*T. regenti*]	59.7	59.7	86 %	5E-08	30.40	122
hypotheticial protein [*S. japonicum*]	59.7	59.7	82 %	2E-07	29.27	235
hypotheticial protein [*S. japonicum*]	59.7	59.7	82 %	3E-07	29.27	241
SJCHGC02069 protein [*S. japonicum*]	59.7	59.7	82 %	3E-07	29.27	237
hypotheticial protein [*S. japonicum*]	59.3	59.3	82 %	3E-07	29.27	233
hypotheticial protein [*S. japonicum*]	59.3	59.3	82 %	3E-07	29.27	241
hypotheticial protein [*S. japonicum*]	59.3	59.3	82 %	3E-07	29.27	233
hypotheticial protein [*S. japonicum*]	59.3	59.3	82 %	4E-07	29.27	232
hypotheticial protein [*S. japonicum*]	59.3	59.3	82 %	4E-07	29.27	241
hypotheticial protein [*S. japonicum*]	59.3	59.3	82 %	4E-07	29.27	241
unnamed protein product [*S. rodhaini*]	58.5	58.5	89 %	3E-07	27.08	156
hypotheticial protein [*S. japonicum*]	58.5	58.5	82 %	7E-07	28.46	233
hypotheticial protein [*S. japonicum*]	58.2	58.2	82 %	9E-07	29.27	241
unnamed protein product [*S. rodhaini*]	57.4	57.4	86 %	5E-07	27.86	147
hypotheticial protein [*S. japonicum*]	57.4	57.4	82 %	2E-06	28.46	241
hypotheticial protein [*S. japonicum*]	57.0	57.0	82 %	2E-06	27.64	241
hypotheticial protein [*S. japonicum*]	57.0	57.0	80 %	3E-06	29.17	241
unknown [*S. japonicum*]	56.2	56.2	82 %	1E-06	26.83	127
hypotheticial protein [*S. japonicum*]	56.2	56.2	82 %	5E-06	26.83	241
hypotheticial protein [*S. japonicum*]	55.8	55.8	82 %	7E-06	26.83	241
unnamed protein product [*S. rodhaini*]	54.7	54.7	57 %	3E-06	35.16	107
hypotheticial protein [*S. japonicum*]	53.9	53.9	80 %	4E-05	26.67	241
unnamed protein product [*S. rodhaini*]	52.8	52.8	48 %	1E-05	38.03	102

unnamed protein product [*T. regenti*]	52.8	52.8	75 %	2E-05	28.83	130
hypotheticial protein [*S. japonicum*]	49.7	49.7	74 %	0.001	25.93	224
unnamed protein product [*T. regenti*]	48.1	48.1	71 %	0.002	28.97	154
MEG 3.2 [*S. mansoni*]	47.0	47.0	14 %	4E-04	100.00	21
unnamed protein product [*T. szidati*]	47.0	47.0	62 %	0.003	34.07	129
MEG 3.2 [*S. mansoni*]	46.6	46.6	14 %	6E-04	100.00	21
hypotheticial protein [*S. j*aponicum]	46.6	46.6	74 %	0.012	24.07	231
SJCHGC02070 protein [*S. japonicum*]	45.8	45.8	74 %	0.019	25.00	187

SWISS-MODEL (Waterhouse et al. 2018) identified in total 16 templates and this list was filtered by a heuristic method down only to two with homology of 21.21 % with neurotoxin Cn11 from *Centruroides noxius* (scorpion toxin acting on sodium channels; 1pe4.1.A) and AphaFold2 predicted model of MEG-3 (Grail) family. Again, none of these models belongs to an experimentally resolved 3D structure.

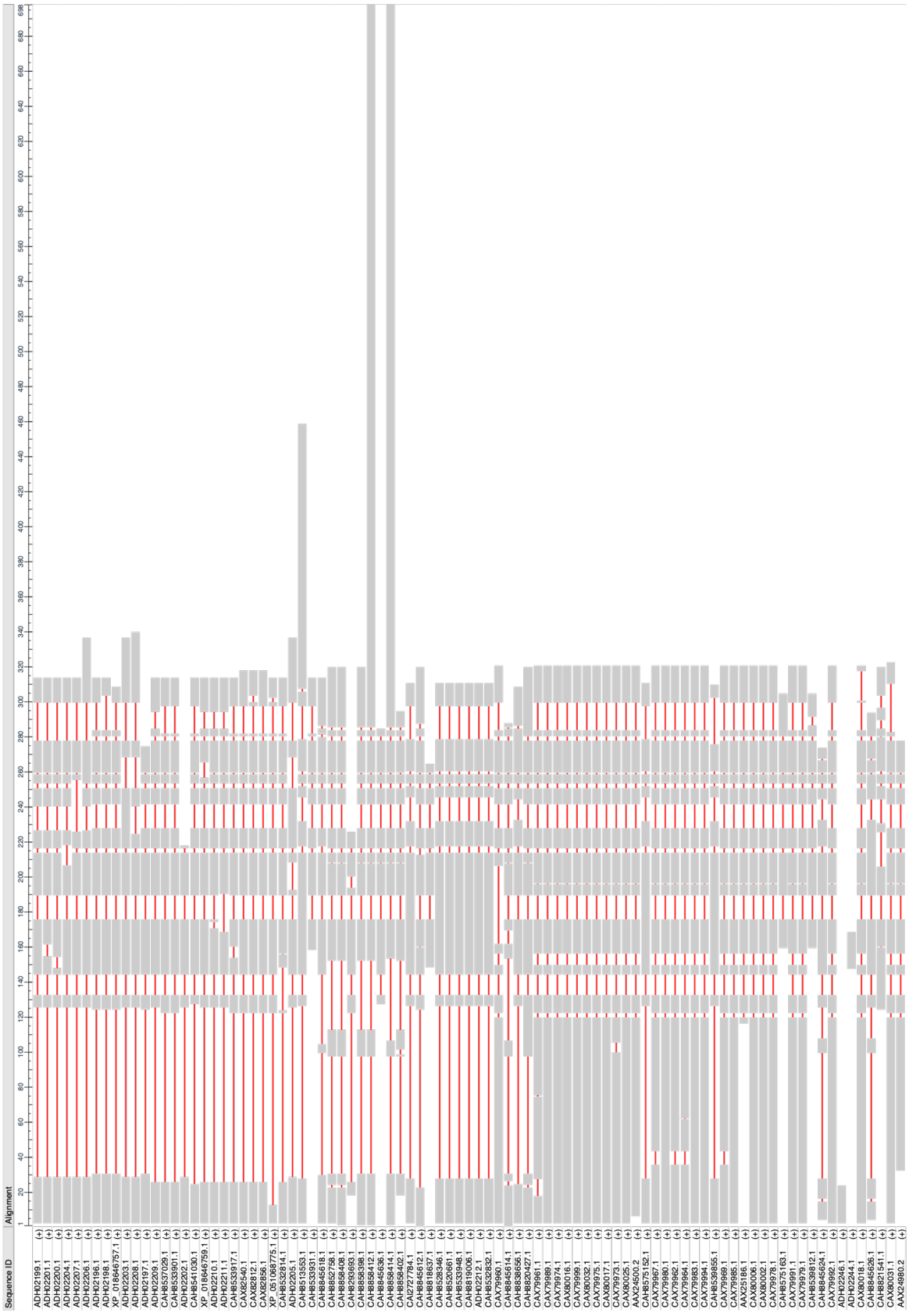

Figure 4.13 - Graphical overview from COBALT for 89 sequences producing significant alignments with MEG 3.2 isoform 1 protein. Positions where the majority of sequences match the MEG 3.2 isoform 1 sequence are colored in grey, while positions that contain a large proportion of mismatches are represented as red lines.

Table 4.5 - Filtered 2 top templates matching MEG 3.2 isoform protein sequence from the SWISS-MODEL.

Template	Seq Identity	Oligo-state	QSQE	Found by	Method	Resolution	Seq Similarity	Coverage	Description
A0A3Q0KMU6.1.A	78.47	monomer	-	AFDB search	AlphaFold v2	NA	0.57	0.99	MEG-3 (Grail) family
1pe4.1.A	21.21	monomer	-	HHblits	NMR	NA	0.35	0.23	Neurotoxin Cn11

MEG 6 is homologous to 11 unnamed proteins of *Schistosoma* genus according to blastp (Table 4.6 and Fig. 4.14).

Table 4.6 - 12 sequences producing significant alignments with MEG 6 protein from the blastp.

Protein Annotation [species]	Max Score	Total Score	Query Cover	E value	% identity	Acc. Len
unnamed protein product [S. rodhaini]	130	130	100 %	9E-37	100.0 %	147
MEG-6 [S. mansoni]	127	127	100 %	1E-36	100.00 %	65
unnamed protein product [S. rodhaini]	115	115	100 %	3E-28	89.39 %	430
unnamed protein product [S. spindale]	85.1	85.1	84 %	4E-19	80.00 %	126
unnamed protein product [S. mattheei]	71.6	71.6	69 %	5E-14	77.78 %	112
unnamed protein product [S. bovis]	77.8	77.8	90 %	1E-14	68.33 %	422
unnamed protein product [S. bovis]	71.2	71.2	92 %	4E-14	65.57 %	91
unnamed protein product [S. bovis]	70.9	70.9	92 %	1E-13	65.57 %	118
unnamed protein product [S. haematobium]	81.3	81.3	98 %	6E-16	64.62 %	415
unamed protein product [S. haematobium]	81.3	81.3	98 %	6E-16	64.62 %	419
unnamed protein product [S. margrebowiei]	43.1	43.1	58 %	0.004	64.10 %	74
unnamed protein product [S. turkestanicum]	65.5	65.5	98 %	1E-11	58.21 %	113

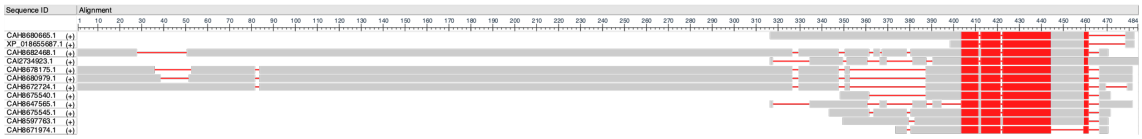

Figure 4.14 - Graphical overview from COBALT for 12 sequences producing significant alignments with MEG 6. Positions where the majority of sequences match the MEG 6 sequence are colored in grey, while positions that contain a large proportion of mismatches are represented as red lines. Red-filled frames indicates highly conserved positions.

The Swiss-Protein database found 22 possible templates for homology modelling from which 18 were selected (Table 4.7).

Table 4.7 - Filtered 18 top templates matching MEG 6 protein sequence.

Template	Seq Identity	Oligo-state	QSQE	Found by	Method	Resolution	Seq Similarity	Coverage	Description
G4VTX1.1.A	100.00	monomer	-	AFDB search	AlphaFold v2	NA	0.61	1.00	MEG-6
3aon.1.B	24.14	monomer	-	HHblits	X-ray	2.00Å	0.32	0.45	V-type sodium ATPase subunit G
5knb.1.H	24.14	monomer	-	HHblits	X-ray	3.25Å	0.32	0.45	V-type sodium ATPase subunit NtpG (F)
3vr4.1.H	24.14	monomer	-	HHblits	X-ray	2.17Å	0.32	0.45	V-type sodium ATPase subunit G
7p48.1.F	19.23	monomer	-	HHblits	EM	NA	0.29	0.40	ABC-F type ribosomal protection protein
6r84.1.A	10.71	monomer	-	HHblits	EM	NA	0.27	0.43	ABC transporter ATP-binding protein ARB1
5zxd.1.A	17.86	monomer	-	HHblits	X-ray	2.29Å	0.30	0.43	ATP-binding cassette sub-family F member 1
5zxd.2.A	17.86	monomer	-	HHblits	X-ray	2.29Å	0.30	0.43	ATP-binding cassette sub-family F member 1
6r84.1.A	7.41	monomer	-	HHblits	EM	NA	0.29	0.42	ABC transporter ATP-binding protein ARB1
7nhk.1.A	3.57	monomer	-	HHblits	EM	NA	0.26	0.43	ABC-F type ribosomal protection protein Lsa(A)
5knd.1.H	24.14	monomer	-	HHblits	X-ray	2.89Å	0.32	0.45	V-type sodium ATPase subunit NtpG (F)
5zlu.1.u	13.79	monomer	-	HHblits	EM	3.60Å	0.29	0.45	Macrolide efflux protein
2d00.1.A	3.45	homo-hexamer	-	HHblits	X-ray	2.20Å	0.27	0.45	V-type ATP synthase subunit F
2d00.1.B	3.45	homo-hexamer	-	HHblits	X-ray	2.20Å	0.27	0.45	V-type ATP synthase subunit F
2d00.1.F	3.45	homo-hexamer	-	HHblits	X-ray	2.20Å	0.27	0.45	V-type ATP synthase subunit F
2d00.1.E	3.45	homo-hexamer	-	HHblits	X-ray	2.20Å	0.27	0.45	V-type ATP synthase subunit F
2d00.1.D	3.45	homo-hexamer	-	HHblits	X-ray	2.20Å	0.27	0.45	V-type ATP synthase subunit F
2d00.1.C	3.45	homo-hexamer	-	HHblits	X-ray	2.20Å	0.27	0.45	V-type ATP synthase subunit F

Only one of these homologous templates showed moderate identity (21.14 %) to the C-terminus of MEG 6, and it was the central axis (NtpD-NtpG) of the catalytic portion of *Enterococcus hirae* V-type sodium ATPase. The remaining templates showed homology below 20%, primarily in the predicted α-helix part at the N-terminus.

From the above-described results, it can be concluded that MEG 2.1, MEG 3.2, and MEG 6 proteins (and their isoforms) show relevant, though low, homology only with other MEG-family proteins and within the genus *Schistosoma*. This not only confirms the uniqueness of these proteins (DeMarco et al. 2010), but also does not allow performing meaningful homology modelling, as there is no X-ray, NMR, or cryo-EM structure on which to build a reliable model. Despite this apparent lack of homology, several comparative *ab initio* predictions of structures have been made. Unfortunately, the results of these predictions vary not only within the chosen methods, date of the run (given the fact that AI neural networks learn from available datasets that change and expand every day), but also across the software used. Despite the imperfection of the predicted models, it is possible to observe some consistent trends across all used software.

4.1.3.1 Ab Initio Protein Structure Prediction

In the last three years, there has been an extreme increase in the use of Artificial Intelligence (AI) for *ab initio* prediction of 3D protein structures. Leaders in this field are AlphaFold2, Robetta, and ESMFold (Baek et al. 2021; Lin et al. 2023; Jumper et al. 2021). These software employ deep learning techniques and very large training data to infer protein structures based on sequence information. While they have shown exceptional accuracy in predicting the structures of folded proteins, their performance with non-homologous Intrinsically Disordered Proteins (IDP) is relatively limited. IDPs are a unique class of proteins that lack a stable 3D structure and exhibit dynamic conformational behavior (Dunker et al. 2013; Lin et al. 2019). IDPs are linked to various diseases and play important roles in many biological processes (Uversky, Oldfield, and Dunker 2008) (Wright and Dyson 2015). Homology modeling, which compares the target protein sequence with recognized structures of related proteins, is the foundation of conventional approaches for predicting protein structures. Homology-based techniques are inadequate for predicting the structures of IDPs because they frequently have little to no substantial sequence similarity to proteins having empirically confirmed structures. The conformational changes and the presence of regions with low sequence homology further complicate the prediction of IDP structures (Wilson, Choy, and Karttunen 2022). In order to supplement AI-based predictions and improve the precision of IDP structure estimation, scientists are looking into additional approaches, such as combining experimental data from methods like NMR, macromolecular crystallography (MX) of complexes or cryo-electron microscopy (cryo-EM). While *ab-initio* software, like AlphaFold 2, use deep learning algorithms and neural networks to learn from known protein structures and predict novel ones, the challenge for non-homologous proteins is that their performance may be constrained in situations where there is no suitable template available for comparison. The structure of non-homologous proteins is often predicted using a combination of experimental, computational, and expert analysis techniques. Despite much effort devoted to advances in AI deep-learning techniques, the prediction of IDP proteins remains a challenging task. For figuring out the precise structures of these proteins, methods like MX, cryo-EM, and NMR spectroscopy remain essential. All three investigated proteins of this manuscript, MEG 3.2 isoform 1, MEG 2.1 isoform 1 and MEG 6, were submitted to AlphaFold2, Robetta, and ESMFold to predict their 3D structures.

To characterize the accuracy of each model, AlphaFold2 (AF2) gives a per-residue confidence metric called the predicted local difference test (pLDDT) (Mariani et al. 2013). A score inferior to 50 indicates a low accuracy or a possible disorder prediction, while a score superior to 70 indicates a plausible prediction of the secondary structure. The best 5-ranked structures are presented below. Robetta also gives multiple models with confidence metric presented as plots, showing corresponding error estimations (in Å) per amino acid position in the sequence. ESMFold is the youngest of the deep learning

AI and uses a color coding based on pLDDT as confidence scoring (blue color represents pLDDT values above 0.9, while red indicates low confidence with pLDDT lower than 0.5).

4.1.3.1.1 MEG 3.2 isoform 1

MEG 3.2 isoform 1 achieves high pLDDT scores for some amino acids, which may seem like a plausible prediction, but there are significant dropouts in these values throughout the length of the sequence. These value drops are mainly in parts of the sequence that do not show any ordered secondary structure. At the same time, the downward trend of pLDDT in the first 20 aa (from the N-terminus) of the AF2 prediction sequence is interesting (Fig. 4.15), because it is the predicted signal peptide that is the most conserved part of the MEG proteins. The motif of the signal peptide is present across almost all of the 89 sequences (Fig. 4.13). Furthermore, it is possible to observe significant differences in the helix length prediction by AF2 of the signal peptide (Fig. 4.15). AF2 predictions indicate the formation of a more compact protein than Robetta's predictions with 7 to 9 helices in the structure.

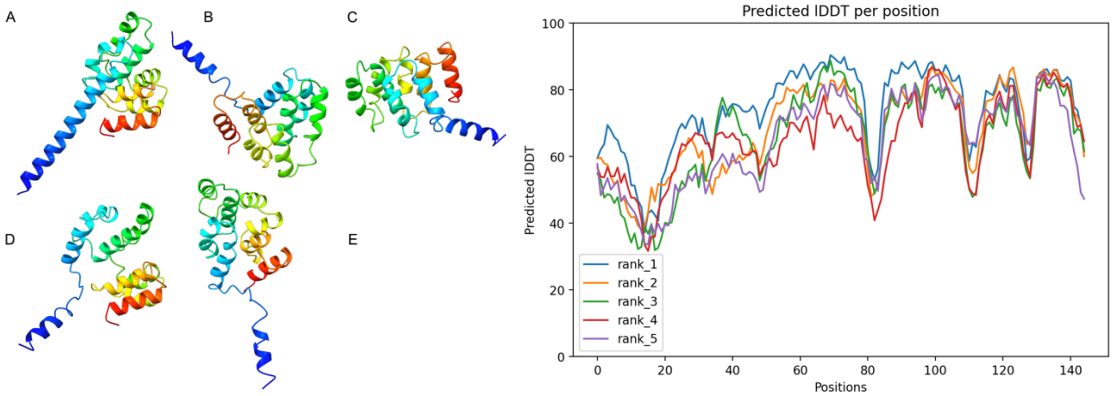

Figure 4.15 - The 5 best-ranked structures of MEG 3.2 isoform 1 protein structure predictions from the AlphaFold2 ColabFold v1.5.2 with default settings (left) with its per-residue confidence scores (pLDDT, right). The order of the model structures corresponds to the order of the plot (A = 1, B = 2, ... E = 5). Models are colored with rainbow scheme - N-terminus (blue) to C-terminus (red).

Robetta also shows spikes in prediction accuracy, which almost replicate the trend of AF2's decreases in the pLDDT values (Fig. 4.16). The N-terminal part of the sequence with the predicted signal peptide and the parts of the protein that show the random coil structure again show the highest error rate. At the same time Robetta predicts fewer helices (4 short helices consisting of 0.5-3 turns), which leads to a less compact structure with a larger part of non-structural elements (Fig. 4.16). Compared to AF2, Robetta does not predict large differences in signal peptide length.

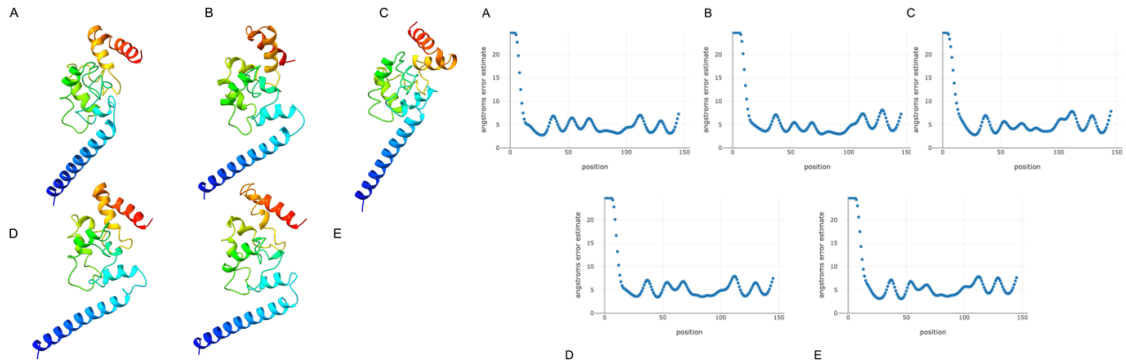

Figure 4.16 - The 5 best-ranked structures of MEG 3.2 isoform 1 protein structure predictions from the Robetta prediction software with RoseTTAFold default settings (left) with their respective plots showing corresponding error estimations in Å per amino acid position (right). Models are colored with rainbow scheme - N-terminus (blue) to C-terminus (red).

ESMFold also predicts a compact structure composed of 9 helices and random coil parts (Fig. 4.17). Again, it points to the lower reliability of the model for the signal peptide region and the non-structural parts. A few turns of some helices are shown to be highly reliable, consistent with previous results of the two AIs mentioned above.

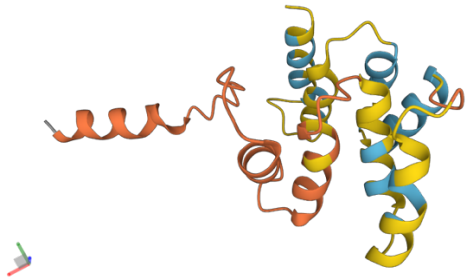

Figure 4.17 - MEG 3.2 isoform 1 protein structure prediction via ESMFold. The predicted structure is colored by local prediction confidence (pLDDT) per amino acid location. Blue indicates confident predictions (pLDDT > 0.9), while red indicates low confidence (pLDDT < 0.5). The amino acid marked in grey is the first from the N-terminus.

4.1.3.1.2 MEG 2.1 isoform 1

MEG 2.1 isoform 1 is the protein for which all three predictive AIs most closely agreed on the resulting structure. All models point to the presence of two helices at the N-terminus and C-terminus, which are connected by a random coil (Fig. 4.18, 4.19, and 4.20). The error rate of AF2 and Robetta is again similar: it increases in the helix regions, while it decreases in the non-structural parts of the protein and at the N-terminus and C-terminus extremities (Fig. 4.18 and 4.19). AF2 in one of the models (Fig. 4.18E) predicts a short region of antiparallel β-sheet inserted in a random coil part linking two helices (Ile29 - Cys31 and Thr34 - Cys36). In this part, Robetta predicts a partial helical turn in four models. Those partial helices, shown in Fig. 4.19, are: model A, from Cys31 to Cys33, model B, from Ile26 to Ile29, models D and E, from Cys48 to Glu50. Model C is the model that is most consistent between Robetta's and the ESMFold predictions, having two long parallel helices and a non-structural part in between.

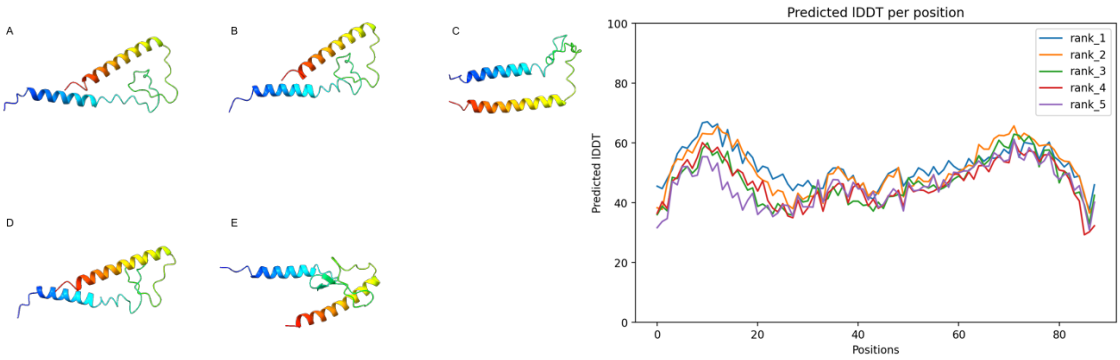

Figure 4.18 - The 5 best-ranked structures of MEG 2.1 isoform 1 protein structure predictions from the AlphaFold2 ColabFold v1.5.2 (left) with default settings with its per-residue confidence scores (pLDDT, right). The order of the model structures corresponds to the rank of the plot (A = 1, B = 2, ... E = 5). Models are colored with rainbow scheme from N-terminus (blue) to C-terminus (red).

Robetta predicts shorter sections of the helix at the N- and C-terminus and the non-structural part in between (Fig. 4.19). However, in this part of the prediction, Robetta again shows the highest error rate of the whole sequence. In the part where AF2 predicts the antiparallel β-sheet, Robetta predicts the partial tours of the helix in four models. Those partial helixes are displayed in Fig. 4.19: model A, from Cys31 to Cys33; model B from Ile 26 to Ile29; model D and E, from Cys48 to Glu50. Model C presents the shortest helix at the C-terminus and thus the largest proportion of the non-structural part of the protein.

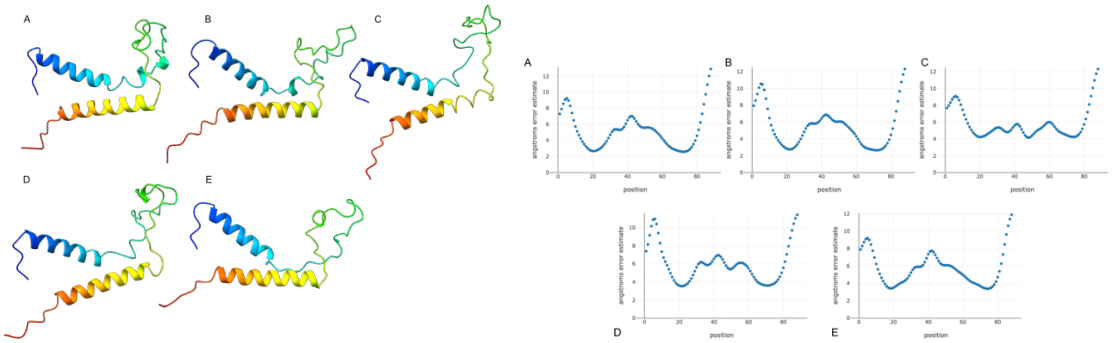

Figure 4.19 - The 5 best-ranked structures of MEG 2.1 isoform 1 protein structure predictions from the Robetta prediction software with RoseTTAFold default settings with its their plots showing corresponding angstrom error estimations per amino acid position. *Models are colored with rainbow scheme - N-terminus (blue) to C-terminus (red).*

EMSFold predictions of MEG 2.1 protein isoform 1 are relatively consistent with both of the above software but indicate low confidences (plDDT below 0.5) all throughout the sequence (Fig. 4.20). All models predict the beginning of the hairpin of the non-structural part around Cys31 region, but they differ considerably in their predictions of its shape and orientation in space. ESMFold and one AF2 model (Fig. 4.18C) predict both helices to be parallel, while the other models present them at an acute angle. Of the three MEG proteins tested, MEG 2.1 isoform 1 is the one whose *ab initio* prediction using the three AIs converges the most.

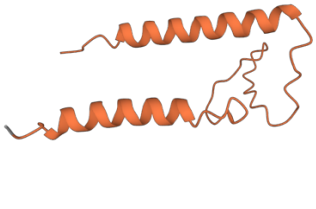

Figure 4.20 - MEG 2.1 isoform 1 protein structure prediction via ESMFold. *The predicted structure is colored by local prediction confidence (pLDDT) per amino acid location.*

4.1.3.1.3 MEG 6

At the other extreme of the spectrum there is MEG 6 protein, whose predictions do not even marginally agree among the three software. MEG 6 is the only MEG protein studied that does not have a predicted signal peptide nor contains cysteines. AF2 predicts a continuous helix from the N-terminus that continues with a random coil. In the region of the predicted helix, the confidence increases, while in the region of the random coil, it is again reduced (Fig. 4.21). In the Arg48 and Ser49 region, one model (Fig. 4.21C) predicts a hairpin. The final part of the other four models is totally unstructured.

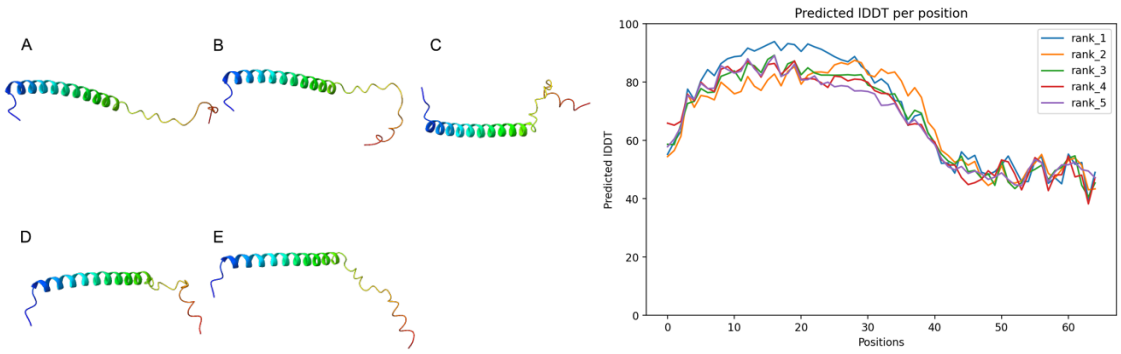

Figure 4.21 - The 5 best-ranked structures of MEG 6 protein structure predictions from the AlphaFold2 ColabFold v1.5.2 (left) with its per-residue confidence scores (pLDDT, right). The order of the model structures corresponds to the order of the plot (A = 1, B = 2, ... E = 5). Models are colored with the rainbow scheme - N-terminus (blue) to C-terminus (red).

Robetta presents its models as a mixture of helices, β-sheets, and random coils. Two models contain antiparallel β-sheets (Fig. 4.22): model B (Val24 - Thr26 and Val29 - Asn31) and model C (Val22 - Val24 and Ile32 - Leu34). All models predict a discontinuous helix at the N-terminus, shorter than the one predicted by AF2, and at the same time another two or three shorter helices spread over the sequence, most of which contain only two turns. The error is significantly high throughout the sequence.

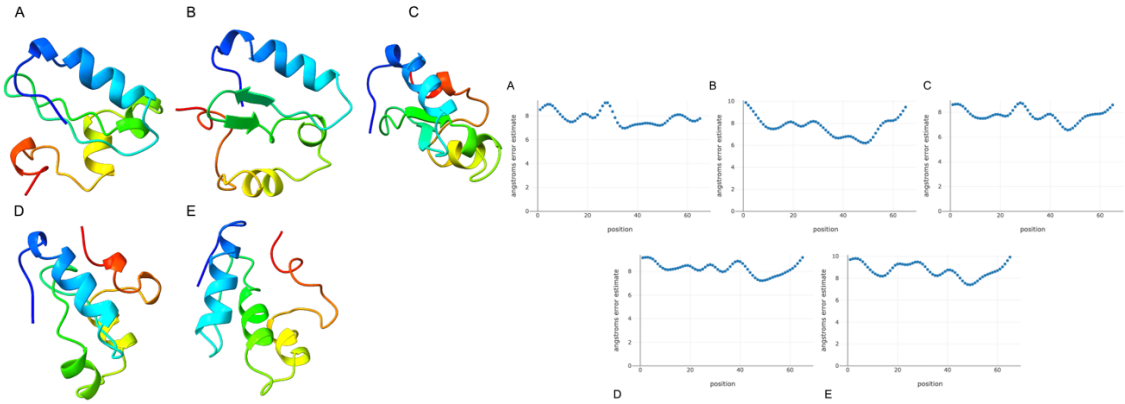

Figure 4.22 - The 5 best-ranked structures of MEG 6 protein structure predictions from the Robetta prediction software with RoseTTAFold default settings (left) with its their plots showing corresponding error estimations Å/aa (left). Models are colored with rainbow scheme - N-terminus (blue) to C-terminus (red).

MEG 6 predicted by ESMFold (Fig. 4.23) has as well a very low confidence (pLDDT below 0.5) along the entire sequence. The whole structure is predicted as a random coil with two sections of very short helices (one turn on the N-terminus and one turn on the C-terminus). The structure is however presented as more compact as compared to the linear AF2 predictions.

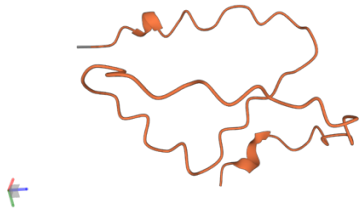

Figure 4.23 - MEG 6 protein structure prediction via ESMFold.

Ab inito prediction of the three MEG proteins confirmed the challenges that predictive deep-learning AIs face today. Shorter non-structural proteins with possible disulfide bonds, without high homology, show significantly low confidence values. In spite of these shortcomings, after experimental measurements (see below), partial agreements were found.

4.2 Recombinant protein expression

In order to resolve the 3D structure of the MEG proteins under study, a large number of essays were performed in many different variations (see also the method section). Not only different expression systems were tested, but also a large scale of tests was performed for each expression system. The expression systems tested were: bacteria [*E. coli* (BL21(DE3), Rosetta(DE3), BL21(DE3)pLysS and Rosetta(DE3)pLysS strains], methanotrophic yeast (*Komagataella phaffii)*, insect cells [S2 (*Drosophila*)] and *in vitro* cell-free expression system. Among these expression systems, only two were suitable for subsequent isotopic labeling for structure determination by NMR: bacterial and cell-free expression. Expression in S2 cells is a very expensive method and subsequent labeling would increase the resulting price many times more; at the same time, this method was performed in the Czech Republic in cooperation with BIOCEV Prague, so

the subsequent transport and handling of the sample implied a complex solution. Moreover, the expression in S2 cells is a very demanding method for sterility and time (minimum 6 weeks of growth of the stable cell line). For these reasons, I could not proceed further. The yeast expression was also performed in Prague, in the IOCB laboratories and the methodology of isotopic labeling was not established in any of the partner laboratories. Thus, it would have been necessary to develop a functional protocol of yeast isotopic labeling first, which was not in the time possibilities of this work and therefore I did not proceed with this expression system after the pilot assays. At the same time, the results of those pilot expressions were not convincing, and it would be necessary to perform several optimizations just for the expression of unlabeled proteins. Cell-free expression was carried out in collaboration with laboratories at the Institute of Structural Biology in Grenoble and isotopic labeling could be performed within the framework of this method, if the method yielded positive results, despite being relatively expensive. The most straightforward, least expensive, and easily accessible expression system for me was the bacterial expression system. I have worked intensively with this system, both in Prague and in Lyon and tried a range of different conditions to express 3 different proteins and their isoforms (see below and also Tables 3.1 - 3.7 in the Methodology).

4.2.1 Bacterial expression

Two plasmids were selected for bacterial expression: the commonly used pET 22b(+) plasmid for periplasmic protein production under the T7 promoter and pET SUMO Champion plasmid, which is meant to produce the highest levels of soluble protein in *E. coli*. Constructs with the 6xHis tag and TEV cleavage site were cloned into the construct within the pET 22b(+) plasmid. All MEG 2.1 isoform 1, MEG 3.2 isoform 1 ("long" and "short" versions) and MEG 6 proteins were cloned into this plasmid. MEG 2.1 isoform 1 and MEG 3.2 isoform 1 proteins were inserted into the pET SUMO Champion plasmid for production of difficult proteins. All detailed information on conditions, plasmids, sequences and expressions have already been given (Methodology, 3.2 Recombinant protein expression, Tables 3.1, 3.2, and 3.3).

4.2.1.1 MEG 2.1 isoform 1

MEG 2.1 isoform 1 (MW = 11.96 kDa, pI = 6.01) in the pET 22b (+) plasmid was not expressed in either the soluble (Fig. 4.24) or insoluble fraction (Fig. 4.25). The tested conditions for BL21(DE3), Rosetta(DE3) and BL21(DE3)pLysS expression strains were: induction temperature (37 °C, 30 °C overnight expression, 18 °C overnight expression), IPTG concentration (0.1 mM, 0.5 mM, and 1 mM at fixed temperature). MEG 2.1 isoform 1 protein did not express even under a special expression protocol for potential toxic proteins in the Rosetta(DE3)pLysS strain (details in Methodology, 3.2.2 Expression in bacteria, Table 3.4, 3.5, 3.6, and 3.7).

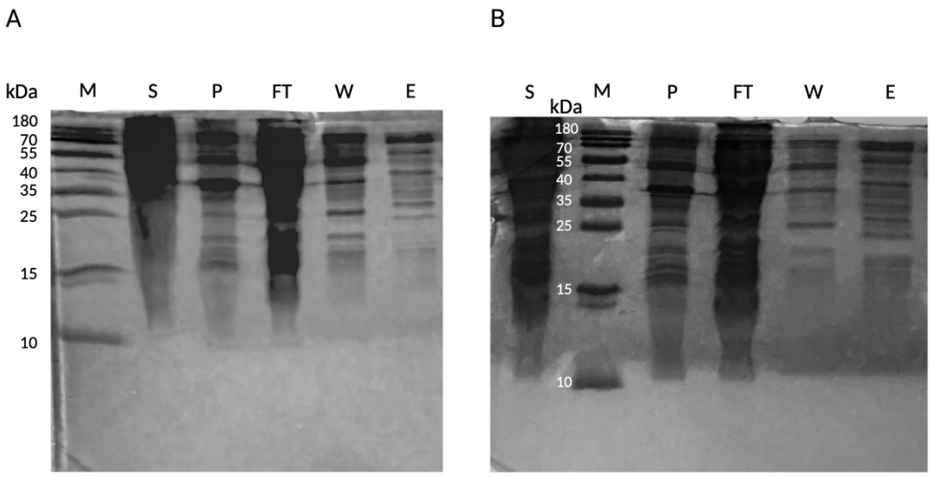

Figure 4.24 - SDS-PAGE gels of MEG 2.1 protein in BL21(DE3) strain – soluble fraction purification with Ni-NTA gravity column. A) in pET-22b(+) plasmid; B) in pET SUMO Champion plasmid. M - protein molecular weight marker, S - supernatant after cell lysis and centrifugation, P - pellet after centrifugation, FT - flow-through after Ni-NTA loading, W - wash of the column (5 column volume with loading buffer), E - elution with 100% of elution buffer. The expected MEG 2.1 isoform 1 molecular weight is 11.96 kDa for the pET-22b(+) plasmid and 21.04 kDa for pET SUMO Champion plasmid.

MEG 2.1 isoform 1 in the pET SUMO Champion plasmid did not express in either the soluble (Fig. 4.24) or insoluble (Fig. 4.25) fraction. The tested conditions for BL21(DE3), Rosetta(DE3) and BL21(DE3)pLysS expression strains were the same as above. Additionally, it was observed that BL21(DE3) cells transformed with pET SUMO Champion plasmid displayed a slower growth rate than when transformed with pET 22b(+) plasmid. For this plasmid in the Rosetta(DE3) strain, it was almost impossible to reach an OD_{600} of 0.6 even after 12 hours of media inoculation starting from a single bacterial colony.

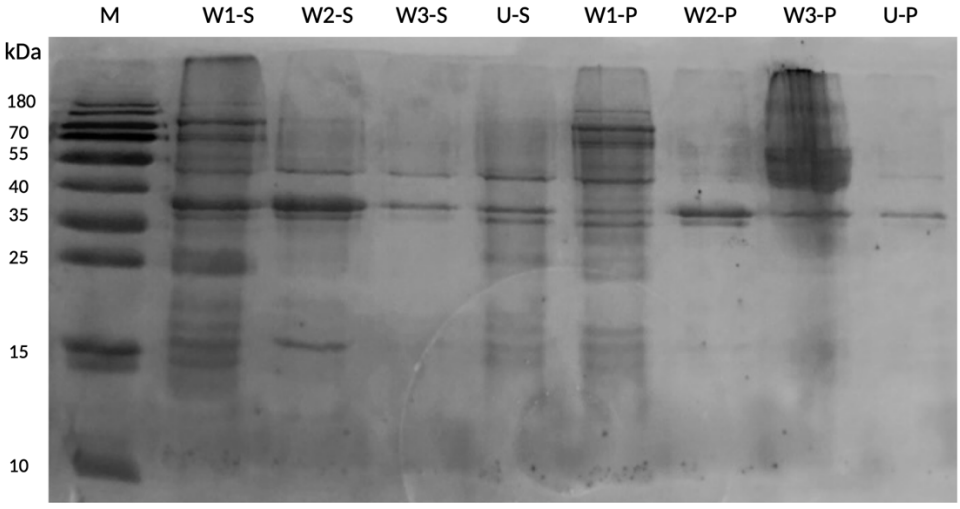

Figure 4.25 - SDS-PAGE gel of MEG 2.1 isoform 1 BL21(DE3) strain – insoluble fraction purification with Ni-NTA gravity column. W1-S; W2-S; W3-S; U-S – expression from pET SUMO Champion, expected MW= 21.04 kDa; W1-P; W2-P; W3-P; U-P – expression from pET 22b(+), expected Mw = 11.96 kDa. M - protein molecular weight marker; W1 - wash and sonication 2 M urea, 20 mM Tris/HCl pH 8, 0.5 M NaCl, 10 mM imidazole, 2% Triton X-100; W2 - wash and sonication in 2 M urea, 20 mM Tris/HCl pH 8, 0.5 M NaCl, 10 mM imidazole; W3 - wash and sonication in 20 mM Tris/HCl pH 8, 0.5 M NaCl, 10 mM imidazole; U - overnight dissolution in 8 M urea.

The above-mentioned tested conditions did not show any relevant change in protein expression, only a slower/faster growth of bacteria was observed. All the pilot expressions were performed in several repetitions, by gradually changing the protocol. The best modification was the addition of 20 mM glucose (final concentration) to the growth medium, which helped to grow the bacterial transformed with pET 22b(+) plasmid, however this had no improvement on MEG expression. In the case of expression according to the rifampicin protocol, glucose was added in generous amounts (10 g/L media, 55.6 mM). However, expression of MEG 2.1 isoform 1 was not detected either.

Furthermore, without the addition of glucose, a trend for a decrease in the OD_{600} of the culture after induction was observed, when induced between OD_{600} = 0.6 - 0.8. In the absence of glucose in the medium, the addition of any concentration of IPTG resulted in a decrease in OD_{600} in the following hours, which started to go back to growth after 2-3 hours (depending on the bacterial strain, plasmid and IPTG concentration, see Table 3.4 in the 3.2.2 Expression in bacteria in Methodology). This decrease was quite radical, in fact, the turbidity decreased from 0.8 to 0.4 one hour after induction. This fact, which was also observed for the MEG 3.2 isoform 1 protein, together with other reasons described below, led to the hypothesis that these two proteins could be toxic to bacteria.

This was the reason for using the BL21(DE3)pLysS, since it permits better control of the pre-induction leakage *via* the suppression of T7 RNA polymerase. The growth of bacteria with both pET plasmids in BL21(DE3)pLysS bacteria was significantly smoother (standard rapid growth, no reduction of OD_{600} after the IPTG induction), but the protein was still not expressed. After all conditions were tested, MEG 2.1 isoform 1 could not be expressed in any of the plasmids in any of the indicated bacterial strains. Given the relatively short sequence of this protein, in order to get experimental structural information by NMR, we decided to chemically synthesize it (see below, 4.3.1 Chemical synthesis of MEG 2.1 isoform 1, 2, and 3).

4.2.1.2 MEG 3.2 isoform 1

MEG 3.2 isoform 1 was also cloned into the above-mentioned two plasmids, pET 22b(+) and pET SUMO Champion. Two versions were cloned into the pET 22b(+) plasmid: a "long" (with the signal peptide for a total of 162 aa) and a "short" (without the predicted signal peptide, for a total of 125 aa). In the case of the MEG 3.2 short version, the transformation into expression strains never occurred, even though the transformation did occur in Stellar bacteria, which are a cloning strain. BL21(DE3) (Fig. 4.26), BL21(DE3)pLysS and Rosetta(DE3)pLysS were unsuccessfully tested for the transformation. This trend of non-transformation of the signal peptide-free protein construct was also observed for the MEG 2.1 construct used in the cell-free expression system (described later). For this reason, I continued expressions of MEG 3.2 isoform 1 protein only with the full-length construct (with signal peptide). From this point on, all

references to MEG 3.2 isoform 1 are only to the complete signal peptide-containing sequence.

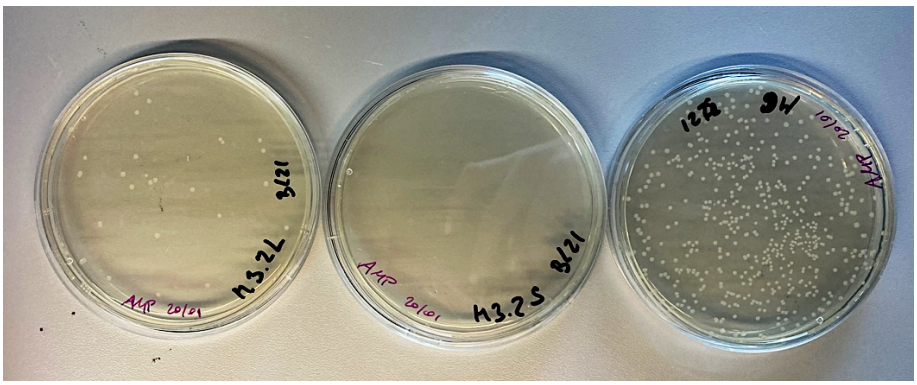

Figure 4.26 - MEG 3.2 isoform 1 in pET 22b(+) plasmid after the transformation into BL21(DE3) expression strain (from left to right): "long"- M3.2L (with the signal peptide; 162 aa), "short" M3.2S (without the predicted signal peptide - 125 aa) and MEG 6 MEG 6 (M6).

MEG 3.2 isoform 1 ("long" version with the signal peptide) cloned in pET 22b(+) plasmid was not expressed in either the soluble (Fig. 4.27) or insoluble fraction (Fig. 4.28), despite using 3 expression strains, 3 different temperatures of expression and 3 different IPTG concentrations (see 3.2.2 Expression in bacteria, Table 3.5 in Methodology).

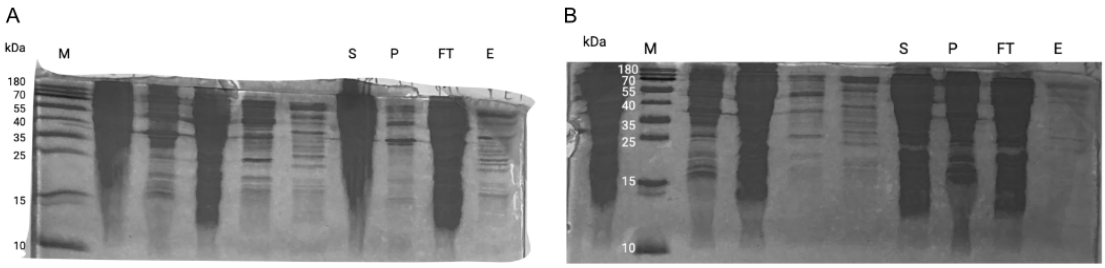

Figure 4.27 - SDS-PAGE gels after Ni-NTA gravity-flow purification of the soluble fraction of MEG 3.2 in the pET 22b(+) plasmid (A) and pET SUMO Champion plasmid (B) BL21(DE3) strain. M - marker, S - supernatant, P - pellet, FT - flow through, E - elution. The expected MEG 3.2 isoform 1 molecular weight is 18 kDa for the pET-22b(+) plasmid and 27 kDa for the pET SUMO Champion plasmid. The unlabeled bands in gel A belong to the MEG 2.1 protein in pET-22b(+) plasmid (next to the marker - supernatant, pellet, flow through, wash, elution); the unlabeled bands in the gel B belong to the MEG 2.1 protein in pET SUMO Champion plasmid (left from the marker - supernatant, right from the marker - pellet, flow through, wash, elution).

MEG 3.2 isoform 1 in the pET SUMO Champion plasmid did not express (Fig. 4.27) despite all the tentatives (see above or methods - 3.2.2 Expression in bacteria, Table 3.5 and 3.6). Even in this case, BL21(DE3) transformed with MEG 3.2 cloned into pET SUMO Champion plasmid grew slower than those transformed with pET 22b(+) plasmid. For this plasmid transformed in the Rosetta(DE3) strain, it was almost impossible to reach an OD_{600} of 0.6 after 11 hours of growth.

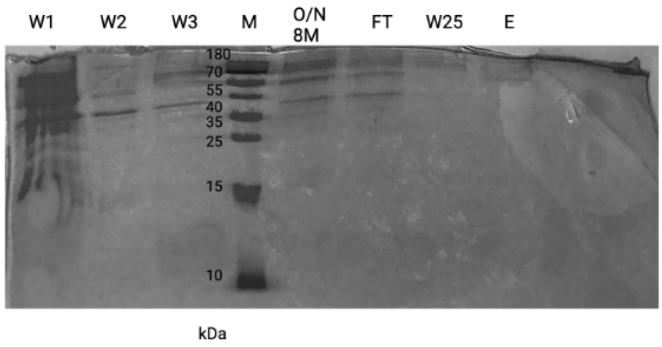

Figure 4.28 - SDS-PAGE gel after Ni-NTA gravity-flow purification of the insoluble fraction of MEG 3.2 in the pET 22b(+) transformed BL21(DE3) strain. W1 - wash and sonication 2 M urea, 20 mM Tris/HCl pH 8, 0.5 M NaCl, 10 mM imidazole, 2% Triton X-100; W2 - wash and sonication in 2 M urea, 20 mM Tris/HCl pH 8, 0.5 M NaCl, 10 mM imidazole; W3 - wash and sonication in 20 mM Tris/HCl pH 8, 0.5 M NaCl, 10 mM imidazole, M - marker, O/N 8M - overnight incubation in 8M urea buffer, FT - flow through, W25 - wash after overnight incubation, E - elution. Expected MEG 3.2 isoform 1 molecular weight is 18 kDa for pET 22b(+) plasmid.

The addition of glucose to the medium helped the growth, as in the case of MEG 2.1 isoform 1. The bacterial growth was faster for the transformed strain BL21(DE3)pLysS. Unfortunately, even the addition of 20 mM glucose to the media did not help to express this isoform. The only functional protocol for the expression of MEG 3.2 isoform 1 was the special expression protocol for potential toxic proteins, in the Rosetta(DE3)pLysS strain. In fact the addition of rifampicin, which is the most effective antibiotic inhibiting the transcription of bacterial RNA polymerase (Lama and Carrasco 1992; Du et al. 2021), blocks the translation of all the bacteria proteins except the protein of interest. Addition of rifampicin after IPTG induction also simplifies protein labeling for subsequent NMR structural studies. The addition of rifampicin results in selective labeling of only the target heterologous protein (Almeida et al. 2001). After successful growth of the bacteria expressing MEG 3.2 iso 1, purification using nickel affinity chromatography on FLPC was performed (Fig. 4.29A). Here, fractions that might contain MEG 3.2 isoform 1 protein were observed on the gel for the first time. Since we used a buffer at pH 8 and given that the theoretical MEG 3.2 isoform 1 pI is 5.8 (Annex A), ion exchange chromatography was tested (Fig. 4.29B). This method of purification did not prove to be suitable because all fractions containing the MEG 3.2 protein were found in the column flow-through. Therefore, purification was followed by size exclusion chromatography (Fig. 4.29C), which successfully separated the fractions. The concentration of pilot expression and purification of this protein was 0.3 mg/mL and the final volume was 2 mL.

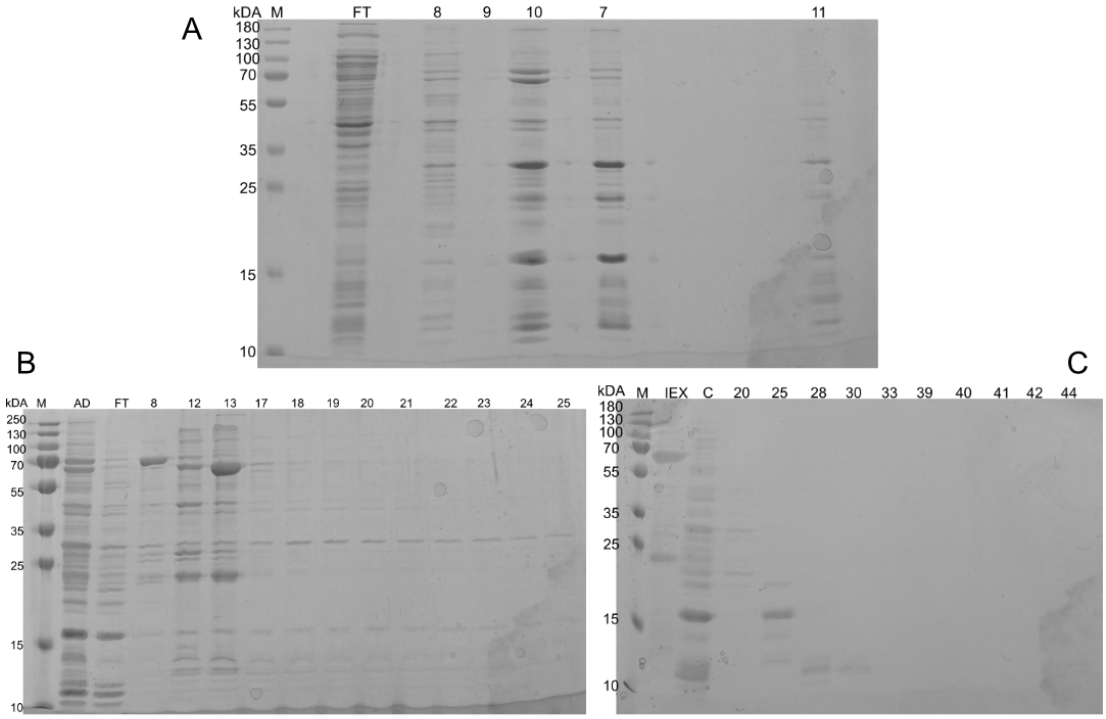

Figure 4.29 - SDS-PAGE gels of after FPLC purification of the soluble fractions of the MEG 3.2 protein the pET 22b(+) and Rosetta/pLysS strain after the addition of rifampicin (see Methodology, 3.2.2 Expression in bacteria). A - Ni-NTA FPLC purification, M - marker, FT - flow through, 7 - 11 - fractions with the protein deposited on the gel; B - Ion Exchange purification of the dialyzed fractions from the fist Ni-NTA purification (A), M - marker, AD - after dialysis, FT - flow through, 8 - 25 - fractions deposited on the gel; C - Size Exclusion purification of the previously purified protein via Ni-NTA and IEX FPLC (A, B), M - marker, IEX - sample loaded to the IEX, C - sample after IEX and concentration, 20 - 44 - fractions deposed on the gel. The expected MEG 3.2 isoform 1 molecular weight is 18 kDa for pET 22b(+) plasmid.

After this pilot expression and purification, a series of optimization expressions were performed to obtain the highest yield of pure protein. The whole process was repeated in 2 liters of LB medium and two different strains of *E. coli* were tested, BL21(DE3) (Fig. 4.30A) and Rosetta(DE3)pLysS (Fig. 4.30B). In the original protocol, Rosetta(DE3)pLysS expression bacteria were used and the test result showed that the growth is almost identical to the growth in the standard BL21(DE3) bacteria (Fig. 4.30A and 4.30B). After performing this test, partially purified protein from both bacterial strains was pooled into a common fraction after initial Ni-NTA purification (Fig. 4.30C), which was then loaded onto size exclusion chromatography (Fig. 4.30D and 4.30E).

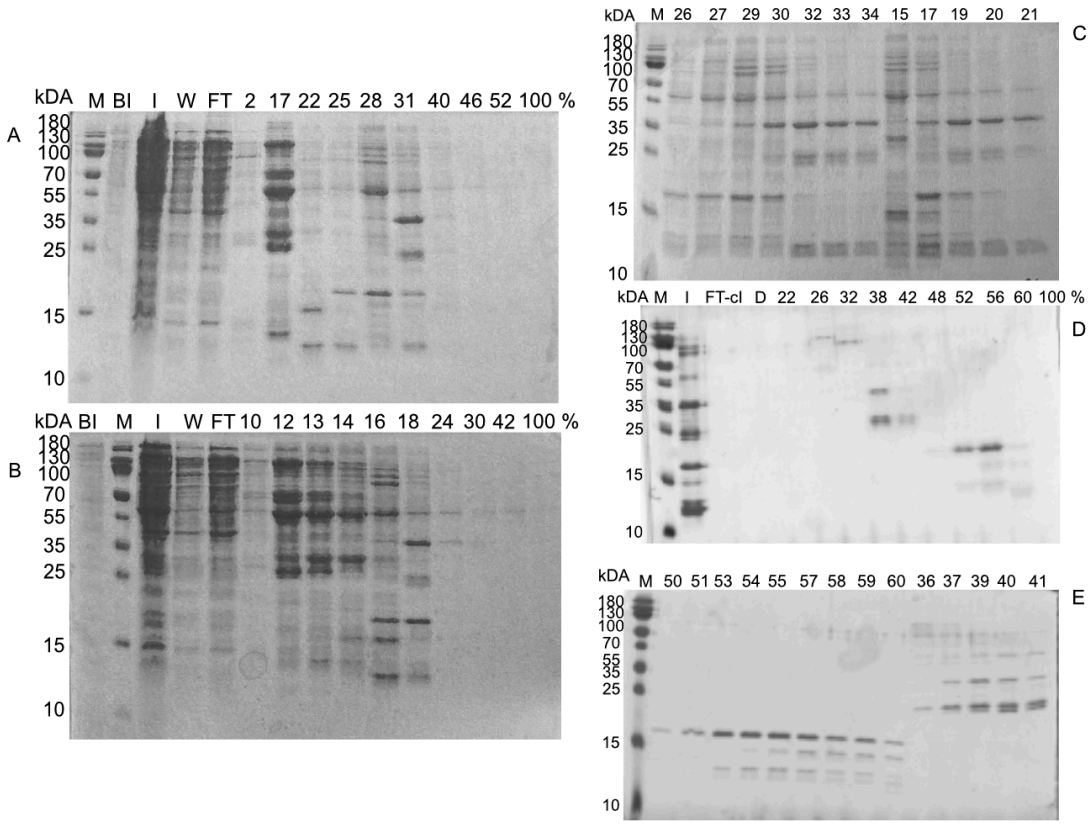

Figure 4.30 - SDS-PAGE gels of after FPLC purification of the soluble fractions of the MEG 3.2 protein expressed under rifampicin (see the methodology) in E. coli BL21(DE3) and Rosetta/pLysS strains. **A**. proteins expressed in *E. coli* BL21(DE3) after FPLC Ni-NTA purification; M - marker, BI - sample before induction, I - injection, W - wash, FT - flow through, 10 - 42 - fractions deposed on the gel, 100 % - elution with 0.5 M imidazole buffer; **B.** proteins expressed in *E. coli* Rosetta/pLysS after FPLC Ni-NTA purification; M - marker, BI - sample before induction, I - injection, W - wash, FT - flow through, 10 - 42 selected fractions deposed on the gel, 100 % - elution with 0.5 M imidazole buffer; **C.** fractions around 25-31 of gel (A) and 16-18 of gel (B) to decide which fractions to pool for subsequent purificaiton; **D.** SEC purification of pooled protein fractions (A, B and C) from both BL21(DE3) and Rosetta/pLysS strains after FPLC Ni-NTA; M - marker, I - injected sample, FT-c - flow through after concentration of the pooled fractions from previous affinity chromatography purifications (A and B), D - dialysis buffer, 22 - 60 - selected fractions deposed on the gel, 100 % - last fraction from the SEC purification; **E.** pooled protein fractions (A, B and C) from both BL21(DE3) and Rosetta/pLysS strains (FPLC Ni-NTA purified) after SEC purification; M - marker, 50 - 41 - selected fractions deposed on the gel. **Expected MEG 3.2 isoform 1 molecular weight is 18 kDa**.

After concentration of all the purest fractions from the purifications described above (lanes 50, 51 and 52 Fig. 4.30), the resulting concentration of non-labeled MEG 3.2 isoform 1 from 4 liters of media was 193 μM (Fig. 4.31 - fraction 2) in a volume of 250 μL. Therefore the final yield was 0.2 mg/L. This was also the highest concentration of MEG 3.2 isoform 1 protein expression achieved. At the same time, less pure fractions containing at least two other smaller proteins were concentrated (Fig. 4.31 - fraction 8), which, however, were very close in molecular weight to MEG 3.2 isoform 1 and were therefore practically impossible to remove from the sample; in addition, each SEC purification resulted in protein loss. A fraction with a presumed dimer (Fig. 4.32 - fraction 5) was also deposited on the gel, which however was ubiquitous in the purification of both MEG 3.2 protein and MEG 2.1 isoform 1 protein expressed in S2 cells (see below).

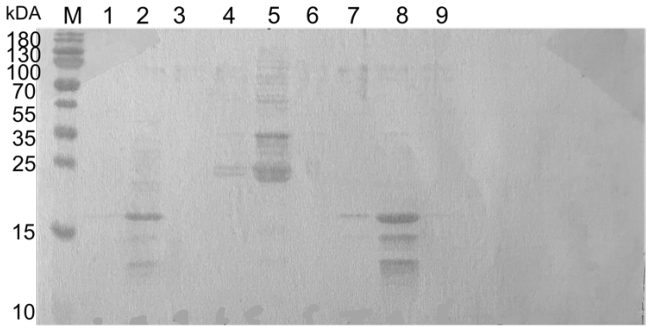

Figure 4.31 - SDS-PAGE gel of after FPLC Ni-NTA and SEC purifications of the soluble fractions of the MEG 3.2 protein after the final concentration of the purest fractions. M - marker, 1 - MEG 3.2 protein all fractions pooled from previous SEC purification (Fig. 4.30D and E) before concentration, 2 - MEG 3.2 protein all fractions pooled from previous SEC purification (Fig. 4.30D and E) after concentration, 3 - flow through from the concentration of MEG 3.2 fractions, 4 - MEG 3.2 hypothetical dimer before concentration, 5 - MEG 3.2 hypothetical dimer after concentration, 6 - flow through from the concentration of MEG 3.2 hypothetical dimer fractions, 7 - fractions 53 - 59 before concentration, fractions 53 - 59 after concentration, 9 - flow through from the concentration of fractions 53 - 59. Expected MEG 3.2 isoform 1 molecular weight is 18 kDa.

Afterwards, the stability of the protein in different buffers was tested. MEG 3.2 isoform 1 does not survive dialysis in MES buffer pH 8 (Fig. 4.32 - fraction 1), regardless of buffer concentration. Already during the pilot expression and first dialyses of this protein it was shown that it is stable only in high salt buffers; in fact, whenever NaCl concentration dropped below 0.5 M, a visible precipitation of the protein occurred. Since most SEC columns operate with a maximum NaCl concentration of 0.5 M, it was necessary to reduce the NaCl concentration in the SEC buffer twice (the buffer for Ni-NTA purification contained 1 M NaCl). This was another reason for the loss of the final protein concentration. Moreover, MEG 3.2 isoform 1 did not resist to one month storage at -20 °C (Fig. 4.32 - fraction 3). Even less pure fractions stored for a month in -20 °C did degrade (Fig. 4.32 - fraction 4). The hypothetical MEG 3.2 isoform 1 dimer had a longer lifetime, even though its concentration dropped after -20 °C storage (Fig. 4.32 - fractions 5 and 6). This dimer formation is only a hypothetical working hypothesis because all lysis and elution buffers that have been worked with contained 5 mM BME and at the same time all gels were run under reducing conditions. It is therefore excluded that the protein dimerized by forming disulfide bonds. On the other hand, this "dimer" formation was observed in both expressed proteins: MEG 3.2 isoform 1 in *E. coli* and MEG 2.1 isoform 1 in the S2 expression system. A progressive band formation of twice the size of the MEG 3.2 and 2.1 proteins was observed on the gels. MEG 2.1 protein even underwent almost complete aggregation after storage at -80 °C, as it is visible on the gels as a decrease in the concentration of the MEG 2.1 protein at the expected MW and an increase in the band at twice its size (described below).

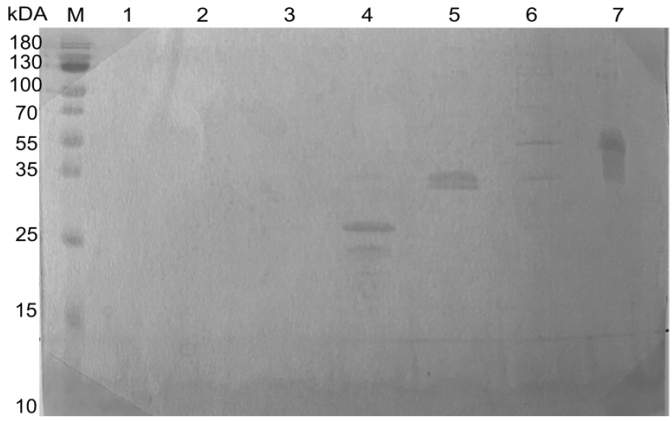

Figure 4.32 - SDS-PAGE gel of protein stability tests of purified MEG 3.2 and MEG 2.1 proteins. M - marker, 1 - MEG 3.2 in 10 mM MES buffer at pH 8, 2 - dialysis buffer after overnight dialysis, 3 - the purest fraction of MEG 3.2 after FPCL Ni-NTA purification – thawed after one month at -20 °C, 4 – partially purified fractions of MEG 3.2 thawed after one month at -20 °C, 5 - hypothetical dimer of MEG 3.2 thawed after two weeks at -20 °C, 6 - hypothetical dimer of MEG 3.2 –thawed after one month at -20 °C, 7 - MEG 2.1 protein from the S2 Drosophila expression system in 10 mM MES buffer 40 days after the dialysis (stored at -20 °C). Expected MEG 3.2 isoform 1 molecular weight is 18 kDa for pET-22b(+) plasmid and 15 kDa for MEG 2.1 protein isoform 1 without signal peptide in the pMT/BiP/SLIN plasmid.

Despite the low concentrations obtained with the rifampicin protocol, I tried to label MEG 3.2 with a ^{15}N isotope. Bacterial growth in minimal medium and nickel affinity chromatography were similar to the unlabeled protein expression. Unfortunately, the SEC purification again resulted in a large loss of protein on the column and its subsequent measured 1D ^{1}H spectrum was indistinguishable from the background signal. In a second attempt, taking advantage of the fact that rifampicin would allow ^{15}N labeling only of MEG 3.2, I performed only the partial purification with Ni-NTA, concentrated only the fractions containing the band at 18 kDa and measured 1D ^{1}H spectra. In this case, it was possible to distinguish the protein signals from the background, but the concentration was still not sufficient for 3D structure determination (Fig. 4.33 and 4.34). It is worth noticing that in order to maximize the concentration and to minimize the loss, we decided to use a high salt buffer, which is suboptimal for cryoprobe measurements.

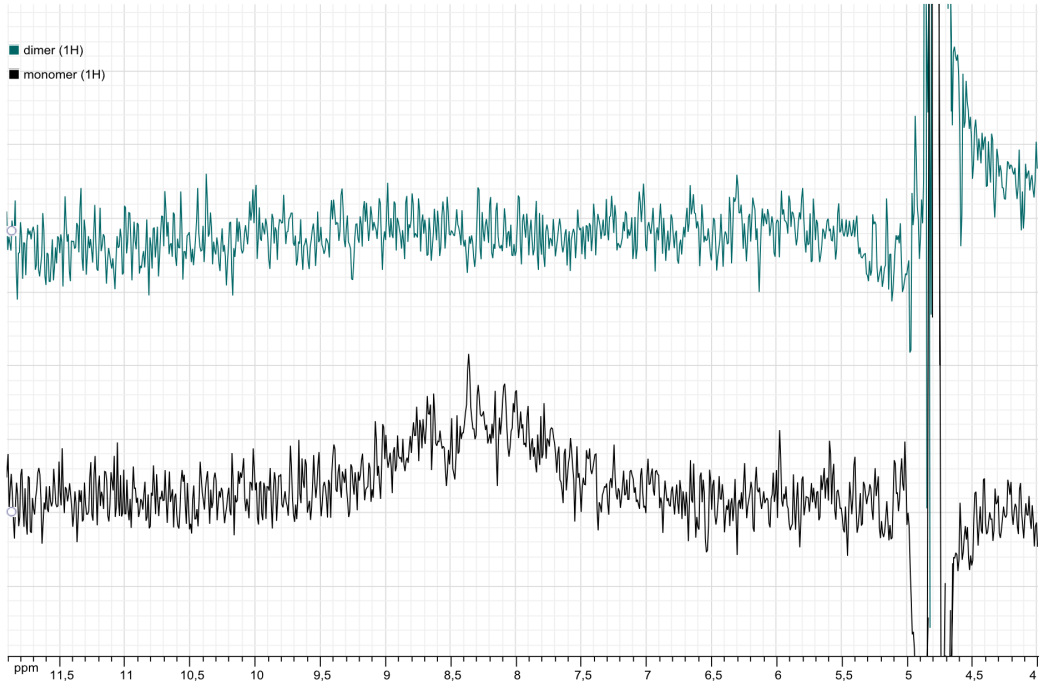

Figure 4.33 - Proton NMR spectra of MEG 3.2 isoform 1. First raw of 2D ^{1}H-^{15}N HSQC recorded for the dimer (cyan) and the monomer (black) MEG 3.2 isoform 1 expressed in *E. coli* expression system at a ^{1}H frequency of 600 MHz with a Varian spectrometer equipped with a triple HCN cryoprobe.

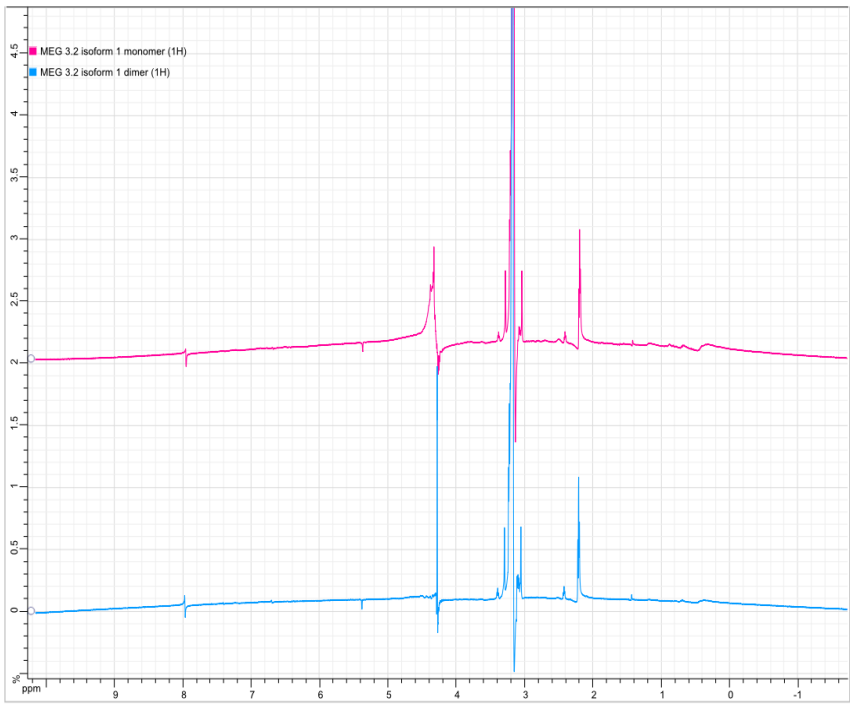

Figure 4.34 - Proton NMR spectra of MEG 3.2 isoform 1. 1D spectra with 128 scans recorded for the monomer (magenta) and dimer (blue) of MEG 3.2 isoform 1 expressed in *E. coli* expression system at a ^{1}H frequency of 600 MHz with a Varian spectrometer equipped with a triple HCN cryoprobe.

4.2.1.3 MEG 6

In my last year of research, MEG 6 protein, which does not contain any cysteine and has no predicted signal peptide, was cloned only into pET 22b(+) plasmid; however, the protein was not expressed in either the soluble or insoluble fractions (Fig. 4.35), despite all the tested conditions and strains, neither in 3 liters of medium nor by using rifampicin and the Rosetta(DE3)pLysS. Even low protein detection by western blot with anti-His-tag antibody confirmed the absence of MEG 6 expression.

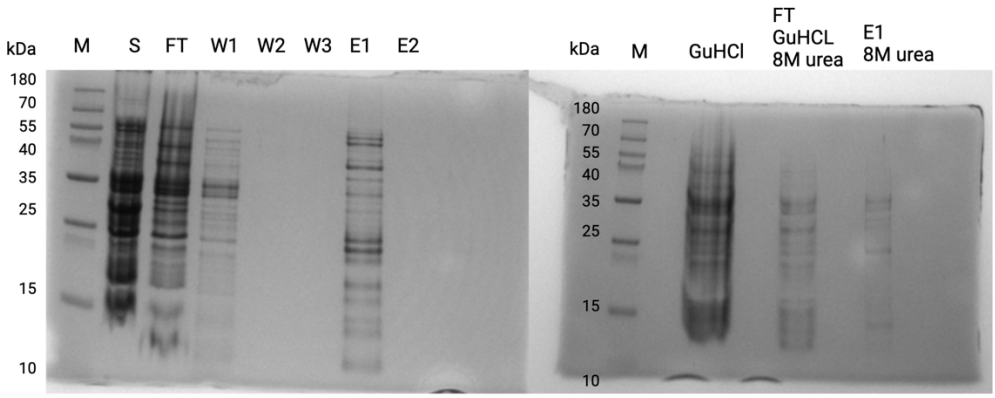

Figure 4.35 - SDS-PAGE gels after Ni-NTA gravity-flow purification of the soluble (A) and insoluble (B) fractions of the MEG 6 protein in the pET 22b(+) and BL21(DE3) strain. (A) M - marker, S - supernatant, FT - flow through, W1, W2, W3 - washes 1 - 3 with 20 mM Tris/HCl pH 8, 500 mM NaCl, 10 mM imidazole, E1, E2 - elutions 1 and 2. (B) M - marker, GuHCl - overnight incubation and sonication in 6 M guanidine hydrochloride, 20 mM Tris/HCl pH 8, FT Gu HCl 8 M urea - wash with 20 mM Tris/HCl pH8, 8 M urea, 0.5 M NaCl, 10 mM imidazole, E1 8M urea - elution with 20 mM Tris/HCl pH 8, 8 M urea, 0.5 M NaCl and 0.5 M imidazole. The expected MEG 6 molecular weight in the pET 22b(+) plasmid is 9.5 kDa.

4.2.2 Cell-free expression

Another procedure chosen to express MEG 2.1 isoform 1 and MEG 3.2 isoform 1 proteins was an *E. coli*-based cell-free expression system. This pilot expression was possible thanks to a grant application to use The Cell Free expression platform of the Institut de Biologie Structurale (Grenoble). The cell-free expression system was chosen because of previous repeated unsuccessful expressions in bacteria and yeast and because of the presumed bacterial toxicity of these proteins. Cell-free is a fast and efficient way of expression, even of challenging proteins such as toxic or membrane proteins, which at the same time allows relatively economical isotopic labeling for subsequent NMR structural analysis. Expression is carried out in bacterial extracts of *E. coli* BL21(DE3) with the addition of the mixture of amino acids, ribonucleotides, cofactors, inhibitors, detergents, and lipids. One advantage to work with cellular extracts is that the transcription and translation machineries are dedicated to the one and only plasmid (pIVEX in our case) inserted in the test tube. Given that the extract is from *E. coli*, the T7 promoter in the vector is essential for expression. Four constructs have been synthesized in the plasmid pIVEX2.4d, for gene fusion to the N-terminal 6xHistag. Two variants of MEG 2.1 isoform 1 and two variants of MEG 3.2 isoform 1 were synthesized (see Methodology chapter 3.2 Recombinant protein expression, Tables 3.1 and 3.2 for

details). Both proteins had a "short" variant (without the signal peptide) and a "long" variant (full length sequence). During transformation of synthetic plasmids into *E. coli* BL21(DE3) expressing bacteria, it was found that pIVEX2.4d-MEG 2.1 isoform 1 with the signal peptide ("M2.1L" - Fig. 4.36A and 4.36E) could not be inserted into the bacteria. This transformation was repeated in three experiments and re-tested with a newly synthesized plasmid. None of these experiments produced transformed colonies. Therefore, only three plasmids, pIVEX2.4-MEG 2.1S (without signal peptide) and both pIVEX4.2-M3.2iso1-S and pIVEX4.2-M3.2iso2-L, were sent to Grenoble for pilot expression. We later verified that the non-transforming pIVEX4.2-M2.1L with the signal peptide can be transfected into Stellar Competent cells, a cloning strain of bacteria. Therefore, this construct is probably so toxic that even minor leakage during the transformation of the plasmid into bacteria causes their death. At the same time, it was evident from all the transformations that even pIVEX4.2-M2.1S was much more difficult to transform because there were significantly fewer colonies on the plates, compared to the pIVEX4.2-M3.2 isoform 1 constructs. This trend was already noticed during bacterial expressions, where MEG 2.1 isoform 1 grew relatively slower than MEG 3.2 isoform 1, regardless of bacterial strain or tested expression conditions (see above and in the Methodology 3.2.2 Expression in bacteria, Table 3.4, 3.5, and 3.6).

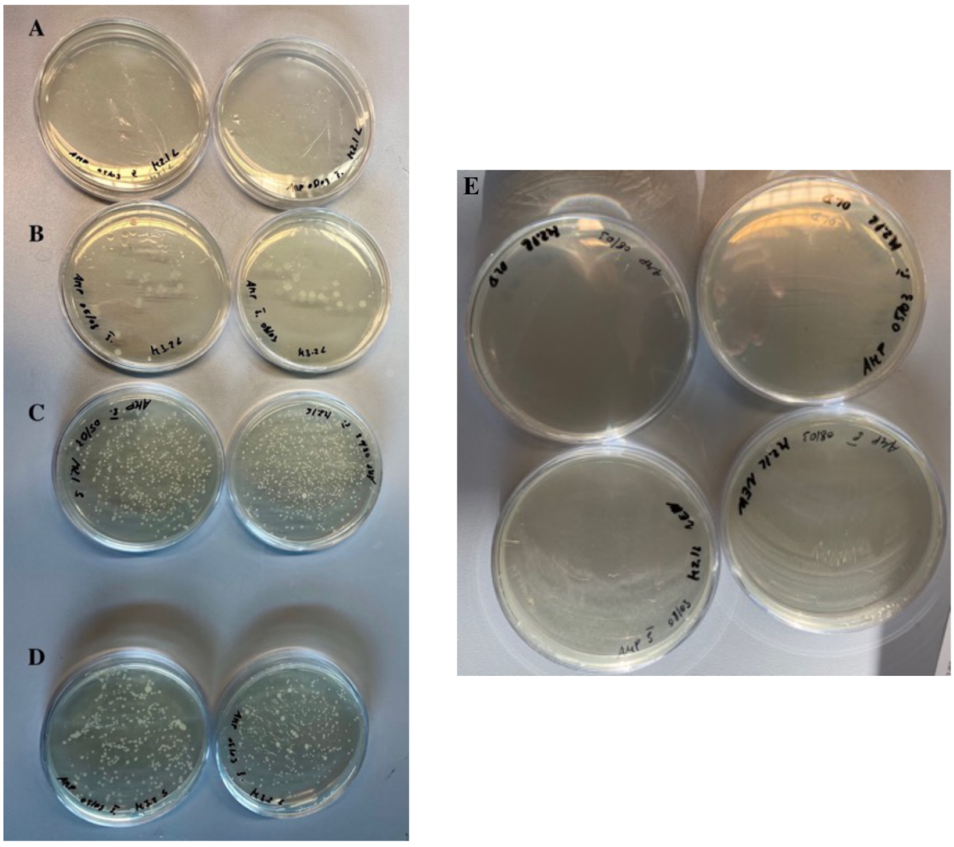

Figure 4.36 - E. coli BL21(DE3) transformed with cell-free expression pIVEX2.4d plasmids with MEG 2.1 isoform 1 and MEG 3.2 isoform 1 inserts (design detail in the supplementary data). A) construct pIVEX2.4d-TEV-PS-MEG2.1 (MEG 2.1 with signal peptide); B) construct pIVEX2.4d-TEV-PS-MEG3.2 (MEG 3.2 with signal peptide); C) construct pIVEX2.4d-TEV-MEG2.1 (MEG 2.1 without signal peptide); D) construct pIVEX2.4d-TEV-MEG3.2 (MEG 3.2 without signal peptide), E) repeated transformation of MEG 2.1 with signal peptide (construct pIVEX2.4d-TEV-PS-MEG2.1).

Pilot expression in the cell-free expression system did not yield positive results. None of the submitted constructs yielded proteins. Expressions were tested simultaneously with the control GFP protein, whose correct expression was confirmed on both gel and western blot (Fig. 4.37).

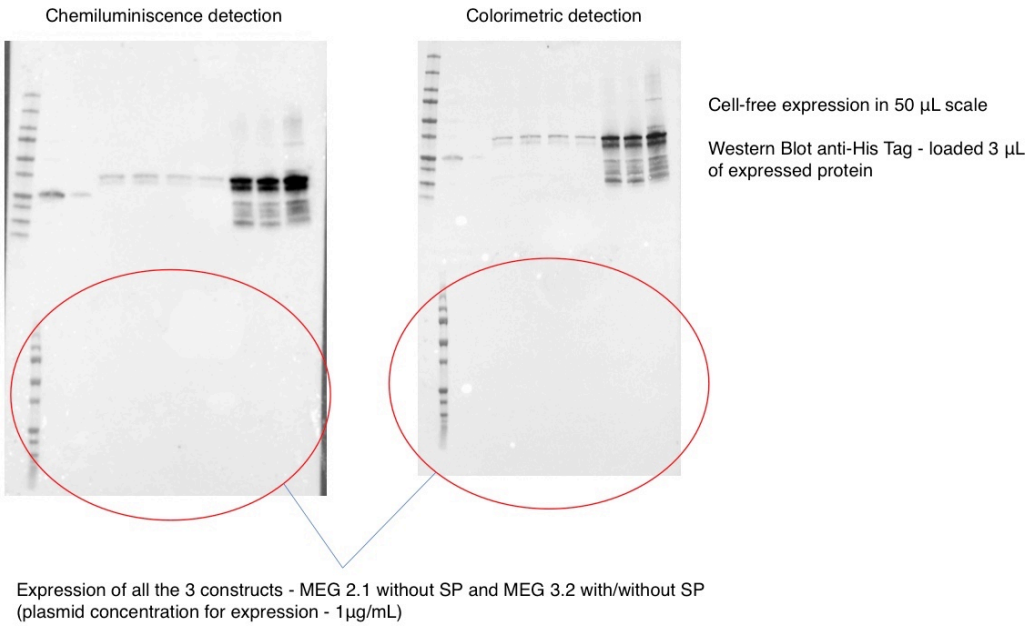

Figure 4.37 - Western Blot gels after cell-free expression of three constructs - MEG 2.1. isoform 1 without signal peptide and MEG 3.2 isoform 1 (with and without signal peptide)

4.2.3 S2 expression

S2 *Drosophila* expression system is an efficient system for large quantities of protein and at the same time it is one of the strategies to produce challenging target proteins, especially eukaryotic ones. Even in this system, there is the possibility to secrete the heterologous proteins into the media during expression, which is another benefit. For expression in S2 cells, three constructs were cloned into the pMT/BiP/SLIN plasmid (Barinka's lab, BIOCEV, Vestec), which contains an insect signal peptide for protein secretion into the medium (BiP) and a SLIN tag (StrepII - FLAG - TEV site - StrepII - TEV site). Three constructs were cloned: MEG 2.1 isoform 1 "short" (without signal peptide), MEG 3.2 isoform 1 with deletion of 16 amino acids (from N-terminus, V1) and deletion of 20 amino acids (from N-terminus, V2). These two different deletions corresponded to two versions of the predicted signal peptide for this isoform 1. Pilot and large volume expressions showed that neither of the short versions of MEG 3.2 isoform 1 protein can be expressed in S2 cells (Fig. 4.38B and 4.38C); however, MEG 2.1 isoform 1 without the signal peptide was indeed expressed (Fig. 4.38A), although at lower concentrations (Fig. 4.38A).

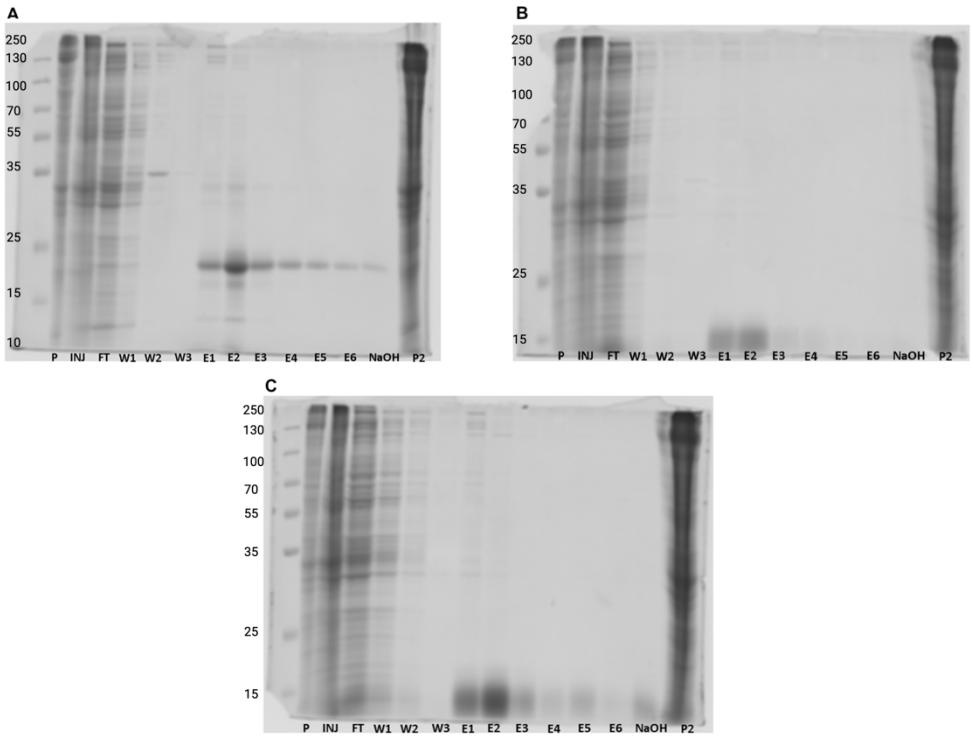

Figure 4.38 - SDS-PAGE gels after the gravity-flow column purification of MEG 2.1 isoform 1 (A), MEG 3.2 V1 (B) and MEG 3.2 V2 (C) proteins expressed in S2 cells. *P - pellet, INJ - injected sample, FT - flow-through, W1, W2, W3 - column washes 1 - 3, E1, E2, E3, E4, E5, E6 - elutions 1 - 6, NaOH - wash with NaOH, P2 - 5x concentrated pellet. The expected size of the MEG 2.1 protein without signal peptide (A) in the pMT/BiP/SLIN plasmid is 15 kDa, the expected size of the first version of MEG 3.2 protein without signal peptide (B) 21 kDa and of the second version of MEG 3.2 protein isoform without signal peptide (C) is 21 kDa.*

After gravity flow affinity chromatography, MEG 2.1 isoform 1 was further purified by size exclusion chromatography (Fig. 4.39), transferred/dialyzed to phosphate buffer and frozen to -80 °C.

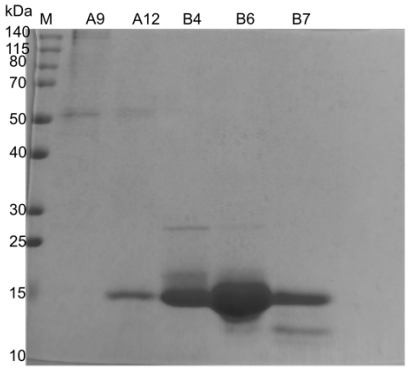

Figure 4.39 - SDS-PAGE gels after triple SEC purification of the MEG 2.1 isoform 1 protein without signal peptide. The purest fractions right after triple SEC purification.

After thawing and loading of all the fractions from the triple SEC purification on the gel, it was found that most of the fractions had aggregated into several forms (Fig. 4.40), the most abundant being a dimeric one. We tried to revert to the monomeric one by adding either DTT or 1 M NaCl (Fig. 4.40A), but degradation of both samples was evident (first two lanes in Fig. 4.40A).

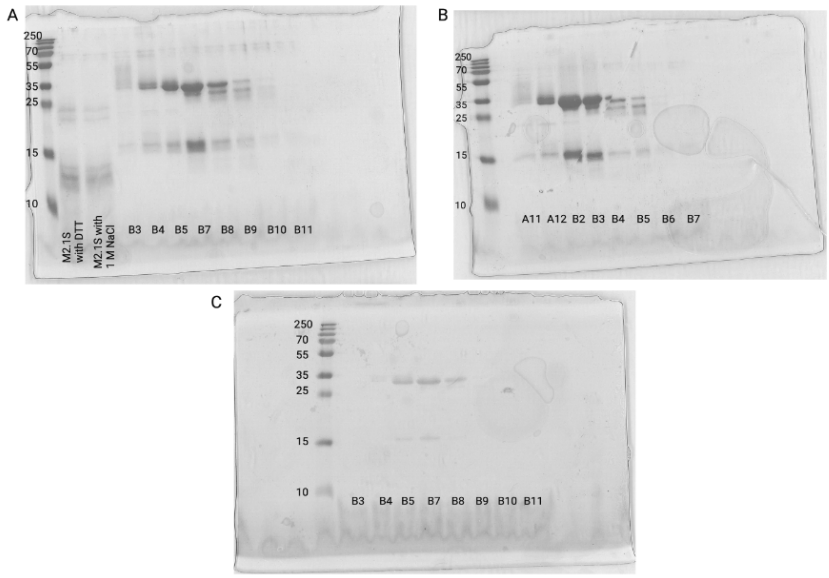

Figure 4.40 - SDS-PAGE gels after triple SEC purification of the MEG 2.1 isoform 1 protein without signal peptide - after -80 °C storage. *A* - MEG 2.1S with DTT, MEG 2.1S with 1 M NaCl and selected potent fractions after SEC purification applied to the gel. The expected size of the MEG 2.1 protein without signal peptide in the pMT/BiP/SLIN plasmid is 15 kDa.

After subsequent concentration of the purest fractions and their loading on the gel, it was evident that all the MEG 2.1 isoform 1 protein had aggregated from its monomeric form (size) to a fraction that was twice the original protein size (Fig. 4.41). Moreover, protein was lost upon concentration, confirming its intrinsic instability.

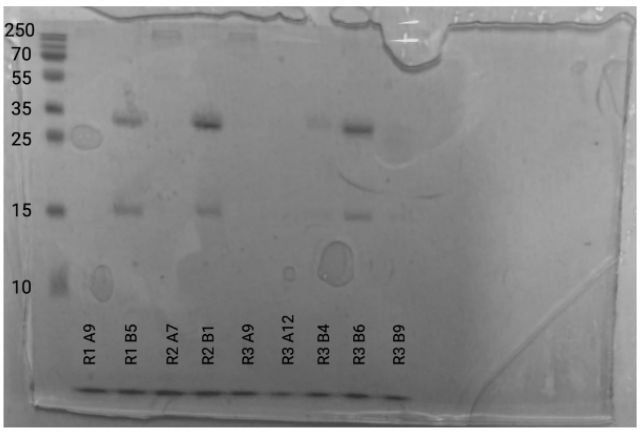

Figure 4.41 - SDS-PAGE gels after triple SEC purification of the purest concentrated fractions with MEG 2.1 isoform 1 protein without signal peptide - after -80 °C storage. *The expected size of the MEG 2.1 protein without signal peptide in the pMT/BiP/SLIN plasmid is 15 kDa.*

We confirmed that both bands were indeed MEG 2.1 isoform 1 protein expressed in insect cells by mass spectrometry (μLC-MS/MS). After the final pooling of all the samples, we obtained 1250 μl at 72 μM. This concentration was not sufficient for NMR structural analyses, but at least secondary structure measurements were made using circular dichroism (see below Fig. 4.44). Part of the sample was dialyzed from 50 mM sodium phosphate buffer at pH 8 to 10 mM MES buffer at pH 6 to perform CD measurements (Fig. 4.44). As can be seen from Fig. 4.32 (fraction 7), prolonged storage

of this protein in this buffer at -20 °C is not possible; complete aggregation occurred (band at about 3 times the expected MW).
Therefore the low yield, the instability of the product and the high cost of this expression system were the reasons for abandoning this strategy.

4.2.4 Yeast expression

MEG 2.1 isoform 1, MEG 3.2 isoform 1 and MEG 6 were cloned into *Komagataella phaffii* pPICZαB cloning vector for secretory methanol induced expression (EasySelect™ *Pichia Pastoris* Expression Kit For Expression of Recombinant Proteins Using pPICZ and pPICZα in *Pichia pastoris*). Pilot expressions did not yield a single positive result for any of the three proteins. Besides the negative results, yeast expression system also did not have an existing protocol for further isotopic labeling, and therefore there was no optimization and continuation with this express system.

4.3 Biophysical analysis

4.3.1 Chemical synthesis of MEG 2.1 isoforms 1, 2 and 3

Due to the above-described complications with obtaining a sufficient concentration of sufficiently pure and isotopically labeled proteins, we chose the strategy of synthesizing shorter peptides that would help in assembling the resulting structure of at least one component of the MEG 2 family. For the production of synthetic peptides by Genosphere Biotechnologies, the MEG 2.1 family was chosen because all its three isoforms have a complete and validated sequence, unlike the proteins of the MEG 3.2 family, and there is also a clear alternative splicing, unlike MEG 6, which has only one isoform. Moreover, the longest MEG 2.1 isoform 1 has a sequence of 88 amino acids, whereas the longest isoform of MEG 3.2 isoform 1 has a sequence almost twice as long, containing 156 aa (A0A5K4EPC8) or 145 aa (D7PD52), which would be considerably more challenging to synthesize.
The primary chosen strategy was the production of three complete synthetic isoforms, i.e., isoforms 1, 2 and 3 of the MEG 2.1 protein (Fig. 4.42). Due to the length of the first isoform with the signal peptide (88 aa), which would be difficult to synthesize, we decided to go for a version without the signal peptide (66 aa in length, hereafter named iso 1). Despite the sequences of MEG 2.1 iso 1 (64 aa), isoform 2 (34 aa) and isoform 3 (26 aa) were sent to be chemically synthesized at the same time, only iso 1 and isoform 3 were successfully synthesized. Synthesis of isoform 2 was repeatedly unsuccessful and after several months of attempts was terminated by Genosphere. For this reason, we decided to split it into two peptides (iso 2a - 18 aa and iso 2b - 16 aa).
After initial NMR analyses of iso 1, it was obvious that it would not be possible to perform a complete assignment of all 66 amino acids. For this reason, it was divided into three peptides: iso 1a (19 aa), iso 1b (17 aa) and iso 1c (31 aa).

Finally, isoform 1c was divided, because it turned out that even 31 amino acids (with significant repeats in the sequence) could not be completely assigned. This isoform was therefore split into two further peptides, namely iso 1f (15 aa) and iso 1g (16 aa).

Figure 4.42 - Design of nine synthesized peptides for structural analysis of the MEG 2.1 family (isoforms 1, 2 and 3) using NMR. Individual peptides are designed named as follows: isoform 1 (iso 1, 66 aa sequence without signal peptide - magenta), isoform 1a (iso 1a, 19 aa sequence - orange), isoform 1b (iso 1b, 17 aa sequence - lime green), isoform 1c (iso 1c, 31 aa sequence - violet), isoform 1f (iso 1f, 15 aa sequence - bubble gum pink), isoform 1g (iso 1g, 16 aa sequence - royal blue), isoform 2a (iso 2a, 18 aa sequence - red), isoform 2b (iso 2b, 16 aa sequence - dark green), isoform 3 (iso 3, 26 aa sequence - dark cyan).

4.3.2 Sample solubility

A major challenge that had to be overcome in the process of biophysical analysis of the synthetic peptides was their very poor solubility in practically all commonly used solvents. In the first place, the buffer used for the preparation was the one used for the purification of recombinant proteins: 20 mM Tris/HCl, pH 8.5 with 500 mM NaCl. To 400 µL of this buffer was added 1 mg of the synthetic iso 1 and 100 µL of D_2O. The peptide was quite visibly insoluble in this buffer and after measuring the first 1D ^{1}H spectrum it was obvious that this buffer would not be suitable for solubilization. Next, a sample with 400 µL of the same buffer was tested with the progressive addition of DMSO-d6 (from 100 to 400 µL). The addition of DMSO-d6 partially helped solubility, but even a 1:1 ratio of Tris/HCl and DMSO-d6 did not result in complete solubility of the peptide. The last test was the dissolution of the synthetic peptide in 100 % DMSO-d6, which led to its complete solubilization. All synthetic peptides were further dissolved in DMSO-d6 to maintain identical conditions of analysis.

Unfortunately, DMSO-d6 is completely unsuitable as a solvent for circular dichroism measurements. For this reason, a variety of solvents were tested that would be suitable for CD analysis and at the same time would keep the synthetic peptides in solution. Solvents that were tested for compatibility with CD analyses were: chloroform, acetone, methanol, ethanol, isopropanol, acetic acid, trifluoroethanol (TFE) and acetonitrile. Acetonitrile, methanol, isopropanol, chloroform, and acetic acid are solvents recommended for the solubilization of poorly soluble peptides. The only solvent that did not interfere with the CD measurement, by not absorbing between 180 and 250 nm, and at the same time that was able to partially solubilize the synthetic peptides was acetonitrile. However, even in acetonitrile, the peptides were not completely soluble (even after stirring overnight), therefore the samples were prepared by leaving them for several hours, then centrifuged and the supernatant was collected for CD analysis. The samples prepared in this way had a concentration between 1 - 10 µM. The addition of 1:1 TFE to acetonitrile increased their solubility.

4.3.3 Circular dichroism

Despite all the above-described challenges with the solubility of the peptides and the search for a suitable solvent, it was possible to record CD spectra of all peptides of all three isoforms of the MEG 2.1 family (Fig. 4.43). The first peptide measurement was performed in 100% acetonitrile and the second measurement was performed in 1:1 acetonitrile and TFE. TFE was chosen because it's known to increase the helicity of the protein/peptide (Myers, Nick Pace, and Martin Scholtz 1998) and we wanted to verify the stability of the secondary structure (or its changes) in both solvents. MEG 2.1 iso 1 in acetonitrile (first graph in Fig. 4.43A) showed a positive ellipticity value at 190 nm, while between 208 and 222 nm it had a slightly negative ellipticity value, indicating the presence of helical regions. The addition of 50% TFE to the sample enhanced this helicity trend (first graph in Fig. 4.43B). The presence of helix in the structure was also observed for isoform 3 dissolved in 100 % acetonitrile (the last graph in Fig. 4.43A). Indeed isoform 3 is a predicted signal peptide of MEG 2.1 isoforms 1 and 2, and simultaneously is strongly hydrophobic (highest GRAVY index of all compared MEG proteins - Annex A. The measured spectra of both peptides of MEG 2.1 isoform 2 (two middle graphs in the Fig. 4.43A) were relatively noisy and both in 100 % acetonitrile showed only a slight trend of one negative peak in the 220 nm region, indicating the nature of the disordered protein. This single negative peak around 220 nm was enhanced by the addition of 50 % TFE (two middle graphs in Fig. 4.43B), which did not push the structure towards the helix, but only enhanced the solubility of these peptides. Thus, there is no doubt that the two derived peptides of MEG 2.1 isoform 2 possess a disordered structure.

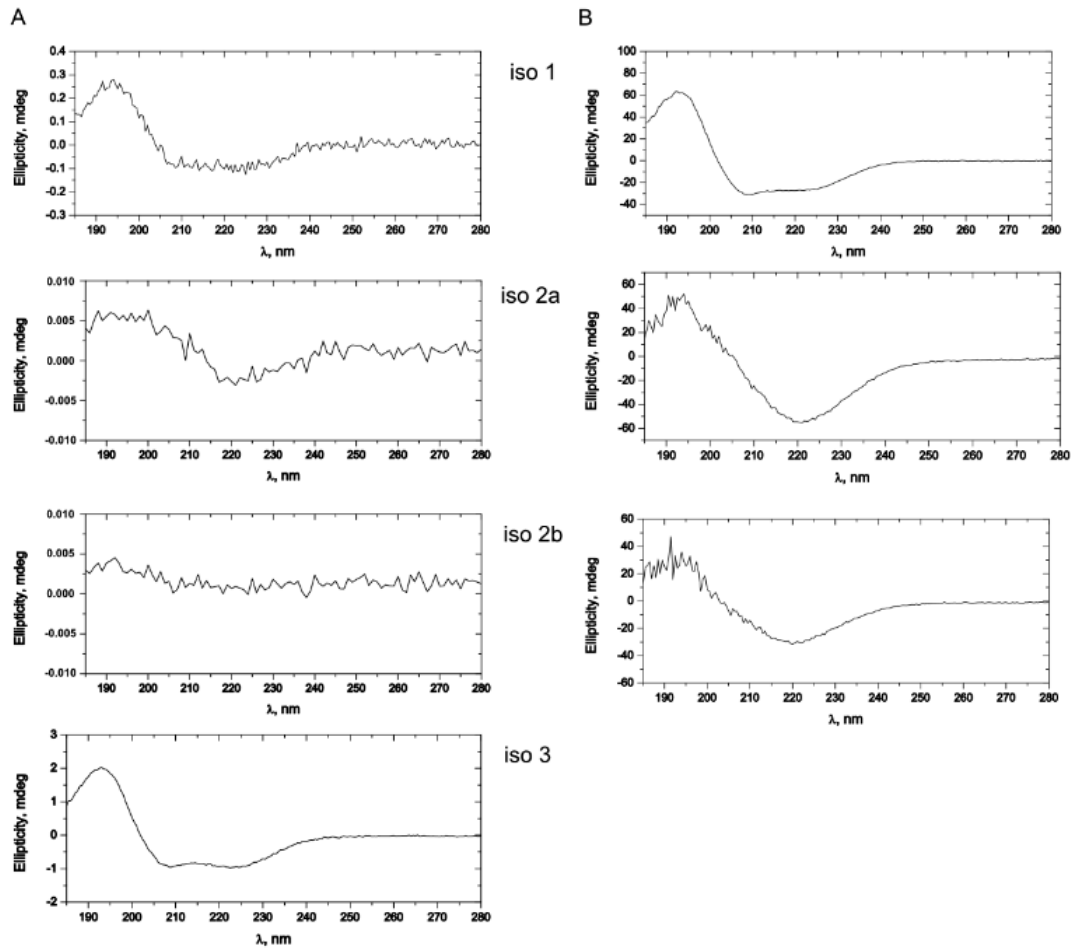

Figure 4.43 - CD spectra of MEG 2.1 iso 1, 2a, iso 2b, and iso 3 recorded at 25 °C in 100 % acetonitrile (A) and 50 % acetonitrile + 50 % TFE (B). Samples were prepared from the chemically synthesized peptides resuspended in acetonitrile and/or acetonitrile/TFE; spectra were recorded for the supernatant after centrifugation. The initial peptide concentration was set to 10 - 20 µM.

As previously mentioned, the expression of isoforms in S2 and *E. coli* with rifampicin did not give a sufficient yield for NMR analysis, but CD spectra could be recorded. Indeed, the CD spectrum of MEG 2.1 isoform 1 protein (without signal peptide but with SLIN tag) expressed in S2 cells shows the presence of alpha helix in the structure (Fig. 44), even though the helix is less obvious than in the case of the CD spectrum for isoform 1 without the signal peptide with 50 % TFE:50% acetonitrile (Fig. 4.43, panel B for iso 1).

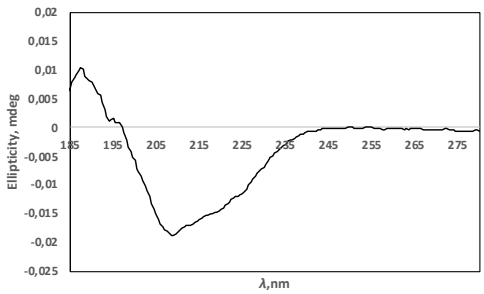

Figure 4.44 - CD spectra of MEG 2.1 isoform 1 protein without signal peptide expressed in S2 cells, purified, and dialyzed in 10 mM MES buffer at pH 6. The protein concentration was 24 µM.

From the spectra measured for MEG 3.2 isoform 1 protein in 50 mM Tris/HCl buffer at pH 8 (Fig. 4.45A) and 10 mM MES buffer at pH 6 (Fig. 4.45B) it was not possible to determine the secondary structure. This could be caused either by a low concentration of the protein or by its instability in buffers without NaCl. In the Tris/HCl buffer, the oversaturation of the signal in the region between 210 - 205 nm is visible and in the MES buffer, there is a noticeable degradation of the sample. Measurement in MES buffer resulted in slightly negative values in the whole spectral range, so it is possible to assume that the concentration of the sample was too low or the protein without the presence of NaCl precipitated and sedimented to the bottom of the CD cuvette.

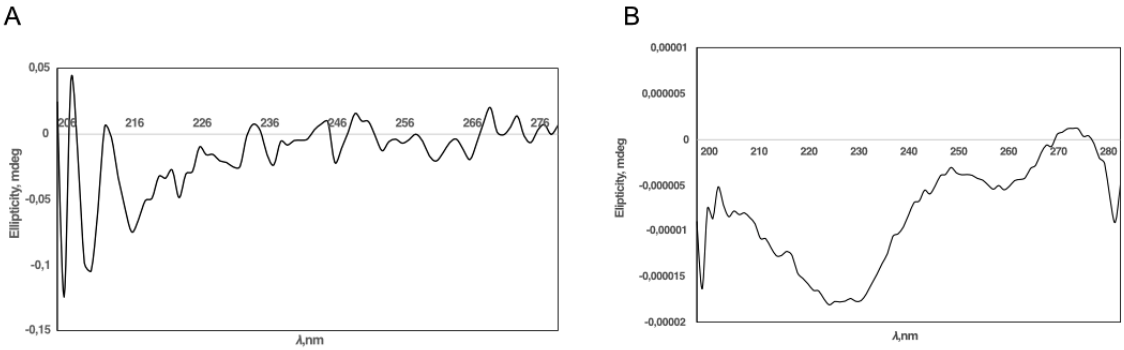

Figure 4.45 - CD spectra of MEG 3.2 recombinant protein in 50 mM Tris/HCl buffer pH 8 (A) 1 µM and 10 mM MES buffer pH 6 (B) 5 µM.

4.3.4 Dynamic Light Scattering (DLS)

We tried to understand the quaternary assembly of MEG 3.2 isoform 1 purified from rifampicin-blocked bacterial cells, by means of DLS. The size distribution profile is presented in Fig. 4.46. These measurements indicated the presence of two populations of particles in the sample. The larger particles with a hydrodynamic diameter of 470.6 nm were significantly more abundant (92.4 % of the volume) and a small fraction of the sample volume was made up of particles with a diameter of 93.74 nm. The polydispersity index (Pdi) has a value of 0.521, indicating that this sample is polydisperse, but still measurable by DLS (Danaei et al. 2018). Before DLS measurement, MEG 3.2 isoform 1 had to be dialyzed from 1 M NaCl buffer to 200 mM NaCl buffer, as previously mentioned, lowering the ionic strength led to almost immediate visible precipitation. Therefore, the result is not surprising: the more abundant population of larger particles is likely to be the precipitated form of an aggregated protein, despite the presence of a reducing agent and of glycerol in the buffer. In fact, the expected MW of this isoform is 21 kDa, which should correspond to a hydrodynamic diameter below 10 nm.

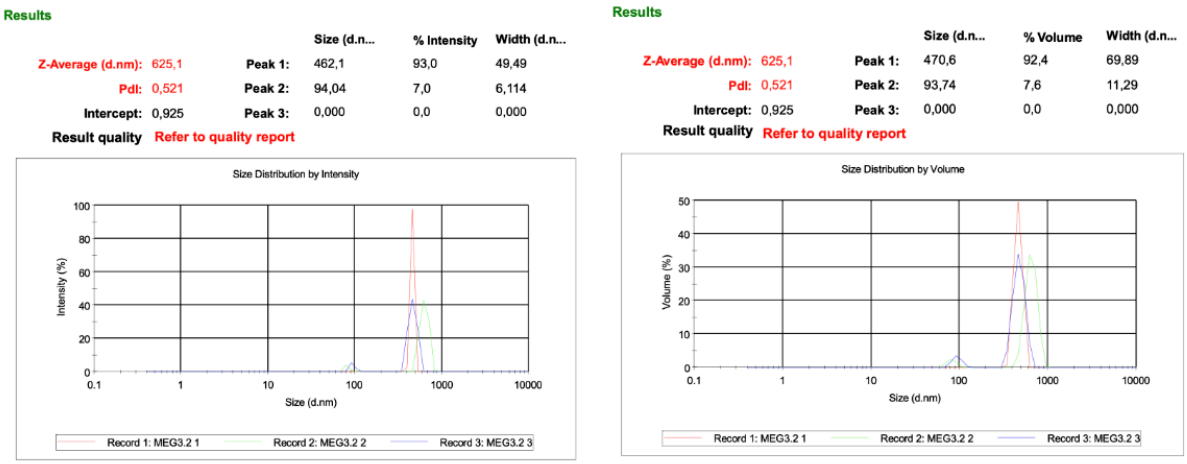

Figure 4.46 - DLS distribution of MEG 3.2 protein expressed in E. coli, purified in 50 mM Tris/HCl pH 8, 200 mM NaCl, 10 % glycerol, 5 mM BME buffer. Protein concentration was 0.32 mg/ml.

4.3.5 Nuclear magnetic resonance

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful technique used in the determination of molecular structures. It provides information about the 3D arrangement of atoms within molecules, the arrangement of molecules in space, molecular dynamic, molecular interactions, etc. NMR spectroscopy exploits the magnetic properties of atomic nuclei containing an odd number of protons and/or neutrons. Proteins are mainly composed of hydrogen, carbon, and nitrogen atoms. To fulfill the NMR rules, isotopic labelling is mandatory for carbon (^{13}C) and nitrogen (^{15}N) since ^{12}C has no spin value (I = 0) and ^{14}N is a quadrupolar nucleus (I = 1) that complexify the NMR data analysis. When placed in a strong magnetic field, these nuclei absorb and emit energy at characteristic frequencies, revealing information about their chemical environment. The same type of nuclei in different electron environment will behave differently because of the shielding effect of electrons and this difference in their environment is defined as the chemical shift.

The chemical shifts observed in an NMR spectrum provide insights into the local electronic environment of each nucleus. In proteins and peptides, these shifts are sensitive to factors such as the amino acid type, secondary structure, hydrogen bonding, and solvent interactions. By analyzing the chemical shifts, it is possible to determine the types of atoms present and gain initial information about the molecular structure. After the transformation of the obtained signals from NMR analyses, it is possible to proceed to the assignment of specific resonances to individual atoms within the molecule. Based on these assignments and on the known geometrical constrains of each individual amino acid, the 3D structure of the molecule is then built; in the case of peptides and proteins, the structure can be reconstructed using automatic structural calculation software (Barron 2015; Jacobsen 2007; Balci 2005).

For the structural analysis of all the synthetic peptides a combination of homonuclear (zTOCSY, NOESY) and heteronuclear (^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, ^{1}H-^{13}C HSQC-TOCSY)

experiments was used. Bidimensional (2D) NMR experiments were interleaved with monodimensional (1D) ^{1}H experiments to check the stability of the peptide. Verification of stability (and possible detection of peptide degradation) was important because of the duration of the measurements. To increase the signal to noise ratio (SNR), of our samples, in which ^{13}C (1.1%) and ^{15}N (0.4%) abundance was the natural one, the measurements could last up to 7 full days. All the synthetic peptides were measured at a concentration of 2 mM in DMSO-d6, at a temperature of 27 °C with the use of Varian Inova spectrometer operated at a ^{1}H frequency of 600 MHz (14.1 T) equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Isoform 1 (iso 1, without signal peptide) and isoform 3 (iso 3) were simultaneously measured on the Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) equipped with a triple HCN cryoprobe.
Another challenge that had to be overcome after the solubilization of the synthetic peptides was the significant signal intensity of DMSO-d6 (2.54 ppm) (Babij et al. 2016) and H_2O in DMSO-d6 (3.33 ppm). A ^{1}H 1D spectrum was recorded for each peptide and the peaks of DMSO-d6 and H_2O in DMSO-d6 were determined. The peaks thus determined were then presaturated with 6 dB intensity for homonuclear experiments and 12 dB for ^{1}H-^{13}C HSQC and ^{1}H-^{13}C HSQC-TOCSY experiments. For some isoforms the signals of DMSO-d6 and H_2O in DMSO-d6 were very intense and made it difficult to assign peaks in the region around 2.7 ppm and 3.3 ppm. This was the case of amino acids with beta protons (HBs) lying in these regions, such as Asn, Asp and Cys. Despite the fact that the spectra were measured in the natural abundance of ^{13}C and ^{15}N, we obtained almost complete assignments for the ^{1}H, ^{15}N and ^{13}C chemical shifts for all peptides except for iso 1 (MEG 2.1 isoform 1 without signal peptide, Fig. 4.47) and isoform 1c (iso 1c, Fig. 4.60). The ^{1}H-^{15}N HSQC spectrum of iso 1 showed only 30 assignable peaks out of 64 aa (Fig. 4.47) and the ^{1}H-^{15}N HSQC spectrum of iso 1c showed only 13 peaks out of 31 aa (Fig. 4.60). For this reason, these two isoforms were divided into several smaller peptides: iso 1 was divided into iso 1a, iso 1b and iso 1c; iso 1c was further divided into iso 1f and iso 1g (Fig. 4.47). Peak lists of all the peptides described below are in the tables in Annex C.

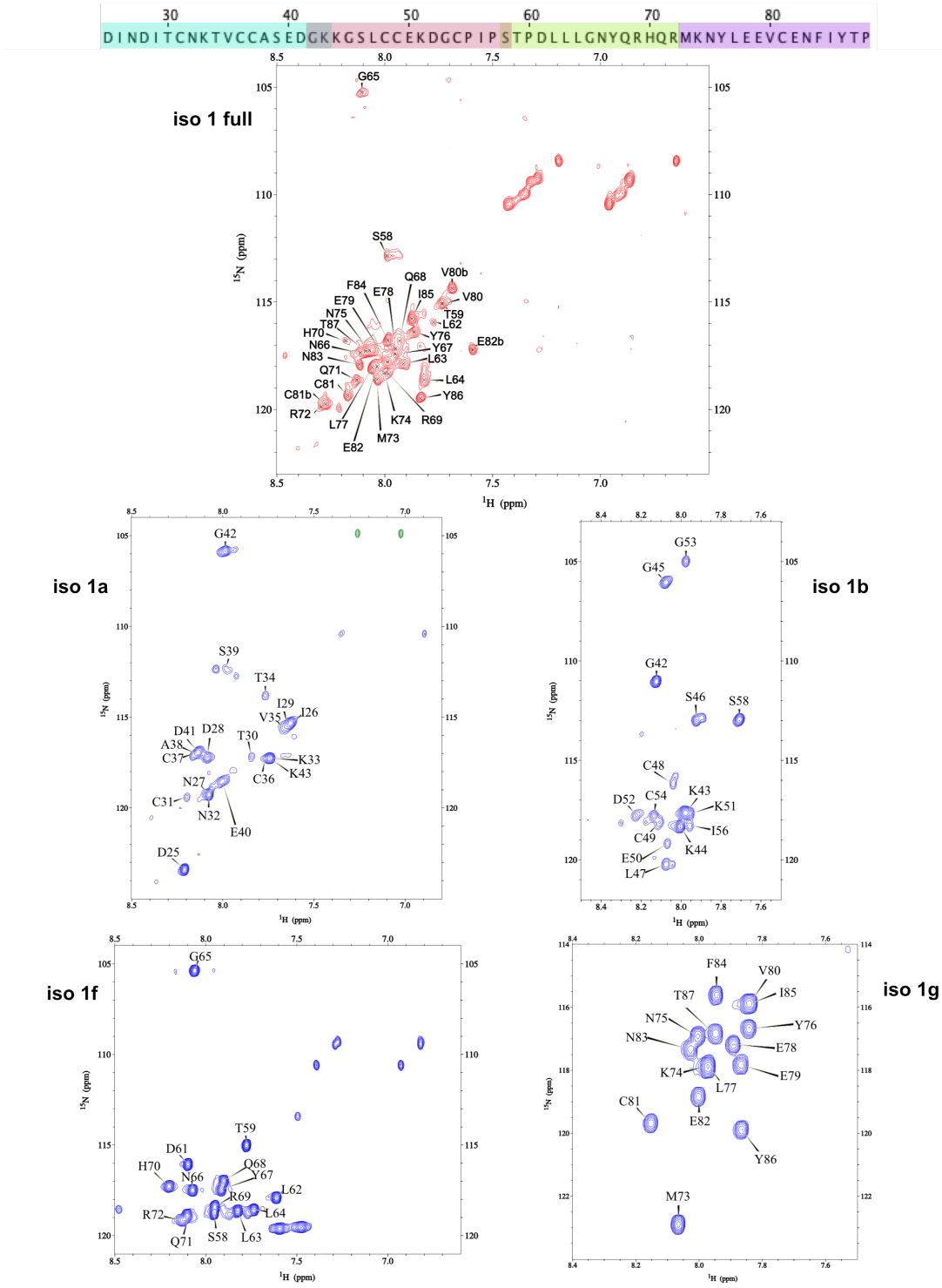

Figure 4.47 - 2D ^{1}H-^{15}N HSQC experiment of MEG 2.1 isoform 1 without SP (64 aa, D25-P88), iso 1a (residues D25-K43, sequence cyan), iso 1b (17 aa, residues G42-S58, sequence red), iso 1f (15 aa, residues S58-R72, sequence green) and iso 1g (16 aa, residues M73-P88, sequence violet) peptides in DMSO-d6 at a concentration of 2 mM. All the experiments have been recorded at 27 °C with a Bruker Neo spectrometer operating at a ^{1}H frequency of 1.2 GHz for the isoform 1 (25-88) and with Varian Inova spectrometer operating at a ^{1}H frequency of 600 MHz for iso 1a, iso 1b, iso 1f, and iso 1g. The two spectrometers are equipped with a triple HCN cryoprobe. The residues numbering is displayed in each spectrum. The sequences are displayed at the bottom of the figure.

As can be seen from Fig. 4.47, the splitting of the uncut isoform 1 into shorter peptides allowed the complete assignment of NH correlations. Thus, for the ^{1}H-^{15}N HSQC assignment, 19 out of 19 aa of iso 1a (Fig. 4.47), 15/17 aa (2 Pro) of iso 1b (Fig. 4.47), 14/15 aa (1 Pro) of iso 1f (Fig. 4.47) and 15/16 aa (1 Pro) of iso 1g (Fig. 4.47) were identified. Indeed proline residues cannot be detected in this experiment. At the same time, it should be noted that all the synthetic peptides resonances are located in a very narrow spectral range from 7.6 to 8.3 ppm, which not only makes their assignment more challenging but also points to their intrinsically disordered protein (IDP) character (Dyson and Wright 2021).

As already mentioned, efforts of the Genosphere company to synthesize isoform 2 in its complete sequence range (without the signal peptide, 52 aa) were terminated after three unsuccessful attempts. This resulted in the division of this isoform into two peptides (iso 2a and iso 2b, Fig. 4.48). The ^{1}H-^{15}N HSQC experiment allowed the complete ^{15}N assignment of 18/18 aa of iso 2a (Fig. 4.48) and 15/16 aa of iso 2b (Fig. 4.48), since here the last amino acid is a proline, that cannot be detected in this experiment. Even these two peptides have resonances in a very narrow spectral range from 7.6 to 8.2 ppm; however, the peaks were less clustered in only one part of the spectrum, as it was the case for all the studied peptides.

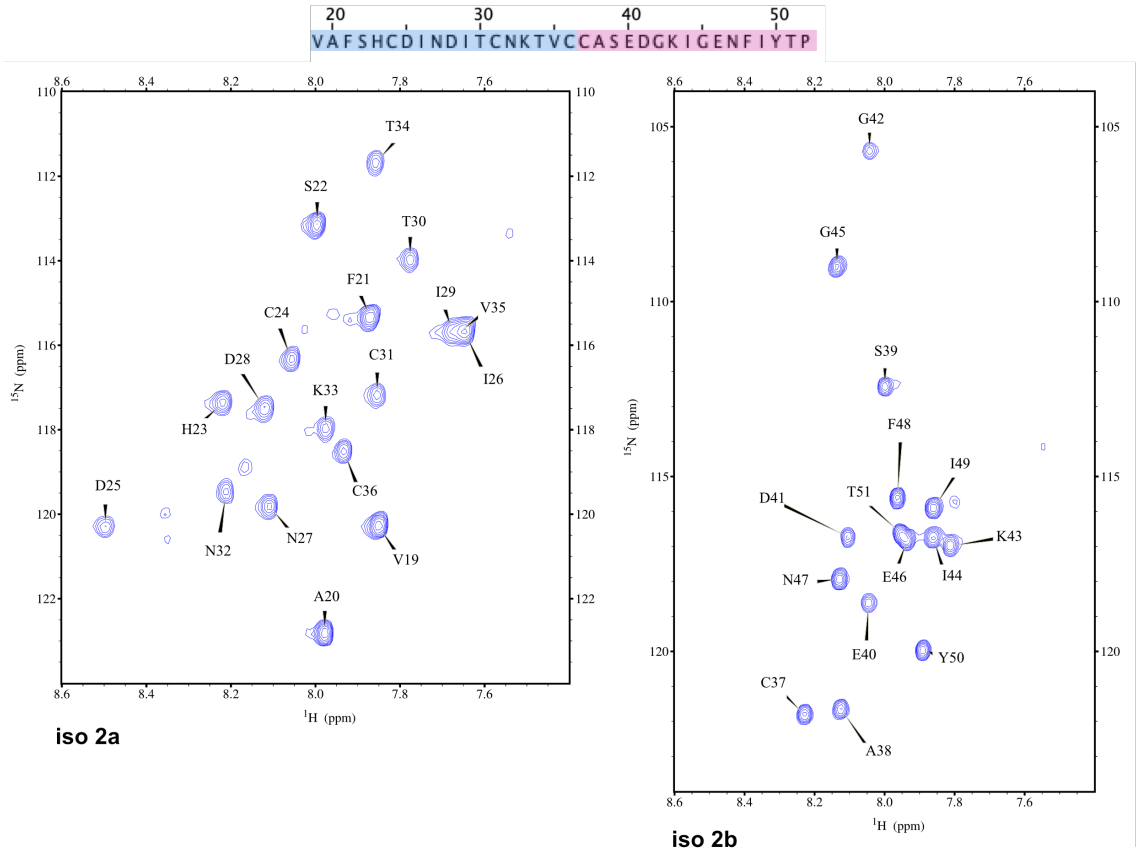

Figure 4.48 - 2D ^{1}H-^{15}N HSQC experiment of MEG 2.1 isoform 2a (18aa, residues V19-C36, sequence blue) and iso 2b (16aa residues C37-P52, sequence pink) peptides in DMSO-d6 at a concentration of 2mM. All the experiments have been recorded at 27 °C with a Varian Inova spectrometer operating at a 1H frequency of 600 MHz and equipped with a triple HCN cryoprobe. The residues numbering is displayed in each spectrum. The sequences are displayed at the bottom of the figure.

Isoform 3, which is composed of 26 amino acids, is predicted as a signal peptide and is common to all the MEG 2.1 isoforms. It contains a high number (7) of leucine residues in the sequence, which increases its hydrophobicity (Annex A). The ^{1}H-^{15}N HSQC spectrum assignment was successfully performed for 25/26 aa since the last amino acid is a proline (Fig. 4.49).

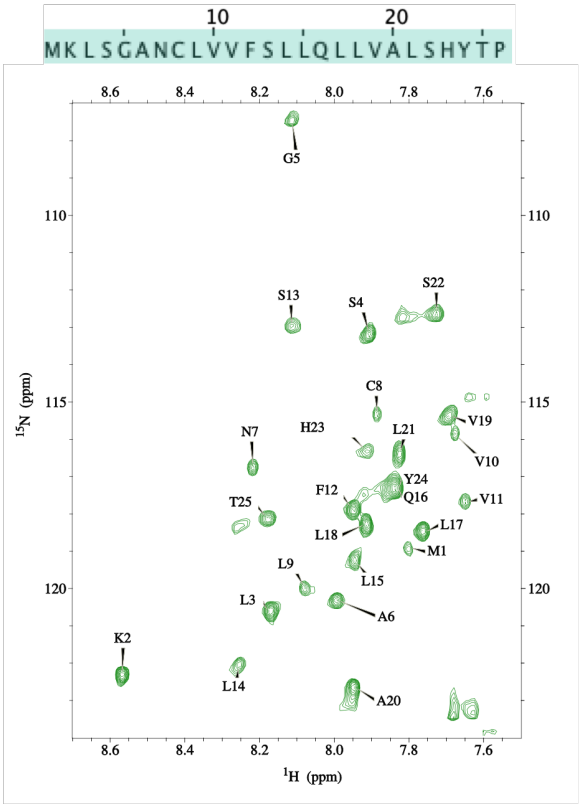

Figure 4.49 - 2D ^{1}H-^{15}N HSQC experiment of MEG 2.1 isoform 3 peptide (26 aa, residues M1-P26) in DMSO-d6 at a concentration of 2mM. The experiment has been recorded at 27 °C with a Varian Inova spectrometer operating at a ^{1}H frequency of 600 MHz and equipped with a triple HCN cryoprobe. The residues numbering is displayed in the spectrum. The sequences are displayed at the bottom of the figure.

The assignment of MEG 2.1 iso 1 (64 aa) was quite challenging. As already mentioned, only 30 amino acids could be identified in the NMR spectra. In order to obtain higher resolution and sensitivity to achieve the most complete possible assignment, NMR experiments were performed not only on Varian Inova spectrometer operating at a ^{1}H frequency of 600 MHz (14.1 T), but also in a magnetic field twice as strong on the Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) and equipped with a triple HCN cryoprobe (Fig. 4.50 and 4.51). From these spectra, overlay is obvious that the measured resonances overlap well, but unfortunately, the increase in the strength of the magnetic field did not lead to the acquisition of a larger number of resonances. All measured spectra contained only half of the theoretical resonances, regardless of the instrument.

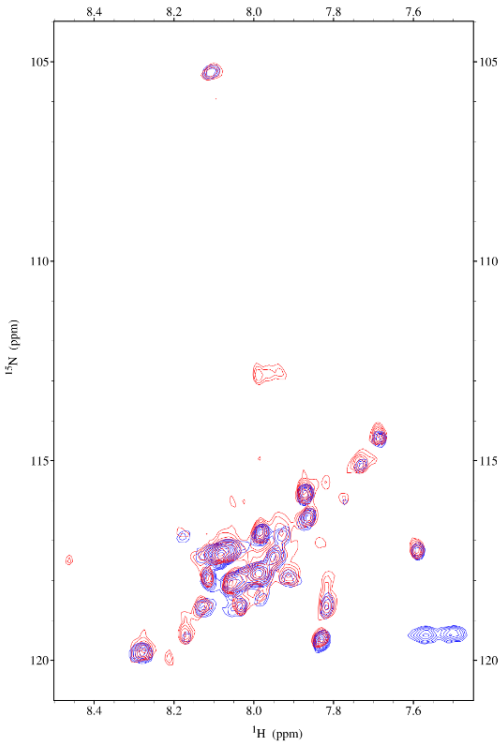

Figure 4.50 - Overlay of ^{1}H-^{15}N HSQC spectra recorded at a 1H frequency of 1.2 GHz (red) and 600 MHz (blue) for the isoform 1 at 2 mM dissolved in DMSO-d6. Experiments have been recorded at 27 °C with Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) and equipped with a triple HCN cryoprobe and with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

At the same time, all resonances were grouped within 0.6 ppm, which makes not only the aforementioned ^{1}H-^{15}N HSQC assignment difficult nor the ^{1}H-^{13}C HSQC (Fig. 4.51), but also the ^{1}H-^{1}H NOESY assignment (Fig. 4.52), which is crucial for determining the 3D structure.

In this narrow region of the spectrum, it was difficult to distinguish individual resonances in the ^{1}H-^{1}H TOCSY spectrum, so it was even more difficult to distinguish resonances and noises in the ^{1}H-^{1}H NOESY spectrum. Therefore, it was decided to proceed with the assignment of the 4 peptides issued from this longer one.

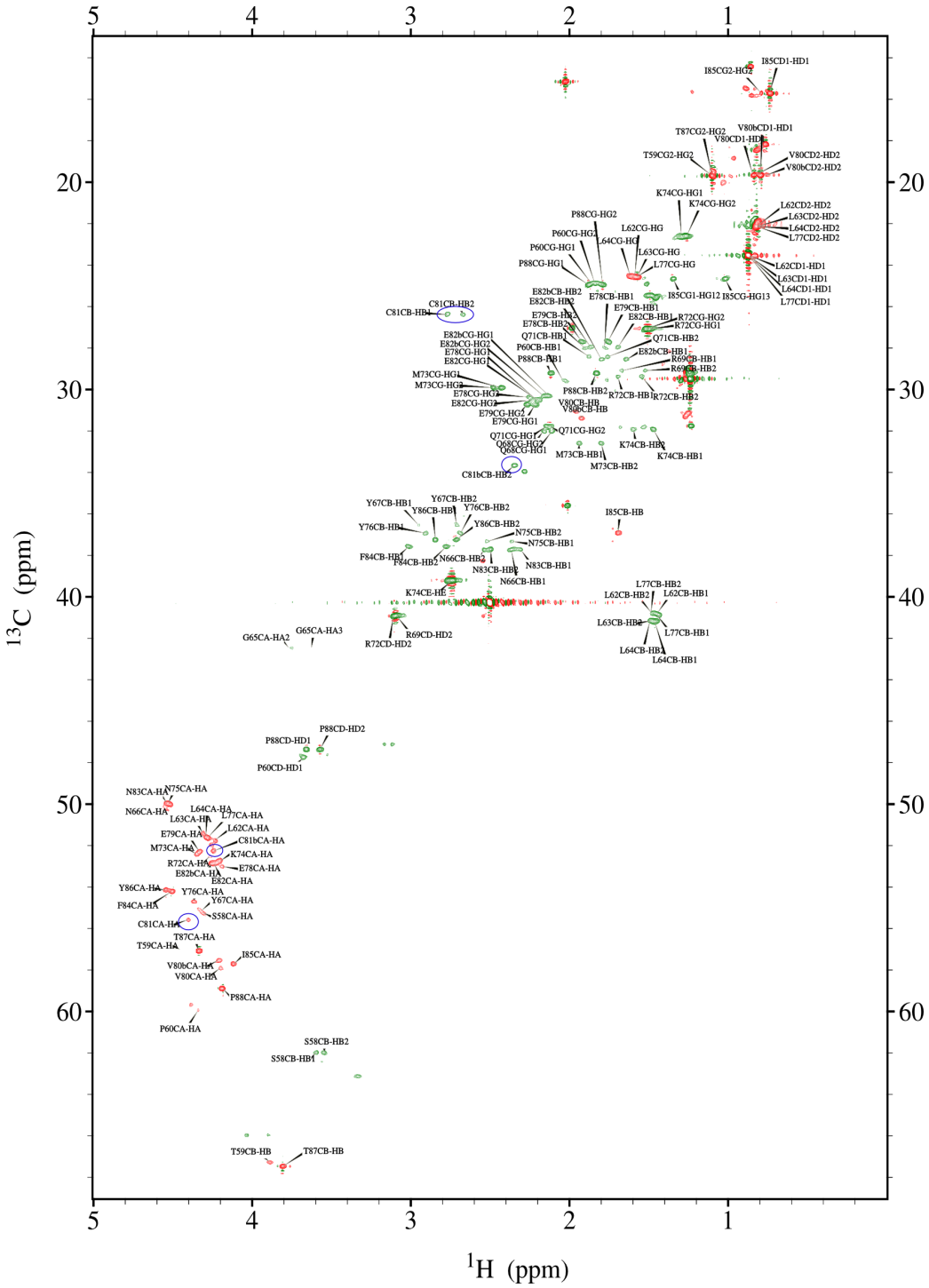

Figure 4.51 - Assignments of 2D ^{1}H-^{13}C HSQC of MEG 2.1 isoform 1 peptide without SP (64 aa, residues D25-P88) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) and equipped with a triple HCN cryoprobe.

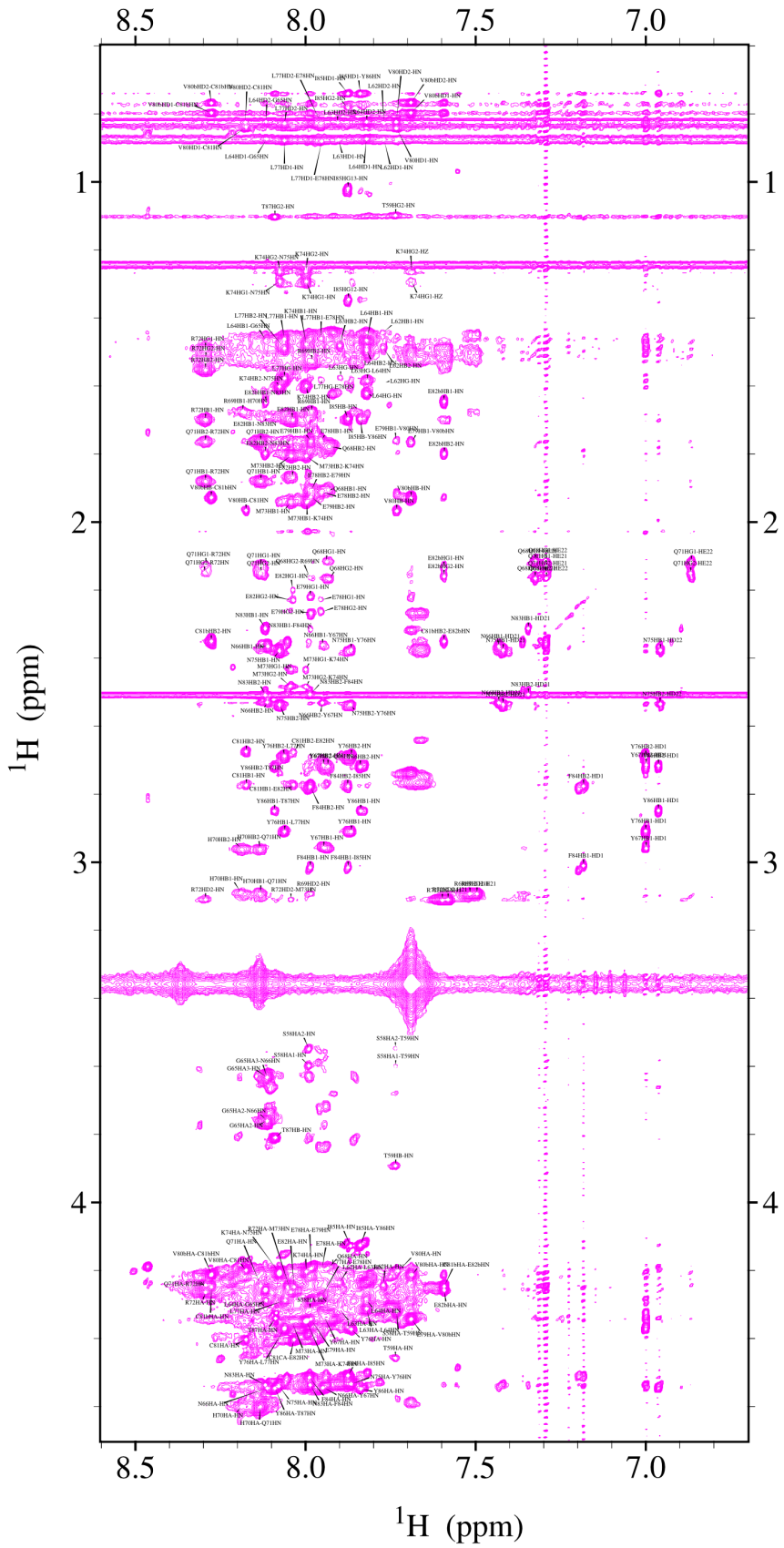

Figure 4.52 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 120 ms) of MEG 2.1 isoform 1 peptide without SP (64 aa, residues D25-P88) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) and equipped with a triple HCN cryoprobe.

Given the high proportion of cysteine in the sequence, it is reasonable to ask if disulfide bonds are formed. We wanted to test this theory by reducing these putative bonds. For this purpose, we added tris(2-carboxyethyl)phosphine (TCEP) to the sample of isoform 1 in DMSO-d6. Unfortunately, in the course of this experiment, the sample degraded, so this hypothesis could not be confirmed or refuted. At the same time, due to the high concentration of the sample, the formation of disulfide bonds between the individual peptides could not be excluded.

The 1.2 GHz spectrometer was also used to test the hypothesis of stabilization of the structure of isoform 1 by Zn^{2+} ions (zinc finger) in the region from C36, C37 to C48, C49. This sequential arrangement of cysteine could resemble the Cys-Cys-X_{10}-Cys-Cys structure of a putative zinc finger motif. Therefore, 3 mM $ZnCl_2$ was added to the 2 mM iso 1 peptide in DMSO-d6 sample and ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC spectra were recorded. These spectra were then overlayed with spectra without the addition of $ZnCl_2$ and both spectra were found to be completely superimposed (Fig. 4.53). From this result, it can be concluded that the formation of the zinc finger in isoform 1 does not occur. Other possible structural elements of isoform 1 are described below (in the structure refinement section).

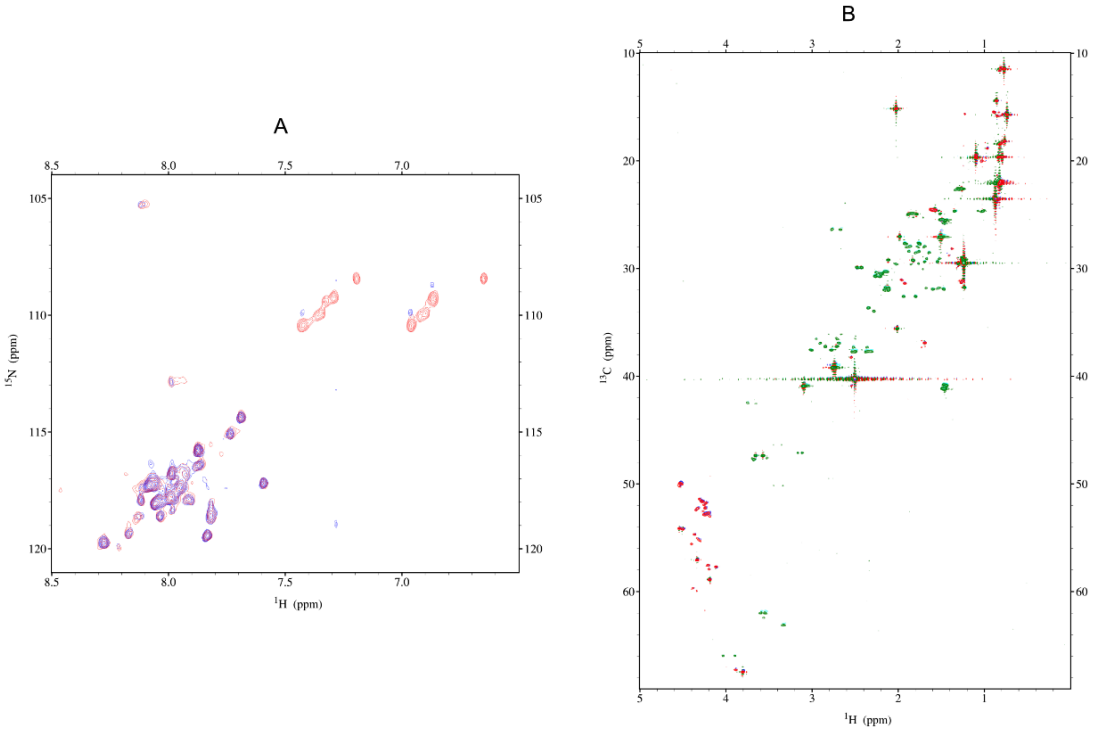

Figure 4.53 - Overlay of 2D ^{1}H-^{15}N HSQC and 2D ^{1}H-^{13}C HSQC of MEG 2.1 isoform 1 peptide without SP (64 aa, residues D25-P88) peptide in DMSO-d6 at a concentration of 2 mM with the addition of 3 mM $ZnCl_2$. A - 2D ^{1}H-^{15}N HSQC of isoform 1 without 3 mM $ZnCl_2$ (red) and with 3 mM $ZnCl_2$ (blue); B - 2D ^{1}H-^{13}C HSQC of isoform 1 without 3 mM $ZnCl_2$ (red/green) and with 3 mM $ZnCl_2$ (blue/cyan); experiments have been recorded at 27 °C with Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) and equipped with a triple HCN cryoprobe.

The first peptide derived from isoform 1 is iso 1a. Its length is 19 aa, and it contains 3 cysteines, 3 aspartic acids and the DINDITCNKTVC motif, which is also present in isoform 2a, due to the nature of the MEG 2.1 alternatively spliced isoforms. The repetitive DINDI motif was one of the challenges of the assignment of this peptide. For this peptide, an

assignment of 75 % ^{1}H, ^{13}C and ^{15}N chemical shifts was achieved (132 assigned resonances out of 176 theoretical ones), as can be seen in Fig. 4.47, 4.54, 4.55, and 4.56. The distribution of resonances in the narrow spectral range again made the assignment of the ^{1}H-^{1}H NOESY (Fig. 4.55) spectrum more time consuming. At the same time, the ^{1}H-^{13}C HSQC-TOCSY experiment was not as helpful as in the case of the other peptides (Fig. 4.56B).

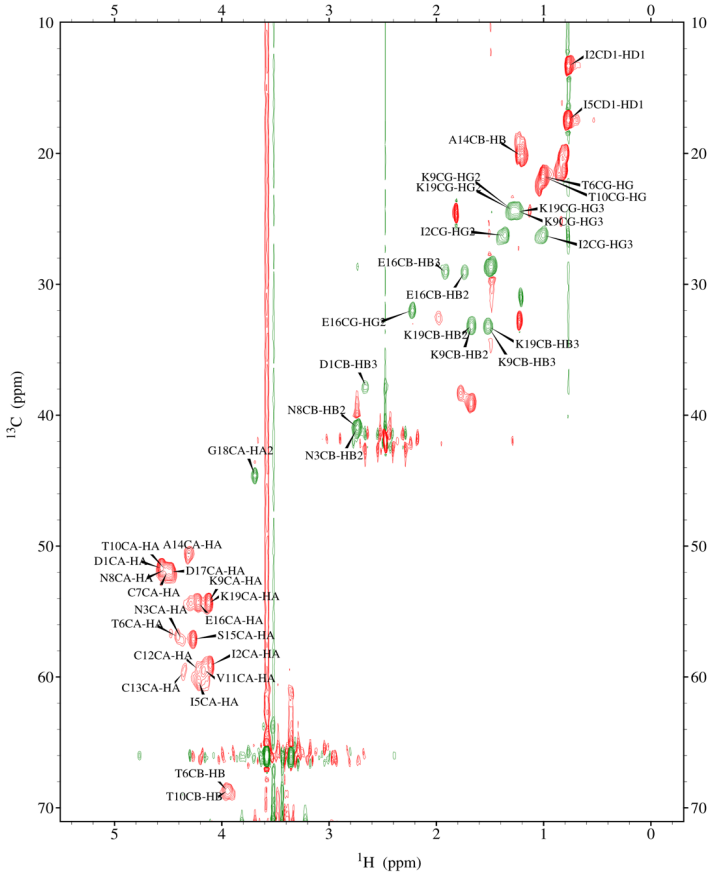

Figure 4.54 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 1a (19 aa, D25 - K43) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-19).

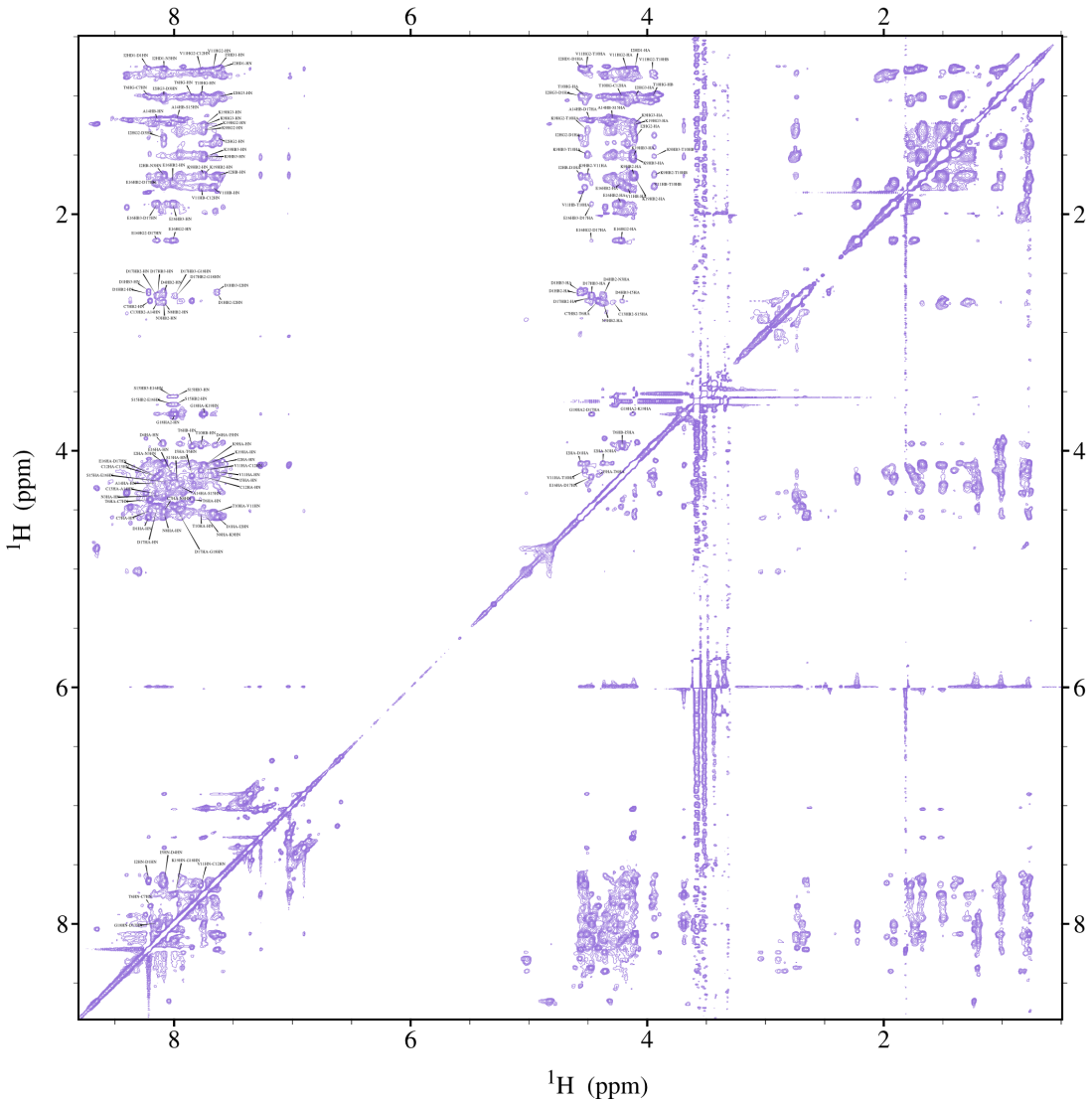

Figure 4.55 - Assignments of 2D ^{1}H-^{1}H NOESY of MEG 2.1 isoform 1a (19 aa, D25 - K43) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-19).

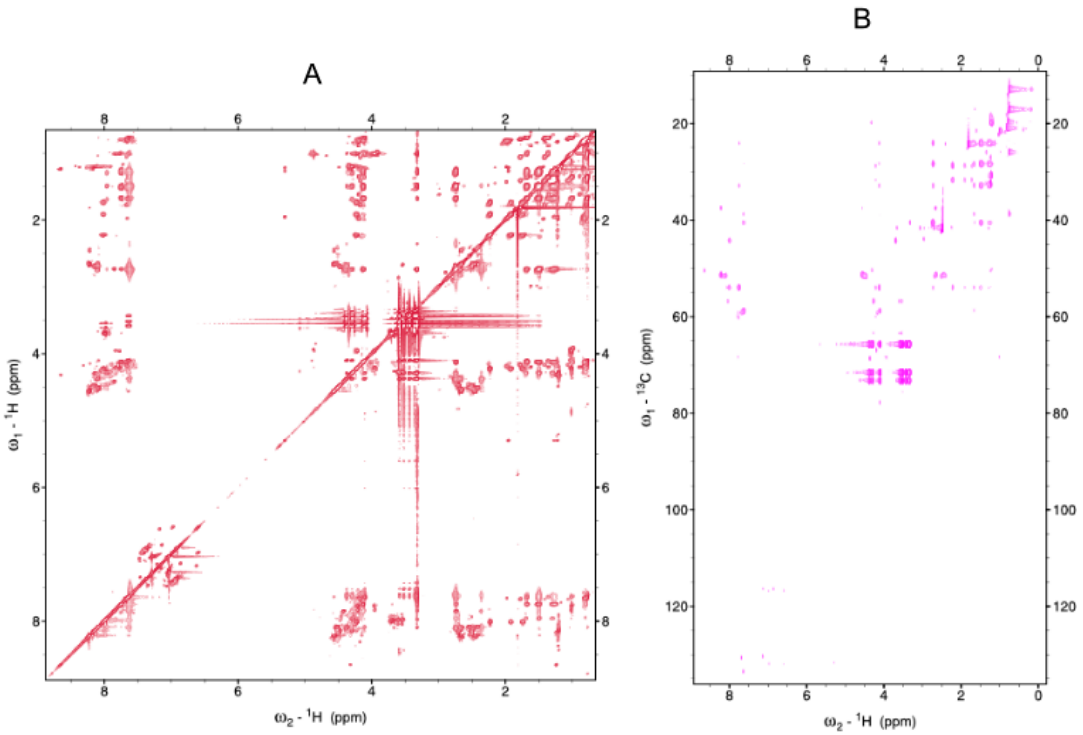

Figure 4.56 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY (A) and ^{1}H-^{13}C HSQC-TOCSY (B), which were also used for the assignment of MEG 2.1 isoform 1a (19 aa, D25 - K43) peptide in DMSO-d6 at a concentration of 2 m. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Isoform 1b contains 17 aa of which three lysines, three cysteines, three glycines and two prolines. The motif of this sequence is unique to this isoform. The resonances of the ^{1}H-^{15}N HSQC spectrum were slightly better separated than for iso 1a, which helped in the assignments. For iso 1b, 83 % of the assignment was achieved for all of the ^{1}H, ^{13}C and ^{15}N chemical shifts: 142 assigned resonances out of 172 theoretical ones (Fig. 4.47, 4.57, 4.58 and 4.59). The more complex parts of assignments in this sequence were the GKKG repetitive motive and the final PIPS part.

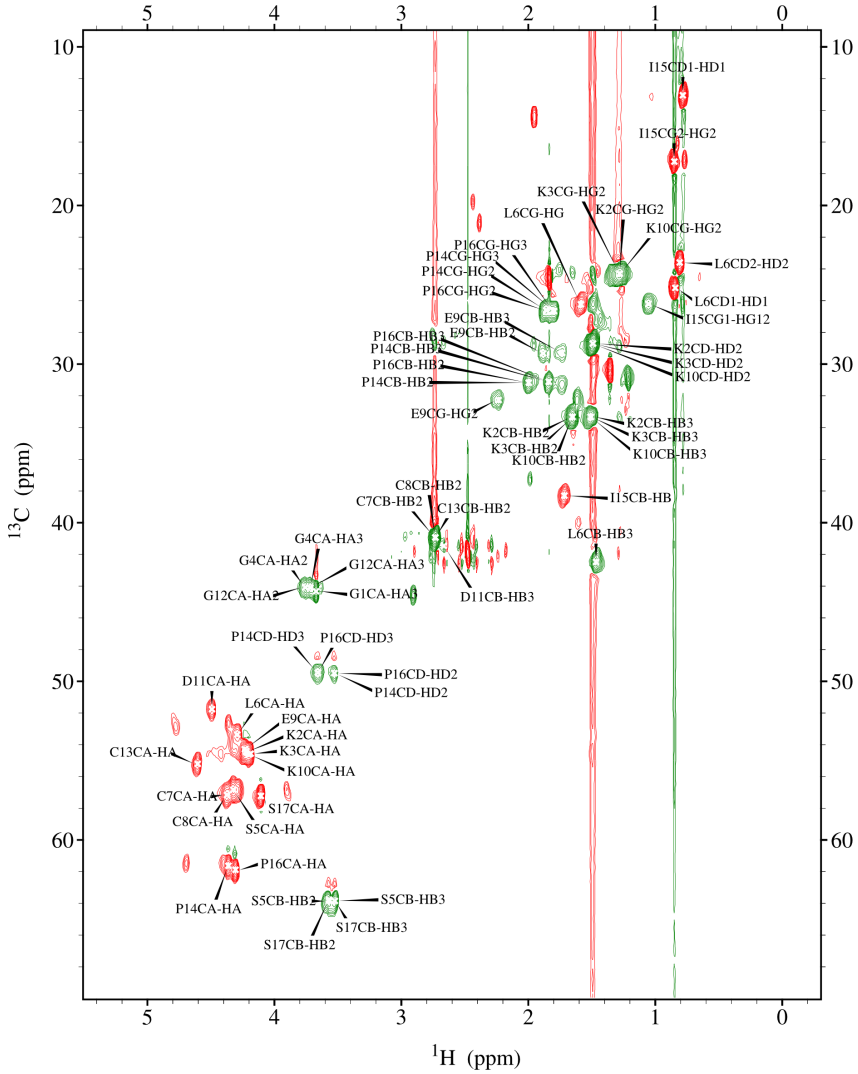

Figure 4.57 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 1b (17 aa, G42 - S58) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-17).

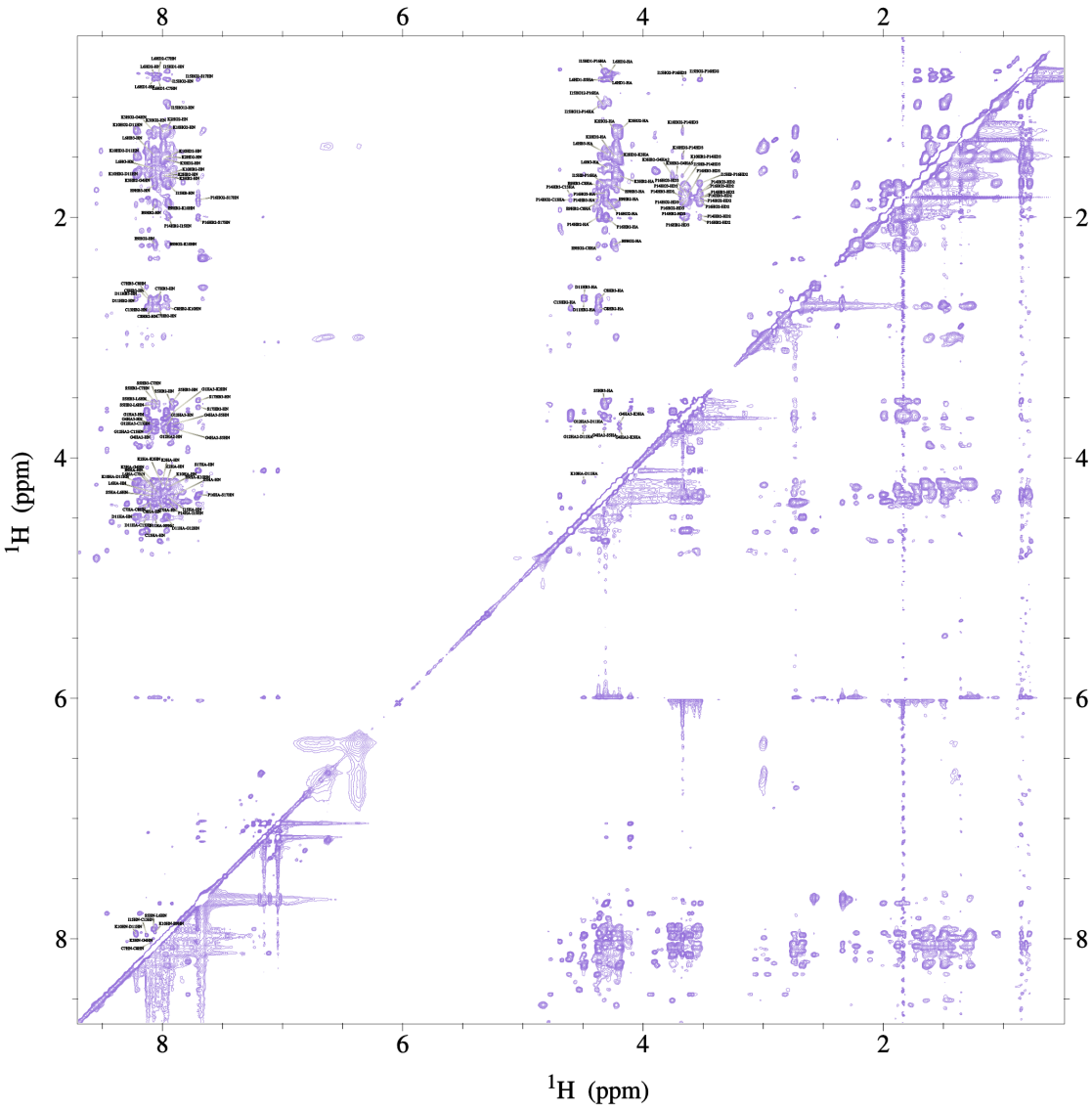

Figure 4.58 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 400 ms) of MEG 2.1 isoform 1b (17 aa, G42 - S58) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-17).

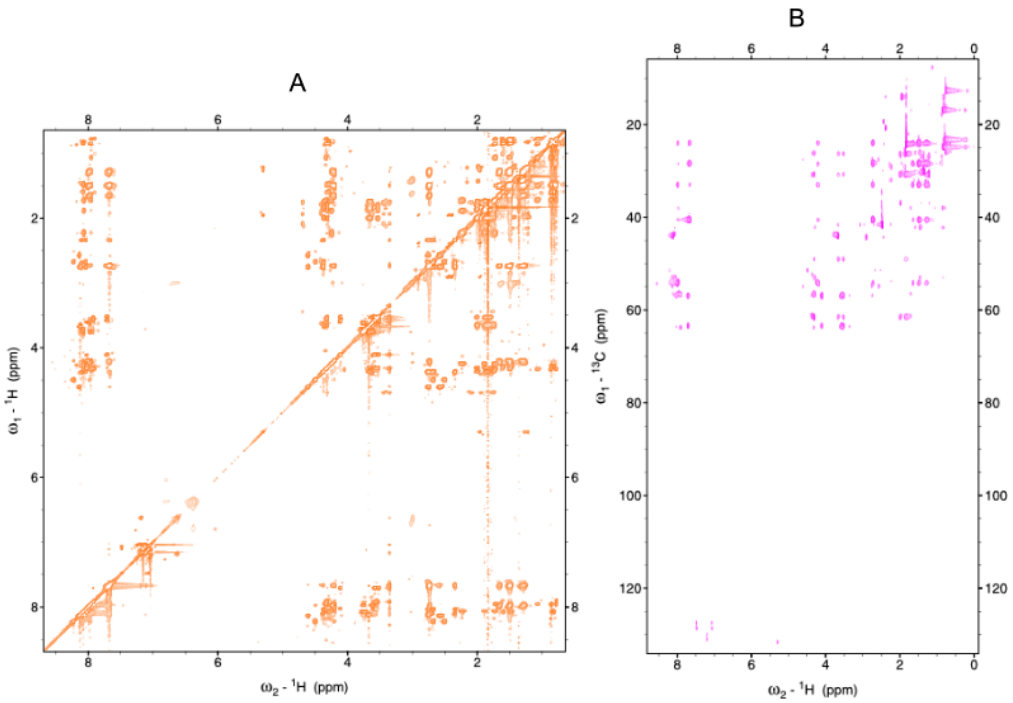

Figure 4.59 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY (A) and ^{1}H-^{13}C HSQC-TOCSY (B), which were also used for the assignment of MEG 2.1 isoform 1b (17 aa, G42 - S58) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Isoform 1c is composed of 31 aa and covers the entire C-terminus of MEG 2.1 isoform 1. As already mentioned, and as can be seen from Fig. 4.60A, the ^{1}H-^{15}N HSQC spectrum does not cover even half of the amino acids. Without a good quality ^{1}H-^{15}N HSQC spectrum the assignment is very difficult and in the case of iso 1c, all resonances were again in a very narrow spectral range (Fig. 4.60C, 4.60D). At the same time, the ^{1}H-^{13}C HSQC-TOCSY (Fig. 4.60E) experiment also brought very few resonances that could help in the assignment. Therefore, we decided to split the 1c isoform into two peptides to achieve a complete assignment of this sequence.

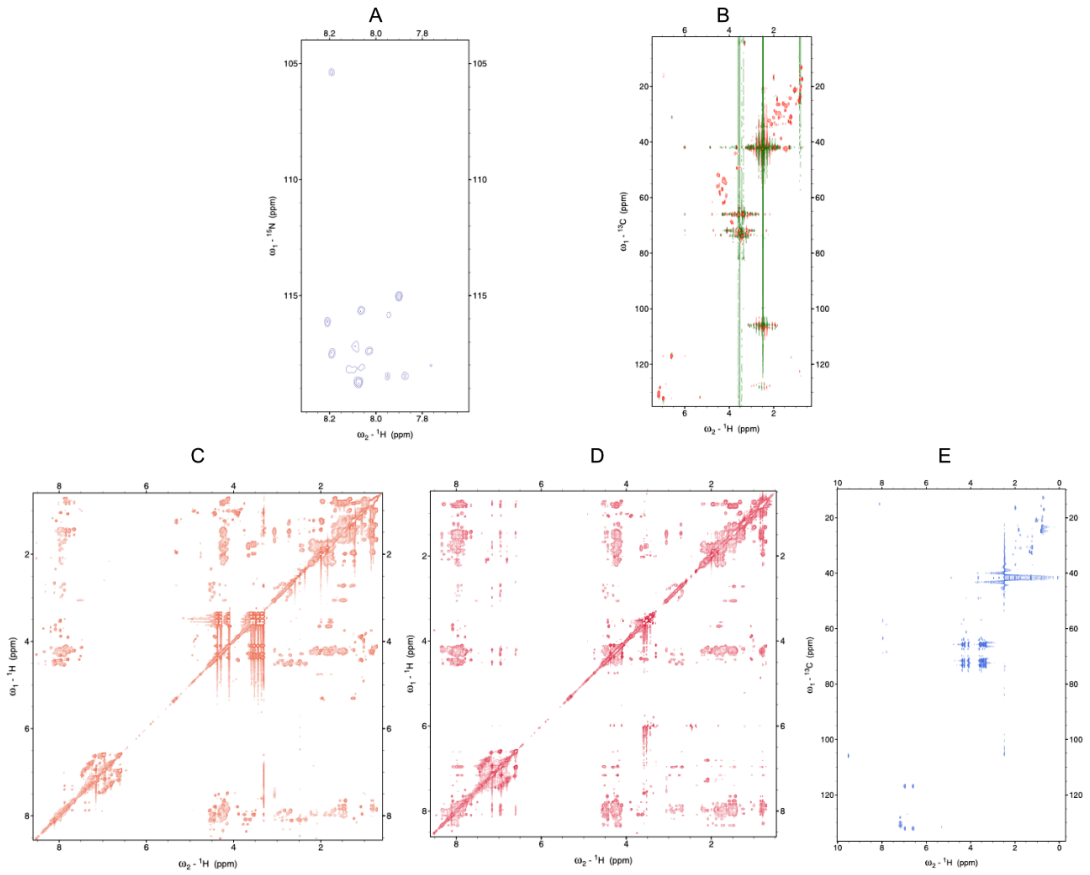

Figure 4.60 - MEG 2.1 isoform 1c - long version of the peptide (31 aa, S58 - P88) in DMSO-d6 at a concentration of 2 mM. A - 2D ^{1}H-^{15}N HSQC; B - 2D ^{1}H-^{13}C HSQC; C - 2D ^{1}H-^{1}H TOCSY (mixing time 80 ms); D - 2D ^{1}H-^{1}H NOESY (mixing time 400 ms); E - ^{1}H-^{13}C HSQC-TOCSY; experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a ^{1}H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Dividing iso 1c into two shorter peptides allowed us to significantly increase the number of visible resonances in all spectra, not only in the ^{1}H-^{15}N HSQC (Fig. 4.47). Isoform 1f is composed of 15 aa and contains a significant number of repeats; it also contains three consecutive leucines, two arginines and two glutamines (in the QRHQR motif). Despite these sequence composition rich in repetitions and three clustered leucines, it was possible to identify 14/15 amino acids in the ^{1}H-^{15}N spectrum (15th aa was proline) and 89 % of the ^{1}H, ^{13}C and ^{15}N chemical shifts (148 assigned resonances out of 166 theoretical ones). ^{1}H-^{13}C HSQC-TOCSY spectrum was only used for checking the assignment because all the necessary information was obtained from ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC (Fig. 4.61), ^{1}H-^{1}H TOCSY (Fig. 4.63) and ^{1}H-^{1}H NOESY (Fig. 4.62) spectra. Indeed, the peaks of iso 1f were well separated even for ^{1}H-^{1}H TOCSY and ^{1}H-^{1}H NOESY experiments.

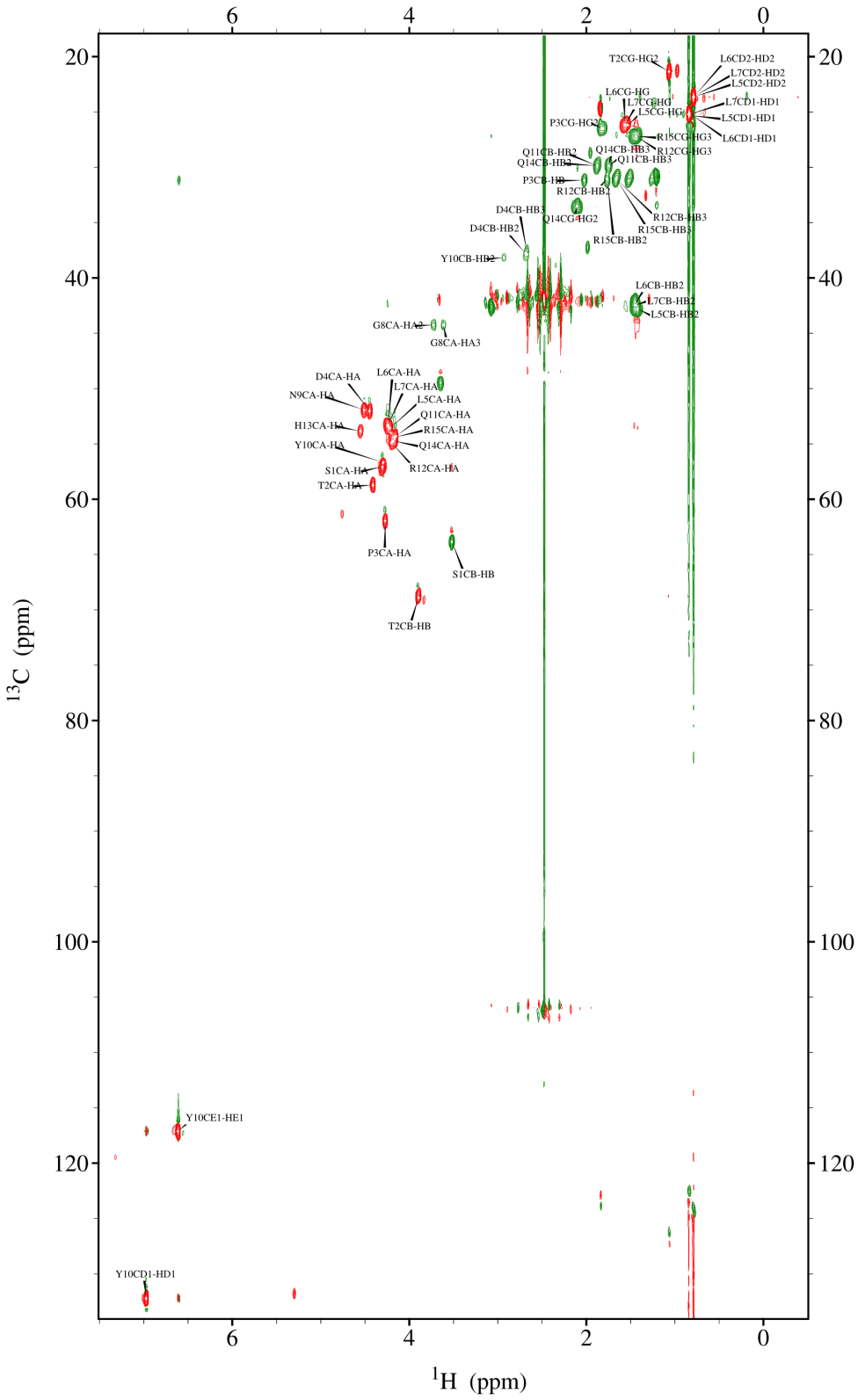

Figure 4.61 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 1f (15 aa, S58 - R72) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-15).

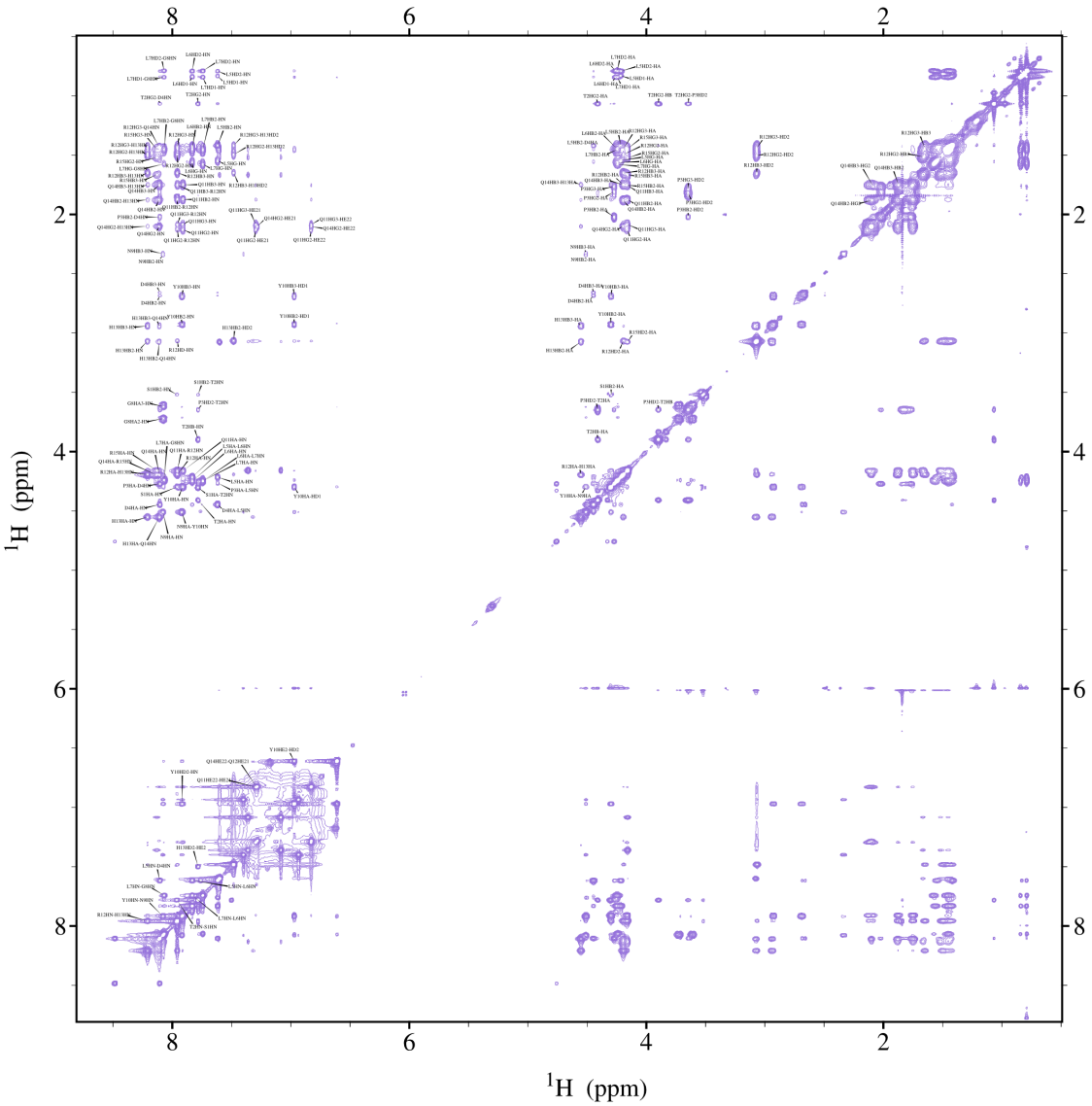

Figure 4.62 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 400 ms) of MEG 2.1 isoform 1f (15 aa, S58 - R72) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-15).

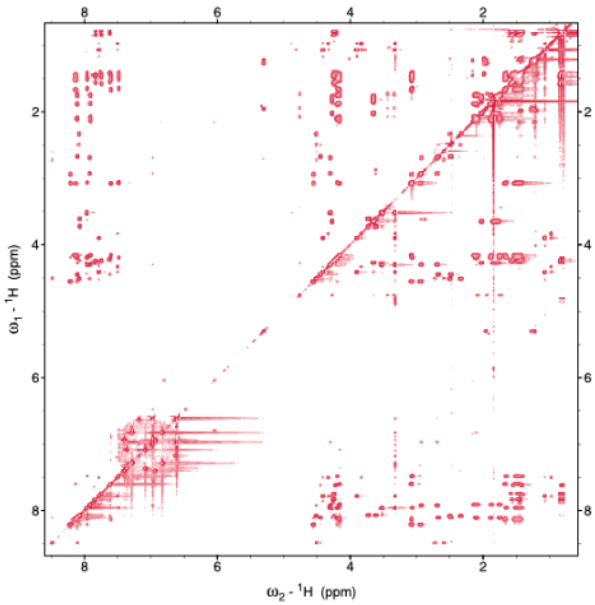

Figure 4.63 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY which was also used for the assignment of MEG 2.1 isoform 1f (15 aa, S58 - R72) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Isoform 1g contains 16 aa and is the C-terminal part of isoform 1; it contains a terminal FIYTP motif common to isoform 2 and a YTP motif common to isoform 3. The sequence contains three glutamic acids (two of which are consecutive), two tyrosines and two asparagines. Of the four peptides synthesized from MEG 2.1 isoform 1, this one gave the best assignment result (91 %) for all of the ^{1}H, ^{13}C and ^{15}N chemical shifts, 160 assigned resonances out of 177 theoretical ones (Fig. 4.64, 4.65 and 4.66). In the ^{1}H-^{15}N HSQC spectrum, all peaks were assigned, as shown in Fig. 4.47.

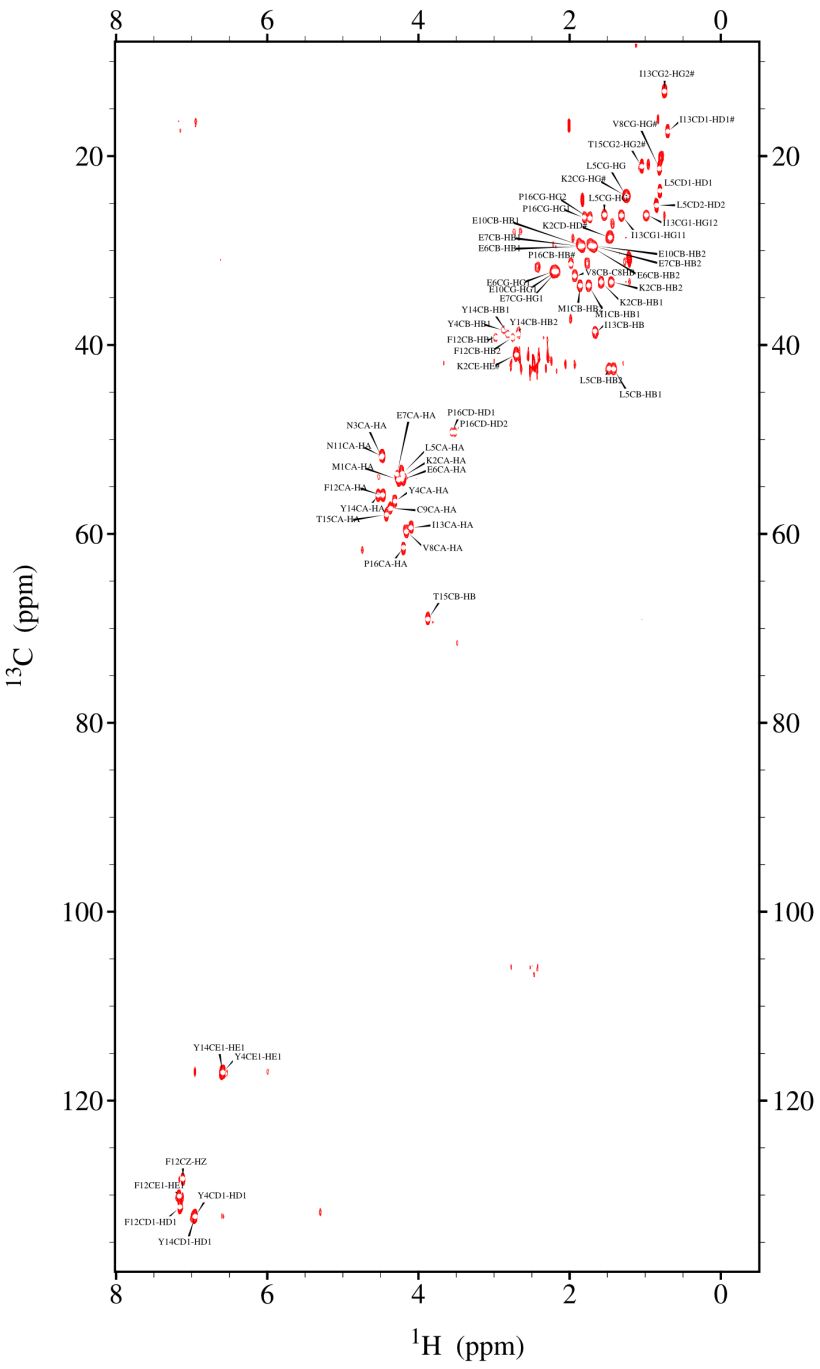

Figure 4.64 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 1g (16 aa, M73-P88) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-16).

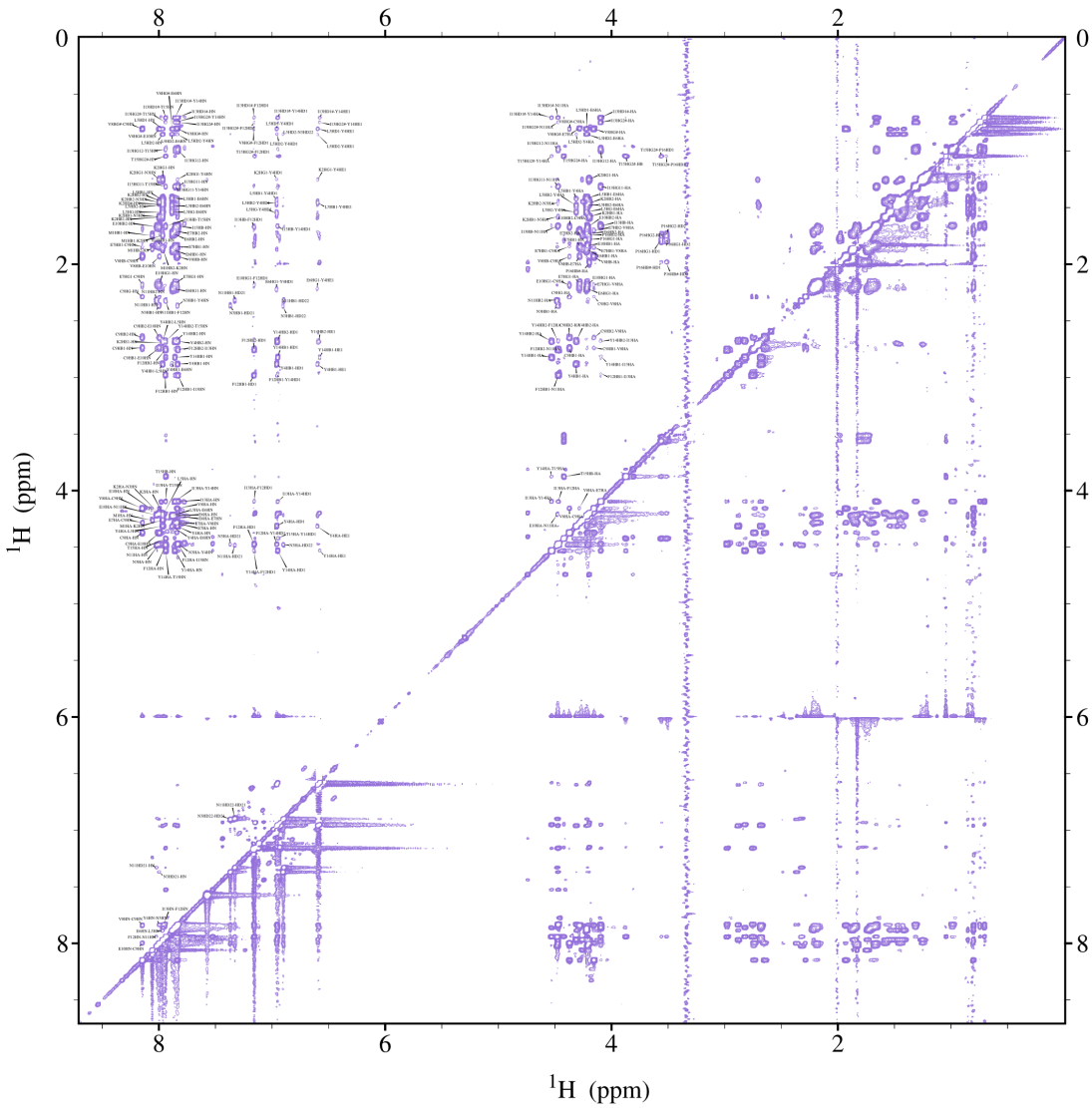

Figure 4.65 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 400 ms) MEG 2.1 isoform 1g (16 aa, M73-P88) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-16).

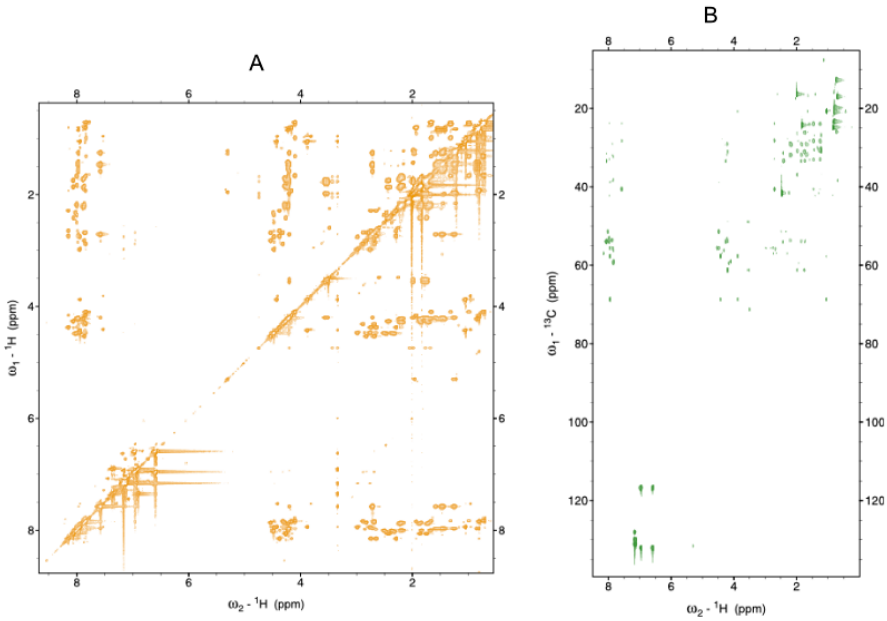

Figure 4.66 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY (A) and ^{1}H-^{13}C HSQC-TOCSY (B), which were also used for the assignment of 1g (16 aa, M73-P88) peptide in DMSO-d6 at a concentration of 2 mM.. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

For a better demonstration of how many resonances were obtained by splitting isoform 1c into two peptides (1f and 1g), the ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC (Fig. 4.67B) spectra of these three peptides were superimposed (Fig. 4.67). In the ^{1}H-^{15}N HSQC spectrum (Fig. 4.67A) not only is the complete superposition of all peaks of isoform 1c with the shorter peptides 1f and 1g is visible, but at the same time it is apparent at first glance that this division of 31 aa into 16 aa and 15 aa resulted in a doubling of the number of assignable resonances. The ^{1}H-^{13}C HSQC spectrum suggests the same conclusion, although it must be noted that the ^{1}H-^{13}C HSQC spectrum of isoform 1c gave us more resonances than the ^{1}H-^{15}N HSQC spectrum; with the shorter peptides strategy we obtained primary carbon peaks in the C-H and aromatic chemical shift regions (Fig. 4.60B).

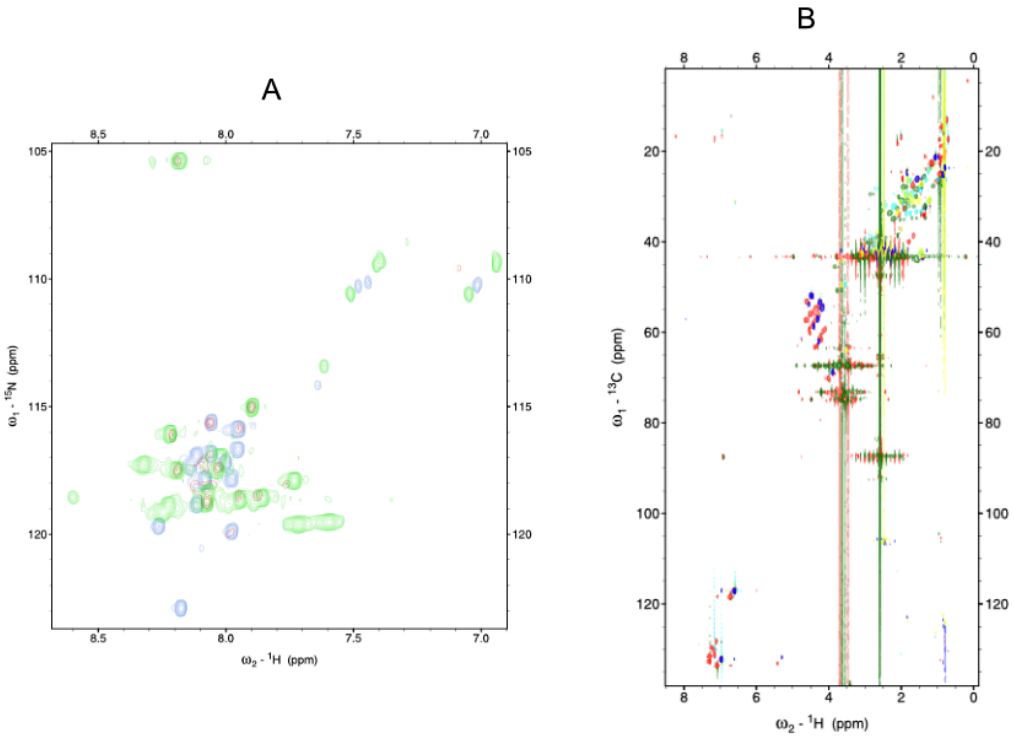

Figure 4.67 - Overlay of three 2D ^{1}H-^{15}N HSQC (A) and 2D ^{1}H-^{13}C HSQC (B) spectra - of the isoform 1c (31 aa) and two isoforms formed by the splitting of this long isoform into two peptides: 1f (15 aa) and isoform 1g (16 aa). A - the long version of isoform 1c (31 aa) - red; peptide 1f (15 aa) - green, peptide 1g (16 aa) - blue. B - the long version of isoform 1c (31 aa) - red/green; peptide 1f (15 aa) - blue/yellow, peptide 1g (16 aa) - crimson/cyan. All the experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a ^{1}H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Finally, the measured ^{1}H-^{15}N HSQC spectra of all four peptides formed from isoform 1 (iso 1a, iso 1b, iso 1f, iso 1g) were superposed with isoform 1 (Fig. 4.68). Even in this spectrum, the peaks of isoforms 1f and 1g are overlapping relatively well with the peaks of isoform 1. On the contrary, iso 1a and iso 1b do not overlap with the resonances of isoform 1 nearly at all. Indeed, resonances of L64-Q68, Q71, K74-L77 and I85 are closely superimposed and resonances of T59, C81, and Y86 show a slight variation of their chemical shifts in the uncut isoform 1 (25-88) compared to the two peptides iso 1f and iso 1g. We also observed that the division into shorter peptides had only a slight effect on the chemical shifts of the residues S58 to T87, which does not imply any drastic change in the electronic environment, such as secondary structural change. Unfortunately, the N-terminal part of MEG 2.1 isoform 1 is part of the peptide, whose resonances are missing in the ^{1}H-^{15}N HSQC spectra measured at both 600 MHz and 1.2 GHz (Fig. 4.47 and 4.52), so it was not possible to perform their assignment. For this reason, no conclusion can be drawn about the N-terminal structure of this MEG 2.1 isoform 1.

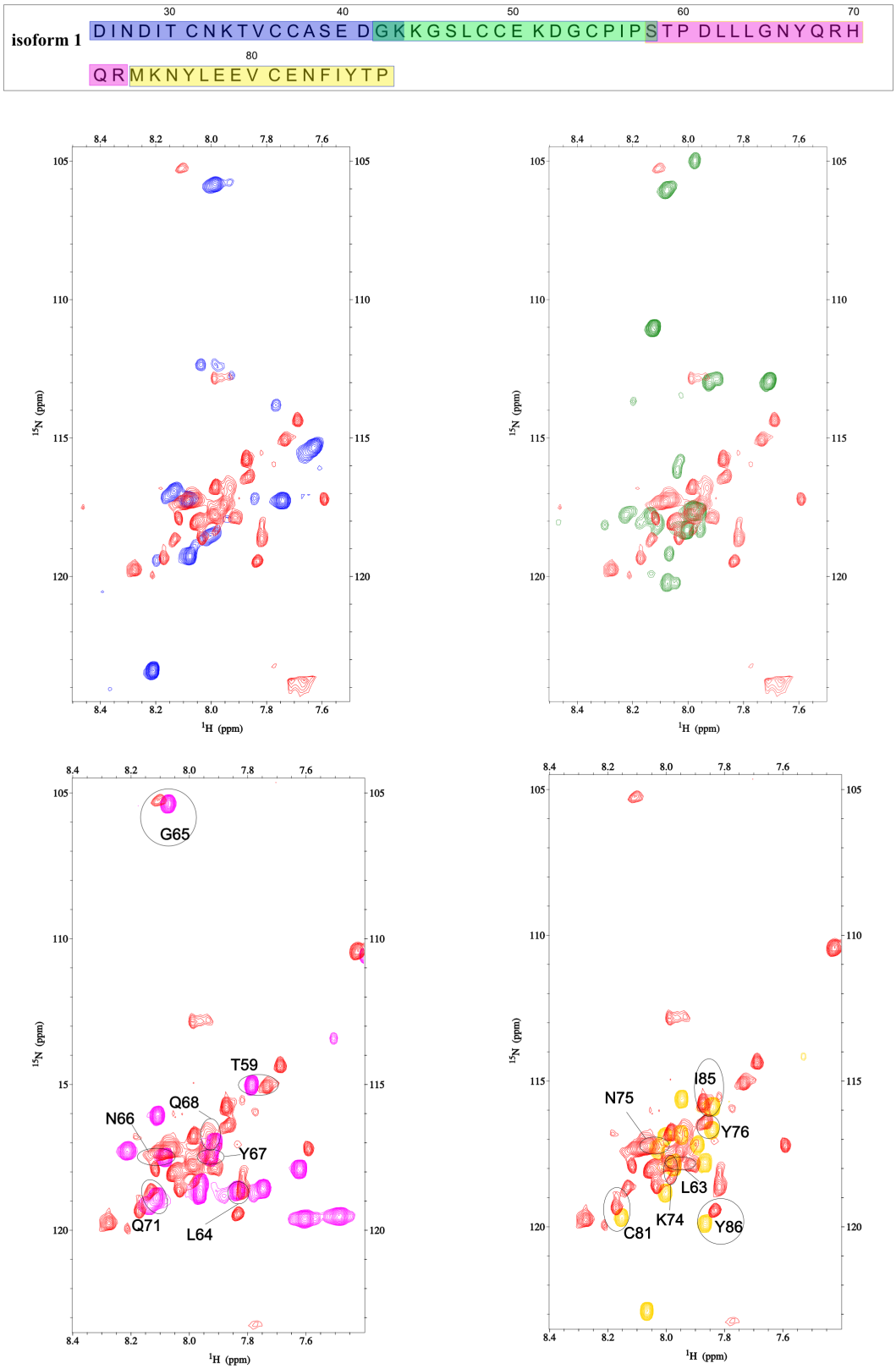

Figure 4.68 - Detailed overlay of all individual peptides of MEG 2.1 isoform 1 (iso 1a - blue, 1b - green, 1f - magenta, and 1g - yellow) with the complete isoform 1 (without SP, red in all the spectra). Peptides in DMSO-d6 at a concentration of 2 mM and all the experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Isoform 2a (Fig. 4.48) was assigned at 83 %, with 142 assigned peaks out of 172 theoretical ones (Fig. 4.70, 4.71, and 4.72). Thanks to the well-separated resonances of

the ^{1}H-^{15}N HSQC spectrum it was possible to assign all amino acids despite the fact that the chemical shifts gathered again in the narrow spectral range. Isoform 2a is 18 aa long, and contains three cysteines, two aspartic acids, two asparagines, two isoleucines and two valines. This isoform is the second most hydrophobic of all synthetic peptides. At the same time, a large part of this isoform corresponds to isoforms 1 and 1a because they share the DINDITCNKTVC motif. This isoform proved to be highly unstable compared to other peptides and its degradation occurred within one week (Fig. 4.69). As part of the degradation, it was possible to observe the loss of some assigned peaks and the appearance of new ones.

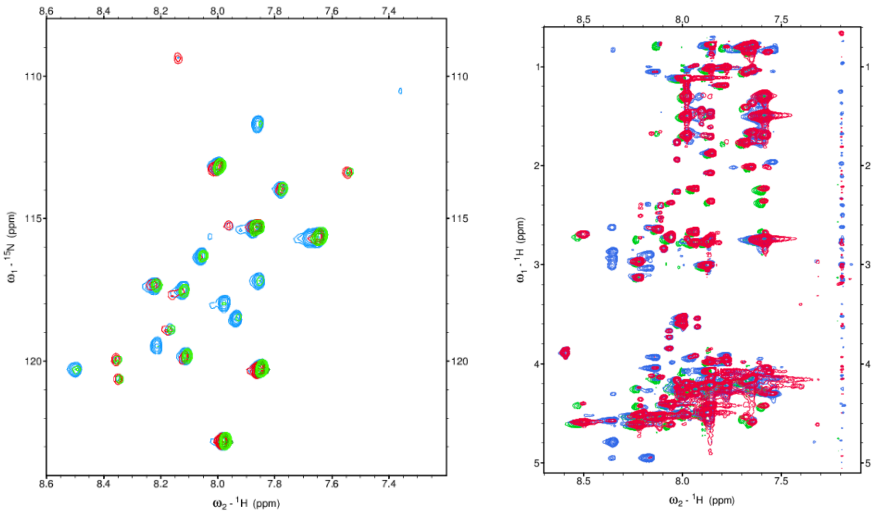

Figure 4.69 - Overlay of three 2D ^{1}H-^{15}N HSQC spectra and 2D ^{1}H-^{1}H TOCSY spectra of isoform 2a collected at different times. Superimposed are the results of the samples prepared on 08/08/22 (lime), 11/10/22 (blue) and 10/11/22 (crimson). All these samples were measured on three consecutive days to determine the rate of degradation. The first sample (lime) was by then 3 months old, the second (blue) was 1 month old and the last sample (crimson) was measured as a freshly prepared "reference".

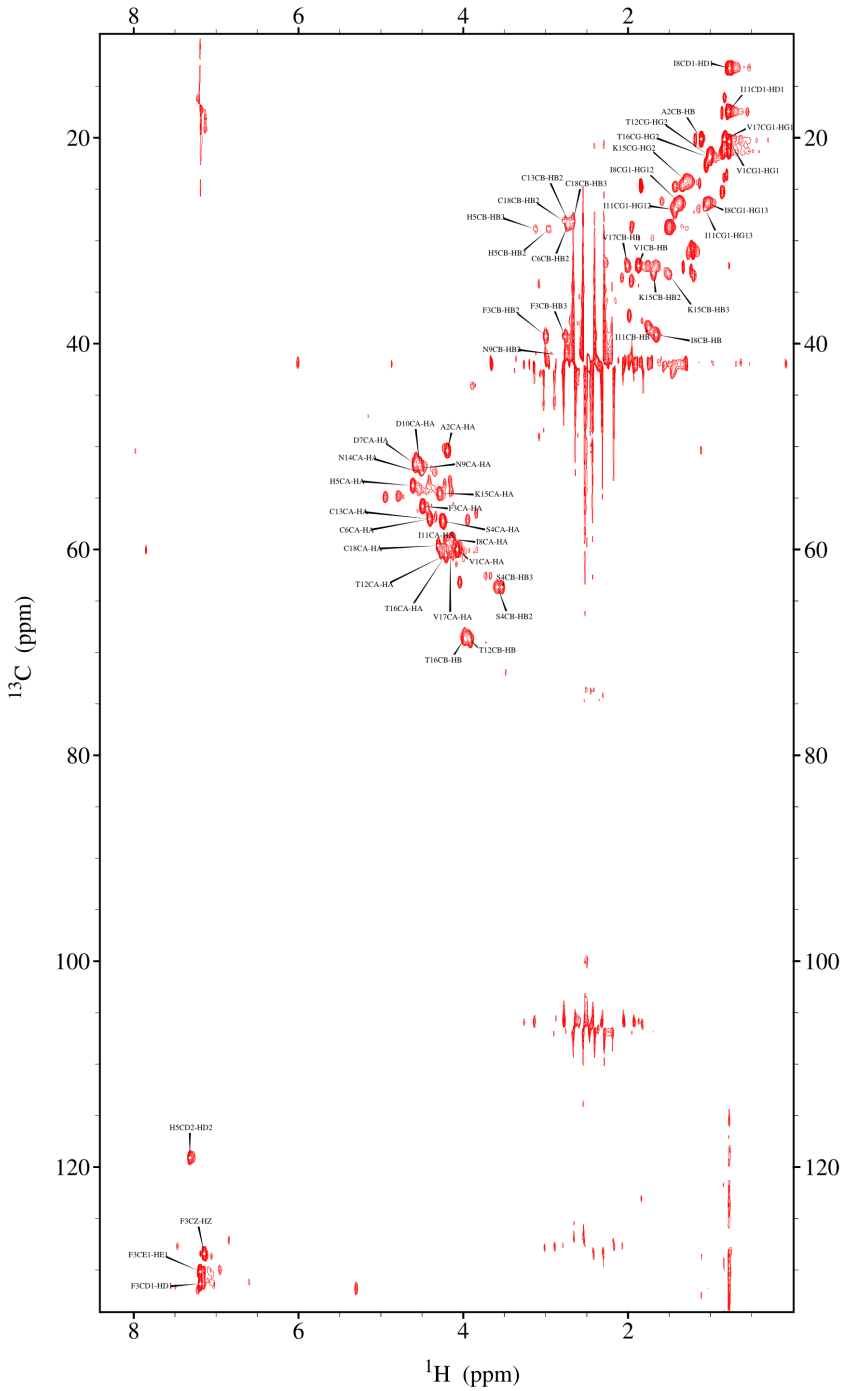

Figure 4.70 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 2a (18 aa, V19 - C36) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-18).

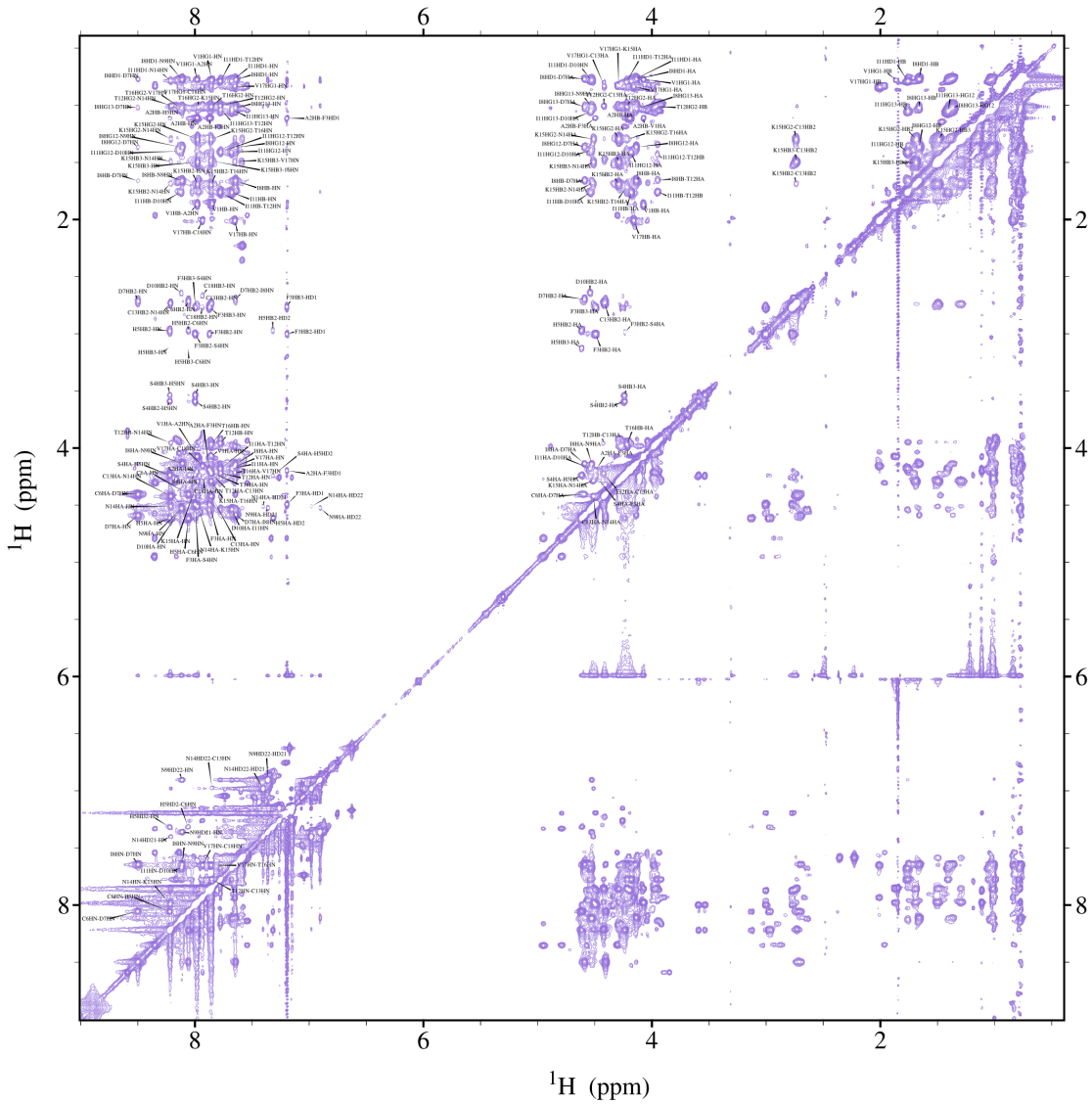

Figure 4.71 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 400 ms) MEG 2.1 isoform 2a (18 aa, V19 - C36) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-18).

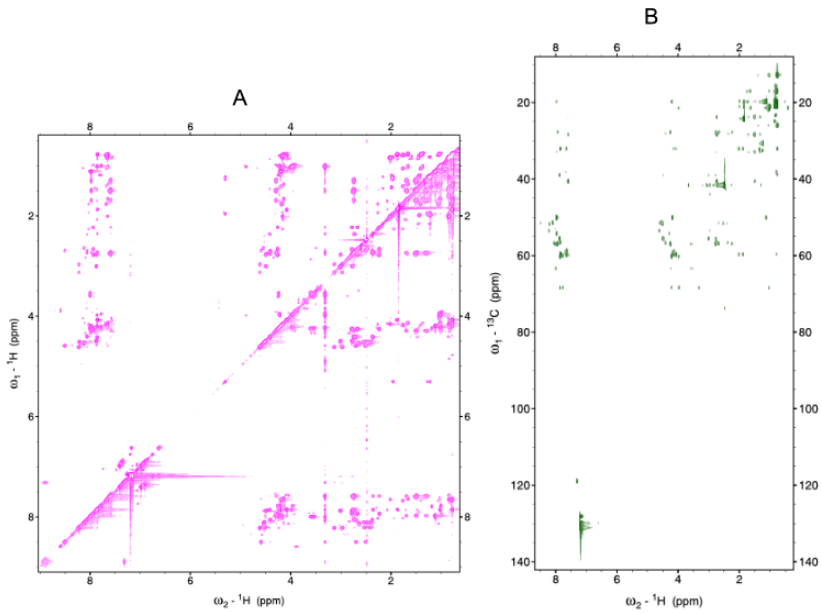

Figure 4.72 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY (A) and ^{1}H-^{13}C HSQC-TOCSY (B), which were also used for the assignment of 2a (18 aa, V19 - C36) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Isoform 2b is part of MEG 2.1 isoform 2, which was subjected to alternative splicing. It contains a part of the motif identical to isoforms 1 and 1a - CASEDGK - and at the same time the terminal part of FIYTP identical to isoforms 1 and 1g (of which the YTP motif is also identical to isoform 3). It was possible to assign 15/16 amino acids in the ^{1}H-^{15}N HSQC spectrum (16th aa is again proline). The overall assignment of all of the ^{1}H, ^{13}C and ^{15}N chemical shifts of the 2b isoform was 96 %, with 151 assigned resonances out of 157 theoretical ones (Fig. 4.48, 4.73, 4.74, and 4.75).

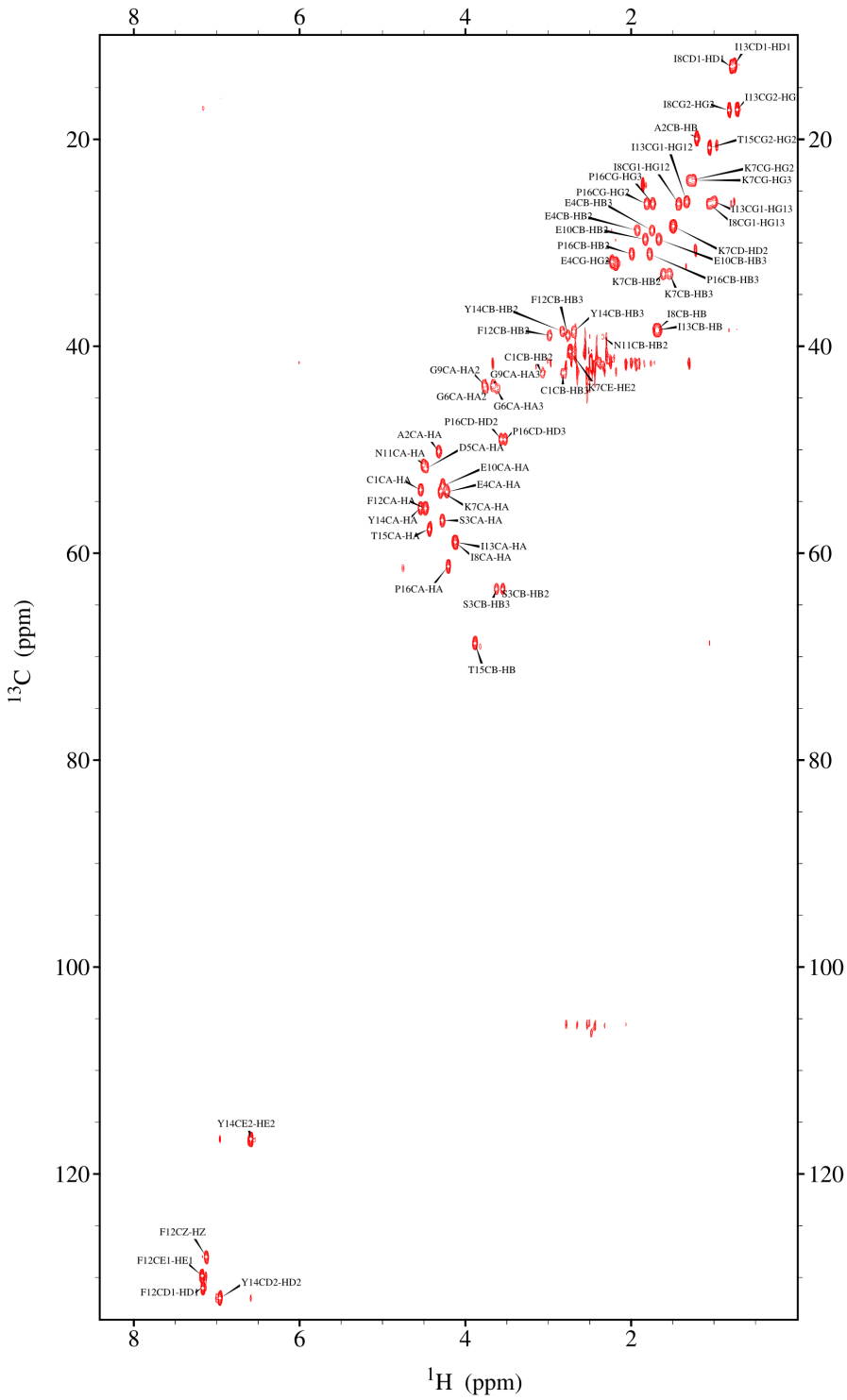

Figure 4.73 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 2b (16 aa, C37 - P52) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-16).

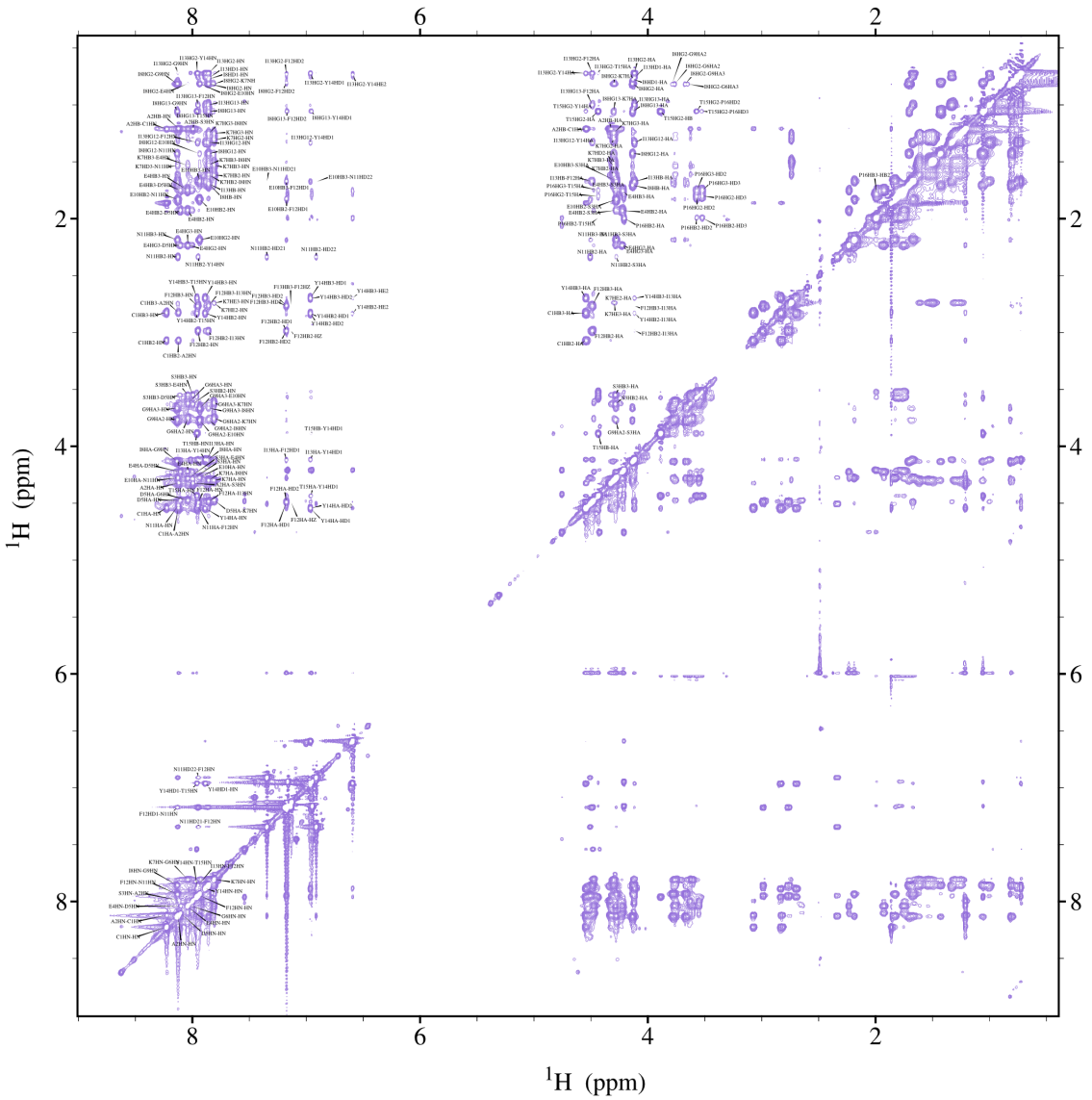

Figure 4.74 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 400 ms) MEG 2.1 isoform 2b (16 aa, C37 - P52) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C. Assignments of amino acids are numbered within the analyzed peptide (1-16).

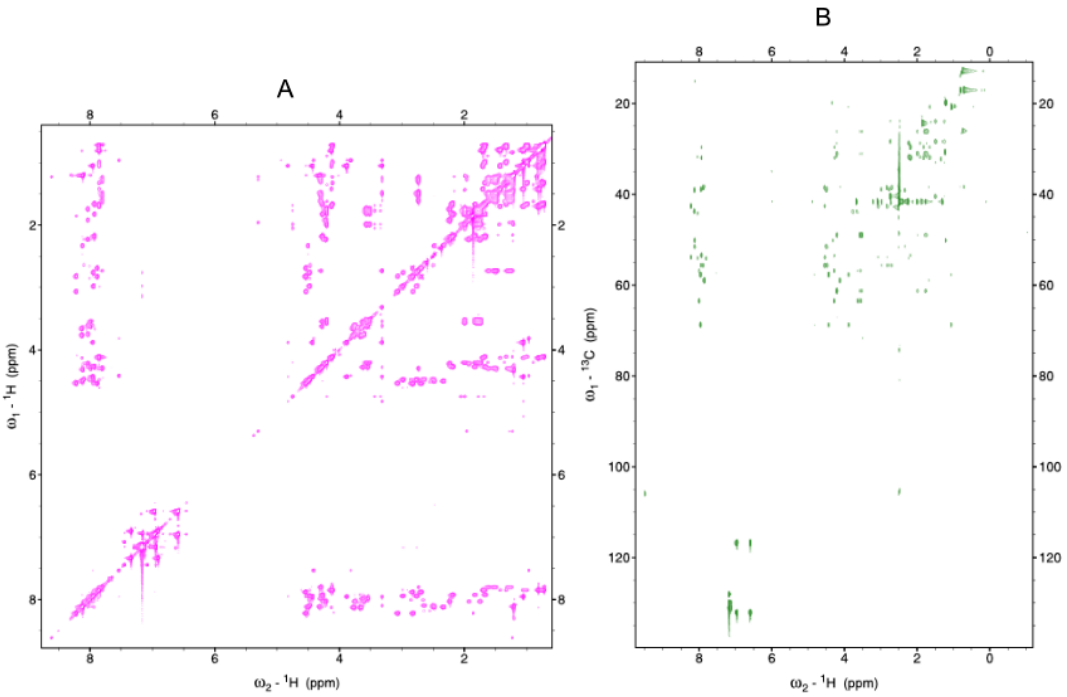

Figure 4.75 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY (A) and ^{1}H-^{13}C HSQC-TOCSY (B), which were also used for the assignment of 2b (16 aa, C37 - P52) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Moreover, we can notice that ^{1}H-^{15}N HSQC spectra of iso 1a and iso 2a, as well as spectra of iso 1a, iso 1d and iso 2b show several chemical shift superpositions (for example ^{26}INDI^{29} and ^{84}FIYT^{87}) indicating a conserved similar structural organization (Fig. 4.76).

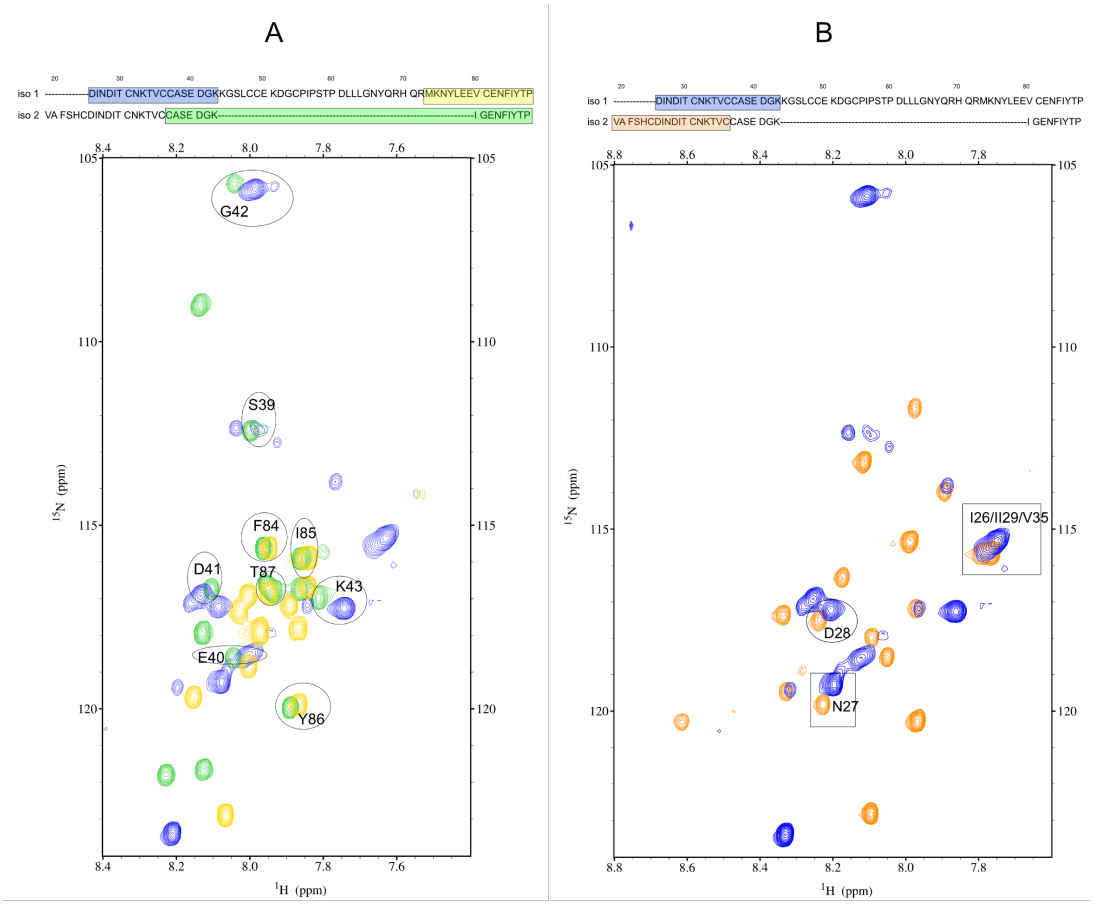

Figure 4.76 - Overlay of 2D ^{1}H-^{15}N HSQC spectra - of the MEG 2.1 isoform 1 and MEG2.1 isoform 2 peptide. *A - overlay of isoform 1a, 1g and 2b; B - overlay of isoform 1a and 2a.* Experiments have been recorded at 27 °C with Varian Inova spectrometer operated at a 1H frequency of 600 MHz (14.1 T) and equipped with a triple HCN cryoprobe enhanced in ^{1}H and ^{13}C.

Isoform 3 is the most dramatically spliced isoform of the MEG 2.1 protein, consisting of 26 aa, which contains, among others, 7 leucines, and 3 valines (both of the residues are often in the sequence immediately after each other). The fact that more than 1/3 of the sequence consists of hydrophobic residues makes it the most hydrophobic isoform of those studied here. Isoform 3 is a predicted signal peptide of the MEG 2.1 family; thus, such high hydrophobicity is not surprising. From CD data analysis, we showed that isoform 3 exhibits the features of α helix, which confirmed the *ab initio* prediction by AI tools. Assignment of isoform 3 was highly challenging not only because of the repetitions of hydrophobic amino acids and motifs such as LLQLLV, but also because all resonances were concentrated within 1 ppm of spectral range. This was a major problem for the assignment of ^{1}H-^{1}H TOCSY and especially ^{1}H-^{1}H NOESY spectra. Despite all the described complications, it was possible to achieve the overall assignment of 88 % of all of the ^{1}H, ^{13}C and ^{15}N chemical shifts of this isoform (236 assigned resonances out of 269 theoretical ones) - Fig. 4.49, 4.77, 4.78, and 4.79. We have also tested different conditions of the NMR experiments, in order to promote a shift/expansion of the spectral width at which the peaks were observed. For this purpose, the higher temperature of the analyses (32 °C) and the addition of europium in the form of EuFOD

to the sample were tested. Neither of these experiments produced a change in the distribution of resonances in the spectral range.

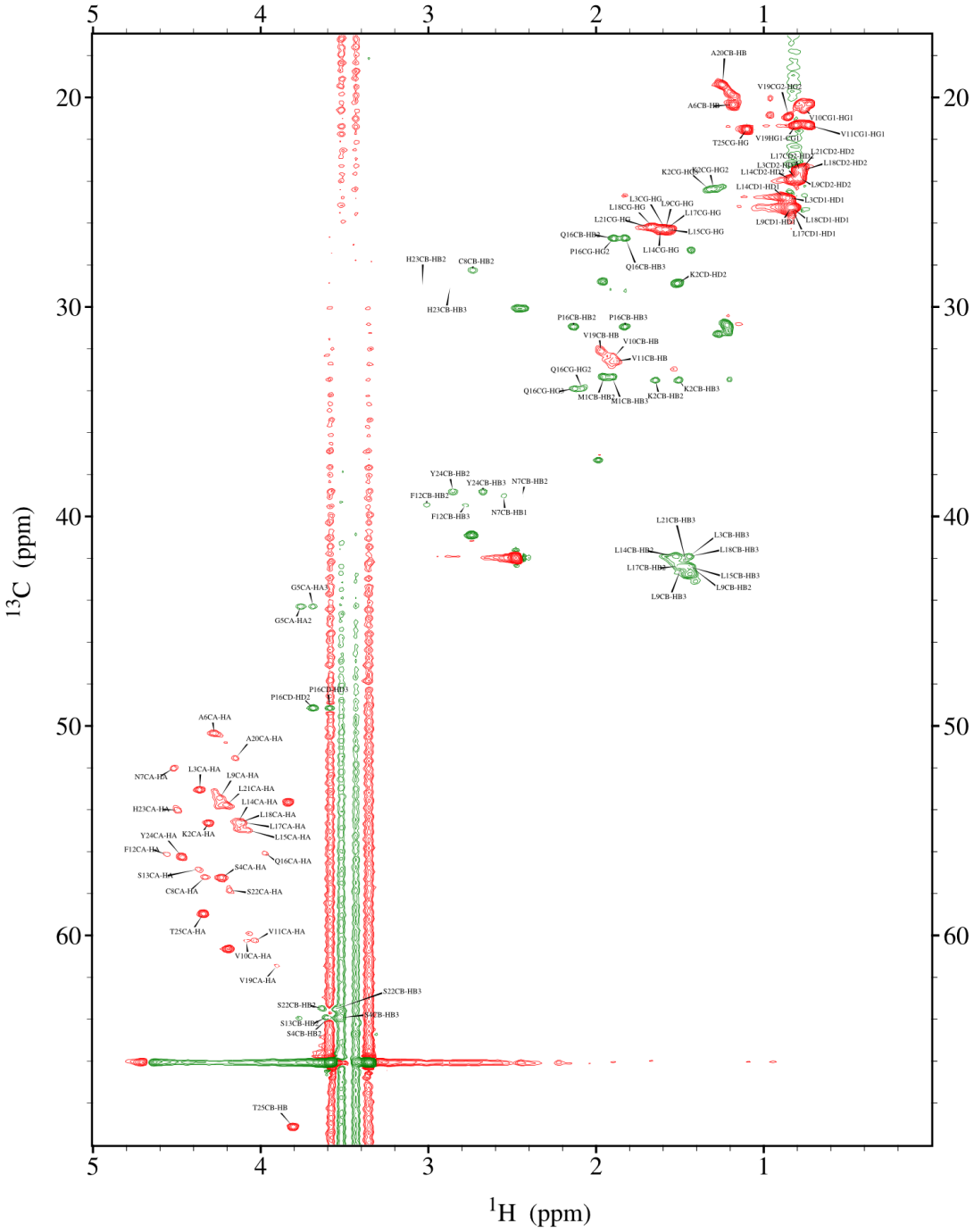

Figure 4.77 - Assignments of 2D ^{1}H-^{13}C HSQC MEG 2.1 isoform 3 (26 aa, M1 - P26) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) equipped with a triple HCN cryoprobe.

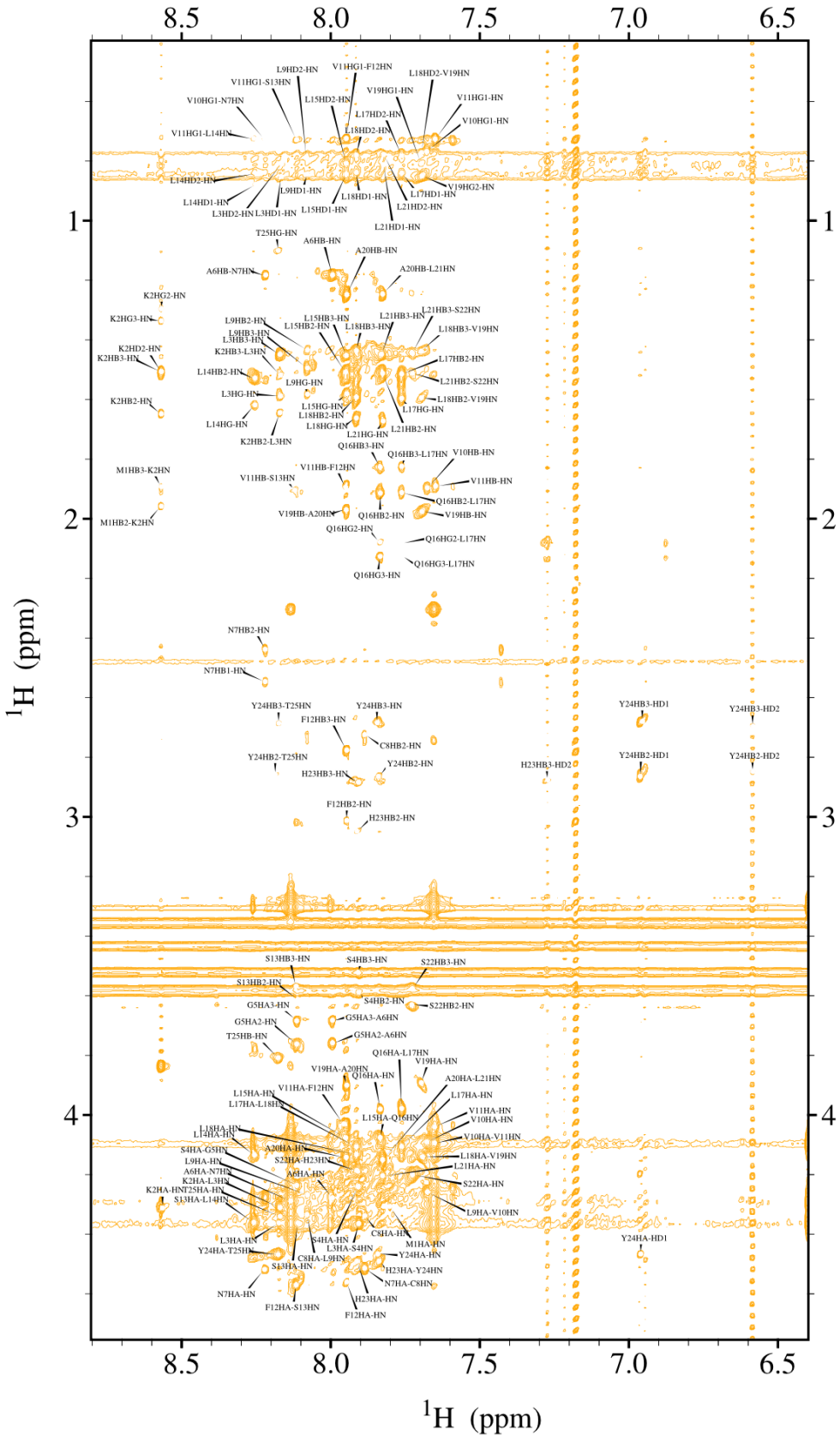

Figure 4.78 - Assignments of 2D ^{1}H-^{1}H NOESY (mixing time 400 ms) MEG 2.1 isoform 3 (26 aa, M1 - P26) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) equipped with a triple HCN cryoprobe.

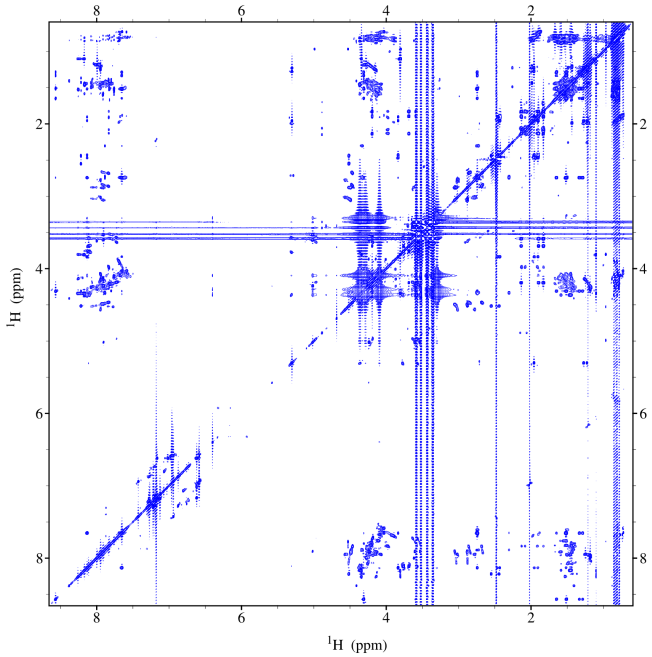

Figure 4.79 - Complementary unassigned spectra of ^{1}H-^{1}H TOCSY (mixing time 60 ms) which was also used for the assignment of 3 (26 aa, M1 - P26) peptide in DMSO-d6 at a concentration of 2 mM. Experiments have been recorded at 27 °C Bruker Neo spectrometer operated at a ^{1}H frequency of 1.2 GHz (28.2 T) equipped with a triple HCN cryoprobe.

4.3.6 Structure refinement and molecular docking

After the assignment of all the above-described isoforms and their spectra, we proceeded to the structural refinement of these peptides with the CYANA software. The dataset CYANA works with contains ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, ^{1}H-^{1}H TOCSY chemical shift assignments and integration of signals measured in ^{1}H-^{1}H NOESY. The number of assigned NOEs, together with the number of ambiguities in their assignment, relative to the other assigned spectra and the standard values of distance constraints, is crucial for the quality of the structural refinement.

None of the isoform 1 peptides (iso 1a, 1b, 1f, and 1g) contain any violated distance constraints after seven cycles of refinement because their Ramachandran plots (Annex D) indicate 0 % in both disallowed and generously allowed regions. Nevertheless, we observed a lower degree of structuration for iso 1a, iso 1b, and iso 1f peptides. Indeed, there are no or very few long-range distances (Fig. 4.80). Consequently, the 10-lowest energy structures don't overlay well, except for a short part of the sequence involving barely 3-4 residues. Those residues exhibit the highest number of NMR constraints, uniquely i/i+1 correlations, without creating any secondary structure. Iso 1a and 1b have the highest proportion of the most favored region NOEs of all peptides (71% for iso 1a and 72% for iso 1b; Table 4.8). On the other hand, we can notice that residues E78 to F84 of iso 1g peptide are well-stacked for the 10-lowest energy structures and we measured a few long-range constraints (Fig. 4.80). Indeed, among 253 measured NOEs, 181 have been used by CYANA to derive the structure (Table 4.8).

Table 4.8 - Total and unambiguous numbers of NOE as well as Ramachandran statistics for MEG 2.1 iso 1a, iso 1b, iso 1c, iso 1d, iso 2a, and iso 2b peptides.

Peptide	# aa	NOE numbers		Ramachandran plots statistics	
		Total	Unambiguous	Most favored region	Additionally allowed region
iso 1a	19	143	80	71.2 %	28.8 %
iso 1b	17	154	83	71.8 %	28.2 %
iso 1f	15	153	81	56.4 %	43.6 %
iso 1g	16	253	181	53.6 %	46.4 %
iso 2a	18	212	138	52.5 %	47.5 %
iso 2b	16	230	141	59.2 %	40.8 %

This is the highest value of NOEs used for structure calculation among the 6 peptides of MEG 2.1 isoform 1 and 2; for this reason, it is also evident that the 10-lowest energy models overlap noticeably better than the other three peptides of isoform 1 (Fig. 4.81). As a result, all four peptides of isoform 1 were determined to be disordered.

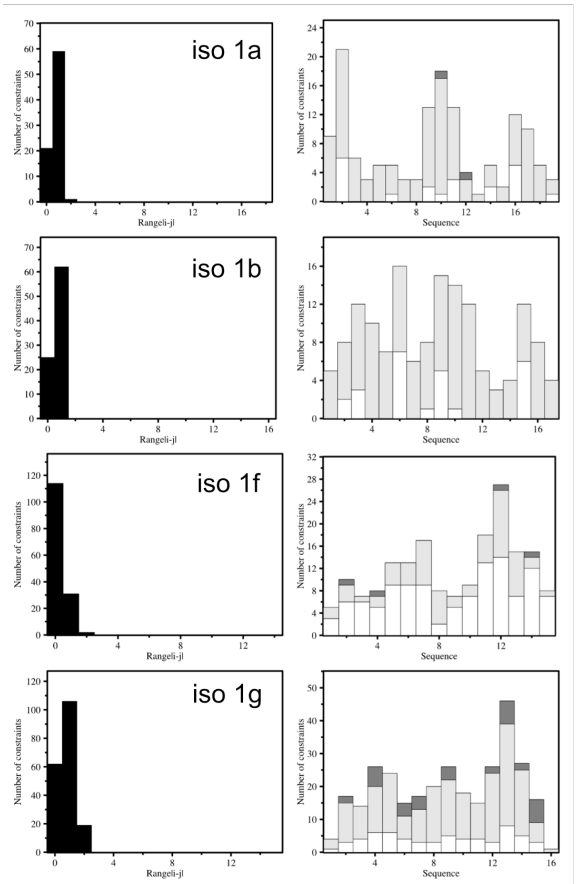

***Figure 4.80 - Number of NMR constraints distributed according to the distance range (left panel) and to the residue number in the sequence (right panel) for*:** MEG 2.1 iso 1a, MEG 2.1 iso 1b, MEG 2.1 iso 1f and MEG 2.1 iso 1g. In the right panel, short distances are displayed in white, medium distances in light grey and long distances in dark grey.

In each model of the 10-lowest energy structures, the section in which the peptide showed the most NMR constraints was highlighted (Fig. 4.81). In these most confident regions, we can observe signs of hairpins, but this structural element is most evident in iso 1a, 1f and especially 1g.

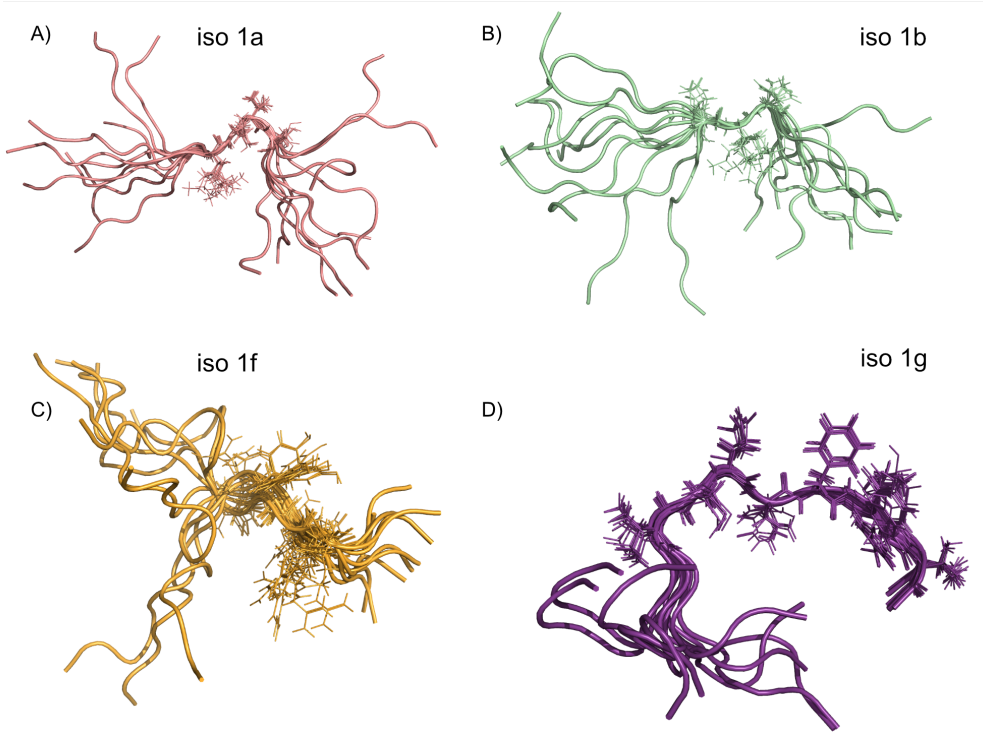

Figure 4.81 - The 10 lowest-energy structures derived by NMR using CYANA for isoform 1. A - MEG 2.1 isoform 1a (D25 - K43); B - MEG 2.1 isoform 1b (G42 - S58); C - MEG 2.1 iso 1f (S58 - R72); D - MEG 2.1 isoform 1g (M73-P88). For each peptide, the side chains of residues exhibiting the highest number of NMR constraints are displayed in lines (namely residues K33 - C36 for isoform 1a, E50 - D52 for isoform 1b, N66 - R69 for isoform 1f and E78 - T87 for isoform 1g).

Peptides of MEG 2.1 isoform 2 contain more i/i+1 and i/i+2 correlations, which leads to a higher number of intermediate and long-range constraints (Fig. 4.82) than for MEG 2.1 isoform 1 peptides. The structural refinement of iso 2a and iso 2b resulted in 10-lowest energy structures that overlap better and contain longer stretches of highlighted residues for which the highest numbers of NMR constraints were found (Fig. 4.82 and 4.83). As for the four peptides of isoform 1, the peptides of isoform 2 do not show any violated distance constraints after structure calculation using CYANA and their Ramachandran plots do not indicate any residues in the disallowed and generously allowed regions.

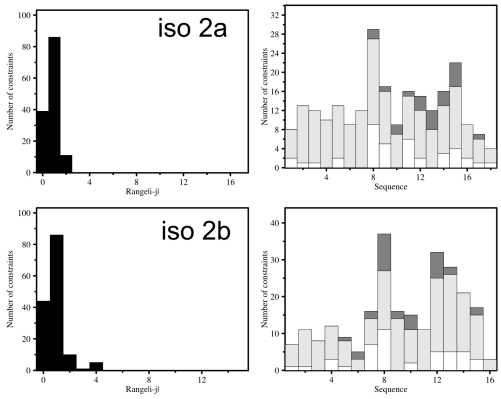

***Figure 4.82 - Number of NMR constraints distributed according to the distance range (left panel) and to the residue number in the sequence (right panel) for*:** MEG 2.1 iso 2a, MEG 2.1 iso 2b. In the right panel, short distances are displayed in white, medium distances in light grey and long distances in dark grey.

MEG 2.1 iso 2a contains 18 residues; among 212 measured NOEs, the structure of the peptide was derived based on the collection of 138 unambiguous NOEs out of 212 NOEs in total (Table 4.8). The statistical distribution in the Ramachandran plot indicates 52.5 % in most favored regions and 47.5 % in additionally allowed regions. Isoform 2a contains the highest number of constraints including short, medium, and long-range constraints for the residues from I26 to K33; in this part of the peptide all 10 lowest energy models overlay very well. At the other end of the peptide (V19-D25) the results are more varied (Fig. 4.83).

Iso 2b peptide consists of 16 residues. For this isoform 230 NOEs have been assigned, and unambiguous 141 NOEs have been used to derive its structure (Table 4.8). There are again no violated distance constraints, and the Ramachandran plot displays 0 % in disallowed and generously allowed regions. For isoform 2b the statistical distribution reveals 59.2 % residues in most favored regions and 40.8 % in additionally allowed regions. Of all the peptides, isoform 2b is the one whose calculated structures based on NMR analyses overlap best over the entire length of its sequence (Fig. 4.83B). Nevertheless, there is no secondary structure such as α-helix or β-sheet, only a turn involving residues G42-F48 is present.

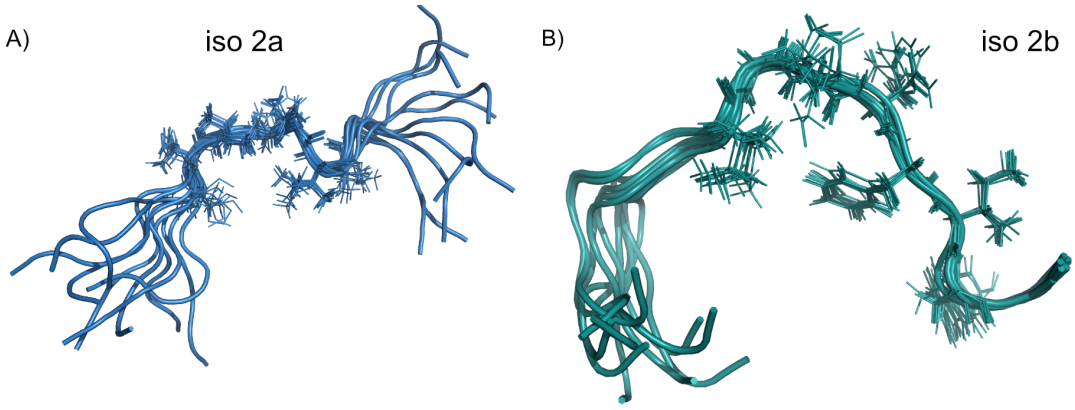

Figure 4.83 - The 10 lowest-energy structures derived by NMR using CYANA for isoform 2. E - MEG 2.1 isoform 2a (V19 - C36); and F - MEG 2.1 isoform 2b (C37 - P52). For each peptide, the side chains of residues exhibiting the highest number of NMR constraints are displayed in lines (namely residues C24-T30 for isoform 2a and I44 - Y50 for iso 2b).

The structure of MEG 2.1 isoform 3 could not be calculated, because its ^{1}H-^{1}H NOESY spectrum was completely overcrowded with H_N chemical shifts of 26 amino acids gathered in an extremely narrow spectral width of 1 ppm. This problem could not be circumvented by increasing the temperature of the NMR experiments, by adding EuFOD to the sample, nor by doubling the magnetic field strength from 14.1 T to 28.2 T.

4.3.7 Test of toxicity

In the previous subsections I described the potential antibacterial toxicity of MEG 2.1 isoform 1 and MEG 3.2 isoform 1 proteins when expressed in bacterial strains with their signal peptide. This toxicity was hypothesized due to repeated failure of transformation of some plasmids as well as a decrease in the optical density of the transformed strains

after induction of expression. Therefore, we wanted to test whether the MEG isoforms supplemented in the growth medium were as toxic as when produced by the bacteria. The tests were performed with synthetic peptides of the MEG 2.1 family (isoforms 1, 2 and 3), due to the yield and stability problem described above.

Even when chemically synthesized, MEG2.1 isoforms are particularly recalcitrant to solubilization in buffers different from DMSO. Therefore, we performed 2 controls in parallel with our tests: one without any addition to the LB medium, and one with the addition of the same quantity of DMSO used to solubilize the proteins.

In the toxicity tests we used 10 nM, 100 nM isoforms in DMSO and approx. 1 mg of lyophilized peptide powder to 10 mL of the LB medium containing the BL21(DE3) culture at logarithmic phase (OD_{600} = 0.6). In the case of the addition of lyophilized powder to the medium, its weak solubility in aqueous solutions and the differences in the molecular weight of individual peptides should be considered. For the longest isoform 1 without signal peptide (7.2 kDa) the final concentration was 14 μM, for the shortest peptide (1.8 kDa) the final concentration was 56 μM.

Three tests were performed (Fig. 4.84), which differed in the preparation of the peptide sample, which was then added to the medium during the growth (OD_{600} at 0.6). It is worth noticing that the chemically synthesized isoforms were not the full length. In order to facilitate their synthesis, we had to split the sequence into several short peptide constructs (see Fig. 4.42), therefore we performed these tests also using a combination of the peptides to mimic a “reconstruction” (see Methodology, 3.4.5 Toxicity test of extracellular MEG on bacterial cells).

However, this method has been unable to demonstrate what happens to a protein when it is expressed at the level of cellular structures. Given the fact that MEG 2.1 families are secreted proteins, it will be very challenging to obtain information about their interactions with the membrane during the secretion process. To prove the actual toxicity of MEG 2.1 family proteins it would be necessary to perform tests with expressed and purified proteins and at the level of cellular structures. Without such tests it is not possible to declare with certainty that MEG 2.1 family proteins are toxic for *E. coli* despite the obvious manifestations of toxicity described above.

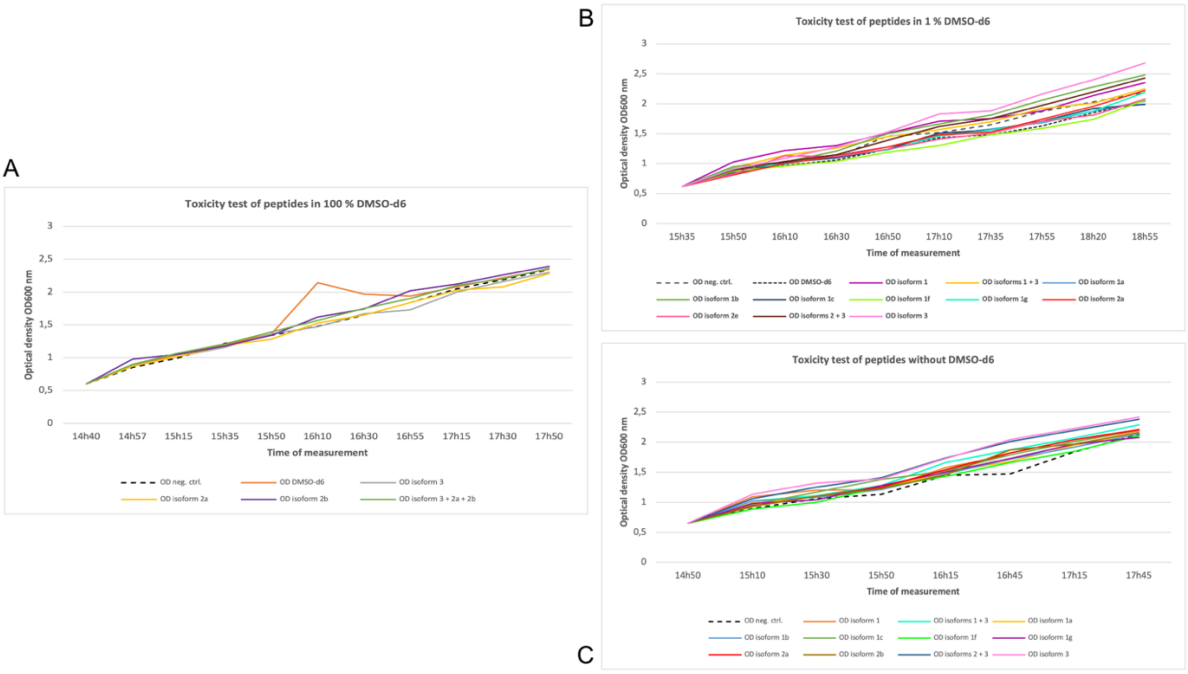

Figure 4.84 - Peptide toxicity test on E. coli BL21(DE3) expression bacteria with different conditions of peptide addition. All measurements included optical density measurements at the displayed time points and all peptide additions were performed at OD_{600} = 0.6.

A) Peptide toxicity test measured with samples prepared for NMR - 450 µL DMSO-d6 and peptide to 2 mM concentration to the final concentration 10 nM). Negative control - LB medium without the addition of peptide (dashed black), LB medium with addition of DMSO-d6 (orange), LB medium with the addition of isoform 3 in DMSO-d6 (grey), LB medium with the addition of isoform 2d in DMSO-d6, LB medium with addition of isoform 2e in DMSO-d6 (purple) and LB medium with addition of isoforms 3 and 2a and 2b in DMSO-d6 (green).

B) Peptide toxicity test measured with peptides in 100 % DMSO-d6 which were added in the culture up to 1 % volume of DMSO-d6 in the final volume of the culture. Negative control - LB medium without addition of peptide (dashed grey), LB medium with addition of 1 % DMSO-d6 (dashed black), LB medium with addition of isoform 1 (without signal peptide) in DMSO-d6 (violet), LB medium with addition of isoforms 1 and 3 in DMSO-d6 (yellow), LB medium with addition of isoform 1A in DMSO-d6 (light blue) and LB medium with addition of isoform 1b in DMSO-d6 (green), LB medium with addition of isoform 1c in DMSO-d6 (dark blue), LB medium with addition of isoform 1f in DMSO-d6 (lime), LB medium with addition of isoform 1g in DMSO-d6 (cyan), LB medium with addition of isoform 2a in DMSO-d6 (red), LB medium with addition of isoform 2e in DMSO-d6 (pink), LB medium with addition of isoforms 2 and 3 in DMSO-d6 (brown), LB medium with addition of isoform 3 in DMSO-d6 (magenta).

C) Peptide toxicity test measured with samples prepared by partial dissolution of lyophilized peptides in in the culture. Negative control - LB medium without addition of peptide (dashed black), LB medium with addition of isoform 1 without signal peptide (orange), LB medium with addition of isoforms 1 + 3 (cyan), LB medium with addition of isoform 1a (dark yellow), LB medium with addition of isoform 1b (light blue), LB medium with addition of isoform 1c (dark green), LB medium with addition of isoform 1f (lime), LB medium with addition of isoform 1g (violet), LB medium with addition of isoform 2a (red), LB medium with addition of isoform 2e (brown), LB medium with addition of isoforms 2 and 3 (dark blue), LB medium with addition of isoform 3 (magenta).

4.3.8 In silico reconstruction of full-length MEG 2.1 isoform 1 and isoform 2

After the structure of the six individual peptides was calculated, they were assembled to reconstruct the entire MEG 2.1 isoform 1 protein structure. To build this model we made use of the best-estimated structures (lowest-energy) for each of the four peptides (Fig. 4.85), and we added at the N-terminus the predicted model of AlphaFold2 signal peptide, whose plDDT was convincingly high. We resorted to the prediction because of the absence of calculated structure of MEG 2.1 isoform 3, which is the predicted signal peptide of all three MEG 2.1 isoforms. In addition, the α-helix present in the structure of isoform 1 was confirmed by CD analysis (Fig. 4.43). At the same time, the last 6 amino

acids (V21 - P26) were removed from this predicted sequence and the -FSHC- tetrapeptide was added, followed by isoforms 1a, 1b, 1f and 1g. Due to the lack of resonances in the N-terminus region of the long peptide of isoform 1 (64 aa) described above, the structure in this region cannot be confirmed nor refuted. On the other hand, at the N-terminal part, the signals of isoforms 1f and 1g overlapped very well with the peaks of the long isoform 1 and at the same time these are the peptides for which the highest number of NMR constraints was measured. For these reasons, I am inclined to conclude that the MEG 2.1 structure of isoform 1 will not differ too much from reality at its C-terminus.

The resulting model of MEG 2.1 isoform 1 (Fig. 4.85) contains structural elements of an α-helix in the region of the predicted signal peptide at the N-terminus (S4-A20) and a random coil throughout the rest of the sequence (F12-P88). This structural arrangement places it in the intrinsically disordered proteins (IDP) family (Duvaud et al.).

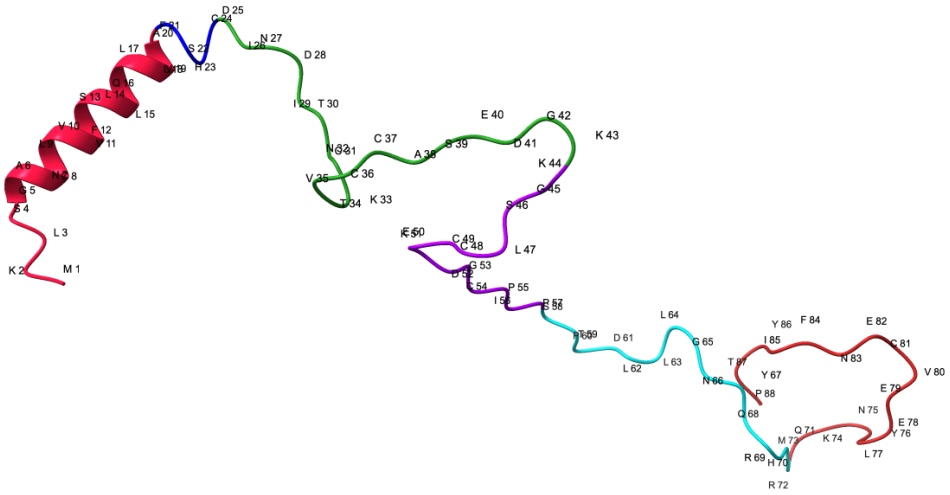

Figure 4.85 - MEG 2.1 isoform 1 complete structure built up from the measured peptides (isoform 1a - green, 1b - violet, 1f - cyan and 1g - red) and AlphaFold2 predicted signal peptide (MEG 2.1 isoform 3 - crimson). The 4 amino acids inserted (FSHC) are marked in blue.

Due to the high content of cysteine in the sequence, it is likely that disulfide bonds could be formed. In the PDB database, the cut off for creation of the disulfide bonds is set to be 3.0 Å (Sun et al. 2017). All possible combinations of distances of the four cysteines potentially involved in the stabilization of the structure by formation of disulfide bonds, Zn^{2+} ions or [2Fe2S] clusters were measured (Fig. 4.86). Despite the fact that this protein contains the majority of the secondary structure in the form of a random coil, we can observe certain regions whose conformation resembles the structural arrangement of a zinc finger type Cys_2-X_{10}-Cys_2 (T34-K51, with Zn^{2+} ions anchored by C36, C37, C48, and C49) or an almost closed loop (Y75-P88). However, the presence of stabilization of the structure by Zn^{2+} was invalidated on the basis of NMR analyses performed on MEG 2.1 isoform 1 (64 aa), thus together with this hypothesis it is possible to consider the presence of another structural motive of [2Fe2S] cluster, anchored by the same four cysteines - C36, C37, C48, C49.

However, in the case of Ferredoxin-1 structure of *Thermosynechococcus vestitus* (UniProt ID - P0A3C9, PDB - 5AUI) solved by X-ray crystallography at a resolution of 1.5 Å, the distance of both cysteine (S-S bonds) in this [2Fe2S] cluster is 3.68 Å (C40-C45) and 3.69 Å (C48-C78). Given that the measured distances between the four potentially stabilizing cysteines are several folds higher than those of ferredoxin-1, it can be concluded that this hypothesis is highly unlikely. Nevertheless, such a hairpin within an IDP is an intriguing hub/platform for putative interactions with partners.
On the other hand, the C-terminal loop is a clear candidate for a potential binding and interaction site of this protein.

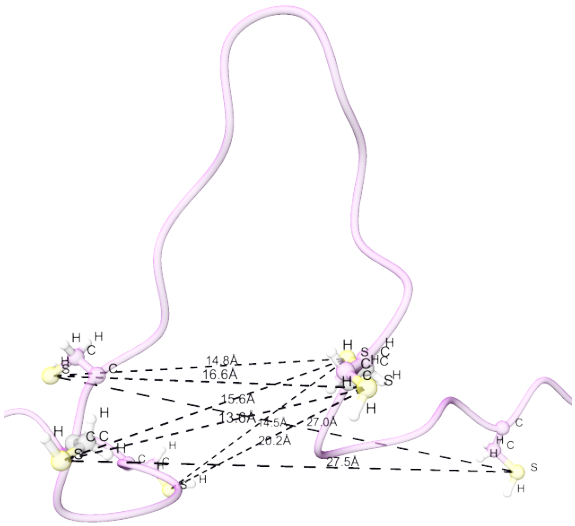

Figure 4.86 - All possible distances measured between six neighboring cysteines (C31, C36, C37, C48, C49, and C54) in the MEG 2.1 isoform 1 protein.

MEG 2.1 isoform 2 was constructed using the same principle as isoform 1, except that the protein structure was created by removing the last 8 amino acids (V19-P26) from the predicted model of isoform 3, and isoform 2a was directly linked to this model, followed by isoform 2b (Fig. 4.87). The structure again contains an α-helix at the N-terminus (predicted signal peptide) and the rest of the protein is again formed by a random coil. However, the structural similarity with MEG 2.1 isoform 1 ends here. The structure of MEG 2.1 of isoform 2 has lost the above-mentioned structural elements and the only apparent motif is the double turn formed by the connection of isoform 2a to the alpha helix and the already described turn of isoform 2b (I44 - Y50).

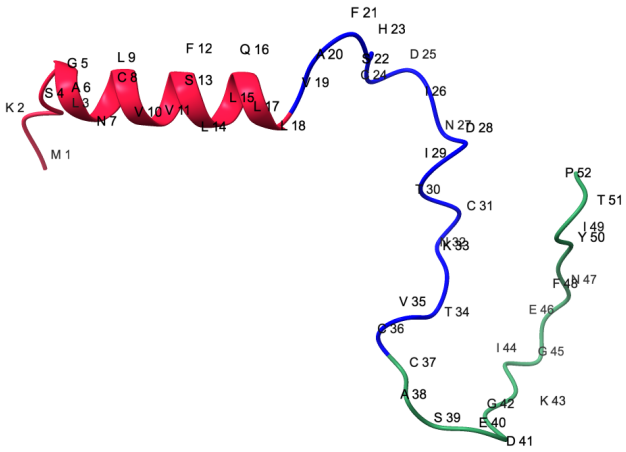

Figure 4.87 - MEG 2.1 isoform 2 complete structure built up from the measured peptides (isoform 2a - blue and 2b - green) and AlphaFold 2 predicted signal peptide (MEG 2.1 isoform 3 - crimson).

In order to verify the relevance of the re-constructed MEG 2.1 isoforms 1 and 2 and also for subsequent molecular docking, energy minimization of both structures was performed (Fig. 4.88). MEG 2.1 isoform 1 (Fig. 4.88A) before energy minimization contained 129 clashes in the model, which were reduced to 16 by minimization, while its total Van der Waals (VdW) repulsion energy was reduced from the initial 247.04 kcal/mol to final 10.43 kcal/mol. For the shorter MEG 2.1 isoform 2 (Fig. 4.88B), the total number of clashes in the model before energy minimization was 9, which was reduced to 8 by the process. Its initial total VdW repulsion energy was 6.73 kcal/mol which was reduced to the final value of 5.14 kcal/mol.

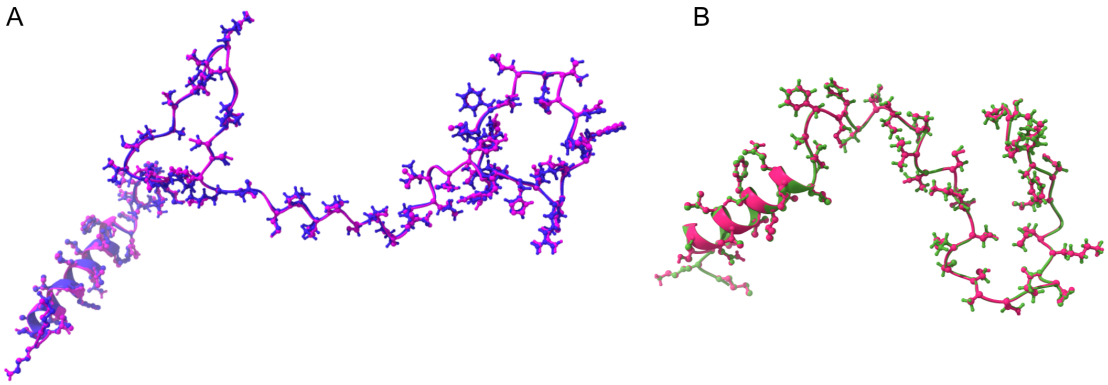

Figure 4.88 - MEG 2.1 isoform 1 (A) and MEG 2.1 isoform 2 (B) proteins. Structures are combined with the structure obtained after energy minimization (A - original structure - magenta, minimized structure - blue; B - original structure - crimson, minimized structure - green).

All subsequent binding site screens and subsequent molecular docking were performed with the minimized structure. Since the synthesis of the whole isoform 2 was not successful and we could not compare the NMR spectra and resonances of the whole isoform with those of the two peptides, further binding and docking screens were continued only with MEG 2.1 isoform 1.

Predictions of the binding pockets of MEG 2.1 isoform 1 were made (Fig. 4.89) for the subsequent molecular docking blind screen. Two possible binding sites were identified

from these predictions. One pocket (Fig. 4.89B) was found at the site described above as a potential [2Fe2S] cluster site (T34-K51); however, due to the absence of resonances in this region, the molecular docking screen was continued with only the first predicted site at the verified C-terminus of this isoform (Fig. 4.89A).

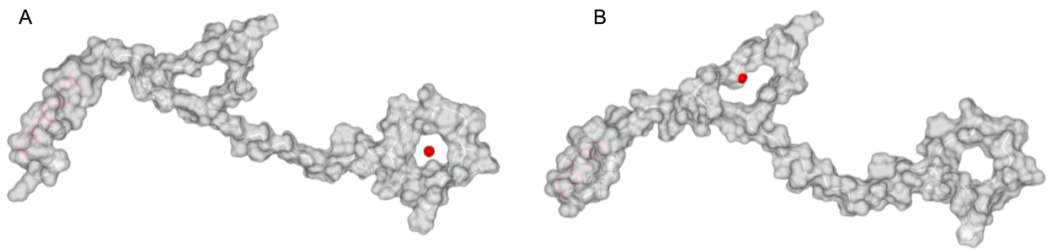

Figure 4.89 - Two predicted ligand binding pockets of MEG 2.1 isoform 1 protein. The spherical interpretation of the protein is complemented by a red ball that marks the predicted binding pocket.

Ligand n. ZINC95543764 was best docked with MEG 2.1 isoform 1 with binding energy of −11.7 kcal/mol. It was followed by a ligand n. ZINC33353312 with binding energy of -10.8 kcal/mol and ligand n. ZINC102407863 with a binding energy of -10.4 kcal/mol (Fig. 4.90). Here I only present results with binding energy of at least -10 kcal/mol.

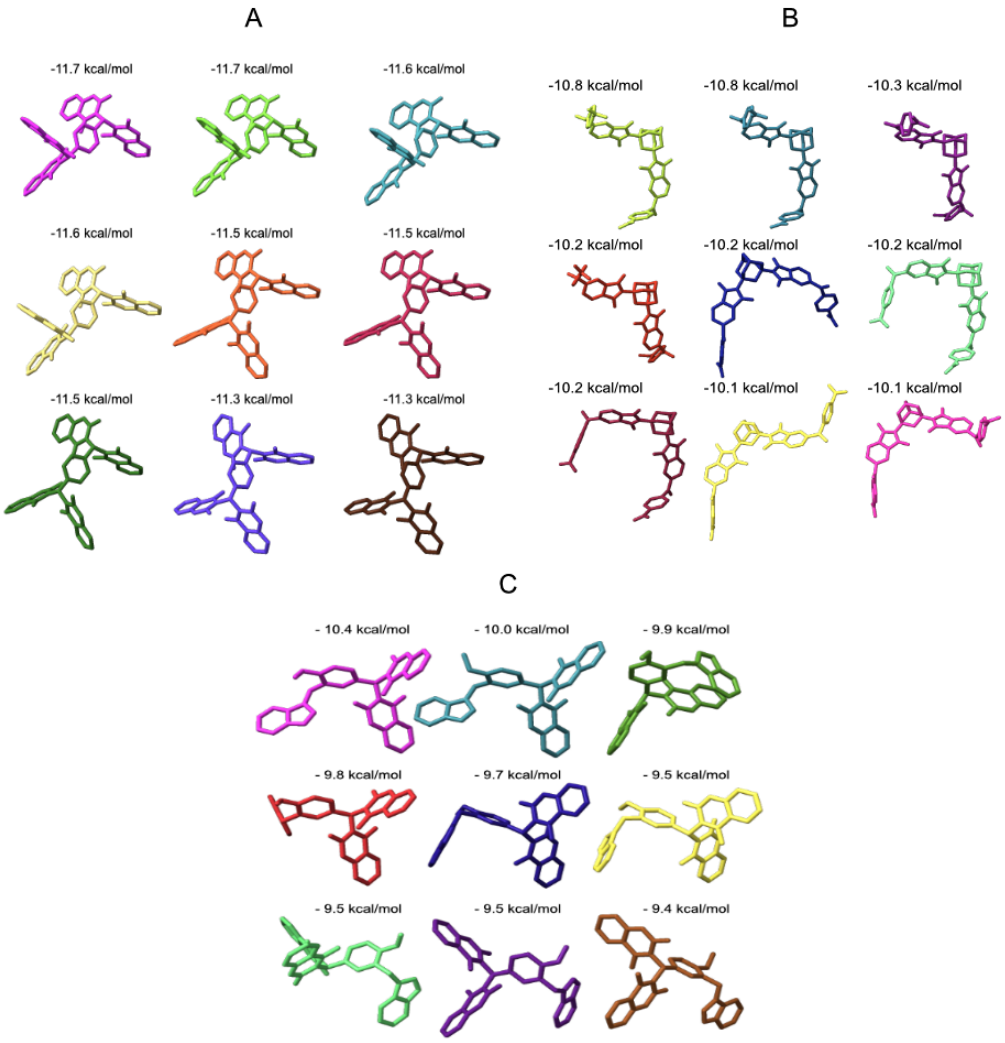

Figure 4.90 - Three molecules (and their conformers) with the best docking scores for the C-terminus binding pocket (image A in Fig. 4.89). A - ligand n. ZINC95543764 (-11.7 kcal/mol); B - ligand n. ZINC33353312 (-10.8 kcal/mol) and C - ligand n. ZINC102407863 (-10.4 kcal/mol).

It is noticeable from the results that the best docking scores are achieved by relatively large ligands with a large number of aromatic rings (Fig. 4.90, 4.91 and Table 4.9). All well docked ligands have molecular weights above 500 g/mol and logP above 5. The lowest number of aromatic rings was 3, the highest was 10, while the ligand with the best docking score contains 9 rings. The most rotatable bonds contained ligand n. ZINC98023120, which simultaneously contains only 4 aromatic rings. The lowest tPSA value was 92 $Å^2$ (ligand n. ZINC13575825) and the highest was 213 $Å^2$ (for the best docked ligand n. ZINC95543764).

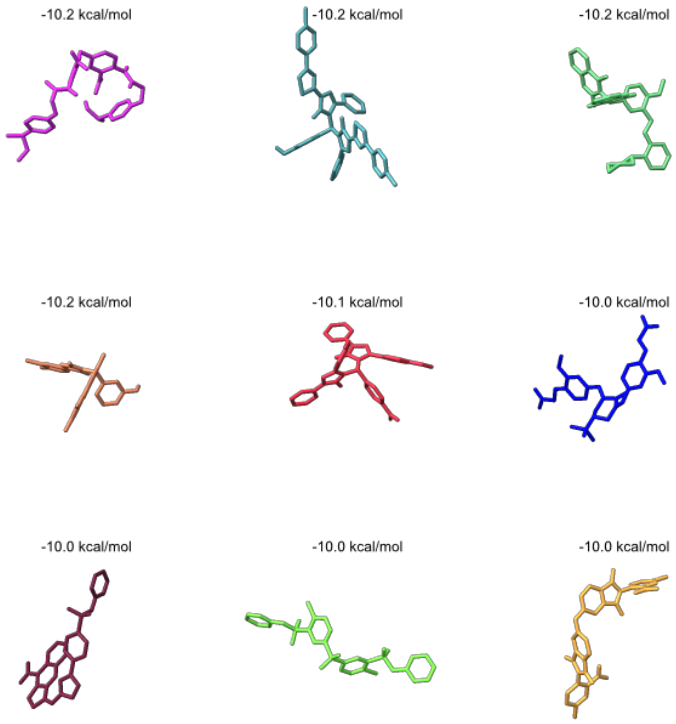

Figure 4.91 - Nine molecules with the best docking scores from the ZINC20 database for the C-terminus binding pocket (image A in Fig. 4.89). Ligand n. ZINC98023120 (magenta), ligand n. ZINC150340706 (dark cyan), ligand n. ZINC102408663 (light green), ligand n. ZINC13575825 (orange), ligand n. ZINC150411294 (red), ligand n. ZINC4418231 (blue), ligand n. ZINC102970415 (brown), ligand n. ZINC2468952 (lime), ligand n. ZINC3143000 (yellow).

In general, it is apparent that all molecules have hydrogen bond donor less than 2, hydrogen bond acceptor less than 12 and no more than 12 heteroatoms (oxygen, nitrogen, and sulphur) (Table 4.9).

Table 4.9 - Overview of the structure properties of ligands with the best docking score. The individual column labels are as follows: Ligand ZINC ID, Molecular Formula (Mol. Formula), docking score (Dock. Sc.), Number of rings (nR), nHtA (number of hetero atoms), MW (molecular weight), HbD (H-bond donors), HbA (H-bond acceptors), nRB (number of rotatable bonds), tPSA (topological polar surface area), logP.

Ligand ZINC ID	Mol. Formula	Dock. Sc. [kcal/mol]	nR	nHtA	MW [g/mol]	HbD	HbA	nRB	tPSA [Å²]	logP
ZINC95543764	C44H26O12	-11.7	9	12	746.68	0	12	6	213	7.295
ZINC33353312	C42H30N2O10	-10.8	10	12	722.706	0	10	8	189	5.529
ZINC102407863	C33H23N3O7	-10.4	7	10	573.561	0	10	6	146	5.292
ZINC98023120	C39H42N2O8	-10.2	4	10	666.771	2	8	16	156	7.736
ZINC150340706	C43H28Cl2N6O3S2	-10.2	9	13	811.776	1	9	9	118	11.243
ZINC102408663	C39H34O8	-10.2	7	8	630.693	0	8	8	125	8.126
ZINC13575825	C31H32N4O3	-10.2	5	7	508.622	1	5	5	92	5.996
ZINC150411294	C40H32N4O6	-10.1	7	10	664.718	0	8	10	130	7.699
ZINC4418231	C30H34O9	-10.0	3	9	538.593	0	9	10	134	5.123
ZINC102970415	C28H20N4O5S	-10.0	6	10	524.558	0	8	6	140	5.271

ZINC2468952	C24H18Cl2N2O6S3	-10.0	4	13	597.523	0	6	8	130	5.428
ZINC3143000	C32H20N2O9	-10.0	6	11	576.517	0	9	6	164	5.093

The best docked ligand n. ZINC95543764 is a hydroxycoumarin derivative containing 9 aromatic rings, 12 oxygen heteroatoms, does not contain any H-bond donors, but provides 12 H-bond acceptors. It has 6 rotatable bonds in its structure, it is the second largest ligand listed here with a molecular weight of 746.7 g/mol, its logP is 7.3 and tPSA is the highest (213 Å^2) of all presented ligands. From the surface view of the ligand docked in the C-terminal pocket of MEG 2.1 isoform 1, it is evident that it fills this pocket relatively well. In the "lower" region of this pocket (E79-I85) the ligand fills the pocket better than in the Y67-L77 region, where the nearly closed loop is formed by T87 and P88 (Fig. 4.92).

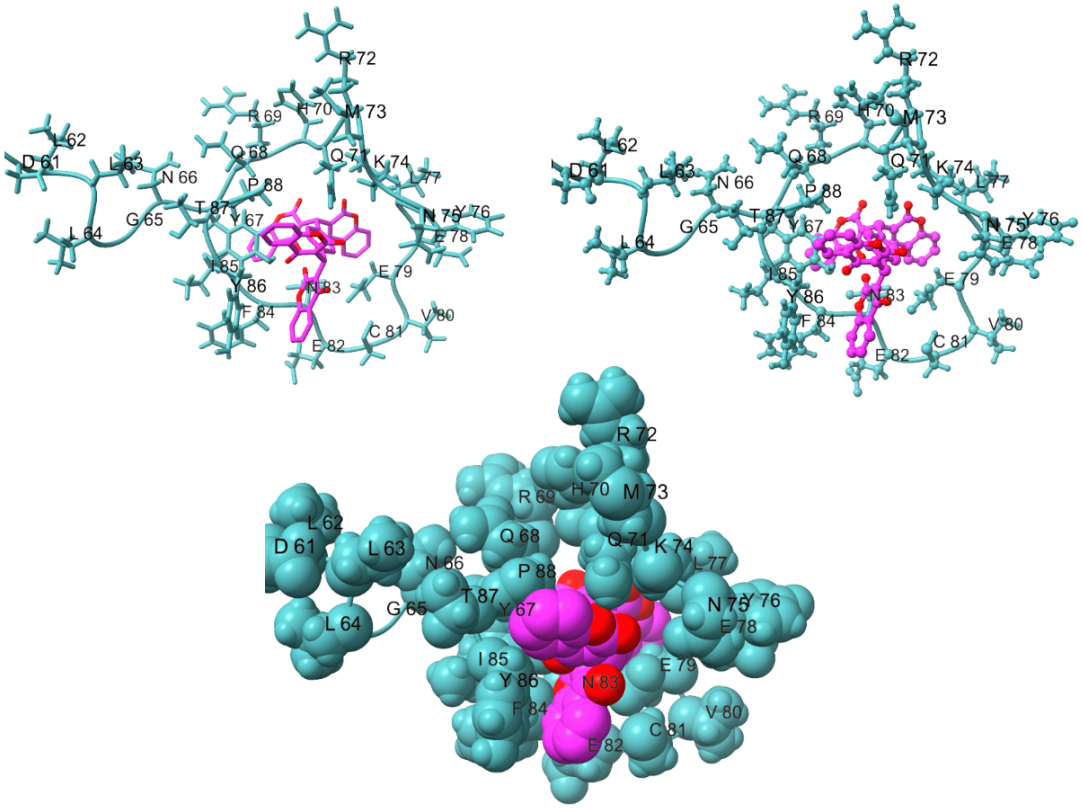

Figure 4.92 - Stick, ball stick and spheric visualization of the ligand ZINC95543764 docked (-11.7 kcal/mol, magenta with colored heteroatoms) in detail of the C-terminus binding pocket (D61 - P88) of MEG 2.1 isoform 1 protein.

Structurally quite distinct, the second-best docked ligand n. ZINC33353312 is a benzoic acid derivative, which contains the highest number of aromatic rings (10) of all the docked molecules listed here and includes an interesting highly symmetrical structure of adamantane (fusion of three cyclohexane rings) in the center of the molecule (Fig. 4.93). This ligand contains 12 oxygen and nitrogen heteroatoms, does not contain any H-bond donors, but provides 10 H-bond acceptors. It has 8 rotatable bonds in its structure, it is the second largest ligand listed here with a molecular weight of 722.7

g/mol, its logP is 5.5, and tPSA is the highest (189 Å^2). Compared to the previous described ligand, this one fills better the space formed by the side chains V80 and E78 and also the one formed by F84 and E82. On the other hand, there are some gaps in the "upper" (E78-P88) and "lower" (I85-V80) parts of the pocket (Fig. 4.93).

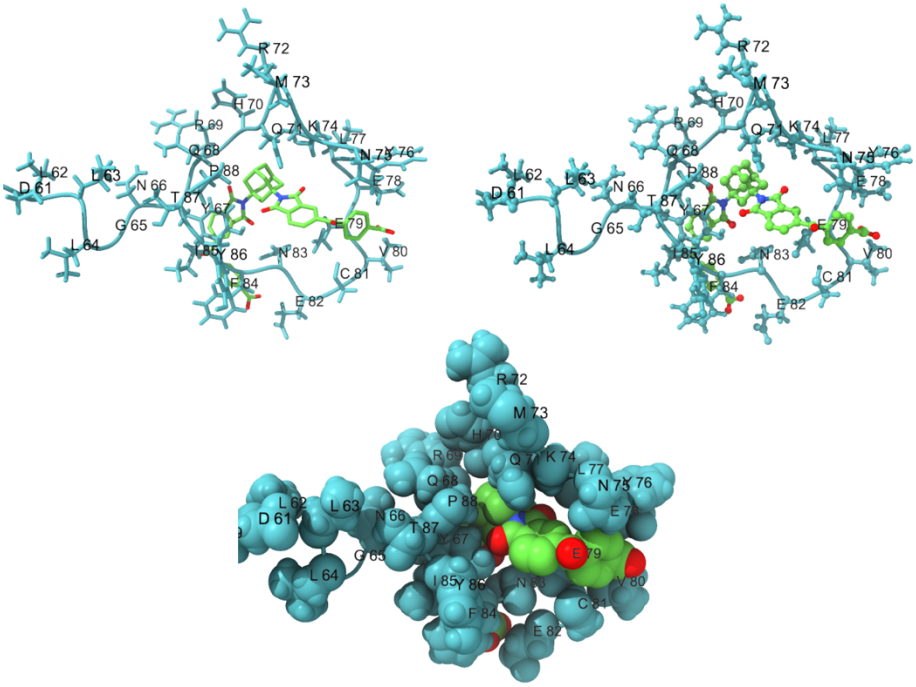

Figure 4.93 - Stick, ball stick and spheric visualization of the ligand ZINC33353312 docked (-10.8 kcal/mol, green with colored heteroatoms) in detail of the C-terminus binding pocket (D61 - P88) of MEG 2.1 isoform 1 protein.

The third best docked ligand n. ZINC102407863 is structurally more similar to ligand n. ZINC95543764. It is a derivative of 4-hydroxy-2H-chromen-2-one, which contains 7 aromatic rings, and 10 nitrogen and oxygen heteroatoms, does not contain any H-bond donors but provides 10 H-bond acceptors. It has 6 rotatable bonds in its structure, it is the second largest ligand listed here with molecular weight of 573.6 g/mol, its logP is 5.3 and tPSA is 146 Å^2. This ligand targets, like the first one described, the region around Y67 - N83 and K74 - N75. In contrast to the first ligand, it lacks two aromatic cycles and does not fill in the space from E82 - E78 and M73 - T87 (Fig. 4.94).

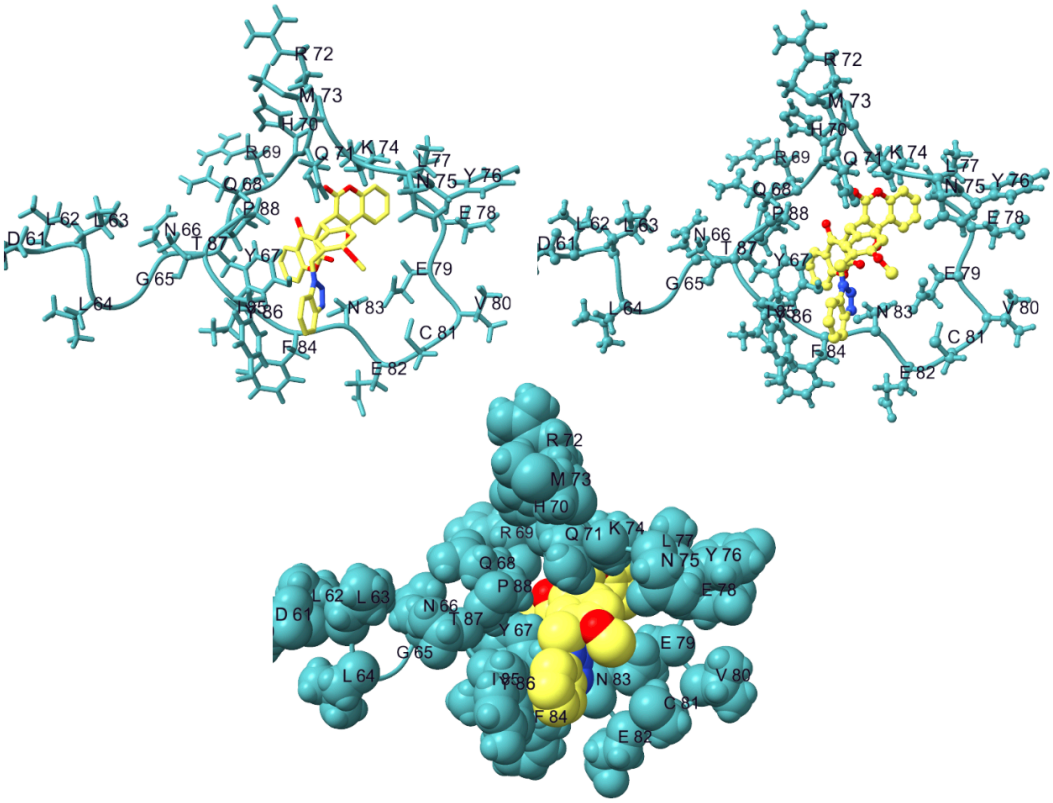

Figure 4.94 - Stick, ball stick and spheric visualization of the ligand ZINC102407863 docked (-10.4 kcal/mol, yellow with colored heteroatoms) in detail of the C-terminus binding pocket (D61 - P88) of MEG 2.1 isoform 1 protein.

4.3.9 Molecular dynamics

Molecular dynamics is a powerful tool to study conformational changes of protein structure, protein folding/unfolding, ligand binding, and behavior in a predefined time (Hollingsworth and Dror 2018). It offers insightful information that can support experimental methods and advance knowledge of the fundamental ideas about protein environment and activity in biological systems.

The structure of MEG 2.1 isoform 1 without the signal peptide (D25-P88) built up from four NMR-analyzed peptides was first minimized and equilibrated during 10 ns using High-Throughput Molecular Dynamics (HTMD) software. Then, a production run of 4 µs was performed on the system and secondary structure evolution was deduced (Fig. 4.95). At the beginning of the simulation, the structure is mainly disordered with some turns (Fig. 4.96, top) and during the MD simulation, we can observe some striking conformational changes. We can observe a formation of multiple small helices involving residues H23 to K33, S39 to L47 and E50 to G53 between 100 and 200 ns of the run. Interestingly, those three helices fold into one long stable helix (H23-G53, Fig. 4.96, bottom) after 1 µs and no conformational change occurs later (Fig. 4.95).

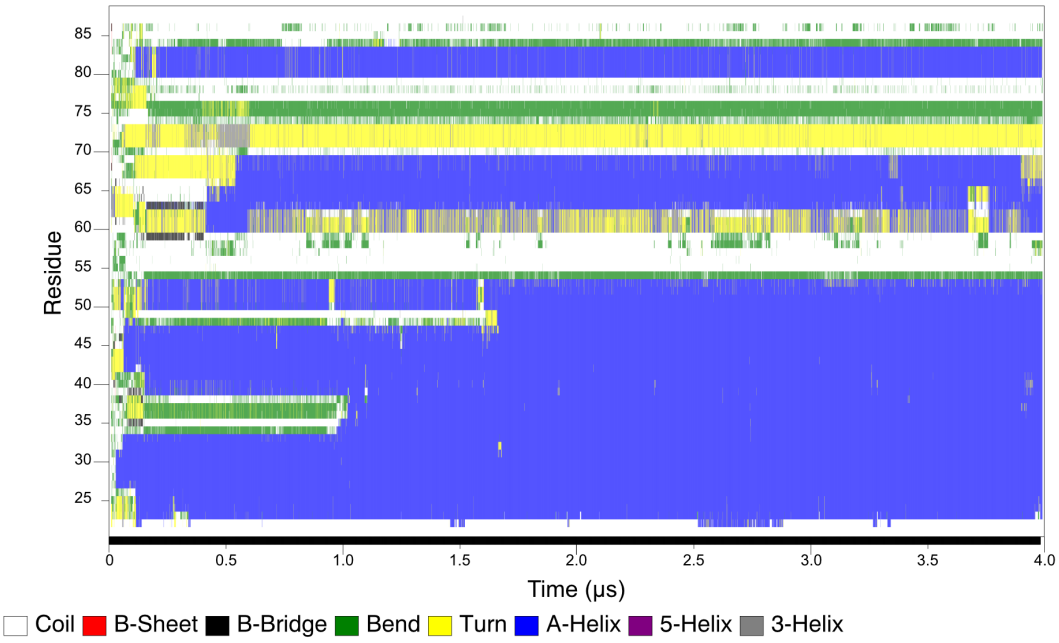

Figure 4.95 - Graphical representation of secondary structure and its changes during 4 µs of molecular dynamics of MEG 2.1 isoform 1 protein.

Also interesting is the C-terminal part, which remains mainly disordered along the MD simulation with the presence of some small helices, bends or turns. It is also possible to observe three parts (C54-P60, H70-V79, and N85-P88) in the sequence that remain random coil/turn/bend throughout the 4 µs run. We can notice that the highest number of NMR constraints were observed for residues N66 to R69 for iso 1f and V80 to T87 for iso 1g and those two regions are part of the two small new helices L64-R69 and V80-N84 along the MD trajectory. The NMR experimental data of the two short peptides, namely iso 1f and iso 1g and the MD simulation converge to similar structural features for the C-terminal part of isoform 1, thus reinforcing our choice of using this part for the virtual screening described above.

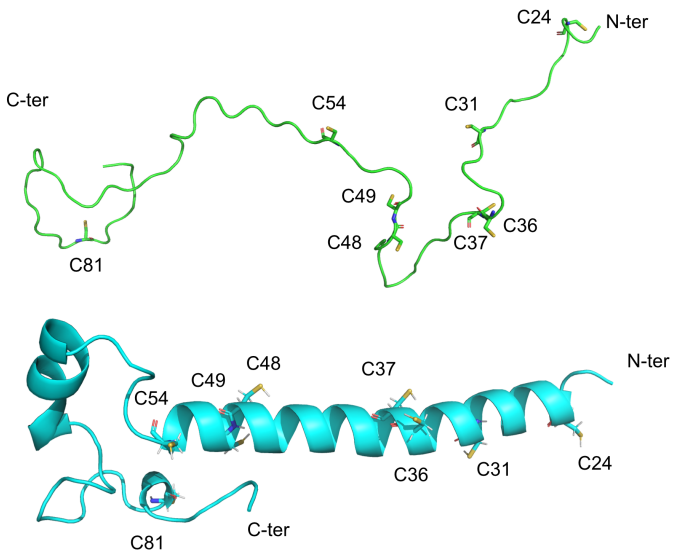

Figure 4.96 - Initial (green) and final (cyan) structure of MEG 2.1 isoform 1 subjected to molecular dynamics simulations. Both models have labelled all cysteines in the sequence.

The distance between the cysteines in the structure was also measured during the 4 µs run to verify the possibility of disulfide bond formation (Fig. 4.97). During the whole simulation no combination of cysteines was found that could form S-S bonds, because the distances were too large (due to the aforementioned 3 Å cut off for disulfide bonds).

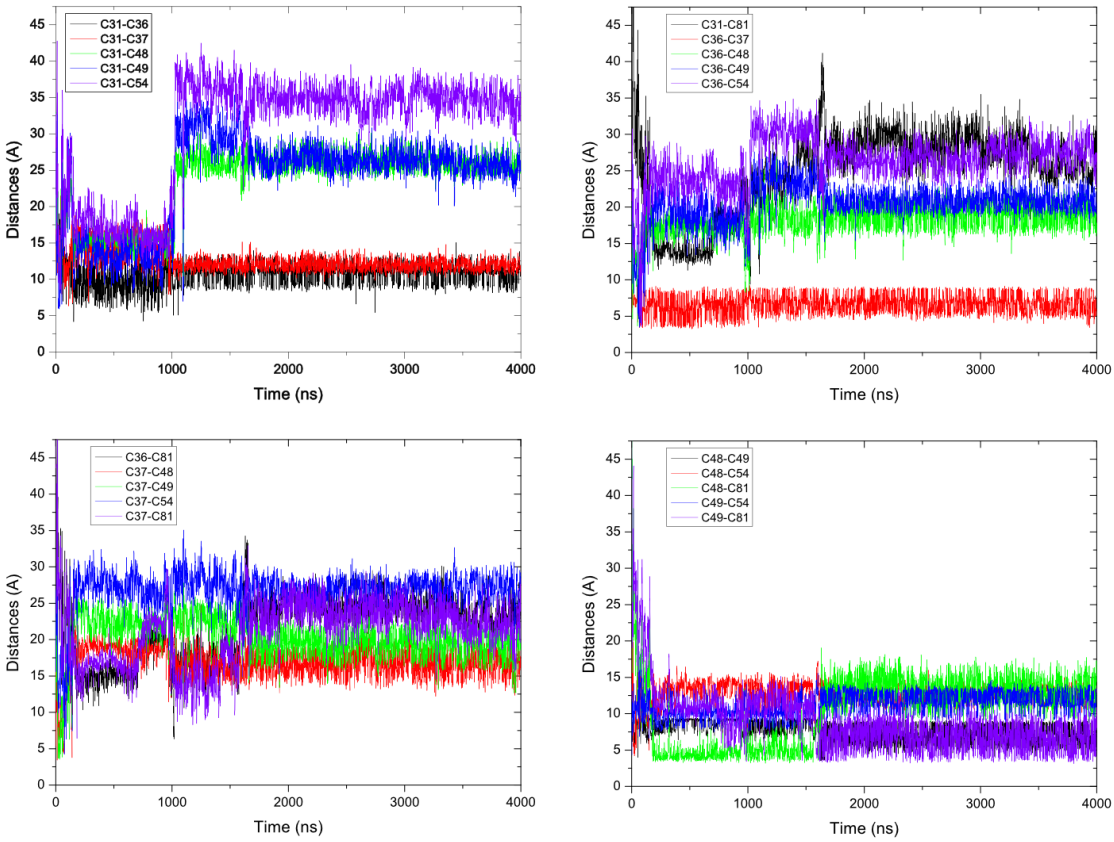

Figure 4.97 - Results of distance measurements of individual cysteines (C31, C36, C37, C48, C49, C54, C81) in the structure of MEG 2.1 isoform 1. Results are shown for all combinations of possible cysteine interactions in the molecule.

References

Anderson, Leticia, Murilo S Amaral, Felipe Beckedorff, Lucas F Silva, Bianca Dazzani, Katia C Oliveira, Giulliana T Almeida, Monete R Gomes, David S Pires, and Joao C Setubal. 2015. 'Schistosoma mansoni egg, adult male and female comparative gene expression analysis and identification of novel genes by RNA-Seq', *PLoS Neglected Tropical Diseases*, 9: e0004334.

Babij, Nicholas R, Elizabeth O McCusker, Gregory T Whiteker, Belgin Canturk, Nakyen Choy, Lawrence C Creemer, Carl V De Amicis, Nicole M Hewlett, Peter L Johnson, and James A Knobelsdorf. 2016. 'NMR chemical shifts of trace impurities: Industrially preferred solvents used in process and green chemistry', *Organic process research & development*, 20: 661-67.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, and R Dustin Schaeffer. 2021. 'Accurate prediction of protein structures and interactions using a three-track neural network', *Science*, 373: 871-76.

Balci, Metin. 2005. *Basic 1H-and 13C-NMR spectroscopy* (Elsevier).

Barron, Andrew R. 2015. 'Physical methods in chemistry and nano science'.

Berriman, Matthew, Brian J Haas, Philip T LoVerde, R Alan Wilson, Gary P Dillon, Gustavo C Cerqueira, Susan T Mashiyama, Bissan Al-Lazikani, Luiza F Andrade, and Peter D Ashton. 2009. 'The genome of the blood fluke Schistosoma mansoni', *Nature*, 460: 352-58.

Danaei, M, M Dehghankhold, S Ataei, F Hasanzadeh Davarani, R Javanmard, A Dokhani, S Khorasani, and MR Mozafari. 2018. 'Impact of particle size and polydispersity index on the clinical applications of lipidic nanocarrier systems', *Pharmaceutics*, 10: 57.

DeMarco, Ricardo, William Mathieson, Sophia J Manuel, Gary P Dillon, Rachel S Curwen, Peter D Ashton, Alasdair C Ivens, Matthew Berriman, Sergio Verjovski-Almeida, and R Alan Wilson. 2010. 'Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts', *Genome research*, 20: 1112-21.

Dunker, A Keith, M Madan Babu, Elisar Barbar, Martin Blackledge, Sarah E Bondos, Zsuzsanna Dosztányi, H Jane Dyson, Julie Forman-Kay, Monika Fuxreiter, and Jörg Gsponer. 2013. 'What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered', *Intrinsically disordered proteins*, 1: e24157.

Duvaud, Séverine, Chiara Gabella, Frédérique Lisacek, Heinz Stockinger, Vassilios Ioannidis, and Christine Durinx. 2021. 'Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users', *Nucleic Acids Research*, 49: W216-W27.

Dyson, H Jane, and Peter E Wright. 2021. 'NMR illuminates intrinsic disorder', *Current Opinion in Structural Biology*, 70: 44-52.

Fneich, Sara, André Théron, Céline Cosseau, Anne Rognon, Benoit Aliaga, Jérôme Buard, David Duval, Nathalie Arancibia, Jérôme Boissier, and David Roquis. 2016. 'Epigenetic origin of adaptive phenotypic variants in the human blood fluke Schistosoma mansoni', *Epigenetics & Chromatin*, 9: 1-13.

Hollingsworth, Scott A, and Ron O Dror. 2018. 'Molecular dynamics simulation for all', *Neuron*, 99: 1129-43.

Howe, Kevin L, Bruce J Bolt, Myriam Shafie, Paul Kersey, and Matthew Berriman. 2017. 'WormBase ParaSite− a comprehensive resource for helminth genomics', *Molecular and biochemical parasitology*, 215: 2-10.

Hull, Rodney, and Zodwa Dlamini. 2014. 'The role played by alternative splicing in antigenic variability in human endo-parasites', *Parasites & vectors*, 7: 1-19.

Jacobsen, Neil E. 2007. *NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology* (John Wiley & Sons).

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and Anna Potapenko. 2021. 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596: 583-89.

Kristjansson, MM, and JE Kinsella. 1991. 'Protein and enzyme stability: structural, thermodynamic, and experimental aspects', *Advances in food and nutrition research*, 35: 237-316.

Lin, Xingcheng, Prakash Kulkarni, Federico Bocci, Nicholas P Schafer, Susmita Roy, Min-Yeh Tsai, Yanan He, Yihong Chen, Krithika Rajagopalan, and Steven M Mooney. 2019. 'Structural and dynamical order of a disordered protein: Molecular insights into conformational switching of page4 at the systems level', *Biomolecules*, 9: 77.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, and Yaniv Shmueli. 2023. 'Evolutionary-scale prediction of atomic-level protein structure with a language model', *Science*, 379: 1123-30.

Lu, Zhigang, Geetha Sankaranarayanan, Kate A Rawlinson, Victoria Offord, Paul J Brindley, Matthew Berriman, and Gabriel Rinaldi. 2021. 'The transcriptome of Schistosoma mansoni developing eggs reveals key mediators in pathogenesis and life cycle propagation', *Frontiers in tropical diseases*, 2: 713123.

Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. 'lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests', *Bioinformatics*, 29: 2722-28.

Myers, Jeffrey K, C Nick Pace, and J Martin Scholtz. 1998. 'Trifluoroethanol effects on helix propensity and electrostatic interactions in the helical peptide from ribonuclease T1', *Protein Science*, 7: 383-88.

Philippsen, Gisele S, R Alan Wilson, and Ricardo DeMarco. 2015. 'Accelerated evolution of schistosome genes coding for proteins located at the host–parasite interface', *Genome Biology and Evolution*, 7: 431-43.

Schwartz, Russell, Claire S Ting, and Jonathan King. 2001. 'Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life', *Genome research*, 11: 703-09.

Sun, Ming-an, Yejun Wang, Qing Zhang, Yiji Xia, Wei Ge, and Dianjing Guo. 2017. 'Prediction of reversible disulfide based on features from local structural signatures', *BMC genomics*, 18: 1-10.

Uversky, Vladimir N, Christopher J Oldfield, and A Keith Dunker. 2008. 'Intrinsically disordered proteins in human diseases: introducing the D2 concept', *Annu. Rev. Biophys.*, 37: 215-46.

Waterhouse, Andrew, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, and Lorenza Bordoli. 2018. 'SWISS-MODEL: homology modelling of protein structures and complexes', *Nucleic Acids Research*, 46: W296-W303.

Wilson, Carter J, Wing-Yiu Choy, and Mikko Karttunen. 2022. 'AlphaFold2: a role for disordered protein/region prediction?', *International journal of molecular sciences*, 23: 4591.

Wright, Peter E, and H Jane Dyson. 2015. 'Intrinsically disordered proteins in cellular signalling and regulation', *Nature reviews Molecular cell biology*, 16: 18-29.

Zwang, Julien, and Piero Olliaro. 2017. 'Efficacy and safety of praziquantel 40 mg/kg in preschool-aged and school-aged children: a meta-analysis', *Parasites & vectors*, 10: 1-16.

List of figures

List of tables

5 DISCUSSION

Thanks to next-generation sequencing technologies, the sequencing of entire genomes of a number of species has become possible, and thanks to RNAseq deduction of actively transcribed genes is being achieved (Mortazavi et al. 2008; Wang, Gerstein, and Snyder 2009; Filichkin et al. 2010; Harr and Turner 2010). In connection with the sequencing of a number of parasite genomes, scientists have started to focus not only on transcriptomic analyses but also on the evolution of mRNA and its post-transcriptional regulation. Thanks to alternative splicing, the diversity of the proteome, which plays an essential role in a number of biological processes, is deployed (Hagiwara 2005). Alternative splicing has played an essential role in evolution (Kelemen et al. 2013), and it seems that in the human parasite *S. mansoni*, there are groups of genes (encoding MEG and VAL proteins) that have undergone accelerated evolution (Philippsen, Wilson, and DeMarco 2015) due to their parasitic lifestyle, during which they try to remain hidden from the host's immune system until successful reproduction. Alternative splicing has been intensively studied in a number of parasites and has been repeatedly pointed out as a means of adapting to the host environment and especially of escaping the attention of the immune system (Piao et al. 2014; Cai et al. 2008; Sorber, Dimon, and DeRisi 2011; Nilsson et al. 2010; Yeoh et al. 2015). A highly interesting group of schistosome-specific proteins are MEG proteins, which are encoded by genes whose sequence is up to 75 % composed of short symmetric exons. This structure gives rise to a number of alternatively spliced variants, which leads to the generation of high protein variability (Berriman et al. 2009; DeMarco et al. 2010; Wilson 2012; Wilson et al. 2015).

The aim of this dissertation was to investigate the function and structure of some candidates of *S. mansoni* MEG-family proteins. For this purpose, it was necessary to express, purify and then analyze selected MEG proteins using several techniques. The final goal of this work was to find potential host interaction partners for these selected proteins.

First, a comprehensive bioinformatic analysis of the primary protein sequence was performed to better understand and categorize the MEG proteins. It also confirmed the unique character of these Schistosomal genes and proteins. Subsequently, various expressions of MEG 2.1, MEG 3.2, and MEG 6 proteins were tested and optimized. These three candidate MEG-family proteins were selected at the beginning of this work on the basis of transcriptomic analyses performed in the laboratory of Dr. Jan Dvorak. The last steps were the structural analysis of selected MEG-family proteins using circular dichroism (CD), dynamic light scattering (DLS), and nuclear magnetic resonance (NMR) spectroscopy. Finally, preliminary tests of molecular docking and molecular dynamics (MD) on the MEG 2.1 isoform 1 were performed.

These objectives proved to be very ambitious, given the very low number of publications on this theme. At the same time, *S. mansoni* immune-evasion/immune modulation has been studied for years and is still not fully understood; thus, it will require much more complex research. All these studies are an important part of research not only on the parasitic lifestyle but also on the human (host) immune system, in which many research gaps still remain.

Bioinformatic analysis pointed out the problem of multiplicative and sometimes not strictly logical nomenclature of schistosome MEG proteins. Therefore, in one of our publications, we proposed a categorization and clarification of the nomenclature of these proteins. The problem with multiple names for a single protein probably occurred during the manual annotation/uploading of genes and proteins to the UniProt and WormBase ParaSite databases without cross-checking. In this thesis, 35 *meg* genes and their positioning on chromosomes were annotated and commented on, which provides information not only about the possible evolution of MEG genes but also explains some structural protein characteristics.

Alignment and subsequent phylogenetic analysis of all 87 validated MEG proteins showed that the MEG family is divided into two large subfamilies based on protein sequence similarities. These two subfamilies differ in the conserved amino acids and motifs, which are better observable after this division into subfamilies. Another marked difference is that one subfamily is specific in its sub-branching (gene duplication leading to the sequence diversification), and the other one is composed of sequences whose diversity was mostly created by alternative splicing. On the contrary, some characteristics are noticeable in the whole MEG superfamily: for example, almost all MEG proteins contain a signal peptide on the N-terminus, and the vast majority contain a high proportion of cysteine in their sequences. Another common feature of all MEG proteins is that they are homologous only to other annotated MEG proteins of the genus *Schistosoma*. Alignment of the studied MEG 2.1 isoform 1, MEG 3.2 isoform 1, and MEG 6 proteins showed a lower percentage sequence identity also with the hypothetical proteins of the genus *Trichobilharzia* for the MEG 3.2 isoform 1 protein. Thus, the relationship remains in the *Schistosomatidae* family of parasites.

None of the three investigated proteins had sufficient homology to proteins with experimentally solved structures to allow their homology modelling. For that reason, *ab initio* predictions were made using three leading deep-learning prediction servers. The results of these predictions varied considerably: the most for MEG 6 and the least for MEG 2.1 isoform 1. This can again be explained by the lack of their homology and the lack of recognizable structural elements and sequence motifs that are necessary for relevant *ab initio* predictions (Ruff and Pappu 2021; Azzaz et al. 2022). The predicted models differed not only among the results of the competing prediction servers but also within multiple models generated by the same prediction server. The prediction error was, therefore, high and reached its maximum in non-structural parts of the sequences.

Recombinant expression of MEG 2.1 isoform 1, MEG 3.2 isoform 1, and MEG 6 was tested in bacterial, yeast, insect, and cell-free expression systems. MEG 2.1 isoform 1 without signal peptide, with SLIN tag and MEG 3.2 isoform 1 with a signal peptide and 6xHis tag and TEV cleavage sequence were expressed, though in very low yield (Methodology 3.2 Recombinant protein expression and Results 4.2 Recombinant protein expression). MEG 6 has never been expressed, despite being a protein without cysteine and without a predicted signal peptide. The absence of the signal peptide is very interesting because it has been repeatedly reported to be a protein secreted from *S. mansoni* eggs (Berriman et al. 2009). At the same time, MEG 6 is the one containing the most probable N-glycosylation markers of the three proteins

studied, suggesting that the protein is indeed secreted (Varki et al. 2022). Moreover, MEG 6 is also interesting because it is encoded by a gene that is the only one located on the sexual chromosome (ZW) of *S. mansoni*. Phylogenetically, MEG 6 belongs to the subfamily whose diversity is due to gene duplication. For MEG 6, a trend of higher expression in mature eggs than in immature ones was observed. This trend was also observed for MEG 2.1 (all three isoforms) and MEG 3.2 (all of ten isoforms) (Introduction 1.5 Schistosomal MEG family proteins, Fig. 1.6). For this reason, it can be assumed that these proteins are involved in the interaction of mature eggs with the host tissue, most likely during the penetration of the intestinal wall before their excretion from the body. This hypothesis has also been previously suggested in the literature (Wilson 2012; DeMarco et al. 2010).

Unfortunately, MEG 2.1 isoform 1 was successfully expressed only in the S2 insect expression system. This one was not a suitable system for the subsequent isotopic labeling for NMR analyses, which was one of the objectives of this dissertation. Analysis of the secondary structure of MEG 2.1 isoform 1 with the SLIN tag indicated the presence of α-helix, but the instability and low quantity of the single sample (shipped to Lyon from Prague) did not allow to repeat this measurement. MEG 3.2 isoform 1 was expressed in the bacterial expression system after long optimizations, but unfortunately, the necessary amount and concentration of this protein, whether isotopically labeled or not, was never obtained. MEG 3.2 isoform 1 was stable only in buffers with a very high NaCl content; below 0.5 M its precipitation was almost immediate and visible. This fact greatly complicated the subsequent biophysical analyses, for which such high salt concentrations are an obstacle. Both expressed proteins have been shown to be recalcitrant to express in all of the tested expression systems, including cell-free. The only expression system that showed decent yields of the expressed protein without the need for further optimization was the S2 insect expression system. One possible explanation could be that this system was developed to express difficult targets (Scotter et al. 2006; Adriaan de Jongh, Salgueiro, and Dyring 2013). At the same time, it was certainly advantageous that this expression system had its own *Drosophila* signal peptide, which mediates the secretion of proteins into the medium and thus does not cause cell death when secreting toxic proteins.

Some MEG 2.1 isoform 1 and MEG 3.2 isoform 1 constructs did not transform into the BL21 expression bacteria, and others showed a rapid decrease in optical density after IPTG induction. This data suggested that both MEG 2.1 isoform 1 and MEG 3.2 isoform 1 could be toxic. This theory was also supported by bioinformatic analyses that showed the homology of MEG 3.2 isoform 1 with the scorpion neurotoxin Cn11. The structure of peptides/small proteins, which is cysteine-rich, is well known for toxic/antibacterial proteins of spiders (Mandard et al. 2002), scorpions (Kobayashi et al. 1991), wasps (Özbek et al. 2019), frog skin (Simmaco et al. 1994; Morikawa, Hagiwara, and Nakajima 1992), snakes (Reeks, Fry, and Alewood 2015), among others (Dimarcq et al. 1998).

Moreover, both these proteins aggregated after multiple days of storage at -20 °C or -80 °C degrees. Bands of stored proteins reloaded to the gels appeared to have approximately twice the molecular weight of the original bands of the expressed proteins. These bands were

evident after the initial purifications of both MEG 2.1 isoform 1 and MEG 3.2 isoform 1 proteins, but with the progressing time and storage at -20 °C there was a visible progressive increase in the band size of the double-sized fraction, compared to the original band of the expressed protein. This trend was observed for both proteins despite the fact that the samples were denatured at high temperature before loading onto the gel, and at the same time, the gels were run under denaturing and reducing conditions. For these reasons, I believe that dimers are formed during the degradation of those two proteins. For MEG 3.2 isoform 1, it is possible that there are not only dimers but multimers formed during its storage; despite the considerably high NaCl concentration, many higher molecular weight proteins remained on gels after the purifications. This theory is only a hypothetical possibility that has not been confirmed by mass spectrometry. However, the overlay of the proton spectra of the monomer and the hypothetical dimer of MEG 3.2 isoform 1 showed that this theory was not entirely false because the spectra overlapped considerably well (Results 4.2.1.1.2, Fig. 4.34).

A two-year effort to express one of the studied proteins in sufficient concentration, purity, and expression system suitable for the following isotopic labeling did not lead to success and therefore, in parallel with the optimization of the expression of MEG 3.2 isoform 1, the strategy of peptide synthesis of the whole MEG 2.1 family (i.e., three isoforms) was chosen. MEG 2.1 family was chosen because of its relatively short sequence, since the longest (all micro-exons covering) isoform 1 has 88 amino acids. At the same time, it was possible to demonstrate on this family how much the individual isoforms differ structurally as a result of alternative splicing, which would not have been possible with MEG 6, for which only one isoform has been reported.

However, this strategy also required solving several challenges: first of all, it was not possible to synthesize MEG 2.1 isoform 2 in its full length (52 amino acids) even after repeated attempts by Genosphere. For this reason, this isoform was subsequently split into two peptides and ordered without the signal peptide. The longest isoform 1 was first ordered as a whole, but without the signal peptide (64 aa from residue 25 to 88), then it was divided into three peptides: iso 1a, iso 1b, and iso 1c. Subsequently, iso 1c was further divided into two shorter peptides: iso 1f and 1g. Isoform 3 was ordered in its entirety. This isoform is the signal peptide of the other two isoforms. This strategy allowed us to obtain assignable NMR spectra of the whole MEG 2.1 family and reconstruct their 3D structure.

Another challenge that had to be overcome when working with synthetic peptides was their extremely poor solubility in physiological buffers. At the same time, they were not soluble in any commonly used organic solvents. The only solvent able to completely dissolve the peptides at a high concentration (2 mM) turned out to be 100% dimethyl sulfoxide (DMSO). DMSO is certainly not the solvent of first choice for the study of biomolecules, but in this case, despite many efforts, no suitable alternative was found.

In order to analyze the secondary structure of peptides, acetonitrile was used, which was suitable for collecting CD spectra because the required final concentration is several times lower than the concentration necessary for NMR measurements of the peptide in the natural abundance of ^{13}C (1.1%) and ^{15}N (0.4%).

On top of those, one of the peptides (iso 2a) proved to be unstable and degraded within one week, which was the time needed to measure the NMR spectra for each peptide.
This time demand of one week per peptide is a specificity of NMR measurement in the natural abundance of carbon and nitrogen isotopes. It should be noted that the optimization and selection of suitable conditions for the measurement of the presented spectra also took several months. For each of the peptides, at least ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, ^{1}H-^{1}H NOESY, and ^{1}H-^{1}H TOCSY were measured. Most of the peptides were also subjected to ^{1}H-^{13}C HSQC-TOCSY experiments. After detecting the instability of one of the peptides, 1D measurements were also inserted between each recorded spectra to check the stability of the samples. The short peptide strategy was successful because it overcame the lack of signal in the undivided longer peptides. At the same time, thanks to alternative splicing of MEG 2.1 family, it was possible to overlay the spectra of isoform 1 and isoform 2 peptides so that some resonances were cross-checked due to sequence overlaps.
In addition to NMR and CD analyses, very basic peptide toxicity tests were also performed on BL21(DE3) bacterial cultures. These tests were performed based on the presumed toxicity of some of the constructs of MEG 2.1 isoform 1 and MEG 3.2 isoform 1, whose plasmids either failed to be transformed into the bacteria or induced cells death after the addition of IPTG. Despite different ways of adding synthetic peptides to the bacterial culture media, no bacterial death was detected. From the results, it is evident that the presence of the signal peptide in these two MEG proteins is possibly the cause of toxicity, and therefore it is very important to choose suitable constructs for bacterial expression.
In addition, MEG 2.1 isoform 1 has never been expressed in any bacterial strain (even with the addition of rifampicin), and during expression in Rosetta(DE3) strain, it was almost impossible to achieve culture multiplication for induction and subsequent harvesting. For this reason, I venture to hypothesize that MEG 2.1 isoform 1 is more toxic for bacteria than MEG 3.2 isoform 1. In order to verify this hypothesis, it would, of course, be necessary to express both proteins and subject them to a series of more adapted toxicity tests.
The last obstacle that stood in the way of a successful assignment of all measured peptides was the fact that their measured resonances were dispersed in a very narrow spectral range, most frequently around 0.6 and at the maximum of 1 ppm. This fact complicated the assignment of all the spectra, especially the ^{1}H-^{1}H NOESY spectra, which are crucial for the subsequent determination of the 3D structure using automated structural calculation. Indeed, the measure of distance constraints (NOE), in combination with the assignment of ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC spectra, allows us to evaluate the conformational constraints followed by structural calculations. Longer peptides NMR spectra were already considerably overcrowded in a narrow spectral range in ^{1}H-^{1}H TOCSY and in ^{1}H-^{1}H NOESY spectra. This led to the overlapping of peaks of hydrogen resonances of individual amino acids with intermolecular/intramolecular NOEs. This problem was overcome for most of the peptides, but unfortunately, the structure of MEG 2.1 isoform 3 was not possible to be determined even after many various attempts to optimize the measurement. According to the CD analyses, it was undoubtedly found that this isoform is in the conformation of an α-helix, as it was

predicted from the beginning by all bioinformatic software for the determination of secondary and tertiary protein structure.

After successful assignment and structural refinement, all MEG 2.1 peptides of both isoform 1 and isoform 2 were confirmed to be IDP. The C-terminus of isoforms 1 and 2 (iso 1g and iso 2b) folds into a hairpin. This structural similarity is due to the partial overlapping of the 7 terminal amino acids. For both peptides, the hairpin structure is formed approximately in the middle of the sequence; for peptide 1g, it is the sequence V80-T87, while for 2b, it is I44-Y50. The other peptides also contain one turn each, but overall they are linear, and their ten lowest-energy models differ in these very flexible parts of the sequences (N-terminus and C-terminus), while in the case of iso 1g and 2b, the models fit very well.

The final proposed structure of MEG 2.1 isoform 1 and MEG 2.1 isoform 2 was assembled from these peptides. To build both of those models, it was necessary to supplement the experimentally obtained structures of four peptides for isoform 1 and two peptides for isoform 2 with a signal peptide. Due to alternative splicing, the signal peptide of these two peptides is MEG 2.1 isoform 3 (which, in addition to the sequence of the signal peptide, also contains 6 amino acids at the C-terminus). Since we could not determine the structure of isoform 3 for the reasons described above, I solved this obstacle by adapting the AlphaFold2 prediction models for this isoform. After adaptation of the signal peptide for both isoforms and assembly of their final structure, the relevance of both models was tested using energy minimization. Isoform 1 contained a higher number of clashes, and its minimization was slightly better than for isoform 2, where minimization saved about 1 kcal/mol. It is worth mentioning that the energetics of non-structural proteins is still a very challenging area of experimental computational strategy (Zou, Simmerling, and Raleigh 2019). Without experimental verification of the whole isoforms, these presented models remain only preliminary suggestions of their 3D structure.

For MEG 2.1 isoform 2, it was not possible to achieve a comparison of the whole isoform structure with that of the two individual peptides (due to the problems with the synthesis of the whole sequence of this isoform). However, this possibility was available for MEG 2.1 isoform 1. Isoform 1 without the signal peptide was again a challenge for the overall assignment because the resonances at the N-terminus were again overlapping in a narrow spectral range. On the other hand, it was possible to confirm that the resonances at the C-terminus overlapped between those of iso 1f and 1g. This confirmed that the strategy of dividing complex IDPs into shorter peptides was, in our case, a suitable method for determining at least the partial structure of these molecules.

Here it can be argued that the structure of the MEG 2.1 proteins studied here contain a significant number of cysteines; therefore the proposed reconstructed structures could not be relevant due to the possibility of disulfide bond formation. This theory has been tested in several ways. Primarily, the distance combinations between all the cysteines that could form these S-S bonds were measured, and it was found that they were too large for any interaction. NMR measurement of the possible disulfide bond reduction by the addition of TCEP to the sample of isoform 1 was tested, but unfortunately, the sample was degraded during the

measurements. The formation of S-S bonds between individual molecules of isoform 1 cannot be rejected nor confirmed. However, the concentration of the measured peptides was high, and no conformational changes of the peptides in the samples were observed, so either all molecules in the sample were bound, or all were unbound.

Apart from disulfide bonds, $Cys_2X_{10}Cys_2$ motifs can be stabilized in other ways, such as Zn fingers or [2Fe2S] cluster formation. Both of these possibilities were considered because the first pocket, just after the α-helix of MEG 2.1 isoform 1 has a conformation strikingly similar to a 4xCys zinc finger. In addition, the zinc finger of 3xCysHis-type protein was found to be one of the best templates found by the bioinformatic programs for homology modelling (despite its very low sequence identity). This hypothesis was verified by adding $ZnCl_2$ to a sample of isoform 1 (without the signal peptide), and by subsequent analysis and accurate overlapping of ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC spectra, it was found that coordination of Zn^{2+} ions does not occur. This is consistent with the results of the measured distances of the proposed structures, which indicated that the distances between the SG atoms of cysteine were too large to allow Zn^{2+} coordination.

After energy minimization treatment, two potential binding pockets were identified in MEG 2.1 isoform 1. Due to the confirmation of the overlapping resonances of iso 1f and 1g with isoform 1 without signal peptide, the pocket at the C-terminus was considered more relevant. Several thousand molecules from the ZINC20 database were tested to identify potential interaction partners. Twelve putative molecules were detected in the blind screen, which showed interesting docking scores suitable for blocking this site. Upon closer examination of these molecules, it is obvious that they are relatively large molecules with a high logP, high number of aromatic rings. At the same time, all potent molecules had high values of topological polar surface area. These aforementioned properties, unfortunately, make those well-docked molecules unlikely to be suitable as oral therapeutics. The molecules listed in the results share these undesirable properties for their potential druggability. At the same time, the hydrophobic nature of the binding pocket of MEG 2.1 is a great challenge for finding druggable molecules that would follow Lipinski's rules. Despite the high docking scores for all these molecules, their druggability is almost impossible, according to Lipinski's rules of five (Lipinski 2000). These rules are guidelines for new molecular entities (NMEs) for orally active drugs. Potential candidate molecules should (i) contain no more than 5 H-bond donors, (ii) contain no more than 10 H-bond acceptors, (iii) have a molecular weight greater than 500 g/mol, and (iv) logP should not exceed 5. (Lipinski 2000; Singh et al. 2022; Benet et al. 2016). Despite the fact that the three best-docked molecules described in the results (Results, Fig. 4.90, 4.91, 4.92, 4.93, and 4.94) do not meet these rules, it must be mentioned that these are molecules that have already been used in biological tests.

The best docked molecule is 3,3',3'',3'''-(1,4-phenylenedimethylidyne)tetrakis(4-hydroxycoumarin) - NSC158393 (Kim et al. 2022). Analogues of this molecule have been tested as HIV integrase inhibitors (Liu et al. 2009; Ma et al. 2010; Chiang et al. 2007; Wang et al. 1996). Its analogues have also been tested for antimicrobial and antioxidant effects (Hamdi, Puerta, and Valerga 2008). 4-hydroxycoumarin derivates are widely used as anticoagulants

(the most well-known from this group is warfarin) (Au and Rettie 2008). 4-hydroxycoumarins are also used as anti-thrombotic agents and also have a number of other effects: for example, analgesic (Adami, UBERTI, and Turba 1959), anti-inflammatory (Luchini et al. 2008), anti-bacterial (Chohan et al. 2006), anti-viral (Kirkiacharian et al. 2008), and anti-cancer (Velasco-Velázquez et al. 2003) properties. 4-hydroxycoumarins have also been tested for antiparasitic activity, specifically against *Trypanosoma cruzi,* but unfortunately, the results were unsatisfactory (Pérez-Cruz et al. 2012). 4-hydroxycoumarin derivatives have also been tested against the protozoan parasite *Leishmania donovani* with convincing *in silico* results (Zaheer et al. 2015).

The other two very well-docked molecules are available synthetic molecules that can be purchased, but neither has been tested for biological effects in this form. 4-(2-{3-[5-(4-carboxybenzoyl)-1,3-dioxo-2,3-dihydro-1H-isoindol-2-yl]adamantan-1-yl}-1,3-dioxo-2,3-dihydro-1H-isoindole-5-carbonyl)benzoic acid is the second best-docked molecule (Results, Fig. 4.93). A very interesting structural element of this molecule is adamantane, which is a commonly used scaffold for biological applications due to its lipophilicity and stabilization of the drug because such a rigid structure formed by the assembly of three cyclohexane rings protects the close functional groups from unwanted metabolic cleavage (Wanka, Iqbal, and Schreiner 2013; Horvat et al. 2006; Roščić, Sabljić, and Horvat 2008; Štimac et al. 2017). It is evident that the adamantane in the middle of the structure strengthens this symmetric molecule, composed of substituted iso-indoles and benzoic acids, in such a way that it fills the C-terminal pocket relatively nicely. Its logP, tPSA, and MW are lower than in the case of the better-docked above-described molecule NSC158393; however, the molecule is more linear and consequently fills the C-terminal pocket less well.

The last found well-docked molecule is 3-({3-[(1H-1,2,3-benzotriazol-1-yl)methyl]-4-methoxyphenyl}(4-hydroxy-2-oxo-2H-chromen-3-yl)methyl)-4-hydroxy-2H-chromen-2-one, very similar to the first one described, with the difference that the fourth hydroxycoumarin has been substituted by benzoazatriol (Results, Fig. 4.94). This makes it more druggable but it fills in the C-terminal pocket of MEG 2.1 isoform 1 the least well of the three molecules.

It should be mentioned that all other ligands docked with a minimum of -10 kcal/mol exhibit a number of similarities: they are symmetrical molecules; they contain a high number of aromatic rings and heteroatoms, most often oxygen, nitrogen, and sulfur; none of them has been tested for biological activity. These preliminary results could be used to refine the search or to generate a more suitable dataset for further ligand search in the blind screen.

The last of the analyses within the framework of this dissertation were tests of conformational changes in protein structure using molecular dynamics (MD). This assay consisted of a long simulation of 4 μs performed on MEG 2.1 isoform 1 without a signal peptide reconstructed from the four NMR-determined peptides. The starting molecule is the same as that used for the docking. Thanks to this *in silico* method, it was possible to confirm that MEG 2.1 isoform 1 is mostly IDP, but at the same time, the N-terminal part can fold into α-helix relatively soon after the beginning of the simulation. At the N-terminus of the sequence, three shorter helices are formed, which merge into one stable long α-helix slightly before the halfway point of the

simulation. MD simulations were performed only on isoform 1 without the signal peptide (D25-P88), so it is possible to assume that this α-helix will be a continuation of the signal peptide α-helix (M1-C24). In the sequence, another three short helices are formed, which are separated from this long N-terminal α-helix and also from each other by a random coil part, that remains unstructured throughout the simulation. This simulation confirmed that the C-terminus (R69-P88) of MEG 2.1 isoform 1 remains unstructured, forming a short α-helix with almost two turns (L64-R69) and another composed of only one turn in the V80-N84 region (Results, Fig. 4.95 and 4.96). As it is evident from the superposition of the structures after MD with peptides iso 1f and 1g in Fig. 5.1, the presence of two short helices slightly changes the structure of the C-terminal pocket by rather reducing and closing it. During MD simulations, it was also confirmed that no disulfide bonds were formed.

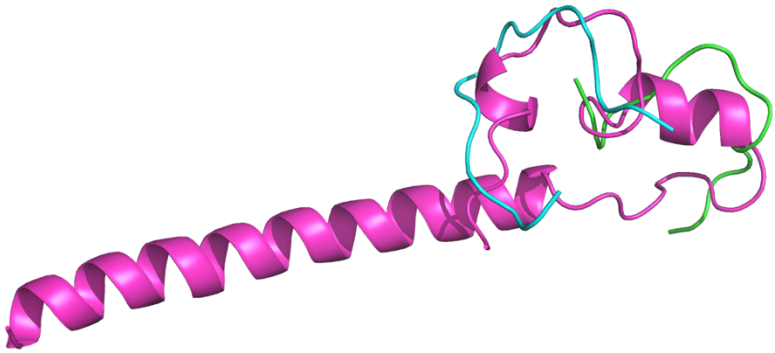

Figure 5.1 - Overlay of the structure obtained after 4 μs run of molecular dynamics with two peptides (iso 1f - green and 1g - cyan) determined by NMR.

These analyses were performed only for the unspliced isoform 1 of the MEG 2.1 family because isoform 2 could not be synthesized in its entirety, so it was not possible to verify whether the resonances of the two peptides (iso 2a and 2b) matched the resonances of the whole isoform. Despite this, it is obvious from the built structure of the NMR-resolved peptides after its energy minimization how important the alternative splicing of MEG proteins is.

In the results, I have presented two different isoforms, which are almost identical in sequence (except for the K/I at position 44), but isoform 2 has 36 amino acids spliced out before the C-terminus. The C-terminus of the two proteins is again identical (ENFIYTP), which is a previously described phenomenon of the whole MEG 2.1 family. Considering the MD analyses of the structure of isoform 1 and the results of the NMR analyses together with the structural refinement of isoform 2, I believe that the model structure presented in the results will not be too far from reality.

The C-terminus is likely to be again in the form of a random coil. Isoform 1 showed the presence of a long stable helix (H23-G53) after MD simulations, so the question arises whether the same effect will occur in isoform 2. Isoform 2 is spliced after G45, so this α-helix would lose almost three turns. At the same time, it was confirmed by CD analysis that both peptides of isoform 2 (2a and 2b) never show features of α-helix, even after the addition of 50 % TFE. The NMR analyses of these peptides showed that these two isoforms contain the highest

number of constraints (short, medium, and long ranges) and, due to this, their ten lowest energy calculated structures displayed a low r.m.s.d., thus better superimposed than those of isoform 1 peptides. Based on these facts, I believe that the structure of isoform 2 will contain an α-helix signal peptide followed by a random coil. Despite alternative splicing of a large part of the sequence, isoform 2 remains very cysteine-rich (up to 10 % of the sequence consists of cysteines). It is, therefore, reasonable to expect stabilization of the structure by the creation of disulfide bonds, but unfortunately, this could not be experimentally verified.

The sequence of isoform 3 is alternatively spliced in such a way that the resulting protein is only a signal peptide and 4 amino acids, three of which are the common C-terminal motif YTP. It is important to mention that all three described isoforms of the MEG 2.1 family have been confirmed only in the form of mRNA in transcriptomic studies. It is, therefore, reasonable to ask, at least for isoform 3, whether this protein really exists (Shyu, Wilkinson, and Van Hoof 2008) and what its biological activity might be.

Based on *ab initio* predictions and bioinformatic analyses, large segments of MEG 2.1 isoforms 1 and 2 proteins were determined to be unstructured. This has been confirmed experimentally in this dissertation.

Those two MEG 2.1 isoforms are not the only MEG family members with IDP character. MEG 14 has been shown to be a morphing IDP (Lopes et al. 2013). Currently, only two other Schistosomal MEG proteins, MEG 24, and MEG 27 have been partially characterized by CD with synchrotron light and differential scanning calorimetry. Both proteins were chemically synthesized in order to be studied *in vitro*. The CD spectra indicated the presence of an amphipathic α-helix, which could interact with membranes (Felizatti et al. 2020).

All the studies presented above are the first biophysical studies of MEG family proteins secreted from *S. mansoni* eggs. Since efforts to express them in sufficient concentration and purity in the bacterial system failed, it was not possible to study their full length by NMR. The strategy of splitting the isoforms into shorter peptides allowed their structural analysis. Of course, this strategy cannot provide complete information about the whole protein, and it is essential that the reconstructed structures presented here need to be verified experimentally.

The expression of MEG proteins has proven to be challenging, and so far, only MEG 8 and MEG 14 proteins have been recombinantly expressed (Lopes et al. 2013; Martins et al. 2014). In the framework of this thesis, MEG 2.1 isoform 1 protein and MEG 3.2 isoform 1 have also been expressed, but at concentrations not suitable for NMR structural determination.

Finally, in this thesis, I have contributed to the rationalization of the MEG superfamily of proteins annotation and to their classification, which is at the core of one paper submitted to PLoS Neglected Tropical Diseases. Moreover, I have tried hard to express in 3 heterologous hosts and in cell-free the isoforms from the MEG 2.1, MEG 3.2, and MEG 6 families; I have contributed to the structural characterization by NMR of the three alternatively spliced isoforms of MEG 2.1, chemically synthesized in shorter pieces. The strategy of dividing, solving, and reconstructing proved successful. These results are the first 3D structural data on any member of the 87 MEG proteins and are part of a second paper submitted to PLoS One.

In parallel, I had also tried to infer their structure by using the servers for homology and *ab initio* modeling, which highlighted the challenges posed by IDP or morphing proteins to prediction software, even if based on Artificial Intelligence or machine learning.
In order to find putative interactors of MEGs, I have carried out a virtual screening of the C-terminal pocket and selected three putative small molecule candidates, which fill in this pocket by displaying a complementarity surface. One of them is a derivative of 3-hydroxycoumarin, a family of known pharmacological effects. It will be interesting to test the three selected molecules *in vitro* in the future.

REFERENCES

Adami, E, E MARAZZI UBERTI, and C Turba. 1959. 'Experimental and statistical data on the analgesic action of 4-hydroxycoumarin', *Archivio italiano di scienze farmacologiche*, 9: 61-69.

Adriaan de Jongh, Willem, Sancha Salgueiro, and Charlotte Dyring. 2013. 'The use of Drosophila S2 cells in R&D and bioprocessing', *Pharmaceutical Bioprocessing*, 1: 197-213.

Au, Nicholas, and Allan E Rettie. 2008. 'Pharmacogenomics of 4-hydroxycoumarin anticoagulants', *Drug metabolism reviews*, 40: 355-75.

Azzaz, Fodil, Nouara Yahi, Henri Chahinian, and Jacques Fantini. 2022. 'The Epigenetic Dimension of Protein Structure Is an Intrinsic Weakness of the AlphaFold Program', *Biomolecules*, 12: 1527.

Benet, Leslie Z, Chelsea M Hosey, Oleg Ursu, and Tudor I Oprea. 2016. 'BDDCS, the Rule of 5 and drugability', *Advanced drug delivery reviews*, 101: 89-98.

Berriman, Matthew, Brian J Haas, Philip T LoVerde, R Alan Wilson, Gary P Dillon, Gustavo C Cerqueira, Susan T Mashiyama, Bissan Al-Lazikani, Luiza F Andrade, and Peter D Ashton. 2009. 'The genome of the blood fluke Schistosoma mansoni', *Nature*, 460: 352-58.

Cai, Pengfei, Lingyi Bu, Jian Wang, Zhensheng Wang, Xiang Zhong, and Heng Wang. 2008. 'Molecular characterization of Schistosoma japonicum tegument protein tetraspanin-2: sequence variation and possible implications for immune evasion', *Biochemical and biophysical research communications*, 372: 197-202.

Chiang, Chih-Chia, Jean-François Mouscadet, Hou-Jen Tsai, Chi-Tsan Liu, and Ling-Yih Hsu. 2007. 'Synthesis and HIV-1 integrase inhibition of novel bis-or tetra-coumarin analogues', *Chemical and Pharmaceutical Bulletin*, 55: 1740-43.

Chohan, Zahid H, Ali U Shaikh, Abdul Rauf, and Claudiu T Supuran. 2006. 'Antibacterial, antifungal and cytotoxic properties of novel N-substituted sulfonamides from 4-hydroxycoumarin', *Journal of enzyme inhibition and medicinal chemistry*, 21: 741-48.

DeMarco, Ricardo, William Mathieson, Sophia J Manuel, Gary P Dillon, Rachel S Curwen, Peter D Ashton, Alasdair C Ivens, Matthew Berriman, Sergio Verjovski-Almeida, and R Alan Wilson. 2010. 'Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts', *Genome research*, 20: 1112-21.

Dimarcq, Jean-Luc, Philippe Bulet, Charles Hetru, and Jules Hoffmann. 1998. 'Cysteine-rich antimicrobial peptides in invertebrates', *Peptide Science*, 47: 465-77.

Felizatti, Ana P, Ana E Zeraik, Luis GM Basso, Patricia S Kumagai, Jose LS Lopes, Bonnie A Wallace, Ana PU Araujo, and Ricardo DeMarco. 2020. 'Interactions of amphipathic α-helical MEG proteins from Schistosoma mansoni with membranes', *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1862: 183173.

Filichkin, Sergei A, Henry D Priest, Scott A Givan, Rongkun Shen, Douglas W Bryant, Samuel E Fox, Weng-Keen Wong, and Todd C Mockler. 2010. 'Genome-wide mapping of alternative splicing in Arabidopsis thaliana', *Genome research*, 20: 45-58.

Hagiwara, Masatoshi. 2005. 'Alternative splicing: a new drug target of the post-genome era', *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1754: 324-31.

Hamdi, Naceur, M Carmen Puerta, and Pedro Valerga. 2008. 'Synthesis, structure, antimicrobial and antioxidant investigations of dicoumarol and related compounds', *European Journal of Medicinal Chemistry*, 43: 2541-48.

Harr, Bettina, and Leslie M Turner. 2010. 'Genome-wide analysis of alternative splicing evolution among Mus subspecies', *Molecular ecology*, 19: 228-39.

Horvat, Štefica, Kata Mlinarić-Majerski, Ljubica Glavaš-Obrovac, Andreja Jakas, Jelena Veljković, Saška Marczi, Goran Kragol, Maja Roščić, Marija Matković, and Andrea Milostić-Srb. 2006. 'Tumor-cell-targeted methionine-enkephalin analogues containing unnatural amino acids: design, synthesis, and in vitro antitumor activity', *Journal of medicinal chemistry*, 49: 3136-42.

Kelemen, Olga, Paolo Convertini, Zhaiyi Zhang, Yuan Wen, Manli Shen, Marina Falaleeva, and Stefan Stamm. 2013. 'Function of alternative splicing', *Gene*, 514: 1-30.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. 2022. 'PubChem 2023 update', *Nucleic Acids Research*, 51: D1373-D80.

Kirkiacharian, B Serge, Erik De Clercq, Raffi Kurkjian, and Christophe Pannecouque. 2008. 'New synthesis and anti-HIV and antiviral properties of 3-arylsulfonyl derivatives of 4-ydroxycoumarin and 4-hydroxyquinolone', *Pharmaceutical Chemistry Journal*, 42: 265-70.

Kobayashi, Yuji, Hiroyuki Takashima, Haruhiko Tamaoki, Yoshimasa Kyogoku, Paul Lambert, Hisaya Kuroda, Naoyoshi Chino, Takushi X Watanabe, Terutoshi Kimura, and Shumpei Sakakibara. 1991. 'The cystine-stabilized α-helix: A common structural motif of ion-channel blocking neurotoxic peptides', *Biopolymers: Original Research on Biomolecules*, 31: 1213-20.

Lipinski, Christopher A. 2000. 'Drug-like properties and the causes of poor solubility and poor permeability', *Journal of pharmacological and toxicological methods*, 44: 235-49.

Liu, Ming, Xiao Jing Cong, Ping Li, Jian Jun Tan, Wei Zu Chen, and Cun Xin Wang. 2009. 'Study on the inhibitory mechanism and binding mode of the hydroxycoumarin compound NSC158393 to HIV-1 integrase by molecular modeling', *Biopolymers: Original Research on Biomolecules*, 91: 700-09.

Lopes, Jose Luiz S, Debora Orcia, Ana Paula U Araujo, Ricardo DeMarco, and Bonnie A Wallace. 2013. 'Folding factors and partners for the intrinsically disordered protein micro-exon gene 14 (MEG-14)', *Biophysical journal*, 104: 2512-20.

Luchini, Ana Carolina, Patrícia Rodrigues-Orsi, Silvia Helena Cestari, Leonardo Noboru Seito, Aline Witaicenis, Claudia Helena Pellizzon, and Luiz Cláudio Di Stasi. 2008. 'Intestinal anti-inflammatory activity of coumarin and 4-hydroxycoumarin in the trinitrobenzenesulphonic acid model of rat colitis', *Biological and Pharmaceutical Bulletin*, 31: 1343-50.

Ma, Xiaowei, Dongliang Wang, Yan Wu, Rodney JY Ho, Lee Jia, Peixuan Guo, Liming Hu, Gengmei Xing, Yi Zeng, and Xing-Jie Liang. 2010. "AIDS treatment with novel anti-HIV compounds improved by nanotechnology." In.: Springer.

Mandard, Nicolas, Philippe Bulet, Anita Caille, Sirlei Daffre, and Françoise Vovelle. 2002. 'The solution structure of gomesin, an antimicrobial cysteine-rich peptide from the spider', *European Journal of Biochemistry*, 269: 1190-98.

Martins, Vicente P, Suellen B Morais, Carina S Pinheiro, Natan RG Assis, Barbara CP Figueiredo, Natasha D Ricci, Juliana Alves-Silva, Marcelo V Caliari, and Sergio C Oliveira. 2014. 'Sm 10.3, a Member of the Micro-Exon Gene 4 (MEG-4) Family, Induces Erythrocyte Agglutination In Vitro and Partially Protects Vaccinated Mice against Schistosoma mansoni Infection', *PLoS Neglected Tropical Diseases*, 8: e2750.

Morikawa, Noriyuki, Ken'ichi Hagiwara, and Terumi Nakajima. 1992. 'Brevinin-1 and-2, unique antimicrobial peptides from the skin of the frog, Rana brevipoda porsa', *Biochemical and biophysical research communications*, 189: 184-90.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature methods*, 5: 621-28.

Nilsson, Daniel, Kapila Gunasekera, Jan Mani, Magne Osteras, Laurent Farinelli, Loic Baerlocher, Isabel Roditi, and Torsten Ochsenreiter. 2010. 'Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei', *PLoS pathogens*, 6: e1001037.

Özbek, Rabia, Natalie Wielsch, Heiko Vogel, Günter Lochnit, Frank Foerster, Andreas Vilcinskas, and Björn Marcus von Reumont. 2019. 'Proteo-transcriptomic characterization of the venom from the endoparasitoid wasp Pimpla turionellae with aspects on its biology and evolution', *Toxins*, 11: 721.

Pérez-Cruz, Fernanda, Silvia Serra, Giovanna Delogu, Michel Lapier, Juan Diego Maya, Claudio Olea-Azar, Lourdes Santana, and Eugenio Uriarte. 2012. 'Antitrypanosomal and antioxidant properties of 4-hydroxycoumarins derivatives', *Bioorganic & medicinal chemistry letters*, 22: 5569-73.

Philippsen, Gisele S, R Alan Wilson, and Ricardo DeMarco. 2015. 'Accelerated evolution of schistosome genes coding for proteins located at the host–parasite interface', *Genome Biology and Evolution*, 7: 431-43.

Piao, Xianyu, Nan Hou, Pengfei Cai, Shuai Liu, Chuang Wu, and Qijun Chen. 2014. 'Genome-wide transcriptome analysis shows extensive alternative RNA splicing in the zoonotic parasite Schistosoma japonicum', *BMC genomics*, 15: 1-12.

Reeks, TA, BG Fry, and PF Alewood. 2015. 'Privileged frameworks from snake venom', *Cellular and molecular life sciences*, 72: 1939-58.

Roščić, Maja, Vanja Sabljić, and Štefica Horvat. 2008. 'In vitro enzymatic stabilities of methionine-enkephalin analogues containing an adamantane-type amino acid', *Croatica Chemica Acta*, 81: 637-40.

Ruff, Kiersten M, and Rohit V Pappu. 2021. 'AlphaFold and implications for intrinsically disordered proteins', *Journal of molecular biology*, 433: 167208.

Scotter, Andrew J, Douglas A Kuntz, Michelle Saul, Laurie A Graham, Peter L Davies, and David R Rose. 2006. 'Expression and purification of sea raven type II antifreeze protein from Drosophila melanogaster S2 cells', *Protein expression and purification*, 47: 374-83.

Shyu, Ann-Bin, Miles F Wilkinson, and Ambro Van Hoof. 2008. 'Messenger RNA regulation: to translate or to degrade', *The EMBO journal*, 27: 471-81.

Simmaco, Maurizio, Giuseppina Mignogna, Donatella Barra, and Francesco Bossa. 1994. 'Antimicrobial peptides from skin secretions of Rana esculenta. Molecular cloning of cDNAs encoding esculentin and brevinins and isolation of new active peptides', *Journal of Biological Chemistry*, 269: 11956-61.

Singh, Madhur Babu, Pallavi Jain, Jaya Tomar, Vinod Kumar, Indra Bahadur, Dinesh Kumar Arya, and Prashant Singh. 2022. 'An In Silico investigation for acyclovir and its derivatives to fight the COVID-19: Molecular docking, DFT calculations, ADME and td-Molecular dynamics simulations', *Journal of the Indian Chemical Society*, 99: 100433.

Sorber, Katherine, Michelle T Dimon, and Joseph L DeRisi. 2011. 'RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts', *Nucleic Acids Research*, 39: 3820-35.

Štimac, Adela, Marina Šekutor, Kata Mlinarić-Majerski, Leo Frkanec, and Ruža Frkanec. 2017. 'Adamantane in drug delivery systems and surface recognition', *Molecules*, 22: 297.

Varki, Ajit, Richard D Cummings, Jeffrey D Esko, Pamela Stanley, Gerald W Hart, Markus Aebi, Debra Mohnen, Taroh Kinoshita, Nicolle H Packer, and James H Prestegard. 2022. 'Essentials of Glycobiology [internet]'.

Velasco-Velázquez, Marco Antonio, José Agramonte-Hevia, Diana Barrera, Alejandro Jiménez-Orozco, María Juana García-Mondragón, Nicandro Mendoza-Patiño, Abraham Landa, and Juan Mandoki. 2003. '4-Hydroxycoumarin disorganizes the actin cytoskeleton in B16–F10 melanoma cells but not in B82 fibroblasts, decreasing their adhesion to extracellular matrix proteins and motility', *Cancer letters*, 198: 179-86.

Wang, Shaomeng, GWA Milne, Xinjian Yan, Isadora J Posey, Marc C Nicklaus, Lisa Graham, and William G Rice. 1996. 'Discovery of novel, non-peptide HIV-1 protease inhibitors by pharmacophore searching', *Journal of medicinal chemistry*, 39: 2047-54.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature reviews genetics*, 10: 57-63.

Wanka, Lukas, Khalid Iqbal, and Peter R Schreiner. 2013. 'The lipophilic bullet hits the targets: medicinal chemistry of adamantane derivatives', *Chemical reviews*, 113: 3516-604.

Wilson, R Alan. 2012. 'Virulence factors of schistosomes', *Microbes and infection*, 14: 1442-50.

Wilson, R Alan, Xiao Hong Li, Sandy MacDonald, Leandro Xavier Neves, Juliana Vitoriano-Souza, Luciana CC Leite, Leonardo P Farias, Sally James, Peter D Ashton, and Ricardo DeMarco. 2015. 'The schistosome esophagus is a 'hotspot'for microexon and lysosomal hydrolase gene expression: implications for blood processing', *PLoS Neglected Tropical Diseases*, 9: e0004272.

Yeoh, Lee M, Christopher D Goodman, Nathan E Hall, Giel G van Dooren, Geoffrey I McFadden, and Stuart A Ralph. 2015. 'A serine–arginine-rich (SR) splicing factor modulates alternative splicing of over a thousand genes in Toxoplasma gondii', *Nucleic Acids Research*, 43: 4661-75.

Zaheer, Zahid, Firoz A Kalam Khan, Jaiprakash N Sangshetti, and Rajendra H Patil. 2015. 'Efficient one-pot synthesis, molecular docking and in silico ADME prediction of bis-(4-hydroxycoumarin-3-yl) methane derivatives as antileishmanial agents', *EXCLI journal*, 14: 935.

Zou, Junjie, Carlos Simmerling, and Daniel P Raleigh. 2019. 'Dissecting the energetics of intrinsically disordered proteins via a hybrid experimental and computational approach', *The Journal of Physical Chemistry B*, 123: 10394-402.

List of figures

6 CONCLUSION AND FUTURE PERSPECTIVES

In conclusion, this dissertation aimed to investigate the function and structure of selected *S. mansoni* MEG-family proteins. The study employed bioinformatic analysis, recombinant protein expression and optimization, and biophysical techniques. The findings revealed the unique character of *Schistosomal* MEG proteins and provided insights into their evolution and structural characteristics. The expression of MEG proteins proved challenging, with low yields and degradation issues observed. However, the S2 insect expression system showed promise for MEG 2.1 isoform 1 protein, while MEG 3.1 isoform 1 was successfully expressed with meager yields in *E. coli*. Despite the difficulties, the study successfully categorized the MEG family proteins and demonstrated their division into two subfamilies based on sequence similarities. The study also highlighted the role of alternative splicing in generating protein diversity of MEGs. Structural analysis using *ab initio* prediction servers was limited due to the lack of homologous proteins. Nonetheless, this work made progress in elucidating the structural and functional aspects of MEG family proteins, paving the way for further investigations in the field of *S. mansoni* immune evasion and immune modulation. In this dissertation, I presented the structure of MEG 2.1 peptides of isoform 1 and isoform 2, confirming their nature as intrinsically disordered proteins (IDPs). The final proposed structures of both isoforms were assembled by combining the experimentally obtained structures of shorter peptides with an AlphaFold2 (AF2) predicted signal peptide. However, the 3D structure of isoform 3 (which coincides with a predicted signal peptide of all the MEG 2.1 family members) could not be experimentally determined. However, the adaptation of AF2 prediction models facilitated the incorporation of the signal peptide into the final structures. The presence of cysteines in *S. mansoni* MEG proteins raised questions about disulfide bond formation, but various measurements and experiments indicated that such interactions were unlikely for MEG 2.1 isoform 1. Potential binding pockets were identified at the C-terminus of this isoform, and blind screens followed by molecular docking simulations revealed several molecules with high binding scores, although their suitability as therapeutics could be limited since they do not conform to Lipinski's rules of five for oral drugs. Molecular dynamics simulations confirmed the predominantly disordered nature of MEG 2.1 isoform 1 and provided insights into its conformational changes. The importance of alternative splicing in the MEG 2.1 family was highlighted, with isoform 2 exhibiting structural differences due to splicing events. Overall, the presented models provide preliminary suggestions for the MEG 2.1 protein structures and open avenues for further research and drug discovery.

Schistosomal MEG family proteins remain to be a very intriguing group of proteins that certainly deserve the attention of scientists due to their unique structure. Several challenges and exciting opportunities lie ahead. It is crucial to combine diverse "omics" datasets, including genomics, proteomics, metabolomics, and transcriptomics, in order to fully comprehend protein structure and function. In the case of MEG proteins secreted by *S. mansoni* eggs, this information is still missing. A significant contribution and advance would be to perform a proteomic analysis of the egg secretome, on the basis of which it would be

possible to determine which of the MEG family proteins are actually present among secreted proteins. This will be a very challenging task, given the fact that the authors of the studies carried out in this area do not agree on how to prepare the secretome sample for proteomic analysis. Obtaining a clean and suitable egg secretome sample for proteomic analysis remains a complex task.

If someone would like to follow up on the presented expressions of recombinant proteins, I believe that for MEG 2.1 isoform 1, a concentration suitable for structural analyses that do not require isotopic labeling (X-ray crystallography or cryo-EM), could theoretically be achieved. This process would, of course, require further financial and time investment in optimizing expression yields. If recombinant MEG 3.2 isoform 1 and MEG 2.1 isoform 1 proteins are expressed in the future, it would be interesting to perform more detailed toxicity tests of these proteins, not only against bacteria but primarily in tissue assays, especially human gut tissue. These tests could provide a final confirmation of the hypothesis that these egg-secreted proteins interact with the host intestinal wall through which they must pass in order to be excreted from the host body. Also, the phylogenetic analysis identified several short linear motifs in the red clade of MEG proteins, one of which was shared with the blue clade (i.e., across the entire MEG family). It would be interesting to experimentally verify whether these peptides are conserved because they are antigenic or because they confer some structural features to the IDPs.

Obtaining a recombinant protein would allow researchers to discover and study real interacting partners within the human host. These interactions could be carried out with selected candidate molecules within defined datasets relevant to the internal environment of the host (intestinal lumen, blood capillaries, ...) using not only *in silico* analyses but also experimental techniques, such as bio-layer interferometry or saturation transfer difference NMR.

Bridging all these gaps between the structure and function of MEG family proteins might unravel the complexity of biological macromolecules, host-parasite interactions, and also our own immune system. All these findings together will pave the way for breakthroughs in medicine, drug discovery, and biotechnology that are so urgent for solving the global problem of schistosomiasis.