Evolution of Interactions Involving Intrinsically Disordered Proteins

EMMA ÅBERG
This thesis describes the evolution of intrinsically disordered proteins and their interaction partners. The work presented is a combination of phylogenetic analysis, ancestral reconstruction and biophysical characterization in order to examine the evolutionary trajectory of protein-protein interactions involving disorder. The intrinsically disordered domains, NCBD and CID are both part of transcriptional co-regulating proteins. In evolution, NCBD existed before the emergence of CID and the most ancient domains display a low affinity complex with many weak contacts and high degree of conformational heterogeneity. Later in evolution, when NCBD and CID co-exists, a few mutations have altered the interaction in a way that the affinity is increased 25-fold and the conformational heterogeneity is decreased. In the same manner, the interaction is further optimized in extant species, resulting in a high affinity complex with less contacts of higher strength and less conformational heterogeneity. The intrinsically disordered transactivation domain of the tumour suppressing protein p53 and its negative regulator MDM2 date back to the beginning of animal life. The interacting domains are either lost or conserved in distinct phyla indicating a tight co-evolution. Phylogenetic trees produced by only including phyla with a conserved interaction domain follow the species evolution. Resurrection of p53 and MDM2 in the vertebrate lineage display an evolution of a high affinity complex in the ancestor of fish and tetrapods to a slightly improved affinity in modern tetrapods but a substantially lower affinity in zebrafish. The p53 protein family, which also includes p63 and p73, diverged from a common ancestor. The individual proteins display altered affinities to MDM2 which is a result of the high sequence divergence between them. The ionic dependence for the interactions is small, and not in line with other studies of disordered proteins. In conclusion, the work in this thesis have contributed with evolutionary analysis and experimental data of interactions involving intrinsically disordered proteins.

Keywords: intrinsically disordered proteins, protein evolution, phylogenetics, ancestral sequence reconstruction, biophysical characterization, NCBD, CID, p53, MDM2

Emma Åberg, Department of Medical Biochemistry and Microbiology, Box 582, Uppsala University, SE-75123 Uppsala, Sweden.

© Emma Åberg 2018

ISSN 1651-6206
urn:nbn:se:uu:diva-336083 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-336083)
To my ancestors, Mum and dad
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>CD</td>
<td>Circular Dichroism</td>
</tr>
<tr>
<td>CREB</td>
<td>Cyclic Adenosine Monophosphate Response Element Binding Protein</td>
</tr>
<tr>
<td>CREBBP</td>
<td>CREB Binding Protein</td>
</tr>
<tr>
<td>CID</td>
<td>CREBBP Interacting Domain</td>
</tr>
<tr>
<td>IDP</td>
<td>Intrinsically Disordered Protein</td>
</tr>
<tr>
<td>ITC</td>
<td>Isothermal Titration Calorimetry</td>
</tr>
<tr>
<td>MDM2</td>
<td>Mouse Double Minute 2</td>
</tr>
<tr>
<td>NCBD</td>
<td>Nuclear Co-Activator Binding Domain</td>
</tr>
<tr>
<td>NcoA</td>
<td>Nuclear Receptor Co-Activator</td>
</tr>
<tr>
<td>PTM</td>
<td>Posttranslational Modification</td>
</tr>
<tr>
<td>TFE</td>
<td>Trifluoroethanol</td>
</tr>
<tr>
<td>WGD</td>
<td>Whole Genome Duplication</td>
</tr>
</tbody>
</table>
Introduction

When people ask me what I do, I usually answer “It’s just like in the movie Jurassic Park, I resurrect the past to examine it in the present”. In this thesis, I will give you the longer version of the answer to that question. I think of myself as a biochemical archaeologist, a person who digs through the past trying to solve evolutionary puzzles. Here, I will present a combination of computational and experimental studies of intrinsically disordered proteins (IDPs) and their evolution. I have utilized a method called ancestral sequence reconstruction with the aim to resurrect proteins that existed hundreds to thousands million years ago. The objective has been to follow an interaction involving intrinsically disordered proteins from an ancestor to present day species and examine biophysical properties such as affinity and (lack of) secondary structure. Intrinsic disorder is a common feature in proteins responsible for regulation and cell signaling, these functions are crucial for maintenance of vital cellular processes. It is therefore of high interest to acquire more knowledge about how intrinsically disordered proteins interact with other proteins and understand the underlying mechanisms. In addition, understanding the evolution of proteins containing disorder could clarify why disorder appears to have become of such importance. The novelty in the content of this thesis is the combination of ancestral reconstruction and intrinsically disordered proteins, which to my best knowledge never been conducted before. The work in this thesis have contributed with evolutionary analysis and experimental data of interactions involving intrinsically disordered proteins.

Evolution of Proteins

Proteins evolve through gene duplication and divergence events (1). Gene duplications are a form of genetic backup, which creates raw genetic material and the possibility for novel protein function. Duplicated genes from a common ancestor are called sister genes and constitute a gene family. Genes present in the same genome are called paralogs while related genes in different organisms separated by speciation events are called orthologs. The fate of a duplicated gene is dependent on random mutations and selection pressure and can therefore take different routes. The most common fate of a duplicated gene is the disappearance from the genome (2). However, retention
of genes leads to functional divergence. There are different kinds of functional divergence, subfunctionalization and neofunctionalization. Subfunctionalization is an event where a division of the ancestral functions can take place. Another route is neofunctionalization, where one of the sister genes acquire a new function not present in the ancestor (3). The preservation of genes can also happen through an initial subfunctionalization, extending the time and possibility for neofunctionalization (4).

Whole Genome Duplication

Whole genome duplications (WGDs) are a result of either autopolyploidy or allopolyploidy. In autopolyploidy two fertilized diploid oocytes fuse and form a single cell which has two sets of chromosomes. Whereas, in allopolyploidy the chromosomes are duplicated but the cell does not divide. WGD is common in plants (5) and has also occurred within the animal lineage (6–8). In an early vertebrate, two WGDs, entitled 1R and 2R, resulted in four copies of the entire genome. The exact time of the WGD events is estimated to be around 500-430 million years ago, however, it is still debated whether the cyclostomes (lampreys and hagfish) have gone through the 2R duplication (9–11). In the teleost fish lineage, a third round of whole genome duplication took place (12, 13) (Fig 1). The retention of genes after a WGD event is higher compared to small scale duplications, this has given an extraordinary opportunity for expansion and divergence of species and new functions (14–17).

Figure 1. A tree depicting the three rounds of whole genome duplications in the vertebrate lineage, entitled 1R, 2R and 3R. The 1R and 2R gave rise to four copies in early vertebrates while the 3R occurred in the fish lineage. Thus, a total set of eight copies of a protein can be found in fish species.
Intrinsically Disordered Proteins

The basis of structural biology is that a function of a protein is determined by its structure, which in turn is determined by its sequence. However, a large fraction of the proteome does not fold into a well-defined structure but instead is disordered, i.e. populates an ensemble of different conformations. Intrinsic disorder was long thought of as non-functional. However, in the past 20 years disorder has been well studied and it has become increasingly clear that functional disorder is a common and essential feature in protein-protein interactions. Proteins exist and can natively act in a structural continuum, ranging from disordered regions of high flexibility, temporary structural elements, dynamic molten globules, disordered linkers to completely folded proteins (18). The discrepancies in structural content can be described by the difference in amino acid composition. Intrinsically disordered regions are enriched in amino acids with charged or polar characteristics and comprise less hydrophobic residues (19, 20). The depletion of hydrophobic residues makes it difficult for IDPs to form a hydrophobic core, which constitutes the basis of structured proteins. Instead of folding into a specific tertiary structure, IDPs sample numerous conformations. This can be displayed in a free energy landscape of an IDP, where the landscape is flatter and more rugged compared to structured proteins (21). Interactions involving IDPs are often mediated through motifs (22). When IDPs bind to an interaction partner, they may undergo a transition from disorder to order (23). This capability allows IDPs to interact with multiple partners, and can even adopt different bound structures to different interaction partners (24). Intrinsic disorder is common in proteins associated with regulation of cellular control mechanisms and signaling proteins such as transcription factors (25, 26). In such regulatory pathways, posttranslational modifications (PTMs) are a way to switch on and off functions. Therefore, it is not surprising that PTMs are enriched in disordered regions (27).

The abundance of IDPs is persistent in evolution and IDPs can be found in all domains of life, however disorder is more common in eukaryotic organisms (28–30). The complexity of the eukaryotic proteome is associated with the presence of disordered proteins. For example, in eukaryotes alternative splicing and, insertions and deletions are frequent in proteins containing disorder (31, 32). New domains which are a result of non-coding regions and have acquired functionality is often disordered (33). In yeast, the gain of disorder facilitates divergence and innovation after whole genome duplication (34). Furthermore, protein enriched in PTMs have a higher retention after WGDs (35) and are often sensitive to gene expression (15) also correlated with disorder (36). IDP are proposed to rewire interactions networks more rapidly than ordered proteins (37).

The constrains on the evolutionary process is less for IDPs as compared to structural proteins (38–40). Though, disordered regions can be divided
into different subcategories: Non-conserved, flexible and constrained (41). For example, disordered regions that can form secondary structure are conserved to a higher extent (42). In less conserved disordered regions, conservation of the chemical composition is however preserved (37), as well as length (43), motif and PTM sites (44).

Background to the Thesis Project

In order to examine the evolution of intrinsically disordered proteins two model systems were chosen to study in detail. The first model system includes the interaction between the intrinsically disordered domains NCBD and CID. Whereas, the second involves the intrinsically disordered transactivation domain in p53 and its structured binding partner MDM2. These systems were selected since they act in transcriptional regulation and are responsible for decision making in cellular processes, where their disordered nature appears to be of importance.

The NCBD-CID Interaction

The nuclear co-activator binding domain (NCBD) is present in the cyclic adenosine monophosphate response element binding protein binding protein (CREBBP) and in the CREBBP paralog, p300. The CREBBP/p300 protein family contain a large proportion of disorder, however they also comprise structurally ordered domains such as Taz1, KIX, Bromo, PHD, HAT, ZZ and Taz2 (Fig 2A). The CREBBP/p300 proteins are hubs and can interact with around 400 proteins (45) and are involved in normal development (46, 47).

The NCBD domain interacts with the CREBBP interacting domain (CID) in the nuclear receptor coactivator (NcoA) protein family. This protein family include NcoA1 (also called SRC-1), NcoA2 (also called TIF2) and NcoA3 (also called ACTR). The NcoA protein family also comprise the following domains: PAS, ST, RID, and HAT (Fig 2B). The CREBBP/p300 and NcoA protein families act together in transcriptional regulation as co-activators (48). These protein families possess histone acetyltransferase activity and therefore has the ability to acetylate histones making DNA more accessible to transcription factors resulting in enhanced gene expression (49, 50).

Figure 2. Domain structure of A) the CREBBP/p300 protein family with NCBD highlighted in blue and B) NcoA protein family with CID highlighted in yellow.
The disordered domain NCBD in CREBBP/p300 and CID in NcoA are interaction partners. Their interaction has been well studied in humans to explore the fundamentals of interactions between intrinsically disordered domains. Previous studies have shown that NCBD adopts a molten globule structure with a small hydrophobic core (51) while CID has a more extended disorder with transient helical structure. The interaction between CREBBP NCBD and NcoA3 CID display a complex with fast association followed by conformational changes (52). The structure between CREBBP NCBD in complex with NcoA1 CID (53) and NcoA3 CID (54) reveal that NCBD forms a three-helix bundle, whereas the two CID domains assume three helixes with slightly different conformations. Though, the first helix in NCBD and the first helix in NcoA1 and NcoA3 CID adopt similar position (Fig 3).

Figure 3. NMR structure of CREBBP NCBD (blue) bound to A) Ncoa1 CID (orange) (PDB-code: 2C52) and B) NcoA3 CID (yellow) (PDB-code: 1KBH).

The interaction of the first helices is mediated through a LxxLL motif, present in both NCBD and the CID domains. This motif is common in protein-protein interactions of transcriptional regulating proteins (55). Mutational studies to increase the helical propensity in the first helix of CID alter the kinetics in binding to NCBD (56).

The p53-MDM2 Interaction

The p53 protein, often referred to as the guardian of the genome, is a transcription factor responsible for regulating the fate of a cell (57, 58). The expression of p53 is regulated through ubiquitination by MDM2, initiating degradation by the proteasome (59–61). The activation of p53 is a result of various stress signals leading to phosphorylation, which in turn will trigger apoptosis or cell cycle arrest (62–64). The p53 protein is mutated in approximately half of all tumors, and therefore fails to execute its tumor suppressing abilities (65, 66).
The p53 proteins contain a disordered transactivation domain (TAD), a DNA binding domain, an oligomerization domain and a C-terminal regulatory domain. The two paralogues of p53, p63 and p73, lack the CRD and instead possess a sterile alpha motif domain (Fig 4A). The MDM2 protein family also includes MDM4, and comprises a p53-binding domain (p53-BD), an acidic domain, a zinc finger domain and a RING finger domain (Fig 4B).

Figure 4. A) Domain structure of the p53 protein family, where the CRD is present in p53 and the SAM in p63 and p73. TAD is highlighted in red. B) Domain structure of the MDM2/MDM4 protein family with the p53-BD highlighted in blue.

The binding of p53 and MDM2 initiates the degradation of p53 in healthy cells. This interaction involves the intrinsically disordered TAD in p53 and the structured p53-BD in MDM2. Upon binding the p53 TAD forms an α-helix that fit into a hydrophobic cleft in MDM2. Three hydrophobic residues in the p53 TAD helix, FxxxWxxL, constitute the binding motif (67, 68) (Fig 5).

Figure 5. The crystal structure of a peptide derived from human p53 TAD (red) and the p53-binding domain in MDM2 (blue). The binding motif (FxxxWxxL) in p53 TAD is highlighted (PBD code: 1YCR).

The N-terminal of the p53-BD in MDM2 is very flexible and in the unbound state acts as a lid in the p53 binding pocket, upon binding the flexible linker is replaced by p53 TAD (69, 70). The binding motif in p53 TAD is crucial to maintain the affinity to MDM2, when mutating only one of these residues the p53 is unable to bind MDM2 (71). The helicity is influenced by prolines adjacent to the motif in p53 TAD, and substitution of the prolines to alanine increase both the helicity and affinity (72, 73). Furthermore, the phosphorylation of Thr18 in p53 TAD abolishes binding to MDM2 initiating the acti-
vation of p53 (74). This imply a well optimized fine-tuning of the binding properties, conformation and regulation in the p53-MDM2 interaction.

Aim

The aim of this thesis is to examine and understand interactions involving intrinsically disordered proteins from an evolutionary perspective. The questions that have been addressed are how disordered regions in proteins have evolved with emphasis on the evolutionary trajectory on a molecular level, and how disorder affects affinity to its binding partner. To answer these questions, I have used evolutionary biochemistry, which is a combination of ancestral sequence reconstruction and biophysical characterization. I have worked with two model systems:

- The first model system involves the dynamic alpha helical molten globule NCBD present in the CREBBP/p300 protein family and its binding partner, the intrinsically disordered CID in the NcoA protein family. The emergence and evolution of the interaction was examined with further biophysical characterization of secondary structure, affinity and dynamics (Paper I).

- The second model system includes the intrinsically disordered trans-activation domain of p53 and its structured negative regulator, the p53-binding domain of MDM2. Three separate studies were conducted. First, the evolution of p53 and MDM2 was examined by studying the conservation in species and their evolutionary relationship (Paper II). Second, ancestral sequence reconstruction and resurrection was performed to elucidate the binding affinity at different evolutionary time points (Paper III). Finally, the binding kinetics at different ionic strength of the human paralogues p53, p63 and p73 binding to MDM2 was examined (Paper IV).
Materials and Methods

Evolutionary Analysis of Proteins

The advances in genome sequencing has led to faster and cheaper analysis. Together with improvements of phylogenetic methods, this have opened up possibilities in the field of evolutionary biology. The increasing number of sequenced genomes can be utilized to retrieve information about the conservation of a gene or gene family in different species and can help improve our understanding of evolutionary trajectories and relationships.

Sequence Identification

The identification of sequences from the same gene family can be achieved by searching in various genome databases, for example the Ensembl genome browser. The Ensembl genome browser (ensembl.org) comprise several vertebrate genomes (75). The EnsemblCompara GeneTree function identifies paralogue and orthologue sequences of the same gene family (76). EnsemblMetazoa (metazoa.ensembl.org) is a sister site to Ensembl and include invertebrate animals (77). To identify sequences within a gene family, the Basic Local Alignment Search Tool (BLAST) can be used (78). BLAST can identify regions with similar amino acid sequence in a genome database. The BLAST algorithm compares the input sequence with sequences in a database and gives a score of the similarity and a statistical value which corresponds to the hits expected to occur by chance in the searched database. In this thesis BLAST was used in the following databases: The National Centre for Biotechnology Information (NCBI) (ncbi.nlm.nih.gov), EchinoBase (echinonbase.org) (79) SkateBase (skatebase.org) (80) and Japanese Lamprey Genome Project (jlampreygenome.imcb.a-star.edu.sg) (81).

Sequence Alignments and Phylogenetic Analysis

Multiple sequence alignment algorithms are used to align protein sequences to evaluate the conservation of different regions and binding motifs. A multiple sequence alignment is required to infer phylogenetic relationships. Phylogenetics is used to study the evolutionary history of a group of species. A phylogenetic tree is a branching diagram that symbolizes the evolutionary
relationships of different sequences. The tips of the tree represent present-day species, while the nodes of the tree represent ancestors.

In this thesis, protein sequences were aligned using Guidance (guidance.tau.ac.il) (82, 83) with the MAFFT algorithm (84). Guidance gives an alignment confidence score for each column, residue and sequence. The score represents the robustness to perturbations in the output alignment. The Guidance algorithm also offers the possibility to remove columns and sequences and mask residues below a certain cut-off value. Gaps in the alignment were removed. The best amino acid substitution model was selected using MEGA6 (85, 86). The alignments were further used to calculate phylogenetic trees where the Phylogenetic Maximum Likelihood was used (87). The tree was supported by a non-parametric SH-like approximate likelihood ration test (88).

Ancestral Sequence Reconstruction and Resurrection

Ancestral sequence reconstruction is a method where the most recent common ancestor of a protein can be predicted. The idea of reconstructing ancestral sequences was suggested already in 1963 (89), however it took about 30 years of technical advancement until reconstruction became reality (90–92). Ancestral sequence reconstruction is based on an alignment of proteins from present-day species and their phylogenetic relationship. The alignment and the phylogenetic tree is based on the entire protein even though a single domain is to be reconstructed. When using this method, it is important that the phylogenetic tree matches the evolution of species. It is also important that the domain subjected to reconstruction is well aligned. Transitive consistency score can be used for evaluation of different alignments (93). The prediction of the ancestral sequences was conducted by using maximum likelihood, which takes branch lengths and relative rates of substitutions into account (94). The output gives the probability for each amino in each position. In the case of low probabilities, several probable amino acids can be resurrected to test different ancestral states. The resurrection part involves conventional methods like cloning, protein expression and purification to obtain a pure protein. The proteins can thereafter be characterized for different biophysical properties (Fig 6).
Figure 6. Illustration of ancestral sequence reconstruction and resurrection. An alignment and a phylogenetic tree of present-day species is used to predict an ancestor in a node. The predicted ancestral sequence is resurrected through synthesis of the coding DNA, followed by protein expression and purification. The pure protein can then be subjected to biophysical characterization (Figure from Paper I).

Experimental Analysis of Protein-Protein Interactions

When two proteins bind to each other a second order reaction occurs, association and dissociation of a complex:

\[
A + B \rightleftharpoons AB
\]

Binding events have thermodynamic components, enthalpy (H) that reflects the heat correlated to the strength of the bonds, entropy (S) which mirror the dynamics and solvent effects. The sum of enthalpy and entropy gives the Gibbs free energy (\(\Delta G\)) of the interaction i.e. the difference in bound and unbound state:

\[
\Delta G = \Delta H - T \cdot \Delta S
\]

where \(\Delta H\) refers to the change in enthalpy, \(T\) the temperature and \(\Delta S\) the change in entropy upon binding.
Gibbs free energy can also be written in terms of the equilibrium dissociation constant:

$$\Delta G = -RT \cdot \ln K_d \iff K_d = e^{\frac{\Delta G}{RT}}$$

where $R$ is the gas constant, $T$ is the temperature and $K_d$ is the equilibrium dissociation constant (Fig 7).

**Figure 7.** Illustration of how the Gibbs free energy is related to different kinetic rate and equilibrium constants.

**Isothermal Titration Calorimetry**

Isothermal titration calorimetry (ITC) is a technique to study protein-protein interactions. When two proteins interact, thermodynamic changes occur and ITC can be used to determine the thermodynamic parameters ($\Delta H$, $\Delta S$, $\Delta G$) of the reaction. However, the primary aim of ITC is to determine the equilibrium dissociation constant ($K_d$) and the number of binding sites ($n$). The ITC will estimate the association equilibrium constant ($K_a$), which is the inverse dissociation equilibrium constant. $K_a$ equals the concentration of the complex divided by the concentration of the unbound proteins:

$$K_a = \frac{[AB]}{[A][B]} = \frac{1}{K_d}$$

The experimental setup involves stepwise injections of a protein placed in a syringe to its binding partner in a calorimetric sample cell. A reference calorimetric cell is kept at constant temperature and after each injection the temperature in the sample cell changes, that afterwards is adjusted back to maintain minimal difference to the reference cell (Fig 8A). For each injection $i$, the heat ($q_i$) of the reaction is measured and then is related to the concentration of the formed complex:

$$q_i = V \cdot \Delta H_a ([AB]_i - [AB]_{i-1})$$
where \( V \) is the calorimetric cell volume, \( \Delta H_a \) the enthalpy of binding and \([AB]\) the concentration of the complex at injection \( i \). As the injections proceed, the complex concentration increases and the heat of each injection decreases until complete saturation (Fig 8B).

\[
\begin{align*}
\text{A) The setup of the experiment is a titration of a protein placed in a syringe to its binding partner in a cell at constant temperature. The change in temperature is controlled via a reference cell. B) The change in heat when two proteins interact is measured in every injection. With every injection, the concentration of the protein complex increases until the protein in the cell is saturated and no change in heat can be observed. C) The change in heat over time is integrated to kcal/mol, plotted against the molar ratio and fitted to a sigmoidal function where the affinity (slope of the curve), enthalpy (difference in energy content of the bound and unbound state) and stoichiometry (midpoint of the transition) of the protein interaction can be assessed.}
\end{align*}
\]

The amount of heat released/absorbed at each injection is integrated and plotted against the molar ratio of the two interacting proteins. The resulting plot is fitted using non-linear regression, where the following parameters can be estimated: The equilibrium association constant \((K_a)\) which is equal to the slope, the number of binding sites \((n)\) which corresponds to the midpoint of the slope and enthalpy of the binding reaction \((\Delta H_a)\) which is the total heat (Fig 8C). From these values, the free energy and the entropy of the reaction can be calculated. ITC can measure binding affinities in the nanomolar to micromolar range. Advantages of the method are that it is label-free and in solution.

**Stopped Flow Fluorimetry**

Stopped flow fluorimetry is a method for studying the kinetics of protein-protein interactions. From this type of experiment the association rate constant \((k_{on})\) and dissociation rate constant \((k_{off})\) of the binding reaction can be determined.

\[
\begin{align*}
A + B & \rightleftharpoons AB \\
& \frac{k_{on}}{k_{off}}
\end{align*}
\]
The experimental system involves rapid mixing of two interaction partners in a cell where monochromatic light is passed through (Fig 9A). To monitor the binding event in the cell, one of the interaction partners requires a fluorescent probe that is affected by the interaction. The monochromatic light excites the fluorescent probe and when binding occurs will give rise to an increase/decrease of the emitted light. This results in a plot of the emitted light over a set time period (Fig 9B). In the case of a two-state mechanism, the trace in the plot can be fitted to a single exponential equation, from which the observed rate constant can be extracted:

\[
\text{Fluorescence emission} = [AB](t) = a \cdot e^{-k_{obs}t} + C
\]

where the fluorescence emission corresponds to the concentration of protein complex over time, \(a\) is the fraction of A bound to B, the observed rate constant, \(k_{obs}\), which is related to the slope of the curve and \(C\), the fluorescence emission signal at equilibrium.

\[\text{Figure 9. A) The experimental setup is two injection syringes with one binding partner in each. The drive ram pushes the binding partners in to a mixer where monochromatic light excites the sample, at the same time the fluorescence emission is recorded. B) The fluorescence emission is measured over time and fitted to an exponential function to extract the observed rate constant of the binding reaction. C) The observed rate constant is plotted against varying concentrations of one binding partner where the association rate constant and the dissociation rate constant can be estimated by fitting.}\]

The association rate constant is determined by the slope of the observed rate constants at different concentration of B. If the experiment is done under true pseudo-first order conditions, the data can be fitted to a straight line:

\[
k_{obs} = k_{on} \cdot [B] + k_{off}
\]
However, the conditions for pseudo-first order is not satisfying at low concentrations of B, which often is the case in experimental setups. Instead, the data can be fitted to an equation for a reversible bimolecular association (95) (Fig 9C).

\[ k_{obs} = \left( k_{on}^2 ([A]_0 - [B]_0)^2 + k_{off}^2 + 2k_{on}k_{off} ([A]_0 + [B]_0) \right)^{0.5} \]

The association and dissociation rate constants can be extracted from the equation by curve fitting. If the dissociation rate constant is close to zero, separate displacement experiments can be conducted. Displacement experiments are performed by mixing a preformed complex (AB) with high concentration of a displacer (C) that has affinity to A. This will result in a competitive binding to A, in which C will displace B and a AC complex will be formed.

\[
\begin{align*}
\text{AB} + C & \rightleftharpoons A + B + C \rightleftharpoons AC + B \\
k_{off_1} & \quad k_{on_2} \\
k_{on_1} & \quad k_{off_2}
\end{align*}
\]

At high concentrations of C, the association back to AB is negligible and the resulting \( k_{obs} \) value is therefore approximately the dissociation of the AB complex. A displacement reaction will result in an observed rate constant that equals:

\[ k_{obs} = \frac{k_{off_1} + k_{on_2} \cdot \frac{K_d}{[C] + K_d}}{k_{on_1} \cdot k_{off_2}} \]

where the \( K_d \) represents the equilibrium dissociation constant for the AC complex.

Experimental Analysis of Protein Secondary Structure

Proteins are built up by polymers of amino acids with different characteristics. The order of the amino acids in a protein is referred to as the primary structure. Proteins can fold into \( \alpha \)-helices and \( \beta \)-sheets defined by different hydrogen bonding, called secondary structure. The organisation of the secondary structure elements within a protein is driven by the hydrophobic effect, termed tertiary structure. Determination of protein structure and how perturbations affect the structure can give information about the binding mechanism and stability of the protein.
Circular Dichroism

Circular dichroism spectroscopy (CD) is a technique to study protein secondary structure and how it changes upon various perturbations. This method utilizes the chiral properties of amino acids and their difference in absorption of left and right polarized light in the far-UV region (Fig 10A). A polypeptide induces an overall chirality that will result in a characteristic shape depending on the secondary structural features of a protein (Fig 10B). The signal from a CD experiment is presented in millidegrees (mdeg), which is the difference in absorbance of left and right-handed polarized light. This unit is often converted to molar ellipticity (deg cm$^2$ dmol$^{-1}$) that corrects for the protein concentration and the path length of the cuvette. Circular dichroism can estimate the overall secondary structure of a protein and how it changes upon different perturbations for example temperature, denaturant or helix-inducing agent. Perturbations of a protein will result in a transition of the native to a non-native state. Experiments with increasing perturbation can be used to determine the midpoint of unfolding/folding. (Fig 10C). The observed signal will be a mix of the native (N) and denatured (D) state:

$$\text{Observed signal} = f N \cdot sN + f D \cdot sD$$

where $f$ is the fraction that is occupied by the state and $s$ is the signal in that state. At equilibrium, the native and denatured state are occupied to the same extent and the equation can be rewritten in terms of thermodynamic parameters:

$$\text{Observed signal} = \frac{(\alpha_N + \beta_N \cdot P) + (\alpha_D + \beta_D \cdot P) e^{-\frac{(\Delta G)}{RT}}}{1 + e^{-\frac{(\Delta G)}{RT}}}$$

where $P$ is the degree of perturbation, $\alpha_{N/D}$ is the signal from the state and $\beta_{N/D}$ is the rate of intrinsic change in the signal from the state with increasing $P$; $\left(\frac{d\alpha_N}{dP}\right)$.

For a chemical perturbation, the free energy of unfolding equals:

$$\Delta G = m([P]_{50\%} - [P])$$

where $m$ is the slope of the transition and $[P]_{50\%}$ is the concentration of the perturbing agent at the transition midpoint.

For a thermal denaturation, the free energy of unfolding equals:

$$\Delta G = \Delta H_m \left(1 + \frac{T}{T_m} \right) + \Delta C_p \left( T - T_m - \left(T \cdot ln \frac{T}{T_m}\right) \right)$$
where $T$ is the temperature, $T_m$ the midpoint of thermal denaturation, $\Delta H_m$ the enthalpy of denaturation at the transition midpoint and $\Delta C_p$ the change of heat capacity of denaturation.

*Figure 10.* A) Circular polarized light is passed through a protein sample where the light is absorbed to different extent at different wavelengths. B) The ellipticity is recorded over a set wavelength or wavelength interval, the curve gives a characteristic shape depending on the secondary structure: $\alpha$-helix (blue), $\beta$-sheet (green) and random coil (red). C) The perturbation, for example urea, at a specific wavelength is plotted against fraction unfolded and fitted to extract the midpoint of unfolding.
Results

The Evolution of the NCBD-CID Interaction

Paper I describe the ancestral reconstruction, biophysical characterization and molecular simulation to shed light on the evolution of the interaction between NCBD and CID. NCBD was found to be present in deuterostome and protostome species, deuterostome species include vertebrates, starfish and acorn worms while protostome species are invertebrates like snails, worms and insects. CID appeared later in evolution and is only present in deuterostome species. Both proteins existed when the two whole genome duplications occurred and resulted in four copies of each protein. The CREBBP/p300 protein family retained two copies: CREBBP and p300 while the NcoA protein family preserved three copies, named NcoA1, NcoA2 and NcoA3. In the fish lineage, the third genome duplication produced two additional copies of the protein which are conserved in both the CREBBP/p300 and NcoA1-3 protein families (Fig 11).

In this study, we resurrected sequences in different nodes of the phylogenetic tree. The resurrected sequences represent the ancestral state. For NCBD the following ancestral sequences where resurrected: The ancestor in the node where deuterostomes diverge from protostomes (D/P), the ancestor in the node for the WGD (denoted 1R/2R) and the ancestor of fish and tetrapods (Fish/Tetrapods) in the CREBBP lineage. Furthermore, NCBD domains from extant species were used in the analysis, namely those from human CREBBP and p300, a fruit fly (*Drosophila melanogaster*), a fish in the CREBBP lineage (*Danio rerio*) and a lamprey (*Petromyzon marinus*). For CID, we resurrected the ancestor in the node of the first and second WGD (denoted 1R and 2R) and the ancestor of fish/tetrapod in the NcoA3 lineage. The CID domains from human NcoA1, NcoA2 and NcoA3 were also included in the analysis. For the resurrected sequences, positions with a low probability were tested as alternative ancestral states.
The affinity of all contemporary complexes including alternative ancestral sequences of NCBD and CID was measured with isothermal titration calorimetry. At the time point of the split of deuterostome and protostome species as well as in fruit fly, the CID domain did not exist. In the most ancient complex, the interaction between NCBD D/P and CID 1R, the equilibrium dissociation constant was estimated to 1.5-18 µM. In fruit fly, similar affinities are observed to different CID variants. This reveals that NCBD and CID are able to interact with each other, even though they are not co-existing in the same proteome. At the time point of the two whole genome duplications, 1R and 2R, NCBD has recruited CID as a binding partner and the affinity is increased around 25-fold. The strength of the interaction is preserved in the ancestor of fish/tetrapod as well as in the extant species of human, fish and lamprey (Fig 12A). This suggests a fast adaptation and a retained affinity for the NCBD-CID complex in evolution.

The NCBD variants were examined for their secondary structure with circular dichroism. The examination showed that all NCBD variants displayed a very similar shape of a helical spectrum and indicate that a similar helical content has been present during evolution. To further test the secondary
structure stability of the NCBD domains, experiments with increasing temperature and urea concentration were conducted. For a globular protein, one expects data points following a sigmoidal curve where the midpoint of the slope corresponds to the equilibrium of the native and denatured state. The temperature stability measurements did not show a sharp transition for any variant. This makes it difficult to determine the thermal midpoint for denaturation. The titration of urea indicates that the NCBD D/P has a slightly lower stability compared to the evolutionary younger NCBD 1R/2R, CREBBP fish/tetrapod and extant variants. However, the curves display a very broad transition (Fig 12B). The difficulty of determining the stability of NCBD is probably because of its small hydrophobic core and dynamic properties.

The CID domains were also subjected to circular dichroism experiments. The spectra display a random coil with a transient helical structure for all variants. Furthermore, examination of the propensity to adopt a secondary structure was performed. To induce a coil-helix transition in the CID variants, different concentrations of 2,2,2-Trifluoroethanol (TFE) was added. The TFE titration displayed no difference in helical tendency for the CID variants (Fig 12C). This imply that the disordered characteristics of CID is preserved from the ancestor in 1R to extant species.

![Figure 12](image_url)

Figure 12. A) The relative affinity between NCBD and CID variants, against the interaction of human p300 NCBD and NcoA2 CID. B) Urea denaturation of NCBD variants. C) TFE titration of CID variants. (Adapted from figure 5 in Paper I)

To further investigate the dynamic properties in the evolution between NCBD and CID bias-exchange metadynamics simulations were performed. The simulations are based on chemical shift determined by NMR. Three complexes were chosen: The most ancient complex (D/P NCBD - 1R CID), the 1R/2R complex (1R/2R NCBD - 1R CID) and one extant complex (the human CREBBP NCBD - NcoA3 CID). The free energy as a function of the fraction helix and the Radius of gyration was calculated and analyzed. The
The most ancient complex displays a higher degree of conformational heterogeneity than the younger 1R/2R complex and the heterogeneity is further reduced in the extant complex. Additional measurements of the root mean square fluctuation display decreasing fluctuations in the younger complexes compared to the more ancient one. To examine the number of interactions the complexes are able to make an interface contact analysis was performed. The number of contacts between ancient and extant complexes demonstrate a reduced number of contacts with a higher strength in the extant complex compared to the ancient complexes where many contacts are made but with less strength.

In conclusion, the molecular simulations and the affinity measurements suggest that a higher affinity is correlated with a reduced conformational heterogeneity. The most ancient complex has more contacts with less strength and over time the optimization of the binding has led to fewer but stronger interactions. The evolution of the interaction has led to a decreased conformational heterogeneity and a higher affinity.

The Evolution of the p53-MDM2 Interaction

In paper II, the evolution of the p53/p63/p73 and MDM protein families with focus on the interaction domains was examined, that is, the transactivation domain (TAD) and the p53/p63/p73-binding domain (p53/p63/p73-BD). The p53/p63/p73 family comprise three proteins, p53, p63 and p73 while the MDM family include MDM2 and MDM4. The emergence of the proteins within the families are a result of whole genome duplication events that took place in an early vertebrate. In the third WGD event in the fish lineage, the second copy was lost for all paralogues of p53 and MDM2. Before the duplication only one copy was present, denoted p53/p63/p73 and MDM. The p53/p63/p73 TAD is consistently detected in deuterostome species within all paralogs. In protostome species, distinct phyla have lost the TAD. The TAD is preserved in the phyla comprising annelids (ringworms), mollusks (snails and mussels) and in the arthropod subphyla myriapods (centipedes) and chelicera (spiders and horseshow crabs). In the arthropod subphyla crustaceans (crabs and cray fish) and hexapods (insects and flies), and nematode (roundworms) a p53/p63/p73 protein is present. However, TAD is missing but the DNA-binding domain and the oligomerization domain is conserved. Interestingly, the MDM p53/p63/p73-BD follows the same pattern as p53/p63/p73 TAD and is found in the same phyla within deuterostomes and protostomes. In a similar manner to p53/p63/p73, the MDM protein can be found in a few arthropod species, where the acidic, zinc-binding and RING domain is conserved. However, the entire protein is completely lost in nematodes (Fig 13).
Figure 13. Schematic species tree displaying the presence of the TAD (red) and p53/p63/p73-BD (blue). Grey lines indicate no presence of TAD and p53/p63/p73-BD (Adapted from figure 2 in Paper II).

In cnidarian species, the interaction domain of both p53/p63/p73 and MDM cannot be found, and in sponges the p53/p63/p73 and MDM proteins are completely missing. Nonetheless, the complete p53/p63/p73 and MDM are present in the multicellular organism *Trichoplax adhaerens*. This suggests that the interaction between TAD and p53/p63/p73-BD dates back to the beginning of animal life and have since then tightly co-evolved. The results show that species in distinct phyla have either both interaction domains conserved or both have disappeared. After this realization, phylogenetic trees were calculated, only including species with a conserved interacting domain. The resulting trees of the p53/p63/p73 and MDM protein families both follow the species evolution. In the case where all sequences with similarity to p53/p63/p73 and MDM were used, the phylogenetic tree did not show an accordance with the evolution of species.
In paper III, ancestral reconstruction and biophysical measurements were used to examine the affinity between the intrinsically disordered transactivation domain of p53 and the p53-binding domain of MDM2 at different time points in evolution. Due to high sequence divergence in the MDM family, reconstruction was only conducted in the vertebrate lineage of p53 and MDM2. The p53-MDM2 complex was examined at five different evolutionary time points: in the ancestor of fish and tetrapods, in the ancestor of reptiles/birds and mammals and in the present-day species *Danio rerio* (zebrafish), *Gallus gallus* (chicken) and *Homo sapiens* (human).

The affinity of the contemporary complexes was calculated from the association and dissociation rate constants determined by stopped flow fluorimetry (Fig 14). The binding affinity in the ancestor of fish and tetrapods was measured to 70-200 nM. Following the evolution in the tetrapod lineage the affinity was preserved at 170-200 nM in the ancestor of reptiles/birds and mammals. The p53-MDM2 complex in the chicken *Gallus gallus* displayed an affinity of 80 nM, which is a 2-fold increase compared to the reptiles/birds and mammals ancestor. In humans, a 10-fold increase in affinity can be measured as compared to the ancestor. However, in the fish lineage represented by the zebrafish *Danio rerio*, the affinity was too high to be measured, with a dissociation rate constant of approximately 500 s⁻¹. This stands in large contrast to the contemporary complexes in the ancestor of fish and tetrapod and in chicken and human, where the dissociation constant is below 10 s⁻¹. Measurements of MDM2 from *Homo sapiens* and the ancestor of fish/tetrapods against p53 from *Danio rerio* displayed a similar dissociation rate constant as the contemporary complex in *Danio rerio*. On the other hand, the MDM2 from *Danio rerio* against p53 from *Homo sapiens* and the ancestor of fish/tetrapods have affinities of 90 nM.
The p53 sequence in *Danio rerio* differs in two ways from the other p53 peptides in this study. First, there is an insertion of an asparagine residue, resulting in a loss of a leucine in the FxxxWxxL motif. Second, there is a deletion of a threonine located just before the motif. To explore the cause of the decreased affinity, three additional *Danio rerio* p53 peptides were tested. One peptide extends the sequence one amino acid and therefore includes the leucine in the motif (*Danio rerio longer*), this resulted in an affinity of 1.8 μM. The other two peptides were modified to mimic the consensus sequence of the other peptides present in the study, one with a substitution of the asparagine to a leucine (350 nM) and the other with an insert of a threonine (2.3 μM) (Fig 15). All the modified peptides increased the affinity.

**Figure 14.** Schematic species tree of vertebrates with binding affinity of contemporary complexes in blue font. The whole genome duplications are denoted 1R and 2R (Figure 2 in Paper III).

**Figure 15.** Alignment of the four *Danio rerio* peptides tested and their affinity to MDM2 in *Danio rerio*.

To shed further light on the evolutionary trajectory, the paralogs p53, p63 and p73 were examined for their affinity against MDM2. These experiments were performed with a 15 residues p53 peptide, instead of the 12 residues peptide used for the ancestral and extant p53 variants. The reason for this was to facilitate comparison with previously published affinities in this ex-
Experimental setup (68, 74, 96, 97). The affinity of the human MDM2 against human p53 was recorded to 100 nM. The discrepancy in affinity of the 12 and 15 residues p53 peptide has been proposed to be a result of the N-terminal flexible lid in MDM2 which occupies the binding pocket in the unbound state of MDM2 (69). The affinity of p73 was measured to 300 nM, while p63 has a much lower affinity of 6.1 μM. This is a good example of how gene duplication result in functional divergence.

In paper IV, the human paralogues, p53, p63 and p73, were further examined for their ionic dependence in the interaction with human MDM2. Kinetic experiments at varying ionic strengths were performed in order to estimate the basal association rate constant, which is in the absence of electrostatic interactions. In contrast to other IDPs, the results displayed a small ionic dependence. The basal association rate constant for p53 and p73 was estimated around $2\cdot5\cdot10^6$ M$^{-1}$s$^{-1}$, and for p63 it is slightly higher, $13\cdot10^6$ M$^{-1}$s$^{-1}$. These values are, however, in accordance with other IDP interactions.
Concluding Remarks and Future Perspectives

The content of this thesis has contributed with an evolutionary perspective of disordered regions and their interactions. The evolution of IDPs are commonly examined by studying sequence data from present-day species. The development of prediction tools for disordered regions have given plenty of useful information about the conservation of disorder in the different domains of life (30) as well as within protein families (98). Our approach to study the evolution of disorder involves ancestral sequence reconstruction, and may give additional information of the time point in evolution important changes has altered a function. Ancestral sequence reconstruction has been utilized to investigate the evolution of structured proteins and has been able to identify several important shifts in function. For example, how promiscuous ancestral enzyme evolved to be selective to certain substrates (99) and the evolution of multimeric complexes (100) as well as the origin of feathers in dinosaurs (101). The fast evolution and high sequence divergence of some disordered regions may not be suitable for ancestral reconstruction. However, when possible it can give useful information of what the past looked like.

In this thesis, ancestral disorder has been resurrected and biophysically characterized. The evolution of the NCBD-CID complex is an example of how an interaction between IDPs emerged and gained a novel function in transcriptional co-regulation. In this case, the disordered features in NCBD and CID may have facilitated the gain of function. We showed that the disordered properties in the individual domains are preserved during evolution, however the complex has evolved to be more rigid resulting in a higher affinity. Molecular simulations indicated that a few residues are responsible for the increase in affinity, therefore ancestral mutations were introduced in human NCBD. However, these results did not remarkably affect the affinity, suggesting that there is more to it than site-specific contacts for the interaction to be altered. This may be the case of several IDP interactions.

In the evolution of p53 and MDM2, the interaction domains have been preserved in all deuterostomes, but only in specific phyla in protostomes species. Additionally, the ancestral reconstruction of MDM2 resulted in many positions with low probability further back in evolution than the vertebrate lineage. This suggests a high substitution rate in species that diverged from the ancestor of deuterostome and protostome, where species that preserved
the domain have gained a beneficial function, and, on the contrary, the species that lost the domain acquired deleterious mutations abolishing the interaction. For p53, similar assumptions can be made, the 12-residue peptide used in the study is well conserved, however the adjacent regions differ substantially in sequence in extant species. A mutation of the phenylalanine or the tryptophan in the binding motif has been shown to abolish binding, and if such a mutation occurred during evolution it may have led to the disappearance of the whole domain. Within the vertebrate lineages, differences in affinity and salt dependence can be explained by functional diversification. Affinities that slightly differ in tetrapod species can be explained in the same manner. However, in the p53 vertebrate lineage the zebrafish has a surprisingly lower affinity compared to contemporary ancestral and extant vertebrate complexes. There are several fish species closely related to the zebrafish with analogous sequence characteristics (Paper III, Suppl. Fig 2). This implies that a large fraction of all fish species has a reduced affinity between p53 and MDM2. If this change in affinity is true, there must either be an altered expression, regulation or other consequence for the cell. Although, studies of zebrafish reveal that p53 and MDM2 have a similar function and are a part of the same pathways (102, 103). The third whole genome duplication in the fish lineage gave additional opportunities for novel function and expansion of species divergence. However, no copies from the p53 or MDM2 family are preserved. On the other hand, copies of downstream targets may have been preserved and may have altered function or expression.

In our studies, the main focus has been to examine the affinity of protein-protein interaction involving disorder. Affinity is a quantitative measurement and can give useful information in order to elucidate binding mechanisms, for example when comparing wildtype protein with mutant or ancestral protein with extant variants. IDPs have important functions in regulatory pathways, and therefore the maintenance of their structure/disorder-function relationship is crucial. It has been proposed that IDPs bind with high specificity and low affinity to account for their regulatory role. Fast association as well as dissociation of a complex is a way of turning on and off a function in a quick manner. However, it is important to put affinity and the interaction in a cellular context. How high are the expression levels, and in which subcellular location and environment does the protein-protein interaction occur? To get the whole picture, interdisciplinary efforts combining bioinformatics, biophysical characterization and cell studies are needed. One good example is a study that examined the interplay between disorder and affinity of p53 and MDM2, and how it affected expression levels and functional differences in human cell lines (72). The results show that increasing residual helicity in p53 leads to higher affinity to MDM2 as well as altered dynamics. In a cellular context, this resulted in reduced expression of target genes and loss of the ability to induce cell cycle arrest upon DNA damage.
Populärvetenskaplig sammanfattning

På samma sätt som man kan lära sig om nutiden genom att studera historien, kan man lära sig om proteiners funktion i kroppen genom att studera evolutionen. I min forskning har jag undersökt hur ostrukturerade proteiner har förändrats under evolutionen och hur de interagerar med andra proteiner i kroppen. Studien har genomförts genom att återuppliva en proteindomän som existerade för hundratals miljoner år sedan för att sedan jämföra den med nutida varianter. Fokus har varit på att undersöka hur starkt två ostrukturerade proteindomäner binder till varandra och följa hur de interagerat genom evolutionen. Genom att studera ostrukturerade proteiner och dess evolution kan man lära sig om hur viktiga processer i kroppen fungerar i detalj, vilket i längden kan leda till upptäckter av nya läkemedel för att bota sjukdomar.


Acknowledgements

In this section I would like to express my gratitude with a rhyme, to those who facilitated the work in this thesis and my conversion to a doctor, just like an enzyme. First and foremost, to my main supervisor Per Jemth: Thanks for making me a part, of a group that have biochemistry in their heart. I’ve been given a good mixture of trust, support and outsmart, in order to improve my skills in the protein science art. It’s soon time for me to depart, with a lot of new knowledge in my cart. To my second supervisor Greta Hultqvist: You taught me all about phylogenetic trees, and helped me gain bioinformatic expertise, good luck with your research on Alzheimer’s disease. To my second supervisor Maria Selmer and my examiners Pernilla Bjerling and Anna-Karin Olsson: It was always nice to know, I could come talk to you if I wanted advice or if someone stepped on my toe, thanks for being around for a friendly hello. To Eva Andersson: You have helped me more than twice, thanks for being super nice, you are as cool as ice. To co-authors of papers: Working with you has been a true delight, I hope your future is bright and you continue writing papers I can cite. To all the people at the IMBIM department I met, both you with and without a pipette. My time in teaching must be addressed, I had loads of fun working with you to aid students on their quest. The administration does nothing but impress, what you do is far better than a beta test. To the iPha association and all the rest, for sharing a beer and telling a good jest. I must have been blessed, because my time at IMBIM has been the best. To former and present members of the lab, without you my time as a PhD student wouldn’t been as fab. You have always given me a happy face, which been helpful after a bad stopped flow trace. To all my old and new friends: Hanging out with you is as great as space, just like playing card with a handful of ace. I like you guys very much, hope we will keep staying in touch. To Erika Melin: For being awesome and a perk, in the making of the cover artwork. To my family: You have always given the support I need, with you by my side I cannot do anything but succeed. Extra thanks to my sister for the proofread, I can always count on you guaranteed. A last big hug to all who contributed to my new degree, this would not have been possible without you and me.
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1424

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)