FINAL REPORT

Comparative analysis of recognition motifs, linear epitopes and amyloidogenic regions in intrinsically disordered proteins

1. Scientific background and main aims

Motif-mediated protein-protein interactions play important roles in many biological processes where transient binding is needed. Such motifs and short binding regions usually take place in disordered regions, where some of them are disposed to adopt a structure upon binding to their partners. Although detailed information about the functional sites of these regions (eg. amino acid patterns, motif boundaries or motif structure) is available, other not so definite amino acid properties (such as physicochemical properties, side-chain volume, charge, or low tendency to form a secondary structure) within the motif and also in the surrounding regions are less investigated yet. During the course of the collaborative grant, we collected all the known disordered motif-mediated protein complexes, and described the motifs and their flanking regions based on different manually sorted amino acid indices. Our main goal was to shed light on the possible role of so far hidden properties in the function of disordered binding motifs and their flanking regions.

We investigated the secondary structure propensities of ten selected motifs in their free form and based upon our in silico analysis chose two interactions, the FnBPA – fibronectin and TNFR5 – TRAF2 complexes to confirm our hypothesis by examining the effect of changing different amino acid properties in the flanking regions of the binding motifs. We mutated residues to alter different properties within the N-terminal flanking regions of the chosen linear motifs to see the importance of these regions in motif function using Discrete Molecular Dynamics (DMD), NMR and CD measurements.

The outline of the work progress and the relation of bioinformatics studies to the experimental work is summerized in **Figure 1.**

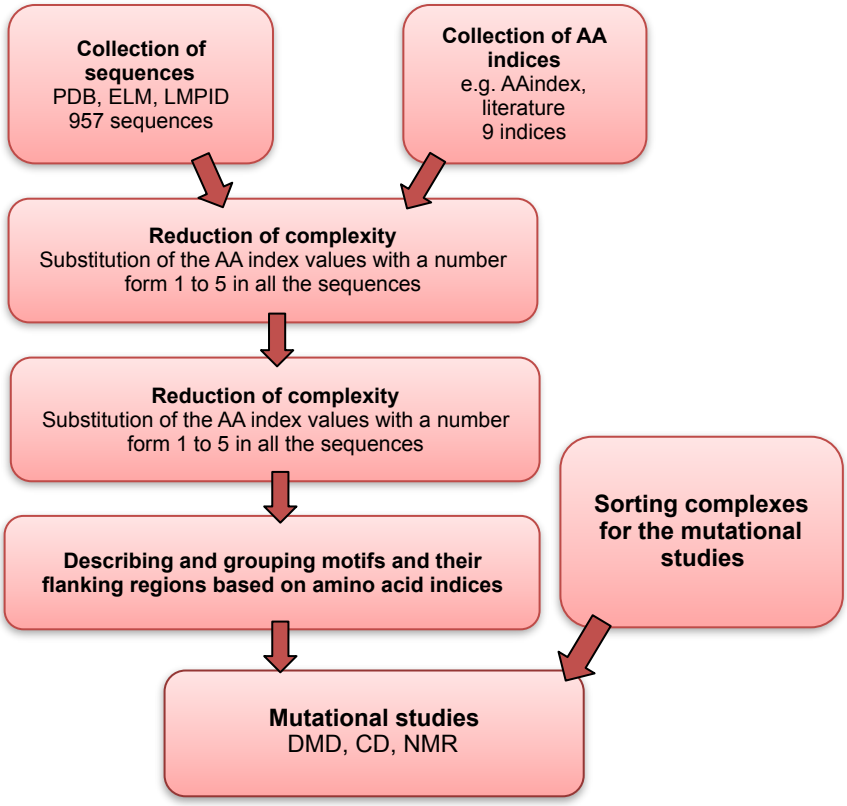

Figure 1. Schematic representation of the work progress during the grant period

2. Results of the bioinformatics analysis

2.1. Collection of known interactions of disrodered binding regions and short linear motifs

To collect all the known motif-mediated protein - protein interactions we downloaded the annotated complexes from the Eukaryotic Linear Motif (ELM, ligand binding sites only), Linear Motif mediated Protein Interaction Database (LMPID) and the Protein Data Bank (PDB) databases.

In the case of the downloaded sequences deriving from the PDB database, determination of the boundaries of binding were also needed. We used 6 angstrom as the cut-off distance to determine the interacting residues (being part of the motif) within the sequences. After the determination of the motif boundaries, we filtered out proteins with too many different partners (eg. ribosomal proteins), and also the known coiled-coil sequences with repeating heptad pattern.

We also collected the 20-amino-acid long flanking regions of the downloaded motif sequences on both sides. Then we only kept motifs occuring in disordered regions: sequences using a 25-amino-acid sliding window (where at least 23 amino acids have a value > 0.60) had been

filtered using IUPred. A representative of the collected motifs with the flanking regions is shown on **Figure 2.**

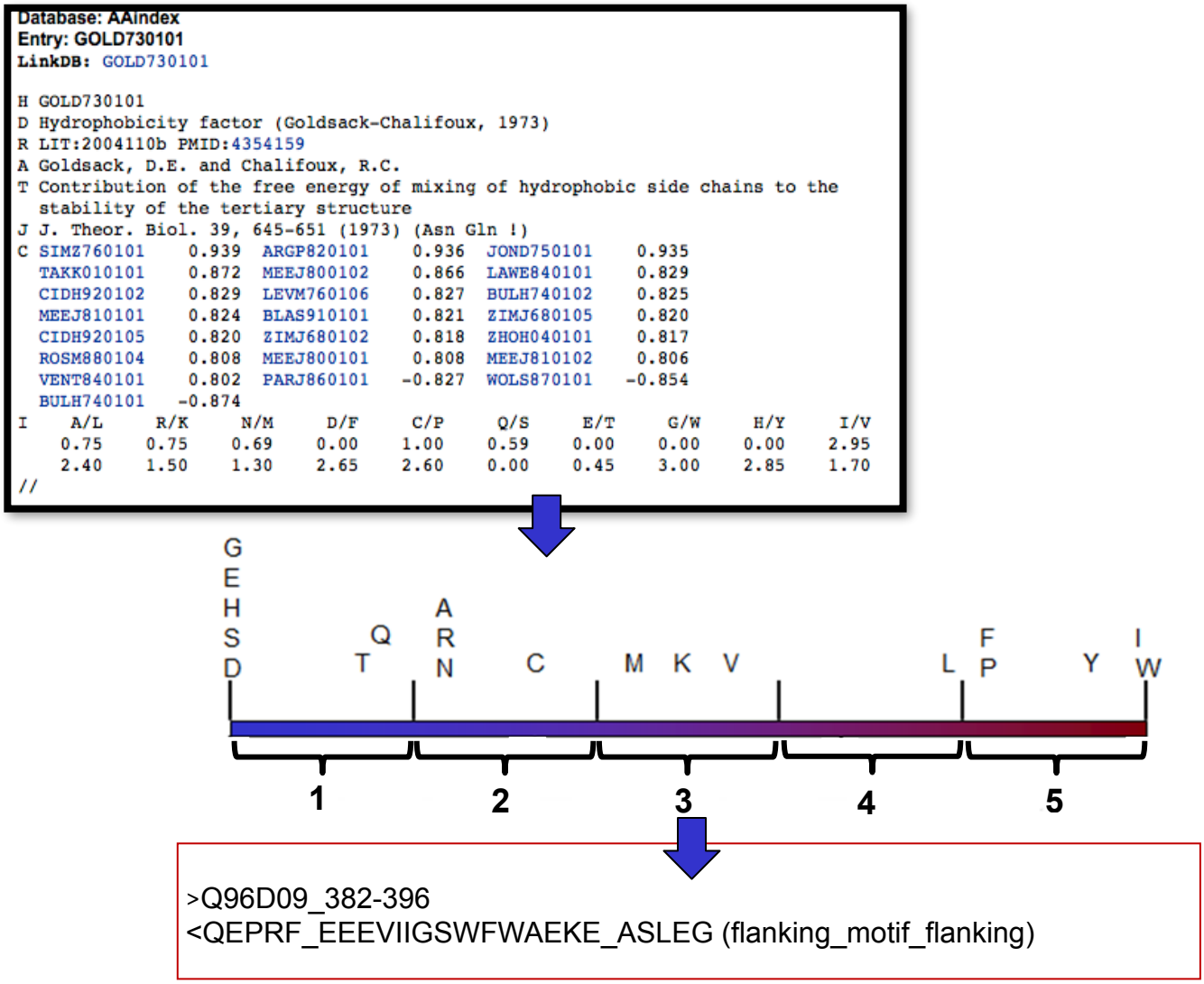

Figure 2. Binding motif and its flanking regions of GPRASP2 protein. One AAindex is shown in the upper panel, the method for reduction of complexity is shown in the middle section

To remove redundant sequences from the dataset, we filtered out sequences with at least 80% identity and we left only one of them. In the case of the same complexes from different databases we always kept the one from the PDB, where the NMR or X-ray structural data is available.

In the end, we started the sequence characterization with 957 collected sequences (motifs with their flanking regions).

2.2. Amino acid indices for the sequence characterization

We manually sorted different physicochemical, interactional and structural indices from the AAindex database, in order to characterize and group the collected sequences based on amino acid properties. We focused on indices with biological relevance and a decisive role in the function of disordered binding regions and linear motifs. Since some properties are represented by more than one AAindices (e.g. side-chain volume), and some of them also have their opposites but refer to the same property (e.g. hydrophobicity- polarity), first we analyzed the correlation between indices. We kept the indices that correlated mostly with the others (the rate of correlations are available on the AAindex database), and also those where the relating publication were available and the amino acid scores seemed biologically relevant.

In the case of the alpha-helix and beta-sheet propensity - where the side-chain values in the AAindex database derive from investigations of structured proteins - we used indices from a newer work of the structural propensity of disordered protein regions [1]. Then we examined the correlation between the different chosen properties again, and in the case of correlation we rejected the index which seemed less reliable or relevant. We kept indices where the side-chain values were located on a linear scale, so that the chosen indices could be used properly in the complexity-reducing step (see the next paragraph).

As a final step, we manually created two indices which were not represented in AAindex, but could play an important role in the function of linear motifs: *the presence of proline* and *charge in the sequences*.

The final 9 indices we used to characterize disordered motifs are the following (in italics):

-*Frequency of alpha-helix exposed* and *Frequency of beta-sheet exposed* from the literature

- The presence of *Charge* and *Proline*

-GRAR740102 *Polarity*, GRAR740101 *Composition*, FAUJ880109 *Number of hydrogen bond donors*, GRAR740103 *Volume* and ZIMJ680104 *Isoelectric point* from the AAindex database.

2.3. Reduction of complexity

Since not only the amino acid values but also the scale of the indices in AAindex are very diverse due to the different methods the authors used, we needed to reduce this complexity to simplify and also normalize our data.

For this reason, first we divided the scales to 5 equal parts, and placed the amino acids on them accordingly to their values of properties (**Figure 2.**). We then substituted the side chains with numbers 1 to 5 based on their location along our simplified scale. Here, amino acids with the lowest values got number 1, while those having the highest values were substituted with number 5.

Following this, we substituted all of the amino acids with numbers 1 to 5 along the collected motif sequences and their flanking regions. This step was performed with every index we selected for this work.

2.4. Sorting complexes for the mutational studies

Our hypothesis was that the characteristics of the amino acids in the flanking regions affect the behavior of the binding motif. To examine how changes in the amino acid properties of the flanking regions influence he structure and the partner-binding of linear motifs, we manually selected complexes for the mutational studies. We chose complexes from our dataset where PDB structure and mutational data of the motif regions were already available, since comparing our results with data from earlier experiments would make the interpretation of the outcome more reliable.

We chose the FnBPA – fibronectin (pdb2rl0) and TNFR5 – TRAF2 (pdb1czz) complexes for the DMD simulations and experimental studies. The linear motifs of proteins FnBPA and TNFR5 both have certain secondary structural tendencies without their respective partners and adopt beta sheet structure upon binding (**Figure 3.**), making them suitable for the NMR and CD measurements.

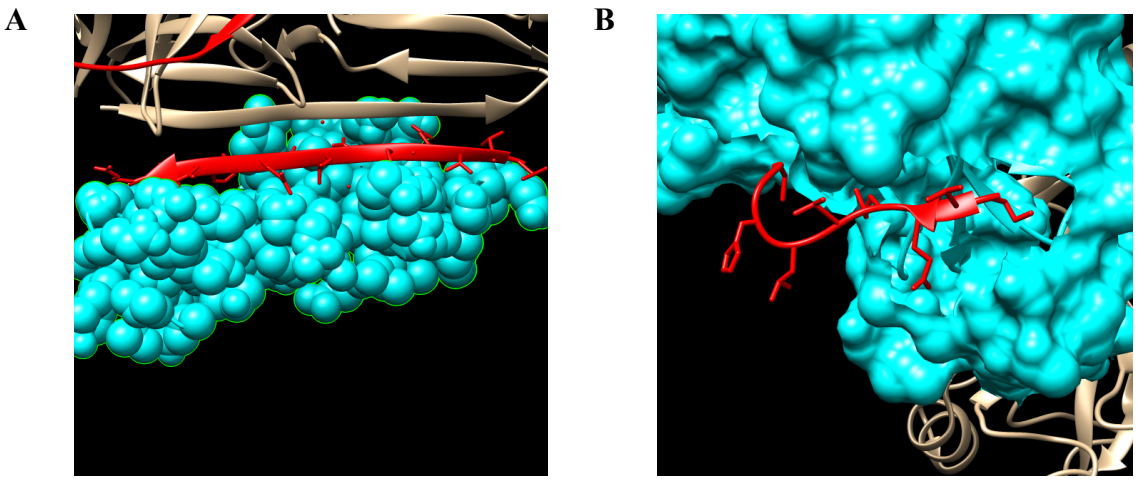

Figure 3. A. FnBPA bound to fibronectin (blue). **B.** TNFR5 bound to TRAF2 (blue). Binding motifs are marked red.

The *Staphylococcus aureus* protein FnBPA is a cell-surface attached protein with multiple fibronectin (Fn) -binding repeats. These repeating motifs, located on the C-terminal region of FnBPA mediate the invasion of *S. aureus* and the activation of the endothelial cells by binding to the N-terminal domain of the human Fn protein, thus triggering various infections, such as infective endocarditis. The repeating Fn-binding regions are separated by short disordered linker regions, which may also affect the evolvement of the tandem beta-zipper structure [2].

The tumor necrosis factor receptor 5 (TNFR5/CD40) and other TNFR family members share a conserved core of motif binding TNFR-associated factor 2 (TRAF2) to initiate intracellular signaling [3]. These motifs take place in an intrinsically disordered region, but a short sequence within the motifs adopt beta-structure upon TRAF2-binding.

The indexed sequences of the two chosen motifs (from 1 to 5) are the following (flanking_**motif**_flanking):

>pdb2rl0.ent_P14738 G 618..674 FnBPA 638-654 motif

<ENSKHHADVVEYEEDTNPGG_GQVTTESNLVEFDEEST_KGIVTGAVSDHTTVEDTK
EY

GRAR740101: 2;3;3;1;2;2;1;3;1;1;2;1;2;2;3;2;3;1;2;2;_;**2;2;1;2;2;2;3;3;1;1;2;1;3;2;2;3;2**;_;
1;2;1;1;2;2;1;1;3;3;2;2;2;1;2;3;2;1;2;1;

GRAR740102: 5;5;3;4;4;4;2;5;1;1;5;1;5;5;5;3;5;2;3;3;_;**3;4;1;3;3;5;3;5;1;1;5;1;5;5;5;3;3**;_;
4;3;1;1;3;3;2;1;3;5;4;3;3;1;5;5;3;4;5;1;

CHARGE01: 5;1;1;5;3;3;1;5;1;1;5;1;5;5;5;1;1;1;1;1;_;**1;1;1;1;1;5;1;1;1;1;5;1;5;5;5;1;1**;_;
5;1;1;1;1;1;1;1;5;3;1;1;1;5;5;1;5;5;1;

Proline01: 1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;5;1;1;_;**1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1**;_;
1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;

AlphaFujexp: 5;2;2;4;3;3;5;3;3;3;5;3;5;5;3;2;2;1;1;1;_;**1;5;3;2;2;5;2;2;4;3;5;3;3;5;5;2;2**;_;
4;1;3;3;2;1;5;3;2;3;3;2;2;3;5;3;2;4;5;3;

BetaFujexp: 2;1;2;2;2;2;1;1;5;5;2;4;2;2;1;4;1;1;1;1;_;**1;2;5;4;4;2;2;1;3;5;2;3;1;2;2;2;4**;_;
2;1;5;5;4;1;1;5;2;1;2;4;4;5;2;1;4;2;2;4;

GRAR740103: 3;2;1;4;3;3;1;2;3;3;3;4;3;3;2;2;2;1;1;1;_;**1;3;3;2;2;3;1;2;4;3;3;4;2;3;3;1;2**;_;
4;1;4;3;2;1;1;3;1;2;3;2;2;3;3;2;2;4;3;4;

FAUJ880109: 2;3;2;3;2;2;1;2;1;1;2;2;2;2;2;2;3;1;1;1;_;**1;3;1;2;2;2;2;3;1;1;2;1;2;2;2;2;2**;_;
3;1;1;1;2;1;1;1;2;2;2;2;2;1;2;2;2;3;2;2;

ZIMJ680104: 1;2;2;5;4;4;3;1;2;2;1;2;1;1;1;2;2;3;3;3;_;**3;2;2;2;2;1;2;2;3;2;1;2;1;1;1;2;2**;_;
5;3;3;2;2;3;3;2;2;1;4;2;2;2;1;1;2;5;1;2;

GRAR740101 Composition

GRAR740102 Polarity

CHARGE01 Charge (R,K,D,E: 5, H: 3, others 1)

Proline01 Proline

BetaFujexp Beta-sheet propensity

AlphaFujexp Alfa-helix propensity

GRAR740103 Volume

FAUJ880109 Number of H-bond donors

ZIMJ680104 Isoelectric point

>pdb1czz.ent_P25942 E 230..277 TNFR5 250-258 motif

<KQEPQEINFPDDLPGSNTAA_PVQETLHGC_QPVTQEDGKESRISVQERQ

GRAR740101: 1;2;2;1;2;2;1;3;1;1;3;3;1;1;2;3;3;2;1;1;_;**1;1;2;2;2;1;2;2;5**;_;2;1;1;2;2;2;3;2;1;2;3;2;1;3;1;2;2;2;2;

GRAR740102: 4;4;5;2;4;5;1;5;1;2;5;5;1;2;3;3;5;3;2;2;_;**2;1;4;5;3;1;4;3;1**;_;4;2;1;3;4;5;5;3;4;5;3;4;1;3;1;4;5;4;4;

CHARGE01: 5;1;5;1;1;5;1;1;1;1;5;5;1;1;1;1;1;1;1;1;_;**1;1;1;5;1;1;3;1;1**;_;1;1;1;1;1;5;5;1;5;5;1;5;1;1;1;1;5;5;1;

Proline01: 1;1;1;5;1;1;1;1;1;5;1;1;1;5;1;1;1;1;1;1;1;_;**5;1;1;1;1;1;1;1;1**;_;1;5;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;

AlphaFujexp: 4;5;5;1;5;5;3;2;3;1;3;3;4;1;1;2;2;2;5;5;_;**1;3;5;5;2;4;3;1;2**;_;5;1;3;2;5;5;3;1;4;5;2;5;3;2;3;5;5;5;5;

BetaFujexp: 2;2;2;1;2;2;5;1;3;1;1;1;3;1;1;2;1;4;1;1;_;**1;5;2;2;4;3;2;1;3**;_;2;1;5;4;2;2;1;1;2;2;2;3;5;2;5;2;2;3;2;

GRAR740103: 4;3;3;1;3;3;4;2;4;1;2;2;4;1;1;1;2;2;1;1;_;**1;3;3;3;2;4;3;1;2**;_;3;1;3;2;3;3;2;1;4;3;1;4;4;1;3;3;3;4;3;

FAUJ880109: 3;3;2;1;3;2;1;3;1;1;2;2;1;1;1;2;3;2;1;1;_;**1;1;3;2;2;1;2;1;1**;_;3;1;1;2;3;2;2;1;3;2;2;5;1;2;1;3;2;5;3;

ZIMJ680104: 5;2;1;3;2;1;3;2;2;3;1;1;3;3;3;2;2;2;3;3;_;**3;2;2;1;2;3;4;3;2**;_;2;3;2;2;2;1;1;3;5;1;2;5;3;2;2;2;1;5;2;

2.5. Mutation of amino acid properties in the flanking regions and their effects on motif structure and partner binding

We compared the indexed sequences of the two chosen motifs to others binding the same partner. We also examined the theoretical structural effect of index-changes in the flanking regions using the web servers Chou & Fasman Secondary Structure Preiction Server (CFSSP) and Agadir (http://agadir.crg.es/). This helped us to decide which amino acid properties are necessary for the proper function of motifs.

The results of our analysis suggested that it is possible to enhance the helical propensity of the FNBPA motif replacing the $GGGQ_{636-639}$ sequence with NAKA. This change reduces the beta sheet preference of the binding motif.

In the case of $TNFR5_{250-258}$ motif, the helical propensity, the polarity and the isoelectric point show the most marked alterations on the flanking region/binding motif border. While the P_{243} and the G_{244} show low structural preferences, the alpha helical tendency rises sharply on the border of the binding region. Polarity rises in the $SNT_{245-247}$ region, while the isoelectric point decreases. As opposed to this, on the border of the flanking/binding region, isoelectric point and structural propensity rises and polarity sharply drops. This tendency can be observed in the N-terminal flanking regions of other TRAF2 binding motifs as well (e.g. CD30 and OX40). Based on the results of the CFSSP server, mutating $PG_{243-244}$ to AA increases helicity as opposed to coil tendencies both in the region preceding the motif and in the motif itself. By exchanging $SNT_{245-247}$ to AAA, the isoelectric point rises and the polarity decreases.

2.6. Results of the DMD simulations

We tested the effect of the mutations on the structure of the motifs using Discrete Molecular Dynamics (DMD) simulations.

DMD simulations of the FNBA motif show that the wild type motif has a significant helical tendency in the N-terminal flanking region (**Figure 4. B** and **D**) and the binding motif itself has an anti-parallel beta propensity. While our preliminary in silico analysis indicated that the mutation would enhance alpha helicity in the motif, the simulations rather resulted in a less pronounced, but wider alpha helical tendency (**Figure 4. E** and **F**) and also a pronounced parallel beta preference.

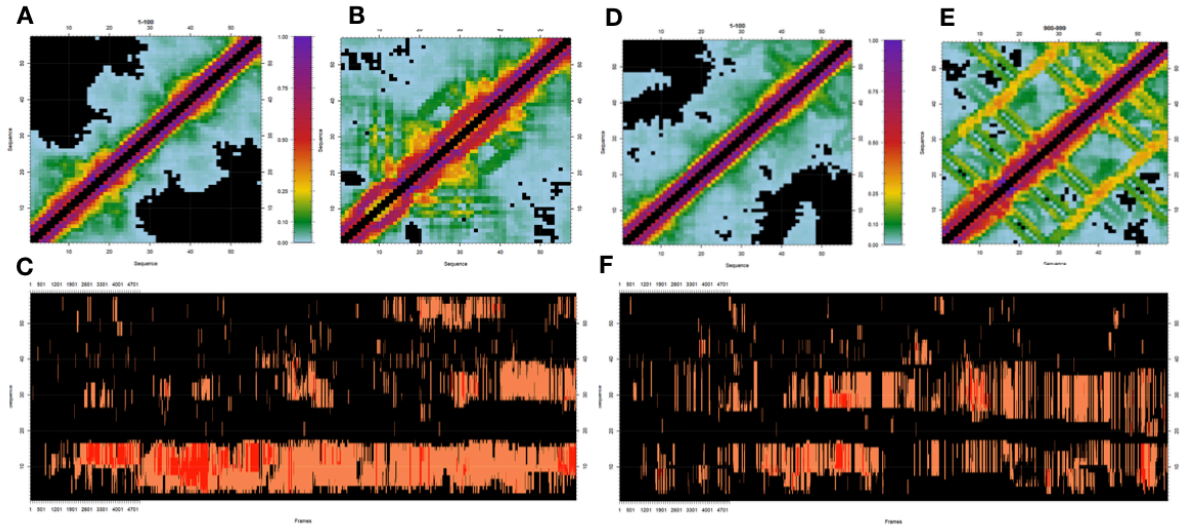

Figure 4. DMD simulation of the FNBPA$_{638-654}$ motif. **A** and **B**: contact maps of the wild type motif at the starting point (A) and at the end (B) of the simulation. **C**: DSSP helix content of the wild type motif. **D** and **E**: contact maps of the NAKA mutant motif at the starting point (D) and at the end (E) of the simulation. **F**: DSSP helix content of the NAKA mutant motif.

Modelling of the TNFR5 motif and its two designed mutant versions are shown in **Figure 5**. The wild type motif does not have a significant alpha helical tendency in the free state (**Figure 5. B** and **C**) and it has a very week anit-parallel beta propensity in the N-terminal flanking region (**Figure 5. B**). Mutating the PG$_{243-244}$ residues to AA significantly increases helix propensity (**Figure 5. E** and **F**), as expected from our in silico predictions. The SNT$_{245-247}$ to AAA replacement mainly influences polarity and the isoelectric point, but structural effects also can be seen in the formation of anti-parallel beta structures in the N- and C-terminal flanking regions and a higher alpha-helical tendency in the motif (**Figure 5. H** and **I**).

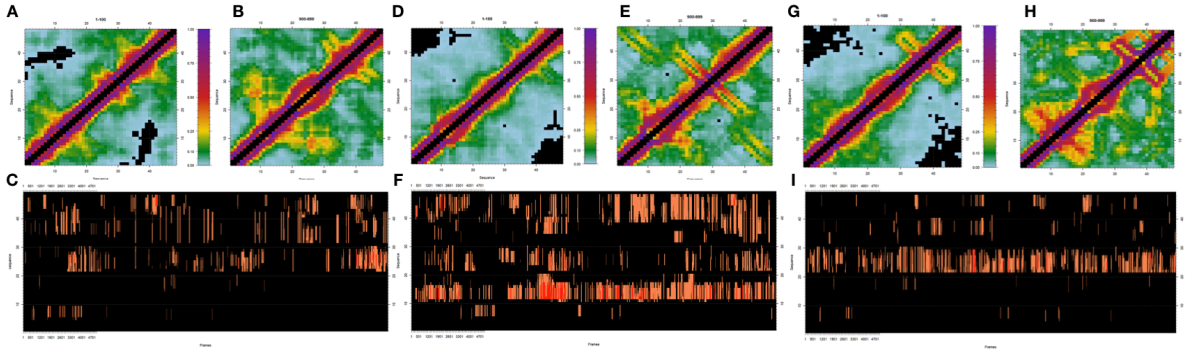

Figure 5. DMD simulation of the TNFR5 motif. A and B: contact maps of the wild type motif at the starting point (A) and at the end (B) of the simulation. C. DSSP helix content of the wild type motif. D and E: contact maps of the AA mutant motif at the starting point (D) and at the end (E) of the simulation. F. DSSP helix content of the AA mutant motif. G and H: contact

maps of the AAA mutant motif at the starting point (G) and at the end (H) of the simulation. I. DSSP helix content of the AA mutant motif.

3. Experimental results

3.1.Expression and purification of the wild type motifs

For the biochemical, biophysical and structural characterization of the selected 10 motifs we first developed a neutral carrier protein sequence to ensure the proper production of the motif in bacterial expression system. This carrier sequence was derived from the plant ERD14 protein, which we analysed in detail in our previous work. The ERD14 protein sequence was randomly scrambled (maintaining the residue composition), structurally characterized in NMR, and a region with no structural preference (mostly random coil structure) was chosen. The 100 residue long region was cloned into a pET22b bacterial expression vector with a C-terminal His-tag and a 14-residue-long cloning site in the middle. This novel vector was used for every motif, the binding region with the flanking regions were cloned into the cloning site of the carrier sequence.

The motifs analyzed were as follows:

Adenomatous polyposis coli protein (APC)

B-cell CLL/lymphoma 9 protein (CLL)

Early E1A protein or Adenovirus early region 1A (E1A)

Fibrinogen alpha chain (Fibrα)

Fibronectin-binding protein A (FnBPa)

RAF proto-oncogene serine/threonine-protein kinase (RAF1)

Mothers against decapentaplegic homolog 3 (SMAD3)

Splicing Factor 1 (SPF1)

Tumor Necrosis Factor receptor 5 (TnFr5)

Trinucleotide repeat-containing gene 6C protein (TNRC)

Expression conditions and isotopic labelling were optimized and the expressed motifs were purified on Ni-NTA resin using the C-terminal His-tag. Purified proteins were dialyzed into water and lyophilized for further use.

Mutant versions of the selected motifs were cloned and purified with the same method as the wild type motifs.

3.2.NMR measurements of the motifs

NMR spectra of the ^{15}N-^{13}C double labeled motifs were recorded by our collaborators.

In the case of CLL and E1A motifs, no NMR signal could be detected, probably due to problems with the labelling process. The spectra of RAF1 and SPF1 were low quality, meaning high background noise and some missing signals from the spectra. Presumably these motifs are capable of interacting with the carrier sequence, causing these anomalies in the NMR results. This observation was underlined by the fact that the spectrum of the empty carrier sequence did not overlay with the corresponding peaks in the case of RAF1 and SPF1 motifs. Due to these complications, we excluded these four motifs (CLL, E1A, RAF1 and SPF1) from further studies.

Full NMR peak assignment could be carried out on the following motifs: APC, Fibrα, FNBPA, SMAD3, TNFR5 and TNRC. Secondary structure calculations based on the NMR spectra showed that most of the studied motifs are fully disordered in their free form, except for FNBPA and TNFR5. These two latter motifs show distinct secondary structural tendencies that resemble the structure they obtain when bound to their respective partners. An example of the peak assignment is shown in **Figure 6.**

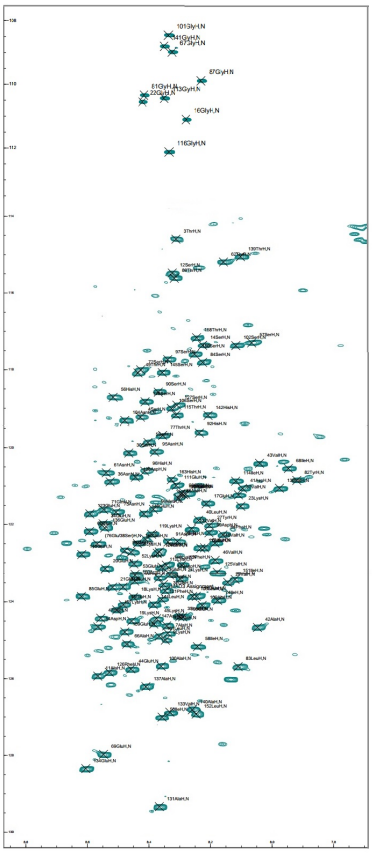

Figure 6. Full NMR peak assignment of the wild-type SMAD3 motif (in the carrier sequence) in absence of its binding partner

The NMR spectra of the mutant forms of the motifs were also recorded. The results show alterations in the detectable NMR resonances, but the detailed analysis of the spectra and the secondary structure calculations are still under way.

3.3.CD measurements of the motifs

CD measurements of the wild type and the mutant motifs were carried out in collaboration with József Kardos (ELTE) at a synchrotron radiation CD facility.

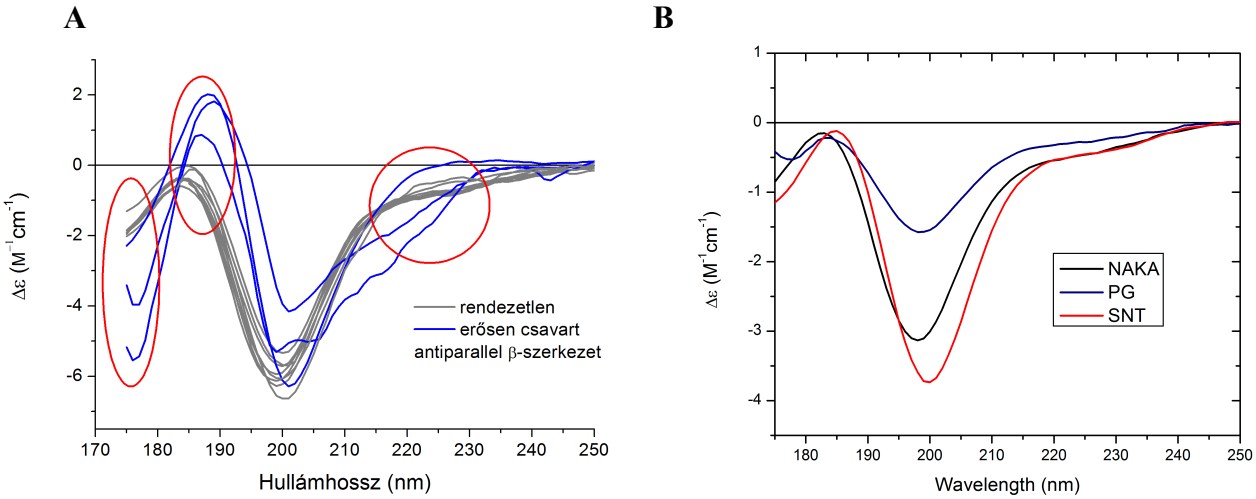

Figure 7. A: CD spectra of the wild type motifs. Grey lines: the 10 measured motifs. Blue lines: representative spectra of twisted beta structures. Red circles: the characteristic differences between random coil and beta sheet strcutrures. **B**: CD spectra of the mutant motifs. NAKA: the FNBP mutant (GGGQ to NAKA), PG and SNT: the TNFR5 mutants (PG to AA and SNT to AAA).

The wild type motifs did not show any significant secondary structure preference in the unbound state (**Figure 7. A.** grey lines). The structural tendencies that were observed in the FNBPA and TNFR5 in the NMR measurements were not seen in the CD spectra, probably due to the limits of the sensitivity of the method.

CD measurement of the mutant versions was also performed and the results are shown in **Figure 7. B.** Detailed analysis of the results is still in progress, but the mutant spectra show some deviation from the random coil spectra observed for the wild type motifs.

3.4.Other structural characterization methods

We also used size exclusion chromatography and diffusion NMR measurements to characterize the structural state of the motifs, but both methods only gave indication of widespread disorder with no indication to any structural propensity. Since the structures adpoted by the unbound motifs are transient, these methods are not capable of capturing these minor structural states.

4. Conclusions

Based on our bioinformatics and experimental results, we summarize the results of the project in the following points.

1. We collected the known motif-mediated interactions
2. We created a unique method to characterize and categorize the motifs and their flanking regions based on the characteristics of the amino acids in the sequences
3. Using our indexing method, we were able to find patterns of amino acid properties in and surrounding the binding motifs
4. We were able to design mutations that alter the secondary structural preferences of the motifs only by changing amino acids in the flanking region
5. We completed the structural characterization of six wild type motifs in the unbound state and started the detailed structural characterization of three mutant versions

Our results give a deeper insight into the structural and sequential determinants of the function of the linear motifs that mediate a plethora of important protein-protein interactions. Our indexing method will be a good starting point for the development of a bioinformatics tool that predicts the function of an unknown motif based on its sequence. The structural information derived from the mutation studies highlight the importance of the regions surrounding the actual binding motifs.

References

1. Fujiwara K, Toda H, Ikeguchi M. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. BMC Struct Biol. 2012;12: 18.

2. Bingham RJ, Rudiño-Piñera E, Meenan NAG, Schwarz-Linek U, Turkenburg JP, Höök M, et al. Crystal structures of fibronectin-binding sites from Staphylococcus aureus FnBPA in complex with fibronectin domains. Proc Natl Acad Sci U S A. 2008;105: 12254–12258.

3. Ye H, Park YC, Kreishman M, Kieff E, Wu H. The structural basis for the recognition of diverse receptor sequences by TRAF2. Mol Cell. 1999;4: 321–330.