Vyacheslav Tretyachenko

Never Born Proteins: Occurrence and characterization of secondary structure motifs

Department of Biochemistry

Supervisor of the master thesis: Mgr. Klára Hlouchová Ph.D

Study programme: Biochemistry

Prague 2015
Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze ............ Vyacheslav Tretyachenko .....................
Poděkování

Rád bych poděkoval především vedoucí své práce Dr. Kláře Hlouchové za neustálou podporu, navenění na správnou cestu ve chvílích kdy jsem se už topil a za domácí maliny. A také velké díky za všechny večerní hodiny nad těmito řádky, díky nímž mohu prezentovat tuto práci a bez kterých by moje angličtina zůstala na kuchyňské úrovni.

Díky patří našemu garantovi a vedouc9mu laboratoře Doc. Janu Konvalinkovi za vřelý přístup, inspirující myšlenky a taky malou místnost na mojí alma mater, která se mi na dva roky stala druhým domovem.

Dr. Lucii Bednárové děkuji za všechnu práci, korekce v předkládáné práci a ochotu spolupracovat i v těžkých chvílích.

Nakonec bych rad podekoval Dr. Václavu Veverkovi a Mgr. Rozalií Hexnerové za pomoc při měření a interpretaci NMR spekter, celému kolektivu laboratoře Doc. Konvalinky a katedry biochemie za přátelskou společnost a ochotu vždy poradit. Poslední poděkování patří rodičům za to, že mi byli vždy oporou v lepších i horších časech.
Contents

Abstract 4

1 Theoretical part 6

1.1 Protein universe .............................................. 6
  1.1.1 Protein sequence universe .......................... 6
  1.1.2 Protein structure universe .......................... 6
      1.1.2.1 Theoretical studies ........................ 7
      1.1.2.2 Protein design and engineering experiments ... 7
      1.1.2.3 Random-sequence experiments .............. 8
  1.2 Protein secondary structure elements ................. 9
      1.2.1 Secondary structure .................................. 9
          1.2.1.1 Helix conformation ......................... 10
          1.2.1.2 Beta structure .......................... 11
          1.2.1.3 Turns and loops ........................ 13
          1.2.1.4 Preference of amino acids in aspect to secondary structures .................. 13
      1.2.2 Tertiary structure .................................. 14
          1.2.2.1 Fold classes ................................ 14
          1.2.2.2 Intrinsically disordered proteins (IDPs) ... 15

1.3 Protein structure prediction .................................... 17

1.4 Methods of protein structure analysis .................. 19

2 Aims of the thesis 23

3 Materials and instruments 24

3.1 Materials .................................................. 24
  3.1.1 Chemicals ............................................ 24
  3.1.2 Buffers ................................................ 25
  3.1.3 Enzymes ............................................. 26
  3.1.4 Competent cells ..................................... 26
  3.1.5 Other materials ..................................... 27
3.2 Instruments ................................................................. 27

4 Methods ................................................................. 28

4.1 DNA manipulation .................................................. 28

  4.1.1 Horizontal agarose gel electrophoresis .................. 28

  4.1.2 Restriction analysis of DNA sample ...................... 28

  4.1.3 Isolation of DNA from agarose gel ....................... 28

  4.1.4 Ligation of protein codin gene into the expression vector 29

  4.1.5 Transformation of bacterial cells ......................... 29

  4.1.6 Isolation of the plasmid DNA from E. coli cells .......... 29

4.2 Protein expression and isolation .................................. 30

  4.2.1 Protein expression .............................................. 30

  4.2.2 Isolation of inclusion bodies ............................... 30

  4.2.3 Renaturation of expressed proteins ...................... 30

    4.2.3.1 Renaturation by dialysis ............................ 30

    4.2.3.2 Renaturation by rapid dilution ...................... 31

4.3 Protein purification .............................................. 31

  4.3.1 Metal – chelate affinity chromatography ................. 31

    4.3.1.1 Purification on Ni-NTA matrix gravity flow column .... 31

    4.3.1.2 Purification on TALON matrix gravity flow column .... 31

4.4 Protein manipulation ............................................. 32

  4.4.1 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) ..................... 32

  4.4.2 Silver staining of proteins in polyacrylamide gel .... 33

  4.4.3 BioSafe Coomassie staining of polyacrylamide gels ... 33

  4.4.4 Western blotting using chemiluminiscent substrates ... 33

4.5 Preparation of selected NBPs .................................. 34

  4.5.1 4090 ............................................................. 34

  4.5.2 2298 ............................................................. 34

  4.5.3 6387 ............................................................. 34

  4.5.4 4712 ............................................................. 35

  4.5.5 6620 ............................................................. 35

4.6 Preparation of protein samples for different methods of structural analysis .................. 35

  4.6.1 CD/FTIR/fluorescence spectroscopy ...................... 35

  4.6.2 NMR spectroscopy ............................................. 36

  4.6.3 DLS ............................................................... 36
Abstract

An experimental study on randomly generated protein sequences can provide important insights into the origin and mechanism of secondary structure formation and protein folding. In this study we bring biophysical characterization of five protein sequences selected from the in silico generated library of random chains. The sequences were selected on the basis of bioinformatic analysis in order to find the candidates with the maximum potential to possess secondary structure. This study shows that the random polypeptide sequences form stable secondary structures and in some show the signs of tertiary structure, such as hydrophobic core formation and distinctive oligomerization pattern.

While the work presented in this thesis is work in progress on a larger study, the data already demonstrate that unevolved protein sequence space provides a lot of potential for secondary and tertiary structure formation that awaits its characterization.
Abstrakt

Studium náhodných proteinových sekvencí skýtá velký potenciál pro pochopení původu a mechanismu tvorby hydrofobního jádra a uspořádání struktury proteinů. V této praci byla provedena biofyzikální charakterizace pěti proteinových sekvencí bez homologie s dosud známými proteiny. Sekvence byly vybrány na základě bioinformatické analýzy, jenž spočívala ve vyhodnocení 10 000 náhodných polypeptidových řetězců pomocí prediktorů sekundárních struktur, rozpustnosti a celkové proteinové neuspořádánosti. Charakterizace vedla k potvrzení přítomnosti sekundárních struktur a v některých případech také k odhalení náznaků nativního uspořádání proteinů - tvorby stabilního hydrofobního jádra a tendencí k specifické oligomerizaci.

Přestože uvedená práce je pouze součástí většího projektu, výsledky již ted naznačují, že náhodné proteinové sekvence mají potenciál k tvorbě sekundární i terciární struktury.
1. Theoretical part

1.1 Protein universe

1.1.1 Protein sequence universe

How large is the protein sequence space? It is well known, that Nature uses only a small fraction of possible sequences to construct protein products. Simple arithmetics show, that for a small protein composed of 100 amino acids from the canonical 20 amino acid set there are $20^{100}$ different combinations. On the other hand, the count of all existing proteins (including isomers and mutated variants) is estimated approximately to $10^{15}$. The ratio between the number of existing vs. possible polypeptide chains has been often compared to the ratio between the radius of the hydrogen atom and radius of the universe[1]. Moreover, if we take shorter polypeptides - 40 amino acids in length - and synthesize all possible sequences at a rate of $10^6$ molecules per second, it would take approximately $10^{44}$ years to complete their syntheses[2]. These thought experiments show the vastness of protein sequence space and force us to question the uniqueness of naturally occurring proteins. Was the actual set of protein sequences "chosen" accidentally or does it represent quite a unique scenario to provide the structural and functional diversity of modern proteins? How frequent and variable are structures in the protein sequence space and could alternative protein sequence selection lead to similar evolutionary optimization? These questions lead us to the subject of this thesis which is an examination of the structural properties of random protein sequences.

1.1.2 Protein structure universe

Although several million unique protein sequences have been deposited in non-redundant databases (e.g. Uniprot) and close to one hundred thousand protein structures have been deposited in the corresponding databases of protein structures (e.g. the PDB), structural classification databases (such as SCOP and CATH) have amassed just 1,500 different domain families that account for 70% of genomic sequences[3]. About 200 of these families are common to all kingdoms of life and they account for approximately 50% of genome structural annotations. Thus, it appears that a finite (and relatively small) library of domains has been created/selected during evolution and yet it gives rise to such an incredible biodiversity and organism
complexity\[4, 5, 6\]. This is reflected by the gradual decrease in the percentage of new structures classified in CATH that are observed to possess a novel fold (Fig. 1.1). Early hypotheses assumed that existing proteins adopt specific three dimensional (3D) folds because natural selection has evolved rare sequences with such ability, while it was implied that unevolved sequences (i.e. random-sequence proteins) would likely be disordered\[7\]. However, more recently, a number of (i) theoretical studies, (ii) protein engineering experiments, and (iii) experiments in synthetic random-sequence polypeptides (i.e. outside the biological data sets) have shown that unique, compact folds may be much more frequent throughout the sequence space than once presumed.

Figure 1.1: Exponential decrease of new fold discoveries over time suggest the limited count of total protein folds\[8\]

1.1.2.1 Theoretical studies

Sampling of random sequence space with computational analysis of obtained sequences was performed by Minervini et al. and Prymula et al.\[9, 10\]. Both random libraries were analyzed by Rosetta prediction software (for tertiary structure prediction) with results suggesting great abundance of folded conformation with similar secondary structure content as is known for natural proteins. These studies may however be of limited reliance for sequences with absent similarity to structurally characterized proteins and may over-score the structural properties of such (see section 1.2.2 for more detail)

1.1.2.2 Protein design and engineering experiments

Over the last 20 years, the group of Michael Hecht has made a great progress in designing pseudo random protein libraries and assessing their structure and function potentials. In short, they performed a systematic classification of 20 amino acid alphabet into two groups based on their hydrophobicity
(a binary pattern). The output of the studies is a set of directions how to construct libraries of semi-random polypeptides (based on a structural template, such as the four helix bundle) with embedded secondary structures (both helices and sheets) defined only by hydrophobic/hydrophilic amino acid pattern. The studies have been mostly successful using alpha helical structural scaffolds so far as the tested beta structures were more prone to aggregation. In general, their studies showed the possibility of constructing folded and functional enzyme-like catalyst from semi-random polypeptide sequence with enzyme activity similar, or even higher than that observed for several enzymes designed using rational design or automated computational methods[11]. Moreover, some studies registered the presence of catalytic promiscuity of these unevolved random sequence enzymes[12]. These results altogether support the thought that random sequences can serve as a vital "feedstock" for evolution, providing broad rudimentary activities.

1.1.2.3 Random-sequence experiments

Studies of solubility of artificial proteins showed that random sequences with reduced amino acid alphabet (composed of evolutionary younger amino acids) showed fairly higher solubility than that of canonical 20 amino acids[13]. Davidson and Sauer conducted research of a library of random sequences constructed using leucine, glutamine and arginine only[14]. By sparse sampling of the library in E. coli, they suggest the presence of structure in roughly 5 % of random sequences based on their high expression in E.coli and further characterization of selected polypeptides. Further survey of random libraries showed the possibility of hydrophobic core formation, a distinct enzymatic activity and subsequent evolvability of random chains[15, 16, 17]. Chiarabelli used phage display technique to construct a library of 50 residue long random sequences (including a thrombin tripeptide substrate site) and suggested secondary structure occurrence in the library by the polypeptide resistance to proteolytic digest (likely due to folding) for more than 20 % sequences[2]. A more recent study by LaBean et al. shows that library of random sequences of 71 amino acids can provide the polypeptides with collapsed conformation and cooperative unfolding behaviour[18]. The frequency of protein function within a random sequence polypeptide library was examined by Keefe and Szostak using an ATP binding activity assay and a library size of $6 \times 10^{12}$[19]. They successfully selected four new ATP-binders from the library that were unrelated to any biological proteins.

Although there is only a limited number of studies of the vast protein
sequence space, these results altogether suggest the evolutionary potential of random polypeptide sequences by means of both structure and function even if only a subset of the amino acid alphabet is used. In general, artificial polypeptides from random libraries are usually less structured and functionally more diverse than natural proteins which got through the eons of evolutionary selection.

1.2 Protein secondary structure elements

1.2.1 Secondary structure

Overall protein folding is accompanied by formation of structurally distinct local amino acid arrangements. These arrangements are conformationally restricted by several factors. First - the peptide bond between two residues is not completely rigid, its dihedral angle ($\omega$) may acquire two values, corresponding to configuration cis and trans. Both configurations are interchangeable, however trans configuration is about $1000 \times$ more stable than cis (with the exception of proline, which allows protein to reach otherwise impossible configurations)[20].

![Ramachandran plot](image)

**Figure 1.2:** Ensemble of all amino acid phi/psi angles in protein is often plotted in form of Ramachandran plot (left). Areas of allowed phi/psi combination are cyan, experimentally measured angles are represented as blue dots. The global conformation of polypeptide chain can be described by dihedral angles of individual amino acids (right). $\phi$ corresponds to the dihedral angle between the $C_\alpha$ and nitrogen of peptide bond. $\psi$ is the angle between $C_\alpha$ and C=O[21].

The second restriction for accidental backbone conformation is in $C_\alpha$-N and $C_\alpha$-C dihedral angles ($\phi$ and $\psi$), which may acquire distinct values. These values were calculated by G.N. Ramachandran by use of van der Waals radii of atoms in different dipeptides[22, 23]. When the allowed $\phi$, $\psi$ and $\omega$ values are applied to a model of protein chain, simple local folds emerge - coils,
loops, helices and pleated sheets (Fig. 1.2). These are the elements of secondary and foundations of protein tertiary structure. Secondary structure strongly correlates with both the overall protein folding and the activity\textsuperscript{24}. When combined, helices and sheets comprise more than half protein residues. Using statistics of 707 nonhomologous protein chains, alpha helices are more abundant (roughly 31\%) than beta sheets (20\%\textsuperscript{25}). When writing this thesis, no similar reference to statistical occurrence of secondary structure motifs in random sequence libraries was found and is probably still unknown.

\subsection{1.2.1.1 Helix conformation}

The helical conformation of amino acid chain was theoretically proposed by Pauling in 1953 despite the fact, that there was no experimental data on globular protein structure\textsuperscript{26, 27}. Alpha helix is a local structure with $\phi/\psi$ angles of -54°/-47°, 3.6 residues per turn and the total length of 5 to 40 residues\textsuperscript{28}. These properties define alpha helix as compact structure, with side chains faced to periphery with a lateral spacing of 100 degrees between them. Another feature of alpha helices is the hydrogen bonding between backbone carbonyl and amide groups located four position from each other. This bonding is of great importance to protein stability although studies suggest that they are not the major stabilizing force\textsuperscript{29, 30}. This issue was studied extensively and it seems that alpha helices are stabilized also by van der Waals interactions of amino acid side chains\textsuperscript{31, 32}. These interactions involve the surrounding environment, since isolated alpha helices are usually unstable\textsuperscript{33}.

Another property of alpha helices is the overall charge distribution through the structure. Since carbonyl oxygen is partially negative and the amide hydrogen is partially positive a peptide bond represents a dipole with strength of 3.5 D. All of these dipoles are arranged in the same nature – carbonyl on the C-end and amide faces the N-end of the helix. Combined all together they create a larger dipole of a strength 5.0 D\textsuperscript{34}. This characteristic of al-
pha helices serves to further functional and structural fine tuning of protein molecule.

Alpha conformation is not the only representative of peptide helical structures. Other, much less common are $3^{10}$ and $\pi$ helices. The name of the $3^{10}$ helix originates from three residues and ten backbone atoms per helix turn. This conformation corresponds to $\phi/\psi$ angles of $-49^\circ/-26^\circ$, with narrower and longer geometry than alpha helix\textsuperscript{28}. $3^{10}$ helix is rarely observed as a separate structure, since narrow shape and $i-i+3$ hydrogen bond pattern are problematic for protein packaging and dipole alignment within the structure.

![Figure 1.4: $\pi$ conformation is shorter and wider than alpha helix with $\phi/\psi$ values of $-57^\circ/-70^\circ$ and 4.4 residues per turn (left). $3^{10}$ helix is more narrow structure with $\phi/\psi$ values of $-49^\circ/-26^\circ$ and 3 residues per turn (right)](image)

$\pi$ helix is shorter and wider than canonical alpha helix, with $\phi/\psi$ values of $-57^\circ/-70^\circ$ and 4.4 residues per turn\textsuperscript{28}. Hydrogen bond formation pattern of $i-i+5$ is energetically unfavorable, so the $\pi$ helix was for many years considered a hypothetical shape. $\pi$ helix tends to occur, like $3^{10}$ conformation, at the edges of alpha helices with a length of a few residues.

1.2.1.2 Beta structure

$\beta$ sheets are known as less compact compared to alpha conformation, with $\phi/\psi$ values $-120^\circ/120^\circ$. Segments of $\beta$ structure are called $\beta$ strands, each strand consisting of typically five to ten residues. The distance between C\textsubscript{\textalpha} atoms in beta strand is about 3.5 Å, thus over twice as long as in alpha helix. The extended conformation of $\beta$ strand does not allow the intramolecular
stabilization, like in alpha helix structure. β strands are usually arranged alongside one another to form an extended flat structure called β sheet. Beta sheets appear in two possible forms. First, when adjacent β strands are separated from each other by a short stretch of amino acids (β turns), allowing the strands to rearrange alongside each other in opposite direction (anti parallel β sheet). Second form is created by β strands, separated within the sequence by much more than a few residues. As a result, the strands have an opportunity to align to each other in the same direction(parallel β sheet).

Contrary to alpha helix, the side chains of amino acids in β structure are faced up and down instead of sideways. The orientation of side chains prevents the formation of characteristic macro dipole, such as in alpha helices. Another difference compared to helix structure is longer distance between the interacting backbone amide nitrogen and carbonyl oxygen due to positioning of these atoms on separated strands. This separation leads to lower stability of partially unpaired strands on the both sides of β structure. One of the elegant solutions that Nature uses to solve this problem is to fold the entire sheet into the larger structure, for example a cylindrical shape called β barrel.

Interestingly, β structures (besides IDPs, chapter 2.2.2) were suggested as the first protein structures that were more accessible even to reduced (prebiotically accessible) amino acids. The theoretical studies reported that β structures are more thermostable than alpha helices and therefore also more resistant to chemical degradation[36]. It was pointed out, that β sheets can form bilayers with a hydrophobic interior and hydrophilic exterior, offering an early potential compartmentalization option[36] and that unlike alpha helices they form only homochiral chains that protect their constituent amino acids from racemization[37].
1.2.1.3 Turns and loops

Segments of polypeptide chain, responsible for changing direction are called turns or loops. The difference between these two lies in structural background – while loop is usually a long stretch of amino acids, without any regular conformation, turn has an ordered structure consisting of four residues. The second and fourth residue of turn are usually cis-proline and glycine respectively. The other two positions of the turn are less conserved, although asparagine and aspartate are common in that position. Another property of turn is van der Waals side chain stabilization between the adjacent strands and characteristic hydrogen bonding between first and fourth residue of the turn.

1.2.1.4 Preference of amino acids in aspect to secondary structures

Prediction of secondary structure relies partly on specific preferences of amino acids to form different secondary structures. Alpha helix generally prefers residues with small or linear side chains, as these have the least chance to clash with each other in side-by-side helix conformation. In addition, folding the amino acid chain into the helix leads to the thermodynamically unfavorable loss of the entropy, so the smallest amino acids provide the lowest starting entropy. The preferred amino acids are Ala, Glu, Leu, Arg, Met and Lys [39, 38]. Beta sheet structure shows strong amino acid preferences on their termini only. At C-termini of beta structure there is a preference for Asp, Asn, Ser and Pro (i.e. beta sheet terminators). Interestingly these amino acids are the same that occur at the N-termini of alpha helices. N-terminus of β sheets is most commonly embraced by Arg and Lys [40]. Deduction of amino acid preferences inside the beta sheet is difficult, as beta sheets do not fold in isolation and cannot be rationalized by structural thermodynamic studies as easily, as alpha helical structure. However, according to Andersen and Rost, Val and Ile are over-represented, while Gly, Asp and Pro are under-represented (Fig. 1.6) [38].
1.2.2 Tertiary structure

The overall three dimensional conformation of proteins, folded by secondary structure elements, is termed tertiary structure. Generally, proteins are divided into the two classes – fibrous and globular. Fibrous proteins (often insoluble, with a structural role) are elongated chains, in which the secondary structure forms the dominant tertiary structure, while globular proteins have compact tertiary folds (section 1.2.2.1). Subsequent chapter is devoted to intrinsically unfolded proteins (IDPs), which acquire their compact conformation only after binding to protein ligand/partner or in a specific environment. The occurrence of IDPs within existing proteins is estimated to 30 % [41].

1.2.2.1 Fold classes

All distinct globular protein folds could be (according to SCOP database) virtually divided to four structural classes – all $\alpha$, all $\beta$, $\alpha/\beta$ and $\alpha+\beta$.

1. all $\alpha$ proteins consist only of alpha helices. The appearance of fold may differ. Typical representatives are RNA binding protein ROP, subtilisin or cytokines.

2. all $\beta$ proteins are folded using only $\beta$ chains. For this class structures like $\beta$ barrels, $\beta$ propellers, and $\beta$ sandwiches are very characteristic. Examples are plastocyanin, transglutaminase or porin

3. $\alpha/\beta$ is a mixed class, which is characterized by repeating $\beta-\alpha-\beta$ units with outer layer composed of alpha helices and central core of parallel $\beta$ sheets. Many enzymes of glycolysis occupy this folding class. The most important example of $\alpha/\beta$ protein class is triose phosphate isomerase with $\alpha/\beta$ barrel structure.

4. $\alpha+\beta$ is another mixed class consisted of significant portion of $\alpha$ and $\beta$ structural elements. In opposite to $\alpha/\beta$, these proteins lack any specific topology. Examples are cysteine proteases, carbonic anhydrase and lysozyme.

The origin of fold classes was subjected to computational analysis which recognized $\alpha/\beta$ protein class as the oldest and the all $\alpha$ and $\beta$ proteins as the newest addition to the protein folding family [42]. In other words, the results suggest that recently born and still evolving proteins belong to all $\alpha$ or all $\beta$ class (as well as their random mixtures, $\alpha+\beta$ class), but the majority of the “mature” (evolved) proteins belong to $\alpha/\beta$ class. Furthermore, it
seems that early evolution produced most of the individual domains whereas subsequent evolution has been mostly re-organizing them into multi-domain structures\[3\].

1.2.2.2 Intrinsically disordered proteins (IDPs)

Proteins are traditionally perceived as more or less rigid structures with restricted degree of flexibility. Throughout 20th century, as the database of solved protein structure expanded, the more complex nature of protein structure has been revealed. Protein disorder, essential for the molecule function, seems to be more abundant, than it was thought before. The growing interest in the field of intrinsic protein disorder is clearly demonstrated by exponential increase in number of publications and citations on the structurally and functionally characterized IDPs (Fig. 1.8). It is estimated that one third of the proteins solved by X-ray crystallography is intrinsically disordered in their nature\[44\]. Within the human proteome, approximately 50 % of all proteins are predicted to contain long disordered segments (longer than 30 residues)(Classification of intrinsically disordered regions and proteins) with
enrichment of 70% of all polypeptide sequences among transcription factors and signalling proteins[45]. Moreover, bioinformatic methods suggest, that the frequency of IDPs in the genome of an organism positively correlates with the organism’s complexity[46].

Interestingly, IDPs have been suggested as logical progenitors of the protein world, because flexible conformation may permit broad (generalist) functional diversity[47]. In agreement, unfolded and flexible structures have been observed, when randomized polypeptide libraries were used for selection of new functions[48] and they have been observed by structural analysis of ancient ribosomal proteins[49].

The representation of individual amino acids in IDPs significantly differs from that of ordered proteins[51, 52, 53]. Primary structure is often enriched in charged residues, structure breakers (Pro, Gly) and alanine (Fig. 1.9).

In addition, these proteins are depleted in hydrophobic and aromatic amino acids. The low hydrophobicity with simultaneous high net charge is an important prerequisite for the absence of compact structure under physiological conditions. This simple empirical observation is applied in disorder prediction methodology by plotting the absolute net charge as a function of the mean normalized hydrophobicity, the final plot being denoted as charge-hydrophathy (CH) or Uversky plot[41]. The disordered proteins are clustered in upper left region of CH plot. Protein disorder is characterized by low protein complexity (repetitive areas and sequences of low diversity[54]. According to earlier record of SwissProt database, 34% of all proteins have at least one such segment of low complexity[51].

When traditional methods, such as SDS-PAGE analysis and gel chromatography, are applied to IDPs, they show unusually low electrophoretic mobility and high hydrodynamic radii[52]. For SDS-PAGE, the mass is typically estimated to 1.2 - 1.8× higher, than expected from amino acid sequence[51]. These observations are explained by enrichment in acidic residues and extension of protein molecule in solution[55]. IDPs are known for abnormal
Figure 1.10: Charge hydropathy plot of protein disorder, Net charge vs mean hydrophobicity is plotted for disordered (red) and ordered (blue) proteins\cite{50}.

sensitivity to proteolysis, thus limited proteolysis is one of the methods of choice for distinguishing ordered and disordered species. Another specialty of IDPs is their low susceptibility to high temperatures and stability toward acidic treatment\cite{56}. These specific conditions are often exploited for purification of intrinsically disordered proteins\cite{57}. For more detailed structural characterization of IDPs, variety of biophysical techniques are used - mostly nuclear magnetic resonance (NMR), small angle X-ray scattering and single molecule FRET (see review \cite{58}). The residual structure of IDPs can also be studied by circular dichroism (CD) which characterize the unfolded polypeptide by a minimum near the 190 and 200 nm\cite{59}.

1.3 Protein structure prediction

1.3.1 Secondary structure prediction

The deduction of three dimensional protein structure, even with the best modern algorithms, is full of contradictions (see section 1.3.2). Secondary structure (SS) prediction produces more reliable results through the application of different approaches. All of modern SS predictors rely on amino acid propensities for occurrence in specific structures, yielded from known protein structures. In addition to calculated propensities, including of evolutionary information, acquired via multiple sequence alignments, to SS content evaluation greatly improved performance of prediction\cite{60}. Another crucial step, especially for parallel beta sheet prediction, is including information on the long-range amino acid interactions in known structures into the predic-
tion model[61]. The way of propensity value generation is different for each predictor. One of the first successful attempts to predict secondary structures from amino acid sequence was carried out by Chou and Fasman[62]. The principle of the method was in an assignment of structural propensity for each of the twenty amino acids and calculation of the probability of secondary structure occurrence by summation and normalization of amino acid stretch propensity values. A very popular prediction method was proposed by Garnier in 1996[63]. Prediction uses the information theory and bayesian approach to calculate the properties and occurrence of amino acids and secondary structures. The accuracy of these early methods was between 60 and 70 %[64]. Another widespread method is fold recognition which is based on an assumption that with correct fold identification, the problem of secondary structure prediction is largely solved. It was shown that the secondary structure in pairs of structurally homologous proteins (sequence identity at least 30 %) was in average 88 % identical[65]. The most successful modern methods of prediction depend on combination of homologous pattern analysis and advanced machine learning techniques (neural networks, nearest neighbour classification, Hidden Markov Modelling etc), reaching 80 % prediction accuracy.

1.3.2 Tertiary structure prediction

Prediction of 3D protein structure is a tough problem. In computational complexity theory such kind of tasks are classified as NP-complete problems[66], i.e. they are the hardest tasks in terms of computational requirements. The performance and reliability of structure prediction algorithms have been annually evaluated in CASP (Critical Assessment of Structural Prediction) competition since year 1994 (http://predictioncenter.org/). The principle of these algorithms is usually a combination of molecular mechanics calculations and knowledge based predictions of the protein structure. According to the data from CASP experiments, ROSETTA and I-TASSER were the most successful predictors during the past years. Both of them first divide the given sequence into shorter oligopeptide fragments, these are compared with known substructures in database of solved protein structures and at last the fragments are merged together using pure ab initio approach, which select the structure with the lowest possible potential energy. The success of the prediction is encouraging, when used for sequences considerably homologous to those with known structure. Because the study presented in this thesis is focused on random, unrelated protein sequences, tertiary structure predictors
are not employed within this work\textsuperscript{67, 68, 69}.

### 1.4 Methods of protein structure analysis

#### 1.4.1 Circular dichroism (CD)

Circular dichroism is one of chiral spectroscopic technique analogues to absorption spectroscopy, depending on the spectral range the electronic (ECD) (in UV/VIS spectral region) and vibrational circular dichroism (VCD) (in infrared spectral region) could be used. In principle, CD is the difference in the absorption of left handed circularly polarized light (LCPL) and right handed circularly polarized light (RCPL) on the wavelength

\[
\Delta A = A(\lambda)_{LCPL} - A(\lambda)_{RCPL}
\]  

(1.1)

where \( \lambda \) is the wavelength. In general, chiral spectroscopies are used to study chiral molecules of all types and sizes. They allow us to study relatively small molecules with or without biological significance with central chirality (described by the general R,S system), optical isomers (enantiomers). They are advantageously used for the structural characterization of biopolymers (peptides/proteins, nucleic acids), because of the more or less regular repetition of their chiral subunits (mostly L-amino acid residues in peptides/proteins, D-sugars in nucleic acids, etc.) which gives rise to segments of chiral secondary structures often displaying intense chiroptical spectra. Chirality of biomolecules and its relationship to biologically important processes is currently being thoroughly investigated and it represents the leading direction in applications of chiroptical spectroscopies\textsuperscript{70}.

Protein conformational analyses are based on evaluation of the changes in secondary structure content because different secondary structure elements (such as alpha helices, \( \beta \) sheets or \( \beta \) turns) typically give rise to unique, characteristic spectral features. The fastest and easiest secondary structure evaluation is usually provided by ECD spectroscopy. ECD bands characteristic for particular types of secondary structures can be found in the spectral region of amide group absorption (far-UV CD below 250 nm). In this region, a typical protein spectrum shows the \( \pi - \pi^* \) (\( \sim 190 \text{ nm} \) and \( \sim 140 \text{ nm} \)) and \( n - \pi^* \) transitions (\( \sim 220 \text{ nm} \)). The region below \( \sim 180 \text{ nm} \) is not accessible by commercial ECD instruments. However, this limitation can be overcome by employing synchrotron radiation circular dichroism (SRCD) which uses an intense synchrotron light beam and allows ECD measurements down to the vacuum-UV
There are several analytical programs (e.g. Dichroweb[71], Protein Circular Dichroism Data Bank[72] or CDPro[73]) which can provide an approximate estimation of secondary structure fractions based on experimental ECD spectra. ECD spectroscopy can contribute to tertiary structure assessment. Information on the tertiary structure can be found in the absorption region of aromatic residues (phenylalanine, tyrosine and tryptophan) in the spectral range 250-290 nm and/or disulfide bonds (near UV CD, above 250 nm)[74].

A few rules must be kept during preparation of protein sample for CD analysis. The solvent allowing the far UV ECD measurements are water and 2,2,2-trifluoroethanol (TFE) (solvent cut off 170 nm). This requirement is however difficult to meet, so more convenient seems to be the phosphate buffer or the low concentrations (up to 50 mM) of Tris, HEPES, PIPES buffers with very low concentration of monovalent salts (NaCl, KCl). Also imidazole and reducing agents (TCEP, DTT) are unsuitable due to their high absorption amide absorption region (190 - 240 nm)[75].

**1.4.2 Infrared spectroscopy (IR)**

Since ECD deals with electronic transitions, it hints on the overall three dimensional arrangement of the molecules, but usually lacks enough structural detail. For the more detailed peptide/protein conformational analysis, it is often beneficial to supplement the electronic CD data with IR. IR spectroscopy is mainly used for chemical structure determination via vibrational frequencies characteristic for distinct functional groups of studied molecules. The same is valid for the amide (peptide) bond, which is characterized by amide A and B (NH stretching vibration (~3300 and ~3170 cm\(^{-1}\)), amide I – the C==O stretching vibration with minor contributions from the out-of-phase CN stretching vibration, the C-C-N deformation and the NH in-plane bend (~1650 cm\(^{-1}\)), amide II – the out-of-phase combination of the NH in plane bend and the C-N stretching vibration with smaller contributions from the
C==O in-plane bend and the CC and NC stretching vibrations (~1550 cm\(^{-1}\)), and amide III - in-phase combination of the NH bending and the C-N stretching vibration with small contributions from the C==O in plane bending and the C-C stretching vibration (~1400 - 1200 cm\(^{-1}\)). For the structural evaluation using IR spectroscopy the amide I in combination with amide II spectral band are used. Although obtaining IR spectra of peptides/proteins in aqueous solution is not an easy task, the combination of IR spectroscopy with ECD can provide a rapid estimation of β-structure content, which could be rather problematic using ECD only especially when α-helical structure is prevailing. Measurement procedure require relatively high concentration of protein (5 mg/ml). Secondary structure estimation by FTIR is especially appropriate in combination with CD spectroscopy, since former method is reported to provide more reliable estimation of beta sheet structure while latter has better performance in helical content determination\[76\][77].

<table>
<thead>
<tr>
<th>Secondary structure</th>
<th>Band position in (^1)H(_2)O/cm(^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-helix</td>
<td>1654</td>
</tr>
<tr>
<td>β-sheet</td>
<td>1633</td>
</tr>
<tr>
<td>β-sheet</td>
<td>1684</td>
</tr>
<tr>
<td>Turns</td>
<td>1672</td>
</tr>
<tr>
<td>Disordered</td>
<td>1654</td>
</tr>
<tr>
<td></td>
<td>1648–1657</td>
</tr>
<tr>
<td></td>
<td>1623–1641</td>
</tr>
<tr>
<td></td>
<td>1674–1695</td>
</tr>
<tr>
<td></td>
<td>1662–1686</td>
</tr>
<tr>
<td></td>
<td>1642–1657</td>
</tr>
</tbody>
</table>

*Figure 1.12: Assignment of amide I band positions to secondary structure based on experimental data and assignments of various authors\[78\]*

### 1.4.3 Dynamic Light Scattering (DLS)

DLS enables analysis of protein size and shape in a process of tracking dynamic brownian fluctuations of protein molecules. Signal acquired during measurement arises from Rayleigh scattering of incidental laser beam on particles of analyte in solution. Each scattered wave has the same intensity but different phase (due to particle movement), which leads to interference of light, thus varying intensity on detector. The intensity of light is proportional to wavelength of incident light and sixth power of diameter of the particle.\[79\]

This very strong dependence on particle size is the reason why monodisperse samples are preferable for measurement - much lower scattering intensity of smaller particle in the mixture causes facile signal loss in presence of larger particles. The instrument measures intensities of scattered light in
different time intervals and correlates them with the intensity in \( t(0) \), autocorrelation function is then calculated from these values. From the shape of autocorrelation curve it is possible to deduct the size of scattering particle[80].

A major requirement in dynamic light scattering measurement is an absence of large sized contaminants (e.g. dust particles), thus one hour of centrifugation (at least \( 15000 \times g \)) is recommended. The quality of DLS output is dependent on protein concentration, therefore the concentration of at least 1 mg/ml is necessary for reliable results. DLS technique is valuable for protein dynamics studies, pre-crystallization control of sample purity and for protein interaction monitoring.

1.4.4 Intrinsic fluorescence of proteins

Fluorescence technique is a useful tool in spectroscopic structural studies. Proteins are unique biopolymers owing intrinsic fluorescence via aromatic amino acids – phenylalanine, tyrosine and tryptophan, where tryptophan has dominant fluorescence. Phenylalanine displays a structured emission with a maximum near 282 nm. The emission of tyrosine in water occurs at 303 nm and is relatively insensitive to solvent polarity. On the other hand tyrosine can undergo excited state ionization, resulting in the loss of the proton on the aromatic hydroxyl group, thus it is pH sensitive. The emission maximum of tryptophan in water occurs near 350 nm and is highly dependent upon polarity and local environment. The most valuable for fluorescence studies is amino acid tryptophan because of the sensitivity to its local surrounding. These changes often occur as results of conformational changes, subunit association, substrate binding or denaturation, as these interaction affect local indole environment. These changes in environment affects tryptophan emission spectra and also its excited state lifetime. Thus steady state fluorescence and kinetic fluorescence spectroscopy is often used for this purpose. Nevertheless complicating factor for the interpretation of experimental spectra could be presence of multiple aromatic amino acids in most protein[81].

Tryptophans buried with some polar residues nearby shows spectra with fluorescence maximum at 316-325 nm. Occurrence of more mobile polar residues, neighbouring tryptophan residue is exhibited by shift of fluorescence maximum up to 330-335 nm. The emission spectrum of tryptophan residue at protein surface in contact with bound water are characterized by fluorescence maximum at 340-345 nm. Totally unfolded proteins with neighboring freely relaxing water molecules have fluorescence spectra with maxima at 350-353 nm[70].
2. Aims of the thesis

1. Recombinant expression of 15 NBP proteins (corresponding to Group 1 NBPs from the bioinformatic screen) in E.coli

2. Analysis of protein solubility and renaturation/purification potential

3. Renaturation/purification of selected NBPs

4. Structural characterization of successfully purified NBPs by MS, DLS, GPC, CD, FTIR, fluorescence and NMR spectroscopy

5. Comparison of secondary structure content of the characterized NBPs with the bioinformatical secondary structure prediction
3. Materials and instruments

3.1 Materials

3.1.1 Chemicals

- 2-(N-morpholino)ethanesulfonic acid Fluka, Germany
- 2-mercaptoethanol Fluka, Germany
- Acetic acid Penta, Czech Republic
- Acrylamide Fluka, Germany
- Agarose Penta, Czech Republic
- Ammonium persulfate Sigma-Aldrich, Germany
- Arginine Roth, Germany
- BioSafe Coomassie stain Bio-Rad, USA
- Bis-Tris Roth, Germany
- Blocker Casein in PBS ThermoFisher Scientific, USA
- Dithiothreitol Sigma-Aldrich, Germany
- Ethanol Penta, Czech Republic
- Ethylenediaminetetraacetic acid Penta, Czech Republic
- Guanidinium hydrochloride Sigma-Aldrich, Germany
- Hydrochloric acid Penta, Czech Republic
- Imidazole Roth, Germany
- LB Broth Sigma-Aldrich, Germany
- Methanol Penta, Czech Republic
- Ni-NTA agarose Qiagen, Germany
- Saccharose Penta, Czech Republic
- Silver nitrate Penta, Czech Republic
• Sodium chloride Sigma-Aldrich, Germany
• Sodium dodecyl sulphate Fluka, Germany
• Sodium hydroxide Penta, Czech Republic
• Sodium phosphate dibasic dihydrate Sigma-Aldrich, Germany
• Sodium phosphate monobasic dihydrate Sigma-Aldrich, Germany
• SuperSignal West Femto ThermoFisher Scientific, USA
• TALON Clontech, USA
• Tetramethylethylenediamine Fluka, Germany
• Tris(2-carboxyethyl)phosphine Thermo Scientific, USA
• Tris-HCl Sigma-Aldrich, Germany
• Urea Penta, Czech Republic
• ethidium bromide Sigma-Aldrich, Germany
• magnesium chloride Penta, Czech Republic

3.1.2 Buffers

• Bis-Tris buffer: 1.5M Bis-Tris (pH 6.5-6.8 with HCl)

• Blotting buffer: 10 ml blotting concentrate, 1 ml 10% (w/v) SDS, water to the final volume of 90 ml and afterwards add 10 ml methanol

• Blotting concentrate: 72.1 g glycine, 15.1 g Tris, water to the final volume of 500 ml

• Buffer A: 25% (w/v) saccharose, 50mM Tris-HCl, 1mM EDTA, pH 7.4

• Buffer B: 0.5% (w/v) Triton X-100, 1mM 2-mercaptoethanol, 100mM NaCl, 50mM Tris-HCl, pH 7.4

• Buffer C: 50mM Tris-HCl, 100mM NaCl, 1mM 2-mercaptoethanol, pH 7.4

• DNA sample buffer (6×): 3 ml glycerol, 25 mg bromophenol blue, water to 10 ml
• **Denaturation buffer**: 50mM phosphate buffer, 300mM NaCl, 30mM 2-mercaptoethanol, 50mM arginine, 8M urea, pH 8

• **Elution buffer**: 20mM phosphate buffer, 100mM NaCl, 250mM imidazole, pH 8

• **MES running buffer**: 250 mM MES, 250 mM Tris, 5 mM EDTA, 0.5% (w/v) SDS

• **Ni-NTA Lysis buffer**: 20mM phosphate buffer, 100mM NaCl, 10mM imidazole, pH 8

• **Ni-NTA Wash buffer**: 20mM phosphate buffer, 100mM NaCl, 30mM imidazole, pH 8

• **PBS (10×)**: 1.37M NaCl, 27 mM KCl, 100 mM Na$_2$HPO$_4$, 18 mM KH$_2$PO$_4$

• **Renaturation buffer**: 50mM phosphate, 100mM NaCl, pH 8

• **SDS sample buffer 10 ml, 10×**: 2.0 ml 1M Tris-HCl pH 6.8, 0.8 g SDS, 4.0 ml 100% glycerol, 0.4 ml 14.7M 2-mercaptoethanol, 1.0 ml 0.5M EDTA, 8 mg bromophenol Blue

• **TALON Lysis buffer**: 20mM phosphate buffer, 100mM NaCl, pH 8

• **TALON Wash buffer**: 20mM phosphate buffer, 100mM NaCl, 10mM imidazole, pH 8

### 3.1.3 Enzymes

- **NdeI**
  - New England Biolabs, USA

- **XhoI**
  - New England Biolabs, USA

### 3.1.4 Competent cells

- **E. coli** BL21 DE3: F$^−$ ompT gal dcm lon hsdS$_B(r_B^− m_B^−)$ λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])

- **E. coli** MACH1: $\Delta$recA1398 endA1 tonA $\Phi$80$\Delta$lacM15 $\Delta$ lacX74 hsdR($r_K^− m_K^+$)
3.1.5 Other materials

3.2 Instruments

- CD spectrometer
  - Jasco 815, Japan
- DLS instrument
  - RiNA laser-spectroscatter 201, Germany
- FTIR spectrometer
  - Nicolet 6700, USA
- centrifuge
  - Eppendorf 5412 R, Germany
- centrifuge
  - Heraeus Biofuge Pico, Germany
- electrophoresis apparatus
  - Consort EV242, England
- fluorimeter
  - Jasco FP-6600, Japan
- incubator
  - Flow Laboratories IR 1500, USA
- mixing block
  - Bioer MB-102, China
- pH meter
  - XS Instruments, Italy
- shaker
  - Infors HT Multitron Standard, Switzerland
- sonicator
  - Bandelin Sonoplus UW 3100, Germany
- transilluminator
  - Quantum ST 4, Canada
4. Methods

4.1 DNA manipulation

4.1.1 Horizontal agarose gel electrophoresis

The gel was prepared by dissolving agarose powder in TAE buffer to final concentration of 1.2% (w/v). The solution was heated to facilitate proper dissolving. Hot agarose solution was cooled to approximately 45-50 °C, then 5 µl of ethidium bromide was added to 50 ml of mixture to final concentration of 0.5 µg/ml, gel solution was poured to gel tray with the well comb in place. Samples of nucleic acid were mixed with DNA sample buffer in prescribed ratio, solidified gel was placed to electrophoresis unit and samples were loaded into the gel wells. Following electrophoresis was performed at 120 V, until the dye line was approximately 70 % of the way down the gel. After that, the DNA bands were visualized by illuminating the gel with UV light.

4.1.2 Restriction analysis of DNA sample

Reaction mixture of 500 ng of plasmid DNA, 1 µl of restriction enzymes XhoI and NdeI, 3 µl of NEB 3 buffer, 0.3 µl of BSA solution was filled with deionized water to the final volume of 30 µl. Mixture was incubated at 37 °C for 2 h, after that 5 ul of solution was used for electrophoresis analysis, the rest was stored at -20 °C for further use.

4.1.3 Isolation of DNA from agarose gel

The DNA samples of interest were recovered from gel after agarose electrophoresis by excision and subsequent procedure using the Zymoclean Gel DNA Recovery Kit according to the manufacturer’s instructions.

Initially, 3 volumes of AD Buffer were added to each volume of agarose excised from the gel (100 ul of buffer to 100 ug of gel). That was followed by dissolving the gel slice (50 °C, 10 min). Melted agarose solution was transferred to a Zymo-Spin column in a collection tube and centrifuged for 60 seconds at 13000×g. The flow-through was discarded and 200 µl of DNA Wash Buffer was loaded to the column and centrifuged for 60 seconds at 13000×g. After that, the column was placed into a 1.5 ml tube, 6 µl of warm water (50 °C) was added directly on the column matrix and tube was centrifuged for a 2 minutes at at 13000×g.
4.1.4 Ligation of protein codin gene into the expression vector

The reaction mixture was prepared as follows:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>plasmid DNA</td>
<td>2 (0.020 pmol)</td>
</tr>
<tr>
<td>NBP DNA</td>
<td>2 (0.10 pmol)</td>
</tr>
<tr>
<td>T4 Ligase buffer</td>
<td>2</td>
</tr>
<tr>
<td>T4 DNA Ligase</td>
<td>1</td>
</tr>
<tr>
<td>water</td>
<td>13</td>
</tr>
</tbody>
</table>

The vector and insert were added to be in 1:6 molar ratio to maximise the yield of successfully ligated product. Reaction mixture was incubated for 12 hours at 16 °C.

4.1.5 Transformation of bacterial cells

Bacterial cells were thawed at 0 °C and held on ice with added plasmid solution (50-150 ng of DNA to 30-50 µl culture) for approximately 20 minutes. Next, cells were incubated for 90 seconds at 42 °C. Suspension was diluted tenfold with sterile LB medium without antibiotics and bacteria were cultivated for one hour at 37 °C. Finally, 150 µl of cell culture was spread on agar plate with kanamycin concentration of 50 µg/ml and incubated at 37 °C for approximately 16 h.

4.1.6 Isolation of the plasmid DNA from E. coli cells

Plasmid DNA was isolated from cells using the Zymo Research Zyppy Plasmid Miniprep Kit.

A cell colony from the agar plate was cultivated overnight in 15 ml of LB media with 50 µg/ml kanamycin at 37°C, 220 rpm. The culture was centrifuged for 15 minutes at 3500×g, the supernatant was discarded, and the pellet was resuspended in 600 µl of LB medium and 100 µl of 7X Lysis Buffer was added. Suspension was mixed by inverting the tube 10 times. When the solution became clear blue, 350 µl of cold Neutralization Buffer was added, suspension was mixed thoroughly and centrifuged at 13000×g after the complete neutralization. Supernatant was transferred into the provided Zymo-Spin column which was placed into a collection tube and centrifuged at 15000×g for 15 seconds. Flow-through was discarded, column placed back into the collection tube and 200 µl of Endo-Wash Buffer was added to the column. After 30 second centrifugation, 400 µl of Zyppy Wash buffer was
added to the column and centrifuged for 1 minute. Plasmid DNA was eluted with 30 \( \mu l \) of *Zyppy Elution Buffer*.

### 4.2 Protein expression and isolation

#### 4.2.1 Protein expression

A starter culture was established overnight from one colony from the agar plate in 5 ml of LB medium containing kanamycin (50 \( \mu g/ml \)) at 37\(^{\circ}\)C, 220 rpm. Large-scale protein production was carried out by adding 1 ml of starter culture to 500 ml of kanamycin-supplemented LB medium. Cells were grown typically at 37 \(^{\circ}\)C while shaking at 220 rpm for approximately 3 h (until the \( \text{OD}_{600} \) reached 0.6). Expression of the recombinant protein was induced by the addition of IPTG to a final concentration of 0.5 mM. After induction, cells were typically grown at 37 \(^{\circ}\)C for 5 h and then harvested by centrifugation (3000 \( \times \) g, 15 min, 21 \(^{\circ}\)C).

#### 4.2.2 Isolation of inclusion bodies

Cells harvested from 0.5 l culture were resuspended in 10 ml of *Buffer A* and subjected to four cycles of freezing (-80 \(^{\circ}\)C) and thawing (45 \(^{\circ}\)C). After that, 400 \( \mu l \) of 1M \( \text{MgCl}_2 \) was added and lysed culture was incubated with DNAse I (10 U/ml) for 30 minutes at 21 \(^{\circ}\)C. Next, 5 short (30 s, 55 % intensity, with 30 s breaks on ice) sonication cycles were used to lyse the cells. Soluble material was separated by centrifugation (10000 \( \times \) g, 15 min, 4 \(^{\circ}\)C) and the pellet was resuspended in 10 ml of *Buffer B*. After the 10 min incubation on ice, the soluble content was removed by centrifugation (10000 \( \times \) g, 15 min, 4 \(^{\circ}\)C) and the pellet was washed with the *Buffer C*. The pellet obtained after the final centrifugation (10000 \( \times \) g, 15 min, 4\(^{\circ}\)C) represented the inclusion bodies (IB) enriched fraction of cells and was stored at -20 \(^{\circ}\)C.

#### 4.2.3 Renaturation of expressed proteins

##### 4.2.3.1 Renaturation by dialysis

Pellet of IB has been dissolved in *Denaturation buffer* for 1 hour at 21\(^{\circ}\)C (0.1 g of IB in 600 ml of buffer). Solution was poured into dialysis tubes and dialysed against 5 l of *Renaturation buffer* for 1 hour at 4\(^{\circ}\)C. The precipitated content of the tubes was centrifuged for 30 min at 20000 \( \times \) g, and pure solution was poured back to the tubes. Dialysis continued for next 12 hours. Then the
solution was centrifuged again or filtered through 0.45 µm membrane and the dialysis tubes were moved to the bucket with fresh Renaturation buffer and the dialysis process continued for the next 24 hours. Final dialysis was performed against Ni-NTA Lysis buffer for the next 48 hours.

4.2.3.2 Renaturation by rapid dilution

Pellet of IB was dissolved in Denaturation buffer (20 ml per 0.1 g of IB). The solution was slowly drop-by-drop added into the 20 volumes of vigorously stirred Renaturation buffer at 4 °C. After that, solution was centrifuged, filtered through 0.22 µm and prepared for purification.

4.3 Protein purification

4.3.1 Metal – chelate affinity chromatography

4.3.1.1 Purification on Ni-NTA matrix gravity flow column

Ni-NTA slurry (2 ml) was poured into prepared column, matrix was washed with 8 ml of distilled water and equilibrated with 8 ml of Ni-NTA Lysis buffer. After the refolding procedure described in 4.2.3.1 or 4.2.3.2, diluted protein solution was loaded onto settled matrix beads in column, flow-through was collected for further SDS-PAGE analysis. Alternatively (when purifying a soluble protein), cells were sonicated (6 × 30s on ice, 55 % intensity) in Ni-NTA Lysis buffer, centrifuged for 15 min at 20000×g, 4°C and the supernatant was loaded onto column matrix. Next, matrix with bound protein was washed with 8 ml of Ni-NTA Wash Buffer and protein was eluted with 2-3 ml of Elution Buffer in 0.5 ml fractions.

4.3.1.2 Purification on TALON matrix gravity flow column

Alternatively to Ni-NTA, TALON affinity resin was used for protein purification. The procedure was the same as for the former technique, with exception of imidazole which was absent in Ni-NTA Lysis buffer and Ni-NTA Wash Buffer buffers.
4.4 Protein manipulation

4.4.1 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

In this work, Bis-Tris polyacrylamide gels were utilized for electrophoresis. The advantage to traditional Tris-Glycine electrophoresis is in use of slightly acidic environment, which suppresses unwanted cysteine reoxidation, but the most important benefit is better separation of low molecular weight proteins\cite{82}. The disadvantage is relatively higher operational expenses.

Resolving gel (12 \%) is prepared by mixing the following components:

\textbf{Table 4.2:} Composition of 12 \% polyacrylamide resolving gel

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bis-Tris buffer</td>
<td>1.75 ml</td>
</tr>
<tr>
<td>40 % acrylamide mix (19:1)</td>
<td>2.1 ml</td>
</tr>
<tr>
<td>distilled water</td>
<td>3.15 ml</td>
</tr>
<tr>
<td>10 % APS</td>
<td>70 ul</td>
</tr>
<tr>
<td>TEMED</td>
<td>7 ul</td>
</tr>
</tbody>
</table>

Gel mixture was poured between two glass plates and overlaid by 20\% ethanol. Gel was incubated for 30 min to polymerise. After that, ethanol solution was disposed and the polymerized separating gel was overlaid by stacking gel mixture composed of the following solutes:

\textbf{Table 4.3:} Composition of 5 \% polyacrylamide stacking gel

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bis-Tris buffer</td>
<td>0.9 ml</td>
</tr>
<tr>
<td>40 % acrylamide mix (19:1)</td>
<td>0.38 ml</td>
</tr>
<tr>
<td>distilled water</td>
<td>1.72 ml</td>
</tr>
<tr>
<td>10 % APS</td>
<td>30 (\mu)l</td>
</tr>
<tr>
<td>TEMED</td>
<td>30 (\mu)l</td>
</tr>
</tbody>
</table>

The well comb was then placed into the polymerizing stacking gel. After 30 minutes, well comb was removed and gel was overlaid with MES running buffer. Protein sample was prepared by mixing with the Sample Buffer in corresponding ratio and boiled for 5 minutes at 100\(^\circ\)C. After that, sample was centrifuged at 13000\(\times\)g for 5 minutes and loaded into the wells. Electrophoresis was performed under 180 V for 75 minutes.
4.4.2 Silver staining of proteins in polyacrylamide gel

Staining procedure consists of washing the polyacrylamide gel in several buffers, described in the table:

<table>
<thead>
<tr>
<th>Table 4.4: Stepwise procedure for silver staining of polyacrylamide gel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Solution</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

4.4.3 BioSafe Coomassie staining of polyacrylamide gels

The gel was washed in distilled water three times for 5 minutes to remove the remnants of SDS, which interferes with staining process. After that, gel was incubated with 20 ml of staining solution for half an hour, rinsed multiple times in water and incubated in water overnight.

4.4.4 Western blotting using chemiluminiscent substrates

SDS-PAGE gel with separated proteins and nitrocellulose blotting membrane were rinsed in Blotting buffer and were placed together with pre-wetted filter paper on the anode of the transfer apparatus in order - 2 wet paper sheets,
blotting membrane, polyacrylamide gel a 2 wet filter paper sheets. The transfer apparatus was then assembled, the blotting container was filled by Blotting buffer and whole apparatus was plugged to power supply. Proteins were transferred at voltage 100 V for 1 hour. Afterwards the membrane was blocked by Blocker solution for at least 2 hours. Protein were visualised by incubation of nitrocellulose membrane with monoclonal anti-polyhistidine peroxidase conjugate (10000x) diluted in blocker solution. The membrane was washed three times with PBS+0.05 % Tween and SuperSignal West Femto chemiluminiscent solution was applied.

4.5 Prepararion of selected NBPs

4.5.1 4090

Protein 4090 was studied both in its soluble and refolded form.

**Soluble** The production-ready cells were prepared according to 4.2.1. After reaching the OD\textsubscript{600} 0.6, culture was placed to another shaker, tempered to 20 °C, culture cooled down and was induced by IPTG at final concentration of 0.5mM. Protein production lasted 20 h, cells were harvested and lysed by sonication in Lysis Buffer. The cell lysate was centrifuged at 18000×g and protein was purified from supernatant according to 4.3.1.1.

**Refolded** Protein in form of inclusion bodies was prepared by procedure 4.2.1 in Methods. The inclusion bodies were isolated by 4.2.2 and protein was refolded (denaturation performed in 250 ml of Denaturation Buffer per 0.1 g of inclusion bodies) and purified by 4.2.3.1 and 4.3.1.1.

4.5.2 2298

Protein 2298 was also studied in both soluble and refolded form. Unlike 4090, protein 2298 does not require lowering of production temperature. Cells were treated according to 4.2.1. Culture was lysed, suspension centrifuged at 18000×g and soluble 2298 in supernatant purified by 4.3.1.2. Pellet was subjected to inclusion body isolation according to 4.2.2 with omitted sonication step. Isolated product was refolded and protein purified by 4.2.3.1(600 ml of Denaturation buffer per 0.1 g of inclusion bodies) and 4.3.1.2.

4.5.3 6387

Protein 6387 was obtained by refolding only. Protein was produced by 4.2.1. Inclusion bodies were refolded by 4.2.3.1(800 ml of Denaturing Buffer per 0.1
g of inclusion bodies) and purified by 4.3.1.1.

4.5.4 4712

Protein 4712 was prepared, isolated and refolded by standard protocols 4.2.1, 4.2.2 and 4.2.3.1 (600 ml of Denaturing Buffer per 0.1 g of inclusion bodies). Dialysate was purified according 4.3.1.2.

4.5.5 6620

Protein 6620 was unable to solubilise by refolding via dialysis. Rapid dilution protocol 4.2.3.2 was used for this protein. The remaining steps of expression and purification were kept standard.

4.6 Preparation of protein samples for different methods of structural analysis

4.6.1 CD/FTIR/fluorescence spectroscopy

After the purification procedure, protein was dialysed overnight against 200× sample volume of 10 mM Tris, 10 mM NaCl, 1 mM BME, pH 8.

For ECD spectroscopy protein concentrations ranged between 0.01 and 0.3 mg/ml for UV spectral region (195-260 nm) and 0.5 and 3 mg/ml for near-UV spectral region (260-350 nm). The experimental set up for both spectral regions was as followed: 1 mm quartz cell, 0.1 nm data pitch, scanning speed 10 nm/min, response time 8 sec, number of scans 2. For quantitative analysis of the CD measurements, the concentration of protein sample was determined by the amino acid analysis (performed by Dr. Radko Souček at IOCB AS CR, v.v.i.). The concentration was calculated using the 4-5 most abundant amino acids in particular sample.

Fluorescence spectroscopy was performed with the sample in the same concentration as ECD in UV spectral region in the quartz fluorescence cell with following experimental set up: spectral region 295-400 nm with excitation wavelength 280 nm, scanning speed 50 nm/min, response time 2 sec, excitation bandwidth 2 nm, emission bandwidth 5 nm). Final fluorescence spectra were normalized.

FTIR spectra were collected in ATR mode using ATR accessory (MIRacle Single Reflection ATR) equipped with diamond crystal and in transmission mode using 6 mm CaF$_2$ home-made cell. The sample concentration was about
3 mg/ml. Spectral contribution of water was eliminated using standard algorithm (Dousseau). Subsequently, the spectrum of water vapors was subtracted and the baseline was corrected using a linear function. Final IR spectra were normalized for constant integrated amide I intensities. CD, fluorescence and FTIR spectra was collected by Dr. Lucie Bednárová at IOCB AS CR, v.v.i.

4.6.2 NMR spectroscopy

Samples for NMR analysis were dialysed after the purification procedure against 200 $\times$ excess of 20mM phosphate buffer, 50mM NaCl, 10mM BME, pH 8 for 12h at 4°C. Protein concentration for the 1D and 2D measurement was in order of a several mg per milliliter. NMR spectra was collected by Dr. Vaclav Veverka at IOCB AS CR, v.v.i.

4.6.3 DLS

Samples for DLS measurements were in the same buffer as for CD (10mM Tris, 10mM NaCl, 1mM BME). Prior to the measurement, the sample was centrifuged at 18000 $\times$ g for at least 30 min to dispose the sample of dust particles interfering with the measurement. The optimum protein concentration for scattering experiment was in range of 1 to 3 mg per milliliter.
5. Results

5.1 Selection of studied NBPs

Prior to this experimental project, a library of random sequences (size of $10^4$) was generated \textit{in silico} by our coworkers at the Bioinformatics group of Jiří Vondrášek at IOCB AS CR, v.v.i.. The generated peptides are 100 residues long and the amino acid composition reflects the amino acid ratios of biological proteins. As this project explores the occurrence of secondary structure elements in random protein sequences, the generated sequences were analysed using five different secondary structure prediction algorithms (GOR4, Jnet, Predator, Simpa, and Psipred) and four disorder predictors (Disopred, DisEmbl, VSL2 and IUpred). According to the predictions, the occurrence of secondary structure in random sequence libraries is surprisingly similar to occurrence of secondary structure in natural proteins. To verify and further analyse these results, $3 \times 15$ candidate NBPs were selected from the library for scarce experimental sampling. The three experimental groups were specified in the following way: GROUP 1: (i) high occurrence of predicted secondary structure (samples from both $\alpha$-helical and $\beta$-sheet rich) and low disorder, (ii) high predicted solubility (using solubility predictors CVsol and Gravy) GROUP 2: random selection GROUP 3: (i) low occurrence of predicted secondary structure and high disorder, (ii) high predicted solubility (see Fig. 5.1). This thesis is devoted to group 1 NBPs only (see Appendix for their label and sequences), while the other two groups are being characterized in parallel.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure51.png}
\caption{Distribution of random sequences on predicted disorder/secondary structure content plot with the illustrative localisation of sequences from groups 1, 2 and 3}
\end{figure}
5.2 Cloning of NBPs

The genes coding for individual NBPs were synthesized commercially (using codons optimized for E. coli expression) as PCR fragments, including NdeI and XhoI restriction sites for the cloning procedure. The supplied DNA fragments (300-1000 ng) were dissolved in distilled water and cleaved by NdeI and XhoI restriction endonucleases. The same procedure was applied to plasmid pET24a to eventually produce NBPs with a C-terminal His-tag. All DNA samples were loaded on agarose gel and the cleaved sequences were purified according to 4.1.3. Each gene for NBP was ligated into pET24a plasmid, yielding NBP-pET24a vector. The results of cloning procedure were verified by DNA sequencing and restriction analysis (Fig. 5.2). The E. coli MACH1 competent bacterial strain was then transformed by DNA in ligation mixture and plasmids were purified using minipreparation protocol. 14/15 NBPs were cloned correctly into pET24a and prepared for protein expression experiments.

5.3 Expression and solubility of NBPs

Individual NBPs were expressed in E.coli BL21 DE3 strain after bacterial transformation with plasmids carrying individual NBPs. The protein overexpression was examined by SDS-PAGE electrophoresis by comparing the cell sample before induction by IPTG and after the cultivation with inducer. 12/14 tested NBPs showed some level of protein expression in E.coli, ranging from very mild to very strong (Fig. 5.3). Some of the proteins seemed to express in two different forms, as reflected by the presence of double-bands. As these
could be of different origins, further characterization was performed to explore the nature of the double-bands (see below). Moreover, the mobility of some overexpressed proteins suggested a possibility of higher oligomeric state. Disulphide formation was studied by electrophoresis in nonreducing SDS-PAGE sample buffer. Although the disulphide formation was not observed after direct processing of cell pellet containing some NBPs, its existence was confirmed later by 2-mercaptoethanol (BME) titration (from 0.08 mM to 4 mM BME) followed by nonreducing SDS-PAGE, as seen on the example analysis of refolded 4090 protein (Fig. 5.3D).

Figure 5.3: SDS-PAGE analysis of production, redox characteristics of all NBPs (panels A-C, theoretical MW is 12-13 kDa) and disulphide bond formation assay of refolded 4090 in dependence of decreasing mercaptoethanol concentration (mM)(panel D). -- Uninduced bacterial culture, N - culture after production in nonreducing sample buffer, R - culture after production in reducing sample buffer.

5.4 Western blot analysis of protein products

In vivo solubility of expressed products was assayed by SDS-PAGE electrophoresis and western blotting of separated cell lysate (soluble and insoluble fractions). The blotted membranes were incubated with AntiHis HRP conjugate antibody and NBPs were visualised by detection of luminol chemiluminescence on CCD camera. All bands on corresponding polyacrylamide gel were
excised and subjected to sequence identification by mass spectrometry at the mass spectrometry laboratory at IOCB AS CR, v.v.i.. All protein sequences (even both of the double bands in the samples where they were present) were confirmed to be correct and to correspond to the uncleaved individual tested NBPs. (Appendix 6.1). As observed on the western blot analysis of NBP.

Figure 5.4: Western blot analysis of crude cell lysates after NPB production. The protein products, in some cases visible as double-bands, are visible in the range of 10-20 kDa. I - insoluble fraction of lysate, S - soluble fraction of bacterial cell lysate (centrifuged supernatant after sonication of cell suspension)

solubility (Fig. 5.4), the majority of NBPs is insoluble and expressed into inclusion bodies. Only 5 proteins also show some degree of solubility, namely 7144, 4090, 2298, 1687 and 6387.

5.5 Isolation, refolding and purification of selected NBPs

Only peptides 2298 and 4090 expressed in soluble forms in sufficient quantities and stabilities, so that they could be purified after cell lysis without the need for refolding. Isolation of inclusion bodies and subsequent refolding was attempted for all the 12 NBPs that were expressed. The isolation of inclusion bodies is described in 4.2.2. The yield of inclusion bodies was typically about 50-100 mg per 0.5 l of bacterial culture.
Isolated inclusion bodies were refolded by different approaches. Dilution in excess of 8M urea and slow dialysis showed as the most successful strategy. Denaturation in GdmCl or small volumes of denaturant and rapid dilution in refolding buffer were also examined, but without convincing results. For some NBPs, none of the mentioned methods was successful as they were very prone to aggregation or had other stability issues. As can be tracked in Table 6.1, only 7/13 NBPs (including 4090 and 2298 that were also purified from the soluble cell lysis fraction) was successfully refolded.

All soluble and refolded proteins were further purified by either on gravity columns with NiNTA or TALON matrix or FPLC column with imidazole gradient elution. The purity of both was roughly similar, with the best results with TALON matrix. The purification protocol sometimes required minor adjustments based on individual NBP behavior. The purification analysis is only shown for 5 NBPs below (2298, 6620, 4090, 6387 and 4712) because 7144 and 1687 were not successfully purified into sufficient purity and stability for further measurements, respectively. 2298 and 4090 were purified both from soluble and insoluble cell fractions for later comparison of secondary structure content.

Figure 5.5: SDS-PAGE analysis of NBP expression, refolding and purification procedure. 1 - 2298, 2 - 6620, 3 - 4090, 4 - 6387, 5 - 4712. – - uninduced culture, + - culture after production, A, B, C - Buffers A, B and C used in inclusion body isolation, p and s subscripts - insoluble and soluble fractions of suspensions in buffers A, B and C, D - contents of dialysis tubes (soluble and insoluble), L - sample loaded onto Ni-NTA purification column, FT - sample that was not bound to Ni-NTA column, W - contents of Ni-NTA wash solution, E - elution fraction
5.6 Characterisation of NBPs mass and dimension

All structurally analyzed samples were subjected to routine preliminary characterization by MALDI MS (except for 6620), DLS and analytical GPC for analysis of protein molecular weight, monodispersity of sample and oligomerization state. MS data were collected at the mass spectrometry laboratory at IOCB AS CR, v.v.i. MALDI spectra confirmed the protein identity most clearly for 2298, 4090 and 6387 (see Fig. 5.6). However, peptides 6387 (in part) and 4712 showed the presence of lower molecular weight fragments in their spectra while no impurities were observed using the SDS-PAGE analysis. This phenomenon is probably a result of partial polypeptide cleavage and will be further described in discussion.

Figure 5.6: MALDI MS spectra of four purified NBPs. A - 4090, expected MW 12785.1 Da; B - 2298, MW 12580.8 Da; C - 6387, MW 13035.6 Da; D - 4712, MW 12786.2 Da

DLS measurements showed a relative monodispersity (with an exception of 6620) and the approximate molecular size ranging from 4 to 30 nm, with the mean around 10 nm (Fig. 5.7). A partial peak around 4 nm was observed significantly only for 2298 and 4090, especially when characterized for the soluble proteins (Fig. 5.8). If our 12-13 kDa proteins were monomeric and globular, their expected size would be around 2 nm, reaching up to 3 nm if
they resembled a random coil. The DLS experiment therefore suggests some higher-order structural arrangement of the analysed NBPs. This was further supported by gel filtration of selected NBPs, which after calibration analysis suggested their molecular weight in the range 70-100 kDa (data not shown). This will be later discussed in more detail.

**Figure 5.7:** DLS analysis of five refolded NBPs. The approximated diameter of NBPs ranges from 4 to 30 nm, with the mean around 10 nm

**Figure 5.8:** DLS analysis of soluble 2298 and 4090. Although mixture is not monodisperse, the proportion of more compact form of NBP is significantly higher comparing to refolded sample.
5.7 CD/IR spectroscopy of chosen NBPs

All the successfully purified NBPs were promptly analysed for secondary structure occurrence using spectral measurements. These were all performed by Dr. Lucie Bednárová at IOCB AS CR v.v.i..

The resulting spectra suggest occurrence of secondary structure in all proteins with the exception of 6387, whose far UV CD spectra corresponds to random coil conformation. Protein 2298 exhibits a dichroism pattern corresponding to distinct tertiary structure. No particular tertiary structure was reported for proteins 4090 and 6387 and it was absent for protein 4712. The occurrence of tertiary structure and unfolding properties were also assayed by measuring protein tryptophan fluorescence in native and denaturating conditions. All five NBPs showed similar behavior (maximum at 340-345 nm with negligible (3-4 nm) red shifts in denaturing conditions). All samples, even the natively unfolded 6387, presented strong tendency to form secondary structure when exposed to 30 % TFE. The results are summarized together with the secondary structure predictions in Table 5.1 (except for the disordered 6387).

![Figure 5.9](image)

**Figure 5.9:** CD spectral analysis of five NBPs. A - far UV CD spectra of NBPs, B - near UV CD spectra of NBPs. Both measurements were performed in 10mM Tris, 10mM NaCl, 1mM BME, pH 8

<table>
<thead>
<tr>
<th>NBP</th>
<th>Predicted secondary structure content</th>
<th>CD data (CONTILL, analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α helix (%)</td>
<td>β sheet (%)</td>
</tr>
<tr>
<td>2298</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soluble</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>Refolded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soluble</td>
<td>15</td>
<td>28</td>
</tr>
<tr>
<td>Refolded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6387</td>
<td>49</td>
<td>6213</td>
</tr>
<tr>
<td>4712</td>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td>6620</td>
<td>54</td>
<td>6</td>
</tr>
</tbody>
</table>
The only soluble proteins - 2298 and 4090, were compared to their refolded variants. Though the typical oligomerization characteristics assayed by DLS and GPC remained the same, they showed differences in CD spectra and thus in the secondary structure content, especially for protein 2298 (Table 5.1). Soluble 2298 shifted towards more α helical structure when compared with the refolded 2298 and the overall secondary structure content was higher (63 vs 44 %). For 4090, the ratio of secondary structure motifs are comparable as well as the overall secondary structure content.

Figure 5.10: CD spectral analysis of native and refolded forms of proteins 4090 and 2298. A - far UV CD spectra of soluble and refolded 2298, B - far UV CD spectra of soluble and refolded 4090

The unfolding behaviour of proteins was studied by CD spectroscopy with and without denaturation agent - 0.4M GdmCl (and/or higher concentrations in some cases). Spectra shown below represent two different responses to denaturing conditions - subtle contrary to well visible change in the CD spectrum for 6387 and 2298 proteins, respectively. This supports the disordered/random nature of 6387. The rest of the NBPs in presence of GdmHCl showed behaviour resembling to 2298. Higher GdmHCl concentration was not standardly used because of its high absorption, nevertheless, to support

Figure 5.11: CD spectral analysis of proteins 6387 (A) and 2298 (B) in native and denatured form (induced by 0.4M GuHCl)
disordered/random nature of 6387 protein the higher GdmHCl concentration or temperature dependence measurement should be complemented.

As a complementary technique to CD for secondary structure quantification IR spectroscopy was used for both soluble 2298 and 4090 samples as these could be prepared in a sufficient amount and concentration for this technique. From the spectra shown below (Fig. 5.12) it seems that besides clear alpha-helical minima both 2298 and 4090 have some beta sheet content. The implications of this data and its comparison to the CD results will be discussed later on.

Figure 5.12: IR spectral analysis of 4090 and 2298 proteins. A - amide I and II regions of protein spectra, B - zoomed view of amide I spectral region

5.8 1D/2D NMR characterization of soluble proteins

Proteins 2298 and 4090 (both soluble and after refolding) were subjected to 1D NMR analysis to inspect if they show any signs of hydrophobic core formation. This property can be recognized by peak dispersion (thus diverse chemical environment) in chemical shift region from 6 to 10 (amide hydrogens) and 0 to 0.5 (hydrogens of methyl groups typically localized in hydrophobic core of folded proteins). Figures 5.13 and 5.14 show spectra of soluble versions of both proteins and suggest that both have a stable hydrophobic core.
Figure 5.13: 1D NMR spectrum of soluble 2298 (top) with zoomed views on amide hydrogens peaks (bottom left) and methyl hydrogens (bottom right). Well observable peak dispersion of these spectral regions indicates the presence of hydrophobic core. The small peak on 10 ppm suggests mobile tryptophan residues.

Figure 5.14: 1D NMR spectrum of soluble 4090 (top) with zoomed views on amide hydrogens peaks (bottom left) and methyl hydrogens (bottom right). Well observable peak dispersion of these spectral regions indicates the presence of hydrophobic core.
The spectra of soluble and refolded forms were compared and overlaid for both proteins. (Fig 5.15) Compared with the soluble proteins, the spectra of refolded proteins show low peak dispersion suggesting the absence of tertiary structure.

Figure 5.15: 1D NMR spectra overlay for soluble (blue) and refolded (red) versions of proteins 2298 (A) and 4090 (B). Overall greater peak dispersion, especially in regions mentioned on Fig. 5.13 and 5.14 is observable in soluble forms.
Because protein 4090 was prepared in a sufficient concentration for more detailed measurements, it was additionally subjected to 2D H-H NOESY characterization. The high count of shuffled peaks indicate larger oligomeric structure with distinct β sheet conformation recognizable from characteristic peak island between 5 and 6 ppm on F2 axis (amide hydrogens).

**Figure 5.16:** 2D H-H NOESY NMR spectrum of soluble 4090. High number of shuffled peaks indicates oligomeric nature of protein. Characteristic peak island between 5 and 6 ppm on F2 axis indicates presence of β sheet structure.
6. Discussion

The goal of this thesis was to prepare and analyze a group of 15 NBPs that were selected from a library of random polypeptide sequences generated in silico preceding this study. The selection was based on a frequent occurrence of secondary structure as predicted by several secondary structure prediction algorithms. The successfully expressed and purified peptides were to be analyzed by several biophysical methods for the secondary structure characteristics.

(i) Analysis of NBP expression and solubility

Synthetic genes of fourteen of the NBPs were successfully cloned and expressed in E.coli with a C-terminal histidine tag. Twelve of these showed some expression level in E.coli while the expression level of individual NBPs varied significantly. For example, it could be hypothesized that the absence of protein 1441 expression could be caused by proteolysis prone primary sequence characteristics - presence of Leu, Phe, Asp, Lys and Arg at the N-terminal end of protein[83]. As could be expected, solubility and expression levels differed depending on the temperature and production volumes. For example, five of thirteen NBPs (2298, 6387, 7144, 4090 and 1687) showed to be soluble in bacterial cytoplasm at 37°C in analytical expression trials. When the expression volume was upscaled at the same temperature, only protein 2298 was soluble and 4090 achieved solubility with lower expression temperature. Therefore to achieve the best solubility/expression level for preparative purposes, the protocol had to be mildly optimized (varying temperature, inducer concentration or transcription promoter) for each NBP and for some, this is still work in progress and at this stage is beyond the scope of this thesis.

Interestingly, a lot of the expressed NBPs appeared as double bands on reducing SDS-PAGE gels (with only a few kDa apparent mass difference). Mass spectrometry confirmed that both bands belonged to the same non-degraded and uncleaved NBPs. This unusual behaviour probably corresponds to two different peptide conformations that are denaturant and temperature resistant and it resembles earlier reports for synthetic polypeptides[14]. We have noticed that the appearance of double bands on SDS-PAGE was more pronounced for proteins that were expressed most in inclusion bodies while it is less significant for the more soluble peptides. It is of interest, that some of the most soluble NBPs (e.g. 6387, 4090, 2298) were among the ones with highest bioinformatically predicted solubility. Moreover, when protein 4090 was purified from both insoluble (after refolding) and soluble fraction, the dou-
ble band was only present for the refolded version while the soluble version appeared as a single band. In addition to double conformations, most NBPs contained at least one cysteine residue in the sequence which had a tendency to cause intermolecular disulfide bridges and so the proteins were expressed and analysed in a reducing environment. Similar to above, the disulphide formation behaviour is often different in insoluble and soluble fractions of produced NBPs (for 2298 and 4090) suggesting different protein conformation in both fractions.

(ii) NBP structural characterization

The main goal of this study was not to evolve an optimized expression, refolding and purification scheme for each NBP but rather to use the most successful protocols to choose the best candidate NBPs from this group for further characterization. Using this strategy, 5 NBPs (2298, 4090, 6620, 6387 and 4712) were successfully refolded and purified from this group, two of which were purified from soluble fractions in parallel (2298 and 4090). SDS-PAGE confirmed the purity of all five NBPs while MALDI MS was only measured for four of them and while conclusive for 2298, 4090 and 6387, in case of 4712 it reported smaller fragments that as hypothesized could belong to a fragmented 4712 protein as a result of Xxx-Pro bond fragmentation\[84\]. The first level of characterization was to evaluate the oligomerization and monodispersity. Using DLS and GPC, the average particle radius of 10 nm and molecular weight of 70-100 kDa were observed for all five candidates, consistently suggesting a higher-order structural organization (of 5-8 monomeric subunits) for all purified proteins. Only for 2298 and 4090 soluble proteins also presented a peak corresponding to particle radius of approximately 4 nm (corresponding to 2 monomeric subunits). The higher-order structural arrangements could have different attributes and reasons for each NBP. First, it could suggest some level of aggregation processes. However, when initial purification protocols were tested, some protocols produced samples that were very prone to immense aggregation as observed using DLS, reaching up to 100 nm (data not shown in the thesis). The most successful purification protocols (that were later used as part of this thesis) contained several chemical agents to prevent aggregation processes and at least apparently blocked major aggregation issues \[4.2.3.1\]. An alternative explanation could be an occurrence of a systematic error in the analyses. This does not seem probable because during the same measurements with Group 2 NBPs (not the subject of this thesis) proteins with approximately correct molecular weight (based on GPC) were discovered. The difference between Group 1 and 2 NBPs is in the selection based
on high predicted secondary structure occurrence for Group 1 while Group 2 is a completely random selection from the library. For example, clusters of hydrophobic residues in primary structure can score negatively during secondary structure prediction while they may actually be advantageous during hydrophobic core formation and possibly form less aggregation prone structures. This is however purely speculative and hopefully will become clearer with the progress of the larger study. Third, protein oligomerization could be present and serve for burial of hydrophobic surface area for polypeptides to fold into a compact structure and stabilize as learnt from the emerging field of de novo protein design. To better resolve the characteristics of Group 1 NBPs, we plan to conduct some analytical ultracentrifuge and NMR experiments.

The next level of characterization was to explore the secondary structure content. This was done by ECD and fluorescence spectroscopy for all five selected NBPs (2298, 4090, 6620, 6387 and 4712) and in addition using FTIR and 1D/2D NMR spectroscopy for the NBPs produced from soluble fractions as they could be prepared in sufficient quantities, concentration and stability (2298 and 4090). In general, the CD experiments showed the presence of all kinds of secondary structure - helices, sheets as well as presumably disordered (6387). In all cases, the spectra was also collected in the presence of GdmHCl and TFE. Except for 6387 (where the denaturant had close to no effect on the already disordered/random structure) GdmHCl showed the loss of secondary structure in the spectra as expected. On the other hand, TFE (used to study peptide structure-forming potential in CD) strongly induced secondary structure formation, mostly alpha helical, for all five NBPs, including the presumably disordered 6387. This kind of behavior reminds the conditional folding of IDPs. The fluorescence measurements were all very similar for all five NBPs and with the maximum at 340-345 and no significant shift with 0.4M GdmHCl they resembled tryptophan residues at protein surface in contact with bound water - i.e. not buried in tertiary structure and not completely unfolded. This result was supplemented by (i) the near UV CD data which at least for protein 2298 suggested some level of tertiary structure arrangement and (ii) 1D NMR which suggested tertiary structure characteristics for both 2298 and 4090 that were expressed as soluble proteins. In addition, FTIR absorption spectroscopy suggested the presence of both helical and sheet content for both 2298 and 4090 that expressed as soluble proteins. An interesting comparison emerged for these proteins prepared from soluble fractions and by refolding, mainly for 2298 which is more helical and has more secondary structure content in total when soluble (according to CD and 1D
NMR data). The difference in refolded and soluble 2298 protein conformation is supported by the observation that unlike refolded 2298, the soluble version exhibited very weak binding to TALON column during purification, perhaps due to partially buried/inaccessible C terminus. 1D NMR spectra comparison for soluble versus refolded 2298 and 4090 proteins supports this behavior showing that the soluble proteins present signs of hydrophobic cores unlike their refolded versions and that is likely the source of difference in their secondary structure content. Shifts of secondary structure propensities have been described before for polypeptides when a more dominant driving force provides a hydrophobic surface area burial solution[85]. One of the goals of this study was to compare the secondary structure characteristics of the selected NBPs with the bioinformatical predictions. With 5 characterized NBPs so far, that is a very scarce sampling. Nevertheless, the outcomes of CD measurements suggest that the overall secondary structure content of the measured NBPs are approximately similar to the predicted values (around 50 % overall content) except for 6387 that exhibits behavior similar to IDPs. However, even 6387 shows some structure-forming propensity as described above. In more detail, the CD data does not correlate well in terms of helical versus sheet content and reports more alpha helical structures when compared with the predictions (Table 5.1). This could result from a phenomenon of CD being more sensitive in distinguishing helix conformations that has been reported before, especially when spectra are not measured down to at least 190 nm, preferably below[70]. Our spectra was collected down to 197nm maximum as our choices of buffers (including reducing agents) did not allow for further wavelength exploration. For two NBPs we could supplement the CD data with FTIR measurement as those are generally better for beta sheet measurement. Interestingly, protein 2298 exhibited some beta-sheet characteristics in the spectrum and yet had no beta sheet structures assigned from the CD analysis. Nevertheless, even that is not a straightforward finding as the sample for FTIR measurement has to be more concentrated and is at a higher risk of aggregation and so the results are not directly comparable. Nevertheless, preliminary 1D/2D NMR spectroscopy inspection of 2298 and 4090 soluble proteins also reported occurrence of beta sheet structures as well as a presence of hydrophobic cores and higher oligomeric arrangement. Taken together, the secondary structure content as analysed by CD and FTIR should be interpreted with caution as each of these method has its own limits and advantages. As much as bioinformatic prediction algorithms have their levels of error, so do these biophysical methods and all that will have to be taken into account when comparing the results as the study proceeds and the
experimental screen becomes less scarce.

So far, only the NBPs that likely have more propensity to form secondary structures (based on the bioinformatical selection) have been scarcely explored experimentally. They all have non-existing homology to any biological proteins and yet secondary structure content was observed for most of them. This finding corresponds to a few previous studies that suggested that the occurrence of secondary structure within random sequence space is not as rare as originally assumed. At least two out of five characterized NBPs also seem to have a tertiary structure and a hydrophobic core which is in accordance with recent studies that found that folds can frequently emerge even in unevolved sequences[18]. None of the previous studies however made a more detailed 3D structural characterization of structured NBPs and it remains unclear what mechanisms are employed in tertiary structure formation of unevolved proteins. Our preliminary NMR spectroscopy data suggest that protein oligomerization may help in assisting burial of hydrophobic surface area. It also remains unclear if the unevolved tertiary structure space hides new protein fold variants, not selected by Nature.
Conclusions

- Fourteen out of fifteen NBP synthesized genes were successfully cloned and twelve of them presented some level of expression in *E.coli*; five of those were partially soluble in the bacterial lysate.

- Five NBPs (2298, 4090, 6387, 4712 and 6620) were successfully purified after renaturation from *E.coli* and two of them were also purified as soluble proteins (2298 and 4090).

- Structural characterization confirmed approximately 50% secondary structure content for most of them (excluding 6387 which resembled a disordered protein in most characteristics) which was in agreement with the bioinformatic secondary structure predictions; Individual secondary structure analyses (CD, FTIR and NMR) varied slightly in the results - mainly CD seemed to over-emphasize alpha-helical content when compared with FTIR and NMR spectroscopy data. CD spectra analysis reported more alpha-helical content when compared to bioinformatic secondary structure predictions.

- Two out of five purified NBPs that were purified as soluble proteins (2298 and 4090) presented characteristics of a stable tertiary structure and a hydrophobic core as deduced from near UV CD and 1D NMR spectra while a higher-order oligomeric arrangement was observed by DLS and GPC.
Bibliography


List of Abbreviations

- ATP: adenosine triphosphate
- BME: 2-mercaptoethanol
- CH: charge–hydropathy
- D: Debye, a non-SI unit of electric dipole moment
- DLS: Dynamic Light Scattering
- DTT: dithithreitol
- ECD: Electronic Circular Dichroism
- FPLC: Fast Protein Liquid Chromatography
- FRET: Förster Resonance Energy Transfer
- FTIR: Fourier Transformation Infrared Spectroscopy
- Fig.: Figure
- GPC: Gel Permeation Chromatography
- GdmCl: Guanidinium Hydrochloride
- HEPES: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
- HRP: Horseradish Peroxidase
- IDP: Intrinsically Disordered Protein
- IPTG: Isopropyl-β-D-galactopyranoside
- MALDI: Matrix Assisted Laser Desorption Ionisation
- MES: 2-(N-morpholino)ethanesulfonic acid
- MS: Mass Spectrometry
- MW: molecular weight
- MWCO: molecular weight cut-off
- NBP: Never Born Protein
- NMR: Nuclear Magnetic Resonance
• NOESY  Nuclear Overhauser Effect Spectroscopy
• Ni-NTA  Nickel-Nitroloacetic acid
• PAGE  polyacrylamide gel electrophoresis
• PCR  Polymerase Chain Reaction
• PIPES  piperazine-N,N-bis(2-ethanesulfonic acid
• RNA  Ribonucleic Acid
• ROA  Raman Optical Activity
• SDS  sodium dodecyl sulphate
• SRCD  synchrotron radiation circular dichroism
• SS  Secondary Structure
• TCEP  tris(2-carboxyethyl)phosphine
• TFE  2,2,2-trifluorethanol
• VCD  Vibrational Circular Dichroism
• VOA  Vibrational Optical Activity
• vs  versus
Appendix

6.1 Sequences of NBPs

>8836|Never Born Protein
MGAGLAQFIDPLYRMAVDLPTIDTWGINRKAEEAFLERTYSANKYDLKALKH
LTGKQYEDVICAHKRLDYAITRDEYLYLGAANDIYNFRGLVSARKAELH

>3027|Never Born Protein
MMMFSALKYVSAYAHYFKRIDDHDSEFLKSAATRTGIREDRALLKVGAAVLEY
SLSITEVAGLRVQYRTEKDPWTDLLNLQAMRTSYNHAYVPQELKVLEHH

>1856|Never Born Protein
MYQIEKADFTDVRRTAATDIEHAFNVMWQLQSWCDVSIIKRTLDAYDEAYDAA
FRLKPAEWAIDDDVASIQRHRHYVAYNLSKIKLVPRLKLSGTLEH

>1441|Never Born Protein
MKKKAGKQPGDECNFRWMFTNTMNVMYRNLMTAGFYGLEAAREYDLYKNRMEE
LERYALSTQRLFHQLMNFWEFIKRIGCLDTYLRDTRILDYETIVYLEHH

>6387|Never Born Protein
MIEHCYSKTVYNLEQQEYDLEVTHIEGWMRAGKDLADNLLEDGSHVFPIEVAL
QENHYREVHAKIGDAERMVYKRELFPFQPEVLETPSQLFFAIEIEHH

>1687|Never Born Protein
MRQYFDLKMVCHELGLHLQARQGMGETYDELNQRWREADVYLMDNGSLLGD
FSHKKYILARYANIRTAEMVGGSTFIAQFAXVEFGDHFSLRSALEHH

>8635|Never Born Protein
MPFLLFQGTLLTIEKVAILKPTLMFHATLLEIKGLLRNPLPHNRATAWQFLYSEN
LLNASRLQDTDEFFKYSKVESSGLYQERVDRKATNRDFQDKLNAYHELEHH

>6620|Never Born Protein
MFITIQEDQVGERIFFPMLIRKEAGNLTDYLLAYWSRSIEGINEDWNFGEI
NQILERGRDRERSILSNNKVAQLIYAIMWLRIAELAPAQISKLKDLEHH

>8600|Never Born Protein
MKVDMEGLNPLHKGMSQVFFADLVAYKFRGITRRAKYYAKGMALYGIKNHFA
EENLCYEIKGLSHKYEALNSKKRLYLLSKAIWARNMIKKDLDQPLEHHHHHH

>4712|Never Born Protein
MFTNRFDYTLTDVTKRVFETYITYLERVPGFRTMAYTDVEYDHGADHPVGGRQG
VDSPVPQEGSMWNCRNLYLVYVQGNTLVQIEGYFVIAKPLDQDWVLEHHHHHH

>2423|Never Born Protein
MTKYRAYVLTANIGYALYERADATVHQHYHQPSEPHWNNQVQIGNKNPYVYKLVVVI
NDEVYSTDKDSHLQEMTYIYTPILNPSDELTWVKYIRDNYGGLDLEHHHHHH

>7144|Never Born Protein
MITMPLLAIEdNDLFYRTPHYLFCVGMYNKQDFKNYVDYTEALVRVNKDKEAK
LTWTSHTHGAEALHTGASRPFLFAKPYEGSGHQFIDNLYPPHEVALEHHHHHH

>4090|Never Born Protein
MVERDKPINWVYDAEPLQGGIVWVHLAALYCANVDDYAPQDHLDTMYGFDFHQ
KTNILSFEDSVNAQSYWQYGIVKFVSHGWEDLQGAIVESWRDSRSLEHHHHHH

>2298|Never Born Protein
MKWYGRGREDFGSPDVEKNCENGVIYGTSEQELYNVSVMWDWAGISEQPTIG
SLTTPNTKDDMLWYRNDAKPGHSILYNLINDYWEATEVSGIGNVVLEHHHHH
<table>
<thead>
<tr>
<th>clonal expression</th>
<th>solubility</th>
<th>MS</th>
<th>WB</th>
<th>purification</th>
<th>DLS</th>
<th>CD</th>
<th>GPC</th>
<th>fluorescence</th>
<th>FTIR</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>cloning</td>
<td>expression</td>
<td>solubility</td>
<td>MS</td>
<td>WB</td>
<td>purification</td>
<td>DLS</td>
<td>CD</td>
<td>GPC</td>
<td>fluorescence</td>
<td>FTIR</td>
</tr>
<tr>
<td>8836</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3027</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1856</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1441</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6387</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1687</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8635</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3451</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6620</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8600</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4712</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2423</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7144</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4090</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2298</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Summary of NBPs experimental characterization