

UNIVERSITÀ DEGLI STUDI DI VERONA

Facoltà di Scienze MM FF NN

Dottorato di Ricerca in Biotecnologie Industriali e Ambientali XVIII CICLO

Dipartimento Scientifico e Tecnologico



“Expression and Site-Directed Mutagenesis  
of Chicken Liver Bile Acid-Binding Protein  
and Structural Studies”

**Dottoranda: dr.ssa Marianna Luppi**

**Supervisore: prof.ssa Henriette Molinari**  
**Coordinatore: prof. Hugo L. Monaco**

# ***Ringraziamenti***

*Desidero ringraziare coloro che hanno contribuito alla realizzazione di questo lavoro di tesi.*

*In particolare sono molto grata alla Prof.ssa Henriette Molinari per avermi dato l'opportunità di lavorare a questo progetto e per avermi continuamente sostenuto.*

*Sono molto grata alla Dr.ssa Raffaella Ugolini per avermi sempre aiutata nella discussione scientifica dei problemi incontrati e per i preziosi consigli.*

*Il mio ringraziamento poi va a tutti i miei colleghi e alle persone che hanno condiviso con me questa esperienza, in particolare a Dr.ssa Silvia Romagnoli, Dr. Federico Fogolari, Dr.ssa Laura Ragona, Dr.ssa Maddalena Catalano, Dr.ssa Lucia Zetta, Massimo Pedò e Prof.ssa Guantieri Valeria.*

*Ringrazio il Dr. Moreno Pagnanelli e Dr. Michael Kaleja per le spiegazioni e per i preziosi consigli sull'uso degli strumenti cromatografici.*

*Ringrazio il Dr. Michael Assfalg, Dott.ssa Guariento Mara, Dott.ssa Zanzoni Serena e Bonfanti Simone per le ultime revisioni di questa tesi.*

*a A. & D.*

## **Publications**

Parts of this thesis will be published in due course:

Laura Ragona, Maddalena Catalano, **Marianna Luppi**, Daniel Cicero, Tommaso Eliseo, Jefferson Foote, Federico Fogolari, Lucia Zetta, and Henriette Molinari

**“NMR dynamic studies suggest that allosteric activation regulates ligand binding in chicken liver Bile Acid Binding Protein”**

*2006, Jan, 26 Journal of Biological Chemistry*

# Contents

Breve riassunto della tesi e scopo del lavoro .....	I
Motivation, objective and summary of the Ph.D. Thesis work.. .....	III
1. Biological Background .....	1
1.1. The Calicin Superfamily .....	1
1.2. The cytoplasmic Fatty acid Binding Protein .....	2
1.2.1 The FABPs fold .....	4
1.2.2 FABPs ligand binding .....	6
1.2.3 FABPs evolution, genes and chromosome location .....	7
1.2.4 Regulation of FABP production .....	9
1.2.5 Functions of FABPs .....	10
1.3. Chichen Liver Bile Acid Binding Protein .....	11
1.4. Bile Acid Formation, Secretion and Enterohepatic Circulation .....	13
1.5. Lipocalins .....	15
1.5.1 Lipocalins sequence and structure relations .....	17
1.5.2 Multiple molecular recognition properties .....	19
1.6. Glycodelin .....	20
1.6.1 Primary sequence of Glycodelin .....	21
1.6.2 Glycodelin gene expression .....	22
1.6.3 Isoforms of Glycodelin .....	23
2. Methods for Structural Studies .....	25
2.1. The Nuclear Magnetic Resonance in Protein Studies .....	25
2.2. The NMR Phenomenon .....	26
2.2.1 Thermal equilibrium .....	27
2.2.2 Radiofrequency pulses .....	27
2.2.3 Free precession, relaxation and NMR spectrum .....	28
2.3. Chemical Shift .....	28
2.4. Scalar Coupling .....	31
2.5. The Relaxation .....	31
2.5.1 Dipole-dipole interaction and rotational motion in liquid .....	32
2.5.2 Spin-lattice relaxation .....	33
2.5.3 Spin-spin relaxation .....	34
2.5.4 Nuclear Overhauser effect .....	35
2.5.5 NOE in Protein structure determination .....	37
2.5.6 Relaxation in Protein NMR .....	37
2.6. NMR experiments .....	39
2.6.1 1D NMR experiments .....	39
2.6.2 2D NMR experiments .....	43
2.6.3 3D NMR experiments .....	47
2.7. Requirements for an NMR sample .....	47
2.8. Strategies for High Level Expression and Labelling of Recombinant Proteins for NMR Studies .....	49
3. Materials and Laboratory Methods .....	58
3.1. Materials .....	58
3.1.1 Microorganisms .....	58
3.1.2 Media and buffers .....	58
3.1.3 Plasmids for protein expression .....	61
3.1.4 Molecular weight marker for SDS-PAGE electrophoresis .....	61
3.1.5 Other chemicals .....	61
3.2. Molecular Biology Techniques .....	64

3.2.1 Protocol for competent bacteria .....	64
3.2.2 Transformation of competent bacteria .....	64
3.2.3 Bacterial cultures .....	64
3.2.4 Site-directed mutagenesis .....	65
3.3. Tools of Biochemistry .....	68
3.3.1 SDS Polyacrylamide Gel Electrophoresis (SDS PAGE) .....	68
3.3.2 Determination of protein concentration .....	69
3.3.3 Lysis of <i>E. coli</i> .....	70
3.3.4 Chromatography Techniques .....	70
3.4. NMR Sample Preparation .....	72
4. Results and Discussion .....	73
4.1. Description of cL-BABP expression system .....	73
4.2. Development of the High Level Expression and Purification Protocols of Unlabeled cL-BABP from Rich Medium .....	76
4.2.1 Cell growth .....	76
4.2.2 Expression Test –Time Course .....	76
4.2.3 Solubility Test .....	77
4.2.4 Purification Procedure .....	78
4.2.5 Delipidation Procedure .....	80
4.2.6 Improvement of cL-BABP yield: other expression tests .....	82
4.2.7 Separation of cL-BABP from higher aggregates .....	84
4.3. NMR Experiments on unlabeled cL-BABP .....	86
4.3.1 Complex of cL-BABP and <sup>15</sup> N-Glycochenodeoxycholate (GDC) .....	87
4.4. Development of the High Level Expression Protocol of cL-BABP in Minimal Medium for the Subsequent <sup>15</sup> N and <sup>13</sup> C Isotopic Enrichment .....	89
4.4.1 Cell growth .....	89
4.4.2 Expression in the time course and solubility tests .....	90
4.5 Production of <sup>15</sup> N cL-BABP .....	92
4.5.1 Large scale expression using the two-stage protocol .....	92
4.5.2 Determination of percentage of enrichment by molecular mass spectrometry .....	93
4.5.3 NMR experiments .....	93
4.6 Production of <sup>15</sup> N <sup>13</sup> C cL BABP .....	94
4.6.1 Large scale expression using the two-stage protocol .....	94
4.6.2 Improvement of cL-BABP expression in minimal media .....	95
4.7. Mutants of cL-BABP .....	97
4.7.1 Rationale, production, NMR analysis of cL-BABP mutants .....	97
4.8. Recombinant expression of Glycodelin .....	113
4.9. Description of Glycodelin expression system .....	113
4.10. Development of expression and purification protocols of Glycodelin .....	114
4.10.1 Cell Growth .....	114
4.10.2 Expression Test-Time course .....	115
4.10.3 Solubility test .....	116
4.10.4 Purification procedure .....	116
5. Conclusions .....	119
6. References .....	121
Appendix A.1 .....	129
Appendix A.2 .....	131

## Abbreviations

$\alpha$ 2-PEG	pregnancy associated secreted $\alpha$ 2 globulin
$\beta$ LG	beta lactoglobulin
AGP	acid glycoprotein precursor
ASBT	apical sodium dependent bile acid transporter
AUP	$\alpha$ -uterine Protein
BA	bile acid
bp	base pair
BSEP	bile salt export pump
CAG-2	chorionic $\alpha$ -2-microglobulin
CG	chorionic gonadotrophin
cL-BABP	chicken Liver Bile Acid Binding Protein
CRABP	cellular Retinol Acid Binding Protein
DCA	deoxycholic acid
DHA	docosahexaenoic acid
DNA	desoxy Ribo Nucleic Acid
EHC	enterohepatic circulation
EP-15	endometrial protein 15
FABPS	Fatty Acid Binding Proteins
FXR	farnesoid X receptor
GdA	Glycodelin of human amniotic fluid
GdM	Glycodelin of bone marrow
GdS	Glycodelin of seminal plasma
HNF1a	Hepatocyte Nuclear Factor 1
hPR	human Progesterone Receptor
HSQC	Heteronuclear Single quantum Correlation spettroscopy
I-BABP	Intestinal Bile Acid Binding Protein
I-FABP	Intestinal Fatty Acid Binding
IMAC	Immobilized metal ion chromatography Protein
LCFA	Long Chain Fatty Acid
L-FABP	Liver Bile Acid Binding Protein
MeF2	Myocyte enhancer Factor 2
MPIs	Metallo Protease Inhibitors
mRNA	messenger Ribo Nucleic Acid
MUP	Major Urinary Protein
NGAL	Neutrophil Gelatinase Associated Lipocalin
NK	Natural killer
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser effect
NOESY	Nuclear Overhauser effect spectroscopy
NTCP	sodium/taurocholate cotransporter
OATP	organic anion-transporting polypeptide
OATP3	organic anion-transporting polypeptide subtype 3
OBP	Odorant Binding Protein
PBMCs	peripheral blood mononuclear cells

PEP or PAEP	progestagen dependent endometrial Protein
pI	Isoelectric point
POU	Pit-Oct-Unc transcription factor family
PPAR	Peroxisome Proliferators Activated Receptor
ppm	parts per million
PPRE	Peroxisome Proliferators Response Elements
RA	retinoic acid
RBP	Retinol Binding Protein
RXR	Retinoid X Receptor
SCR	structurally conserved regions
tASBT	truncated Apical Sodium Dependent Bile Acid Transporter
TOCSY	Total correlated spectroscopy
TP	lipocalin-Type Prostaglandin
ZIF	Zona Inhibitory Factor

## Breve riassunto della tesi e scopo del lavoro

La presente tesi descrive l'attività svolta durante il periodo del mio Dottorato di Ricerca in Biotecnologie Industriali ed Ambientali.

La parte principale del mio lavoro ha riguardato lo sviluppo di protocolli per l'over-espressione delle seguenti proteine per successivi studi strutturali tramite la tecnica spettroscopica di Risonanza Magnetica Nucleare (NMR): *Bile Acid Binding Protein* da fegato di pollo (cL-BABP) e Glicodelina umana (GdA). Queste proteine appartengono alla superfamiglia delle Calicine e questo lavoro si colloca in un più ampio progetto di ricerca, a cui da molti anni si sta lavorando nel nostro laboratorio, che riguarda lo studio comparativo della struttura, della stabilità e delle proprietà di interazione di proteine appartenenti alla superfamiglia delle Calicine. L'analisi dettagliata delle sottili variazioni conformazionali introdotte dall'evoluzione all'interno di una stessa famiglia proteica è fondamentale per poter approfondire le specifiche relazioni tra sequenza, struttura e funzione, e ottenere informazioni importanti su proprietà generali, quali il meccanismo di *folding* e modalità di interazione.

La spettroscopia NMR è una tecnica molto potente non solo per la determinazione della struttura tridimensionale, ma anche per la caratterizzazione delle modalità di interazione e delle proprietà dinamiche di molecole biologiche in soluzione. La biologia molecolare è un importante strumento per la tecnica NMR in quanto sono richieste quantità rilevanti (0.5-1 mM in 700  $\mu$ L) di campioni doppiamente arricchiti ( $^{15}\text{N}$  and  $^{13}\text{C}$ ) che presentino elevata purezza (> 95%).

I principali scopi di questo lavoro di tesi hanno riguardato i seguenti punti:

1. La funzione e il meccanismo di azione delle *Bile Acid Binding Proteins* (BABPs) non sono attualmente stati chiariti, sebbene per questa famiglia di proteine sia riportata principalmente la funzione di trasporto di molecole idrofobiche all'interno della cellula. Lo studio strutturale e la determinazione delle proprietà di legame di cL-BABP può perciò chiarire il ruolo funzionale di queste proteine nell'ambito del trasporto degli acidi biliari all'interno della cellula. Gli acidi biliari sono coinvolti nella circolazione enteroepatica, un processo collegato all'omeostasi del colesterolo. Sono perciò stati messi a punto i protocolli di over-espressione e purificazione per la produzione di  $^{15}\text{N}$  e  $^{13}\text{C}$  cL-BABP al fine di effettuare studi sulla struttura, sulle proprietà dinamiche e sull'interazione.
2. Sono successivamente stati disegnati degli opportuni mutanti al fine di identificare determinanti di legame e di *folding* nonché di ottenere informazioni sulla funzione specifica di questa famiglia di proteine.

3. Per Glicodelina, una proteina glicosilata, sono riportati diversi ruoli biologici, sia nell'inibizione delle cellule *natural killer*, sia nell'angiogenesi. Considerati questi importanti ruoli funzionali, non sono disponibili studi strutturali su questa proteina. Una parte di questo lavoro perciò ha riguardato lo sviluppo dell'espressione e della purificazione di Glicodelina espressa nella forma non glicosilata, come punto preliminare per la caratterizzazione strutturale.

La presente tesi è così organizzata:

- **Capitolo 1** riporta l'introduzione sulle proteine oggetto di studio, che appartengono rispettivamente alle famiglie delle Fatty Acid Binding Proteins (FABPs) e delle Lipocaline.
- **Capitolo 2** descrive i principi della risonanza magnetica nucleare applicata allo studio di molecole biologiche. E' riportata una breve introduzione del fenomeno fisico assieme ad una descrizione delle procedure sperimentali e delle sequenze degli impulsi usate per lo studio di macromolecole.
- **Capitolo 3** contiene la descrizione dei materiali e dei metodi impiegati in questo lavoro di tesi.
- **Capitolo 4** riporta la descrizione dettagliata e la discussione dei risultati. I paragrafi 4.1 e 4.2 trattano la descrizione del sistema di espressione di cL-BABP e gli esperimenti effettuati per lo sviluppo dei protocolli di espressione; il paragrafo 4.3 riporta gli esperimenti NMR effettuati sulla proteina apo e olo complessata con gli acidi biliari; i paragrafi 4.4, 4.5 and 4.6 descrivono gli esperimenti effettuati per l'over-espressione in mezzo minimo ed il successivo arricchimento di cL-BABP in  $^{15}\text{N}$  e  $^{13}\text{C}$ . Il lavoro pubblicato è qui incluso; il paragrafo 4.7 descrive il disegno e la produzione di specifici mutanti di cL-BABP e riporta gli esperimenti NMR preliminari con i relativi risultati; i paragrafi 4.8, 4.9 and 4.10 riportano le considerazioni introduttive sull'espressione ricombinante di Glicodelina e la descrizione i risultati preliminari.
- **Capitolo 5** riporta le conclusioni
- **Capitolo 6** riporta le referenze.

## Motivation, objective and summary of the Ph.D. Thesis work

The main subject of my thesis work has been the development of over-expression protocols for the production of the following proteins: chicken Liver Bile Acid Binding Protein (cL-BABP) and Glycodelin (GdA) for structural studies through Nuclear Magnetic Resonance. This proteins belong to the Calycin superfamily and this project is part of a wider research work, going on in our laboratory, involving the comparative study of structure, stability and interaction properties of members of the Calycin superfamily.

The rationale underlying this project is that the detailed analysis of the subtle structural changes brought about by evolution within the same protein family is relevant to the understanding of the sequence-function specificity relationships and of the main folding determinants.

NMR spectroscopy is a powerful technique, not only for determining the three-dimensional structure, but also for the characterization of interactions and dynamics of biological molecules in solution. Molecular biology is an important tool for NMR as it is mandatory not only to be able to produce samples of high purity (> 95%) but also to obtain relevant amounts of doubly enriched ( $^{15}\text{N}$  and  $^{13}\text{C}$ ) proteins (0.5-1 mM samples, in 700  $\mu\text{L}$  solution are typically required).

The main aims of this thesis work are related to the following points:

1. The function and the mechanism of action of BABPs are still unclear. The structural studies and the determination of binding properties can therefore provide insights for the understanding of the functional role played by this proteins, in connection with bile acids transport within the cell. Bile acids are involved in enterohepatic circulation, a process related to cholesterol homeostasis. Over-expression and purification protocols have been therefore set up for the production of  $^{15}\text{N}$  and  $^{13}\text{C}$  cL-BABP in order to perform structural dynamics and interaction studies.
2. Different cL-BABP mutants have been designed and produced in order to identify the determinants of protein binding and dynamics, thus providing functional information.
3. Glycodelin, a glycosylated protein, has been reported to play an important biological role both in the inhibition of natural killer cells activity and in angiogenesis. In spite of its important biological functions no structural studies are available. In the present work the main focus has been the development of expression and purification protocols for non glycosylated Glycodelin as a preliminary step towards the structural characterisation.

The present thesis is organised as follows:

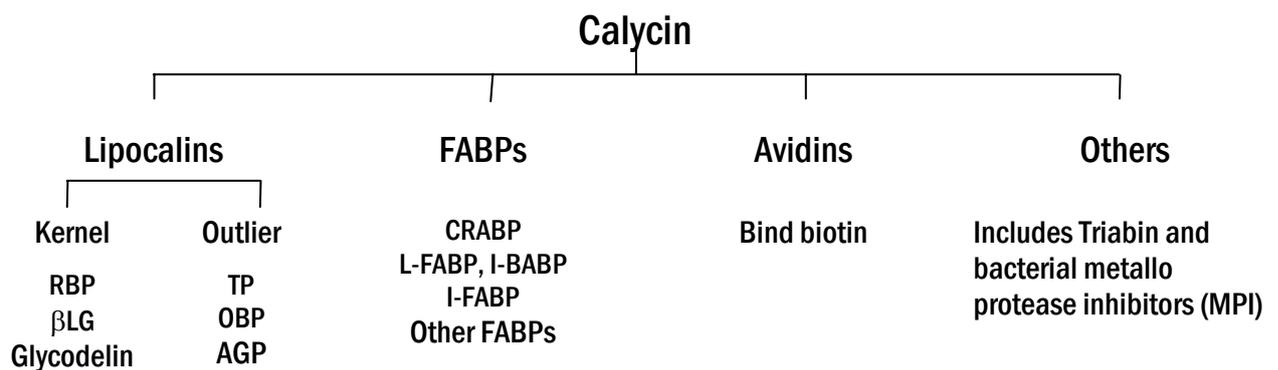
- **Chapter 1** provides an overview of the proteins under study, belonging to the Fatty Acid Binding Proteins (FABP) and Lipocalins families.
- **Chapter 2** provides an overview of the NMR spectroscopy techniques applied to bimolecular systems. A brief introduction to the physical phenomena is given together with a description of the experimental procedures and pulse sequences used to study biological macromolecules.
- **Chapter 3** contains the description of materials and methods employed for the present work.
- **Chapter 4** reports a detailed description and discussion of the results. Paragraphs 4.1 and 4.2 deal with the description of the cL-BABP expression system and the experiments performed for the development of over-expression protocols; paragraph 4.3 reports the NMR experiments performed on apo and holo cL-BABP complexed with bile acids; paragraphs 4.4, 4.5 and 4.6 describe the experiments performed for the development of over-expression protocol in minimal medium for  $^{15}\text{N}$  and  $^{13}\text{C}$  labeling and includes the paper published; paragraph 4.7 describes the design and production of specific cL-BABP mutants; paragraph 4.8, 4.9 and 4.10 report the introductory considerations on recombinant expression of Glycodelin, the description of the Glycodelin expression and purification and the preliminary results for Glycodelin.
- **Chapter 5** reports the conclusions.
- **Chapter 6** reports the references.

# 1. Biological Background

## 1.1. The Calycin Superfamily

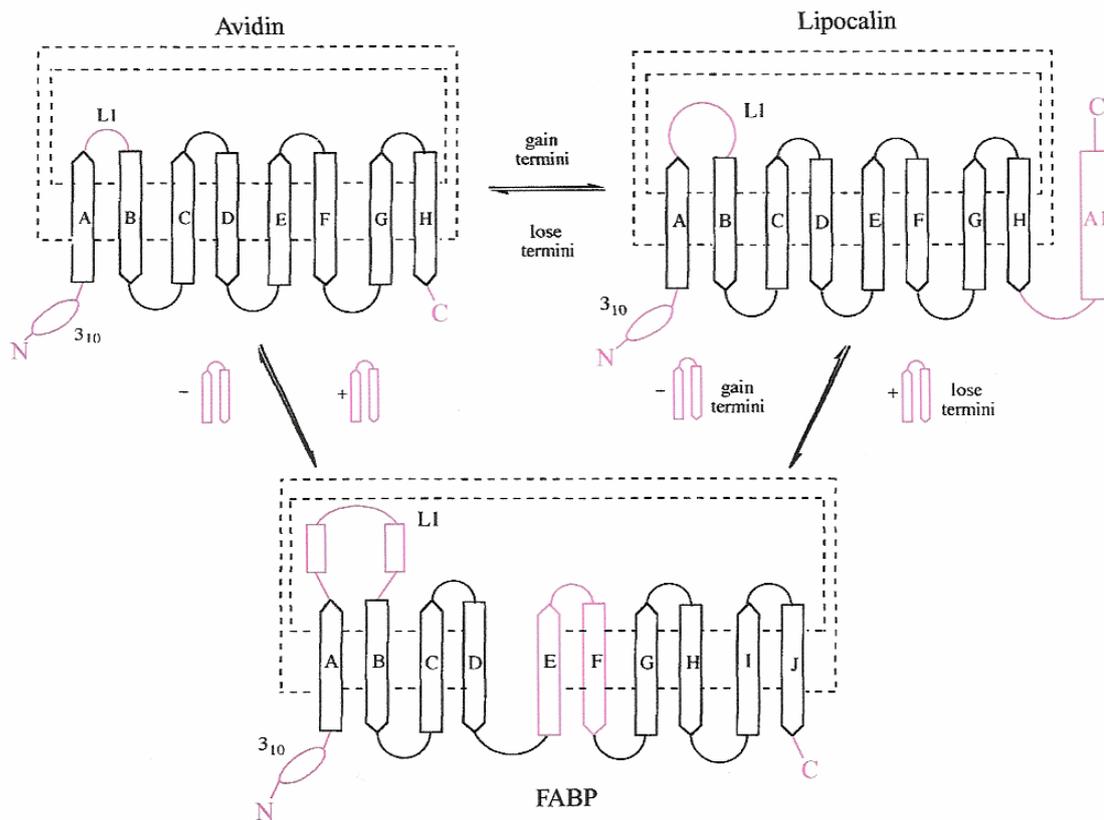
The main subject of this thesis work is about proteins belonging to the structural superfamily of Calycins: **chicken liver Bile Acid Binding Protein (cL-BABP)** and **Glycodelin (Gd)**. As the principal levels of classification for proteins are superfamily, family, and fold, the Calycin superfamily is formed by three families of ligand-binding proteins, the **cytoplasmic Fatty Acid-Binding Proteins (FABPs)**, the **Lipocalins** and the **Avidins**, together with Triabin and a group of bacterial metalloprotease inhibitors (MPIs) (Table 1.1). In particular, cL-BABP belongs to the FABPs family and Glycodelin to the Lipocalins. This classification reflects both structural and evolutionary relatedness.

Table 1.1 The Calycin superfamily



The Calycin superfamily is characterized by (Figure 1.1.) (Flower *et al.*, 2000):

- Similar folding pattern  $\Rightarrow$  an antiparallel  $\beta$  barrel with repeated +1 topology (Triabin has a similar barrel geometry but a different topology).
- Little global sequence similarity signature  $\Rightarrow$  it groups together proteins that have low sequence identities ( $< 30\%$ ), but whose structural and functional features suggest that a common evolutionary origin is probable.
- Conserved sequence motif  $\Rightarrow$  the short  $3_{10}$ -helix leading into the  $\beta$ -strand A.
- Distinct structural signature  $\Rightarrow$  an arginine or lysine residue which forms several hydrogen bonds with the main-chain carbonyl groups of the N-terminal  $3_{10}$ -helix and packs across a conserved tryptophan from the  $\beta$ -strand A.
- Function similarity  $\Rightarrow$  the binding of hydrophobic ligands.



**Figure 1.1.** Relationship between FABP, Avidin, and Lipocalin fold.  $\beta$ -strands are shown as arrows and labelled by letter. The N-terminal  $3_{10}$ -helix and the C-terminal Lipocalin  $\alpha$ -helix (labelled A1) are marked. The hydrogen-bonded connection of two strands is indicated as a pair of dotted lines between them. Connecting loops are shown as continuous lines; loop L1 is labelled and marked for each fold (reproduction from Flower, 1996).

## 1.2. The cytoplasmic Fatty Acid Binding Proteins family

FABPs are small intracellular proteins (14-15 kDa, about 130 residues) isolated from different tissues and species which include mammals, fish, birds and insects. These proteins are products of a large multigene family of unlinked loci distributed throughout the genome. They are able to bind a variety of hydrophobic ligands such as long chain fatty acids (LCFA), retinoids and bile salts. The first FABPs were discovered in mammalian tissues in 1972 (Ockner *et al.*) and since then many other members of this family have been isolated (Table 1.2). The proposed nomenclature (Glatz *et al.*, 1990) provides for the use of the general abbreviation X-FABP, where X is the tissue of predominance type (liver, L; adipose, A; heart, H; muscle, M; brain, B; intestinal, I; epidermal, E; keratinocyte, K).

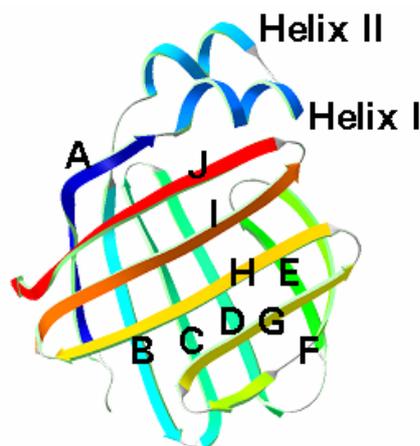
Table 1.2: Nomenclature and expression pattern for cytoplasmic Fatty Acid Binding Proteins				
FABP-type	Alternative name	Gene name (human)	Mammalian expression	Non-mammalian expression
L-FABP (liver)		<i>fabp1</i>	Liver, intestine, kidney, lung, pancreas	
I-FABP (intestinal)		<i>fabp2</i>	Intestine	
H-FABP (heart)	M-FABP (muscle), MDGI	<i>fabp3</i>	Heart, skeletal muscle, kidney, lung, mammary placenta, testis, stomach, ovary	Fish muscle, bird muscle, insect muscle, fish ovary
A-FABP (adipocyte)	ALBP, aP2	<i>fabp4</i>	Adipose tissue	
E-FABP (epidermal)	E-FABP, KLEP	<i>fabp5</i>	Skin, adipose tissue, lung, brain, heart, skeletal muscle, testis, retina, kidney	
I-BABP (intestinal)	ILBP, Gastrotopin	<i>fabp6</i>	Ileum	
Brain FABP	B-FABP	<i>fabp7</i>	Brain, neurons	Bird brain, retina
	R-FABP			
M-FABP (myelin)	mP2, myelin P2	<i>fabp8</i>	Schwann cells	
T-FABP (testis)	T-FABP	<i>fabp9</i>	Testis	
Lb-FABP (liver basic)	L-FABP	<i>fabp10</i>		Fish, chicken, iguana liver
Midgut FABP				Insect midgut
RABP	CRABP	<i>crabp</i>		African frog
RABP1	CRABP-I	<i>crabp1</i>	Skin, Adrenal glands, testis	Chicken, Japanese pufferfish
RABP2	CRABP-II	<i>crabp2</i>	Placenta, colon, mammary gland, corpus luteum	Chicken embryo
RETY	CRABP-IV	<i>rbp7</i>	Pancreas, mammary gland, embryo	

Data from Swiss-Prot and TrEMBL databases (<http://www.expasy.org/sprot>)

### 1.2.1 The FABPs fold

The primary structure of over 60 members of the FABPs family has been determined. The tertiary structure of all FABPs is highly conserved, despite the considerable differences in their primary structure. Sequence identities in this protein family range from 25% for some paralogous member to over 90% for some orthologous (for example, human, bovine, murine and rat H-FABP show between 92-96% of identity and 92-96% similarity, while human H-FABP and L-FABP have only 27% identity and 42% similarity).

The structure of FABPs is characterized by 10 antiparallel  $\beta$  strands ( $\beta A-\beta J$ ) with  $\beta A-\beta E$  and  $\beta F-\beta J$  forming the typical barrel structure (Figure 1.2). The strands of the two sheets ( $\beta A-\beta E$  and  $\beta G-\beta J$ ) are nearly orthogonal to each other. All  $\beta$  strands are hydrogen-bonded to the preceding and succeeding strands except  $\beta D$  and  $\beta E$ . This rift in the continuous surface of the barrel allows flexibility of the barrel structure without significant disruption of the H-bonded strands around the rest of its circumference. At the 'top' of the gap region (the structure is usually oriented with the helix-turn-helix motif at the top and back, figure 1.2) is the portal region where the base of the helix-turn-helix and right turns of  $\beta CD$  and  $\beta EF$  are proximal. It is through the portal region that the ligand probably enters and exits the internal binding site. The internal cavity within the barrel contains a number of hydrophilic side chains that form charge/charge interactions and a H-bonded lattice that interacts directly with the carboxylate moiety of the ligand. Table 1.3 lists the FABPs structures that have been reported.



**Figure 1.2.** Three-dimensional structure of apo cL-BABP (PDB code 1MVG). All FABPs have the characteristic  $\beta$ -barrel structure, in which 10 antiparallel  $\beta$ -strands form the barrel shaped ligand binding site, framed by the helix-turn-helix domain as part of the portal.

**Table 1.3** A list of 3D structures of FABPs deduced from X-ray crystallography or NMR solution studies. Studies with recombinant FABP are indicated by (recomb). When the structure was determined by crystallography the resolution (in Ångstroms) is given in parentheses. Whether the structure solution was carried out with protein bearing ligand is indicated by (holo or apo). The table does not include various mutated FABP structures.

Year	Organism	FABP type	Method	PDB Reference
1992	Rat	I (recomb) apo	X-ray (1.96)	1IFB
1992	Rat	I (recomb) holo	X-ray (2.0)	2IFB
1998	Bovine	H (recomb)apo	NMR	1BWY
1993	Mouse	A apo	X-ray (2.5)	1ALB
1994	Locust	M apo	X-ray (2.2)	1FTP
1995	Human	M holo	X-ray (1.4)	1HMR/1HMS/1HMT
1997	Rat	I apo	NMR	1AEL
1997	Rat	I holo	NMR	1URE
1997	Rat	L (recomb) holo	X-ray (2.3)	1LFO
1997	Pig	Ileal (recomb)apo,holo	NMR	1EAL/1EIO
1999	Human	E (recomb) apo*	X-ray (2.05)	1B56
2001	Human	H (recomb) apo	NMR	1G5W
2001	Human	B apo, holo	X-ray (2.10,2.80)	1FDQ/1FE3
2002	Human	E (recomb) apo*, holo	NMR	1JJJ/1JJX
2003	Human	I (recomb) apo	NMR	1KZW
2003	Chicken	L	NMR	1MVG
2003	Toad	L	X-ray (2.5)	1P6P
2004	Human	A	X-ray (2.0)	1TOW
2004	Chicken	L apo, holo	X-ray (2.0)	1TVQ/1TW4
2006	Chicken	L (recomb) apo	NMR	1ZRY

\* There is an unusual disulfide linkage in the epidermal protein between Cys120 and Cys127 not seen in other FABPs.

FABPs are classified in four subfamilies (Spener & Haunerland, 2004), as outlined by the phylogenetic tree obtained from the multiple sequence alignment performed among all FABPs using Clustal W (Clustal W, <http://www.ebi.ac.uk/clustalw/>), (Figure 1.3):

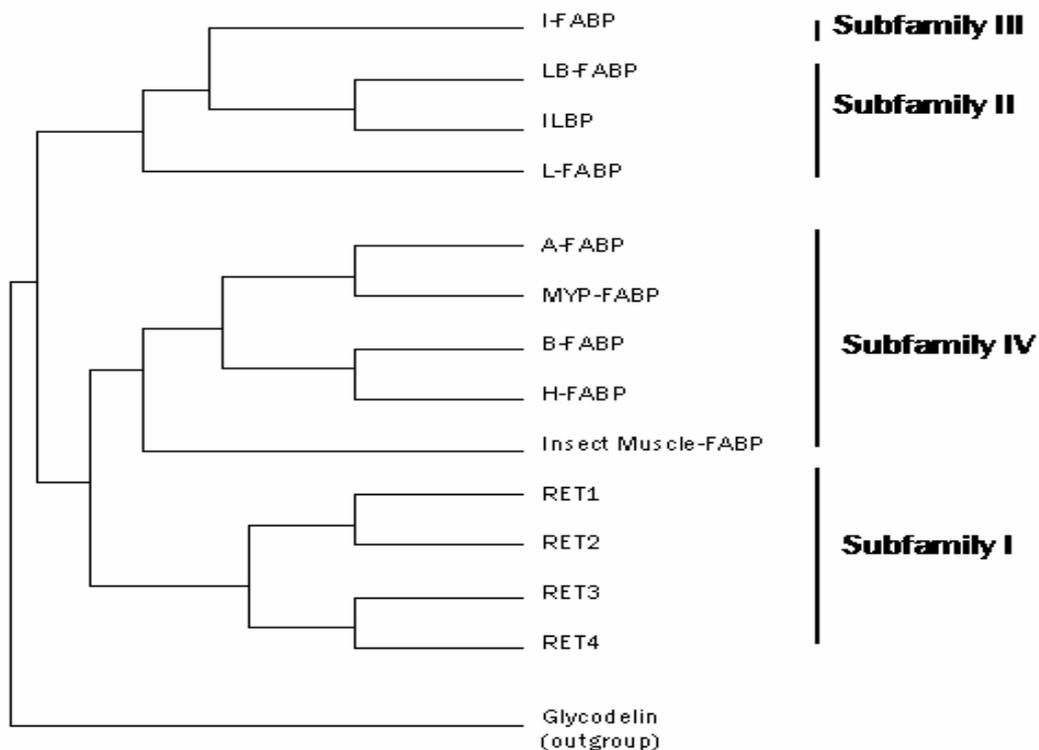
*Subfamily I:* Cellular Retinoic Acid Binding Proteins (CRABPs) bind retinoic acid (RA) with high affinity, exhibit distinct patterns of expression during embryonic development and are thought to play important roles in the RA signaling pathway.

*Subfamily II:* Liver Fatty Acid-Binding Proteins (L-FABPs) and Intestinal Bile Acid-Binding Proteins (I-BABPs) are closely related based on sequence homology and both stand out because of their unusual ligand binding specificities. This subfamily of proteins binds a broad range of ligand molecules (acyl-CoAs, heme, squalene, bile acids, bilirubin and certain eicosanoids); in particular L-FABP are the only FABPs that form a complex with two fatty acid molecules at the same time.

*Subfamily III:* Intestinal Fatty Acid-Binding Proteins (I-FABP) is rather singular in sequence

characteristics and binds only one fatty acid molecule.

*Subfamily IV:* This subfamily comprises the largest number of different type of FABPs, i.e. H- (heart-type), A- (adipocyte-type), E- (epidermal-type), M- (myelin-type), T- (testis-type) and B- (brain-type) FABP. They all bind only a single fatty acid molecule.



**Figure 1.3.** Evolutionary tree of the FABPs. Sequences for the vertebrate and invertebrate members of FABPs gene family were aligned with Clustal W. The tree was constructed with the neighbor joining method, using Glycodelin as an out group.

### 1.2.2 FABPs ligand binding

In the binding pocket of FABPs the deprotonated carboxyl group of the bound ligand is generally buried inside the cavity for electrostatic interaction with one or two arginine residues, in addition to be hydrogen bonded by a tyrosine- or serine-OH and an ordered water molecule (Zanotti *et al.*, 1992). Nonetheless, important differences between individual FABP-type exist, which influence binding kinetics and affinity as well as the mechanism of ligand transfer (Hanhoff *et al.*, 2002; Zimmerman *et al.*, 2002).

A close-up inspection of ligand conformation in the resolved protein structures fosters the above-mentioned subfamily concept for FABPs (Figure 1.4):

*Subfamily I:* the conformation of the characteristic isoprenoid tail of the retinoid ligands is extended and the  $\alpha$ -ionone ring located close to the helix-turn-helix domain, whereas the functional group is always deeply immersed into the binding cavity.

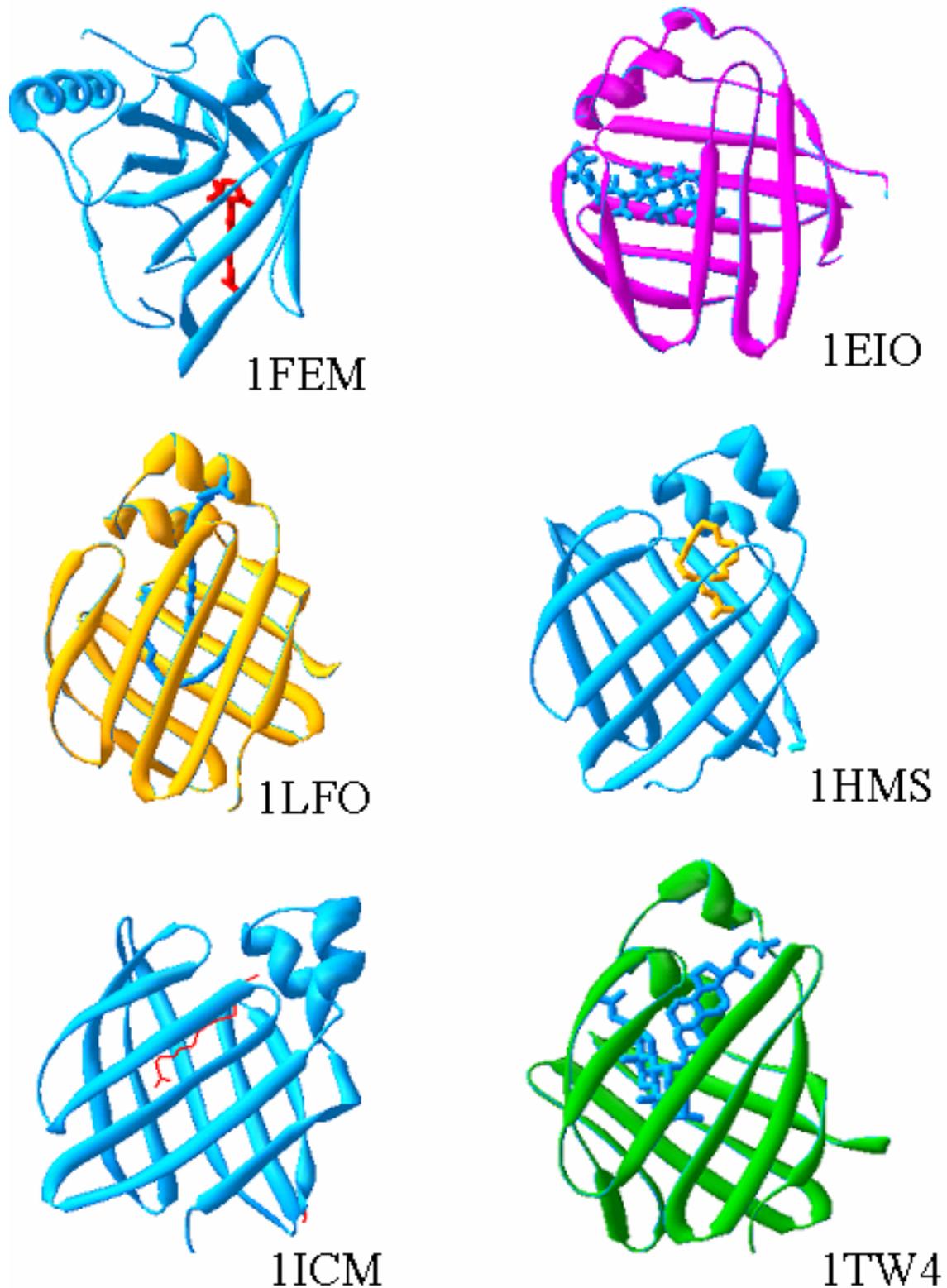
*Subfamily II:* of the two fatty acids bound by L-FABP, one is coordinated in a bent conformation electrostatically via Arg121 and an extensive hydrogen-bonding network involving Ser123 and 38 located at the bottom of the protein cavity, which again is reminiscent of fatty acid binding in subfamily IV. The second fatty acid in L-FABP adopts a rather linear shape, with the acyl chain in the cavity extending down towards the centre of the other fatty acid molecule and the carboxylate sticking out of the fatty acid portal, thus being solvent exposed and pH sensitive. Interestingly, although I-BABP contains the respective residues (Arg121, Ser123 and 38), it binds fatty acid only weakly, while it binds a bile acid molecule with high affinity. Again, the bulk steroid molecule is inside the cavity and the carboxylate group at the protein–solvent interface.

*Subfamily III:* The fatty acid bound by I-FABP adopts a slightly bent conformation, reverse in direction to the second fatty acid in L-FABP, thus the carboxylate group is located deep inside the protein cavity directly coordinated to the side-chain of Arg106 similar to the ligands carboxylate bound by proteins belonging to subfamilies I and IV.

*Subfamily IV:* the FABP-types of this subfamily all bind only a single fatty acid molecule in a U-shaped conformation. While the carboxylate group is bound electrostatically and hydrogen bonded via Arg106 and 125 as well as Tyr127 (H-FABP numbering), the hydrocarbon chain is located close to Phe57 (Leu60 in E-FABP) at the fatty acid portal. Several unique features in this FABPs subfamily have been reported only recently. First, human E-FABP contains six cysteine residues, of which C120 and C127 form a disulfide bridge inside the protein cavity. Secondly, human B-FABP binds oleic acid in the common U-form conformation, but very long-chain docosahexaenoic acid (DHA) in a helical conformation. The three-dimensional structure of insect muscle FABP has been solved for the apo-protein only. It is remarkably similar to mammalian H-FABP, although steric limitations seem to predict a somewhat different shape of the ligand in the binding pocket.

### 1.2.3 FABPs evolution, genes and chromosome location

Analysis of FABPs evolution emerged as soon as amino acid sequences and cDNA sequences became available (Lowe *et al.*, 1985, Schleicher *et al.*, 1995). The picture drawn is of an ancient precursor that was probably the initial solution to solubilizing and mobilizing hydrophobic resources at least  $10^9$  years ago. The precursor that gave rise to the liver/intestinal/ ileal grouping emerged from the heart/adipose/myelin P2 lineage about 700



**Figure 1.4.** Ligand conformation in different FABPs complexes. The PDB code is reported for each complex. 1FEM: human CRABP complexed with retinoic acid; 1EIO: pig ileal binding protein complexed with one molecule of chenodeoxycholic acid; 1LFO: rat liver FABP complexed with two molecules of oleic acid; 1HMS: human muscle FABP complexed with oleic acid; 1ICM: rat intestinal FABP complexed with miristic acid; 1TW4: Chicken liver FABP complexed with two molecules of cholic acid.

millions of years ago, prior to the vertebrate/invertebrate divergence. While muscle type FABPs has been found in locust, no liver/intestinal types have been reported for invertebrates. The retinoid binding proteins appear to have diverged from the liver/intestinal line about 500 millions of years ago.

It has been suggested that the multiple CRABP genes of mammals arose from gene duplication events after the split with amphibian lines: *Xenopus* has only one CRABP gene (Matarese *et al.*, 1989). The structures of fatty-acid-binding protein genes are remarkably similar, with four exons separated by three introns and with positioning of the introns very similar relative to the coding regions, although there is considerable variability in the intron length. The exception to this organization is actually the two-intron, three-exon structure of locust flight muscle FABPs (Haunerland *et al.*, 1994).

### 1.2.4 Regulation of FABP production

All FABP promoters contain a classical TATA box. The elements that control the tissue-specific expression of FABP are currently only poorly understood, but potential enhancer sequences have been characterized for several genes. These include two hepatocyte nuclear factor 1 (HNF1a) regulatory elements in the L-FABP promoter (Akiyama, *et al.*, 2000), a fat-specific enhancer required for A-FABP expression in adipocytes (Ross *et al.*, 1990), and several binding sites for members of the Pit-Oct-Unc (POU) transcription factor family that control B-FABP expression (Josephson *et al.*, 1998). A concise promoter region that contained an atypical myocyte enhancer factor 2 (MEF2) binding site was shown to be responsible for the muscle-specific expression of H-FABP (Qian *et al.*, 1999). Better understood is the up-regulation of various FABP genes by fatty acids. It has long been known that the induction of FABP expression in response to lipid-rich diet (Malewiak *et al.*, 1988) or endurance training (Vanbreda *et al.*, 1992) is the result of increased intracellular concentrations of fatty acids, which in turn activate nuclear transcription factors (Duplus *et al.*, 2000; Wolfrum *et al.*, 2000). The best known of such transcription factors are the subtypes of the peroxisome proliferators activated receptor (PPAR a, b, g), so called because of their activation by xenobiotic peroxisome proliferators in rodents (Desvergne *et al.*, 1999); LCFA and certain eicosanoids are considered as their natural ligands. PPARs bind as heterodimers with the subtypes a, b, g of the retinoid X receptor (RXR) to direct-repeat elements (peroxisome proliferators response elements, PPREs) in the promoter region of the genes that they regulate.

Analogously, recent studies report for Bile Acids (BA) a role in FABP gene expression control (Makoto *et al.*, 1999), since they are physiological ligands for the farnesoid X receptor (FXR), which repress transcription of the gene encoding cholesterol 7 $\alpha$ -hydroxylase

and activate the gene encoding intestinal BABP, which is a candidate in bile acid transport (see section 1.4 for more details).

### 1.2.5 Functions of FABPs

Since the discovery of FABPs with a strong affinity for LCFAs and their presence at large concentrations in tissues that produce large fluxes of LCFAs, it has been clear that they perform a transport function. Since that time, there have been many refinements of this hypothesis as information became available (Glatz *et al.*, 1996; Storch *et al.*, 1996; Veerkamp, *et al.*, 1993; Vork *et al.*, 1993; Borchers *et al.*, 1994). It was suggested that the expression of more than one type of FABP in a tissue is a strong predictor that FABPs perform functions other than bulk transport. Additionally, the presence of true isoforms (with amino acid substitutions and distinct mRNAs such as seen in bovine H-FABP, (Schröder *et al.*, 1998) may indicate either very subtle functional refinements or the development of two homologous genes in progress. The functions suggested for the cytoplasmic FABP family are:

- increasing transport across plasma membrane;
- solubilizing LCFAs and facilitating their diffusion;
- protecting from toxic and/or detergent effects of LCFAs and mediating partitioning of LCFAs between membrane and solubilized pools;
- targeting LCFAs to intracellular sites and specific enzyme systems;
- mediating cell signaling metabolites (eicosanoids, prostaglandins, leukotrienes and thromboxanes);
- behaving as an LCFAs sensor in regulation of FABP mRNA expression.

Recent and still tentative additions to the list of possible functions include modulation of aspects of protein synthesis and integrating carbohydrate and fatty acid metabolism but clearly, in contrast to the very detailed knowledge of the structure and binding characteristic of FABPs, much less is known about their biological function.

### 1.3. Chicken Liver Bile Acid Binding Protein

In the liver, two paralogous groups of FABPs have been described: liver fatty acid binding protein (L-FABP) (Thompson *et al.*, 1999) type, extensively characterized in mammals, and liver (basic) fatty acid-binding proteins (Lb-FABP) that have not yet been found in mammalian liver but have been described in several other vertebrates such as chicken (Scapin *et al.*, 1988), axolotl, toad, iguana (Di Pietro *et al.*, 1999), catfish (Di Pietro *et al.*, 1997), zebrafish (Denovan-Wright *et al.*, 2000), lungfish (Di Pietro & Santomè, 2001) and shark (Cordoba *et al.*, 1999). The word “basic” was added to the acronym FABP to name the first member of this family identified in chicken liver because the protein turned out to have an isoelectric point (pI) of 9.0. This protein displays very low sequence similarity with the other liver-type FABPs (between 28 and 43%), while the percentage of identity between Lb-FABP-types range from 79-90%. Based on sequence homology, the Lb-FABP-type is evolutionary closely related to the I-BABP subgroup. This observation reflects differences in binding properties of these FABP-types: the mammalian L-FABP-type differs from most other members of the FABP family in that it binds two fatty acid molecules (Thompson *et al.*, 1997), whereas the chicken Lb-FABP-type, the best known “basic” type, have been shown to bind a single fatty acid molecule (Schievano *et al.*, 1994; Beringhelli *et al.*, 2001) and to bind two molecules of bile acids (Vasile *et al.*, 2003); this fact supports the proposal that the main function of the Lb-FABPs is more likely to be binding bile acids and not fatty acids. The fact that this protein also binds fatty acids is not unexpected since a similar lack of specificity has also been observed in other members of the FABP family. From these observation chicken liver basic FABP has been renamed chicken liver Bile Acid Binding Protein (cL-BABP).

The alignment of the amino acid sequences of cL-BABP with those of other Lb-FABP-types, of mammalian type L-FABPs and of the five ileal BABPs that are currently available (human, mouse, pig, rabbit and rat) is here reported (Figure 1.5).

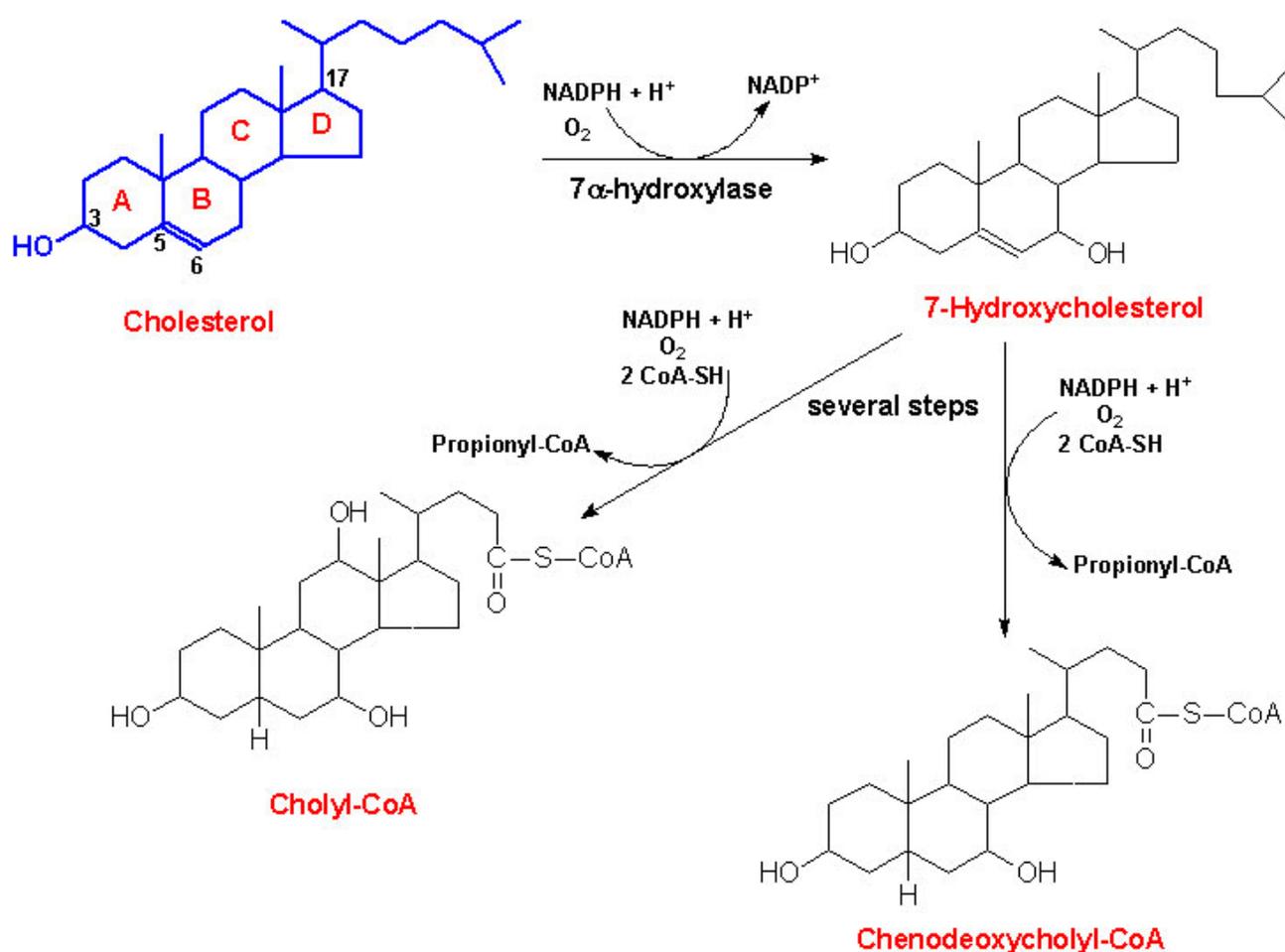
The sequences were aligned using CLUSTALW. The identity scores between cL-BABP and the other sequences are given.

The solution and crystal structures of cL-BABP are known. By NMR, the structure of the apo protein has been solved at pH 5.6 and 7.0 (Vasile *et al.*, 2003, Ragona *et al.*, 2006, PDB entries 1MVG, 1ZRY). By crystallography, the structure of apo protein and holo complexed with cholate refined to 2.0 Å has been resolved (Nichesola *et al.*, 2004, PDB entries 1TVQ and 1TW4).



## 1.4. Bile Acid Formation, Secretion and Enterohepatic Circulation

Bile acids are the major constituents of bile, and in mammals compose approximately 67% of this secretion. Bile acids synthesis in mammals and birds occurs in the liver and is considered a primary pathway of cholesterol metabolism (Center *et al.*, 1993). Cholesterol is the precursor molecule for bile acid synthesis (Figure 1.6, shown in blue). This multistep process involves the conversion of cholesterol to chenodeoxycholic acid, the most abundant bile acid in human bile (45%) and cholic acid (31%) (Figure 1.6). The rate limiting step of this process is catalyzed by the enzyme 7 $\alpha$ -hydroxylase. The enzymatic conversion of cholesterol to bile acids is regulated through feed-forward activation by oxysterols and feedback repression by bile acids (Russel *et al.*, 1992, Gong *et al.*, 1994). The activity of this enzyme is influenced by fasting, cholestasis, glucocorticoid exposure, liver failure, and lymphatic drainage (bile acid wasting),



**Figure 1.6.** Synthesis of the 2 primary bile acids, cholic acid and chenodeoxycholic acid. The reaction catalysed by the 7 $\alpha$ -hydroxylase is the rate limiting step in bile acid synthesis. Conversion of 7 $\alpha$ -hydroxycholesterol to the bile acids requires several steps not shown in detail in this image. Only the relevant co-factors needed for the synthesis steps are shown.

which can affect bile acid production. Within the liver, bile acids are conjugated by the addition of glycine or taurine (Figure 1.7), which are polar amino acids; these conjugation reactions yield glyco-coniugates and tauro-conjugates.

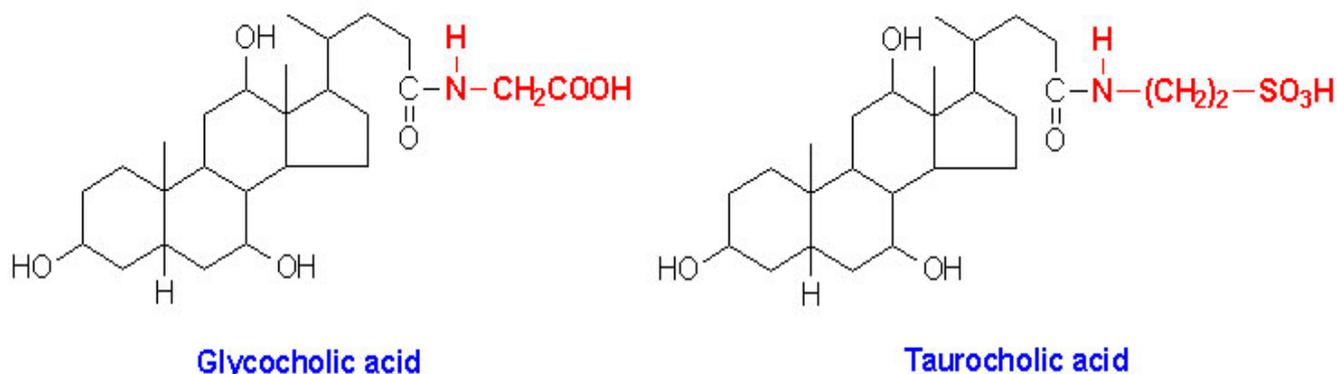


Figure 1.7. Structure of the conjugated cholic acids.

Carrier protein(s) bound bile salts for transport to the bile ducts for immediate secretion into bile by an ATP-dependent transporter, the bile salt export pump (BSEP), located in the canalicular membrane. The transport across this membrane is the rate limiting step in the transfer of bile salts from blood to bile. In the gallbladder bile salts are stored for future use. Presence of food in the duodenum stimulates the release of hormones such as secretin, vasoactive intestinal peptide and/or cholecystokinin. In species that have a gallbladder, the effects of these hormones include relaxation of the sphincter of Oddi, gallbladder contraction, and expulsion of bile acids through the bile duct and into the small intestine (Jones *et al.*, 1999). The ultimate fate of bile acids is secretion into the intestine, where they aid in the emulsification of dietary lipids. In mammals bile acid circulation from the gut lumen to the ileum is mediated by apical sodium dependent bile acid transporter (ASBT) and organic anion-transporting polypeptide subtype 3 (OATP3) proteins, present on the brush boarders, and by ILBPs which bring them, through the cytoplasm, to the basolateral ileal membranes. Here, a truncated apical sodium dependent bile acid transporter (tASBT) secretes bile salts into portal capillaries, where they bind to albumin and flow to the liver, where they are recognized by sodium/taurocholate cotransporter (NTCP) and organic anion-transporting polypeptide (OATP). This process of secretion from the liver to the gallbladder, to the intestine and finally reabsorbtion is termed the enterohepatic circulation (EHC) (Figure 1.8). During the EHC deconjugation (the removal of glycine and taurine residues) occurs regularly in normal humans. Rewiews (Lewis *et al.*, 1972) indicate that a number of species of bacteria normally found in the gastrointestinal tract can deconjugate bile acids and that the majority of bacteria capable of deconjugation were strict anaerobic. Bacterial modification of secondary bile acids



## 1.5. Lipocalins

Lipocalins are a large group of small (18 kDa, about 160 residues), mostly extracellular proteins that are found in vertebrates and invertebrate animals, plants and bacteria (Table 1.4), (Brew *et al.*, 1985). The collective name “Lipocalin” derived from the greek words 'lipos', meaning fat, and 'calyx', meaning cup (Pervais and Brew, 1987). They are characterized by a range of different molecular recognition properties: the binding of small, principally hydrophobic molecules (such as retinal and fatty acids), binding to specific cell surface receptors, and the formation of covalent and non covalent complexes with other soluble macromolecules (Flower, 1996).

**Table 1.4** List of different Lipocalins grouped according to their known functions. Molecular mass (Mr.) oligomeric status, glycosylation status and ligand binding properties are indicated

Protein	Mr (kDa)	Oligomeric status	Glycosylation	Ligand
<i>Small hydrophobic molecule transporter</i>				
Retinol binding protein	21	Monomer	–	+
$\beta$ -lactoglobulin	18	Dimer/Monomer	–	+
<i>Pheromones</i>				
Major urinary protein	17.8	Monomer	–	+
Aphrodisin	17	?	+	?
$\alpha_{2u}$ -globulin	18.7	Dimer	–	+
<i>Olfaction &amp; Gustation</i>				
Odorant binding protein	17	Dimer	–	+
Tear prealbumin	18	Dimer	–	+
<i>Cryptic coloration</i>				
Bilin binding protein	19.6	Tetramer	–	+
$\alpha$ -crustacyanin	350	Octamer <sup>4</sup>	–	+
<i>Enzyme</i>				
Prostaglandin-D-synthase	19-32	Monomer/Dimer	+	+
<i>Immune modulation</i>				
Glycodelin	27-30	Dimer	+	–
$\alpha$ 1-microglobulin	26-33	Monomer/Dimer	+	+
Complement factor 8 $\gamma$	22	Complex	–	+
$\alpha$ 1-acid glycoprotein	41-43	Monomer	+	+
<i>Cell homeostasis</i>				
QSP <sup>5</sup>	21	?	?	?
Apolipoprotein D	29-32	Dimer	+	+

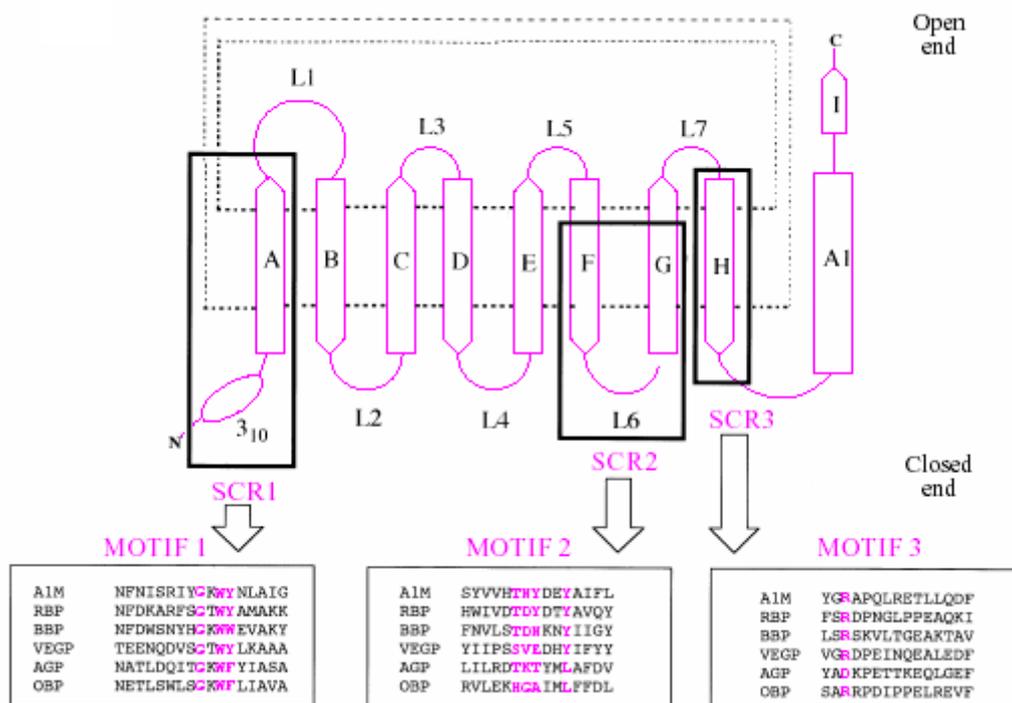
According to their known function, members of the Lipocalin family are clustered as retinol binding proteins (transport of retinoids in the circulation; RBP, purpurin), beta-lactoglobulins (major protein component of whey from the milk of many mammals), pheromones transport proteins ( $\alpha$ -2u globulin, major urinary proteins (MUP)), smell reception proteins (bovine pyrazine binding protein, odorant binding protein (OBP)), colourant proteins (Bilin-binding proteins, Crustacyanins); immune related proteins ( $\alpha$ -1-microglobulin, complement C8 $\gamma$ , neutrophil gelatinase associated Lipocalin (NGAL), Glycodelin) (Flower, 1996).

### 1.5.1 Lipocalin sequences and structure relations

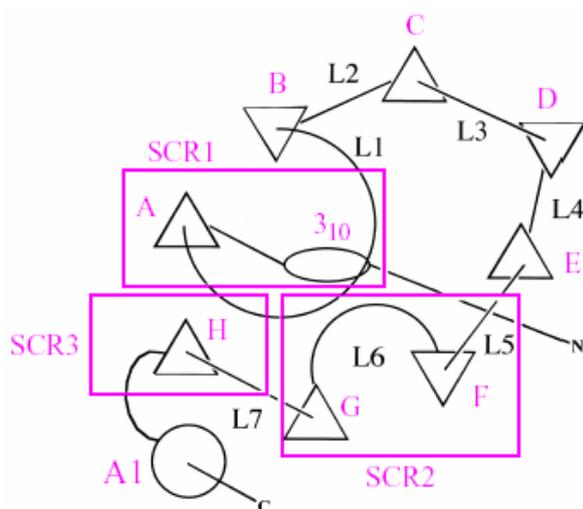
At the sequence level, the Lipocalin family is remarkably diverse. Despite many common characteristics and common functions, the Lipocalin family has been defined largely on the basis of sequence similarity. Within this the Lipocalins display unusually low levels of overall sequence conservation, with pairwise comparisons often falling well below 20 %, the threshold for a reliable alignment. However, all Lipocalins share sufficient similarity, in the form of short characteristic conserved sequence motifs (structurally conserved regions, SCRs, signed as MOTIF 1, MOTIF 2 and MOTIF 3 in Figure 1.9), to form the basis of a useful definition of family membership (Flower *et al.*, 1991, 1993): the “kernel” Lipocalins share the three characteristic conserved sequence motifs while others, defined “outlier” Lipocalins, are even more divergent family members and typically share only one or two SCR. All Lipocalins share the first of these three characteristic motifs and this feature can be used as a diagnostic of family membership.

Despite lacking high sequence similarity, Lipocalin structures are well conserved. They comprise a single eight-stranded continuously hydrogen-bonded antiparallel  $\beta$ -barrel. The common structure of the Lipocalin protein fold is now well-described (Flower *et al.*, 1993; Flower, 1995, 1996). The Lipocalin fold is a highly symmetrical all- $\beta$  protein dominated by a single eight-stranded antiparallel  $\beta$ -sheet closed back on itself to form a continuously hydrogen-bonded  $\beta$ -barrel (Figure 1.9). In cross-section, this has a flattened or elliptical shape (Figure 1.10). The  $\beta$ -barrel encloses a ligand-binding site composed of both an internal cavity and an external loop scaffold. It is this diversity of cavity and scaffold that gives rise to a variety of different binding modes each capable of accommodating ligands of different size, shape, and chemical character. The eight  $\beta$ -strands of the barrel, labelled A-H, are linked by a succession of + 1 connections, giving it the simplest possible  $\beta$ -sheet topology. These seven loops, labelled L1 to L7, are all typical of short  $\beta$ -hairpins, except loop L1: this is a large  $\Omega$  loop. Loop L1 forms a lid folded back to close partially the internal ligand-binding site found at this end of the barrel. Between strands H, and the short terminal strand I, is an  $\alpha$ -helix; this

is an ever present feature of the Lipocalin fold but is not conserved in its position relative to the axis of the  $\beta$ -barrel nor in its length.



**Figure 1.9. Structure of the Lipocalin fold** Characteristic features of the Lipocalin fold. An unwound view of the Lipocalin fold orthogonal to the axis of the barrel. The nine  $\beta$ -strands of the antiparallel  $\beta$ -sheet are shown as arrows and labelled A-I. The N-terminal 310-like helix and C-terminal  $\alpha$ -helix (labelled A1) are also marked. The hydrogen-bonded connection of two strands is indicated by a pair of dotted lines between them. Connecting loops are shown as solid lines and labelled L1-L7. One end has four  $\beta$ -hairpins (L1, L3, L5 and L7); the opening of the internal ligand-binding site is here and so is called the Open end of the molecule. The other has three  $\beta$ -hairpin loops (L2, L4 and L6); the N-terminal polypeptide chain crosses this end of the barrel to enter strand A via a conserved 310 helix affecting closure of this end of the barrel: the Closed end of the molecule. Those parts which form the three main structurally conserved regions (SCRs) of the fold, SCR1, SCR2 and SCR3, are marked as boxes. Three sequence motifs which correspond to these SCRs are shown (MOTIF 1, MOTIF 2 and MOTIF 3). The first three sequences are from kernel Lipocalins and the second three from outlier Lipocalins. Note that MOTIF 1 is well conserved in all sequences, whereas the other two, particularly MOTIF 2, are only well conserved in kernel Lipocalin sequences. This figure has been adapted from Flower, 1996.



**Figure 1.10.** The Lipocalin  $\beta$ -barrel in cross section.  $\beta$ -strands are shown as triangles.

Triangles pointing downwards (shaded gray) indicate a strand direction into the plane of the paper and those pointing upwards (empty) indicate a strand direction out of the plane of the paper. The view shown, down the axis of the barrel, is orthogonal to that in Figure 1.9. Connecting loops are shown as continuous lines. Labelling and features shown are as in Figure 1.9. This figure has been adapted from Flower, 1996.

### 1.5.2 Multiple molecular recognition properties

There are three common molecular recognition properties of Lipocalins: ligand binding, receptor binding and macromolecular complexation.

#### - Ligand binding

Lipocalins are best known for their ability to bind small hydrophobic molecules (see Appendix A.1 for a list of molecules bound by different Lipocalins), including both endogenous and exogenous ligands. This list contains molecules with critical biological functions: retinoids (retinol and retinoic acid), arachidonic acid, and various steroids. Thus the broad selectivity of binding exhibited by some Lipocalins, such as  $\beta$ lg or acid glycoprotein precursor (AGP), may reflect a general transport role, such as the clearance of unwanted endogenous, or exogenous, compounds. The structural features of the Lipocalin fold are well adapted to the task of ligand binding; the amino acid composition of the pocket and loop scaffold, as well as its overall size and conformation, determine selectivity. To accommodate ligands of different sizes and shapes, the binding sites of different Lipocalins can be quite different.

#### - Receptor binding

There is experimental evidence to show that a number of Lipocalins are bound by specific cell-surface receptors and may be internalized by receptor-mediated endocytosis, for example, it has been shown that, in the liver, the retinol-RBP complex is taken up by receptor-mediated endocytosis in parenchymal and stellate cells (Senoo *et al.*, 1990), probably involving potocytosis (Malaba *et al.*, 1995), whereas AGP is endocytosed via clathrin-coated pits (Senoo *et al.*, 1990). There is increasing evidence, from a wide variety of different tissues, that RBP binding to its target cells occurs via specific surface receptors (Bavik *et al.*, 1993; Smeland *et al.*, 1995). A cell-surface receptor for  $\alpha$ -1-microglobulin has also been identified (Fernandez-Luna *et al.*, 1998; Akerstrom *et al.*, 1990), and there is additional evidence to suggest the existence of receptors for MUP (Bocskei *et al.*, 1992),  $\beta$ lg (Papiz *et al.*, 1986; Sivaprasadarao *et al.*, 1993), and OBP (Sivaprasadarao *et al.*, 1993). Epididymal secretory protein has been shown to bind to the plasma membrane of spermatozoa (Morel *et al.*, 1993), and may be another Lipocalin to act via a specific surface receptor. It has been hypothesized that the three conserved sequence motifs characteristic of the family, which lie next to each other forming a surface patch at the closed end of the Lipocalin fold, constitute a common cell-surface receptor binding site (Flower *et al.*, 1993; North, 1989). Thus the structural determinants of binding to different membrane-bound receptors, although still poorly understood in themselves, clearly reside in different parts of the Lipocalin fold for different family members.

### - Macromolecular complexation

Lipocalins form complexes with soluble macromolecules and their molecular-recognition property are well known (Flower, 1995). Some examples of macromolecular complexation of Lipocalins are RBP with transthyretin (noncovalent), (Goodman, 1984; Monaco, 2002); purpurin with adherons, large extracellular multi-component macromolecular complexes present in cultured chick retina growth medium, (noncovalent) (Schubert & LaCorbiere, 1985); NGAL with neutrophil gelatinase (covalent), (Treibel *et al.*, 1992; Kjeldsen *et al.*, 1993); complement 8 gamma (C8) with C8 $\gamma$  (covalent, disulphide), (Haefliger *et al.*, 1991); there is a report of macromolecular complexation of Glycodelin with  $\alpha$ -2-macroglobulin (Riely *et al.*, 2000). The protein–protein interactions take place at the Open end loop scaffold of the Lipocalins. The length, conformation and variability of amino acids present in these loops determines the affinity and selectivity of these interactions.

## 1.6. Glycodelin

Glycodelin was first reported by Petrunin *et al.* in 1976. The isolated protein was named as chorionic  $\alpha$ -2-microglobulin (CAG-2). In 1980 (Joshi *et al.*, 1980) it was reported the isolation of progestagen dependent endometrial protein (PEP or PAEP) which was shown to be serologically identical to  $\alpha$ -uterine protein (AUP) (Suttcliffe *et al.*, 1982) During the same period two other groups (Bohn *et al.*, 1982; Bell *et al.*, 1982) isolated the same protein but named it placental protein 14 (PP14) and endometrial protein 15 (EP 15) respectively. The protein was then renamed as pregnancy associated secreted  $\alpha$ 2 globulin ( $\alpha$ 2-PEG) (Bell *et al.*, 1985). The name, PP14 was replaced by Glycodelin (Gd) in the year 1996 (Morris *et al.*, 1996) when it was discovered that the contraceptive function of the protein is dependent on the specific glycosylation present on the protein. According to the present nomenclature Glycodelin A (GdA) refers to the isoform present in human amniotic fluid, Glycodelin S (GdS) refers to that in human seminal plasma and Glycodelin M (GdM) refers to the protein found in hematopoietic cells (Seppälä *et al.*, 1998) in the bone marrow; the other tissues expressing Glycodelin include fallopian tubes, ovary, breast, seminal vesicle, and eccrine glands. PP14 which is equivalent to GdA is still used in the literature.

### 1.6.1 Primary sequence of Glycodelin

The N-terminal amino acid sequence of the protein was reported (Huhtala *et al.*, 1987) and it was noted that the protein is similar to  $\beta$ lg and plasma retinol binding protein (RBP). The complete amino acid sequence was reported (Julkunen *et al.*, 1988) after cloning the Glycodelin gene from endometrial complementary DNA library. The protein was called a  $\beta$ lg homologue due to its 70 % similarity to  $\beta$ lg and hence it was placed under the Lipocalin superfamily.

Structure of the Glycodelin gene was reported (Vaisse *et al.*, 1990); the gene is 5.05 kb long having seven exons separated by six introns, the exact organization found in  $\beta$ lg. Sequencing of the promoter region of the Glycodelin gene revealed that there is consensus glucocorticoid response element (Bolton *et al.*, 1987) adding support to the progesterone inducible expression of the protein. Further, there is a short gene duplication of 400 base pair (bp) lying at position -2,660. This duplication is homologous to 100 bp of exon 4 and 300 bp of intron 4, including 180 bp corresponding exactly to the right arm of an Alu element lying on the complementary strand. This homology suggests that this duplication may have arisen through a retrotransposition event.

The putative functions of Glycodelin reported are:

- a. immunosuppressive (Bolton *et al.*, 1987); the protein was reported to inhibit phytohemagglutinin induced proliferation of human peripheral blood mononuclear cells (PBMCs) as well as in mixed lymphocyte reaction. Subsequently the protein was reported to suppress natural killer (NK) cell activity (Okamoto *et al.*, 1991).
- b. contraceptive (Oehninger *et al.*, 1995); the protein was reported to inhibit sperm binding to zona pellucida.

Two more functions have been ascribed to this protein recently e.g. morphogenic (Kamarainen *et al.*, 1997) and angiogenic (Song *et al.*, 2001). In 1996 Morris *et al.* reported that the glycosylation present on GdA is responsible for its contraceptive function and that GdS is not contraceptive. It is generally assumed that GdS is also immunosuppressive and adds to the immunosuppressive nature of seminal plasma.

Glycodelin is also reported to be an endometrial marker, as there is a clinical correlation between Glycodelin level in the endometrium and habitual abortion (Tulppala *et al.*, 1995). As the protein is synthesized under the control of progesterone the level of this protein can indicate endometrial sufficiency. Apart from general endometrial health the level of the

protein has been reported to be elevated in the serum in case of endometrial malignancies (Tatarinov *et al.*, 1990) and in ectopic pregnancy (Stabile *et al.*, 1994).

### 1.6.2 Glycodelin gene expression

Glycodelin is synthesized by the endometrium under the control of progesterone. As expected the protein is expressed by the endometrium during the secretory phase of the cycle and during pregnancy in concert with progesterone secretion.

The protein is also detectable in the serum during these periods. The circulatory level of Glycodelin is two orders lower than the endometrial or amniotic fluid level, thus it never reaches the concentration that would cause systemic immunosuppression in the mother (Figure 1.11).

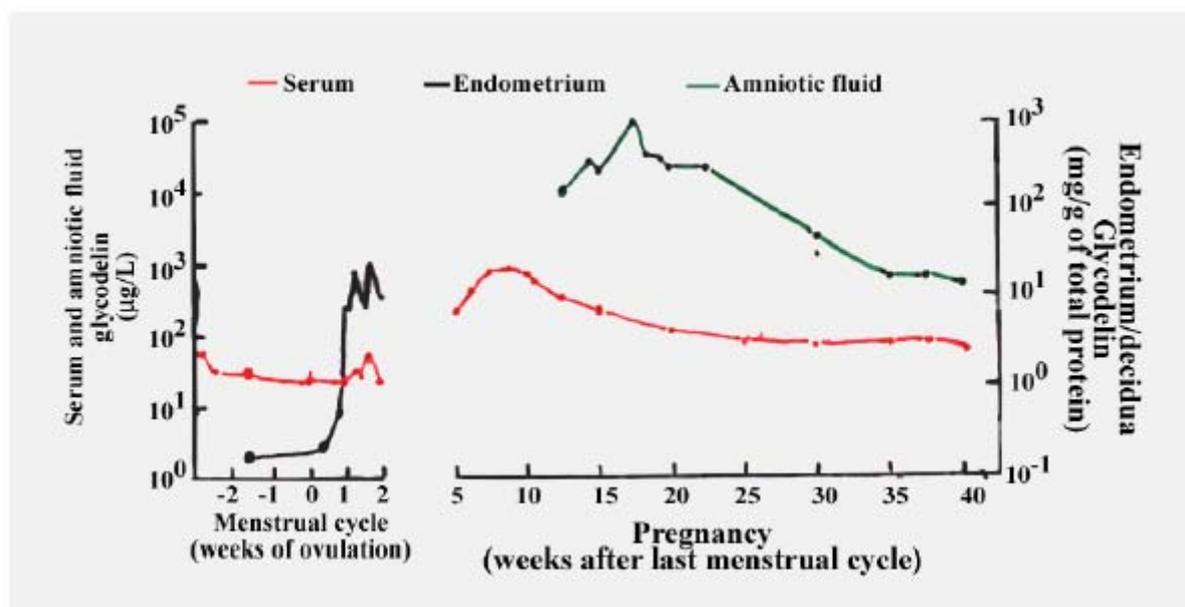


Figure 1.11. GdA levels in the serum, amniotic fluid and endometrial tissue during normal cycle and pregnancy. This figure has been adapted from Seppälä, 1997.

Molecular details of the gene expression by progesterone are still unclear. A recent study has reported that two of the three putative Sp1 sites present in Glycodelin promoter region (-1900 bp to +30 bp) are important for human progesterone receptor (hPR) mediated GdA expression (Gao *et al.*, 2001). There is a glucocorticoid response element in the Glycodelin promoter which can be modulated directly by progestins and antiprogestins (Song *et al.*, 2001). Estrogen has no effect on the synthesis of Glycodelin and the effect of relaxin on the production of this protein is controversial (Taylor *et al.*, 2000; Stewart *et al.*, 1997). There is a report on regulation of Glycodelin gene expression by chorionic gonadotrophin (CG) (Hausermann *et al.*,

1998). Higher level of GdA expression in case of endometrial malignancies can be a useful marker of the disease (Horowitz *et al.*, 2001).

### 1.6.3 Isoforms of Glycodelin

There are two kinds of isoforms possible for Glycodelin, splice variants and differentially glycosylated full length protein. Though splice variants are reported at the mRNA level for hematopoietic Glycodelin (Morrow *et al.*, 1994) and GdS (Koistien *et al.*, 1997), they are not detected at the protein level for GdA. More remarkable are the isoforms generated by differential glycosylation of the protein as in the case of GdA, GdS and zona inhibitory factor (ZIF, a recently identified variant of GdA, Chiu *et al.*, 2003). The nature of complex glycosylation present on Glycodelin A and S has been studied extensively (Dell *et al.*, 1995, Koistinen *et al.*, 1996) (see Figure 1.12 for a comparative list of glycosylations). Native Glycodelin is glycosylated at N28 and N63 (another putative glycosylation site N85 is not glycosylated in the native protein). Both the glycosylation sites in GdA have complex glycosyl groups whereas only N63 in GdS has complex glycosylation. N28 glycosylation in GdS contains branched 5–7 mannose residues. All the different forms of complex glycosylation on GdA are sialylated and contain 1–2 fucose residues, whereas, GdS glycosylation (N63) does not contain any sialic acid. Instead it is rich in fucose residues.

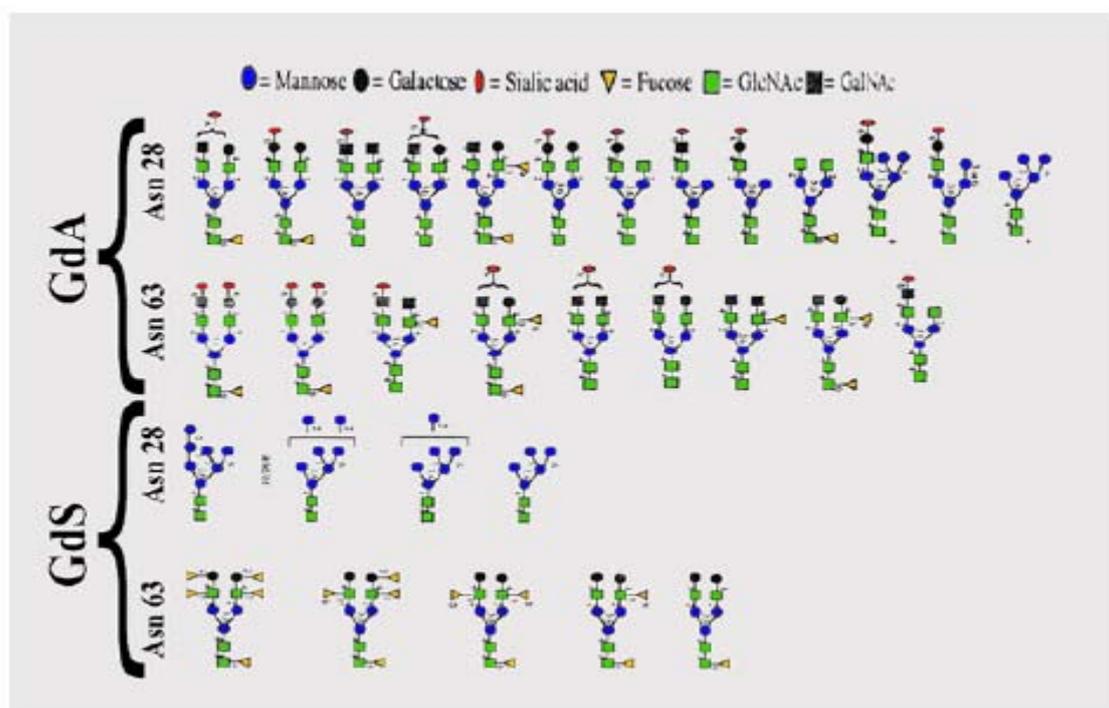


Figure 1.12. Different glycosylation patterns observed in GdA and GdS. All the oligosaccharides present on GdA have sialic acid residues but few fucose residues. N28 glycans on GdS are composed of 5–7 mannose residues only, whereas N63 is rich in fucose residues. This figure has been adapted from Dell, 1995.

It was proposed on Glycodelin (Koistinen *et al.*, 1999), that native folding of Glycodelin-A and Glycodelin-S is not influenced by the differences in glycosylation or by ligand binding, based on thermodynamic parameters of reversible denaturation. Recently, a close analyses of the glycans associated with the Gd molecules from various sources suggested that the activity (e.g. apoptogenic) of Gd lies in the protein backbone and the glycans modulate the activity by masking (as in case of GdS) or unmasking (as in case of GdA), the functional region of the molecule (Karande *et al.*, 2005).

## 2. Methods for Structural Studies

In this chapter will be shortly presented the Nuclear Magnetic Resonance spectroscopy (NMR), a powerful technique for structural studies of proteins in solution. NMR can solve protein structures in three dimensions and it can be useful for dynamics studies and for providing many other useful information.

### 2.1 The Nuclear Magnetic Resonance in Protein Studies

Nuclear Magnetic Resonance (NMR) spectroscopy is based on the fact that atomic nuclei oriented by a strong magnetic field (2-14 Tesla) absorb radiation at characteristic frequencies (typically a few hundred megahertz). The usefulness of NMR to the chemist and biologist results largely from the fact that nuclei of the same element in different environments give rise to distinct spectral lines. This makes it possible to observe signals from individual atoms even in complex biological macromolecules in solution. The parameters that can be measured from the resulting spectra can be interpreted in terms of molecular structure, conformation, and dynamics.

A very wide range of different elements have nuclei that are amenable to be studied by NMR spectroscopy. The nucleus which is most sensitive to detection by NMR is that of hydrogen ( $^1\text{H}$ , the proton) and this is by far the most important nucleus for the study of biological molecules.

The first published NMR spectrum of a biological macromolecule was the 40 MHz  $^1\text{H}$  spectrum of pancreatic ribonuclease reported in 1957 (Saunders *et al.*, 1957). The subsequent years, perhaps the last ten years, have seen astonishing development in instrumentation and methodology which have enormously increased the power of NMR, notably in its application to studies of conformations and interactions of biological molecules.

In structural and dynamical characterization of macromolecules by NMR, the four most important physical phenomena are:

- chemical shift
- scalar coupling
- relaxation
- Nuclear Overhauser effect

A brief overview of these four phenomena is given below together with the basic 1D, 2D and 3D experiments used in protein NMR. Included is also a description of the sequential assignment procedure. The assignment of resonances to individual nuclei is an essential first step in any NMR study, and this has been made much easier by the advent of multidimensional NMR experiments. Nonetheless, for all but the smallest macromolecules this is still the rate-limiting step. The main problem in the NMR analysis of proteins bigger than 120 residues, is the overlap of the signals in the  $^1\text{H}$  spectrum. The most powerful approach to overcome the overlap problem is the isotope labelling with  $^{13}\text{C}$  and  $^{15}\text{N}$ , which, in conjunction with multidimensional heteronuclear NMR, is crucial in extending the usefulness of NMR to larger molecules. The isotope labelling, together with the provision of the quantity of proteins (in the order of ten milligrams) required for NMR spectroscopy, has been possible thanks to the developments in molecular genetics and the ability to construct over-expression systems. In 3D NMR, the two-dimensional  $^1\text{H}$  spectra are “spread out” in a third dimension by the  $^{15}\text{N}/^{13}\text{C}$  chemical shift. Individual “slices” at a particular  $^{13}\text{C}$  or  $^{15}\text{N}$  resonance are thus much less crowded and easier to analyse.

In the last years several NMR research groups, including our group in Verona, have equipped a molecular biology laboratory and devoted a lot of energy to production of recombinant labelled proteins, as nowadays molecular and structural biology are strongly interconnected.

## 2.2 The NMR Phenomenon

When a magnetic nucleus of spin  $I$  is placed in a magnetic field  $B_0$ , it adopts one of the  $2I+1$  number of allowed orientations of different energy. For a  $\frac{1}{2}$ -spin nucleus ( $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ), the permitted orientations are two: the magnetic moment can point in the same direction as the field or in the opposite direction. These two states are separated by an energy  $\Delta E$ , which depends on the strength of the interaction between the nucleus and the field:

$$\Delta E = h\nu = \hbar\gamma B_0 \quad (2.1)$$

where  $B_0$  is the strength of the magnetic field,  $h$  is the Planck's constant and  $\gamma$  is the gyromagnetic ratio. The frequency

$$\nu_0 = \frac{\gamma B_0}{2\pi} \quad \text{or} \quad \omega_0 = \gamma B_0 \quad (2.2)$$

is the Larmor frequency.

$\Delta E$  may be measured by applying a radiofrequency of frequency  $\nu_0$ , which causes nuclei to flip from the lower level to the upper one. This is the nuclear magnetic resonance spectroscopy.

There are two possible representation of the NMR phenomenon: the vector model and the product operator formalism. The vector model is a description of basic NMR experiments based on the classical physics that has applicability for simple spin systems. A more powerful formalism is based on the quantum mechanical product operator formalism, that allows the description of more complex experiments and spectra. In the next paragraphs, the vector model will be used.

### 2.2.1 Thermal equilibrium

When placed in a magnetic field  $B_0$ , a collection of magnetic nuclei spread themselves among the  $2I+1$  available levels according to the Boltzmann distribution. For a  $\frac{1}{2}$ -spin nucleus, the populations  $\eta$  of the two levels are:

$$\frac{\eta_{upper}}{\eta_{lower}} = e^{-\Delta E/KT} \quad (2.3)$$

where  $K$  is the Boltzmann's constant and  $T$  is the temperature.

At the thermal equilibrium, since the lower level is slightly more populated than the upper level, the sample has a net magnetization  $\mathbf{M}$  in the direction of the magnetic field  $\mathbf{B}_0$ . Usually  $\mathbf{B}_0$  is assumed to be directed as the  $z$  axis, therefore the net magnetization  $\mathbf{M}$  is oriented along  $z$ , with magnitude  $M_0$ :

$$M_0 = \frac{1}{2} \gamma \hbar \Delta n_{eq} \quad (2.4)$$

where  $\Delta n_{eq}$  is the population difference at the thermal equilibrium.

In the perpendicular direction  $x$  and  $y$ , the phases of the individual nuclear magnetic moments are random, because there is no transverse magnetic field to align them and their vector sum vanishes.

In the vector model, the motion of  $\mathbf{M}$  is a precession around the field direction of  $\mathbf{B}_0$  at the Larmor angular frequency  $\omega_0$ .

### 2.2.2 Radiofrequency pulses

The simplest NMR experiment involves applying a single, short, intense radiofrequency radiation, equivalent to an oscillating magnetic field  $\mathbf{B}_1$ , to a sample, previously at thermal equilibrium. The transmitter frequency  $\omega_{rf}$  is set at the Larmor frequency. Choosing the

phase of the radiofrequency such that  $\mathbf{B}_1$  lies along the  $x$  axis, the pulse cause  $\mathbf{M}$  to precess in the  $yz$  plane at angular frequency  $\gamma B_1$ . The angle through which the magnetization turns is called flip angle. Different flip angles may be produced by setting appropriately the duration of the radiofrequency pulse. The most common used pulse have  $90^\circ$  or  $180^\circ$  flip angles: a  $90^\circ$  pulse rotates  $\mathbf{M}$  from the  $z$  axis to the  $y$  axis, while a  $180^\circ$  pulse inverts  $\mathbf{M}$  leaving it along the negative  $z$  axis.

In terms of population difference, a  $180^\circ$  pulse inverts the population difference  $\Delta n$ , while a  $90^\circ$  pulse equalises the two populations, at the same time converting the equilibrium magnetization entirely into  $y$  magnetization,  $M_y$ . That is, the radiofrequency pulse induces a coherence amongst the spins such that the orientations of the individual magnetic moments in the  $xy$  plane are no longer random.

### 2.2.3 Free precession, relaxation and NMR spectrum

Once the pulse has been switched off, the only field remaining is the static magnetic field  $\mathbf{B}_0$  along the  $z$  axis.  $\mathbf{M}$  therefore precesses in the  $xy$  plane around the  $z$  axis and the spins return to the equilibrium condition through the relaxation processes. There are two distinct relaxation processes. First, the recovery of the  $z$  magnetization to its equilibrium value, known as spin-lattice relaxation which occurs with an exponential time constant  $T_1$ . Second, the decay to zero of the  $xy$  magnetization, known as spin-spin relaxation, which occurs with an exponential time constant  $T_2$ .

The oscillating, decaying transverse magnetization is detected by the NMR spectrometer. The signal, known as the free induction decay (FID), is the sum of individual frequencies from the various nuclei in the sample, each with characteristic amplitude and  $T_2$ . It contains all the information necessary to obtain an NMR spectrum. All these oscillating components in the FID are unravelled by means of the Fourier transform.

## 2.3 Chemical Shift

The NMR frequency of a nucleus in a molecule is determined by its gyromagnetic ratio  $\gamma$  and the strength  $B_0$  of the magnetic field it experiences (equation 2.2). Thus proton and  $^{13}\text{C}$  nuclei resonate respectively at 500 and 125 MHz in a 11.7 Tesla field. But not all protons, nor all  $^{13}\text{C}$  nuclei, have identical resonance frequencies:  $\nu$  depends on the local electron distribution. This effect is called *chemical shift*. The chemical shift is generally defined as the difference, in resonance frequencies, between the nucleus of interest ( $\nu$ ) and a reference nucleus ( $\nu_{\text{ref}}$ ),

$$\delta = \frac{\nu - \nu_{\text{ref}}}{\nu_{\text{ref}}} \quad (2.5)$$

The frequency difference  $\nu - \nu_{\text{ref}}$  is divided by  $\nu_{\text{ref}}$  so that  $\delta$  is a molecular property, independent of the magnetic field used to measure it;  $\delta$  values are quoted as parts per millions, or ppm.

In NMR terminology, a resonance is at high field (up-field) relative to another resonance if it has a smaller  $\delta$  in ppm values.

The chemical shift is a fundamental parameter in protein NMR as it gives separately detectable signals for the hundreds of protons that can therefore be distinguished and assigned. Chemical shift dispersion arises because interior peptide segments in globular proteins are shielded from the solvent and are nearest neighbours to other peptide segments, so that different residues experience different microenvironments. Chemical shift dispersion can also arise for protons within the same residue, due to different local electron distribution.

The characteristic  $^1\text{H}$  chemical shifts for amino acid residues of random extended chain structures are given in Table 2.1 (Wütrich, 1986). The data were obtained from NMR measurements in aqueous solution of 20 tetrapeptides H-Gly-Gly-Xxx-Ala-OH, where in each peptide Xxx was a different one of the 20 common amino acid. The hydrogen atoms with similar shifts are grouped in Table 2.2.

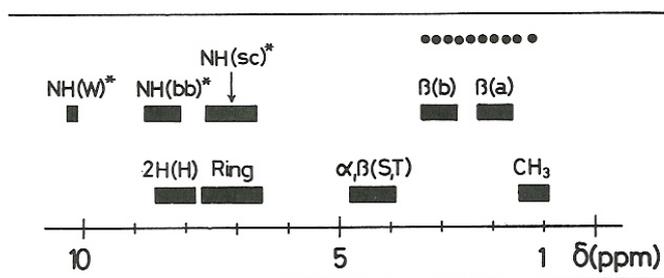
It is well established (Wishart *et al.*, 1991) that chemical shifts deviations from the random coil values are strongly correlated with protein secondary structures. In particular, the chemical shifts of NH and  $\text{H}_\alpha$  can be indicative of regular secondary structures. In helices NH and  $\text{H}_\alpha$  have chemical shifts up-field ( $\approx 0.1$  ppm for NH;  $\approx 0.39$  ppm for  $\text{H}_\alpha$ ) from the average for extended chain conformations, and in  $\beta$ -sheets resonate at lower field.

**Table 2.1** Random coil  $^1\text{H}$  chemical shifts for the 20 amino acid residues.

Residue	NH	$\alpha\text{H}$	$\beta\text{H}$	others
Gly	8.39	3.97		
Ala	8.25	4.35	1.39	
Val	8.44	4.18	2.13	$\gamma\text{CH}_3$ 0.97, 0.94
Ile	8.19	4.23	1.90	$\gamma\text{CH}_2$ 1.48, 1.19 $\gamma\text{CH}_3$ 0.95 $\delta\text{CH}_3$ 0.89
Leu	8.42	4.38	1.65, 1.65	$\gamma\text{H}$ 1.64 $\delta\text{CH}_3$ 0.94, 0.90
Pro		4.44	2.28, 2.02	$\gamma\text{CH}_2$ 2.03, 2.03 $\delta\text{CH}_2$ 3.68, 3.65
Ser	8.38	4.50	3.88, 3.88	
Thr	8.24	4.35	4.22	$\gamma\text{CH}_3$ 1.23
Asp	8.41	4.76	2.84, 2.75	
Residue	NH	$\alpha\text{H}$	$\beta\text{H}$	others

Glu	8.37	4.29	2.09, 1.97	$\gamma$ CH <sub>2</sub> 2.31, 2.28
Lys	8.41	4.36	1.85, 1.76	$\gamma$ CH <sub>2</sub> 1.45, 1.45 $\delta$ CH <sub>2</sub> 1.70, 1.70 $\epsilon$ CH <sub>2</sub> 3.02, 3.02 $\epsilon$ NH <sub>3</sub> 7.52
Arg	8.27	4.38	1.89, 1.79	$\gamma$ CH <sub>2</sub> 1.70, 1.70 $\delta$ CH <sub>2</sub> 3.32, 3.32 NH 7.17, 6.62
Asn	8.75	4.75	2.83, 2.75	$\gamma$ NH <sub>2</sub> 7.59, 6.91
Gln	8.41	4.37	2.13, 2.01	$\gamma$ CH <sub>2</sub> 2.38, 2.38 $\delta$ NH <sub>2</sub> 6.87, 7.59
Met	8.42	4.52	2.15, 2.01	$\gamma$ CH <sub>2</sub> 2.64, 2.64 $\epsilon$ CH <sub>3</sub> 2.13
Cys	8.31	4.69	3.28, 2.96	
Trp	8.09	4.70	3.32, 3.19	2H 7.24 4H 7.65 5H 7.17 6H 7.24 7H 7.50 NH 10.22
Phe	8.23	4.66	3.22, 2.99	2,6H 7.30 3,5H 7.39 4H 7.34
Tyr	8.18	4.60	3.13, 2.92	2,6H 7.15 3,5H 6.86
His	8.41	4.63	3.26, 3.20	2H 8.12 4H 7.14

**Table 2.2** Groups of hydrogen atoms in the common amino acid residues with similar random coil  $^1\text{H}$  chemical shifts. (The figure has been reproduced from Wütrich, 1986.)



Code	$\delta$ (ppm)	Comments
$\text{CH}_3$	0.9–1.4	
$\beta$ (a)	1.6–2.3	$\beta\text{H}$ of V, I, L, E, Q, M, P, R, K
$\beta$ (b)	2.7–3.3	$\beta\text{H}$ of C, D, N, F, Y, H, W
•••••	1.2–3.3	Other Aliphatic CH
$\alpha, \beta$ (S, T)	3.9–4.8	All $\alpha\text{H}$ , $\beta\text{H}$ of S and T
Ring	6.5–7.7	Aromatic CH of F, Y, W; 4H of H
2H(H)	7.7–8.6	2H of H in the pH range 1–11
NH(sc) *	6.6–7.6	Side Chain NH of N, Q, K, R
NH(bb) *	8.1–8.8	Backbone NH
NH(W) *	10.2	Indole NH of W

\* In model peptides the labile protons (identified by \*) are only observed in  $\text{H}_2\text{O}$  solution.

## 2.4 Scalar Coupling

The scalar coupling, known also as spin-spin coupling, or J-coupling, is a magnetic interaction between nuclei linked via a small number of covalent bonds in a chemical structure. The scalar coupling is characterised by the spin-spin coupling constant  $J$ .  $J$  values are field independent and quoted in Hz. Spin-spin coupling is manifested in the spectrum by the partially fine structure of the individual resonance lines.

For example, a nucleus coupled to a neighbour with spin  $\frac{1}{2}$  has its resonance line split into a doublet. The space between the two lines equals the  $J$  constant. The general rule is that the number of lines in a multiplet is  $2I+1$ , where  $I$  is the spin of the neighbour.

## 2.5 The Relaxation

In NMR experiments, radiofrequency pulses disturb the equilibrium of the spin systems, as described previously. Relaxation is the process by which equilibrium is regained through interactions of the spins with the thermal molecular environment.

The two relaxation processes, the spin-lattice relaxation and the spin-spin relaxation, are reviewed in this chapter together with a description of the applications of relaxation experiments in the protein study.

### 2.5.1 Dipole-dipole interaction and rotational motion in liquid

The mechanism of nuclear spin relaxation lies in magnetic interactions, the most important being dipolar coupling. Every nucleus with non-zero spin quantum number has a magnetic dipole and therefore behaves like a small bar magnet, producing a local magnetic field,  $B_\mu$ . The interaction of a nucleus with neighbours nuclei through the local magnetic fields produced by the magnetic moments is known as dipolar coupling. The dipolar coupling between two nuclei depends on the separation  $r$  and on  $\theta$ , the angle between the internuclear vector and the static field. As the molecules translate, rotate and vibrate,  $r$  and  $\theta$  vary in a complicated way causing the interaction to fluctuate rapidly. Thus the dipolar coupling, modulated by molecular motions, causes nuclear spins to experience time-dependent local magnetic fields, which, if they contain a component at the Larmor frequency, can induce the transition which return spins to equilibrium.

The frequency with which the local magnetic fields  $B_\mu$  fluctuate depends on the correlation time  $\tau_c$ . The correlation time indicates how long  $B_\mu$  maintains the same position and orientation before changing them due to the random collisions among the molecules in the solution. Rapid fluctuations have a small value of  $\tau_c$ , while slow fluctuations have a large value of  $\tau_c$ . The spectrum of the molecular motions, that depends on the value of  $\tau_c$ , is described by the spectral density  $J(\omega)$  (Figure 2.1) ( $\omega$  is the angular frequency in radian/s).

$J(\omega)$  can be interpreted as the probability of finding a component of the random motion at a particular frequency. The integral of  $J(\omega)$  over all frequencies is a constant, independent of  $\tau_c$ . The relation between  $J(\omega)$  and  $\tau_c$  is:

$$J(\omega) = \frac{2\tau_c}{1 + \omega^2\tau_c^2} \quad (2.6)$$

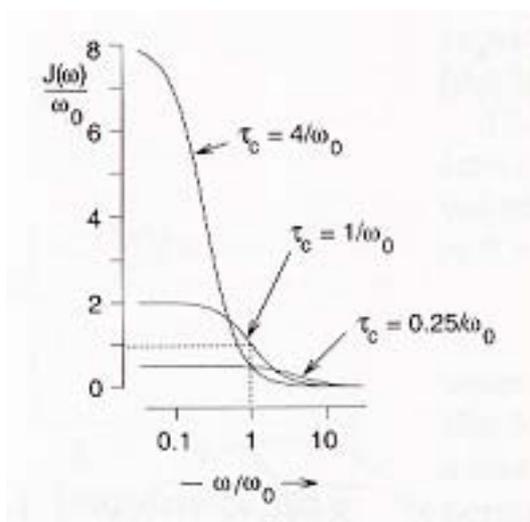


Figure 2.1. The spectral density  $J(\omega)$  drawn for three values of the correlation time  $\tau_c$ . (The figure has been reproduced from Hore, 1995.)

## 2.5.2 Spin-lattice relaxation

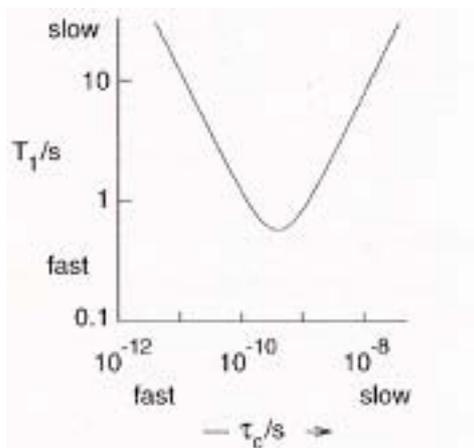
Spin-lattice relaxation is caused by fluctuating local fields which induce nuclei to flip amongst their available spin states. The rate of this process,  $T_1^{-1}$ , depends on the probability that the local field has a component oscillating at the appropriate frequency, namely the Larmor frequency  $\omega_0$ .  $T_1^{-1}$  is proportional to the spectral density  $J(\omega_0)$  as follows:

$$\frac{1}{T_1} = \gamma^2 \langle B_{\mu}^2 \rangle J(\omega_0) \quad (2.7)$$

where  $\langle B_{\mu}^2 \rangle$  is the mean square value of the local field.

Figure 2.1 shows that  $J(\omega_0)$  is small for  $\tau_c^{-1}$  much smaller than  $\omega_0$ , or much larger than  $\omega_0$  and reaches a maximum when  $\tau_c^{-1}$  matches the Larmor frequency ( $\omega_0\tau_c = 1$ ). This behaviour is summarised in Figure 2.2. For rapidly tumbling molecules with  $\omega_0\tau_c \ll 1$  (left-hand side of Figure 2.2),  $J(\omega_0) \approx 2/\tau_c$  and the relaxation gets slower as the mean tumbling rate is increased. Conversely, slowly tumbling molecules have  $\omega_0\tau_c \gg 1$  (right-hand side of Figure 2.2) and  $J(\omega_0) \approx 2/\omega_0^2\tau_c$ , so that the relaxation accelerates as the tumbling speeds up. The maximum relaxation rate occurs for  $\omega_0\tau_c = 1$  (minimum  $T_1$ ), at which point  $J(\omega_0) = 1/\omega_0$ . For macromolecules that rotate slowly relative to  $\omega_0$ , so that  $\omega_0\tau_c \gg 1$ ,  $T_1$  increases proportionally to  $\tau_c$ .

Physically, spin-lattice relaxation couples the spins (very weakly) to the motion of the molecule that carry them and so provides a pathway for the exchange of energy between the spin system and its surroundings. The energy absorbed or released in the course of the spin relaxation is transferred from or to the motions of the molecules, causing a slight cooling or warming of the lattice. Since the spin energies are small compared to the rotational, vibrational and translational energy of the molecules in solution, nuclear spins are relaxed with a non measurable small change in the temperature of the sample.



**Figure 2.2.** The dependence of the spin-lattice relaxation time  $T_1$  on the correlation time  $\tau_c$ . The regions of the graph corresponding to fast and slow tumbling and fast and slow relaxation are indicated. (The figure has been reproduced from Hore, 1995.)

### 2.5.3 Spin-spin relaxation

The spin-spin relaxation is the process through which the  $xy$  magnetization created by a radiofrequency pulse decay to zero by the randomisation of the individual spins. The local fluctuating magnetic fields  $B_\mu$  are responsible of the  $T_2$  relaxation. These fields indeed produce small time-dependent variations in the precession frequencies of individual spins which lead to loss of phase-coherence in the sample.

The relation between  $T_2$  and the spectral density  $J(\omega)$  is:

$$\frac{1}{T_2} = \frac{1}{2}\gamma^2 \langle B_\mu^2 \rangle J(\omega_0) + \frac{1}{2}\gamma^2 \langle B_\mu^2 \rangle J(0) \quad (2.8)$$

The motional dependence of  $T_2$  is shown in Figure 2.3, together with the  $T_1$  behaviour.  $T_2$  increases as the tumbling gets faster, while in the slow motion limit ( $\omega_0\tau_c \gg 1$ ), typical of macromolecules, is simply inversely proportional to the correlation time. The two relaxation times,  $T_1$  and  $T_2$ , are identical in the extreme narrowing conditions  $\omega_0\tau_c \ll 1$ .

In general,  $T_1$  is greater than  $T_2$  and this is because there are additional causes of loss of transverse magnetization, the main of whose is that the static magnetic field  $B_0$  is not uniform throughout the sample. Thus, if we divide the sample in small regions such that the field is uniform in regions known as isochromats, then the total magnetization is the sum of all these regions, each of which contributes a precessing vector, which differs slightly in frequency. Consequently the transverse magnetization of the sample as a whole is reduced and the transverse relaxation is faster ( $T_2$  is shorter).

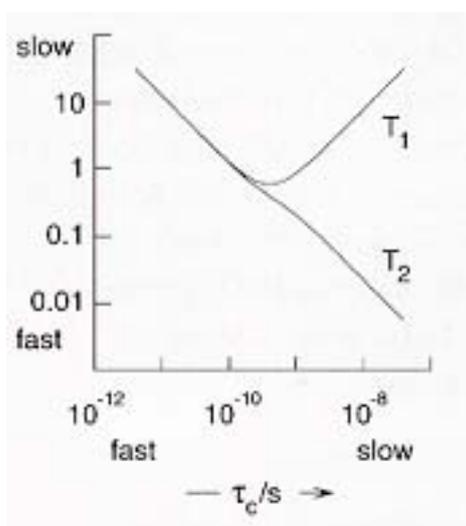


Figure 2.3. Dependence of  $T_1$  and  $T_2$  on correlation time  $\tau_c$ . The regions of the graph corresponding to fast and slow tumbling and fast and slow relaxation are indicated. (The figure has been reproduced from Hore, 1995.)

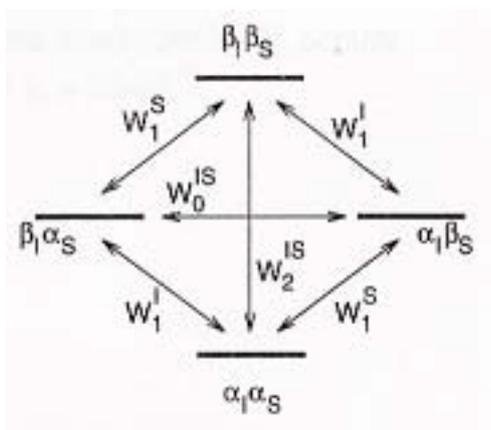
### 2.5.4 Nuclear Overhauser effect

The dipole-dipole interaction between nuclei close together in the space causes a particular relaxation path known as cross relaxation that is the origin of the Nuclear Overhauser effect (NOE). The NOE is the fractional change in intensity of one NMR line when another resonance is selectively perturbed. A description of the phenomenon is given below.

Consider a pair of dipolar coupled  $\frac{1}{2}$  spins, I and S. Denoting the  $+1/2$  and  $-1/2$  spin state of each spin with  $\alpha$  and  $\beta$  respectively, the four possible combination of the two spins I and S are usually represented as  $\alpha_I\alpha_S$ ,  $\alpha_I\beta_S$ ,  $\beta_I\alpha_S$  and  $\beta_I\beta_S$  (Figure 2.4). At thermal equilibrium, the relative population of the four states can be described, to a good approximation, by solving equation 2.3, as  $1+2\Delta$  ( $\alpha_I\alpha_S$ ),  $1-2\Delta$  ( $\beta_I\beta_S$ ),  $1$  ( $\alpha_I\beta_S$ ) and  $1$  ( $\beta_I\alpha_S$ ), where  $\Delta = \frac{1}{2}\hbar\omega_0/kT$ .

In Figure 2.4, six relaxation pathways are indicated:

- $W_1^I$  and  $W_1^S$  correspond to the single spin flipping  $\alpha_I \leftrightarrow \beta_I$  and  $\alpha_S \leftrightarrow \beta_S$  and are the spin-lattice relaxation processes.
- $W_0^{IS}$  and  $W_2^{IS}$  correspond to the simultaneous flipping of both I and S  $\alpha_I\alpha_S \leftrightarrow \beta_I\beta_S$  (both spins flipping in the same direction) and  $\alpha_I\beta_S \leftrightarrow \beta_I\alpha_S$  (I and S flipping in opposite directions). These relaxation processes are known as cross relaxation.



**Figure 2.4.** Energy levels for a pair of  $\frac{1}{2}$  spins nuclei I and S, showing the six possible relaxation pathways. (The figure has been reproduced from Hore, 1995.)

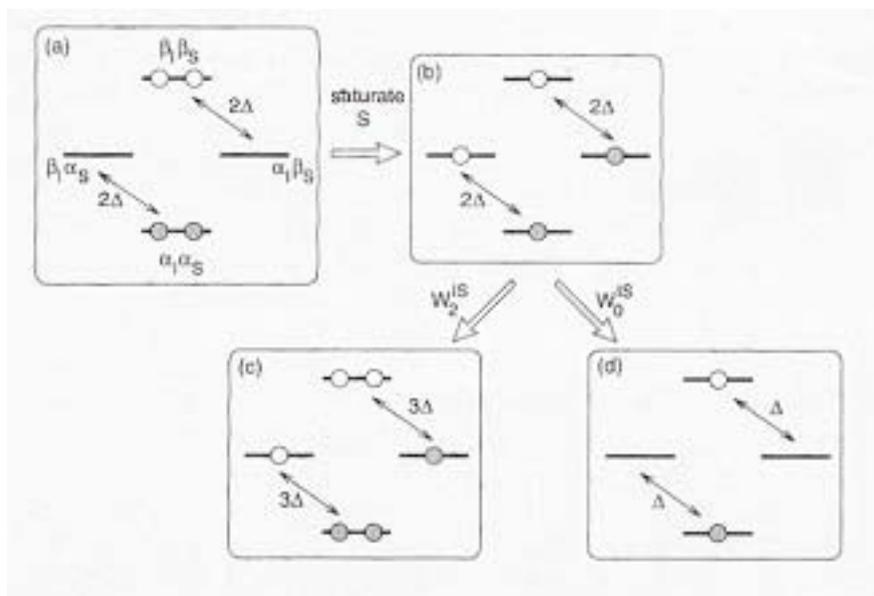
Cross relaxation comes about because the chaotic molecular motion, combined with the mutual dipolar interaction, causes the fluctuating local fields experienced by I and S to be correlated, with the results that the nuclei can undergo simultaneous spin-flips. The  $W_0^{IS}$  and  $W_2^{IS}$  processes are extra pathways that allow the spin state populations to return to equilibrium following some disturbance.

The NOE effect can be explained through the following experiment. Imagine that the S transitions are saturated, i.e. the population  $\alpha_I\alpha_S$  and  $\alpha_I\beta_S$ , and the population  $\beta_I\alpha_S$  and  $\beta_I\beta_S$

are equalised, by the application of an appropriate radiofrequency field. This has no effect on the population difference across the I transitions ( $\alpha_I\alpha_S \leftrightarrow \beta_I\alpha_S$  and  $\alpha_I\beta_S \leftrightarrow \beta_I\beta_S$ ).

Two different processes are now analysed (Figure 2.5):

1. All the relaxation pathways are insignificant except  $W_2^{IS}$ . This relaxation route transfers population between  $\alpha_I\alpha_S$  and  $\beta_I\beta_S$  and restores the equilibrium population of these two states,  $1+2\Delta$  and  $1-2\Delta$ . The population difference across the I transitions is now  $3\Delta$ , and therefore the intensity of the I signal has increased by 50%. Cross relaxation has transferred magnetization from the saturated spin S to its dipolar-coupled spin I.
2. All the relaxation pathways are insignificant except  $W_0^{IS}$ . This relaxation route transfers population between  $\beta_I\alpha_S$  and  $\alpha_I\beta_S$  and restores the equilibrium population of these two states (both unity). The population difference across the I transitions is now  $\Delta$ , and therefore the intensity of the I signal has decreased by 50%.



**Figure 2.5.** Spin state populations for a pair of neighbouring  $\frac{1}{2}$  spin nuclei I and S. Shaded circles indicate a population excess of  $\Delta$ ; open circles, a population deficit of  $\Delta$ . (a) Thermal equilibrium. (b) Effect of saturating both transitions of spin S. (c) Effect of  $W_2^{IS}$  cross relaxation. (d) Effect of  $W_0^{IS}$  cross relaxation. (The figure has been reproduced from Hore, 1995.)

The NOE can be quantified by the parameter  $\eta$ :

$$\eta = \frac{i - i_0}{i_0} \quad (2.9)$$

where  $i$  is the perturbed NMR intensity of the spin I and  $i_0$  its normal intensity.

The maximum homonuclear NOE can be  $\frac{1}{2}$  and the minimum  $-1$ . In reality, neither  $W_2^{IS}$  nor  $W_0^{IS}$  dominates the other relaxation pathways and  $\eta$  is somewhere between the extremes. The parameter  $\eta$  has the same sign as  $W_2^{IS} - W_0^{IS}$ .

Both the relaxation rates  $W_2^{IS}$  and  $W_0^{IS}$  are related to the spectral density  $J(\omega)$  as follows:

$$W_2^{IS} \approx J(2\omega_0) \quad \text{and} \quad W_0^{IS} \approx J(0) \quad (2.10)$$

Without giving all the mathematical details, it occurs that the proton-proton NOE,  $\eta$ , should be positive for fast motions ( $\omega_0\tau_c \ll 1$ ) and negative for slow motions ( $\omega_0\tau_c \gg 1$ ). The change of sign occurs when  $W_2^{IS} = W_0^{IS}$ , at which point the effect of the two cross relaxation pathways cancel; this happens when  $\omega_0\tau_c \approx 1$ .

NOE are also observable for heteronuclear pairs of spins.

### 2.5.5 NOE in protein structure determination

The NOE  $\eta$  is exceedingly useful as a source of molecular structure information as its intensity is correlated with the inverse sixth power of the internuclear distance between the two dipolar coupled spins. The NOE is the base of the protein structure determination by NMR.

In the NOESY-type experiment, that will be described below, many hundreds of NOEs between pairs of protons distant in space less than  $5\text{\AA}$  are detected. Some of these will link protons in the same residue, others protons of neighbouring residues and others will connect protons in very different parts of the molecule. From the intensity of the signal, it is possible to extract the distance between the two interacting spins. If there are enough NOEs signals distributed throughout the protein, then the three-dimensional structure can be defined by using computer algorithms that search for the conformations that satisfy all the NOEs constraints.

The whole procedure relies on having first assigned the  $^1\text{H}$  NMR spectrum, i.e. as many resonances as possible must be resolved and attributed to specific protons in the sequence.

### 2.5.6 Relaxation in protein NMR

Protein backbones and side chains display varying degrees of flexibility, which allows many slightly different but related conformational sub-states to occur. Such fluctuations are known to differ in both timescale and magnitude, from rotation of methyl group (nanoseconds) to the flipping of buried tyrosine rings (seconds). Movement at molecular level is essential for many biological processes, including protein folding, protein-protein interaction and protein-ligand interaction (Stock, 1999).

The NMR relaxation is strictly related to the internal motions of a molecule and is sensitive to motions over a wide range of timescale. In the last 10 years, NMR relaxation experiments have become an essential instrument in the study of protein dynamics as it

can provide information about the internal motions at many sites throughout the protein. The development of heteronuclear and multidimensional NMR and the isotopic labelling methods, were fundamental in the growth of the relaxation studies applied to proteins. In particular,  $^{15}\text{N}$  and  $^{13}\text{C}$  relaxation data are widely used for studying respectively backbone and side chains dynamics (Yang and Kay, 1996). The relaxation properties of protonated heteronuclei such as  $^{15}\text{N}$  and  $^{13}\text{C}$  are typically dominated by the dipole-dipole interaction with the attached proton(s); therefore, the relaxation data can be interpreted in terms of the motions of the  $^{15}\text{N}\text{-}^1\text{H}$  or  $^{13}\text{C}\text{-}^1\text{H}$  bond vector (Peng and Wagner, 1994). Heteronuclear relaxation experiments typically provide heteroatom  $T_1$ ,  $T_2$  and heteronuclear steady-states NOE. These data are subsequently interpreted in the context of motional models. The most used approach is the “model-free” analysis described by Lipari & Szabo (1982a,b), which in its simplest form allows extraction, from NMR relaxation measurements, of an overall rotational correlation time characterising internal dynamics and of an order parameter, describing the amplitude of the internal motions from NMR relaxation measurements.

Heteronuclear relaxation experiments can be used to study several biological process.

In the characterization of ligand-binding properties of a protein, the analysis of how the relaxation parameters vary upon binding can highlight the regions of the protein that show different flexibility in the apo and holo form and that are more affected by the presence of the ligand. Moreover, under some assumptions, changes in the NMR relaxation parameters can be linked to changes in thermodynamic quantities so that the entropic contributions to the site-specific binding can be delineated. Therefore, the information obtained from relaxation parameters, combined with structural data, can help in the comprehension of the mechanism of binding giving also an estimate of the role of the dynamics in driving molecular interactions (Forman-Kay, 1999).

In protein folding the NMR relaxation experiments are usually applied to unfolded or partially folded states. The variations of the NMR parameters are measured between the native and the denatured or partially folded states. This can highlight the regions of the protein that have a different backbone flexibility in the different stability states. The regions that are motionally more restricted in the unfolded or partially folded state are probably involved in some kind of structural interactions and can therefore play an important role in the stability of the protein. Constriction of backbone motions can occur through formation of local hydrophobic clusters, through formation of elements of secondary structure or through long-range interactions in compact folding intermediate (Dyson & Wright, 2001).

## 2.6 NMR Experiments

### 2.6.1 1D NMR experiments

The general scheme of a 1D NMR experiment is shown in Figure 2.6.

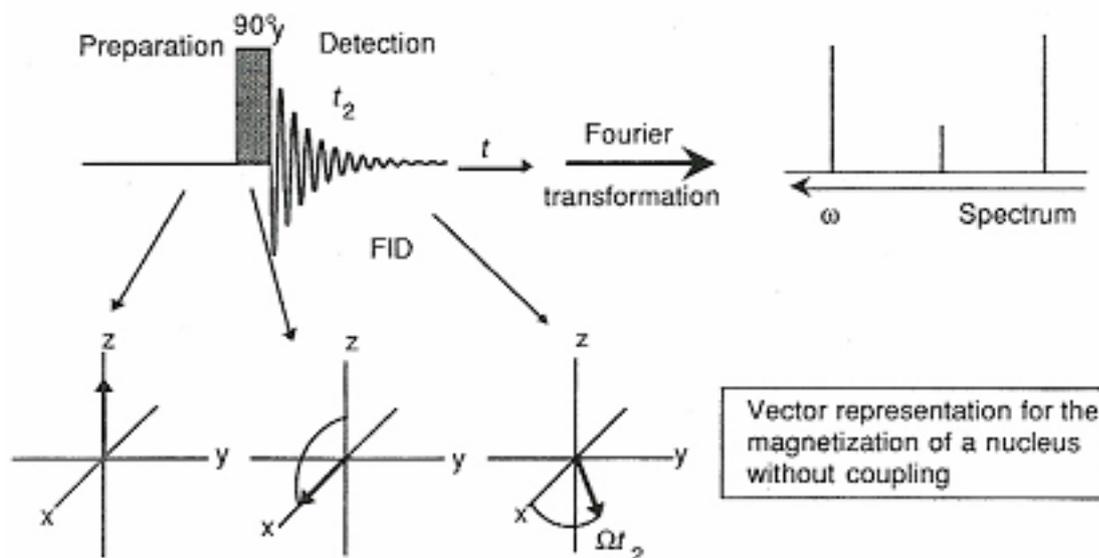


Figure 2.6. 1D NMR experiment, vector model for a  $90^\circ$  pulse. (The figure has been reproduced from Evans, 1995.)

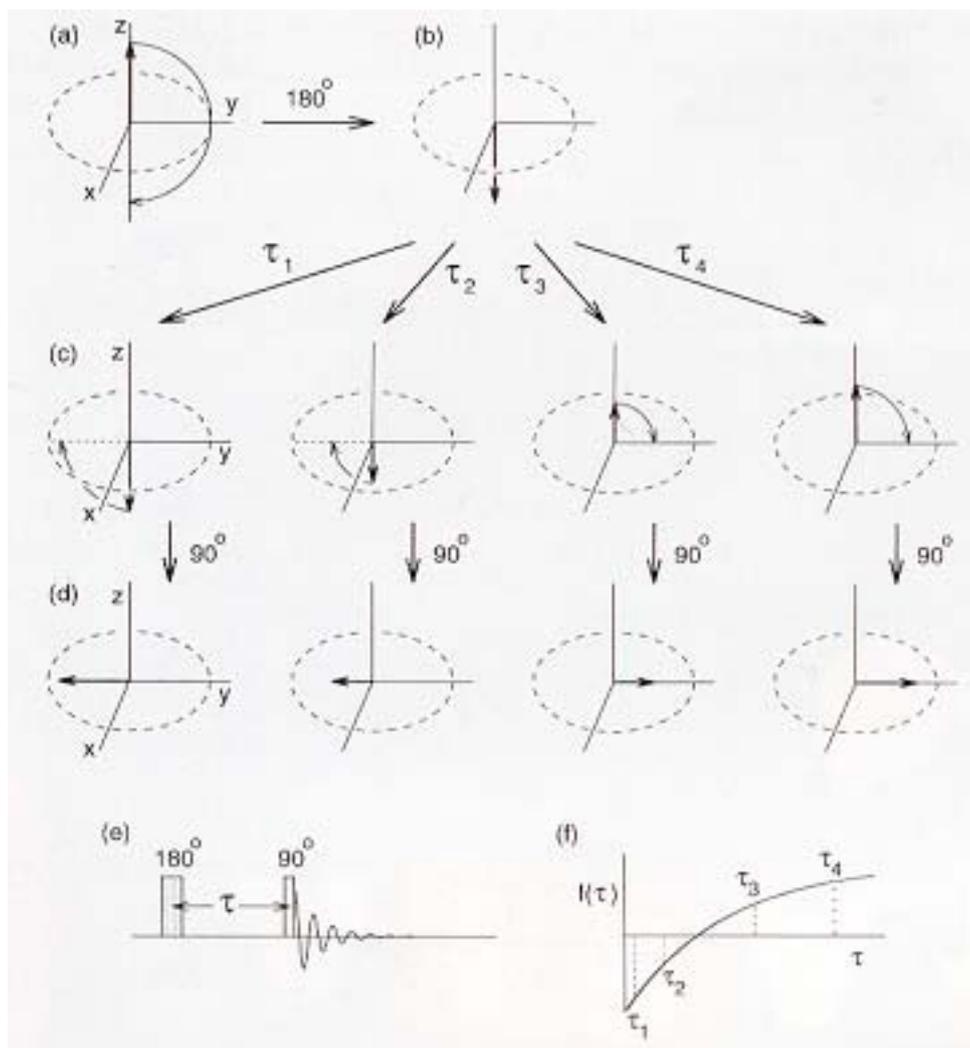
After a preparation time during which an equilibrium magnetization  $\mathbf{M}$  parallel to  $\mathbf{B}_0$  has been established by placing the sample in a static magnetic field  $\mathbf{B}_0$ , a radiofrequency pulse is applied that rotates  $\mathbf{M}$  away from the  $z$  axis by a flip angle  $\beta$ , generally  $90^\circ$ . Immediately after the pulse, the free induction decay (FID) is recorded during  $t_2$ . The NMR spectrum is then obtained by Fourier transformation of these data.

#### - Inversion recovery-measurement of $T_1$

Spin-lattice relaxation time may be measured using the pulse sequence  $180^\circ - \tau - 90^\circ$  (Figure 2.7). The equilibrium magnetization (a) is inverted by the first pulse, leaving  $\mathbf{M}$  along the negative  $z$  axis (b). During the delay  $\tau$ ,  $\mathbf{M}$  undergoes partial spin-lattice relaxation (c) to give a  $z$  magnetization  $M_z(\tau)$  which the  $90^\circ$  pulse rotates onto the  $y$  axis (d). The FID is recorded and after Fourier transformation the spectrum contains peaks whose intensities  $I(\tau)$  are proportional to  $M_z(\tau)$ . The whole process is repeated for different values of  $\tau$  so to map out the recovery of the inverted magnetization (f). Assuming exponential relaxation:

$$M_z(\tau) = M_0 \left[ 1 - 2 \exp(-\tau / T_1) \right] \quad (2.11)$$

the  $T_1$  of each peak can be obtained.



**Figure 2.7.** Inversion recovery experiment. (a) equilibrium. (b) After the  $180^\circ$  pulse. (c) After four different delays  $\tau_1 < \tau_2 < \tau_3 < \tau_4$ . (d) After the  $90^\circ$  pulse. (e) The pulse sequence. (f) the observed NMR intensity  $I(\tau)$  as a function of the delay  $\tau$ . (The figure has been reproduced from Hore, 1995.)

### - Spin echo-measurement of $T_2$

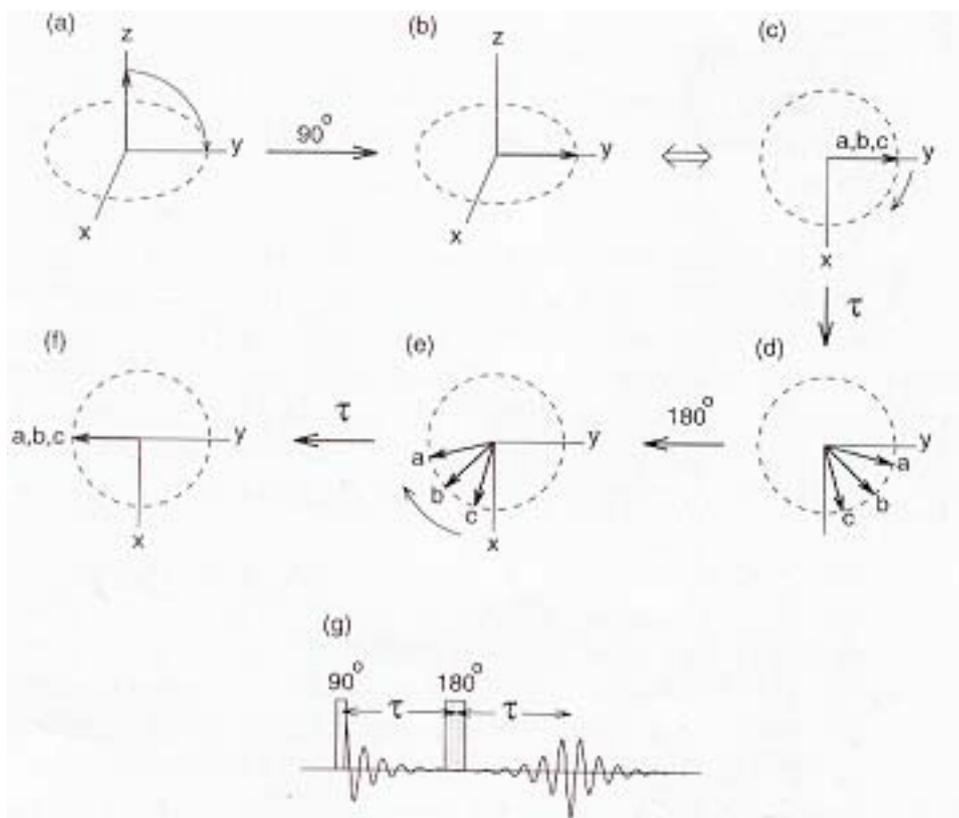
The width at half height of the peaks in an NMR spectrum is given by  $1/(\pi T_2)$ . This suggests that  $T_2$  may be estimated simply by measuring the width of the spectral peaks. Unfortunately, this rarely gives a reliable estimate of  $T_2$ . The width of the NMR peaks tends to be larger than  $1/(\pi T_2)$  because of the inhomogeneity of the static magnetic field, which spreads the peaks out, thus giving a larger overall line-width.

The spin-echo experiment allows to measure the spin-spin relaxation times  $T_2$  independently on the inhomogeneity of the static magnetic field (Figure 2.8).

After the  $90^\circ$  pulse (b, c), the inhomogeneity of the static magnetic field causes the isochromats to fan out in the  $xy$  plane (d) (spin-spin relaxation is ignored). After a period  $\tau$ , the  $180^\circ$  pulse flips the magnetization of each region around the  $x$  axis to symmetrical positions in the  $xy$  plane (e), where precession continues for a further time  $\tau$ . Whatever the

precession frequency and  $\tau$ , all regions at the end of this second delay come back into phase perfectly, giving an echo. The signal is recorded and the spectrum contains NMR lines whose amplitude are independent of the field inhomogeneity.

Now, consider the effect of relaxation on the echo amplitude. During both  $\tau$  delays, spin-spin relaxation destroys the phase-coherence created by the  $90^\circ$  pulse, and causes the transverse magnetization to decay at a rate  $T_2^{-1}$ . This dephasing, which is produced by the fluctuating magnetic fields arising from random molecular motions, is not refocused after the  $180^\circ$  pulse.



**Figure 2.8.** Spin-echo experiment. (a) Equilibrium. (b) After the  $90^\circ$  pulse. (c) Same as (b), viewed from vertically above the  $xy$  plane. (d) After a delay  $\tau$ . (e) After the  $180^\circ$  pulse. (f) after the second delay  $\tau$ . (g) The pulse sequence. (The figure has been reproduced from Hore, 1995.)

The NMR intensity of each line in the spin-echo spectrum is then given by

$$I(2\tau) = I(0) \exp(-2\tau / T_2) \quad (2.12)$$

The whole experiment is repeated with different  $\tau$  delays, and  $T_2$  is obtained.

The 1D NMR spectra of biopolymers are dramatically crowded and complex, as it is possible to see, for example, in Figure 2.9 where the 1D  $^1\text{H}$ -NMR spectrum of cL-BABPABP, a 14000 Da protein, is reported.

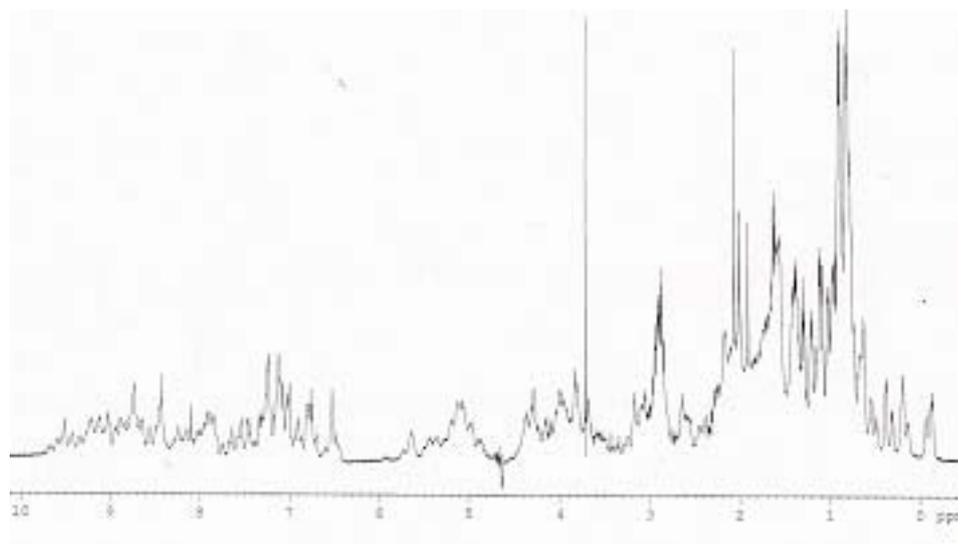
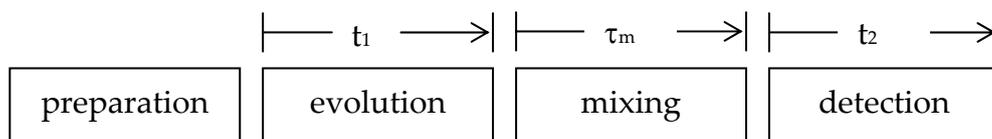


Figura 2.9. 500 MHz 1D  $^1\text{H}$ -NMR spectrum of cL-BABP

The large overlap in the proton resonances exclude any chance for a detailed structural study. To overcome the natural limitation of the 1D NMR spectroscopy, multidimensional 2D and 3D, experiments have been developed in the last twenty years. A brief description of the 2D and 3D NMR spectroscopy is reported below.

## 2.6.2 2D NMR experiments

A general scheme for 2D NMR includes four successive time periods: preparation, evolution, mixing and detection.



The evolution time  $t_1$  is an incrementable delay and a characteristic feature of all 2D pulse sequences. When a 2D experiment is carried out, the sequence is run with a range of  $t_1$  values evenly spaced from near zero to a value  $t_{1\max}$ . With each  $t_1$  value an FID is recorded in the detection period. In this context the real time variable that defines the FID itself is given by the symbol  $t_2$ . The pulse sequences are designed so that the signals detected in  $t_2$  are modulated in amplitude or phase as a function of the delay  $t_1$ . The different correlations between the spins are realised during the mixing time. Therefore, the frequencies of the signals present during  $t_1$  can be measured indirectly through their effect on the real-time signals detected directly in  $t_2$ . Once the complete data set has been acquired, it is Fourier transformed with respect to  $t_1$  and  $t_2$  to give a two dimensional spectrum that is function of two frequencies variables  $F_1$  and  $F_2$ . From this point of view the various 2D experiments differ in the type of signals that is present during  $t_1$  and in the interaction employed during the mixing time to transmit information from  $t_1$  to  $t_2$ .

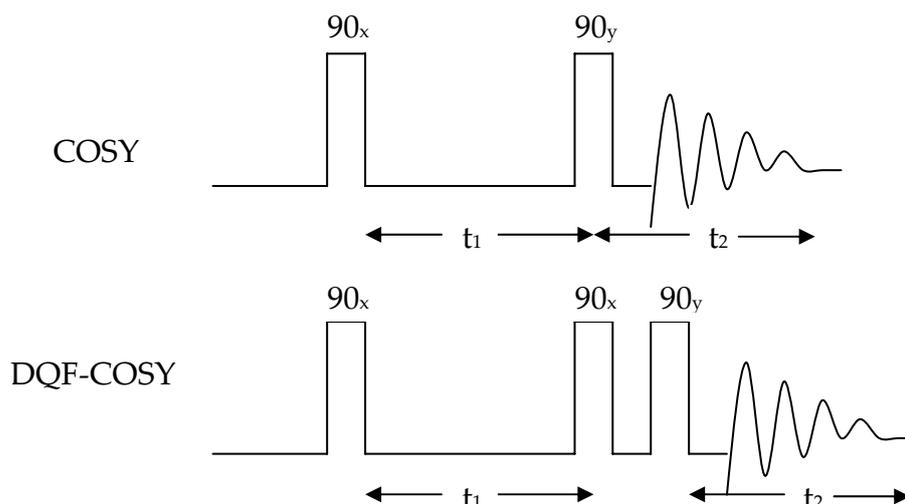
A brief descriptions of the 2D NMR experiments used in this work is given below.

The majority of these experiments have a common structure: the diagonal corresponds to the conventional “one-dimensional” spectrum, while the off-diagonal peaks, or cross-peaks, contain information about the connections between resonances on the diagonal. The nature of these connections depend on the kind of two-dimensional experiment being carried out: one can observe scalar connections between the resonances of the nuclei separated by two or three bonds, and dipolar connections (through-space), between the resonances of nuclei which are close together in space.

### - $^1\text{H}$ - $^1\text{H}$ scalar connection experiments: COSY and TOCSY

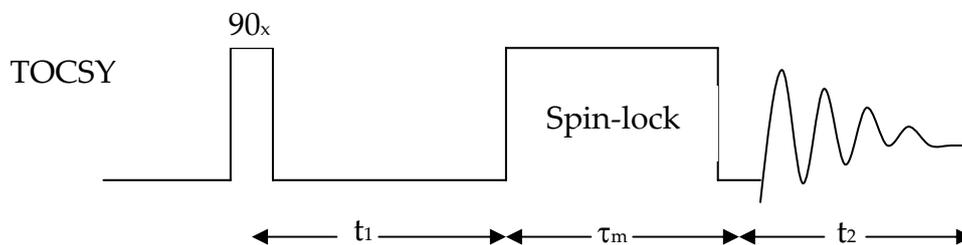
COSY (CORrelated Spectroscopy) is the simplest and oldest 2D experiment that can be used to identify pairs of protons that have a mutual scalar coupling. In a COSY experiment only cross-peaks between protons connected by three or less covalent bonds are detectable.

Due to the poor sensitivity of the COSY, other sequences have been implemented. The most used of these is the DQF-COSY in which the pulse sequence has been modified by adding a third  $90^\circ$  pulse before the detection period.



A TOCSY (TOtal Correlation SpectroscopY) experiment allows to observe through an “isotropic mixing” all the scalar connections within a spin system. A spin system is a group of spins that are connected by spin-spin coupling. In a polypeptide chain, since the peptide bond interrupts the scalar coupling between the protons, each amino acid residue constitutes an independent spin-system. Therefore the cross peaks in the TOCSY spectrum of a protein are produced by the protons of a single amino acid. Considering, for example, the scalar connections of the NH backbone proton, for each residue cross peaks arising from all the scalar coupling between the H<sub>N</sub>-H<sub>α</sub> and H<sub>N</sub>-sidechain protons are present. Differently, a COSY spectrum shows only the H<sub>N</sub>-H<sub>α</sub> cross-peaks.

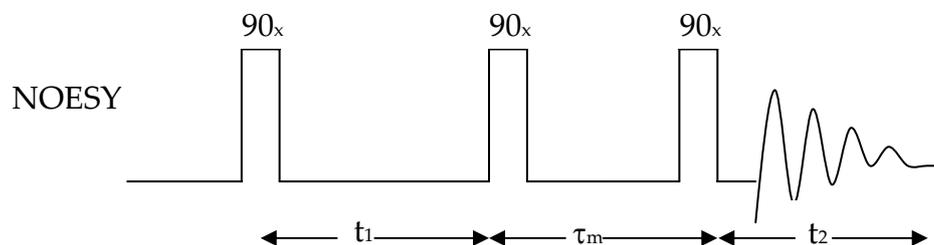
The key feature of the TOCSY experiment is that it uses a period of spin-locking for the magnetization transfer. The spin-locking field is a long, strong radio-frequency pulse along a specified axis. If this field is strong enough, the effect is that all the chemical shift differences become irrelevant and the spins become equivalent. In this condition, it is possible to observe the scalar connections between directly and remotely connected spins.



### - <sup>1</sup>H-<sup>1</sup>H dipolar connection experiments: NOESY

In a 1D experiment for measuring the NOE between two dipolar-coupled nuclei A and B, the resonance of A, for example, is subjected to a weak, selective irradiation for a time period, which causes saturation of this resonance. The NOE resulting is manifested by a fractional change of the resonance intensities of B. Looking at the complex and crowded 1D spectrum of

Figure 2.9, it is easy to deduce that for a protein selective excitation of individual resonances is impossible due to the overlapping lines. The NOE measurement for macromolecules is realised through the 2D NOESY experiment where the NOEs are manifested by cross-peaks and are measured simultaneously.



The characteristic feature of NOESY sequence is the mixing time  $\tau_m$  during which magnetization is transferred between dipolar-coupled nuclei. From the intensities of the cross-peaks, it is possible to calculate the distance between the two dipolar-coupled nuclei.

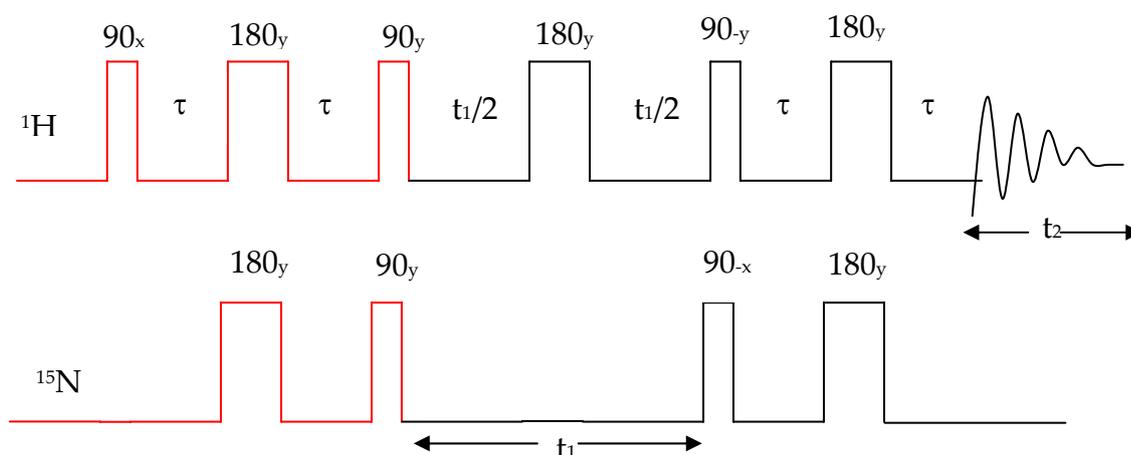
#### - Heteronuclear scalar connection experiment: HSQC

In the heteronuclear 2D spectrum the proton resonances are spread out according to the shifts of the heteronuclei to which they are coupled. The experiments generally used in protein NMR are based on the large coupling constant between proton and the bonded heteronucleus ( $J^{1H-13C}$  140Hz,  $J^{1H-15N}$  90Hz).

The common scheme starts with the proton magnetization and ends up with proton detection. In the first part of the sequence, proton magnetization is transferred with an appropriate pulse sequence to the heteronucleus that evolves during the  $t_1$  period. In this way, the proton spin is labelled with a modulation arising from the heteronucleus. After  $t_1$ , the magnetization is transferred back to the proton and detected. Information on both proton and heteronucleus frequencies are thus present in the FID and in the resulting NMR spectrum.

One of the most used pulse sequence is the HSQC (Heteronuclear Single Quantum Correlation). The basis of the HSQC experiment is the INEPT sequence in which the magnetization of the more sensitive nucleus ( $^1H$ ) is used to enhance the weak NMR signal from the less sensitive nucleus, such as  $^{15}N$  or  $^{13}C$ . The INEPT sequence is marked in red in the HSQC pulse sequence scheme. The final element of the HSQC is a reversal of the initial INEPT (all the pulses and delays are applied in reverse order), except that the first pulse has been removed. In this part of the sequence the  $t_1$  frequency-labelled magnetization is transferred back to the proton for detection.

## HSQC pulse sequence

**- Heteronuclear relaxation experiments**

Backbone dynamics are most commonly investigated by measurements of  $^{15}\text{N}$   $T_1$ ,  $T_2$  relaxation times and the  $^1\text{H}$ - $^{15}\text{N}$  NOE in uniformly  $^{15}\text{N}$ -labelled protein (for a review, Peng & Wagner, 1994).

Pulse sequences for measuring spin relaxation usually consists of five building blocks: preparation, relaxation, frequency labeling, mixing and acquisition. The preparation period normally consists of a INEPT transfer step from protons to  $^{15}\text{N}$ . The coherence present after the preparation period provides the initial condition for the relaxation delay  $T$ . Chemical shifts are recorded during the  $t_1$  frequency-labelling period to generate the indirect dimension of the two-dimensional NMR spectrum. The desired heteronuclear coherence is transferred to proton magnetization using reverse INEPT during the mixing period. The relaxation-encoded, frequency-labelled transverse proton magnetization is recorded during the  $t_2$  acquisition period.

The  $T_1$  relaxation time is measured in a series of experiments that combine the inversion recovery applied to the  $^{15}\text{N}$  and the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC. Several spectra are acquired with different relaxation delay  $T$  (usually, 10 values are enough to well characterize  $T_1$ ). The resulting 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra differ only in the intensity of the peaks. For each residue  $T_1$  can be determined by measuring the intensity of the peak for each  $T$  value and fitting the data with the exponential decay (2.11).

Similarly, the  $T_2$  relaxation time is measured in a series of experiments that combine the spin-echo applied to the  $^{15}\text{N}$  and the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC. Several spectra are acquired with different relaxation delay  $T$  (usually, 10 values are enough to well characterise  $T_2$ ). The resulting 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra differ only in the intensity of the peaks. For each residue  $T_2$  can be determined by measuring the intensity of the peak for each  $T$  value and fitting the data with

the exponential decay (2.12). Usually, the spin-echo used is the Carr-Purcell-Meiboom-Gill (CPMG) sequence, which is:

$$90^\circ - \tau - 180^\circ - 2\tau - 180^\circ - 2\tau - 180^\circ - \dots$$

The CPMG sequence allows for compensation of cumulative errors due to an incorrectly set pulse length.

The  $^1\text{H}$ - $^{15}\text{N}$  NOE is measured by recording one  $^1\text{H}$ - $^{15}\text{N}$  HSQC with saturation of  $^1\text{H}$  magnetization and one spectrum without saturation. The saturation is realised by applying a radiofrequency field to the proton so that the population of the two  $^1\text{H}$  energy levels is equalised. During the time in which the saturation is applied, the cross-relaxation can occur between the  $^{15}\text{N}$  and the  $^1\text{H}$  that are dipolarly coupled. The resulting  $^{15}\text{N}$  magnetization is read out using an  $^1\text{H}$ - $^{15}\text{N}$  HSQC through which the NOE information is transferred from the  $^{15}\text{N}$  to the  $^1\text{H}$ . The  $^1\text{H}$ - $^{15}\text{N}$  NOE for each residue is obtained by comparing the intensity of each peak between the spectrum with  $^1\text{H}$  saturation and the spectrum without  $^1\text{H}$  saturation.

### 2.6.3 3D NMR experiments

One of the problems with  $^1\text{H}$  NOESY NMR of proteins is that all the NOEs must be resolved. This becomes increasingly difficult with higher molecular weight proteins due to the overlapping signals. To solve this problem, a whole family of pulse sequences has been introduced that involve three time variables ( $t_1$ ,  $t_2$ ,  $t_3$ ) instead of two, so that three-dimensional Fourier transformation leads to a cube rather than a contour map (for a review, see Clore & Gronenborn, 1994). One important group of experiments combines an HSQC sequence with a homonuclear technique such as NOESY or TOCSY. This can be achieved by replacing the first  $90^\circ$  pulse of the homonuclear experiment with an HSQC sequence. In a 3D HSQC-NOESY for example, one axis represents the  $^1\text{H}$  chemical shift of the amide  $\text{H}_\text{N}$ , one axis represents the  $^{15}\text{N}$  chemical shift of the amide NH and the third represents the chemical shifts of the protons that show NOE connectivities to  $\text{H}_\text{N}$ . Rather than following the assignment pathway in three dimensions, strips along the  $^{15}\text{N}$  dimension are sorted according to their sequential connectivities.

## 2.7 Requirements for an NMR Sample

The fundamental problem in the biological applications of NMR is its intrinsic poor sensitivity. The intensity of the NMR signal is proportional to the amount of material and for structural studies concentrations of at least 0.5 mM in 0.5 ml are required. These concentration and volume quantities means that amounts of the order of milligrams are necessary for a small

protein. The large amount of material may be sometimes a limiting factor specially for recombinant labelled proteins.

At concentration of 0.5-1 mM, several proteins may aggregate. The aggregation must be avoided since it decreases the resolution of the spectrum through the broadening of the signals. Therefore, a crucial role is played by the pH, the ionic strength and the buffer type. These three parameters should guarantee high solubility and stability for the protein that must be stable during the course of the NMR experiments.

## 2.8 Strategies for High Level Expression and Labeling of Recombinant Proteins for NMR Structural Studies

A well-known pivot of biology is that protein structure determines function. Nuclear Magnetic Resonance spectroscopy (NMR) plays a central role in dissecting the relationship between sequence, structure, dynamics and molecular recognition in biology, a key for achieve insight into biological function and into the nature of fundamental biological processes.

Molecular biology represents a very important tool for NMR for two main reasons:

- NMR needs **over-expression**; the intensity of the NMR signal is proportional to the amount of material and so for structural studies concentration of at least 0.5 mM of protein are required; in most cases the target protein represents a small percentage of the total cellular proteins in the original tissue and sometimes the biological material is difficult to rescue (need of large quantities, authorizations, etc..).
- NMR needs **isotopically enriched proteins**; increasing molecular size leads to slower tumbling and correspondingly shorter spin-spin relaxation times ( $T_2$ ) and also leads to increasingly complex spectra. Short  $T_2$  values severely limit the power and flexibility of multiple-pulse NMR experiments in at least two ways: (i) the signal-to-noise ratio of a Lorentzian line degrades with decreasing  $T_2$ , and the effectiveness of the currently available library of multidimensional and multinuclear NMR experiments is exponentially sensitive to  $T_2$ ; (ii) accordingly, the standard triple resonance experiments become unreliable at room temperature for proteins larger than 30 kDa and largely fail for proteins above 35 kDa in the absence of elevated temperature and/or extensive deuteration. So majority of biomolecular NMR techniques require isotopic labelling ( $^{15}\text{N}$ ,  $^{13}\text{C}$ ,  $^2\text{H}$ ) of recombinant proteins that are generally required for optimising NMR solution conditions (Bagby *et al.*, 1997, Lepre & Moore, 1998), NMR assignments and structure determination (Sattler *et al.*, 1999), performing heteronuclear spin relaxation studies, biological screening and high throughput structural genomics studies (Montelione *et al.*, 2000). Recombinant over-expression is needed since labeling is almost impossible to achieve in vivo.

So, as described in section 2.1, for proteins of 100-120 amino acids, the realistic limit of homonuclear NMR is the overlap of the signals in the  $^1\text{H}$  spectrum (Figure 2.10, A). For



	available. Gene expression easily controlled. Easy grown with high yields (product can form up to 50% of total cell protein).	Disulfide bonds difficult to achieve (only in periplasma).
Yeasts ( <i>S. cerevisiae</i> , <i>P. pastoris</i> )	Fermentation relatively inexpensive. Facilitates glycosilation and disulphide bonds.	Gene expression less easily controlled. Glycosilation not identical to mammalian systems.

### Choice of vector for protein amplification

In order to clone the gene of interest all engineered vectors have a selection of unique restriction site downstream of a transcription promotor sequence. The choice of vector family is governed by the host. Once the host have been selected, many different vectors are available for consideration, from simple expression vectors to those that secrete fusion proteins. However, as for the selection of a suitable host system, the final choice of vector should take into consideration the specific requirements of the application and will, of course, be influenced by the behaviour of the target protein. One key factor that has led to the increased use of fusion protein vector is that amplification of a fusion protein containing a tag of known size and biological function can greatly simplify subsequent isolation, purification and detection. Fusion proteins have advantages and disadvantages:

Advantages	Disadvantages
<b>Fusion proteins</b>	
Cell compartment can be targeted. Provide a marker for expression. Simplification of purification using affinity chromatography under denaturing and non-denaturing conditions. Easy detection. Refolding achievable on chromatography column. Ideal for secreted proteins as product is easily isolated from growth media.	Tag may interfere with protein structure and affect folding and biological activity. Cleavage site is not always 100% specific if tag needs to be removed.
<b>Non-fusion proteins</b>	
No cleavage step is necessary	Purification and detection not simple Problems with solubility may be difficult to overcome, reducing potential yield

### Choice of growth conditions

To improve protein yield it is necessary to optimize the culture conditions that are the effect of cell strain, medium composition, incubation temperature and induction conditions. Exact conditions will vary for each fusion protein expressed.

Recombinant proteins may be produced as insoluble inclusion bodies. The alteration of the growth conditions in order to slow the rate of translation could allow the production of soluble protein:

- Lowering of the growth temperature (within the range of +20° to +30°C) to improve solubility.
- Decreasing IPTG concentration to <0.1 mM to alter induction level.
- Altering time of induction.
- Inducing for a shorter period of time.
- Inducing at a higher cell density for a short period of time.
- Increasing aeration. High oxygen transport can help preventing the formation of inclusion bodies.

#### - Choice of host strain

Several strains should be tested to check their different ability to growth and survive to minimal media conditions and to express the target protein (Figure 2.11; example of growth curves for different strains *E. coli* grown in M9 in the same conditions).

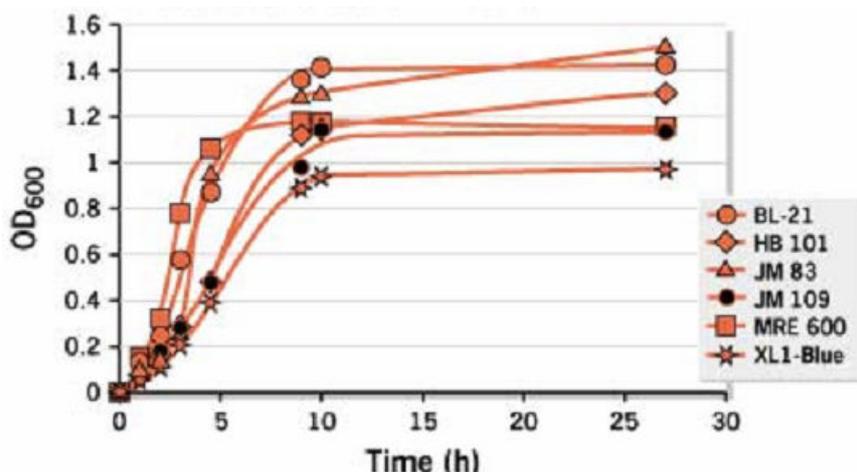


Figure 2.11. Growth curves of *E. coli* on unlabeled minimal media M9.

#### - Choice of culture medium

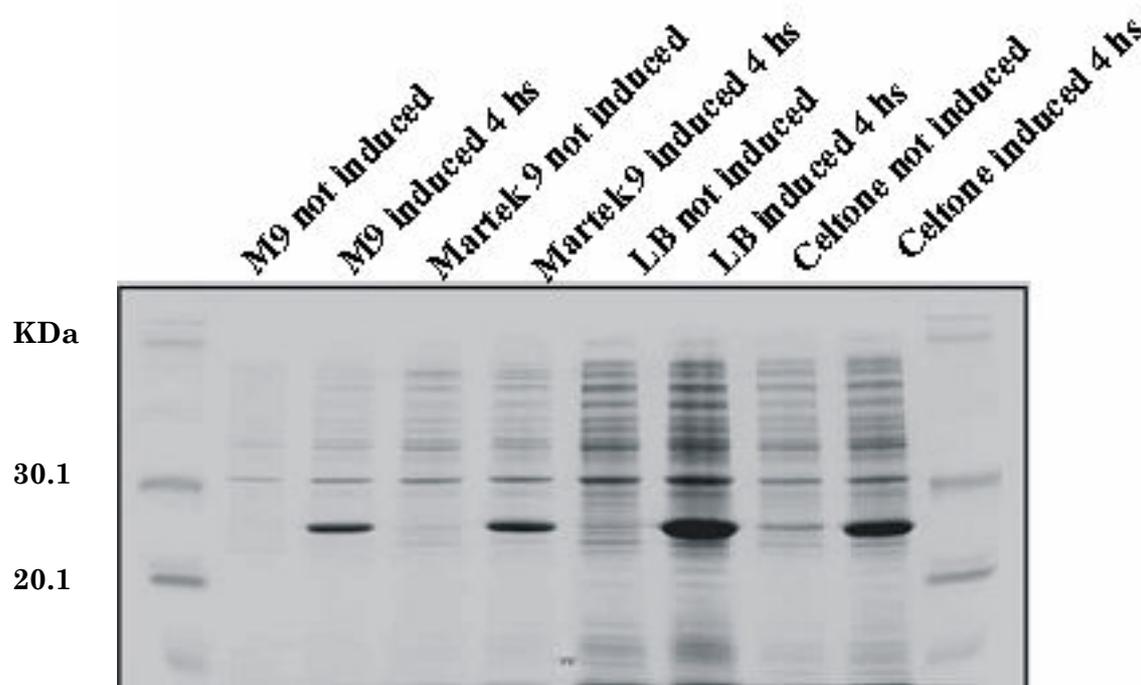
Two main types of culture media can be tested for labeling:

- Ready-to-use media like algae or bacteria hydrolysate.

Minimal media added with  $^{15}\text{N}$  nitrogen source or/and  $^{13}\text{C}$  carbon source.

Minimal media are made of nutrients like C and N source, salts, buffering substances, traces elements and vitamins; the carbon source can be glucose, glycerol, acetate, succinate, methanol; the nitrogen source can be  $\text{NH}_4\text{Cl}$  or  $(\text{NH}_4)_2\text{SO}_4$ ; salts are  $\text{NaCl/KCl}$ ,  $\text{MgSO}_4$ ,  $\text{CaCl}_2$ ; buffer usually is phosphate, pH 7.5; the trace elements are constituted by a mixtures of metal ions like  $\text{Co}^{2+}$ ,  $\text{Cu}^{2+}$ ,  $\text{Zn}^{2+}$ ,  $\text{Mn}^{2+}$ ,  $\text{Fe}^{2+}$ ; the vitamins commonly used are thiamine, biotin, folic acid, niacinamide, pantothenic acid, pyridoxal, riboflavin.

Expression tests must be done in order to decide the best conditions achievable for a specific expression system (Figure 2.12).



**Figure 2.12.** Example of expression test for glutathione S-transferase: comparison of the growth conditions that give the higher yield; lane 1: Molecular weight marker; lane 2: M9 not induced; lane 3: M9 induced after 4 hours; lane 4: Martek 9 broth not induced; lane 5: Martek 9 broth induced after 4 hours; lane 6: Luria broth not induced; lane 7: Luria Broth induced after 4 hours; lane 8: Celtone broth not induced; lane 9: Celtone broth induced 4 hours; lane 10: molecular weight marker. (M9 = 4 g/L glucose).

### Extraction of recombinant proteins

Samples should be clear and free from particles before beginning any isolation or purification. Extraction procedures should be selected according to the source of the protein, such as bacterial, plant or mammalian, intracellular or extracellular. Selection of an extraction technique is dependent as much upon the equipment available and scale of operation as on the type of sample. Common extraction processes are cell lysis (osmotic shock), enzymatic

digestion, Grinding with abrasive (e.g. sand), ultrasonication, use of Manton-Gaulin homogeniser, french press and fractional precipitation.

Extraction should be performed quickly, at sub-ambient temperatures, in the presence of a suitable buffer to maintain pH and ionic strength and to stabilize the sample.

### **Choice of the procedure for $^{13}\text{C}$ and $^{15}\text{N}$ proteins enrichment**

For the obtainment of labelled proteins two different expression protocols are currently used: the first employs a minimal medium used both for cellular growth and induction; the second, a two-stage protocol (Marley *et al.*, 2001), employs a rich unlabeled media for rapid cellular growth, then the cellular biomass is harvested, re-suspended in isotopically labelled minimal media which volume is reduced to one fourth, incubated for 30 min-1 hour to allow for the recovery of growth and clearance of unlabeled metabolites and finally induced.

The first protocol has the advantage of producing uniform labelling; the disadvantages are that is very difficult to get reasonable OD (the cellular growth is slower than in rich medium with following low yield of protein) and very often pre-lysis of the cells occurred using even number of different protocols.

The two-stage protocol has the advantages of reducing isotope consumption by generating the majority cell mass using unlabeled medium and of increasing protein expression level by removing bio-product inhibitory to growth and expression; the disadvantage is that if control of expression is not really stringent it could be expressed a mixture of labeled and unlabeled protein. In order to reduce basal expression, commercial expression systems are available to overcome this problem (e.g. presence of an additional plasmid for the expression of control element/s); the basal expression level associated with the upstream *lac* promoter can be decreased in presence of 2% glucose in the growth medium; this addition should not significantly affect overall expression following induction with IPTG. The overall two-stage protocol is summarized in figure 2.13.

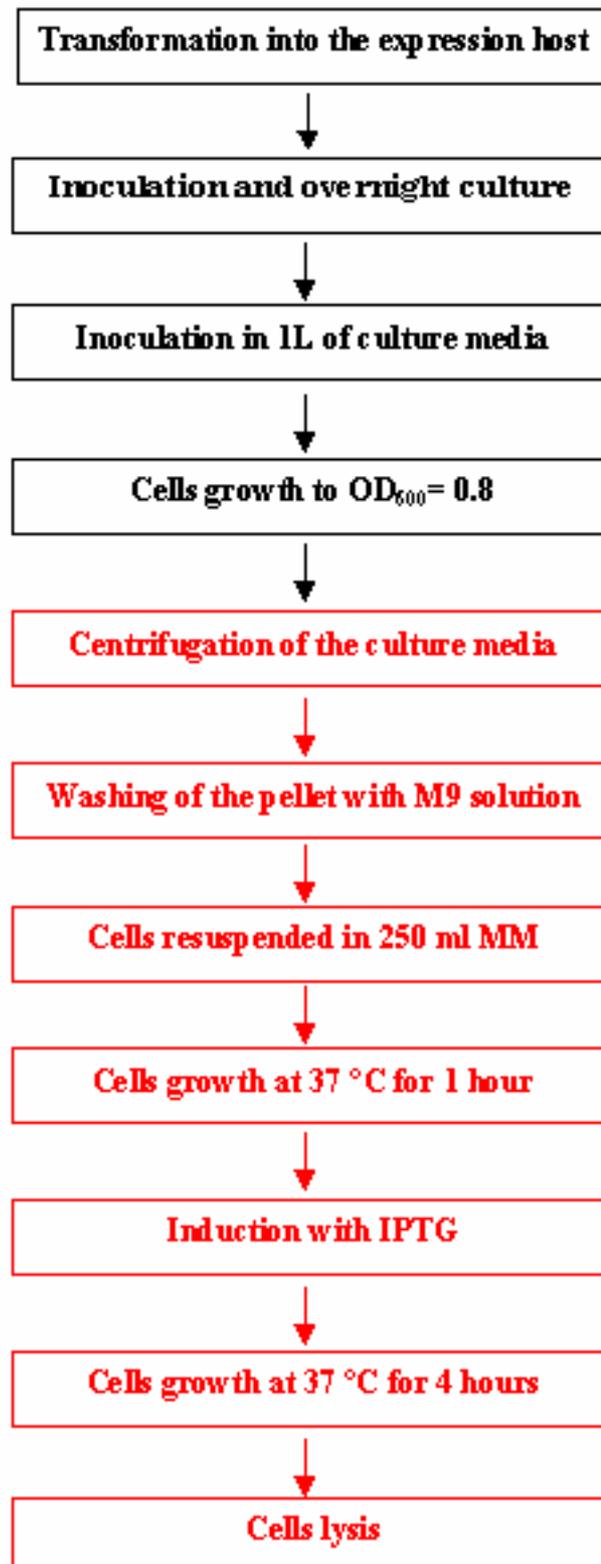


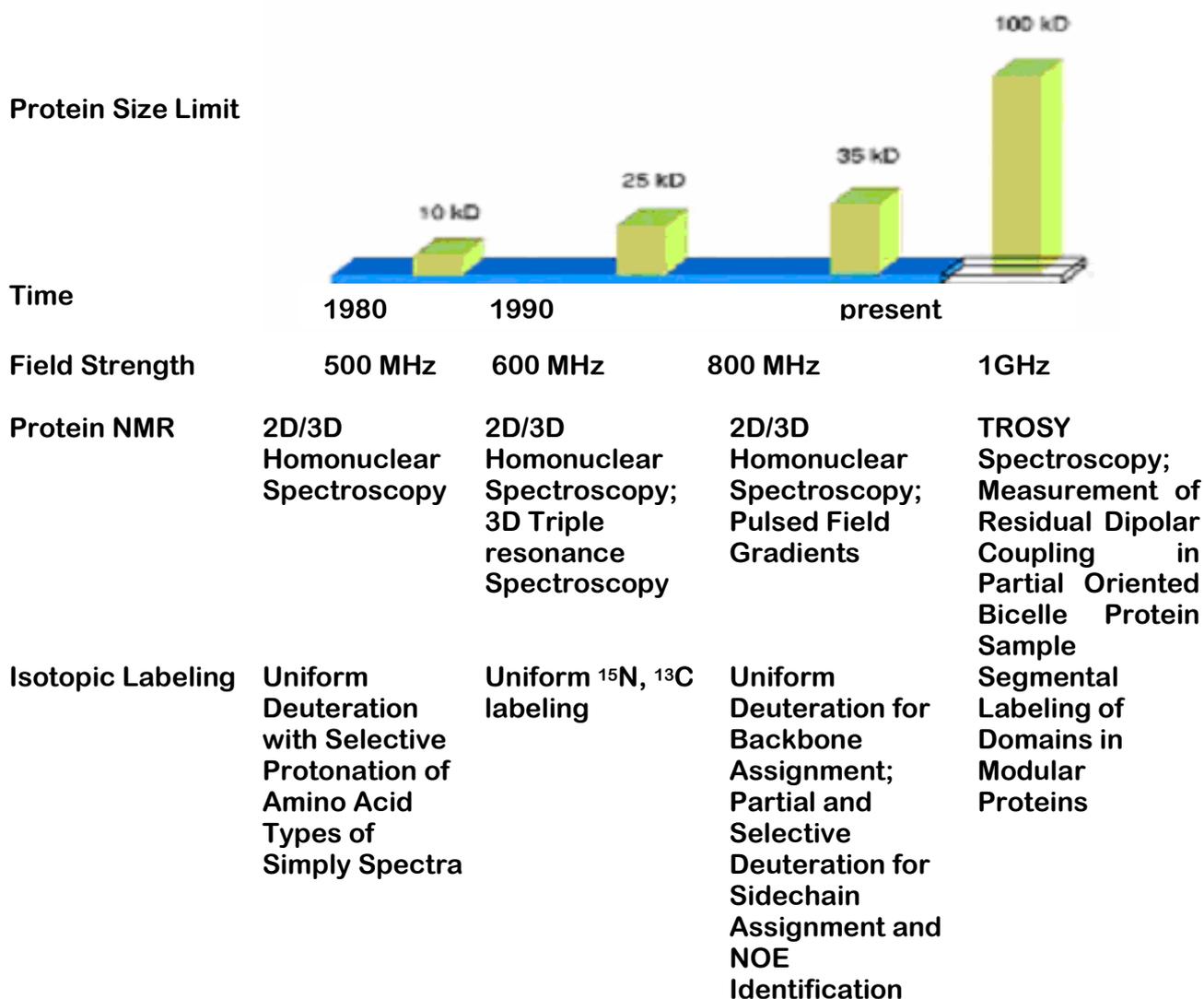
Figure 2.13. Schematic representation (in red) of the two-stage protocol

A labelling procedure that need special considerations is represented by the deuteration of proteins; this part has not been object of this thesis work but it is worth to mention that using cellular systems, the  $^2\text{H}$  incorporation of proteins require different considerations because the higher density of  $\text{D}_2\text{O}$  affect cellular methabolism leading to inhibition of cell growth. The deuteration of proteins is useful for achieving better spectra resolution and suppress spin diffusion.

An alternative system for protein enrichment is represented by high-throughput bacterial *cell-free system*.

Bacterial cell-free protein synthesis is a simple process where extraneously added DNA is transcribed and translated in vitro to produce protein. Efforts from different laboratories in the past few years led to design of protocols to generate highly synthetic bacterial cell extracts capable of producing hundreds of micrograms of protein in batch reactions. However, the short lifetime of the extract in batch reactions, consequently leading to low yield of protein is a limitation of the cell-free translation systems. Nevertheless the cell-free protein synthesis has several advantages over cell-based systems particularly in the expression of toxic proteins, labeling (uniform and/or partial) of amino acids for structural studies and expression of mutants of a protein for rapid analysis. Cell-free protein synthesis enables addition of detergents, chaperones and appropriate ligands during the process of protein synthesis, which may aid in proper folding of the proteins. Most of the genes cloned into bacterial expression vectors with T7 promoter, can also serve as templates for bacterial cell-free expression, obviating the need for sub-cloning. Cell-free protein synthesis requires several ingredients such as tRNA, amino acids, nucleotides, components of energy regenerating system, small molecules and T7 RNA polymerase in optimum proportions. Use of this complex mixture requires extensive optimization to produce proteins in a reproducible manner. Commercial extracts for protein synthesis are highly expensive, not practical for high-throughput studies and are not amenable to modifications, as the composition is not disclosed. It can be adopted the strategies of preparation of bacterial cell extracts for protein synthesis (Kigawa *et al.*, 2004). With the production of bacterial cell extract, it can be possible to find out the optimized conditions for protein synthesis. Many proteins belonging to different species were expressed in the cell-free system (Murthy *et al.*, 2004, Endo & Sawasaki, 2005).

The current advances in NMR spectroscopy and isotopic labeling have extended the size limit of protein NMR; in Figure 2.14 this concept is summarized.



**Figure 2.14.** Advances in NMR spectroscopy and isotopic labeling have extended the size limit of protein NMR. (The figure was adapted from Yu, 1999).

## 3. Materials and Laboratory Methods

### 3.1. Materials

All chemicals used in the work were supplied from Sigma, Carlo Erba, Pharmacia Biotech, Merck, unless otherwise indicated.

DE52 gel was from Whatman, Sephacryl S-100 HR and fast-flow Sepharose (Immobilized-Metal Affinity Chromatography, IMAC) were from Amersham Biosciences Ltd. (Uppsala, Sweden), Lipidex 1000 was from Packard. Protein markers for gel filtration and amino acids were from Sigma-Aldrich Co.

#### 3.1.1 Microorganisms

- **Bacterial strains:** *Escherichia coli* BL21(DE3),  
*Escherichia coli* XLIBLUE supercompetent cells were from Stratagene.

#### 3.1.2 Media and buffers

All buffers, stock solutions and media, if not mentioned here, were prepared as described in Sambrook & Russell (2001). During all biological transformations and work with *E. coli* cultures, standard biological safety procedures were followed. All large scale culture media were added with 2  $\mu$ L antifoam. All glassware, other containers, pipettes and media were autoclaved over a 121°C heating cycle before use. All the solutions of antibiotics, isopropyl- $\beta$ -D-thiogalacto-pyranosid (IPTG), glucose and vitamins were filtered into sterile falcon tubes with Millipore filters set on a sterile syringe (pore size 0.22  $\mu$ m). After use, all glassware, centrifuge tubes and other containers were sterilised by soaking either in bleach or in a freshly prepared 1% solution of Virkon disinfectant.

#### LB Medium:

Tryptone 10 g/L

Yeast Extract 5 g/L

NaCl 5 g/L

For the preparation of agar plates the medium was supplemented with 1% agar.

Antibiotic were added after the medium has been cooled to 50°C.

LB/Kan media contains 50  $\mu$ g/mL kanamicine

LB/Amp media contains 50  $\mu$ g/mL ampicilline

**SLBH Medium:**

Tryptone 10,75 g/L  
Yeast Extract 22,5g/L  
NaCl 5 g/L  
5 mL of glycerol 50%  
88 mL/L of 1M K<sub>2</sub>HPO<sub>4</sub>  
22 mL/L of 1M KH<sub>2</sub>PO<sub>4</sub>

**SOC Medium:**

For 1 L of SOB,  
20 mL of 1M glucose was added.

**SOB Medium:**

Tryptone 20 g/L  
Yeast Extract 5 g/L  
NaCl 0.5 g/L  
10 mL/L of 250 mM KCl  
pH adjusted to 7.0  
Before using 5 mL/L of sterilised 2 M MgCl<sub>2</sub> was added.

**Minimal Medium:**

For a final 500 mL  
400 mL of water,  
100 mL of M9 salts,  
1 mL of 1 M MgSO<sub>4</sub>,  
1 mL of 50 mM ZnSO<sub>4</sub>,  
5 µl of 10 mM FeCl<sub>3</sub>,  
500 µl of 0.1 M CaCl<sub>2</sub>,  
10 mL of glucose (40 % stock solution),  
3 mL of vitamin solution were added.

**M9 salts:**

Na<sub>2</sub>HPO<sub>4</sub> 33.9 g/L  
KH<sub>2</sub>PO<sub>4</sub> 15 g/L  
NaCl 2.5 g/L  
NH<sub>4</sub>Cl 5 gr/L

**Vitamin solution:**

For a final 200 mL solution,  
Thiamine 100 mg,  
d-biotin 20 mg,  
Choline Chloride 20 mg,  
Folic acid 20 mg,  
Niacinamide 20 mg,  
d-Panthenate 20 mg,  
Pyridoxal hydrochloride 20 mg,  
Riboflavin 2 mg were added.

**Lysis Buffer:**

50 mM Tris  
10% sucrose  
1 mM EDTA  
0.2 mM PMSF  
10 mM  $\beta$ -mercaptoethanol  
pH adjusted at 8.0

**Triton wash solution:**

0.5% Triton X100  
50 mM Tris/HCl pH 8.30  
100 mM NaCl,  
0.02% NaN<sub>3</sub>

**Denaturing solution:**

6M Guanidine Chloride (or 8 M Urea)

**Refolding buffer:**

50 mM Tris/HCl pH 8.30  
100 mM NaCl  
10 mM EDTA  
10 mM DTT

### 3.1.3 Plasmids for protein expression

#### - cL BABP

Plasmid for protein over-expression in *E. coli* were a kindly gift from Dr. Jefferson Foote (Fred Hutchinson Cancer Research Center, Seattle, WA98109, USA).

#### pET24d

#### - Glycodelin

Plasmids for protein over-expression in *E. coli* were provided from Dr. Anjali A. Karande (Dept. of Biochemistry Indian Institute of Science, Bangalore 560012 India).

**1<sup>st</sup> construct: pRSETA**, glycodelin full length of the mature protein cDNA with N-term fusion peptide containing His<sub>6</sub>-tag.

**2<sup>nd</sup> construct: pET22**, glycodelin full length of the mature protein cDNA with C-term fusion peptide containing His<sub>6</sub>-tag.

**3<sup>rd</sup> construct: pRSETA**, glycodelin full length of the mature protein cDNA with N-term fusion peptide containing His<sub>6</sub>- tag.

### 3.1.4 Molecular weight marker for SDS-PAGE electrophoresis

#### Sigma Marker Low Range

Protein (mol. wt.)

Aprotinin, bovine lung (6,500)

Ribonuclease A (13,700)

Trypsin inhibitor, soybean (20,000)

Trypsinogen, bovine pancreas (24,000)

Carbonic anhydrase, bovine erythrocytes (29,000)

Glyceraldehyde-3-phosphate dehydrogenase, rabbit muscle (36,000)

Ovalbumin, chicken egg (45,000)

Albumin, bovine serum (66,000)

### 3.1.5 Other chemicals

#### A. IPTG stock solution:

IPTG was dissolved in water (2.38g/10mL) to the end concentration of 1M. The stock solution was sterile filtered and stored in aliquots at -20°C until used. The stock solution was diluted 1:1000 when added to the medium, unless otherwise indicated.

**B. Kanamycin stock solution:**

Kanamycin was dissolved in water (0.5g/10mL) to the end concentration of 50 mg/mL. The stock solution was sterile filtered and stored in aliquots at  $-20^{\circ}\text{C}$  until used. The stock solution was diluted 1:1000 when added to the medium.

**C. Ampicillin stock solution:**

Ampicillin was dissolved in water (0.5g/10mL) to the end concentration of 50 mg/mL. The stock solution was sterile filtered and stored in aliquots at  $-20^{\circ}\text{C}$  until used. The stock solution was diluted 1:1000 when added to the medium.

**D. Rifampicin stock solution**

Rifampicin was dissolved in methanol (340mg/10mL) to the end concentration of 34 mg/mL. The stock solution was stored in aliquots at  $-20^{\circ}\text{C}$  until used. The stock solution was diluted 1:195 when added to the medium.

**E. Protease Inhibitors:**

- Complete Protease Inhibitors Cocktail

**F. Isotopically Enriched Chemicals:**

- Deuterium oxide, D<sub>2</sub>O 99%, 99.99% (Spectra 2000)
- Unlabeled and <sup>15</sup>N-Celtone was from Nalgene.
- <sup>15</sup>N-Ammonium chloride, NH<sub>4</sub>Cl 99.9% (ISOTEC)
- <sup>13</sup>C-Glucose 99,9% (Spectra 2000)

**G. Other Chemicals:**

- Acetic acid
- Acrylamide
- Ammonium chloride, NH<sub>4</sub>Cl
- Ammonium persulfate, APS
- Antifoam : Sigma Antifoam 289', CAT. n° A-8436.
- Bacto-agar
- d-Biotin
- Calcium chloride, CaCl<sub>2</sub>
- Choline Chloride
- Coomassie Brilliant Blue R-250
- Disodium hydrogenphosphate, Na<sub>2</sub>HPO<sub>4</sub>
- Dithiothreitol, DTT

- Ethanol
- Ethylenediaminetetraacetic acid, disodium salt, EDTA pH 8.00
- Folic acid
- d-Glucose
- L-Glycine
- Guanidine hydrochloride
- Hydrochloric acid, HCl
- Imidazole
- Isopropanol
- Isopropyl- $\beta$ -D-thiogalactopyranoside, IPTG
- Magnesium chloride,  $MgCl_2$
- Magnesium sulfate,  $MgSO_4$
- $\beta$ -Mercaptoethanol,  $\beta$ -ME
- Methanol
- N,N'-Methylenbisacrylamide
- Niacinamide
- Sodium azide,  $NaN_3$
- Sodium chloride, NaCl
- Sodium dihydrogenphosphate,  $NaH_2PO_4$
- Sodium dodecylsulphate, SDS
- Sodium hydroxide, NaOH
- N,N,N',N'-Tetramethylenethylenediamine, TEMED
- Thiamin
- Tris-(hydroxymethyl)-aminomethane, TRIS
- Triton X-100
- Tryptone
- Urea
- Yeast Extract

## 3.2. Molecular Biology Techniques

All employed molecular biology protocols, if not mentioned here, were as described in Sambrook & Russell (2001).

### 3.2.1 Protocol for competent bacteria

1. Bacteria were streaked on an LB agar plate, and incubated at 37°C overnight.
2. 100 mL of LB medium in a 500mL flask were inoculated with a single colony from the LB fresh plate and incubated at 37°C with shaking (200 rpm). The culture was grown in shaking (200rpm) incubator at 37°C until the OD600 was between 0.5 – 0.6 (approximately 3 hours).
3. The culture was aseptically transferred to two chilled, sterile 50 mL polypropylene tubes and incubated on ice for 10 min. Thereafter centrifugation followed at 4 000 g for 10min. at 0 – 4°C.
4. Supernatant was decanted, and tubes put back on ice. Each cell pellet was resuspended in approximately 10mL of cold (0 – 4°C) sterile 0.1 M CaCl<sub>2</sub> and subsequently centrifuged like before.
5. The cell pellet was then resuspended in 2mL of cold sterile 0.1 M CaCl<sub>2</sub> (2 mL for each 50 mL of original culture).
6. Using a pre-chilled pipette the cell suspension was aliquoted (50 µL) to pre-chilled 1.5mL tubes and frozen immediately in liquid nitrogen. The aliquots were kept at – 80°C ready for use.

### 3.2.2 Transformation of the competent bacteria

1µL of plasmid DNA solution in water was mixed together with the 50 µL aliquot of competent bacteria and stored on ice for 30min. After ice incubation they were heated in a 42 °C water bath for 45-60sec, and then placed on ice for 2 minutes. 0.5 mL of SOB broth preheated to 42°C was added to each tube. The tubes were then shaken at 200 rpm at 37 °C for 90min. 50µL of each transformation was spread on a Lb agar plate with the appropriate antibiotic and inoculated overnight at 37 °C.

### 3.2.3 Bacterial cultures

#### Bacterial Culture in rich medium:

In this thesis work the rich media used were LB, SLBH, unlabeled and <sup>15</sup>N labelled -Celtone (Nalgene).

1. 50mL LB with the appropriate antibiotic were inoculated with a fresh single bacterial colony and incubated overnight at 37°C with vigorous shaking (240 rpm) in a 500mL flask.
2. 1L LB with the appropriate antibiotic was inoculated with 10mL of the overnight culture, supplemented with appropriate antibiotic, and incubated at 37°C with shaking (180 rpm) in a 5L flask until the OD<sub>600</sub> reached the appropriate value.
3. When the expected OD was reached, induction by IPTG addition (refer to the text for final concentration) was followed. After time induction (for the time of induction refer to the text) cells were harvested by centrifugation at 8000 g for 20min and stored at -20°C.

#### **Bacterial Culture in minimal medium:**

For growth of bacteria in minimal medium were used two procedures: the first resembles the protocol used for rich media in which minimal medium solutions were used instead of the rich medium, the second is the so called “two-stage protocol” (Marley *et al.*, 2001).

The general protocol implemented is the following: the cells were collected from the 50 mL overnight culture, re-suspended in 15 mL of supernatant and inoculated in 1L of LB/Kan. When OD<sub>600</sub> of 0.6-0.8 was reached, cells were pelleted by centrifugation at 3000 g for 20 min. The cells were then washed and pelleted using an M9 solution with non-labeled NH<sub>4</sub>Cl. The cell pellet was re-suspended in isotopically labelled medium and then incubated to allow for the recovery of growth and the clearance of unlabeled metabolites. The volume of the fresh minimal medium solution was then reduced to one fourth as it is reported to give the highest protein yields (e.g. starting from 1L LB, 250 mL minimal medium were used). Protein expression was induced after 1 hour by addition of 0.7 mM IPTG in the same way as in rich media.

Cells are collected after 4-5 hours by centrifugation at 3000 g for 20min and stored at -20°C.

#### **3.2.4 Site-directed mutagenesis**

For site-directed mutagenesis was used the Stratagene Kit. The QuickChange site-directed mutagenesis method is performed using *PfuTurbo* DNA polymerase and a temperature cycler. The basic procedure utilises a supercoiled double-stranded DNA vector with an insert of interest and two synthetic oligonucleotide primers containing the desired mutation (step 1 in Figure 3.1). The oligonucleotide primers, each complementary to opposite strands of the vector, are extended during temperature cycling by *PfuTurbo* DNA polymerase (step 2).

Incorporation of the oligonucleotide primers generates a mutated plasmid containing staggered nicks. Following temperature cycling, the product is treated with *Dpn* I (step 3). The *Dpn* I endonuclease is specific for methylated and hemimethylated DNA and is used to digest

the parental DNA template and to select for mutation-containing synthesized DNA. DNA isolated from almost all *E.coli* strains, including the BL21(DE3) strain, is dam methylated and therefore susceptible to *Dpn* I digestion. The nicked vector DNA containing the desired mutations is then transformed into XL1-BLUE supercompetent cells (step 4).

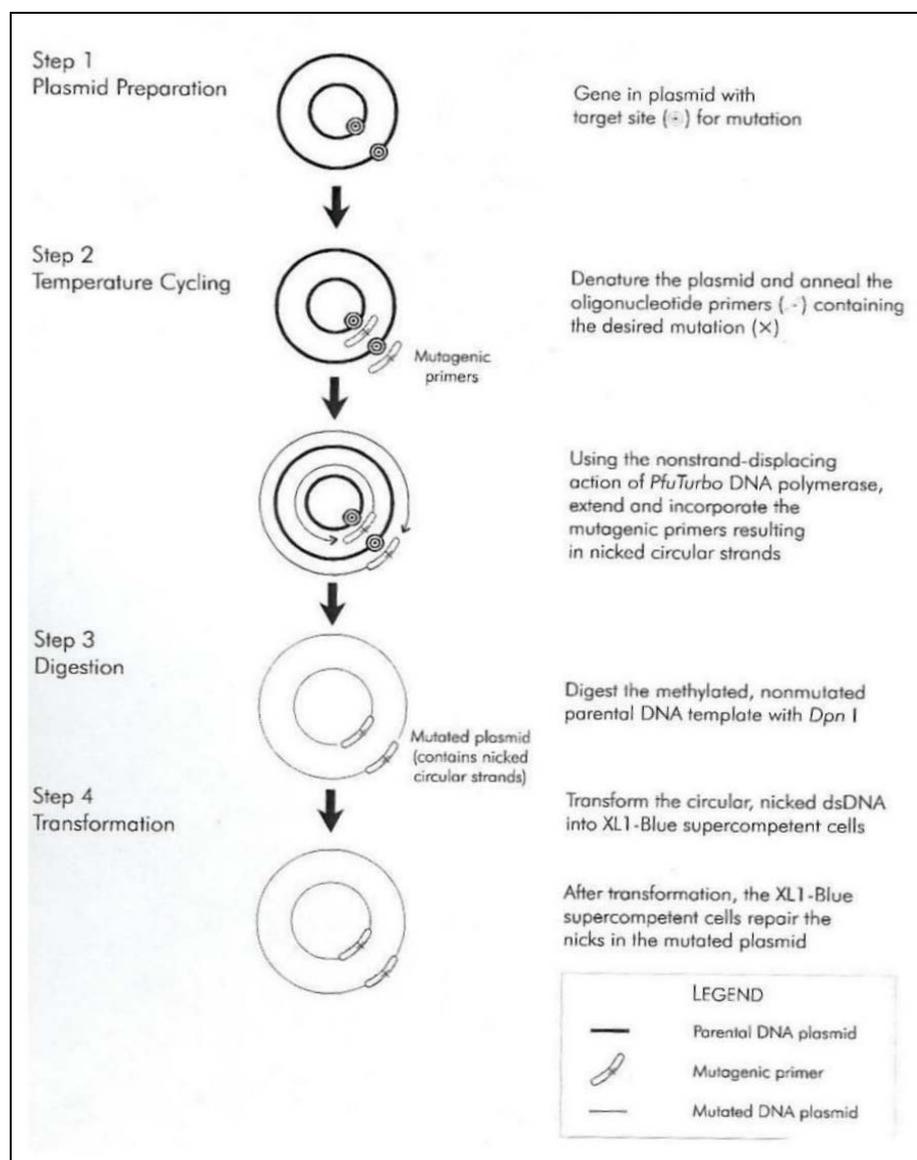
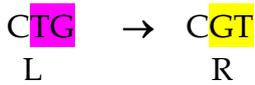


Figure 3.1. Overview of the QuickChange site-directed mutagenesis method (from the Stratagene manual 2002).

### Primer design

The mutagenic oligonucleotide primers used in the mutagenesis were designed according to the mutations. Here are reported the designed primer sequences and the relative terminal parameters.

**L21R :**

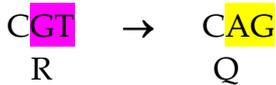
Fw: 5' GAA GAA TTC CTG AAA GCT **CGT** GCT CTG CCG GAA GAC CTG – 3'  
 Rev : 5' CAG GTC TTC CGG CAG AGC **ACG** AGC TTT CAG GAA TTC TTC – 3'

$$\%GC = 21/39 = 53.85\%$$

$$N = 39$$

$$\%Mis = 2/39 = 5.13\%$$

$$T_m = 81.5 + 0.41(53.85) - 675/39 - 5.13 = 81,14$$

**R120Q :**

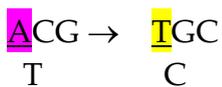
Fw : 5' GTG GTG TTA CCC TGA TCC AGC GTT CTA AAC GTG TTT GAG – 3'  
 Rev : 5' CTC AAA CAC GTT TAG AAC GCT GGA TCA GGG TAA CAC CAC – 3'

$$\%GC = 19/39 = 48,7\%$$

$$N = 39$$

$$\%Mis = 2/39 = 5.13\%$$

$$T_m = 81.5 + 0.41(48.7) - 675/39 - 5.13 = 79$$

**T91C :**

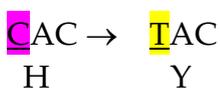
Fw : 5' CC AAA TCT GAA AAG TTC TCT **TGC** GAA CAG GAA GTT AAA GGT AAC  
 Rev : 5' GTT ACC TTT AAC TTC CTG TTC GTA AGA GAA CTT TTC AGA TTT GG

$$\%GC = 16/44 = 36,36\%$$

$$N = 44$$

$$\%Mis = 1/44 = 2,27\%$$

$$T_m = 81,5 + 0,41 \cdot 36,36 - 675 / 44 - 2,27 = 78,79$$

**H98Y :**

Fw : 5' CC AAA TCT GAA AAG TTC TCT **TAC** GAA CAG GAA GTT AAA GGT AAC  
 Rev : 5' GTT ACC TTT AAC TTC CTG TTC GTA AGA GAA CTT TTC AGA TTT GG

% GC =  $16/44 = 36,36\%$

N = 44

% Mis =  $1/44 = 2,27\%$

$T_m = 81,5 + 0,41 \cdot 36,36 - 675 / 44 - 2,27 = 78,79$

#### PCR MIX:

Reaction buffer 10X	5 $\mu$ l	5 $\mu$ l
pIDNA template	50 ng	20 ng
Primer forward	125 ng	125 ng
Primer reverse	125 ng	125 ng
dNTP mix	1 $\mu$ l	1 $\mu$ l
Distilled and steril water till a final volume of 50 $\mu$ l		
<i>Pfu</i> Ultra DNA polimerase (2.5 U/ $\mu$ l)	1 $\mu$ l	1 $\mu$ l

#### Termal cycle:

1° segment	1 cycle	95°C for 2 minutes
2° segment	18 cycles	95°C for 30 seconds
		55°C for 1 minutes
		68°C for 12 minutes

### 3.3. Tools of Biochemistry

All biochemical methods that are not mentioned here were performed exactly according to Sambrook & Russell (2001).

#### 3.3.1 SDS Polyacrylamide Gel Electrophoresis (SDS PAGE)

The glycine SDS PAGE and native PAGE was performed exactly like described in Sambrook & Russell (2001).

##### Cellular pellet sample preparation

For electrophoresis were taken 1,5 mL from the culture solution, centrifuged at maximum speed for 5min. The surnatant was discarded accurately and the cellular pellet was resuspended in 100  $\mu$ L 2x SDS PAGE. Samples were heated for 20min and 15 $\mu$ L from every

sample was loaded onto the gel walls. For the quantitative comparison of the overexpressed bands on SDS PAGE, it were loaded on each gel wall an equivalent concentration of cells, calculated by OD<sub>600</sub> measures doing dilutions when necessary.

### Protein sample preparation

For electrophoresis were taken 20 $\mu$ L of samples from the solution to test. 20 $\mu$ L samples were mixed with 10 $\mu$ L of the 3x SDS PAGE loading buffer and heated for 5 min. 15 $\mu$ L from every sample was loaded onto the gel walls.

### Staining and destaining of Proteins

Staining of SDS PAGE and native PAGE proteins was performed with Coomassie Blue G250 like described in Sambrook & Russell (2001).

### Protein concentration

The concentration devices used in this work were Amicon ultra-filtration devices (400mL, 50 mL, 10 mL) supplied by YM-3 membranes with a nominal cut-off of 3000 Da.

## 3.3.2 Determination of protein concentration

The concentration of proteins in solution was estimated by UV spectroscopy. Extinction coefficient corresponding to 1 mg/mL solution of all proteins were obtained using composition analysis performed with the Protparam tool program (<http://www.expasy.org/tools/protparam>)

cl-BABP  $\Rightarrow$  theoretical extinction coefficient  $\epsilon_{280}$  8250 M<sup>-1</sup> cm<sup>-1</sup>

L21R cl-BABP  $\Rightarrow$  theoretical extinction coefficient  $\epsilon_{280}$  8250 M<sup>-1</sup> cm<sup>-1</sup>

R120Q cl-BABP  $\Rightarrow$  theoretical extinction coefficient  $\epsilon_{280}$  8250 M<sup>-1</sup> cm<sup>-1</sup>

T91C cl-BABP  $\Rightarrow$  theoretical extinction coefficient  $\epsilon_{280}$  8250 M<sup>-1</sup> cm<sup>-1</sup>

H98Y cl-BABP  $\Rightarrow$  theoretical extinction coefficient  $\epsilon_{280}$  9530 M<sup>-1</sup> cm<sup>-1</sup>

Glycodelin  $\Rightarrow$  the theoretical extinction coefficients relative to the three constructs are:

Molecular mass	Ext. coefficient 280 nm
24495,9 Da	29400 M <sup>-1</sup> cm <sup>-1</sup>
19921 Da	22710 M <sup>-1</sup> cm <sup>-1</sup>
22985,3 Da	29700 M <sup>-1</sup> cm <sup>-1</sup>

After thoroughly mixing of the sample, the absorbance at 280 nm was measured. As a reference the same buffer of the protein sample was used.

### 3.3.3 Lysis of *E.coli*

Ten grams of wet cells from 1 liter culture was pelleted and re-suspended with 25 mL of fresh Lysis Buffer (2.5 mL/gr of cell) and maintained at 0 °C in an ice-bath to prevent heating of the solution during cell disruptions. The solution was subjected to sonication at maximum sonicator power for 10x10s, causing cells to undergo lysis. The lysed cell suspension (total fraction) was centrifuged at 20min at 4°C with 10000 G, causing the larger cellular debris to be pelleted. For recombinant protein expressed in soluble form (cl-BABP) the supernatant was used for subsequent purification procedure; for recombinant protein expressed as inclusion bodies the supernatant was discarded and the pellet was washed 2-3 times with Triton wash solution. Then the inclusion bodies were dissolved in the denaturing solution and used for subsequent purification procedure.

### 3.3.4 Chromatography Techniques

All chromatography procedures were carried out at 4-10 °C, all buffers were 0,45 µm filtered, degassed and contained 0.02% NaN<sub>3</sub>.

#### **Anion-exchange chromatography (cl-BABP purification)**

After lysis the supernatant was applied to a XK 26/40 column. The column was washed with 50 mM Tris-CH<sub>3</sub>COOH buffer pH 7.8 at the flow rate of 1 mL/min and cl-BABP is contained in the unbound fraction. The active fractions were pooled and concentrated by means of the centrifugal ultrafiltration with MW cut off at 3,000 and kept in an ice bath for the next step of purification. The flow-through was collected in 5 mL fractions.

#### **IMAC chromatography (Glycodelin purification)**

A XK16/20 column packed with 10 mL of fast-flow Sepharose was packed using 5 column volumes of distilled water, charged with 0.5 column volumes of 200 mM NiSO<sub>4</sub>, washed with five column volumes of water and equilibrated with binding buffer (50 mM Tris/HCl, 500 mM NaCl, 20 mM imidazole, pH 8.0). Then the solubilized sample containing a total of 1-10 mg of protein with a mean concentration of 1 mg/mL was applied on to the column. The flow rate was of 1 mL/min and the flow-through was collected in 5 mL fractions. The column was washed till UV<sub>280</sub> reaches the value of baseline. Then the bound protein were eluted by using a linear imidazole gradient from 20-400 mM.

**Gel filtration chromatography (cl-BABP and Glycodelin purifications)**

Gel filtration chromatography was used for the last step of purification. The Sephacryl 200 HR 26/100 column with 90 cm packed bead size volume was pre-equilibrated with 50 mM Tris/HCl, 200 mM sodium chloride buffer pH 7.2. The concentrated protein solution from the previous steps (maximum volume of 5 mL) were applied. The proteins were eluted from the column with the same buffer at a flow rate of 1 mL/min. Fractions containing cl-BABP were determined by SDS PAGE and stored at  $-20^{\circ}\text{C}$ .

**Delipidation procedure (cl-BABP)**

A column (Pharmacia, XK16-20) was packed with Lipidex 1000 and equilibrated at  $37^{\circ}\text{C}$  with 10 mM  $\text{Na}_2\text{HPO}_4/\text{NaH}_2\text{PO}_4$  at pH 7.4, following the procedure reported in Glatz & Veerkamp (1983). A sample of maximum 15 mg, diluted to 1 mg/mL, was loaded onto the column at 12 mL/h and eluted at the same flow rate. The resin was regenerated with methanol and washed extensively with buffer before a new purification.

**Molecular weight determination by gel filtration chromatograph**

The relative molecular weight ( $M_r$ ) of the native enzyme was determined by using Sephacryl 200 HR column. Elution was done at the flow rate of 0.25 mL/min with an elution buffer comprising 50 mM sodium phosphate buffer pH 7.0 and 0.15 M NaCl. The calibration curve was constructed using protein markers: cytochrome C (12,400), carbonic anhydrase (29,000), bovine serum albumin (66,000), alcohol dehydrogenase (150,000) and b-amylase (200,000). Dextran blue (2,000,000) and vitamin B<sub>12</sub> (1,355.4) were used to determine the void volume ( $V_0$ ) and total volume ( $V_t$ ), respectively. A calibration curve between log molecular weights of protein markers and the partition coefficient values,  $K_{av}$ , was constructed.

**Storage of plDNA**

All the extracted plasmids were stored in distilled sterile water at  $-20^{\circ}\text{C}$ . Transformed *E. coli* cells were kept in glycerol stock at  $-80^{\circ}\text{C}$ .

**Storage of purified proteins**

Purified proteins were lyophilised on Edwards Lio 5P under 0.05 atmospheres; the protein powders were stored at  $0^{\circ}\text{-}4^{\circ}\text{C}$ .

### 3.4. NMR Samples Preparation

If not otherwise indicated, the samples for NMR spectroscopy were concentrated and dialyzed against PBS buffer. Typically, the sample concentration varied from 0.3 to 1.0 mM.

Before measuring, the sample was centrifuged in order to sediment aggregates and other macroscopic particles. 450 $\mu$ L of the protein solution were mixed with 50 $\mu$ L of D<sub>2</sub>O (5-10%) and transferred to an NMR sample tube.

#### **Phosphate-Buffered Saline (PBS) Buffer:**

10mM Na<sub>2</sub>HPO<sub>4</sub>\*7H<sub>2</sub>O/ NaH<sub>2</sub>PO<sub>4</sub>, pH 7.0

30 mM Na<sub>2</sub>HPO<sub>4</sub>\*7H<sub>2</sub>O/ NaH<sub>2</sub>PO<sub>4</sub>, pH 7.0

0.02 % NaN<sub>3</sub> g

#### **Complex of cL-BABP with [1-<sup>13</sup>C]palmitic acid**

The complex of cL-BABP with palmitic acid was performed as follow: 1 mg of [1-<sup>13</sup>C]palmitic acid was dissolved in chloroform and dispensed in a glass tube. After the organic solvent was evaporated under nitrogen flow, 800  $\mu$ L 1 mM solution of cL-BABP (20 mM Na<sub>2</sub>HPO<sub>4</sub>/NaH<sub>2</sub>PO<sub>4</sub> buffer at pH 7) was added to the tube and the mixture of cL-BABP/palmitic acid was incubated overnight at 37°C.

#### **Complex of cL-BABP with glycochenodeoxycholic acid**

The complex of cL-BABP with palmitic acid was performed as follow: it was prepared a 0.022 M stock solution of glycochenodeoxycholic acid dissolved in tetrahydrofuran Then the appropriate amounts of stock solution was aliquoted and the solvent was evaporated under under nitrogen flow. The bile salt was solubilized with 1:1 equivalent of 1M NaOH and brought up to a volume of 60  $\mu$ L in a buffer containing 30mM potassium phosphate at pH 7.0 300  $\mu$ L of cL-BABP protein solution (0.5 mM) in 30mM potassium phosphate at pH 7 were then added and incubated overnight at room temperature.

## 4. Results and Discussion

The subject of my thesis work has been the development of high level expression protocol of recombinant cL-BABP and relative site-directed mutants for structural characterization by NMR spectroscopy. A second part of this work has been concerned with the expression and purification of GdA. In this chapter the main steps involved in the achievement of these objectives are described:

- a. the development of high level expression and purification protocols of unlabeled cL-BABP;
- b. the development of the high level expression protocol of cL-BABP in minimal medium for the subsequent  $^{15}\text{N}$  and  $^{13}\text{C}$  isotopic enrichment;
- c. the production of  $^{15}\text{N}$  and  $^{13}\text{C}$  enriched protein;
- d. the production of several site-directed mutants of cL-BABP in order to clarify the role of specific residues;
- e. identification of the proper conditions for correct refolding of proteins expressed in inclusion bodies;
- f. development of the high level expression and purification protocols of unlabeled GdA.

### 4.1. Description of cL-BABP Expression System

CL-BABP clone was kindly provided by Dr. Jefferson Foote (Fred Hutchinson Cancer Research, Seattle, USA). Dr. Foote cloned cL-BABP in pET24d (Figure 4.1); in this work, the *E. coli* strain BL21(DE3) was used as host expression system.

The nucleotide sequence is reported (Figure 4.2); here are described some preliminary considerations that were done on the main features of the expression vector, in order to evaluate if it was suitable for high level expression.

- a. Vector type: pET-plasmid type is an high-level expression vector; it was originally developed by Studier *et al.* (1986). The target gene is under the control of the T7 promoter, that is not recognised by *E. coli* RNA polymerase. *E. coli* strain BL21(DE3) carries a chromosomal copy of the T7 RNA polymerase gene under the control of the lacUV5 promoter. Addition of IPTG induces the expression of the T7 RNA polymerase and the subsequent transcription of the recombinant protein.

b. The cDNA sequence of the original clone coded for a mutant of cL-BABP in which the tryptophan 6 of the wild type protein was modified in tyrosine (cL-BABP W6Y). During my previous work the cDNA sequence was restored to the wild type.

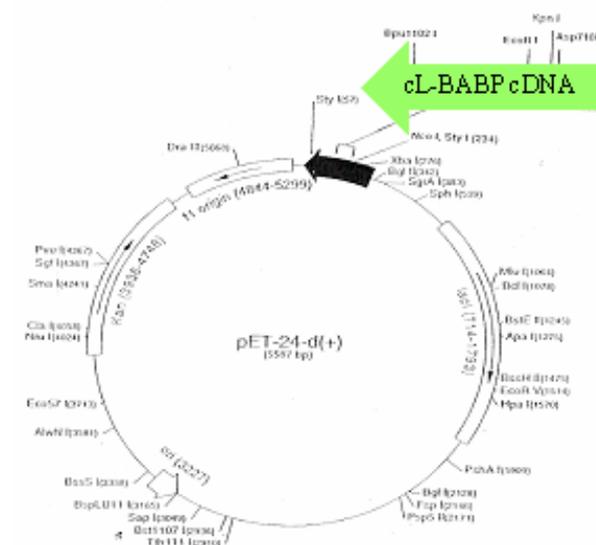


Figure 4.1. Map of the expression vector pET24d and position of the cL-BABP cDNA.

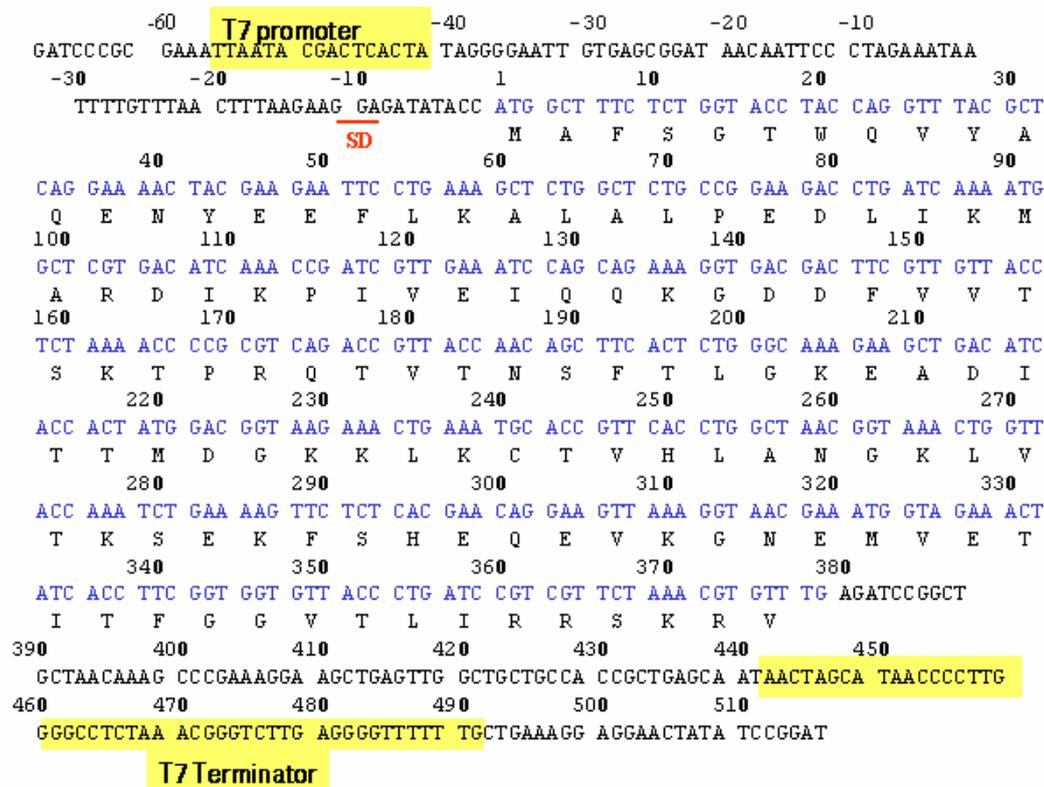


Figure 4.2. Nucleotide sequence of the full length cL-BABP cDNA.

c. It was observed that the cL-BABP cDNA coding sequence was obtained from gene synthesis, since each amino acid is codified respectively by a single type of codon sequence. The coding sequence was therefore “codon-optimized”, a procedure that refers to the alteration of the original gene sequence to make codon usage match the available tRNA pool within the cell/species of interest, in the present work *E. coli*.

d. The spacing between the Shine-Dalgarno (SD) sequence and the sequence initiation codon AUG has been detected to be important for the efficient translation by *E. coli* (Gold *et al.*, 1981, Chen *et al.*, 1994): the best efficiency is reached when this distance resemble that found in natural mRNAs of *E. coli*, as reported in Table 4.1. In our expression vector the distance between the SD and the start codon is of 8 nucleotides, observed to be in the range of optimal spacing.

**Table 4.1** <sup>1</sup>Optimal spacing and <sup>2</sup>aligned spacing between the SD and the initiation codon in different systems. The SD sequence present in cL-BABP vector is evidenced in the red square (reproduction from Chen *et al.*, 1994).

SD	Spacing variation	Optimal spacing	Optimal aligned spacing
. AAGGG . . .	2–15	7	4
. . <span style="border: 1px solid red;">GGA</span> . . .	5–8	8	5
. . AGGA . . .	8–34	8–10	5–7
. AAGG . . .	6–23	8–13	4–9
. AAGGA . . .	4–12	7	4
UAAGGAGG .	4–12	7–8	6–7
. . AGGA . . .	7–17	7–9	4–6
. AAGGGU . .	2–15	9	7
<b>UAAGGAGGU*</b>	3–11	5	5
. AAGGAGGU	3–11	7	7
. . AGGA . . .	8–20	8–9	5–6
. AAGGGU . .	7–13	8–9	6–7
. AAGG . . . .	7–14	9	5

<sup>1</sup>The SD-AUG spacing is the number of nucleotides separating each SD sequence in the given mRNA from the Adenine of the AUG initiation codon.

<sup>2</sup>The SD-AUG aligned spacing is the number of nucleotides separating the nucleotide corresponding to SD<sub>ref</sub> (indicated by \*) to the Adenine of the AUG.

From these observations the plasmid vector was considered suitable for the high level expression of recombinant cL-BABP.

If not mentioned, the procedures described in the next paragraphs were performed exactly as described in chapter 3.

## 4.2 Development of the High Level Expression and Purification Protocols of Unlabeled cL-BABP from Rich Medium

In order to decide the conditions for expression of consistent quantities of recombinant cL-BABP the following preliminary tests were done:

- Monitoring cell growth
- Expression test in the time course
- Solubility test

### 4.2.1 Cell growth

The growth of the cells was monitored measuring the  $OD_{600}$  as a function of time to establish the logarithmic and the stationary phase. A crucial point in protein expression is that induction must occur into the first half of the logarithmic phase.

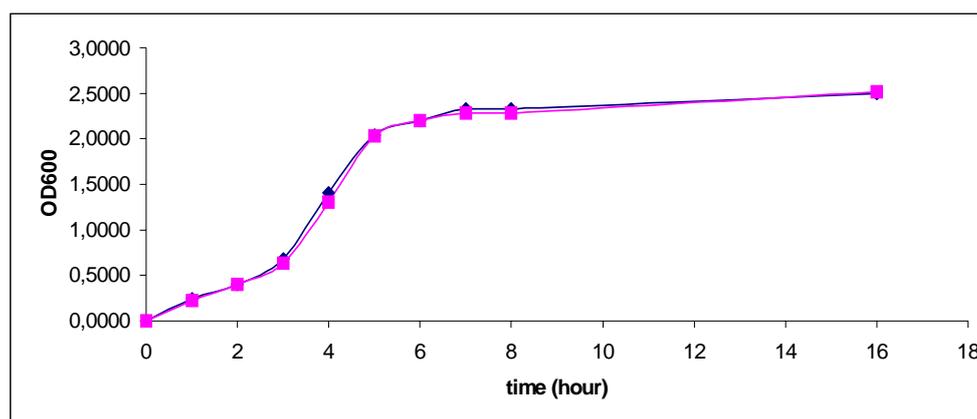
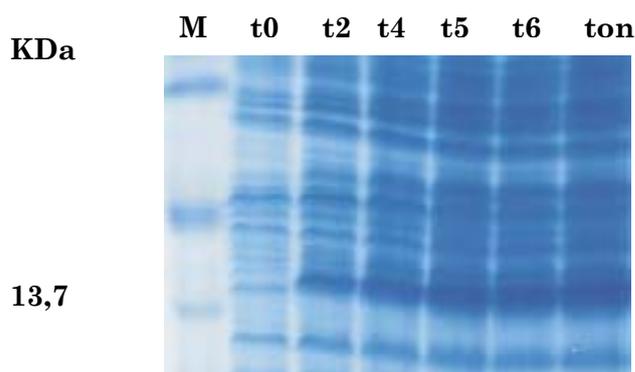


Figure 4.3. Growth of *E. coli* cells at 37°C. LB/kan medium was inoculated with a single bacterial colony from a fresh LB/Kan agar plate and incubated with shaking (180 rpm). In the figure is shown  $OD_{600}$  as a function of time. Measurements on a 200 ml culture sample in a 1L flask are plotted in pink, measurements on a 700 ml in a 3 L flask culture sample are plotted in blue.

From the observation of the graph in figure 4.3, the identified  $OD_{600}$  range for a correct induction was between 0.5 and 0.8.

### 4.2.2 Expression Test –Time Course

Cell growth was monitored until the  $OD_{600}$  of 0.7 was reached. At that time ( $t = 0$ ) the culture was induced with 1 mM IPTG (end concentration). The culture was grown overnight. 1 mL samples for electrophoresis were taken before induction ( $t = 0$ ) and after 2, 4, 5, 6 hours and overnight induction ( $t_2, t_4, t_5, t_6, t_{on}$ ). Cells were centrifuged and the pellet was prepared for SDS PAGE (Figure 6.4).



**Figure 4.4.** Expression test for cL-BABP by coomassie stained SDS-PAGE. Protein samples were taken after 2hs 4hs, 5hs, 6hs, on induction. Lane 1: molecular weight marker; lane2: t0; lane 3: t2 lane 4: t4; lane 5: t5 lane 6: t6; lane 7: t overnight.

It was observed that protein expression increases in the time course till overnight induction.

### 4.2.3 Solubility Test

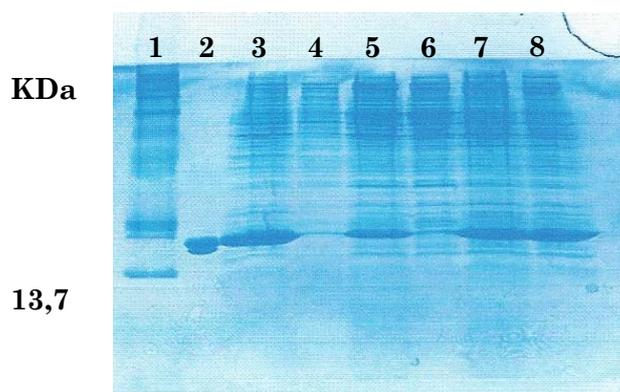
It was done a preliminary screening of the expression conditions. Cultures of bacteria containing tested construct for protein expression were grown similar to the conditions of the previous paragraphs. For every test the varied parameters are presented in table 4.2; two temperatures were tested (20 and 37°C). For each of the presented set 10 mL samples were collected by centrifugation for 10min at 4°C, 6000 g. The pellets were resuspended in 3 mL Lysis Buffer and sonicated with a maximum sonicator power for 3x10s keeping in ice-bath, in order to disrupt the cells.

The suspension, representing the “total fraction”, was then centrifuged for 20min at 4°C, 10000 g, and the supernatant represents the “soluble fraction”. Cells were centrifuged and the pellet was prepared for SDS PAGE.

<b>Table 4.2</b> Conditions tested for the optimization of the expressed protein solubility. All given sets of parameters were tested both for 20°C and 37°C.			
parameters set number	culture induced at OD <sub>600</sub>	induction with IPTG end concentration [mM]	time from induction to harvest [h]
1	0.6	0.7	4
2	0.75	0.7	4
3	1	0.7	4
4	0.75	0.3	4
5	0.75	0.7	4
6	0.75	1.5	4

parameters set number	culture induced at OD <sub>600</sub>	induction with IPTG end concentration [mM]	time from induction to harvest [h]
7	0.75	0.7	2
8	0.75	0.7	4
9	0.75	0.7	overnight

The total and soluble fractions of each set of parameter were compared in SDS PAGE, the most meaningful results are shown in SDS PAGE of Figure 4.5:



**Figure 4.5.** SDS PAGE of set parameter n°8 and 9 at 20°C and 37°C (only 4 hours induction at 37°C is shown). Lane 1: molecular weight marker; lane 2: cL-BABP marker; lane 3 and 4: total and soluble fraction of LB, 37 °C after 4 hours; lane 5 and 6: total and soluble fraction of LB, 37 °C overnight; lane 7 and 8: total and soluble fraction of LB, 20 °C overnight.

The best conditions of over-expression were found to be in the range of OD<sub>600</sub> = 0.6-0.75, 0.7 mM IPTG, at 20 °C overnight; induction at 37°C already after 4 hour lead to the formation of inclusion bodies.

The work proceeded with a large scale cL-BABP over-expression in 1 liter of LB/Kan using the determined parameters. After overnight induction at 20°C cells were harvested by centrifugation, the supernatant was discarded and the pellet was resuspended in lysis buffer for cell disruption. The solution was then centrifuged and the supernatant used for the subsequent purification procedure.

#### 4.2.4 Purification Procedure

The purification procedure was determined during my previous work and during this thesis work it has been well standardized. After lysis the soluble fraction containing cL-BABP was loaded onto a DE52 anion exchange column (40 x 2,6 cm) equilibrated with 50 mM

Tris/CH<sub>3</sub>COOH, pH 7.8. The same buffer was used for protein elution. cL-BABP was eluted in the unbound fraction (Figure 4.6).

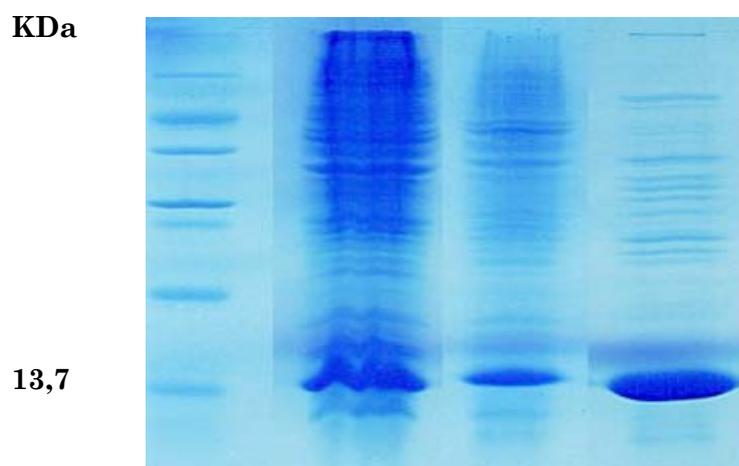


Figure 4.6. SDS-gel showing: lane 1: molecular weight marker; lane 2: total fraction; lane 3: soluble fraction; lane 4: unbound cL-BABP fraction.

The fractions containing cL-BABP were pooled, concentrated to 5 mL, and applied on a Sephacryl S-100 HR column (90 x 2,6 cm) equilibrated with 50 mM Tris/HCl, 0.2 M NaCl at pH 7.2 and eluted with the same buffer. In Figures 4.7 and 4.8 the details of gel filtration chromatography are shown.

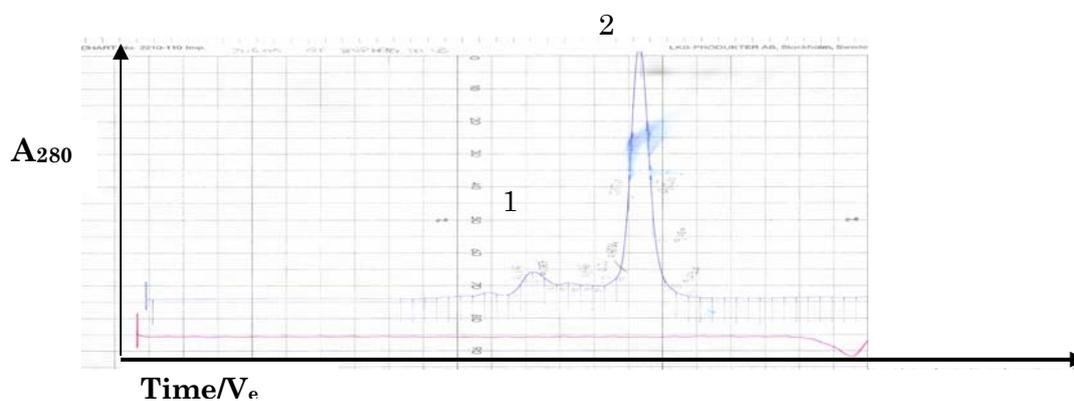


Figure 4.7. Gel filtration of the pooled fractions eluted from AEX chromatography containing cL-BABP. (Gel Sephacryl S-100 HR, flow rate 1 mL/min). cL-BABP is eluted in the peak n° 2, samples have been loaded in SDS PAGE (Figure 4.8).

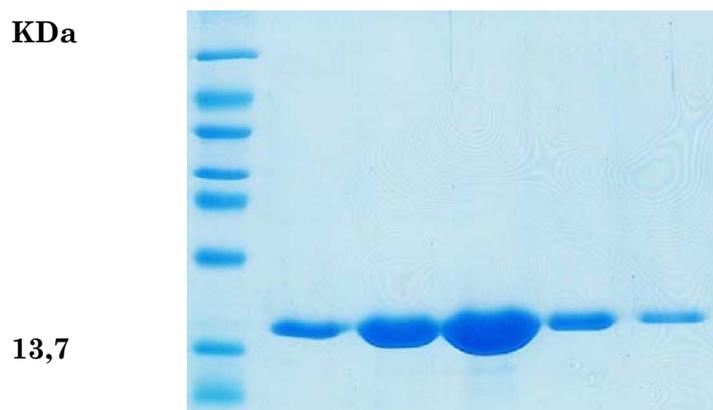
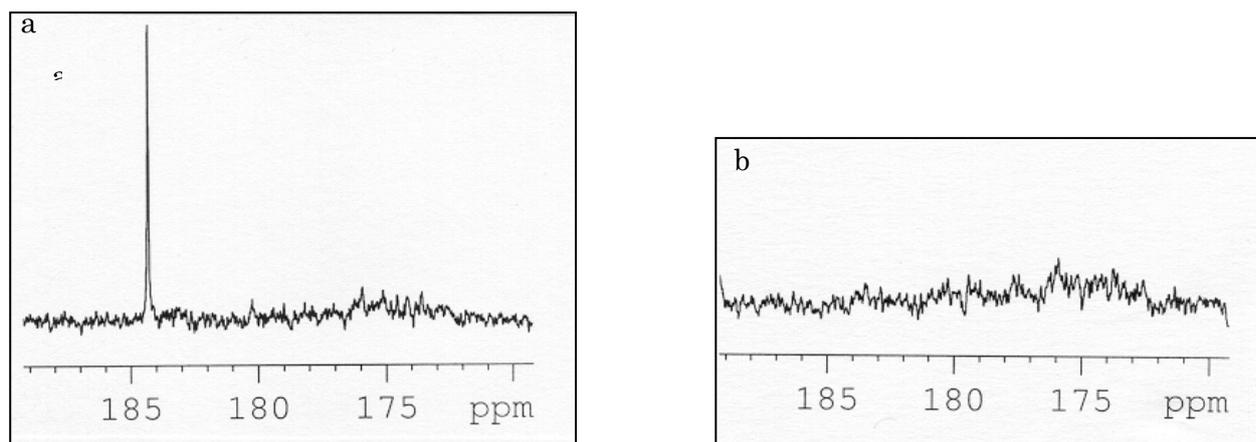


Figure 4.8. SDS-PAGE lane 1: molecular weight marker; lanes 2-6: fractions of peak n°2 (Figure 4.7) containing cL-BABP

### 4.2.5 Delipidation Procedure

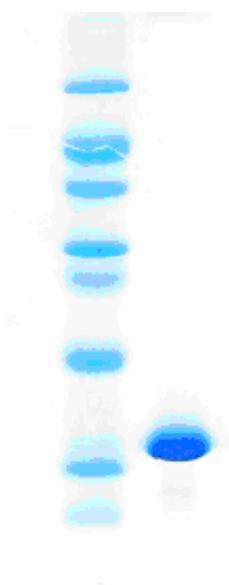
The last step was the delipidation of recombinant cL-BABP. The recombinant protein needs therefore to be delipidated since we are interested in the pure *apo* cL-BABP. It is reported in the literature that recombinant FABPs are bound to a mixture of *E. coli* fatty acids (palmitate and stearate are usually the dominant species, representing 50-55% and 23-29% of the ligand population, Sacchettini *et al.*, 1987). The delipidation procedure described by Glatz & Veerkamp (1983) was followed. It was tested the validity of the procedure on bacterial extracts through an NMR experiment: a complex of non-delipidated cL-BABP with [1-<sup>13</sup>C]palmitic acid was prepared in 20 mM PBS buffer at pH 7.4 following the procedure described in chapter 3.

In these conditions, it is reported that cL-BABP can bind palmitic acid (Beringhelli *et al.*, 2001). The <sup>13</sup>C spectrum of the complex acquired at 298 K is reported in Figure 4.9.a. The NMR signal at 184 ppm arises from the carboxyl group of the bound palmitic acid. (In the absence of cL-BABP, palmitic acid is insoluble in aqueous solution and does not give rise to any NMR signal). After recording the spectrum, the complex cL-BABP/[1-<sup>13</sup>C]palmitic was subjected to delipidation as described in 3.4.4. A second <sup>13</sup>C NMR spectrum was acquired (Figure 4.9.b) on the protein sample eluted from the Lipidex column and concentrated to 1 mM in 20 mM PBS buffer at pH 7. The absence of NMR signal indicates that palmitic acid has been correctly removed and confirms the validity of the delipidation procedure for bacterial extract.



**Figure 4.9.** Carboxyl region of the 125.7 MHz 1D proton decoupled  $^{13}\text{C}$  spectrum acquired at 298 K on the complex prepared with non-delipidated cL-BABP and  $[1-^{13}\text{C}]$ palmitic acid, in 20 mM PBS at pH 7, (a) before delipidation; (b) after delipidation.

The final yield, after the purification, was quantified as **90 mg /L** . The protein purity was checked by the presence of a single band on SDS-PAGE (Figure 4.10) and by MALDI analysis (14 081.2 Da, Figure 4.11), in agreement with the theoretical molecular mass calculated with the Protparam tool program (<http://www.expasy.org/tools/protparam.html>).



**Figure 4.10.** SDS PAGE of cL-BABP after the purification procedure

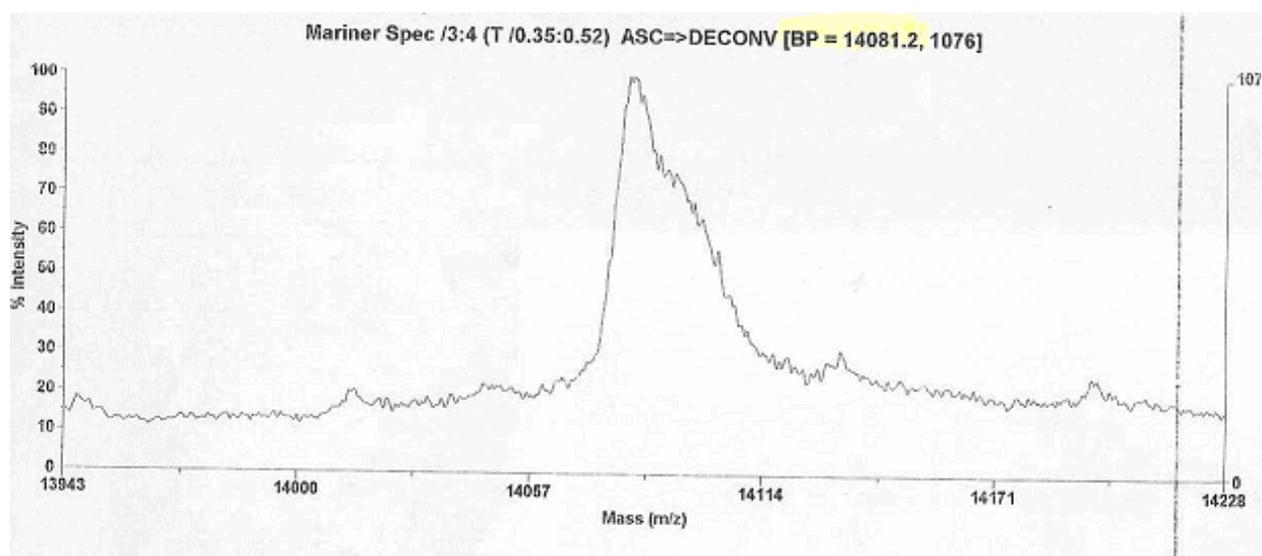


Figure 4.11. MALDI-TOF spectrum of cL-BABP wild tipe.

In order to evaluate the described protocol for over-expression, methods described in the literature for other recombinant FABPs were analysed (Table 4.3).

Table 4.3 A summary of the literature reviews of recombinant FABPs.		
FABP type	Yield	References
Rat I-FABP	7 mg	Lowe 1984
rat H-FABP	30-40 mg/l	Schaap 1996
(L-FABP) (I-FABP)	15 mg (L-FABP) 25 mg (I-FABP)	Storch 1996
Human E-FABP	30 mg/L	Hohoff1999
Human ileal BABP	1,25 g /L (fermenter)	Tochtrop 2003

On the basis of the comparison of the yields reported in literature for proteins of the same family, it could be stated that an high yield over-expression protocol has been developed.

#### 4.2.6 Improvement of cL-BABP yield: other expression tests

##### *Other media*

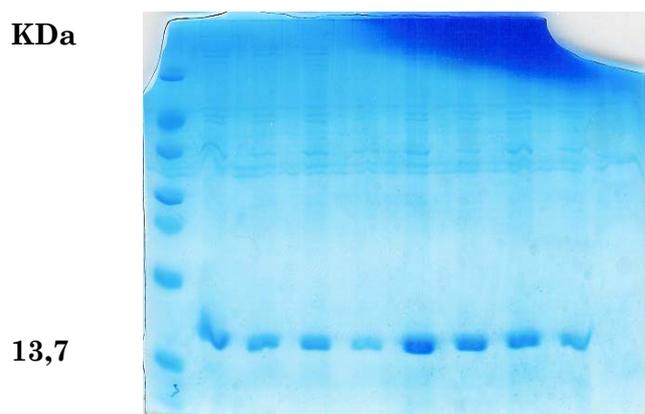
It was evaluated the expression using SLBH medium at the same conditions till now reported; it was not seen any appreciable increase in protein yield.

*Addition of rifampicin*

As reported by Maatman *et al.*, (1994), it was tested, subsequently to induction (45min), the addition to the culture 175 µg/mL rifampicin. Rifampicin and other compounds of the ansamycin group specifically inhibit DNA-dependent RNA polymerase; that is, they prevent the transcription of RNA species from the DNA template. Rifampicin is an extremely efficient inhibitor of the bacterial enzyme while T7 phagic RNA polymerase is not affected. In this way expression of recombinant protein could be favoured. Samples at different conditions were tested on SDS PAGE (Figure 4.11): 5 mL cultures were grown till OD<sub>600</sub> reached 0.7 and induced at final concentration of 0.7 mM IPTG; after 45 minutes rifampicin was added to two samples, one of them was incubated for 2 hours at 37 °C and then at 20 °C overnight, the other straight at 20 °C overnight. For comparison other two samples were induced in the same conditions without addition of antibiotic. The next day the OD<sub>600</sub> was measured in order to evaluate the cell growth.

It was observed that cell growth was inhibited by rifampicin addition.

Growth Conditions	2 h 37 °C + 20°C on	Straight 20 °C on	2 h 37 °C + 20°C on	Straight 20 °C on
OD <sub>600</sub>	1.670	1.425	2.400	1.775



**Figure 4.11.** Expression test with rifampicin. lane 1: molecular weight marker; lane 2 and 3: total and soluble fraction with addition of rifampicine 175 µg/mL, induction 2 hs at 37°C and then at 20°C on; lane 4 and 5: total and soluble fraction of sample with addition of rifampicine 175 µg/mL, induction 20 °C on; lane 6 and 7: total and soluble fraction of sample with no addition of rifampicin, induction 2 hs at 37°C and then at 20°C on; lane 8 and 9: total and soluble fraction of sample with no addition of rifampicin, induction 20°C on.

From SDS PAGE it was observed that the eventual increase in protein expression was negatively compensated by inhibition in cell growth and so the addition of the antibiotic rifampicin cannot be considered a good way for yield improvement.

#### 4.2.7 Separation of cL-BABP from higher aggregates

It is frequently found that proteins pure solutions could contain dimers and higher aggregates. A particular regard was taken for gel filtration procedure, for which it has been determined the calibration curve: gel filtration provide a gentle method for separating the monomer from the aggregates and provides a means of determining the molecular weight or size (Stokes radius) of native or globular proteins under a variety of conditions of pH, ionic strength, temperature and so on. The calibration curve was obtained following the procedure described in chapter 3 (Figure 4.12); the resulting molecular size for cL-BABP resulted of about 18 KDa.

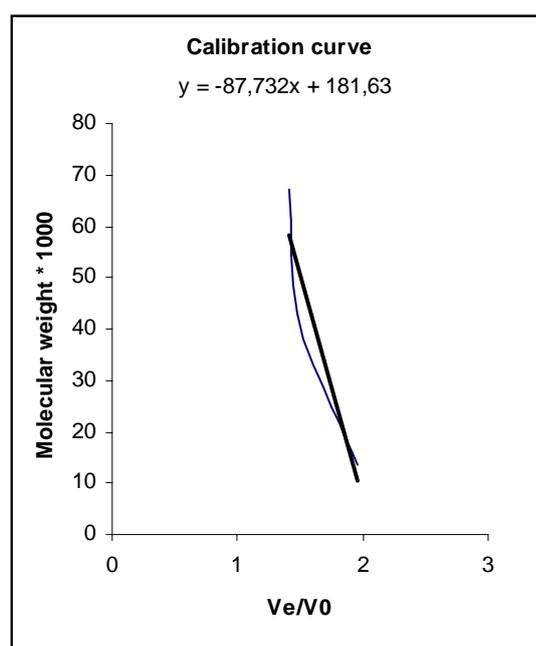
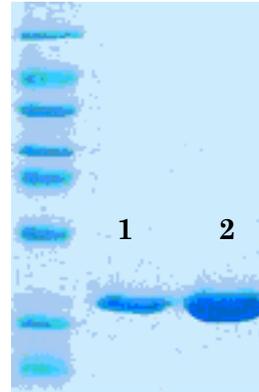
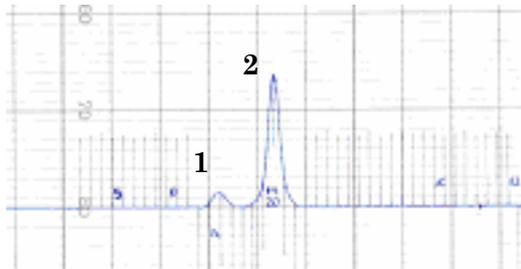


Figure 4.12. Calibration curve obtained with for molecular weight markers (see chapter 3 for details) run on Sephacryl S-100 HR.

Through the determination of the cL-BABP time elution it was possible to observe that under some experimental conditions a small fraction (about 5%) of the purified cL-BABP protein solution aggregates. The elution diagrams of Figure 4.13 is an example showing the elution of peaks of a pure cL-BABP protein solution after gel filtration chromatography.

The conditions favouring aggregation are:

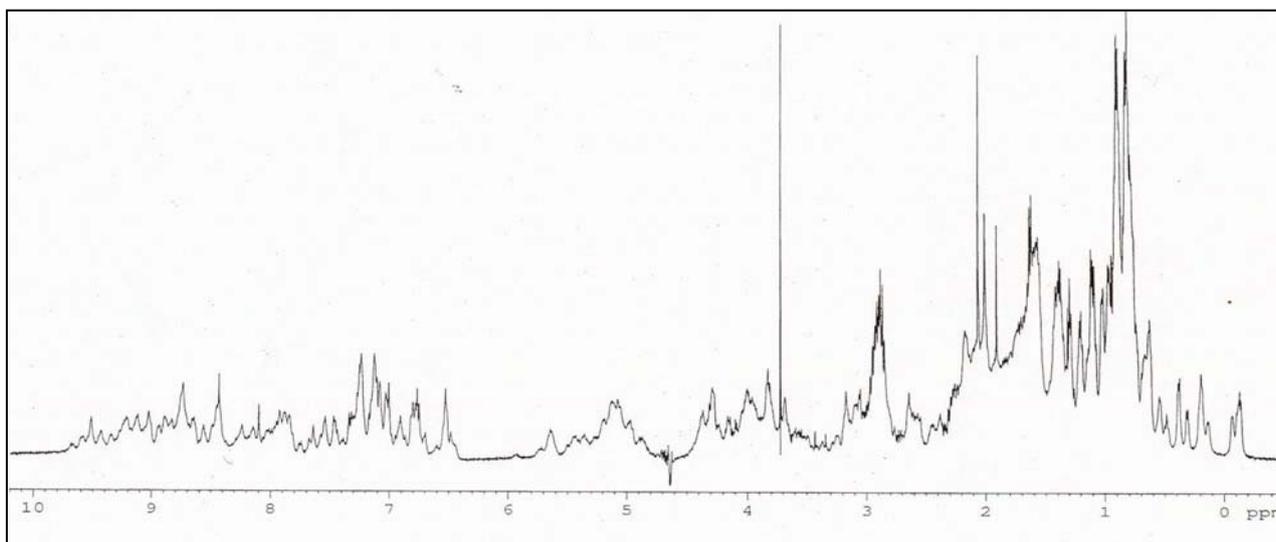
- a single cycle of freeze-thaw at  $-20^{\circ}\text{C}$ ;
- one week of storage at  $0-4^{\circ}\text{C}$  of the purified delipidated cL-BABP solution.
- 



**Figure 4.13.** Gel filtration chromatography of pure cL-BABP : the two peaks (1,2) eluted at different volumes, indicating presence of aggregated since cL-BABP is present in both peaks (detection by SDS PAGE).

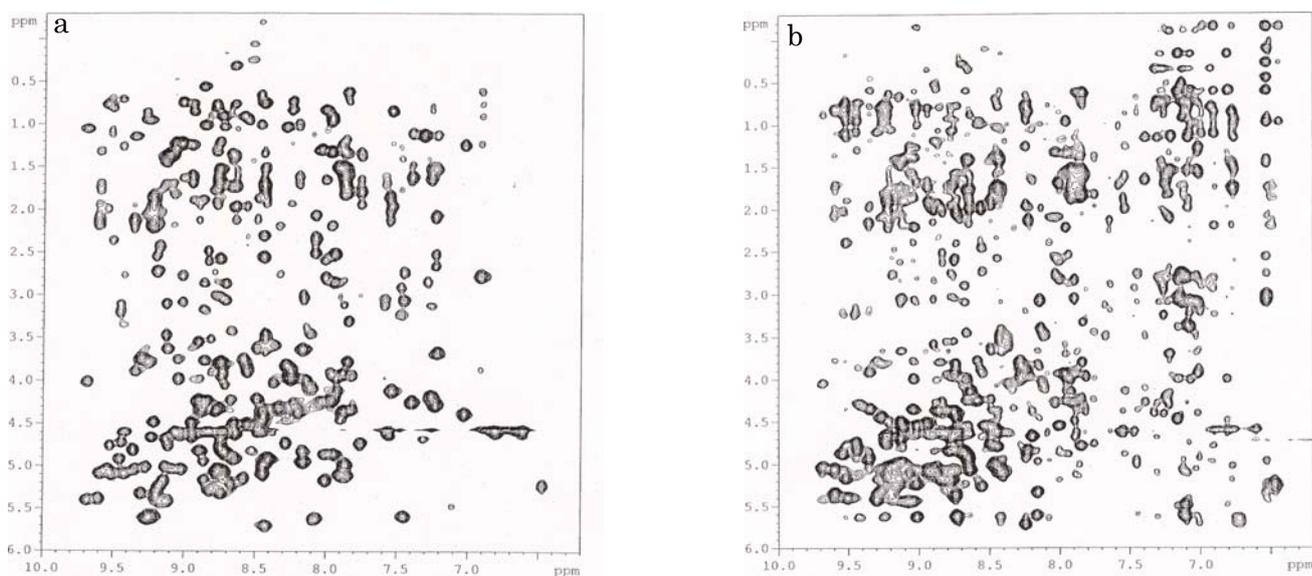
### 4.3. NMR Experiments on unlabeled cL-BABP

The 1D  $^1\text{H}$ -NMR spectrum of the delipidated wt cL-BABP is reported in Figure 4.14. The large chemical shift dispersion and the line width of 5-7 Hz are reliable indicators of folded and non-aggregated protein.



**Figure 4.14.** 1D 500MHz  $^1\text{H}$ -NMR spectrum of 1 mM delipidated wt cL-BABP in 20 mM PBS buffer at pH 5.6 (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), 37 °C.

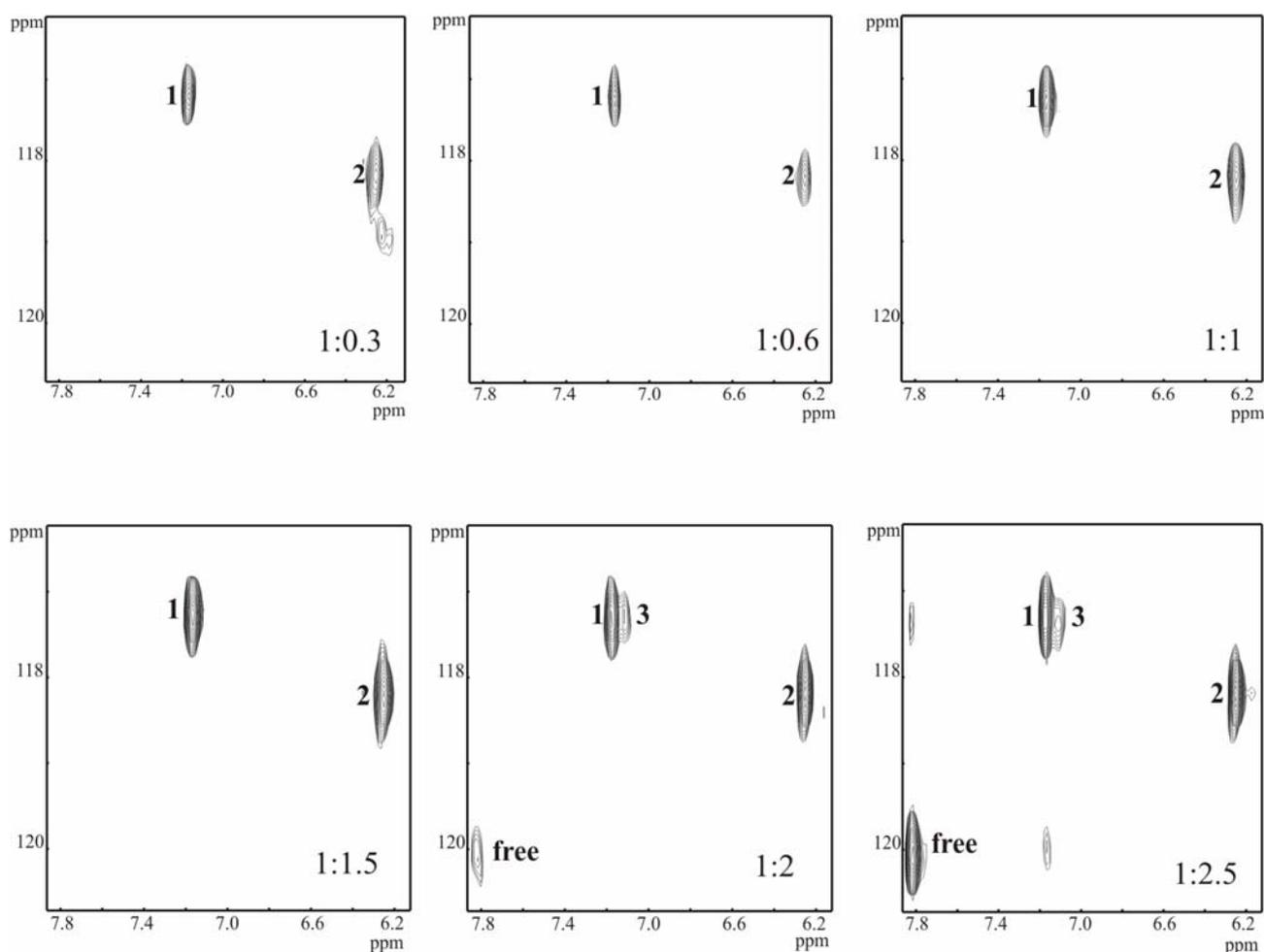
The fingerprint regions of the 2D  $^1\text{H}$ -NMR spectra TOCSY and NOESY of the delipidated cL-BABP are reported in Figure 4.15. The large chemical shift dispersion observed in the 1D  $^1\text{H}$ -NMR spectrum is here confirmed.



**Figure 4.15.** Fingerprint regions of 2D 500 MHz  $^1\text{H}$ -NMR spectra of 1 mM delipidated wt cL-BABP in 20 mM PBS buffer at pH 5.6 (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), 37 °C. (a) TOCSY spectrum, mixing time 70 ms; (b) NOESY spectrum, mixing time 150 ms.

### 4.3.1 Complex of cL-BABP and $^{15}\text{N}$ -Glycochenodeoxycholate (GDC)

The NMR analysis of the interaction of cL-BABP with a bile acid,  $^{15}\text{N}$ -Glycochenodeoxycholate (GDC), showed that the ligand can bind the protein in two different sites, named “site 1” and “site2” (Figure 4.16). Cross peak volumes give an indication of the population of each binding site.



**Figure 4.16.** Contour plot representation of  $^1\text{H}$ - $^{15}\text{N}$ -HSQC spectra of  $^{15}\text{N}$ -Glycochenodeoxycholate in complex with cL-BABP at 298 K and pH 7.1. Different protein:ligand molar ratios are shown. For 1:0.3 molar ratio the spectrum with 32 scans is shown whereas spectra with 8 transients are shown for the other ratios.

In Figure 4.17 A is reported the plot of the molar fraction (cross peak volume of each peak divided by the sum of all cross peak volumes), vs protein:ligand molar ratios. The graph suggests that the two binding sites are equally populated at this temperature. At protein:ligand ratios equal or higher than 2 an upfield shoulder for the amide resonance named “site 1” appears. The new peak, “site 3”, may be an indication of two slightly different populations of GDC bound at site 1. This hypothesis is confirmed by the decrease

of molar ratio observed for site 1 and the concomitant increase of site 3 molar ratio (Figure 4.18 A). In Figure 4.18 B the plot of the sum of all cross peak volumes except the volume of cross peak relative to unbound GCD is shown as a function of protein-ligand ratio. The analysis of this plot suggests that the stoichiometry of the complex is 1:2 cl-BABP:GCD.

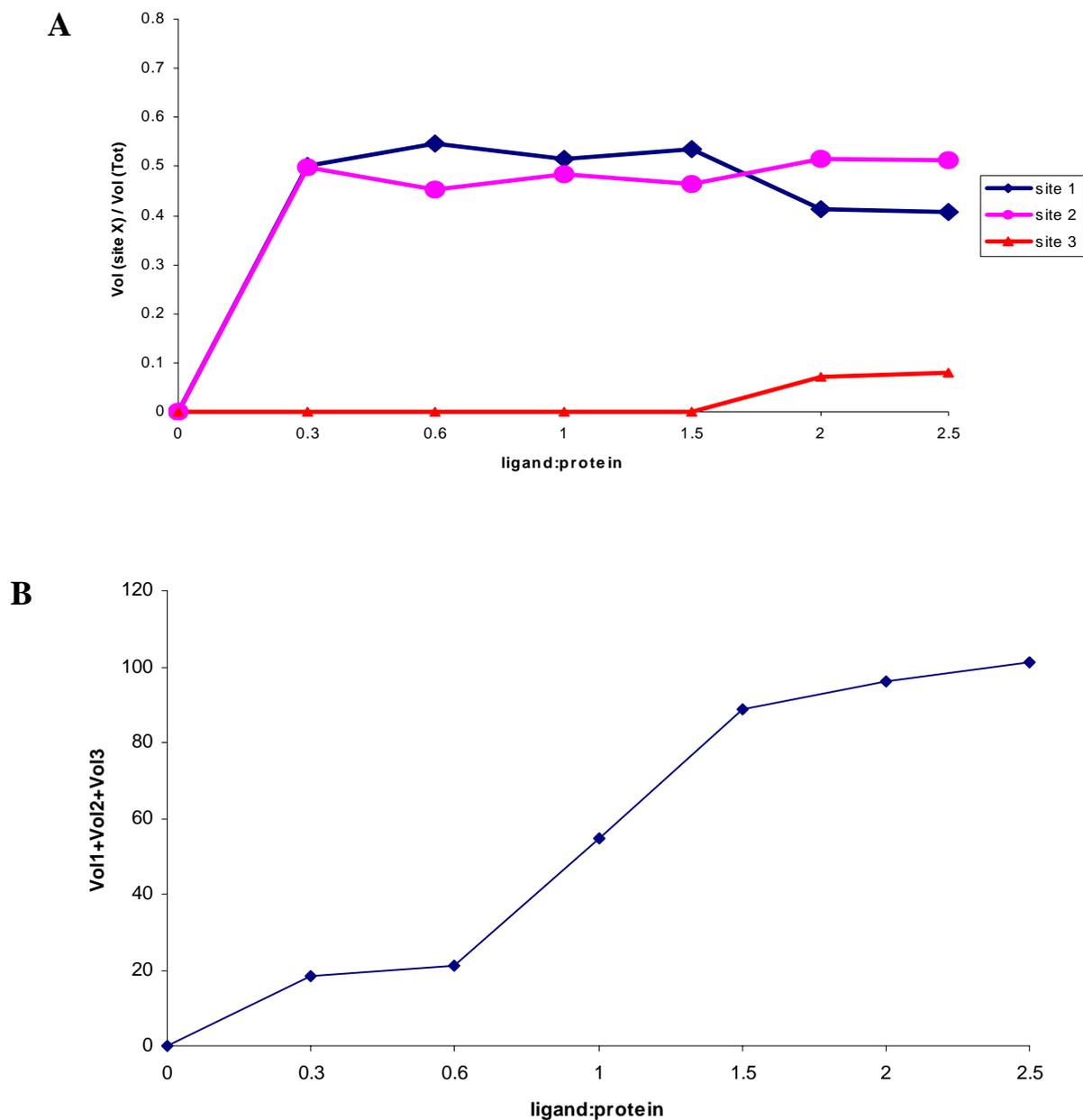


Figure 4.18. Plot of molar ratio (panel A) and stoichiometry (panel B) as deduced from the analysis of NMR cross peak volumes.

## 4.4. Development of the High Level Expression Protocol of cL-BABP in Minimal Medium for the Subsequent $^{15}\text{N}$ and $^{13}\text{C}$ Isotopic Enrichment

As discussed in section 4.8, NMR techniques rely on the availability of proteins highly enriched with stable isotopes ( $^{15}\text{N}$  and  $^{13}\text{C}$ ). To achieve this objective several procedures were tested and compared in order to decide the best expression protocol, taking in account the costs of the labelled chemicals:

- expression from minimal medium (minimal media is composed in the laboratory and it contains nutrients like C and N source, salts, buffering substances, traces elements and vitamins);
- expression from commercial enriched media;
- expression using the two-stage protocol developed by Marley *et al.* 2001: as described in section 4.8 it divides the cell growth and induction in two different steps: the accumulation of cellular biomass in rich unlabeled medium, harvest of the biomass and the massive inoculum in labelled minimal medium which volume is decreased of one quarter and subsequent induction.

The expression and purification procedures were carried out exactly like described in chapter 3; initially unlabelled media were used.

In order to decide the conditions for expression of consistent quantities of recombinant cL-BABP, preliminary tests were done using unlabeled media that were: M9 with supplements, a commercial media (Celtone, Nalgene), and LB and M9 with supplements for the two-stage protocol. For these three type of procedures the following data were collected:

- monitoring cell growth
- expression test in the time course and solubility test

### 4.4.1 Cell growth

#### *M9 with supplements*

It is well known that cell growth in minimal medium is slower than in rich medium. Cultures of *E. coli* containing tested construct for protein expression were grown in 10 mL LB/Kan at 37° C overnight and 500  $\mu\text{L}$  were inoculated in 50 mL of LB and in 50 mL of M9 with supplements. The comparison of the respective growth curves is reported in Figure 4.19.

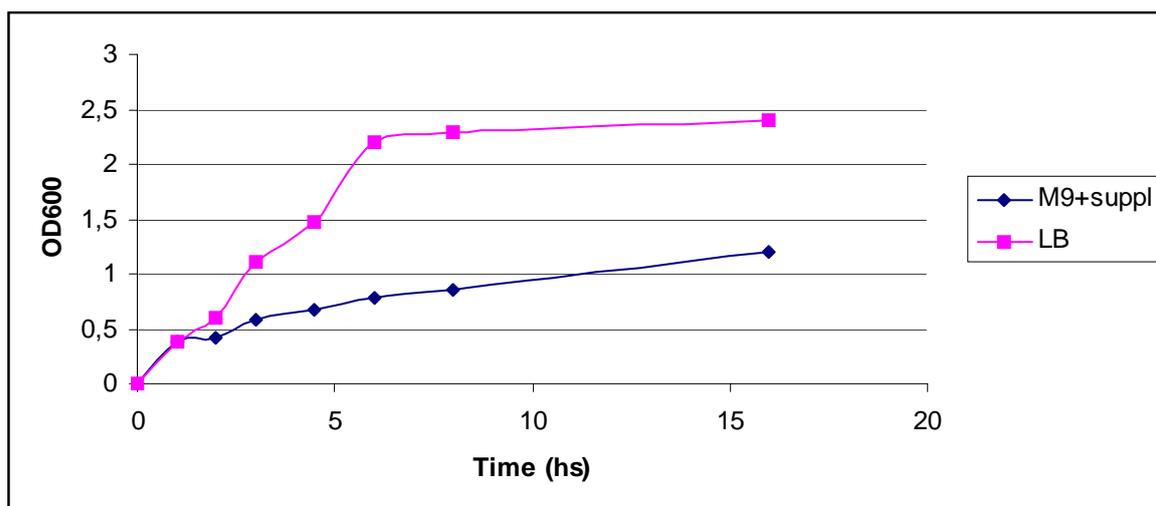


Figure 4.19. Growth of *E. coli* cells in LB and M9+suppl at 37°C: in blue cell growth in LB/Kan 37°C overnight; in pink cell growth in M9+suppl/Kan 37°C overnight .

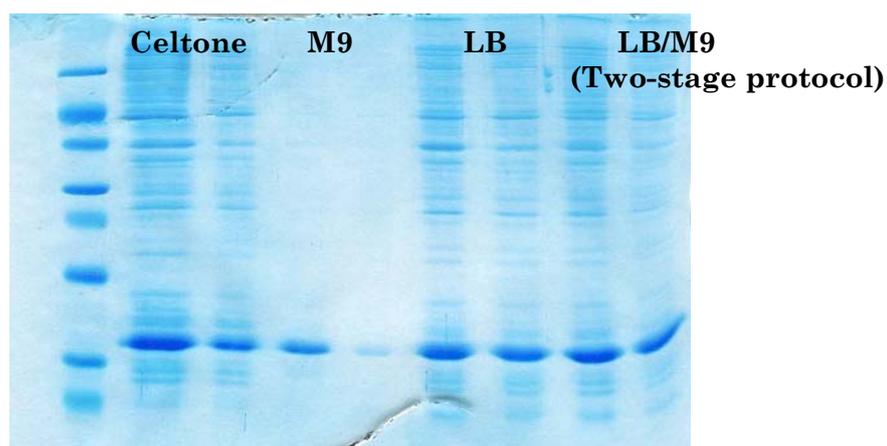
Long time induction can lead to premature cell lysis with lost of protein in the culture media (see section 4.8), so cellular harvesting was done after 16 hours induction; the relative cellular concentration was lower than that reached in LB, and so also the expected protein yield had to be lower.

#### *Ready-to-use medium (Celtone-Nalgene)*

The cellular growth curve using Celtone resembled that reported in Figure 4.3 for LB medium.

#### **4.4.2 Expression in the time course and solubility test**

The conditions described in section 4.2.2 were used for growth and induction tests (cell growth till  $OD_{600} = 0.6-0.75$ , 0.7 mM IPTG, at 20 °C overnight). Cellular sample were compared on SDS PAGE (Figure 4.20). From the comparison of the total and the soluble fractions of each test it was observed that the minimal medium showed the worst yield of soluble protein, while the yields from Celtone and LB and the two-stage protocol were comparable (Celtone slightly lower).



**Figure 4.20.** Lane 1: molecular weight marker; lane 2 and 3 : Total fraction and soluble fraction from Celtone ; lane 4 and 5 : total and soluble fraction from M9, 37 °C 4 h ; lane 6 and 7: total and soluble fraction of LB, 37 °C 4 h; lane 8 and 9: total and soluble fraction from LB two-stage protocol (see text for details).

Four expression and purification tests were carried out starting from 200 mL of unlabeled media, respectively 200 mL of M9 with supplement, 200 mL of Celtone; for the two-stage protocol 800 mL of LB for the first stage and 200 mL of M9 with supplements for the second stage. Expression and purification were carried out as described previously. The final yields from the different procedure were quantified by UV spectroscopy (after gel filtration chromatography), the results are summarized in Table 4.4.

Table 4.4 Comparison between minimal and commercial media in terms of yield			
Celtone	LB	MM	Mixed Protocol
14 mg	18 mg	2 mg	20 mg

From the SDS PAGE in Figure 4.20 the dimensions of the over-expressed bands reflect the obtained yields. From the collected data it was chosen for cL-BABP labelling the mixed protocol.

## 4.5. Production of $^{15}\text{N}$ cL-BABP

### 4.5.1 Large scale expression using the two-stage protocol

The expression was carried out for a total of 1 liter of minimal media (starting from three cell growth of 1.35 liters LB/Kan). In the LB stage it were added 2% glucose for basal expression inhibition (see section 4.8) and 8 g/L of glucose were added to M9. The over-expression of a 14000Da band after induction is visible from the analysis on SDS PAGE (Figure 4.21).

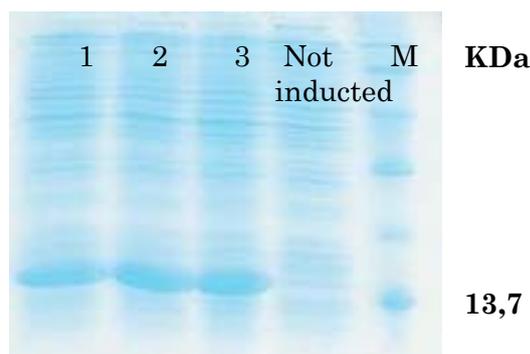


Figure 4.21. SDS PAGE; lane 1,2,3: harvested cell after induction ; lane 4 : cell before induction, lane5: Molecular Weight Marker

The purification procedure was carried out exactly as previously described.

The final yield obtained was **100 mg/L**, purity was checked by the presence of a single band on SDS PAGE and by MALDI mass spectrometry analysis (Figure 4.22) that also confirmed uniform labelling.

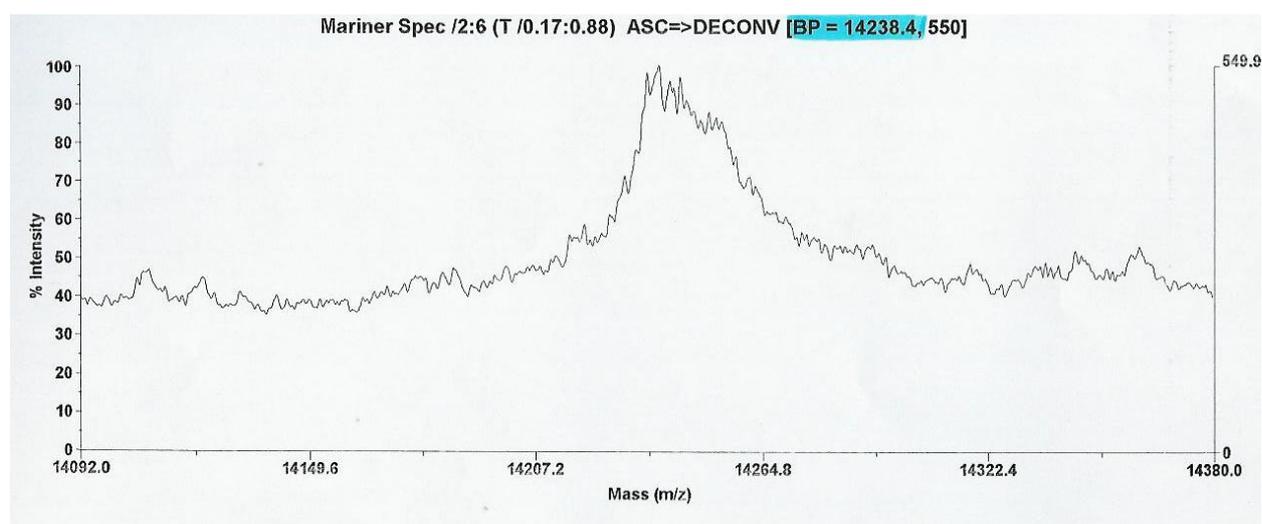


Figure 4.22. MALDI-TOF spectrum of  $^{15}\text{N}$ -cL-BABP

### 4.5.2 Determination of percentage of enrichment by molecular mass spectrometry

The molecular mass obtained from a MALDI analysis for the recombinant  $^{15}\text{N}$ -cL-BABP was of 14238,4 Da. The percentage of  $^{15}\text{N}$  -labelling was calculated as follow:

total number of sites available for  $^{15}\text{N}$  incorporation: 169

cL-BABP wild type MW = 14080,1 Da

The theoretical MW of the enriched protein, accounting the incorporation of 169  $^{15}\text{N}$  nitrogen atoms results: 14249,1 Da

The experimental MW for  $^{15}\text{N}$  enriched cL-BABP resulted of 14238,4 Da

The obtained percentage of enrichment was **93,67 %**

### 4.5.3 NMR experiments

An  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectrum (Figure 4.23) was collected on the  $^{15}\text{N}$  cL-BABP under native conditions in 20 mM PBS, pH7 ad 298 K. The good dispersion of the signals is consistent with a substantially folded protein. The spectrum shows the expected number of cross peaks.

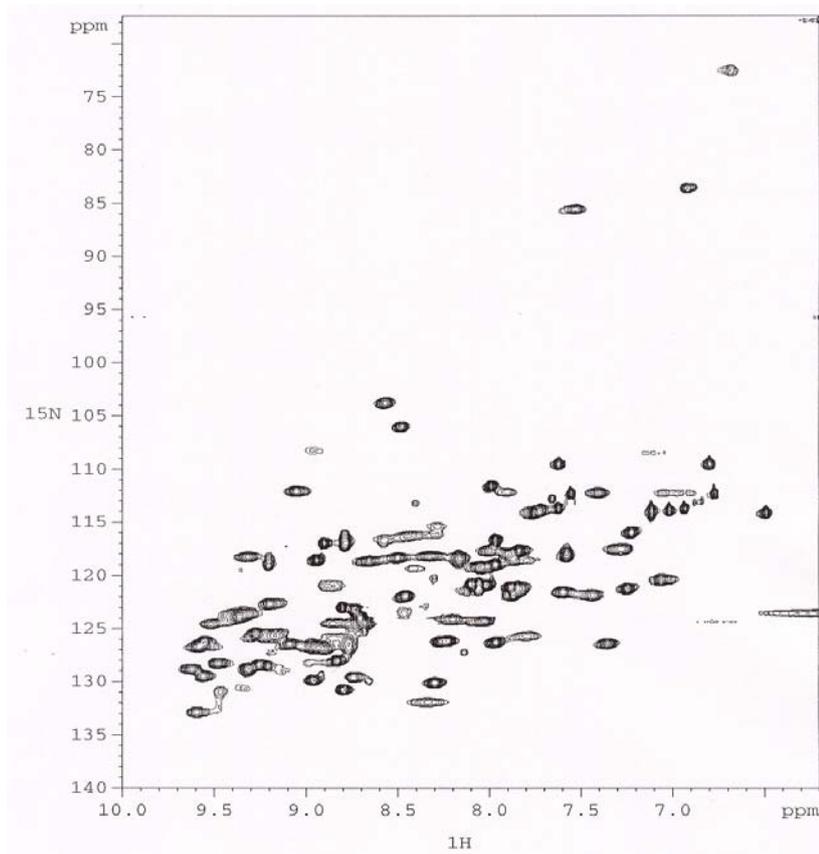


Figure 4.23.  $^1\text{H}$ - $^{15}\text{N}$  HSQC of 1 mM delipidated  $^{15}\text{N}$  cL-BABP in 20 mM PBS, pH 7, 298 K (95%  $\text{H}_2\text{O}$ , 5%  $\text{H}_2\text{O}$ ).

## 4.6 Production of $^{15}\text{N}$ $^{13}\text{C}$ cL-BABP

### 4.6.1 Large scale expression using the two-stage protocol

The same procedure for labelling in  $^{15}\text{N}$  was used. The two-stage protocol was carried out from 1 liter LB and 250 mL M9 with supplements. Initially 2% glucose was not added to LB because it was thought that it could interfere with the final percentage of  $^{13}\text{C}$  incorporation. Given the high cost of labelled D-glucose, 4 g/L instead of 8 g/L were added to minimal medium; the final yield resulted halved (24 mg/250 mL, about 50 mg/L). Unfortunately from MALDI mass spectrometry analysis (Figure 4.24) it was observed that a mixture of three partially labelled proteins was expressed.

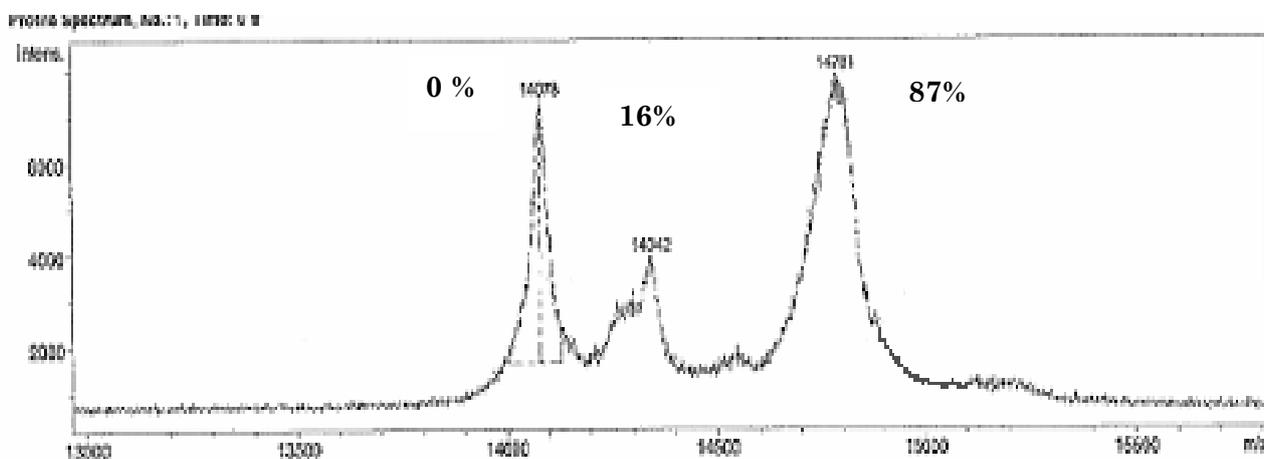


Figure 4.24. MALDI-TOF spectrum of first preparation of  $^{13}\text{C}$   $^{15}\text{N}$ -cL-BABP.

The molecular mass values obtained from a MALDI analysis for the recombinant  $^{15}\text{N}$ - $^{13}\text{C}$ -cL-BABP were of 14078 Da, 14342 Da and 14781 Da. The percentage of double labelling was calculated as follow:

total number of sites available for  $^{13}\text{C}$   $^{15}\text{N}$  incorporation:  $625 + 169 = 794$

cL-BABP wild type MW = 14080,1 Da

The theoretical MW of the enriched protein, accounting the incorporation of 169  $^{15}\text{N}$  nitrogen and 625  $^{13}\text{C}$  carbon atoms results: 14874,1 Da

The experimental MW resulted respectively of

The respective percentage of enrichment obtained for the partially labeled  $^{15}\text{N}$ - $^{13}\text{C}$ -cL-BABP was of **0 %**, **32,8 %**, **88,3 %**.

The two-stage protocol had not worked well when used for the production of the double  $^{13}\text{C}$   $^{15}\text{N}$  enriched cL-BABP, so it was necessary to improve the expression in minimal media.

### 4.6.2 Improvement of cL-BABP expression in minimal media

Several experimental conditions were tested in order to improve the protein expression yield in minimal media. It was first taken in consideration the type of culture used for the inoculum. The cellular growth curves of *E. coli* in 50 mL M9 with supplements were inoculated with: a) a cell culture grown in LB at 37°C overnight and b) a cell culture grown in M9 at 37°C overnight. The results were then compared (Figure 4.25).

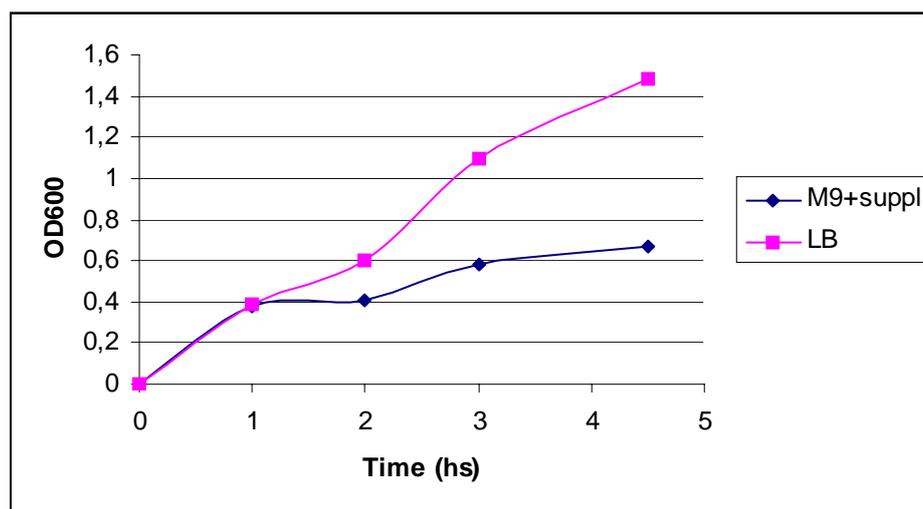


Figure 4.25. Growth of *E. coli* in 50 mL M9 with supplements at 37°C using two different type of pre-inocula: in blue growth curve using, as pre-inoculum, 500  $\mu$ L of LB/Kan 37°C overnight; in pink using, as pre-inoculum, 500  $\mu$ L of M9+suppl/Kan 37°C overnight.

The strategy was to observe if growth of cells in minimal medium resulted improved.

It was observed that the lag-phase of the two inocula has the same course, while the exponential-phase is more rapid for the inoculum derived from cells grown in minimal medium. We could observe a substantial improvement in cell growth rate inoculating M9 with cells “adapted” overnight in the same medium.

It was then considered the agitation speed, that was changed from 180 rpm to 240 rpm in order to increase aeration; then a further selection (testing over-expression bands on SDS PAGE) of freshly transformed bacteria was done, in order to use colonies with the higher plasmid copy number; finally induction was carried out with IPTG at the final concentration of 2 mM.

All these improvements led to the obtainment of a final yield of 40 mg/L of pure protein in M9 with supplements using 4 g/L of glucose.

This procedure was so used for the production of uniformly double  $^{13}\text{C}$  and  $^{15}\text{N}$  enriched cL-BABP with a percentage of isotopic incorporation of 92,3% (MALDI spectrum not shown).

$^{13}\text{C}$   $^{15}\text{N}$  enriched cL-BABP samples were used for *apo* cL-BABP assignment, as reported in the enclosed article (Ragona *et al.*, 2006) for the structure and dynamic determination of cL-BABP (PDB entry 1ZRY).



**NMR DYNAMIC STUDIES SUGGEST THAT ALLOSTERIC ACTIVATION  
REGULATES LIGAND BINDING IN CHICKEN LIVER BILE ACID BINDING  
PROTEIN**

**Laura Ragona<sup>1</sup>, Maddalena Catalano<sup>1,2</sup>, Marianna Luppi<sup>2</sup>, Daniel Cicero<sup>3</sup>, Tommaso  
Eliseo<sup>3</sup>, Jefferson Foote<sup>4</sup>, Federico Fogolari<sup>5</sup>, Lucia Zetta<sup>1</sup>, and Henriette Molinari<sup>2\*</sup>**

<sup>1</sup>Laboratorio NMR, ISMAC, CNR, via Bassini 15, 20133 Milano, Italy; <sup>2</sup>Dipartimento Scientifico e Tecnologico, Università di Verona, Strada Le Grazie 15, 37134 Verona, Italy; <sup>3</sup> Dipartimento di Scienze e Tecnologie Chimiche, Università di Roma Tor Vergata, Via della Ricerca Scientifica, 00133 Roma, Italy; <sup>4</sup> Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; <sup>5</sup>Dipartimento di Scienze e Tecnologie Biomediche, Università di Udine, Piazzale Kolbe, 43100 Udine, Italy.

Running title: NMR dynamic studies of apo and holo cl-BABP

Address correspondence to: Henriette Molinari, Dipartimento Scientifico e Tecnologico, Strada Le Grazie 15, University of Verona, 37134 Verona, Italy. Phone: +390458027097; FAX:+390223699620; E-mail [molinari@sci.univr.it](mailto:molinari@sci.univr.it)

**Apo chicken liver bile acid binding protein has been structurally characterised by NMR. The dynamic behaviour of the protein in its apo and holo forms, complexed with chenodeoxycholate, has been determined *via* <sup>15</sup>N relaxation and steady state heteronuclear <sup>15</sup>N(<sup>1</sup>H) NOE measurements. The dynamic parameters were obtained at two pHs (5.6 and 7.0) for the apo protein and at pH 7.0 for the holo protein, using the model free approach. Relaxation studies, performed at three different magnetic fields, revealed a substantial conformational flexibility, on the microsecond to millisecond time scales, mainly localised in the C-terminal face of the beta-barrel. The observed dynamics are primarily caused by the protonation/deprotonation of a buried histidine residue, H<sub>98</sub>, located on this flexible face. A network of polar buried side-chains, defining a spine going from E to J strand, is likely to provide the long range connectivity needed to communicate motion from H<sub>98</sub> to the EF loop region. NMR data are accompanied by MD simulations, suggesting that H<sub>98</sub> protonation equilibrium is the triggering event for the modulation of a functionally important motion, i.e. the opening/closure at protein open end, while ligand binding stabilises one of the pre-existing conformations (the open form). The results presented here, complemented with an analysis of proteins belonging to the**

**intracellular lipid binding protein family, are consistent with a model of allosteric activation governing the binding mechanism. The functional role of this mechanism is thoroughly discussed within the framework of the mechanism for bile acids enterohepatic circulation.**

Recent studies have shown that bile acids not only serve as the physiological detergents that facilitate absorption, transport and distribution of lipid-soluble vitamins and dietary fats, but also are the signalling molecules that activate nuclear receptors and regulate bile acid and cholesterol metabolism. In addition, bile acids induce the cytochrome P450 3A family of cytochrome P450 enzymes that detoxify bile acids, drugs and xenobiotics in the liver and intestine, induce hepatocyte apoptosis and activate the gene encoding a candidate bile acid transporter protein (1). Given the important role of bile acids the study of their transport at a molecular level is of special medical and pharmacological interest. In this line it is essential to gain insight into the three-dimensional (3D) structures and dynamical behaviour of proteins, in their free and complexed forms, involved in bile acid recycling.

Interestingly bile acids have been suggested to be the putative ligands of a group of intracellular

Lipid Binding Proteins (iLBPs) or Fatty Acid Binding Proteins (FABP), expressed in the liver of non mammalian species, and previously referred to as Liver basic FABP. FABPs have been classified and described on the basis of the organ that they were initially isolated from, but several instances are known in which more than one FABP type has been shown to be produced by a single tissue. We have previously reported on the higher similarity of liver basic FABPs from non mammalian species with ileal lipid binding protein (ILBP) rather than with mammalian liver FABP (2). In agreement with this observation bile acid binding and transport is emerging as the specific function of the liver non mammalian subfamily, hence called liver bile acid binding protein (BABP) (2,3). At variance, the paralogous proteins expressed in the same tissue but in mammals, play a role in fatty acid binding and transport (4). A multiple alignment of all the known sequences of non mammalian liver BABPs with ILBPs is reported in Figure 1.

It has been proposed that internal protein dynamics in iLBPs could be intimately connected with ligand recognition and interaction (2,5-8). We report here a structural and dynamic study on chicken liver BABP (cl-BABP), in its apo and holo form, combining heteronuclear NMR experiments and  $^{15}\text{N}$  NMR relaxation measurements with MD simulations. We investigate the role of the protonation state of a buried histidine on protein dynamics. We discuss here the observed change in dynamics upon ligand binding in terms of an allosteric activation mechanism, i.e a shift between inactive and active conformations (9). The proposed mechanism for ligand binding in cl-BABP is further analysed in light of data reported for other members of iLBP family and discussed as functional to bile acid enterohepatic circulation.

## Materials and Methods

### *Protein expression and purification.*

Recombinant cl-BABP was expressed as soluble protein in *E. coli* BL21 (DE3) bearing the recombinant plasmid pET24d. Transformed cells were grown on plates containing 50  $\mu\text{g}/\text{ml}$

kanamycin. One liter LB was inoculated with an overnight culture, and incubated at 310 K until cells reached an  $\text{OD}_{600}$  of 0.8. Protein expression was induced by addition of 0.7 mM isopropylthiogalactopyranoside (IPTG) and incubation continued overnight at 293 K. The cells were harvested and re-suspended in lysis-buffer (50 mM Tris, 10% sucrose, 1 mM EDTA, 10 mM  $\beta$ -mercaptoethanol, pH 8.0). After lysis, the supernatant, containing cl-BABP, was loaded on a DEAE cellulose (Whatman) anion exchange column equilibrated with 50 mM Tris-acetate at pH 7.8. The same buffer was used for protein elution. Fractions containing cl-BABP were concentrated and resolved on a Sephacryl S-100 HR (Amersham Biosciences) column equilibrated with 50 mM Tris-HCl, 0.2 M NaCl at pH 7.2. cl-BABP was delipidated as described (10). The protein purity was checked by the presence of a single band on SDS-PAGE and by mass spectrometry. The protein yields were 90 mg/L of bacterial culture.  $^{15}\text{N}$  isotope labelling was achieved using M9 minimal media containing 1gr/L  $^{15}\text{NH}_4\text{Cl}$ , following protocols reported in the literature (11). The extent of  $^{15}\text{N}$  labelling was verified by MALDI mass analysis and the isotope incorporation was found to be more than 92%.  $^{15}\text{N}$  cl-BABP was obtained in a yield of 50 mg/L of minimal media.  $^{13}\text{C}$ ,  $^{15}\text{N}$  double labelling was obtained with the same procedure using M9 minimal media containing 1gr/L  $^{15}\text{NH}_4\text{Cl}$  and 4gr/L  $^{13}\text{C}$  enriched sucrose. The extent of labelling, verified by MALDI mass analysis was > 90 % and yields of 25 mg/L of minimal media were obtained.

Commercial chenodeoxycholic acid (Sigma-Aldrich) was employed for the preparation of holo cl-BABP with a ligand to protein ratio 5:1, as previously described (8).

*NMR experiments.* NMR data were recorded on Bruker Avance 500, 600 and 700 MHz spectrometers equipped with pulse field gradient triple-resonance probes. 0.5 mM protein samples in phosphate buffer at pH 7.0 and 5.6 and 298 K were employed for structure determination and relaxation measurements.

2D homonuclear TOCSY (mixing 70 ms) and NOESY (mixing 150 ms) were performed at 500 and 700 MHz on cl-BABP sample at pH 7.0, 298 K. Water suppression was achieved using the excitation sculpting sequence (12) for TOCSY

and WATERGATE (13) for NOESY. 2D homonuclear TOCSY and NOESY were also performed at 298 K and pH 5.6, i.e. in the same conditions as those previously reported for the purified protein (2).

On the  $^{15}\text{N}$  enriched apo (pH 5.6 and pH 7) and holo (pH 7.0) cl-BABP sample  $^{15}\text{N}$  HSQC-TOCSY (14) (mixing 85 ms),  $^1\text{H}$ - $^{15}\text{N}$  HSQC-NOESY (14) (mixing 150 ms), HNHA (15) were recorded.

The following triple resonance experiments, using standard parameter sets (16), were recorded on the doubly labelled [ $^{15}\text{N}$ ,  $^{13}\text{C}$ ] apo and holo cl-BABP in  $\text{H}_2\text{O}$  at 700 MHz and pH 7.0: HNCA, HN(CO)CA, HNCO, CBCANH and CBCA(CO)NH. For the sample dissolved in  $\text{D}_2\text{O}$  HACACO, (H)CCH-COSY, (H)CCH-TOCSY, H(C)CH-COSY and H(C)CH-TOCSY experiments were performed (17). Two NOESY-type 3D experiments (mixing 100 ms), one optimised for aliphatic and one for aromatic residues, were acquired.

A series of 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiments were performed for the apo protein at different pHs (in the range 4.2-7.4) to allow for measurement of the midpoint of the chemical shift pH driven titration. Spectra were assigned on the basis of the assignments obtained at pHs 7.0 and 5.6.

The  $^{15}\text{N}$  chemical shift titration data were fitted to eq. [1] in order to evaluate pKa values (18):

$$\delta_{obs} = \delta_d + \frac{\delta_p - \delta_d}{1 + 10^{[pH - pK_a]}} \quad [1]$$

where  $\delta_p$  and  $\delta_d$  are the chemical shifts of the protonated and the deprotonated state, respectively. Calculation of  $^1\text{H}$  and  $^{15}\text{N}$  secondary shifts was performed according to  $\delta = [(\Delta\delta_{\text{HN}})^2 + \Delta\delta_{\text{N}}^2/25]^2$  (19).

$^{15}\text{N}$  relaxation experiments (20), run as water flip-back version, were acquired at 600 and 700 MHz both at pH 7.0 and 5.6. Eleven delays (2.5, 20, 60, 100, 150, 200, 300, 400, 600, 800, 1000 ms) were used for  $T_1$  measurements and nine delays (16.96, 33.92, 50.80, 67.84, 101.76, 135.68, 169.6, 220.48, 237.44 ms) for  $T_2$  measurements. The delay in the CPMG pulse train was set to 0.45 ms.  $^1\text{H}$ - $^{15}\text{N}$  NOE experiments were acquired with an overall recycling delay of 6 s (20). In order to analyse the exchange contribution to relaxation at pH 7,  $T_1$ ,  $T_2$  and  $^1\text{H}$ - $^{15}\text{N}$  NOEs were also measured

at 500 MHz, in the same conditions as described for higher field measurements

Relaxation measurements were identically performed at pH 7.0 for holo cl-BABP complexed with chenodeoxycholate.

Data were processed with XWINNMR and NMRPipe (21) and analysed with NMRView 5.0.3 software package (22).

*Structure calculation of apo cl-BABP.* Volume integration was performed on the 3D  $^{15}\text{N}$ - $^{13}\text{C}$  NOESY and  $^1\text{H}$ - $^{15}\text{N}$  HSQC-NOESY spectra using NMRView (22). Peak volumes calibration was performed using the median method, a routine of NMRView program, and the obtained list of distances was used as input for DYANA (23) calculations.  $\phi$  angle restraints were derived from  $J_{\text{HN},\text{H}\alpha}$  coupling constants estimated from 3D HNHA experiments (15).  $\phi$  angle restraints of  $139^\circ \pm 30^\circ$  for  $J_{\text{HN},\text{H}\alpha}$  coupling constants greater than 8.0 Hz and  $60^\circ \pm 30^\circ$  for  $J_{\text{HN},\text{H}\alpha}$  coupling constants smaller than 5.0 Hz were used as restraints.

Amide proton exchange rates were estimated from a series of  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra performed at different times after dissolving the protein in  $\text{D}_2\text{O}$  (data not shown). The partners for all hydrogen bonds were assigned on the basis of preliminary structures obtained by imposing only NOE restraints. Each hydrogen bond was introduced as a restraint on O-N distance of 3.00 Å and HN-O distance of 2.00 Å. The decision was taken to introduce in the calculation only totally unambiguous restraints, i.e. those correlations which were not affected by overlap in any spectra.

The restraints were re-examined to check for consistent violations. One hundred calculations were run employing DYANA (23) and the 20 conformers with the lowest residual target function were analysed. The 20 final DYANA structures were further refined using the AMBER force field, as implemented in the program DISCOVER (Molecular Simulations, San Diego, CA). A dielectric constant of  $4 \times r$  was used and a scaling factor of 10 was used for out-of-plane interactions. Each structure was minimised performing 100 steps of steepest descent and 300 steps of conjugate gradient. The

10 structures with the lowest potential energy were selected for further analysis.

The structures were deposited in the PDB with id 1zry.

*Relaxation data analysis*. Relaxation times were calculated *via* least squares fitting of peak intensities, using the rate analysis routine of NMRView program (22). The heteronuclear NOE effects were calculated from the ratio of cross-peak intensities in spectra collected with and without amide proton saturation. The principal components of cl-BABP inertia tensor were calculated using Pdbinertia (A. G. Palmer III, Columbia University). The principal moments of inertia of apo cl-BABP at pH 7 were calculated on the basis of our NMR structure (PDB id 1zry), while at pH 5.6 the representative coordinates from MD simulations were used (see later). For holo protein the X-ray structure (PDB id 1tw4) was employed. Isotropic and anisotropic model were tested for apo and holo cl-BABP.

An initial estimate of the overall correlation time and of principal components and orientation of the diffusion tensor can be reliably determined from the angular dependence of the relaxation rates of a subset of NH vectors assumed to have a negligible component of internal motion and/or exchange contribution to  $^{15}\text{N}$  relaxation. The selection of the subset of residues was made following the procedures described in the literature (24): residues with NOE < 0.65 were removed from the data set and residues with low  $T_2$  values ( $T_2 \leq \langle T_2 \rangle - \sigma_{T_2}$ ) were removed from data set unless their corresponding  $T_1$  values were high ( $T_1 \geq \langle T_1 \rangle + \sigma_{T_1}$ ), indicating that they could be affected by anisotropic tumbling.

For the axially symmetric model  $D_{\parallel}$ ,  $D_{\perp}$ ,  $\theta$  and  $\phi$  initial estimates were evaluated using the Quadric Diffusion program (A. G. Palmer III, Columbia University) that uses the quadratic representation approach (25). Relaxation of amide  $^{15}\text{N}$  nuclear spins were analysed using the standard equations assuming, for a diamagnetic protein, dipolar coupling with directly attached

protons and a contribution from the  $^{15}\text{N}$  chemical shift anisotropy (26) evaluated as  $\Delta\sigma = -170$  ppm. The experimental data were fitted to the Lipari-Szabo model (27) using the program MODELFREE (version 4.0). The extended Lipari-Szabo formalism proposes five spectral density functions which depend upon  $S^2$  (the generalised motional order parameter),  $\tau_m$  (the overall correlation time of rotational diffusion),  $\tau_e$  (the effective correlation time) and  $R_{ex}$  (the rate of conformational exchange). The five models of motion were iteratively tested in order of increasing complexity and the model which best fitted the data was selected as described elsewhere (28). After model selection the overall rotational diffusion model parameters and the internal motional parameters for each spin were optimised simultaneously.

At pH 7 the exchange contributions were extracted from the relaxation data at three frequencies using the approach described in (29). The parameter  $R_2 - (R_1/2)$  can be expressed as :

$$R_2 - \frac{R_1}{2} \approx \left( \frac{c_1^2}{3} J(0) + A \right) B_0^2 + \frac{d_1^2}{3} J(0) \quad [2]$$

in the assumption of an exponentially decaying autocorrelation function, where  $c_1 = (8\pi/15)^{1/2} \gamma_N \Delta\sigma$ ,  $A = R_{ex}/B_0^2$  and  $d_1 = (6\pi/5)^{1/2} d$ . Using a plot of  $R_2 - (R_1/2)$  versus  $B_0^2$  the spectral density function  $J(0)$  can be calculated from the intercept,  $I_0 = (d_1^2/3)J(0)$  and the exchange constant  $A$  can be deduced from the slope,  $m = [J(0)c_1^2/3] + A$ , of the line. In principle any spin for which the data has a slope  $m > c_1^2 I_0 / d_1^2$  will have an exchange contribution. However taking in consideration experimental and fitting errors a threshold of  $1.3 \cdot \langle m \rangle$  was used to determine residues subject to exchange (29), where  $\langle m \rangle$  is the average slope.

*Theoretical pKa calculations*. All pKa calculations have been performed as previously described (30,31). The linear Poisson-Boltzmann equation was solved for different charge states and the electrostatic free energy was used to estimate pKa shifts. The mid-point of the titration for each site is taken as its pKa. All Poisson-Boltzmann calculations have been performed using the program UHBD (32).

*Molecular dynamics simulations.* Molecular dynamics simulations were performed using the program GROMACS (version 3.2.1) employing the Gromacs forcefield (ffgm2) (33). The protocol used was essentially as previously described for beta-lactoglobulin (34). The structure of the bile acid binding protein was taken from PDB (pdb id. 1zry, model 1). Protons were added using the program pdb2gmx, in the GROMACS suite of programs, for optimisation of the hydrogen bond network. The protein was first minimized by 200 steepest descent minimization steps, followed by 200 conjugate gradients steps. Due to lack of solvent in this step the dielectric constant used was 10. The Poisson Boltzmann equation was used to compute the electrostatic potential around the molecule. The lowest potential region at 0.7 nm from any protein atom was chosen for placing a counterion. The procedure was repeated on the protein and ion(s) until the net charge of the system was 0. The minimized protein and ions were then solvated in a box of SPC water with boundaries at least 1.6 nm away from any protein or ion atom. After addition of solvent molecules and ions to the system, long-range electrostatic interactions were treated by particle mesh Ewald method with the following parameters: distance for non-bond interaction cutoff 12 Å, spacing for the fast Fourier transform grid 1.2 Å.

The solutes were fixed and water was energy minimized by 100 steepest descent minimization steps. A short molecular dynamics run (50 ps) keeping the solutes fixed was performed to let the water soak the system. During this run the timestep was set to 1 fs. Finally the unrestrained system was energy minimized by 200 steepest descent steps and equilibrated in the NTP ensemble for 100 ps.

In all molecular dynamics simulations the system was in equilibrium with a temperature bath at 300 K, with relaxation

time constant of 0.1 ps. The system compressibility was that of water,  $4.5 \times 10^{-5}$  bar<sup>-1</sup>. The relaxation time for pressure equilibration was 0.5 ps. The initial velocities were set to 0. Two 3.6 ns MD simulations were performed for the low pH form (with the two histidines protonated) and the neutral pH form (with both histidines deprotonated) of cl-BABP. In both cases 100 ps equilibration time were employed.

The RMSD from starting structure could be fitted by an exponential with time constant of 150 ps for both simulated forms, although for the protonated form a much slower, very small but detectable, increase in RMSD is observed throughout the run. The backbone RMSD from native, including protein ends and loops, is fluctuating around 2.2 Å after few hundred ps. In order to make sure that the system was equilibrated (at least in this time range) we repeated all analyses of local fluctuations for the same trajectories truncated at 1.8 ns. No significant difference was found.

Snapshots were taken at 100 ps intervals along the simulations and these 37 snapshots were used for structural analysis. The snapshot exhibiting the smaller average RMSD with respect to all other snapshots has been taken as the most representative structure in the ensemble.

All structural analysis have been performed using the program Molmol (36) and the analysis programs of GROMACS. Pairwise superposition has been performed using the program ProFit (Martin, A.C.R., <http://www.bioinf.org.uk/software/profit/>).

## RESULTS

*Apo cl-BABP NMR assignment and structure calculation.* Recombinant cl-BABP has been characterized by <sup>1</sup>H, <sup>13</sup>C, and <sup>15</sup>N NMR. The choice of working at pH 7.0 was dictated by the need to perform structural and dynamical comparisons with the protein in its holo form at neutral pH. Backbone assignment,

performed by a combination of classical 3D NMR experiments, was not straightforward especially for the C-terminal region of the protein corresponding to strands F, G, H and I. In this region, breaks in the process of assignment were caused by missing correlations due either to fast exchange of amide protons with solvent and/or to conformational exchange (see later). It was therefore necessary to combine the standard 3D backbone assignment strategy with the sequential assignment strategy. 3D  $^1\text{H}$ - $^{15}\text{N}$  TOCSY/NOESY, performed at pH 5.6, guided the assignment of those amide resonances in fast exchange with solvent at pH 7.0. In this way the assignment was possible for all but six residues, namely M<sub>73</sub>, V<sub>90</sub>, S<sub>93</sub>, K<sub>95</sub>, E<sub>99</sub>, Q<sub>100</sub>, located in a region of the protein mostly affected by conformational exchange, as revealed by  $^{15}\text{N}$  relaxation analysis (see later). The  $^1\text{H}$ ,  $^{13}\text{C}$  and  $^{15}\text{N}$ , assignments of apo cl-BABP have been deposited in the BioMagResBank (entry code 6642).

3D  $^1\text{H}$ - $^{15}\text{N}$  TOCSY/NOESY spectra obtained at pH 5.6 revealed the presence of double peaks for several residues: unambiguous assignment was possible for S<sub>3</sub> (A strand), G<sub>44</sub> (BC loop), F<sub>47</sub> (C strand), D<sub>74</sub> (EF loop), A<sub>85</sub> (FG loop), L<sub>89</sub> (G strand) and G<sub>104</sub> (HI loop). The small difference in chemical shift of major and minor peaks of approximately 20-120 Hz indicated a time scale of exchange of the order of 0.001-0.01s. These double peaks provide an indication of slow exchange processes affecting the protein backbone.

Only totally unambiguous restraints, i.e. those correlations which were not affected by overlap in any spectra, were used for structural calculation. In this way a set of 1000 non redundant NOEs was supplemented by (i) 26 distance restraints for 13 backbone hydrogen bonds defined on the basis of deuterium hydrogen exchange studies (data not shown) and (ii) 48  $\phi$  angle

constraints derived from  $J_{\text{HN-H}\alpha}$  coupling constants. It should be stressed that this protein is highly flexible, as revealed both by H/D exchange and relaxation measurements, and several residues did not exhibit long-range NOE correlations (see later). The superposition of the ten best NMR structures, as obtained after DYANA molecular dynamics simulations followed by energy minimisation, reported in Figure 2, affords an RMSD<sub>bb</sub> (3-125) value of  $2.02 \pm 0.26$  Å. The structural quality of the minimised structures was examined with the PROCHECK-NMR (35). Analysis of the backbone dihedral angles showed that 95 % of all non-glycine non-proline residues in apo cl-BABP fall within the additional allowed regions of conformational space. Considering that this analysis includes some poorly defined regions located in the C-terminal end, this result can be considered reasonable. The NMR structures have been deposited in the Protein Data Bank as 1zry. A survey of the quality of structure determination is reported in Table 1.

The distribution of distance restraints per residue accounts for the observed distribution of average global displacement (Figure 1 of Supplementary materials). The high backbone dispersion of certain segments of cl-BABP essentially corresponds to residues that showed fewer distance restraints because of either conformational dispersion/mobility or lack of assignment.

Even if a few more amides could be detected at pH 5.6, the number of collected restraints did not exceed the 5% of the total restraints obtained at pH 7.0, thus reinforcing the picture of a highly flexible molecule.

*Holo cl-BAB NMR assignment.*  $^1\text{H}$  and  $^{15}\text{N}$  assignments of cl-BABP complexed with chenodeoxycholate were obtained following the same strategy described for the apo protein. The assignments of holo protein are

reported in Table 1 of Supplementary Materials. The comparison of apo and holo  $^1\text{H}$  and  $^{15}\text{N}$  chemical shifts indicated that regions mostly affected by binding are located in the C terminal FGHIJ strands (Figure 3).

*$^{15}\text{N}$  relaxation data and Model Free analysis for apo cl-BABP.* The  $R_2/R_1$  ratios and heteronuclear NOEs at 700 and 600 MHz for apo protein at pH 7 are reported in Figure 4a-b. The same data obtained at pH 5.6 are reported in Figure 2 of Supplementary Materials. At both pHs significantly high  $R_2/R_1$  ratios, indicative of conformational exchange processes, were found for residues located in the C-terminal end of the protein. Heteronuclear NOE values lower than 0.65, indicative of protein regions with fast internal mobility, were detected mostly for helix II, the loop connecting helix II to strand B, CD and FG loops.

The principal moments of inertia of apo cl-BABP were in the ratio 1.0: 0.94: 0.59 (pH 7), and 1.0: 0.85: 0.66 (pH 5.6) suggesting that the shape of the molecule does not deviate appreciably from the sphere. However  $D_{\parallel}/D_{\perp}$  values obtained from Quadric Diffusion program suggested a slightly different degree of anisotropy for apo (1.2 at both pHs) and holo (1.4) proteins (see later). The relaxation data were therefore analysed both with the isotropic and axially symmetric model. The results were substantially unchanged for the two models: in the text, the data obtained with the axially symmetric diffusion model are presented, to take into account even minor effects due to anisotropy.

At neutral pH correlation times ( $\tau_m$ ) of  $6.9\pm 0.4$  ns,  $7.2\pm 0.3$  ns and  $7.0\pm 0.5$  ns were estimated (20) at 700, 600 and 500 MHz, respectively. The data sets at the three magnetic fields were simultaneously used to perform Lipari-Szabo model-free analysis for 86 residues. The final optimised values

were  $\tau_m = 7.1$  ns,  $D_{\parallel}/D_{\perp}=1.2$  and the values for internal motion parameters of the single spins are reported in Table 2 of Supplementary Materials. A calculated  $S^2$  average value of  $0.90\pm 0.04$  was observed.

At pH 5.6 values of  $7.7\pm 0.8$  ns and  $7.1\pm 0.7$  ns were obtained at 700 and 600 MHz, respectively. The final optimised values were  $\tau_m=7.2$  and  $D_{\parallel}/D_{\perp}=1.2$ , while the values for internal motion parameters of the 91 analysed spins are reported in Table 3 of Supplementary Materials. A calculated  $S^2$  average value of  $0.91\pm 0.04$  is observed. Residues  $D_{33}$ ,  $T_{57}$  and  $D_{74}$  could not be fitted to any model.

$S^2$ ,  $\tau_e$  and  $R_{ex}$  contributions, obtained from model free analysis of the available data at two fields (600 and 700 MHz) and at two pHs are reported in Figure 5 and residues affected by motions are mapped in colour onto the protein structure (Figure 6).

*Validation of  $R_{ex}$  contributions.* In order to evaluate possible artefacts on  $R_{ex}$  estimate, the  $R_{ex}$  figures obtained from the model-free approach, using model 3 of the spectral density function, were compared with data obtained from two strategies. In the first approach,  $R_1$  and NOE data of residues showing large  $R_2$  values, were fitted to Lipari-Szabo model 1.  $R_{ex}$  contribution were derived as  $R_{ex} = R_2(\text{experimental}) - R_2(\text{fitted})$  and the obtained data are reported in Table 4 of Supplementary Materials. In the second approach additional relaxation experiments were acquired at 500 MHz and  $R_2 - (R_1/2)$  was plotted as a function of the static magnetic field (29) to determine  $dR_{ex}/dB_0$ . The advantage of this approach is that no model-based assumption is made for the spectral density function. The described analysis was possible for a total of 64 residues and exchange contributions were detected for twelve residues located in the C-terminal half of the protein, namely in DE

and EF loops and FGHIJ strands (Figure 3 and Table 5 of Supplementary Materials). A summary of conformational exchange contributions obtained for cl-BABP at pH 7 with all the discussed approaches is presented in Figure 7.

*<sup>15</sup>N relaxation data and Model free analysis of holo-clBABP.* The  $R_2/R_1$  ratios and heteronuclear NOEs measured at 700 and 600 MHz for the holo protein at pH 7.0 are reported in Figure 4 of Supplementary Materials. Interestingly only residues R<sub>124</sub> and V<sub>125</sub>, at the C-terminal end of the protein, exhibited a high  $R_2/R_1$  ratio, while no residue exhibited NOE values lower than 0.65.

The principal components of holo cl-BABP inertia tensor were 1.0:0.91:0.68. Correlation times of  $7.3\pm 0.4$  ns and  $7.1\pm 0.3$  were derived from  $R_2/R_1$  ratio at 700 and 600 MHz, respectively. The final optimised values of  $\tau_m=7.3$  ns and  $D_{\parallel}/D_{\perp}=1.4$  were obtained. Internal motion parameters of 90 residues were determined and reported in Table 6 of Supplementary Materials. A  $S^2$  average value of  $0.90\pm 0.06$  was obtained. Residues D<sub>33</sub> and K<sub>95</sub> did not fit to any model.

Residues affected by  $\tau_e$  and  $R_{ex}$  contributions are mapped in colour onto the protein structure (Figure 6).

*Histidine protonation equilibrium.* A series of <sup>15</sup>N-<sup>1</sup>H HSQC spectra recorded in the pH range 4.2-7.4 allowed the determination of the midpoint of the chemical shift pH driven titration for some residues highly influenced by pH (Table 7 of Supplementary Materials). The average titration midpoint of L<sub>89</sub>, F<sub>96</sub>, S<sub>97</sub>, I<sub>111</sub> (close to H<sub>98</sub>) is  $5.1\pm 0.1$  and that of G<sub>65</sub>, I<sub>84</sub> and A<sub>85</sub> (close to H<sub>83</sub>) is  $6.2\pm 0.1$ . It was not possible to obtain data relative to H<sub>83</sub> and H<sub>98</sub> themselves, due to broadening and overlap of their resonances upon lowering pH below 5.5.

Titration curves relative to the mentioned residues are reported in Figure 5 of Supplementary Materials. Theoretical pKa calculations suggested that only H<sub>98</sub> exhibited a shifted mean pKa of 4.7, while a mean pK<sub>a</sub> value of 5.7 was calculated for H<sub>83</sub>.

*Molecular dynamics (MD) simulations.* Molecular dynamics simulations were performed for the low pH and neutral pH forms of cl-BABP in order to investigate the possible role of the equilibrium between protonated and deprotonated form of the two histidines (H<sub>83</sub> and H<sub>98</sub>) in affecting the observed dynamics, as reported for other proteins (38) (see later). The limit of 3.6 ns was chosen in order to sample protein movements taking place in times of the order of one ns.

The most representative structures derived from MD simulations performed at acidic and neutral pHs have been superimposed globally. The largest differences involve residues 72-76 and 114-117 (Figure 8). These two stretches of the protein partially hinder access of ligands to the cavity of the protein (Figure 9a). The creation of a net charge inside a protein, as is the case for protonation of the buried H<sub>98</sub>, is not favourable and it is usually accompanied by solvent exposition of the charged group. Here the charged H<sub>98</sub> remains buried, however it is involved in a salt bridge with E<sub>109</sub> which in turn loosens to some extent its salt bridge with R<sub>120</sub>. This is consistent with the evidence that buried salt bridges mostly occur within salt bridges networks which favour charge dispersal (36).

One striking difference between protonated and deprotonated structures is a hydrogen bond between hydroxyl of T<sub>72</sub> and carboxyl of D<sub>74</sub>, which is conserved in all snapshots in the deprotonated simulation but it is never found in the protonated simulation. Residue D<sub>74</sub> is instead loosely interacting through a

salt bridge interaction with K<sub>95</sub> in most of the protonated simulations (Figure 9b). In addition to the major conformational change observed for D<sub>74</sub>, a further change at residues E<sub>94</sub> and K<sub>95</sub> is observed after 1.3 ns of simulation. This transition does not alter the overall direction of the main chain but enables different interactions for the side-chains of K<sub>95</sub>. Moreover, upon protonation, a rearrangement of H<sub>98</sub> H-bonds with E<sub>109</sub> and R<sub>120</sub> takes place, concomitant with the movement of D<sub>74</sub> and K<sub>95</sub> enabling the formation of a loose salt bridge.

Root mean square fluctuation analysis of backbone atoms, after superposition on the starting (reference) structure, has been performed using the program `g_rmsf` in GROMACS (<http://www.gromacs.org>). For both simulations the first two N-terminal residues and loops FG, HI and IJ in the C-terminal part of the molecule are not conformationally well defined. The largest differences in conformational flexibility between the two simulations are observed for the segment 72-80 (entailing loop EF) which shows very large fluctuations only in the protonated simulation. These results do not depend on the choice of reference structure. Indeed, almost identical results are obtained by the analysis of the average contribution to global RMSD in pairwise superposition of all snapshots on each other, performed using the program MOLMOL (37).

Average distances and computed J-couplings were compared with the available experimental data (which were not used in MD simulations). For both simulations less than 10% of the J coupling constants were found to differ more than 2 Hz from the corresponding experimental restraints. Similarly, only 5% of interatomic distances showed violations of upper bounds derived from NOE larger than 2 Å. The average upper bound violation is rather limited (0.38 Å and 0.28 Å for the deprotonated and

protonated simulation, respectively) and it is mostly contributed by very large violations involving atoms in most mobile regions and/or involving longer distance bounds. It is worth noting that these violations are greatly reduced when using third power averaging.

## DISCUSSION

The 3D structure, obtained for the apo protein on the basis of NMR data (Figure 2), is typical of all the proteins of iLBP family, and consists of ten antiparallel  $\beta$ -strands (A–J) organised in two nearly orthogonal  $\beta$ -sheets that form a  $\beta$ -clam type structure with a gap between D and E strands. Helices I and II, inserted between A and B strands, close the protein cavity where bile acids are bound.

Protein dynamics was investigated at two pH values characterising the two functional state of cl-BABP: pH 7, where the binding can take place (active conformation), and pH 5.6, where a substantial decrease of the bound ligand is observed (inactive conformation). The dynamic analysis afforded average order parameters ( $S^2$ ), viewed over the entire protein sequence, substantially unchanged on going from pH 5.6 ( $0.91 \pm 0.04$ ) to 7.0 ( $0.90 \pm 0.04$ ). At both pHs the same protein segments, namely helix II, all loops and E strand, experience fast internal perturbations (ps-ns timescale) while  $R_{ex}$  contributions are observed only for residues located in the C-terminal half of the protein (Figure 6). It is important to stress here that the conformational exchange contributions were obtained for the same protein regions applying both Lipari-Szabo and model-independent approaches (Figure 7). Upon changing pH a few differences in the dynamical behaviour of cl-BABP were observed, mostly located in the C-terminal half of the protein, where two histidines, i.e. the buried H<sub>98</sub> (H strand) and the more exposed H<sub>83</sub> (F strand), are located. The

observed protein dynamics might be therefore coupled to the exchange between their protonated and deprotonated states.

Dramatic pH dependent variations of the  $R_2$  rates are observed for three residues close to H<sub>98</sub>, namely F<sub>96</sub>, S<sub>97</sub> and I<sub>111</sub> (Figure 6 of Supplementary Materials), indicating large changes in their  $\mu$ s-ms dynamics in the investigated pH range. These residues have large  $R_2$  rates at pH around 5 where the interconversion between the protonated and deprotonated forms of H<sub>98</sub> takes place, as inferred from NMR titration experiments (Figure 5 of Supplementary Materials) and theoretical pKa calculations. These results strongly suggest that the observed conformational exchange in cl-BABP is closely related to the protonation state of buried H<sub>98</sub>. Further support to this hypothesis comes from the observation that missing <sup>1</sup>H-<sup>15</sup>N HSQC cross-peaks, namely M<sub>73</sub>, V<sub>90</sub>, S<sub>93</sub>, K<sub>95</sub>, S<sub>97</sub>, E<sub>99</sub>, Q<sub>100</sub> are all from the same flexible C-terminal half of the protein and most of them are close to H<sub>98</sub>. In addition T<sub>91</sub> and T<sub>110</sub>, close to H<sub>98</sub>, exhibit an  $R_{ex}$  contribution at low pH, which was not detected at neutral pH.

It is worth mentioning that H<sub>98</sub>, differently from H<sub>83</sub>, is highly conserved in the iLBP family (Figure 1).

The role of histidine protonation on protein conformational change was further investigated by 3.6 ns MD simulations. It is indeed true that, in the presence of conformational exchange and pronounced flexibility, conformational sampling provided by molecular dynamics may be by far inadequate for explaining NMR experimental data obtained by sampling processes on much longer timescales. Even the processes taking place on the picosecond to nanosecond timescales may not be sampled by MD simulation simply because the conformation that enables those motions is not sampled. Nevertheless it is worthwhile examining molecular dynamics trajectories

in order to understand possible conformational trends. In this line it is worth mentioning that two molecular dynamics simulations of bovine beta-lactoglobulin (which belongs to the same superfamily of cl-BABP) were able to sample a pH-driven transition in even shorter simulation times (34,38).

Both simulations indicate a larger flexibility in the C-terminal half of the protein compared to the N-terminal half, in agreement with experimental data. The extent of conformational variability in the 10 NMR derived structures is much higher than that emerging from MD simulations, but this is linked to different timescale sampling and might also be related to lack of experimental data.

The most important suggestion coming from MD simulation is that protonation of H<sub>98</sub> has a rather dramatic effect on interactions involving residues close to residue D<sub>74</sub> which are able to trigger (at least in the simulation) the large conformational change involving the open end of cl-BABP (Figure 9) This large rearrangement is consistent with the pKa shift computed for H<sub>98</sub>, which points out the energetic cost for the neutral pH protein environment to accommodate the titration event. Within the simulation a clear closure movement of the EF loop at the open end of the protein is observed upon protonation. This conformational rearrangement finds experimental ground in the appearance of double peaks for D<sub>74</sub> at low pH, as shown in Figure 10. The structural basis for EF loop opening/closure mechanism can be identified in the presence of a network of H-bonds and salt bridges involving buried residues defining a sort of continuous polar “spine” going from E to J strand (T<sub>72</sub>, C<sub>80</sub>, S<sub>93</sub>, H<sub>98</sub>, E<sub>109</sub>, R<sub>120</sub>) (Figure 9c). Upon lowering the pH the first residue changing its protonation state is H<sub>98</sub> and the presence of this new charge could induce side-chain reorientations of the cited

residues, transmitting motion to the EF loop region, across the whole C-terminal  $\beta$ -sheet. This is further confirmed by  $^{15}\text{N}$  chemical shift changes  $> 0.2$  ppm, observed upon lowering the pH, for residues T<sub>72</sub>, C<sub>80</sub>, H<sub>98</sub> and neighbours of S<sub>93</sub> and E<sub>109</sub> pointing to a conformational change even at the level of backbone.

In order to correlate the dynamic data obtained for apo cl-BABP to a biological function, the dynamic behaviour of the protein complexed with a physiological ligand was investigated. Cholate, deoxycholate and their glyco conjugated derivatives are the most abundant bile salts, as they constitute the 80% of the natural pool (7). Interaction studies of cl-BABP with chenodeoxycholate (1:2 stoichiometry ratio) indicated that protein regions mostly affected by binding are located in the C-terminal FGHIJ strands, as deduced by significant secondary  $^1\text{H}$  and  $^{15}\text{N}$  shift changes (Figure 3). Interestingly, resonances of residues M<sub>73</sub>, V<sub>90</sub>, S<sub>93</sub>, K<sub>95</sub>, E<sub>99</sub>, Q<sub>100</sub>, not present in  $^1\text{H}$ - $^{15}\text{N}$  HSQC of apo protein, appeared in the spectra of the holo protein, thus suggesting a change in the dynamical behaviour in this region of the protein.

The comparison of the dynamical behaviour of apo and holo cl-BABP indicated that while fast motions were similarly observed in the helical regions of the protein, conformational exchange contributions, observed for apo cl-BABP at the level of EFGHI strands, substantially disappeared upon binding. (Tables 2-3, 6 of Supplementary Materials). In holo cl-BABP all the residues showing vanished  $R_{\text{ex}}$  contribution map to the regions that exhibited the highest secondary  $^1\text{H}$  and  $^{15}\text{N}$  shifts ( $> 0.5$  ppm) upon chenodeoxycholate binding (Figure 3). These results indicate that: i) the ligand is capable of stabilising one conformation and ii) in the apo protein exchange takes place between the active and inactive conformations, having high and low

affinity for the ligand, respectively. Such a mechanism is consistent with an allosteric activation, where the histidine protonation state modulates a functionally important motion, i.e. the opening/closure of loops at the entrance of the cavity, and ligand binding shifts a preexisting equilibrium. It has already been suggested, in a study on nitrogen regulatory protein (9), that the stabilisation of pre-existing conformations may be a fundamental paradigm for ligand binding. Our model nicely parallels the results on interactions of human ILBP where the binding of glycocholate has been reported to be characterized by two intrinsically weak binding sites, and strong positive cooperativity, i.e. by an allosteric mechanism where the binding of the first ligand is energetically communicated to the second site through a conformational change in the protein (7,8,39).

To investigate whether the allosteric mechanism proposed for cl-BABP could be extended to the other liver and IBABPs, an analysis of the conservation of residues involved in this allosteric mechanism (T<sub>72</sub>, D<sub>74</sub>, K<sub>95</sub>, H<sub>98</sub>, E<sub>109</sub>, R<sub>120</sub>, S<sub>122</sub>) was performed. From the sequence alignment (Figure 1) it is clear that these residues are always conserved in liver proteins, pointing to a common binding mechanism. When the comparison is extended to ILBPs, it appears that residues 74 and 95 are mutated to glycine and asparagine, respectively, while H<sub>98</sub> is conserved only in pig and rabbit species, even if a histidine is present at position 97 in human, mouse and rat species. Structures of both apo and holo proteins have been resolved for an isoform of cl-BABP (T91C)<sup>2</sup> (PDB id. 1tvq, 1tw4) (3) and for human (PDB id. 1o1u, 1o1v) (40) and pig (PDB id. 1eal, 1eio) (41,42) species. Average global RMSD differences obtained for these apo and holo structures have been compared with the average global RMSD differences between our protonated and

deprotonated simulations (Figure 11). It is clear from this comparison that the EF loop (residues 72-76) is similarly affected by ligand binding or histidine deprotonation in all the structures examined, suggesting that cl-BABP and ILBPs share the same conformational switch upon binding. Interestingly, in human ILBP another loop is strongly affected by binding, namely CD loop (52-57) where two histidines (H52, H57) are located. In this line an analysis of the dynamic properties of human ILBP, together with NMR titration experiments, could clarify whether a similar pH dependent triggering mechanism is at work for liver and ILBP proteins.

The mechanism described in this paper can be regarded as an extension of the “dynamic portal hypothesis” model (6,43,44) which implies that residues in the portal region exhibit large movements enabling the opening or closing of the portal. In the present view the event triggering this functional rearrangement is correlated with histidine protonation equilibria and a network of polar buried side-chains is likely to provide the long range connectivity needed to allosterically communicate motions from H<sub>98</sub> to the EF loop region. Further NMR titration and relaxation experiments, together with dynamics simulations, are in progress in our laboratory on selected cl-BABP mutants and other ileal lipid binding proteins to study in detail the conformational switch mechanism controlling protein activity and function.

It should be noted that a pH driven conformational change, functional to ligand binding, has been reported for other proteins belonging to the same calycin superfamily, namely beta-lactoglobulins, where the closure of the binding cavity lid is triggered by the protonation of a highly conserved glutamic acid residue (45).

*Biological implications.* In mammals bile acid circulation from the gut lumen to the ileum is mediated by ASBT and OATP3 proteins, present on the brush borders, and by ILBPs which bring them, through the cytoplasm, to the basolateral ileal membranes. Here, a truncated ASBT (tASBT) secretes bile salts into portal capillaries, where they bind to albumin and flow to the liver, where they are recognized by NTCP and OATP. In the liver cytosol, bile salts are bound by carrier protein(s) and shuttled to the canalicular membrane. Bile salts conjugated with taurine or glycine are directed for immediate secretion into bile by an ATP-dependent transporter, BSEP, located in the canalicular membrane. The transport across this membrane is the rate limiting step in the transfer of bile salts from blood to bile. Bile salts finally pass down the biliary ducts into the gallbladder for storage and ultimate expulsion into the duodenum. Over 90% of bile salts are efficiently reabsorbed from the small intestine (46). In Figure 12 a schematic representation of the described enterohepatic circulation is given. There are ample evidences that the liver of lower vertebrates such as chicken, frog, turtle, little skate, rainbow trout, has evolved specific transport proteins for mediating bile salt uptake and excretion, although the molecular basis of this transport remains to be clarified. In chicken, as in mammals, chenodeoxycholic acid is the predominant primary bile acid deriving from cholesterol catabolism, followed by cholic acid, both conjugated with taurine. Interestingly, orthologues of Slc10a1 genes of Ntcp have not been found in lower vertebrates, hence only Oatps must mediate the bile salt uptake in these species (47,48). Moreover, the functional similarity in ATP-dependent taurocholate transport between the liver membrane vesicles of rat and those of the lower vertebrates, including chicken, indicates that an evolutionary conserved

protein may be mediating the mechanism for bile acid excretion into bile (49,50). Very little is known about how bile salts are shuttled from one subcellular compartment to another. Conjugated bile acids, which carry a negative charge at physiological pH, require carrier-mediated transport to move within the enterohepatic tissues. Photoaffinity labelling experiments of ileal brush border membrane vesicles (51-53) identified a 14 kDa peripheral membrane protein, the ILBP, as component of the ileal  $\text{Na}^+$ /bile acid transporter system. In summary in the bile acid enterohepatic circulation three key steps are mediated by ASBT, ILBP and tASBT in the ileum and three are mediated by NTCP, cytosolic carrier(s) and BSEP in the liver (54). Hence similar molecules might constitute the active players in both ileocytes and hepatocytes, i.e. i) a receptor system, that binds bile salts on one surface and translocates them into the cell; ii) a cellular bile salt binding protein, that moves them across the cell and iii) an exit system, which moves bile salts out of the other side of the cell. Our hypothesis is that cl-BABP is the cytosolic protein carrying bile salts in liver in the same way as ILBP does in ileocytes. Structure and amino acid sequence of ILBPs are very similar to those of the liver BABPs (Figure 1). In particular, the identity of the bile acid binding residues shown by the two kinds of proteins (3) suggests that the two families could share a common function in ileocytes and hepatocytes, respectively, related to bile acid intracellular trafficking and targeting towards membranes (2,3).

If this hypothesis is correct, the release of bile salts by cl-BABP at acidic pH, triggered by  $\text{H}_{98}$  protonation, could be thought to occur at the canalicular membrane where a pH gradient can be generated by the  $\text{H}^+$ -pumping ATPase operating at the level of the bile acid export pump BSEP (49,55). In several cases, protein pH-dependent

conformational changes involve histidine residues which, having a pKa of 6.3 in the free form in solution, are likely to participate in structural changes around the physiological pH. A cascade of electrostatic interactions can be induced by their pH-dependent protonation/deprotonation equilibrium mechanisms. As an example, the nuclear receptor FXR, that transcriptionally regulates production, movement and absorption of bile acids (Figure 12), upon binding a bile acid molecule, is activated by His 464 that acts as a molecular switch through a  $\pi$ -cation interaction with the orthogonally oriented Trp 466. In the case of cl-BABP, the protonation/deprotonation mechanism seems to be strongly related to the modulation of the opening and closure at the protein open end and hence to the bile acid release/uptake process.

## REFERENCES

1. Makishima, M., Okamoto, A. Y., Repa, J. J., Tu, H., Learned, R. M., Luk, A., Hull, M. V., Lustig, K. D., Mangelsdorf, D. J., and Shan, B. (1999) *Science* **284**, 1362-1365
2. Vasile, F., Ragona, L., Catalano, M., Zetta, L., Perduca, M., Monaco, H., and Molinari, H. (2003) *J Biomol NMR* **25**, 157-160
3. Nichesola, D., Perduca, M., Capaldi, S., Carrizo, M. E., Righetti, P. G., and Monaco, H. L. (2004) *Biochemistry* **43**, 14072-14079
4. Thompson, J., Reese-Wagoner, A., and Banaszak, L. (1999) *Biochim Biophys Acta* **1441**, 117-130
5. Bakowies, D., and van Gunsteren, W. F. (2002) *J Mol Biol* **315**, 713-736
6. Krishnan, V. V., Sukumar, M., Gierasch, L. M., and Cosman, M. (2000) *Biochemistry* **39**, 9119-9129
7. Tochtrop, G. P., Bruns, J. L., Tang, C., Covey, D. F., and Cistola, D. P. (2003) *Biochemistry* **42**, 11561-11567
8. Tochtrop, G. P., DeKoster, G. T., Covey, D. F., and Cistola, D. P. (2004) *J Am Chem Soc* **126**, 11024-11029
9. Volkman, B. F., Lipson, D., Wemmer, D. E., and Kern, D. (2001) *Science* **291**, 2429-2433
10. Glatz, J. F., and van der Vusse, G. J. (1996) *Prog Lipid Res* **35**, 243-282
11. Marley, J., Lu, M., and Bracken, C. (2001) *J Biomol NMR* **20**, 71-75
12. Prost, E., Sizun, P., Piotto, M., and Nuzillard, J. M. (2002) *J Magn Reson* **159**, 76-81
13. Piotto, M., Saudek, V., and Sklenar, V. (1992) *J Biomol NMR* **2**, 661-665
14. Marion, D., Driscoll, P. C., Kay, L. E., Wingfield, P. T., Bax, A., Gronenborn, A. M., and Clore, G. M. (1989) *Biochemistry* **28**, 6150-6156
15. Kuboniwa, H., Grzesiek, S., Delaglio, F., and Bax, A. (1994) *J Biomol NMR* **4**, 871-878
16. Pelton, J. G., Torchia, D. A., Meadow, N. D., Wong, C. Y., and Roseman, S. (1991) *Biochemistry* **30**, 10043-10057
17. Powers, R., Clore, G. M., Bax, A., Garrett, D. S., Stahl, S. J., Wingfield, P. T., and Gronenborn, A. M. (1991) *J Mol Biol* **221**, 1081-1090
18. Hass, M. A., Thuesen, M. H., Christensen, H. E., and Led, J. J. (2004) *J Am Chem Soc* **126**, 753-765
19. Cicero, D. O., Melino, S., Orsale, M., Brancato, G., Amadei, A., Forlani, F., Pagani, S., and Paci, M. (2003) *Int J Biol Macromol* **33**, 193-201
20. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G., Shoelson, S. E., Pawson, T., Forman-Kay, J. D., and Kay, L. E. (1994) *Biochemistry* **33**, 5984-6003
21. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) *J Biomol NMR* **6**, 277-293
22. Johnson, B. A. (2004) *Methods Mol Biol* **278**, 313-352
23. Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997) *J Mol Biol* **273**, 283-298
24. Pawley, N. H., Wang, C., Koide, S., and Nicholson, L. K. (2001) *J Biomol NMR* **20**, 149-165
25. Bruschweiler, R., Liao, X., and Wright, P. E. (1995) *Science* **268**, 886-889
26. Abragam, A. (1961) *Principles of Nuclear Magnetism*, Clarendon Press, Oxford, UK
27. Lipari, G., and Szabo, A. (1981) *Biochemistry* **20**, 6250-6256
28. Mandel, A. M., Akke, M., and Palmer, A. G., 3rd. (1995) *J Mol Biol* **246**, 144-163

29. Phan, I. Q. H., Boyd, J., and Campbell, I. D. (1996) *J Biomol NMR* **8**, 369-378
30. Antosiewicz, J., McCammon, J. A., and Gilson, M. K. (1994) *J Mol Biol* **238**, 415-436
31. Fogolari, F., Esposito, G., Viglino, P., and Molinari, H. (2001) *J Comput Chem* **22**, 1830-1842
32. Wade, R. C., Luty, B. A., Demchuk, E., Madura, J. D., Davis, M. E., Briggs, J. M., and McCammon, J. A. (1994) *Nat Struct Biol* **1**, 65-69
33. van Aalten, D. M., Findlay, J. B., Amadei, A., and Berendsen, H. J. (1995) *Protein Eng* **8**, 1129-1135
34. Fogolari, F., Moroni, E., Wojciechowski, M., Baginski, M., Ragona, L., and Molinari, H. (2005) *Proteins* **59**, 91-103
35. Laskowski, R. A., Moss, D. S., and Thornton, J. M. (1993) *J Mol Biol* **231**, 1049-1067
36. Musafia, B., Buchner, V., and Arad, D. (1995) *J Mol Biol* **254**, 761-770
37. Koradi, R., Billeter, M., and Wuthrich, K. (1996) *J Mol Graph* **14**, 51-55, 29-32
38. Eberini, I., Baptista, A. M., Gianazza, E., Fraternali, F., and Beringhelli, T. (2004) *Proteins* **54**, 744-758
39. Tochtrop, G. P., Richter, C., Tang, C., Toner, J.T., Covey, D.F., and Cistola, D.P. (2002) *Proc Natl Acad Sci U S A* **99**, 1847-1852
40. Kurz, M., Brachvogel, V., Matter, H., Stengelin, S., Thuring, H., and Kramer, W. (2003) *Proteins* **50**, 312-328
41. Lucke, C., Zhang, F., Ruterjans, H., Hamilton, J. A., and Sacchettini, J. C. (1996) *Structure* **4**, 785-800
42. Lucke, C., Zhang, F., Hamilton, J. A., Sacchettini, J. C., and Ruterjans, H. (2000) *Eur J Biochem* **267**, 2929-2938
43. Hodsdon, M. E., and Cistola, D. P. (1997) *Biochemistry* **36**, 2278-2290
44. Hodsdon, M. E., and Cistola, D. P. (1997) *Biochemistry* **36**, 1450-1460
45. Ragona, L., Fogolari, F., Catalano, M., Ugolini, R., Zetta, L., and Molinari, H. (2003) *J Biol Chem* **278**, 38840-38846
46. Trauner, M., and Boyer, J. L. (2003) *Physiol Rev* **83**, 633-671
47. Boyer, J. L., Hagenbuch, B., Ananthanarayanan, M., Suchy, F., Stieger, B., and Meier, P. J. (1993) *Proc Natl Acad Sci U S A* **90**, 435-438
48. Marin, J. J., Mangas, D., Martinez-Diez, M. C., El-Mir, M. Y., Briz, O., and Serrano, M. A. (2003) *Biochim Biophys Acta* **1611**, 249-257
49. Ballatori, N., Rebbeor, J. F., Connolly, G. C., Seward, D. J., Lenth, B. E., Henson, J. H., Sundaram, P., and Boyer, J. L. (2000) *Am J Physiol Gastrointest Liver Physiol* **278**, G57-63
50. Gerloff, T., Stieger, B., Hagenbuch, B., Madon, J., Landmann, L., Roth, J., Hofmann, A. F., and Meier, P. J. (1998) *J Biol Chem* **273**, 10046-10050
51. Gong, Y. Z., Everett, E. T., Schwartz, D. A., Norris, J. S., and Wilson, F. A. (1994) *Proc Natl Acad Sci U S A* **91**, 4741-4745
52. Kramer, W., Wess, G., Bewersdorf, U., Corsiero, D., Girbig, F., Weyland, C., Stengelin, S., Enhsen, A., Bock, K., Kleine, H., Le Dreau, M. A., and Schafer, H. L. (1997) *Eur J Biochem* **249**, 456-464
53. Kramer, W., Corsiero, D., Friedrich, M., Girbig, F., Stengelin, S., and Weyland, C. (1998) *Biochem J* **333 ( Pt 2)**, 335-341
54. Small, D. M. (1997) *J Clin Invest* **99**, 1807-1808
55. Scharschmidt, B. F., and Van Dyke, R. W. (1987) *Annu Rev Physiol* **49**, 69-85

#### FOOTNOTES

\* CIRMMP (Consorzio Interuniversitario di Risonanze Magnetiche di Metalloproteine Paramagnetiche), Fabio Calogiuri and Massimo Lucci (CERM, Florence) are gratefully acknowledged for the <sup>15</sup>N relaxation measurements performed at 600 and 700 MHz. Fulvia Greco is acknowledged for technical assistance. This research was supported by FIRB 2001, FIRB 2003 (grant RBNE03B8KK) from the Italian Ministry for Education, University and Research MIUR 2004.

<sup>1</sup> The abbreviations used are: ASBT, apical sodium-dependent bile salt transporter; cl-BABP, chicken liver bile acid binding protein; FABP, fatty acid binding protein; FXR, farnesoid X receptor; het-NOE, heteronuclear Nuclear Overhauser Effect; iLBP, intracellular lipid binding protein; ILBP, ileal lipid binding protein; IPTG, isopropylthiogalactopyranoside; MD, molecular dynamics; NTCP, sodium-taurocholate cotransport protein; OATP3, organic anion transport protein.

<sup>2</sup> Monaco, HL., personal communication.

## FIGURE LEGENDS

Fig. 1. ClustalW multiple alignment of proteins belonging to iLBP family. The alignment includes the thirteen “liver basic” fatty acid binding proteins from non mammalian species and the five known ILBPs. Secondary structure elements are highlighted on the top of the sequences.

Fig. 2. Sausage representation (MOLMOL) (37) of the superimposed ten final minimised structures of apo cl-BABP.

Fig. 3. Chemical shift changes upon chenodeoxycholate binding at pH 7 and 298 K. (a)  $^1\text{H}$  and  $^{15}\text{N}$  shifts ( $\Delta\delta(\text{HN},\text{N}) = [(\Delta\delta_{\text{HN}}^2 + \Delta\delta_{\text{N}}^2/25)/2]^{1/2}$ ) bar diagram versus residue number. A dotted line at 0.4 ppm is drawn to easily identify residues exhibiting the highest shifts. In (b) residues mostly influenced by binding are drawn onto the protein structure in black. Unassigned residues are shown in dark grey.

Fig. 4  $^{15}\text{N}$  relaxation parameters for apo cl-BABP at pH 7.0 and 298 K.  $R_2/R_1$  ratios (a), and het-NOE effects (b) obtained at two different fields, 600 MHz (triangle) and 700 MHz (circle), are plotted as a function of residue number. Error bars are shown.

Fig. 5. Comparison of  $S^2$ ,  $R_{\text{ex}}$  and  $\tau_e$  values obtained from Lipari-Szabo analysis of apo cl-BABP at pH 7 (circle) and 5.6 (triangle) vs residue number.

Fig. 6. Color-coded representation of backbone dynamics of apo cl-BABP (left panel) and holo cl-BABP (right panel) at pH 7.0. Residues affected by  $R_{\text{ex}}$  (red) and by  $\tau_e$  (blue) are highlighted. Residues whose amide signals were broadened beyond detection are reported in orange, while dark grey stretch of ribbon are related to those residues which could not be included in the analysis due to resonance overlap.

Fig. 7.  $R_{\text{ex}}$  contributions of apo cl-BABP at pH 7.0, 298 K derived with different approaches. Results of Lipari-Szabo approach using model 3 for data fitting (circle);  $R_{\text{ex}}$  derived as  $R_2(\text{experimental}) - R_2(\text{fitted})$ , where data fitting was done with model 1 of Lipari –Szabo approach (grey square, 600 MHz; white square 700 MHz);  $R_{\text{ex}}$  values deduced from data at three frequencies following the approach described in (reference Phan) (triangle). Only upper part of error bars is shown for clarity purposes.

Fig. 8. Comparison of the most representative structures derived from MD simulations performed at acid and neutral pH: RMSD obtained from global superposition is plotted as a function of residue number.

Fig.9. Backbone and side-chain arrangement in the most representative structures of MD ensembles obtained at acidic (red) and neutral pH (blue) for cl-BABP. (a) Differences of backbone conformation at the open end of the protein are highlighted. (b) Residues involved in the pH driven conformational change and their H-bond patterns are shown: labels are in colour when necessary, to distinguish the different side-chain orientations at acidic and neutral pH. Green and grey dotted lines represent H-bonds and salt bridges, respectively. Ribbon regions 97-98 and 108-122 were made invisible for clarity. (c) Side-chain and Van der Waals surface of residues T<sub>72</sub>, C<sub>80</sub>, S<sub>93</sub>, H<sub>98</sub>, E<sub>109</sub>, R<sub>120</sub> belonging to buried polar “spine” encompassing E-J strands are shown on cl-BABP structure.

Fig. 10. Expansion of  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiments performed on cl-BABP at different pHs: double peaks for NH cross-peak of residue D<sub>74</sub> appear upon lowering the pH.

Fig. 11. Average global displacement vs residue number as obtained for the superposition of the most representative structures of protonated and deprotonated MD simulations of cl-BABP (blue line) compared with the differences between apo and holo structures of human ILBP (PDB id.: 1o1u, 1o1v) (red line), pig ILBP (PDB id.: 1eal, 1eio) (green line) and the T91C isoform (PDB id. 1tvq, 1tw4) of apo cl-BABP. A dotted line evidences EF loop region.

Fig. 12. Bile acid enterohepatic circulation (see paragraph *Biological implications* for details).

**Table 1.** Analysis of the 10 best structures obtained for apo cl-BABP at pH 7.0 and 298 K.

---

(a) Restraints	
----------------	--

---

Number of upper limit distance restraints	1000
Number of hydrogen bond restraints	13
Number of torsion angle ( $\phi$ ) restraints	48

---

(b) DYANA	
-----------	--

---

Target function ( $\text{\AA}^2$ )	$2.07 \pm 0.46$
Average number of upper restraint violation $> 0.25 \text{\AA}$ per structure	0
Maximum violation ( $\text{\AA}$ )	0
Average number of angle restraint violations $> 5^\circ$ per structure	0
Maximum violation (degrees)	0
RMSD (backbone atoms) (3-125)	$1.47 \pm 0.22$
RMSD (heavy atoms) (3-125)	$2.16 \pm 0.18$

---

(c) DISCOVER (AMBER forcefield)	
---------------------------------	--

---

Total energy (kcal/mol)	$-402 \pm 18$
Bond energy (kcal/mol)	$21 \pm 1$
Angle energy (kcal/mol)	$141 \pm 2$
Torsion angle (kcal/mol)	$160 \pm 6$
Out of plane energy (kcal/mol)	$3.9 \pm 0.4$
Hydrogen bond energy (kcal/mol)	$-44 \pm 2$
Lennard-Jones energy (kcal/mol)	$-362 \pm 14$
Coulomb energy (kcal/mol)	$-321 \pm 12$
Restraining potential energy (kcal/mol)	$67 \pm 12$
Average number of upper restraint violations $> 0.25 \text{\AA}$	0
Maximum violation ( $\text{\AA}$ )	0
Average number of angle restraint violations $> 5^\circ$ per structure	0
Maximum violation (degree)	0
RMSD (backbone atoms) (3-125)	$2.02 \pm 0.26$
RMSD (heavy atoms) (3-125)	$3.07 \pm 0.25$

---

	A	I	II	B	C
	10	20	30	40	
FABPL_CHICK	-AFSGTWQVYAQENYEEFLKALALPEDLIK	MARDIKPIVEIQQK	GDDFVV		
Q8UWE2_ANAPL	MAFSGTWQVYAQENYEEFLKALALSEDI	IKVARDIKPVVEIQQK	GDDFVV		
FABPL_RHASA	-AFSGTWQVYAQENYEEFLRAISLPEDVI	KLAKDVKPVTEIQQT	GNDFVI		
FABPL_ANOPU	-AFNGTWQVYSQENYEDFLKAIALPDDI	IKAAKDVKPVTEIRQT	GNTFVV		
FABP2_AMBME	-PFNGTWQVYSQENYEAFLRAVGLPEDI	INVAKDINPIIEIQQ	NGDNFVV		
Q9I8L5_BRARE	MAFSGTWQVYAQENYEEFLRAISLP	EEVIKLAKDVKPVTEIQQ	NGSDFTI		
Q7LZK9_LATJA	MDFSGTWQVYAQENYEEFLRAMELPAD	VIKMAKDIKPI	EIKQSGNDFVI		
Q7T1A3_PLAFE	MEFNGTWQVYSQENYEEFLRAMDLP	EDVIKMAKDIKPI	EIKQSGKDFVI		
Q8JJ05_ACASC	---GTWQVYAQENYEEFLRAMDLP	ADVIKMAKDIKPI	EIKQSGNDFVI		
Q645P9_FUNHE	MDLSGTWQVYSQENYEAFLRAMELP	EDVIKMAKDIKPI	EIKQNGHDFVI		
FABPL_BUFAR	-AFNGTWNVYAQENYENFLRTVGLP	EDIKVAKDVPVIEIEQ	NGNEFVV		
FABPL_HALBI	-AFSGTWQVYSQENIEDFLRALSLP	EEVIKIGKDIKPV	DIKQTEGFVI		
FABPL_LEPPA	-AFSGTWQVYAQENYEAFLKVI	GVAEDIIPHAKEIKPTIEIQ	QSGNSFTV		
ILBP_PIG	-AFTGKYEIESEKNYDEFMKRLAL	PSDAIDKARNLKIIEV	KQDQGNFTW		
ILBP_RABIT	-AFTGKFEMESEKNYDEFMKLL	GLPSDVVEKSRNIKIVTE	EIKQDQDFTW		
ILBP_HUMAN	-AFTGKFEMESEKNYDEFMKLL	LGISSDVIEKARNFKIVTE	VQDQDQDFTW		
ILBP_MOUSE	-AFSGKYEFESEKNYDEFMKRL	GLPGDVIERGRNFKIITE	VQDQDQDFTW		
ILBP_RAT	-AFTGKYEFESEKNYDEFMKRL	GLPDEVIERGRNFKIITE	VQDQDGENFTW		
	*::: : : * : * : : . . : : : . : : . : : . : : * * *				

	C	D	E	F	G	H
	50	60	70	80	90	
FABPL_CHICK	TSKTPR-QVTNSFTLGKEADIT	TMDGKKLKCTVHL	LANGKLVTKSEKFSH			
Q8UWE2_ANAPL	TSKTPK-QSVTNSFTLGKEADIT	TMDGKKLKCTVNL	VNGKLVCKSDKFSH			
FABPL_RHASA	TSKTPG-KSVTNSFTIGKEAEIT	TMDGRKLCIVKLEGGKLI	SETEKFSH			
FABPL_ANOPU	TSKTPN-KSVTNSFTLGKEADMT	TMDGKKVKCTVNL	VDGKLVAKSDKFIH			
FABP2_AMBME	TSKTPN-QSVTNSFTIGKEAEIT	SMGGKKIKCTVVLEGGKLV	SKTDQFSH			
Q9I8L5_BRARE	TSKTPG-KVTNSFTIGKEAEIT	TMDGKKLKCI	VLDGGKLVCRDTRFSH			
Q7LZK9_LATJA	TSKTPG-KVTNSFTIGKEADIT	TMDGKKIRCVVNLEGGKLV	CNTGKFSH			
Q7T1A3_PLAFE	TSKTPG-KSVTNSFTVGKEADIT	TMDGKKIKCIVNMEGGKLV	CNTGKFSH			
Q8JJ05_ACASC	TSKTPG-KVTNSFTIGKEAEIT	TMDGKKLKCI	VNMEGGKLVCKTGKFSH			
Q645P9_FUNHE	TFKTPG-KFVTKSFTIGKEAEIT	TMDGKKLKCI	VNMEGGKLVCKTGKFSH			
FABPL_BUFAR	TSKTPK-QTHSNSFTVGKESEIT	SMGKKIKVTVQLEGGKLI	CKSDKFSH			
FABPL_HALBI	VVKTSQ-QVTNFTVGKEAEIT	SMGKKLKCTVQLEDGKLV	AKLKFTH			
FABPL_LEPPA	TSTAQK-KSTNTFTIGKEAEIT	TMNGNKL	RCTINMEDGKLVCKTEKFSH			
ILBP_PIG	SQQYPGGHSITNTFTIGKECDI	ETIGGKKFKATVQMEGGKVV	VNSPNYHH			
ILBP_RABIT	SHHYSGGQIMTNKFTIGKESEI	QTFGGKKFKAVVNMEGGKVV	ANFPNYQH			
ILBP_HUMAN	SQHYSGGHTMTNKFTVGKESNI	QTMGGKTFKATVQMEGGKLV	VNFPNYHQ			
ILBP_MOUSE	SQSYSGGNIMSNKFTIGKECEM	QTMGGKFKATVKMEGGKVV	AEFPNYHQ			
ILBP_RAT	SQSYSGGNIMSNKFTIGKECEM	QTMGGKFKATVKMEGGKVV	ADFPNYHQ			
	: : : * : * * : : : . : * . : : : : : * * : : : . : : :					

	H	I	J	
	100	110	120	
FABPL_CHICK	EQEVKGNEMVETITFGGVTL	IRRSKR	V-	
Q8UWE2_ANAPL	EQEVNGNEMVETITFGGVTL	VRRSKR	V-	
FABPL_RHASA	KQEI KGGEMIE	TLTVAGTTMVRK	SKV-	
FABPL_ANOPU	EQEIVGNEMVETITSGSAT	FTRRSK	KI-	
FABP2_AMBME	IQEVKGNEMVETLTVGGAT	LIRRSKR	V-	
Q9I8L5_BRARE	IQEI KAGEMVETLTVGGT	TMIRSK	KI-	
Q7LZK9_LATJA	IQELRGEMVETLTMGST	TLIRKSK	MM-	
Q7T1A3_PLAFE	IQELQGGEMVETLTMGSV	SLIRKSK	MM-	
Q8JJ05_ACASC	IQELKGGEMIE	TLTMGST	LVRKSKMM-	
Q645P9_FUNHE	TQELKGGEMVETLTVGST	TLVRKSR	KI-	
FABPL_BUFAR	IQEVNGDEMVEKITIGS	STLTRKSR	V-	
FABPL_HALBI	IQEVQGNEMIEKLTAG	NATMIRKSR	RM-	
FABPL_LEPPA	IQEVQGEEMIE	TLTSGSAT	LIRRSKR	V-
ILBP_PIG	TAEIVDGKLV	EVSTVGGV	TYERVSKKLA	
ILBP_RABIT	TSEIKGDKLV	EVSSI	GGVTYERVSKRLA	
ILBP_HUMAN	TSEIVGDKLV	EVSTI	GGVTYERVSKRLA	
ILBP_MOUSE	TSEVVGDKLV	EISTIG	DVTYERVSKRLA	
ILBP_RAT	TSEVVGDKLV	EISTIG	DVTYERVSKRVA	
	* : : : * : : . : : * * : : :			

Figure 1

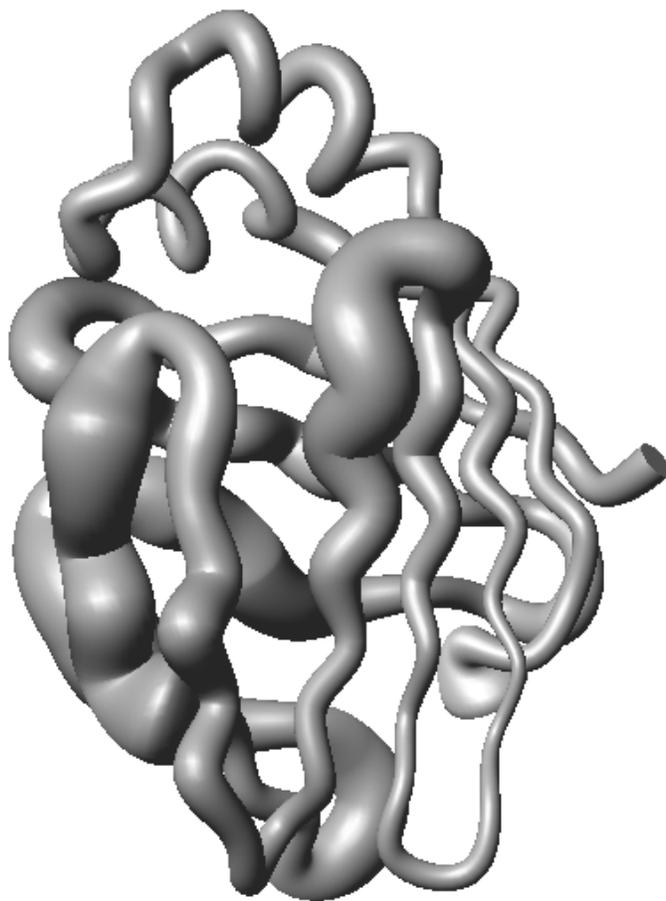


Figure 2

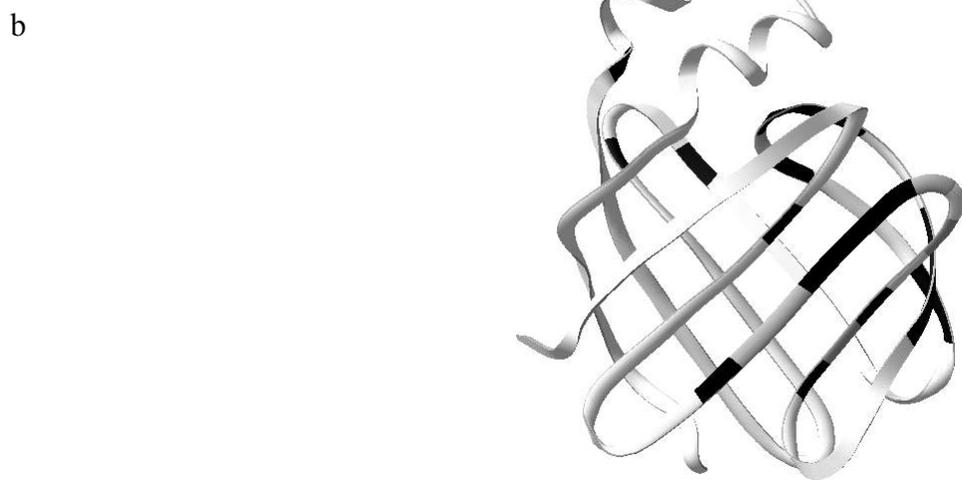
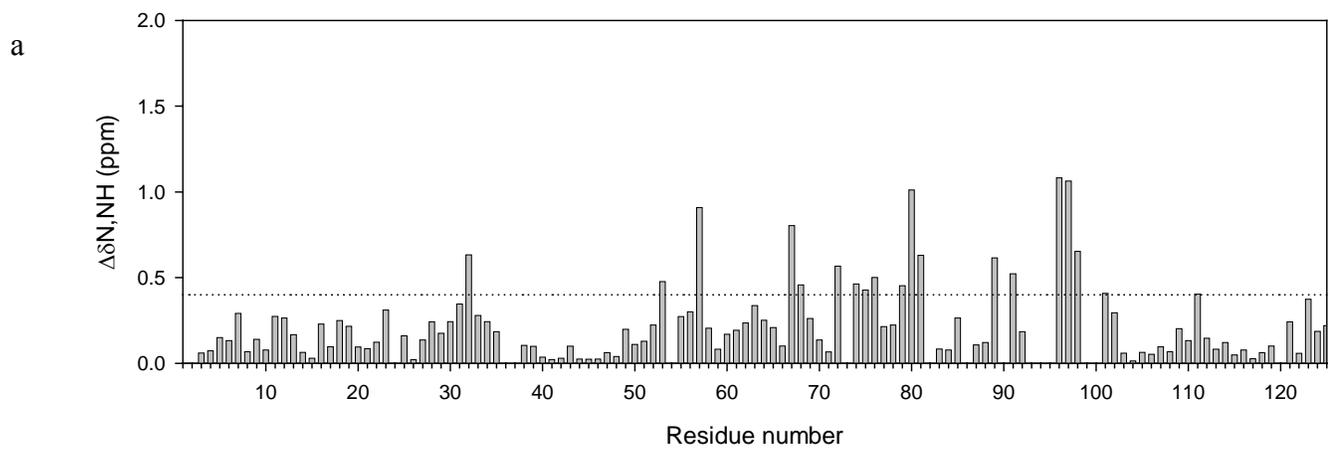


Figure 3

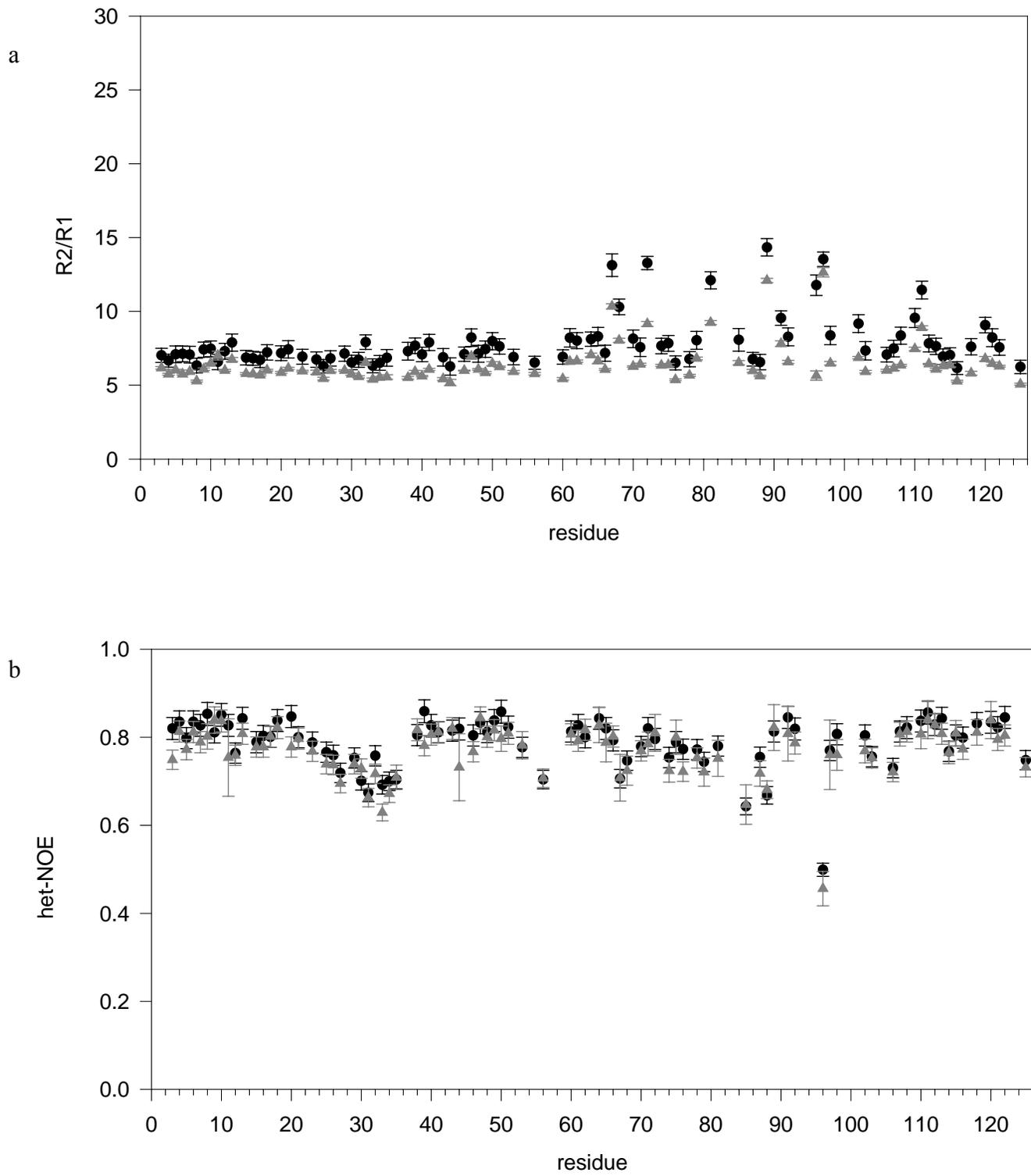


Figure 4

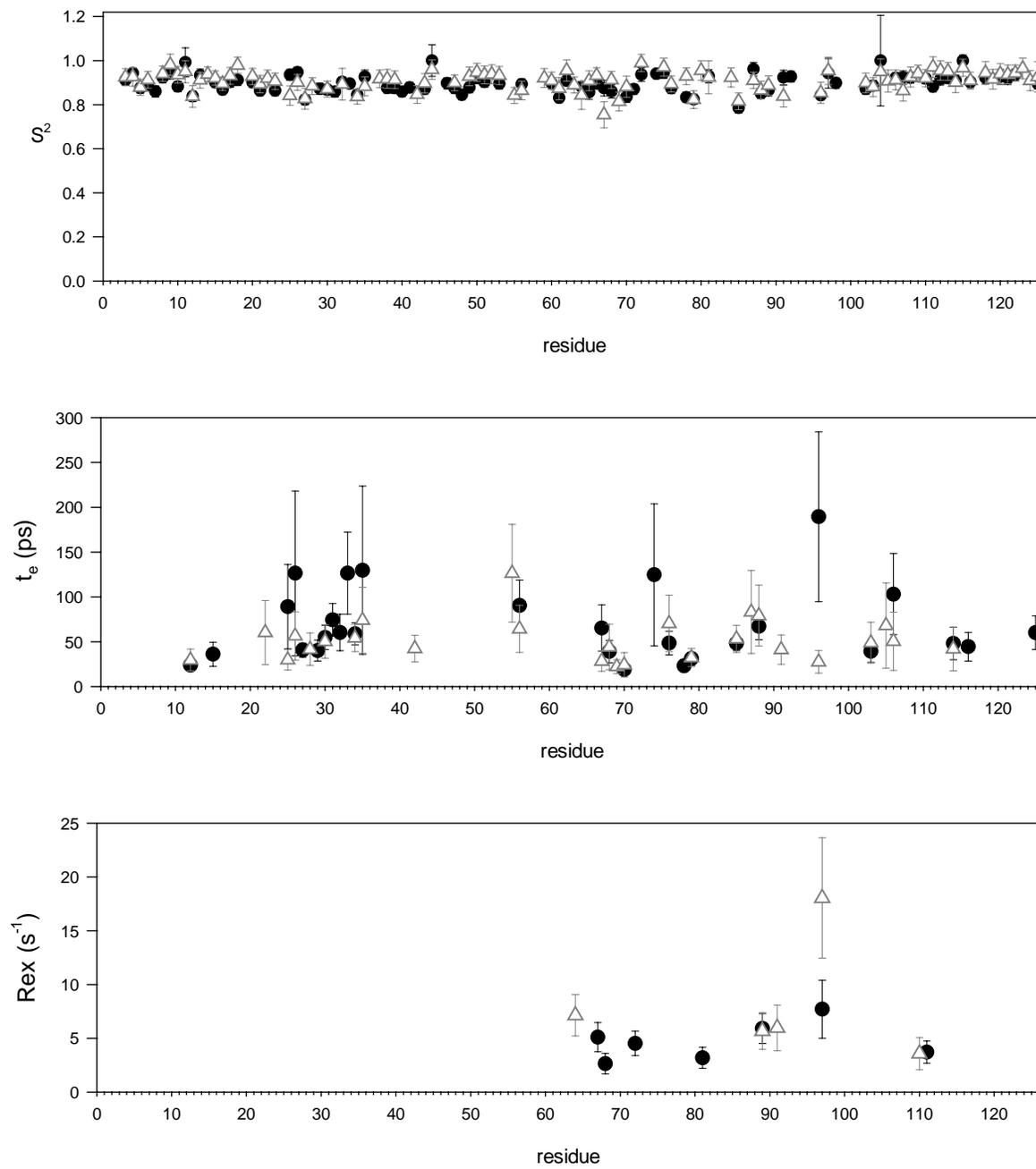


Figure 5

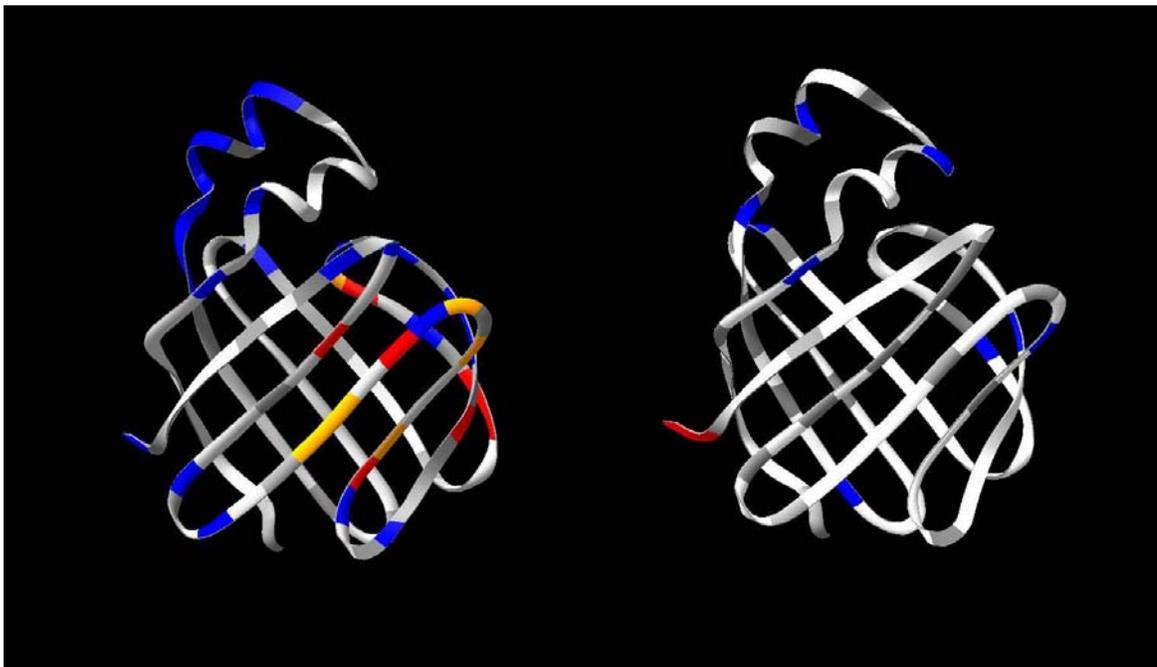


Figure 6

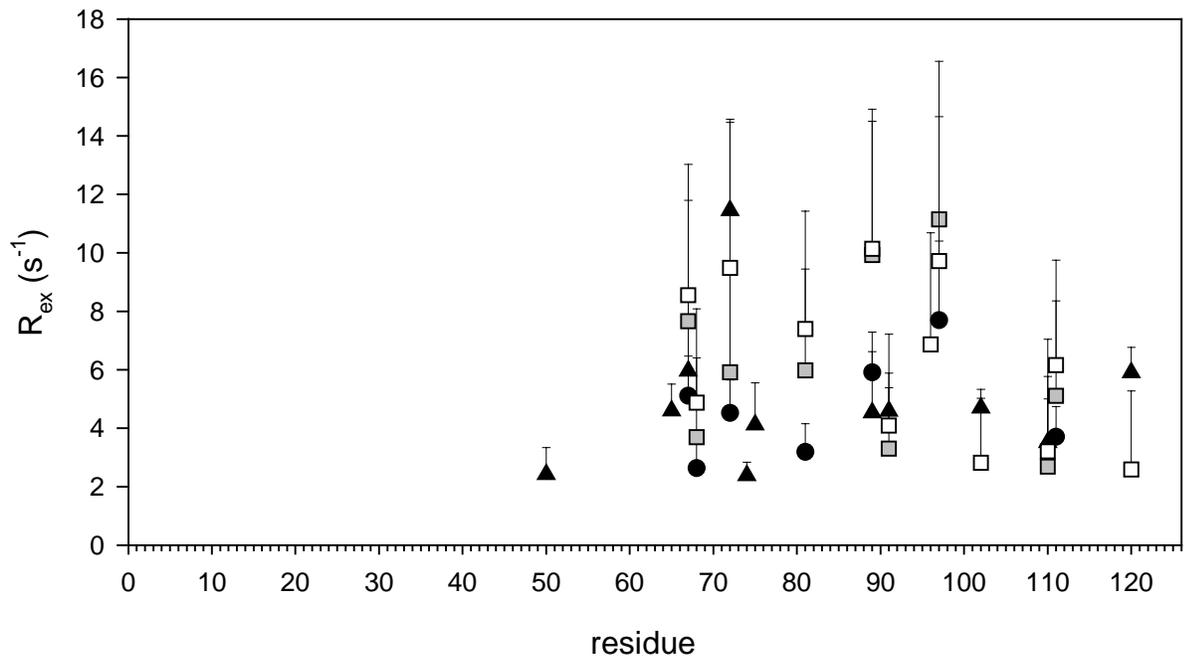


Figure 7

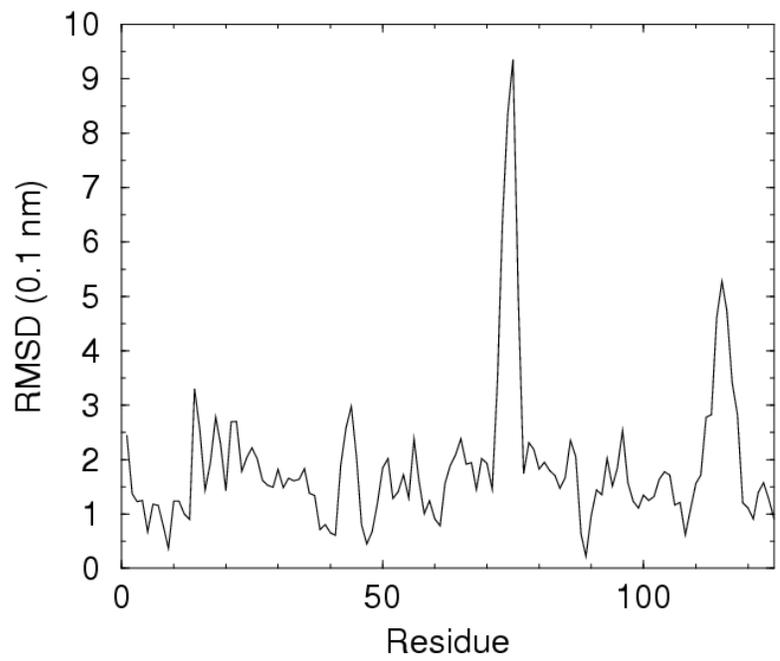


Figure 8

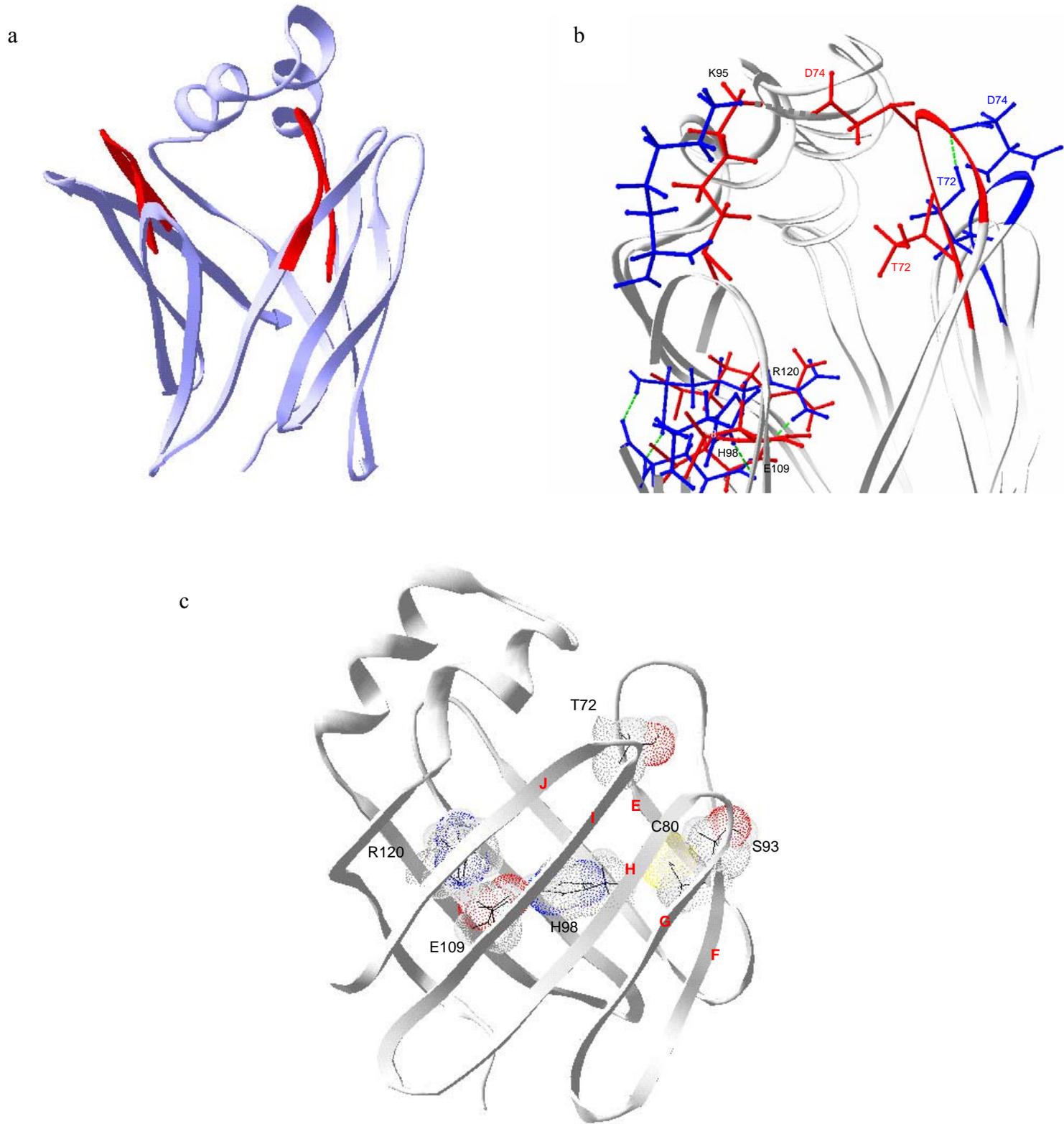


Figure 9

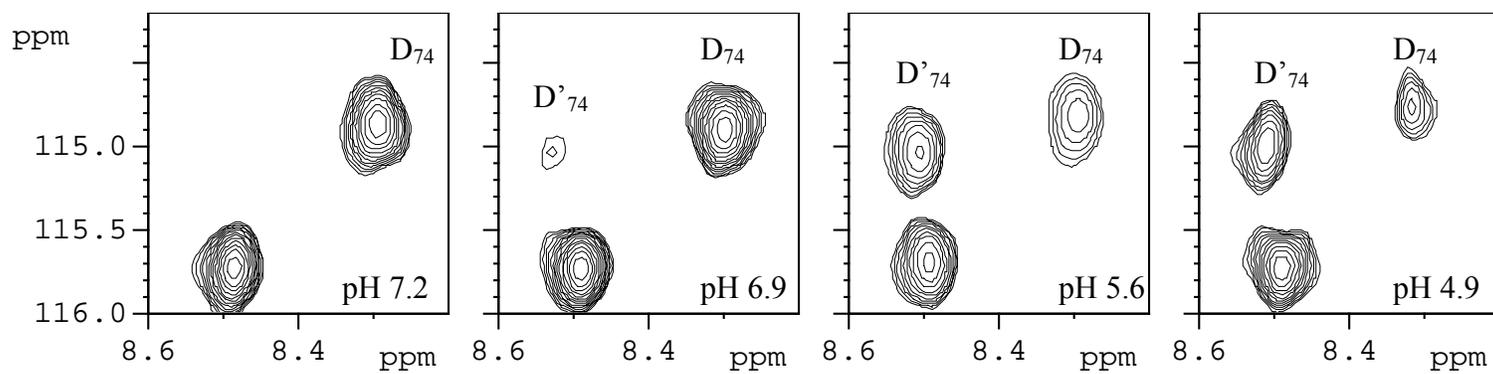


Figure 10

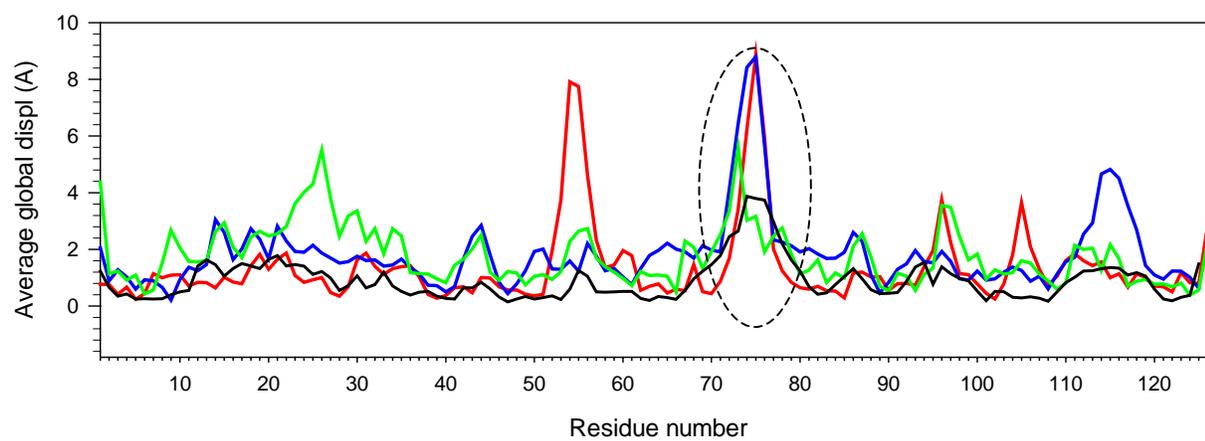


Figure 11

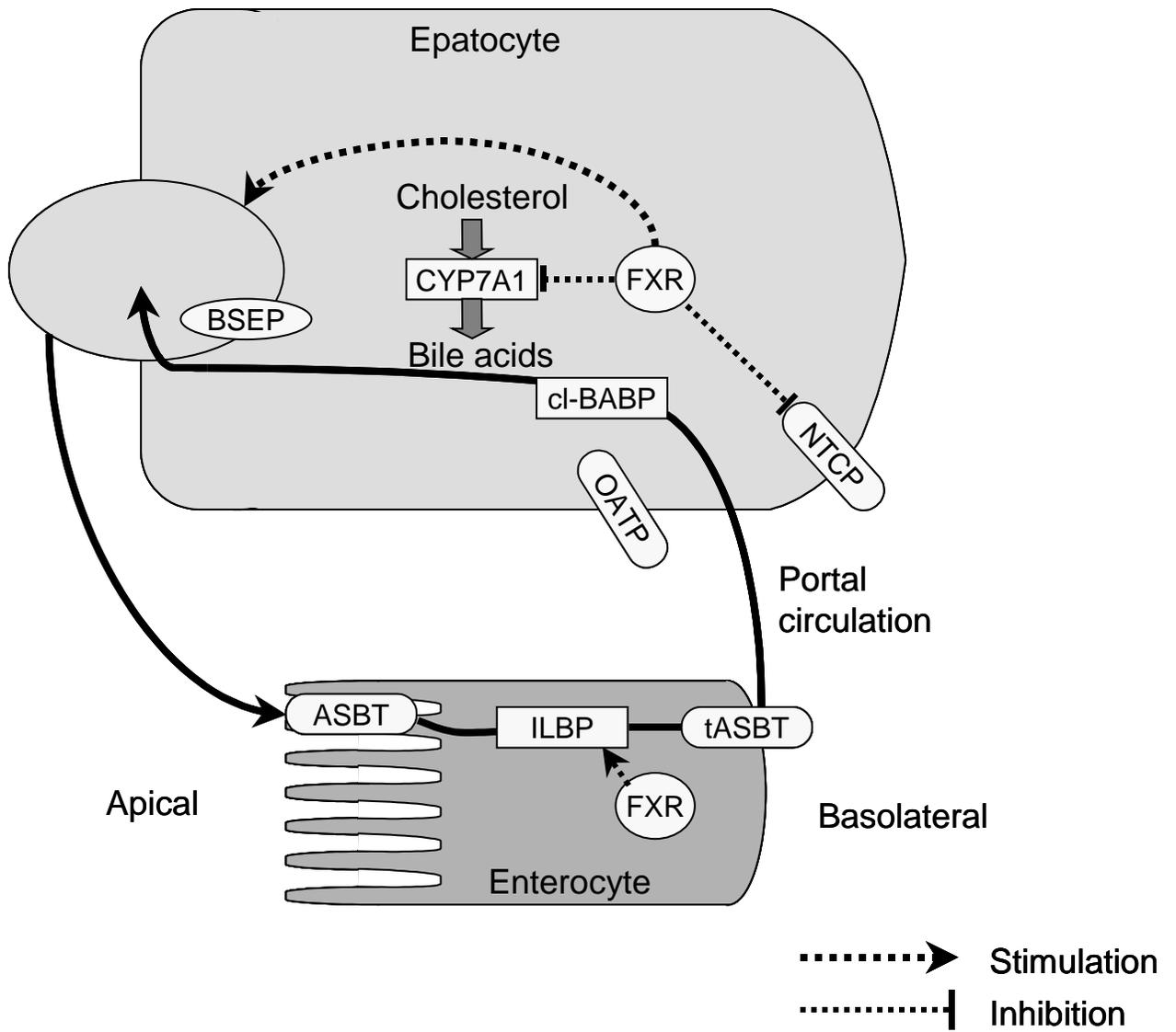


Figure 12

## 4.7 Mutants of cL-BABP

Site-direct mutagenesis is a common technique for establishing the importance of specific residues in protein function, in particular the role of residues involved in the binding properties of cL-BABP was investigated. The cL-BABP mutants produced are: L21R, R120Q, T91C, H98Y. For each mutation, the cDNA sequence was checked by sequencing and resulted correct. For each mutant, expression and purification procedures followed the procedure described for the production of unlabeled cL-BABP. Modified procedures are reported in the following sections; in Appendix the main biochemical characteristics of each mutant are reported.

### 4.7.1 Rationale, production, NMR analysis of cL-BABP mutants

#### L21R

*The portal domain:* The substitution of leucine 21 with arginine was designed to form an arginine<sup>21</sup>-aspartic<sup>74</sup> salt bridge across the region thought to be the preferential pathway for ligand entry in to the binding cavity.

The expression was performed at 24°C overnight and L21R cL-BABP was present in the soluble fraction; from the comparison of the total fraction and soluble fraction by SDS PAGE, it was observed that almost all the expressed protein was soluble (Figure 4.27).

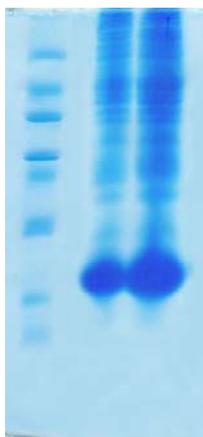


Figure 4.27. Lane 1: Marker ; lane 2: soluble fraction, lane 3: total fraction

#### Purification procedure

The theoretical isoelectric point (pI) for L21R resulted 9.02.

The final yield for 1 litre of LB/Kan was 85 mg/L. The protein purity was checked by the presence of a single band on SDS-PAGE and by mass spectrometry (Figure 4.28). The experimental molecular mass resulted correct (14122,2 Da).

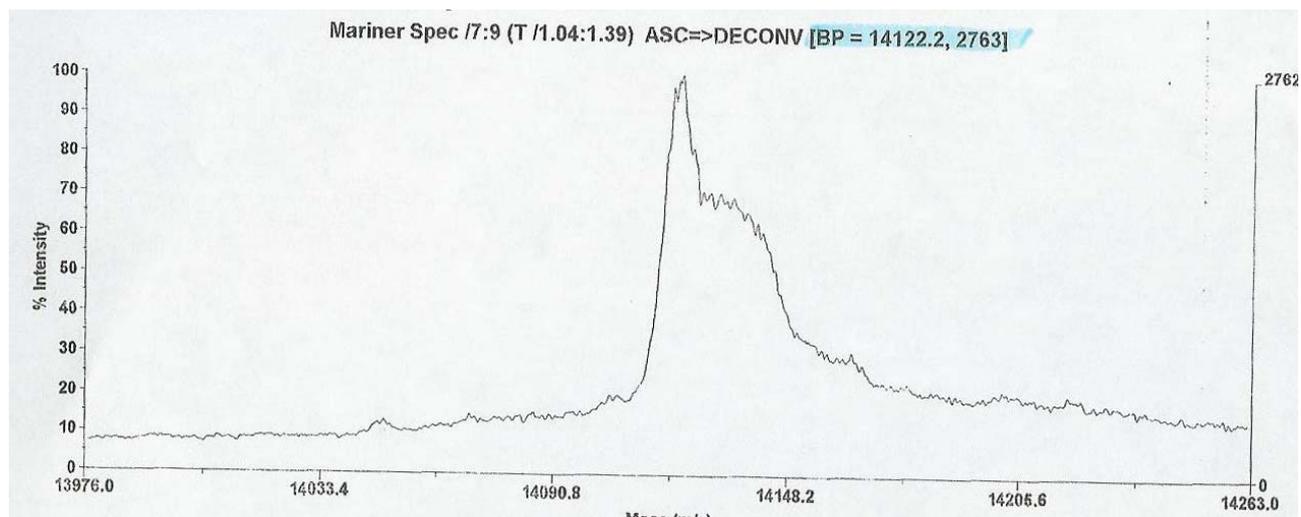
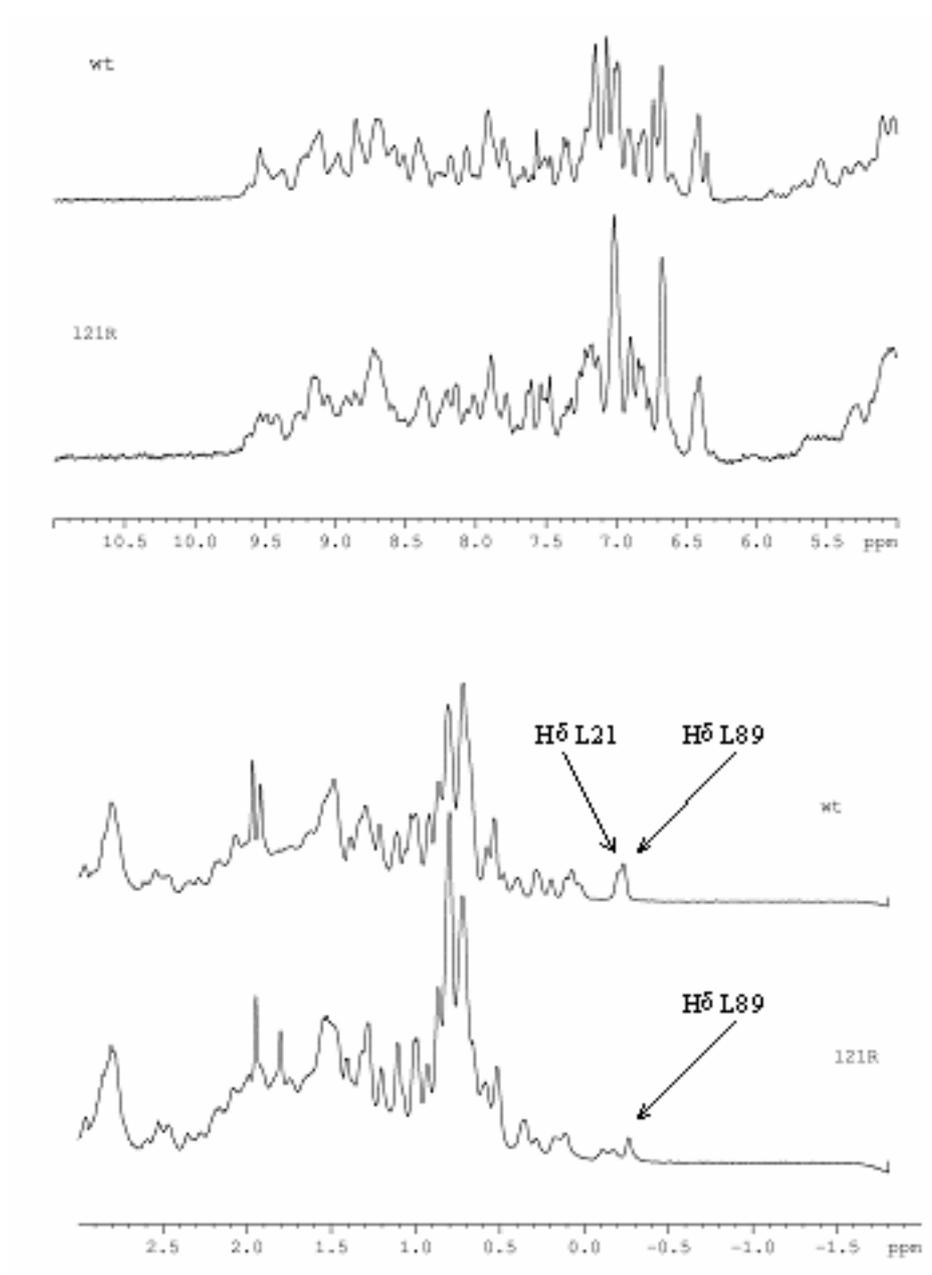


Figure 4.28. MALDI TOF spectrum of L21R cL-BABP

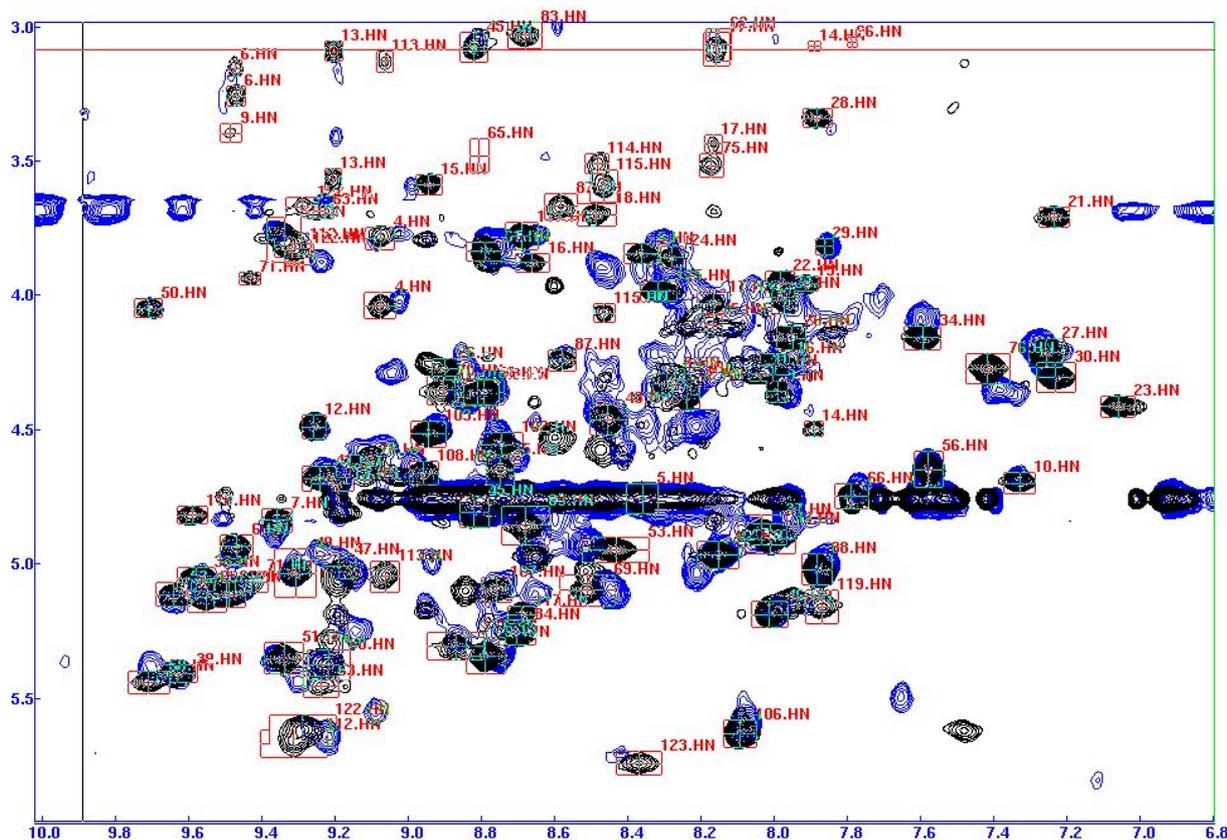
### NMR experiments

NMR analysis of the L21R cL-BABP mutant indicated that the protein was folded and stable for at least one week at room temperature. From the comparison of the 1D spectra of wild type protein and L21R at pH 7.0, 30 mM PBS at 298 K many differences were observed in the amide and aromatic regions, as well as in the high field region containing the methyl signals. The side-chain methyl resonances of L21 e L89 are indicated in Figure 4.29; the signal of L21 disappears, as expected, from the spectrum of the mutant protein.



**Figure 4.29.** 1D 500 MHz <sup>1</sup>H-NMR spectra of delipidated wild type and L21R cL-BABP at pH 7.0, 30 mM PBS buffer (95% H<sub>2</sub>O, 5% D<sub>2</sub>O), at 298 K; the chemical shift of L21 disappears.

Figure 4.30 shows the overlap of the fingerprint region of 2D-TOCSY spectra of wild type cL-BABP (black) and of L21R cL-BABP (blue) obtained at 298 K. In red, the assignments of the wild type protein are reported.



**Figure 4.30.** Fingerprint regions of 2D 500 MHz  $^1\text{H}$ -NMR spectra of 1 mM delipidated wt cL-BABP in 30 mM PBS buffer at pH 7.0 (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), 298 K (black) TOCSY spectrum of wild type cL-BABP; (blue) TOCSY spectrum of L21R cL-BABP.

The mutation in position 21 has led to many chemical shift variations that influence amino acid residues distant in the space. About 30% of the residues show meaningful variations suggesting that the mutation induces a conformational change that involves the region at the open end such as the  $\alpha$  helix, the A- and J-strands and the CD, EF and IJ loops. Long-distance effects localised in the  $\beta$ -sheet was also observed. The characterization of the structural changes induced by the mutation requires the production of  $^{15}\text{N}$  labeled protein and the acquisition and assignment of the tri-dimensional spectra. The analysis of the proton two-dimensional spectra gives a preliminary indication of which regions are affected by the mutation. Those regions are indicated in red in Figure 4.31. The amino-acid chain in position 21 is shown in red.

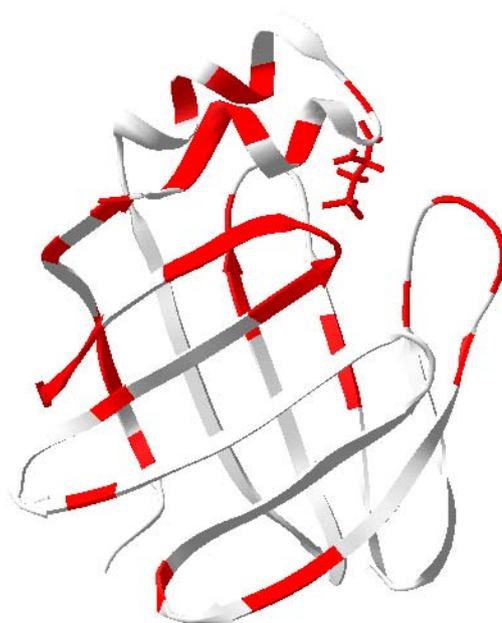


Figure 4.31. L21R cL-BABP; regions influenced by the point mutation.

## R120Q

*Cavity binding domain:* as described in chapter 1, arginine 120 is involved in electrostatically interaction with the carboxylic group of fatty acids; it was decided to invert the charge introducing a glutammic acid residue in order to observe possible changes in the binding properties of the protein.

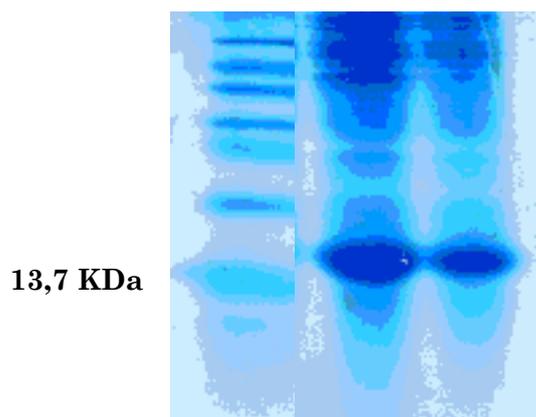


Figure 4.32. Lane 1: Marker ; lane 2: total fraction, lane 3: soluble fraction

The expression was performed at 24°C overnight and R120Q cL-BABP was present in the soluble fraction; from the comparison of the total fraction and soluble fraction by SDS PAGE, it was observed that not all the expressed protein was soluble (Figure 4 .32).

### Purification procedure

The theoretical isoelectric point (pI) for R120Q resulted 8.05 the pH of the buffer for anion exchange chromatography was therefore changed to 7.3.

The final yield for 1 litre of LB/Kan was 50 mg/L. The protein purity was checked by the presence of a single band on SDS-PAGE and by mass spectrometry (Figure 4.33). The experimental molecular mass resulted correct (14052,5 Da).

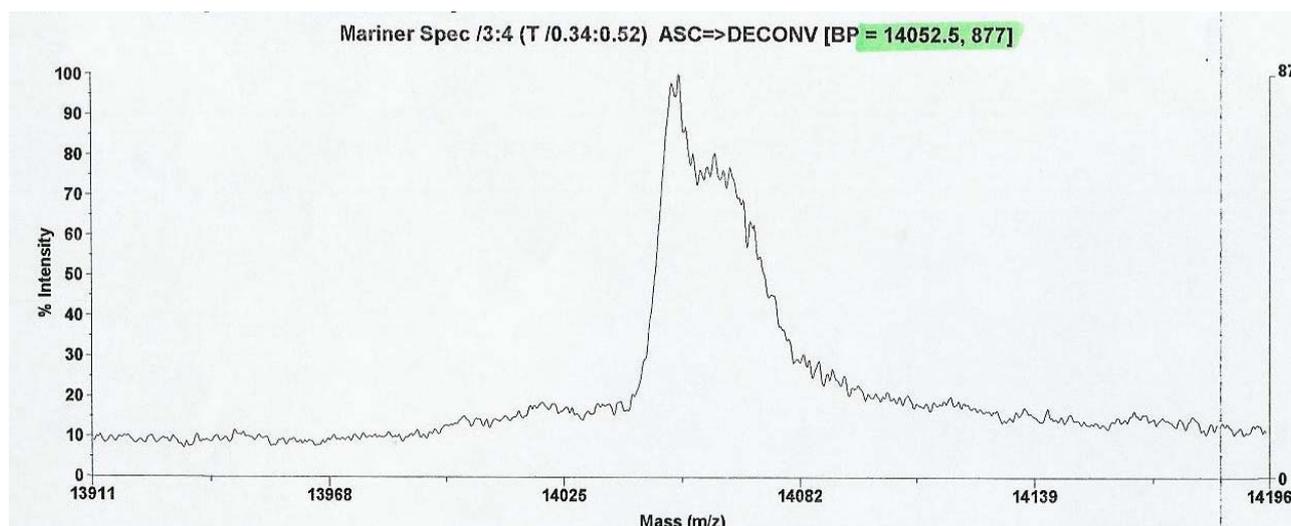


Figure 4.33. MALDI TOF spectrum of R120Q cL-BABP

### NMR experiments

NMR analysis of the R120Q cL-BABP mutant indicated that the protein was folded and stable for at least one week at room temperature (Figures 4.34, 4.35, 4.36).

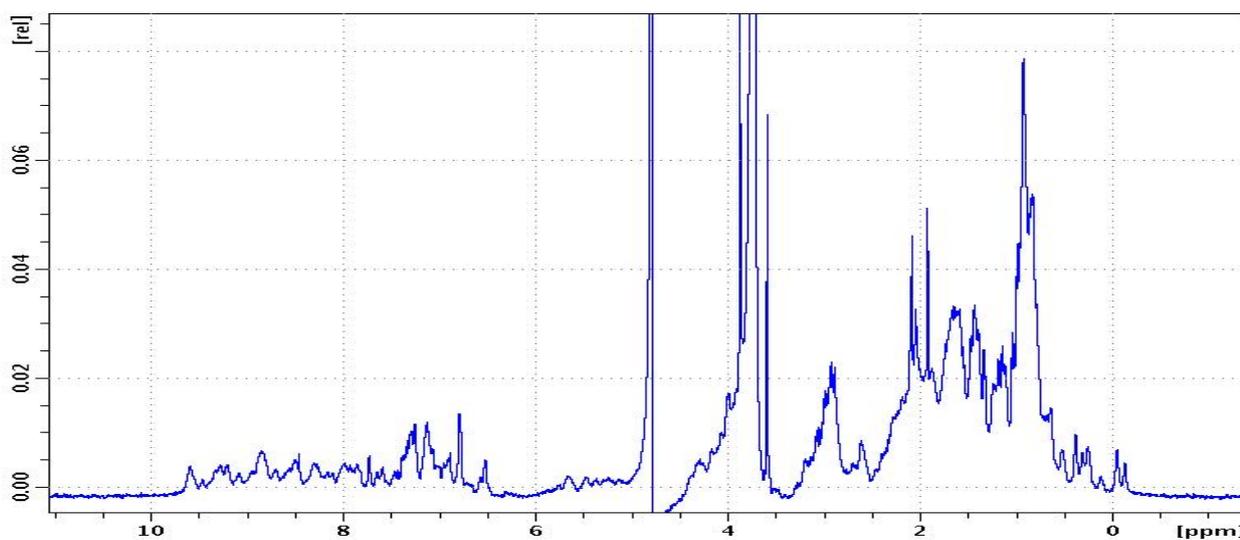


Figure 4.34. 1D 500 MHz  $^1\text{H}$ -NMR spectra of 0.5 mM delipidated R120Q cL-BABP at pH 7.0, 30 mM PBS buffer (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), at 298 K.

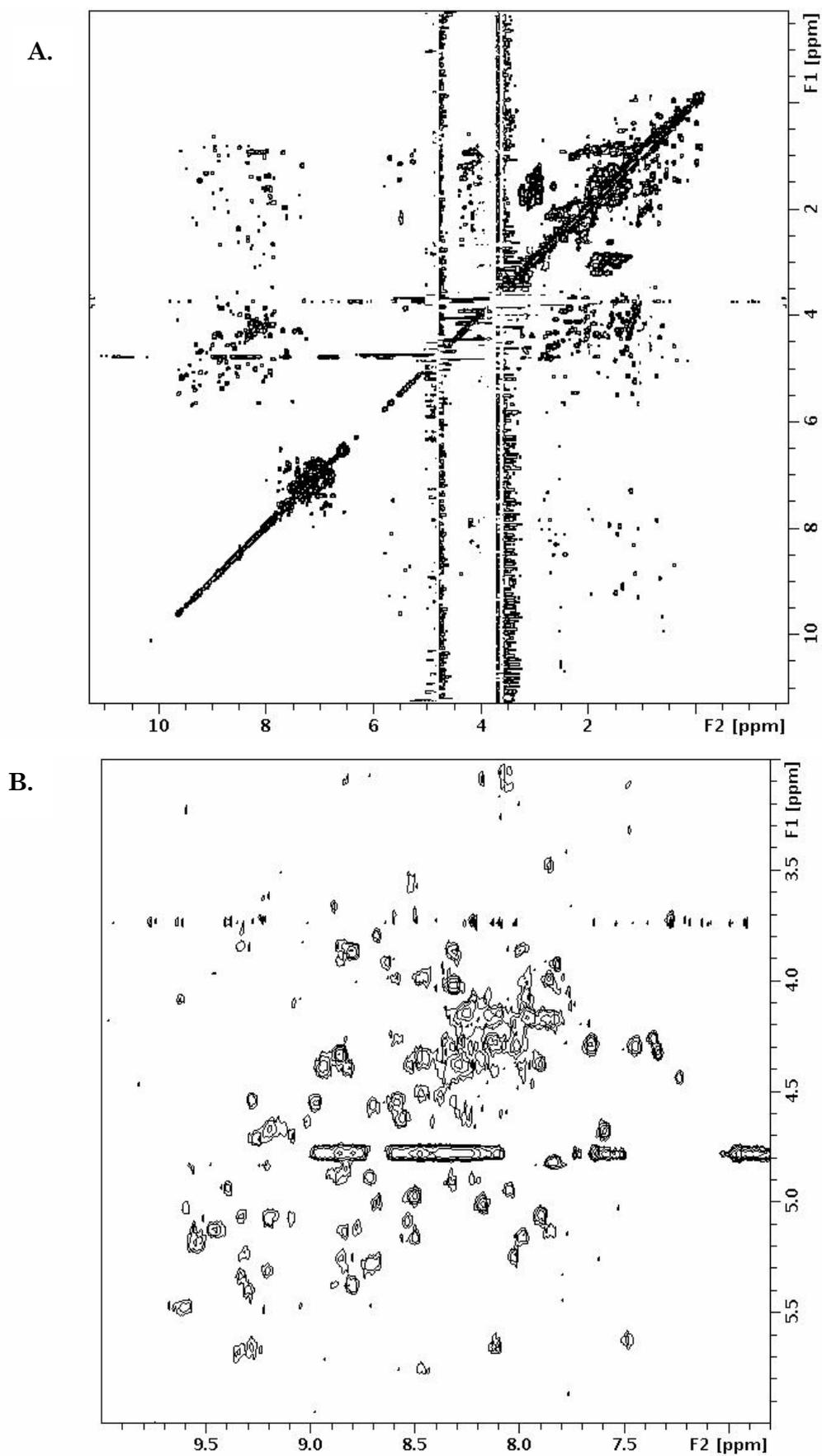


Figure 4.35. A. Fingerprint region of 2D 500 MHz  $^1\text{H}$ -NMR TOCSY spectrum of 0.5 mM delipidated R120Q cL-BABP at pH 7.0, 30 mM PBS buffer (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), at 298 K. B. Detail of the spectrum.

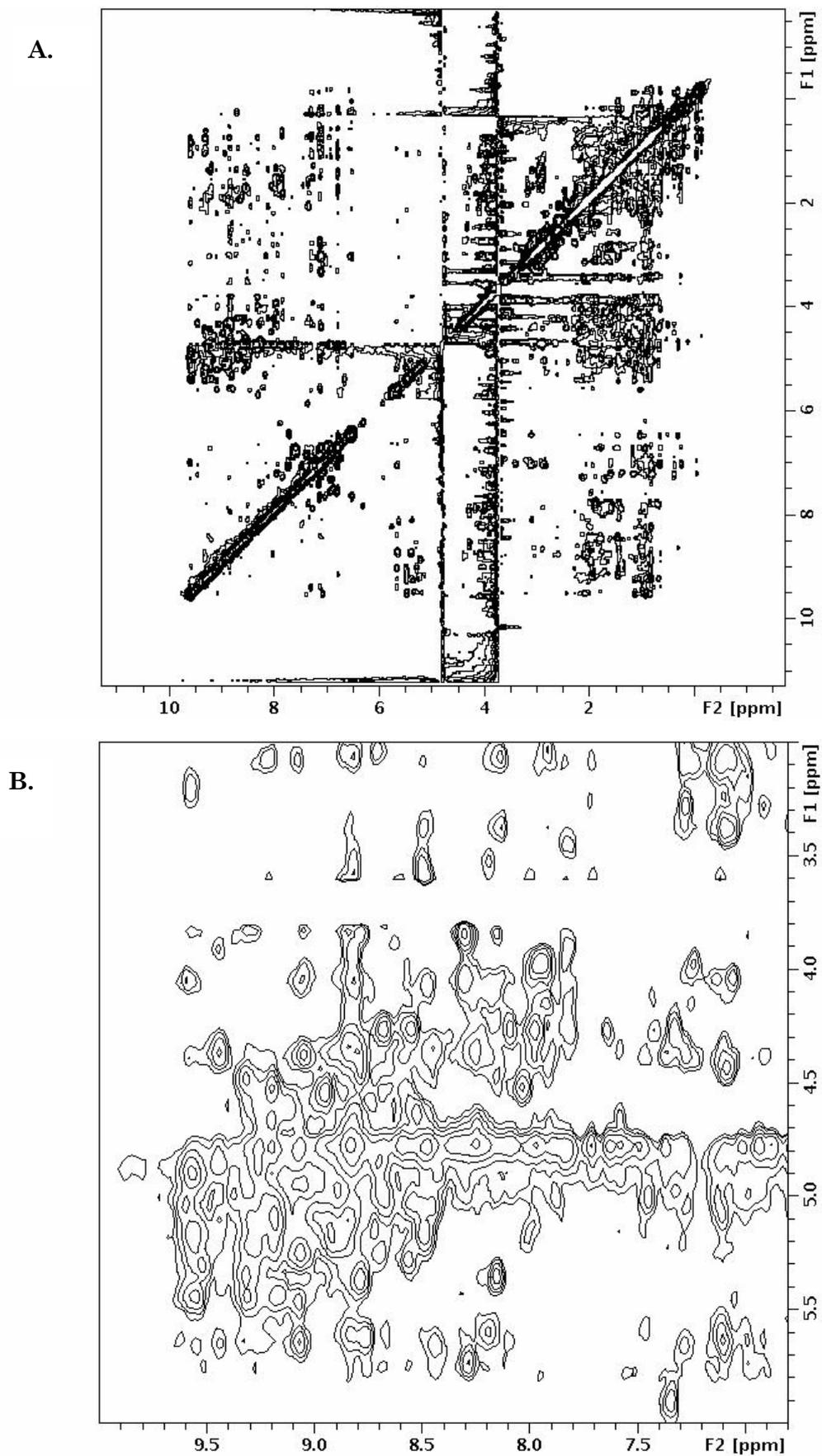


Figure 4.36. A. Fingerprint region of 2D 500 MHz  $^1\text{H}$ -NMR NOESY spectrum of 0.5 mM delipidated R120Q cL-BABP at pH 7.0, 30 mM PBS buffer (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), at 298 K. B. Detail of the spectrum.

## T91C

The 61,5% of BABPs belonging to the liver non-mammalian species show in position 91 a residue of cysteine that form a disulphide bridge with a cysteine in position 80 (Figure 4.37). It was decided to investigate whether the introduction of the S-S bridge in cL-BABP could influence its stability and the stochiometry of the binding.

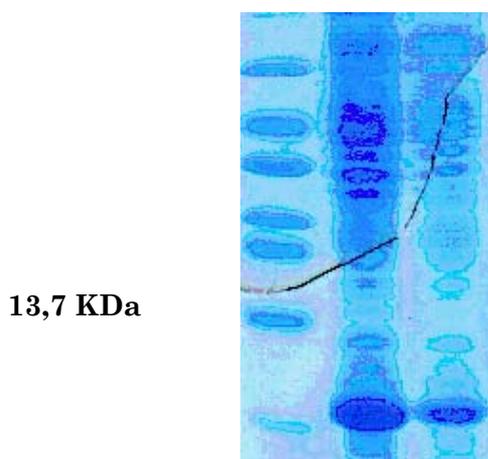
```

FABPL_CHICK      -QTVTNSFTLGKEADITTM DGKKLKCTVHLANGKLVTKSEKFSH-
Q8UWE2_ANAPL    -QSVTNSFTLGKEADITTM DGKKLKCTVNLVNGKLVCKSDKFSH-
FABPL_RHASA     -KSVTNSFTIGKEAEITTM DGRKLLKCI VKLEGGKLISET EKFSH-
FABPL_ANOPU     -KSVTNSFTLGKEADMTTM DGKKVKCTVNLVDGKLVAKSDKFIH-
FABP2_AMBME     -QSVTNSFTIGKEAEITSM GKKIKCTVVLEGGKLVSKTDQFSH-
Q9I8L5_BRARE    -KTVTNSFTIGKEAEITTM DGKKLKCI VKLDGGKLV CRTDRFSH-
Q7LZK9_LATJA    -KTVTNSFTIGKEADITTM DGKKIKCVVNLEGGKLV CNTGKFCH-
7T1A3_PLAFE     -KSVTNSFTVGKEADITTM DGKKIKCI VNMEGGKLV CNTGKFCH-
Q8JJ05_ACASC    -KTVTNSFTIGKEAEITTM DGKKLKCI VNMEGGKMV CKTGKFCH-
Q645P9_FUNHE    -KFVTKSFTIGKEAEITTM DGKKLKCI VNMEGGKLV CKTGKFSH-
FABPL_BUFAR     -QTHSNSFTVGKESEITSM DGKKIKVTVQLEGGKLI CKSDKFSH-
FABPL_HALBI     -QTVTNEFTVGKEAEITSM DGKKLKCTVQLEDGKLV AKKLFTH-
FABPL_LEPPA     -KSTTNTFTIGKEAEITTM NGNKLKCTINMEDGKLV CKTEKFSH-

```

**Figure 6.37.** Multiple sequence alignment performed with ClustalW among the Liver BABPs of non-mammalian species: Chick=chicken, Anapl= *Anas platyrhynchos*, Rhasa= *Rhamdia Sapo*, Anopu= *Anolis Pulchellis*; Ambe= *Ambistoma mexicanum*, Brare= *Branio rerio* Latja= *Lateolabrax japonicus* Plafe= *Platichthys flesus* Acasc= *Acanthopagrus schlegeli* Funhe= *Fundulus heteroclitus* Bufar= *Bufo arenarum* Halbi= *Halaelurus bivius* Leppa= *Lepidosiren paradoxa*.

The expression was performed at 24°C overnight and T91C cL-BABP was present in the soluble fraction; from the comparison of the total fraction and soluble fraction by SDS PAGE, it was observed that not all the expressed protein was soluble (Figure 4 .38).



**Figure 4.38.** Lane 1 : Marker ; lane 2: total fraction, lane 3: soluble fraction

### Purification procedure

The theoretical isoelectric point (pI) for T91C resulted 8.05; the pH of the buffer for anion exchange chromatography was changed to 7.3.

The final yield for 1 litre of LB/Kan was 70 mg/L. The protein purity was checked by the presence of a single band on SDS-PAGE and by mass spectrometry. The correct formation of the disulfide bridge was confirmed by mass spectrometry: the protein was partially digested with trypsin and it was detected the presence of the peptide fragment relative to the two peptides covalently linked by disulfide bridge (Figure 4.39, 4.40).

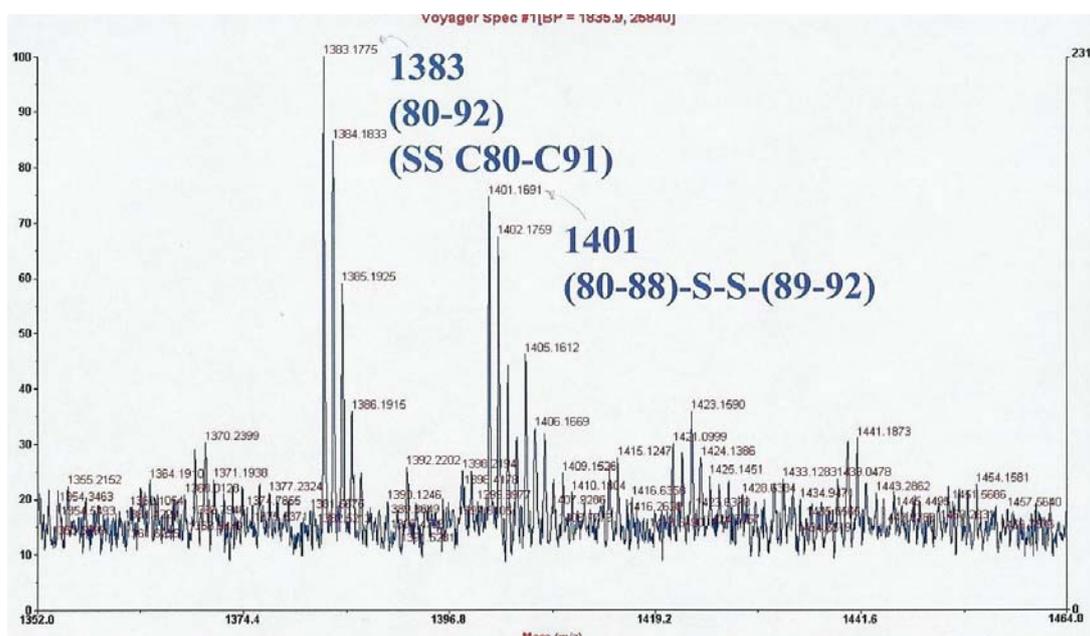


Figure 4.39 : MALDI TOF spectra of T91C cL-BABP.

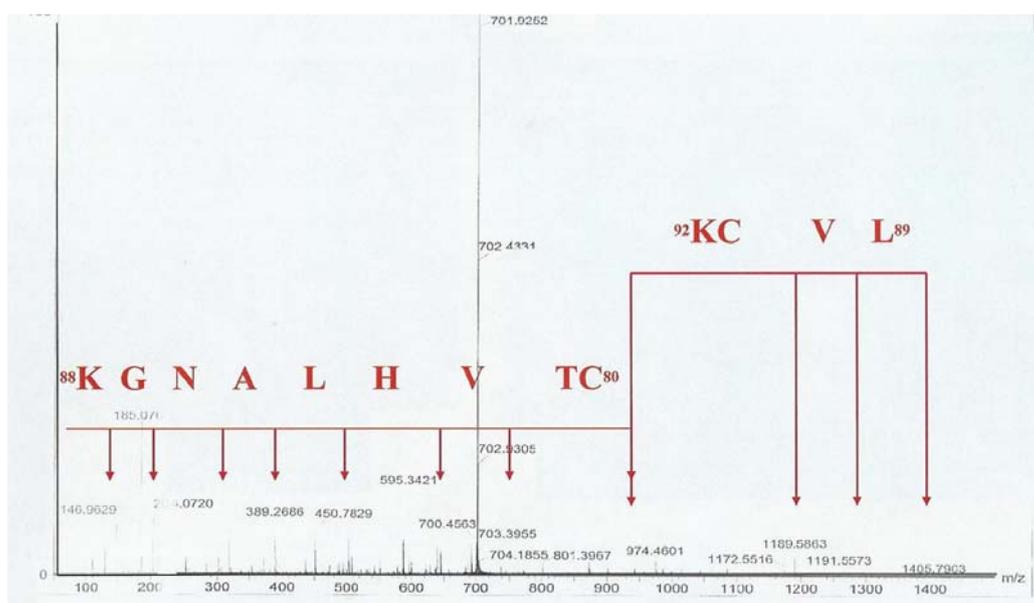
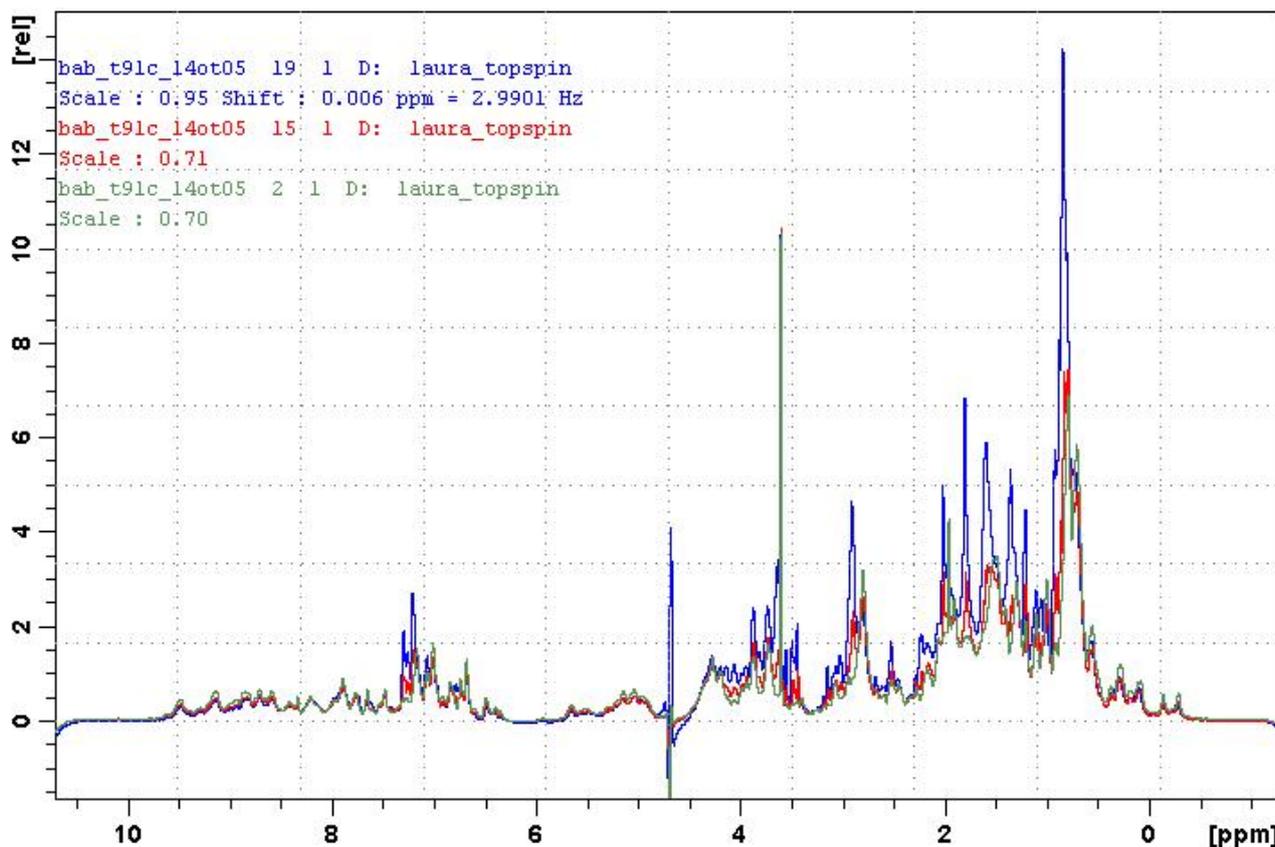


Figure 4.40 : Detailed analysis of the digested di-peptide linked by disulphide bridge.

## NMR experiments

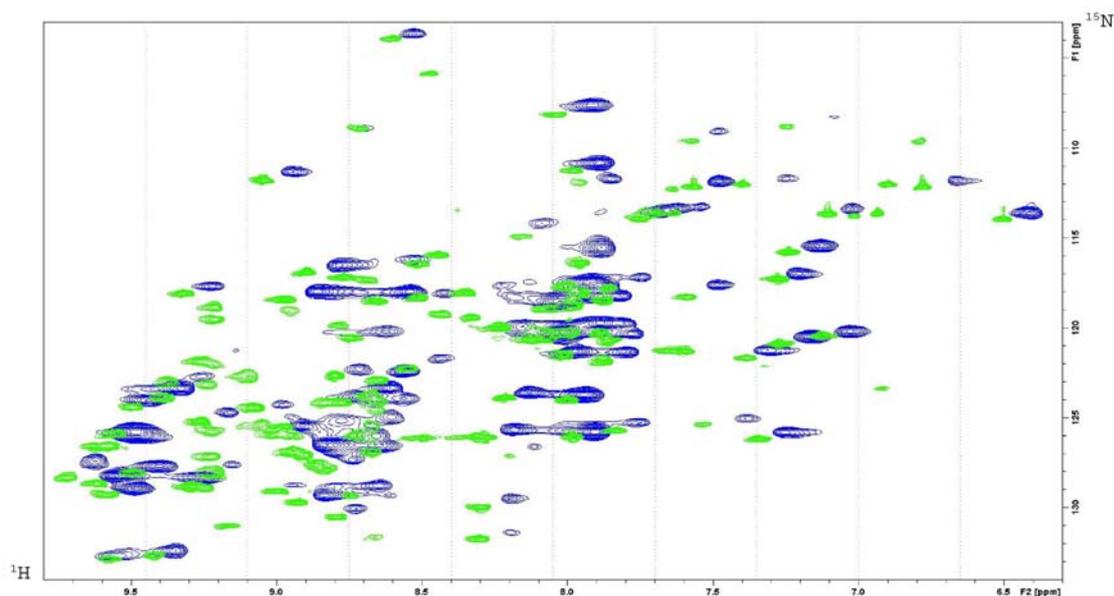
NMR analysis of the T91C cL-BABP mutant indicated that the protein was folded. The stability was observed during a week at room temperature with the acquisition of 1D spectra; the spectrum collected after immediate dissolution of the protein is shown in green in figure 4.41; the spectrum collected after four days is shown in red, after a week in blue. It was observed that just after four days the protein began to unfold, as evidenced by the disappearance of some signals in the high field region.



**Figure 4.41.** 1D 500 MHz <sup>1</sup>H-NMR spectra of 0.6 mM delipidated T91C cL-BABP at pH 7.0, 30 mM PBS buffer (95% H<sub>2</sub>O, 5% D<sub>2</sub>O), at 298 K, in green the spectra collected after immediate dissolution of the protein, in red after four days, in blue after a week.

As the protein exhibited partial aggregation at room temperature, it was decided to perform the delipidation at 25 °C. In order to check whether the delipidation was effective at this temperature, the same procedure reported in section 4.2.5 was followed. The same NMR experiment indicated that [1-<sup>13</sup>C] palmitic acid was correctly removed and confirmed the validity of the delipidation procedure at this temperature.

<sup>15</sup>N T91C cL-BABP was then produced for further NMR relaxation studies which analysis is actually in progress. The comparison of 2D-<sup>1</sup>H-<sup>15</sup>N HSQC spectra of the protein dissolved in buffers with different salt concentration is reported in figure 4.42.

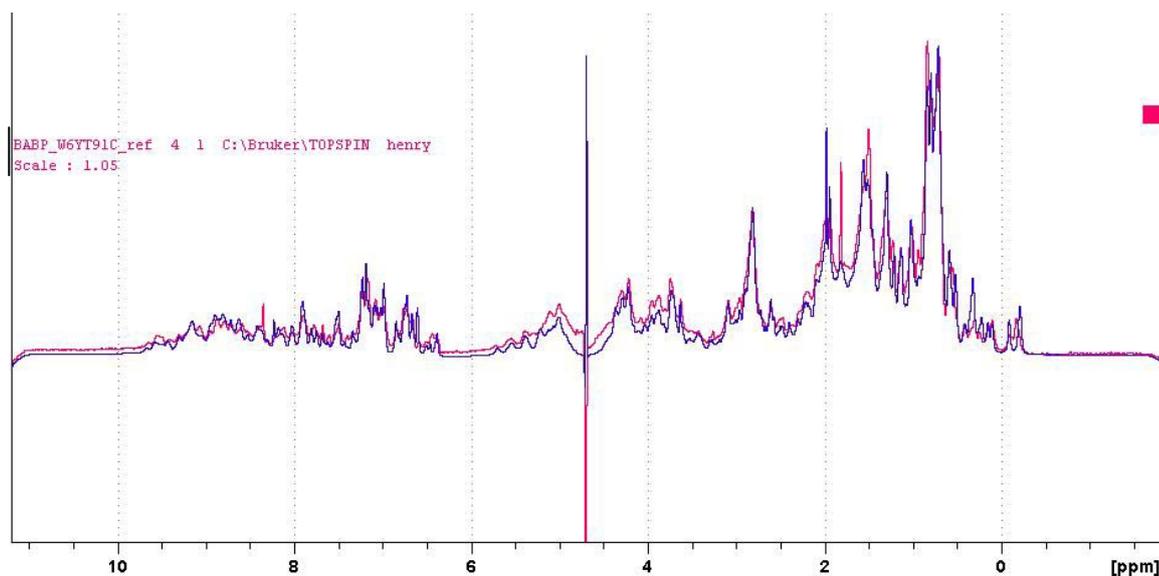


**Figure 4.42.**  $^1\text{H}$ - $^{15}\text{N}$  HSQC 0.5 mM delipidated  $^{15}\text{N}$  T91C cL-BABP , pH 7.0, 298 K, comparison of the the protein spectra in two different conditions of salts concentration, in blue buffer TRIS/HCl, 200mM, in green PBS 50mM.

From the spectra it is visible that high salts concentration increase the line width, an indication of an aggregated protein sample.

### Refolding of T91C cL-BABP

In order to rescue the *apo* T91C cL-BABP, instable in solution, the aggregated protein was subjected to refolding experiments (for buffers composition see Chapter 3): the protein concentration was firstly quantified by absorbance at 280 nm, the protein solution was then centrifuged at maximum speed at 4 °C for 30min and the supernatant was taken. The precipitated fraction containing the protein was washed 5-6 times with the Triton wash solution and dissolved in 8 M urea reaching a final protein concentration of 1 mg/mL (considering that all the protein was present in the precipitated fraction). The solution was kept at 4°C in low agitation for 15min, the insoluble material was removed by centrifugation and the unfolded protein was renatured by diluting 100-fold into the refolding buffer. The solution was kept at 4°C in low agitation overnight and the following day it was concentrated to about 50 mL using a YM-3 membrane. Aggregated material was removed by centrifugation (10000 g, 30min). The sample was then loaded on gel filtration chromatography using the same procedure as previously reported. About 20% of the protein resulted refolded. The comparison of the 1D spectra of the protein derived from the *E. coli* over-expressed soluble fraction (blue), with that derived from the refolding procedure (red) is reported in figure 4.43.



**Figure 4.43.** 1D 500 MHz  $^1\text{H}$ -NMR spectra of 0.6 mM delipidated T91C cL-BABP at pH 7.0, 30 mM PBS buffer (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), at 298 K, in blue the protein sample derived from the *E coli* over-expressed soluble fraction, in red the protein sample derived from the refolding procedure.

We could observe that the refolding procedure was successful.

## H98Y

Relaxation studies (Ragona *et al.* 2006) revealed the substantial conformational flexibility of cL-BABP mainly localised in the C-terminal face of the  $\beta$  barrel. The observed dynamics is suggested to be related to the protonation/deprotonation equilibrium of the buried histidine 98, located in this flexible face. NMR data suggest that the triggering event for the modulation of a functionally important motion, i.e. the opening/closure at the protein open end could be related to the protonation equilibrium of His 98. The mutation was done in order to clarify the role of this buried charged residues. The substitution was done with a tyrosine in order to maintain a similar steric hindrance.

The expression was performed in LB at 24°C overnight; H98Y cL-BABP was present in the soluble fraction but from the comparison of the total fraction and soluble fraction by SDS PAGE, it was observed that a small fraction of the expressed protein was soluble (Figure 4 .44).

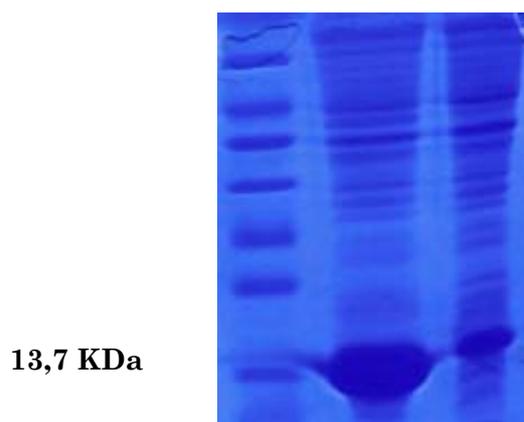


Figure 4.44. Lane 1: Marker; lane 2: total fraction, lane 3: soluble fraction

### Purification procedure

The theoretical isoelectric point (pI) for H98Y resulted 8.05.

The final yield for 1 litre of LB/Kan was of 24 mg/L, lower than wild type and the other mutants. Most of the protein gone to inclusion bodies, The protein purity was checked by the presence of a single band on SDS-PAGE and by mass spectrometry (Figure 4.45).

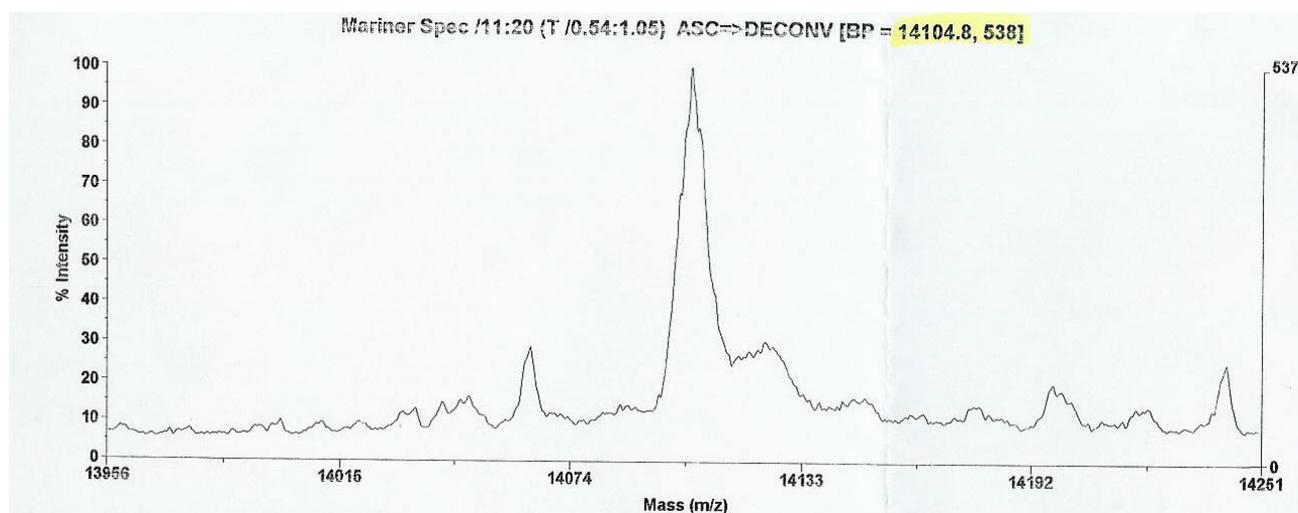


Figure 4.45 MALDI TOF spectrum of H98Y cL-BABP

Most of the protein gone to inclusion bodies; a refolding procedure was performed as described for T91C cL-BABP.

The analysis of the gel filtration chromatography elution indicated that the refolding solution contained aggregated protein with a wide range of molecular size.

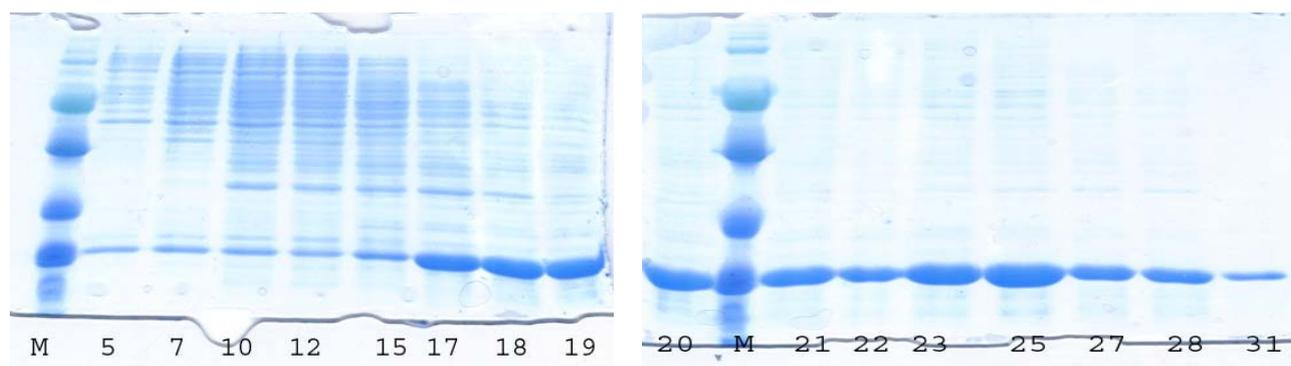


Figure 4.45 SDS PAGE of the fractions eluted from gel filtration chromatography; H98Y cL-BABP was eluted in a wide range of elution volumes

For this mutant, the refolding procedure was not successful and so other refolding experiments have to be done to improve the procedure.

### NMR experiments

The analysis of the 1D  $^1\text{H}$ -NMR spectrum of the delipidated H98Y cL-BABP is reported in figure 4.46, reporting the superposition of the 1D spectrum obtained after dissolving the protein and after one week. The increase of the resonance at 1 ppm is an indication of the presence of some mis-folded species.

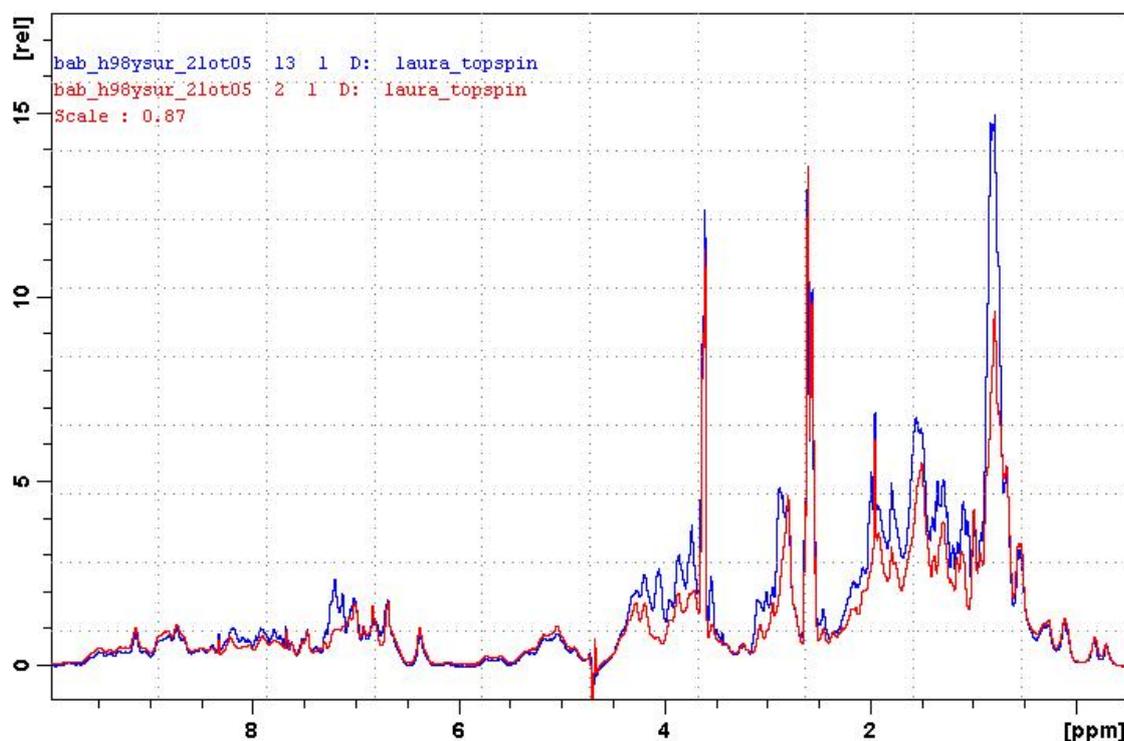
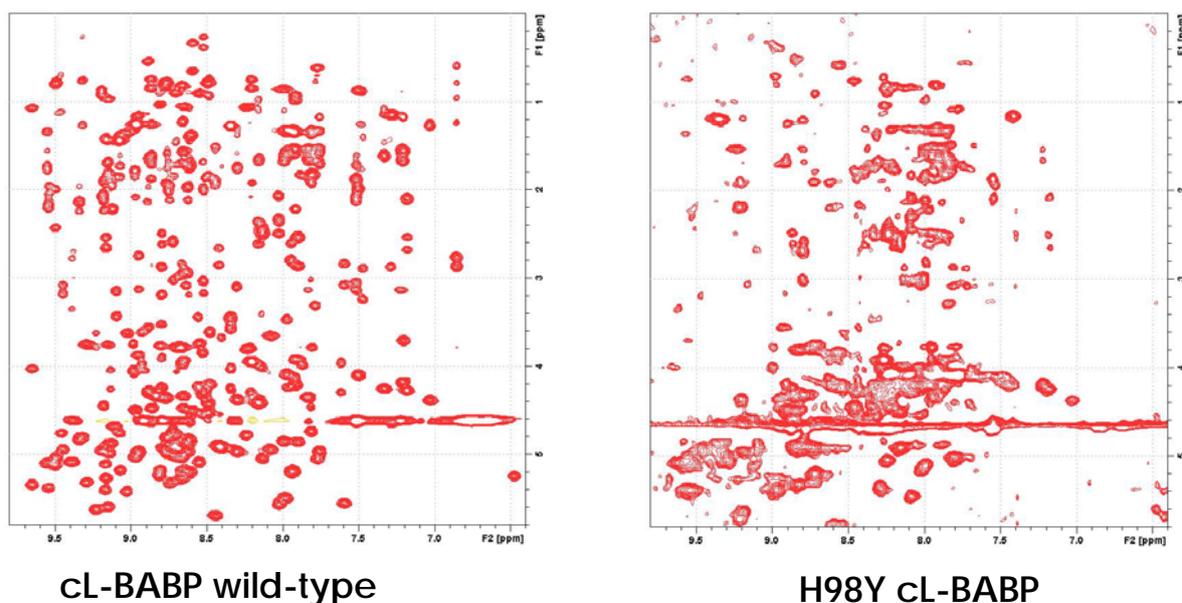


Figure 4.46. 1D 500 MHz  $^1\text{H}$ -NMR spectra of 0.6 mM delipidated H98Y cL-BABP at pH 7.0, 30 mM PBS buffer (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), at 298 K, in red the spectra collected after immediate dissolution of the protein, in blue after three days.

From the analysis of the 2D TOCSY spectra in figure 4.47 was visible the increase of the line width of the H98Y cL-BABP, if compared with the cL-BABP wild type spectra; the increase of line width is a strong indicator of the presence of mis-folded and/or aggregated species.

Our hypothesis is that the H98Y mutation may prevent closure of the  $\beta$ -barrel leaving it vulnerable to aggregation; this is supported by biological evidences, a large fraction of the expressed protein goes in inclusion bodies and by the increasing of line width of the mono and bi-dimensional NMR spectra.



**Figure 4.47.** Fingerprint regions of 2D 500 MHz  $^1\text{H}$ -NMR TOCSY spectra of 0.5 mM delipidated cL-BABP wt (left) and H98Y cL-BABP (right) at pH 7.0, 30 mM PBS buffer (95%  $\text{H}_2\text{O}$ , 5%  $\text{D}_2\text{O}$ ), at 298 K.



```

      10           20           30           40           50           60
MDIPQTKQDL  ELPKLAGTWH  SMAMATNNIS  LMATLKAPLR  VHITSLLPPT  EDNLEIVLHR

      70           80           90           100          110          120
WENNSCVEKKV LGEKTENPKK  FKINYTVANE  ATLLDTDYDN  FLFLCLQDTP  TPIQSMCCQY

      130          140          150          160
LARVLVEDDE  IMQGFIRAFR  PLPRHLWYLL  DLKQMEEPCRF

```

c. All the three constructs contained a fusion peptide with an His<sub>6</sub> tag in different positions:

- First: pRSETa plasmid with N-terminal His<sub>6</sub> tag presents 50 extra amino acids, 19 amino acids after cleavage with enterokinase.
- Second: pET22b plasmid with C-terminal His<sub>6</sub> tag presents 8 extra amino acids with no cleavage site.
- Third: pRSETa N-terminal His<sub>6</sub> tag presents 36 extra amino acids, 19 amino acids after cleavage with enterokinase.

d. The spacing between the SD and the start codon is in the range of optimal space (see section 4.1)

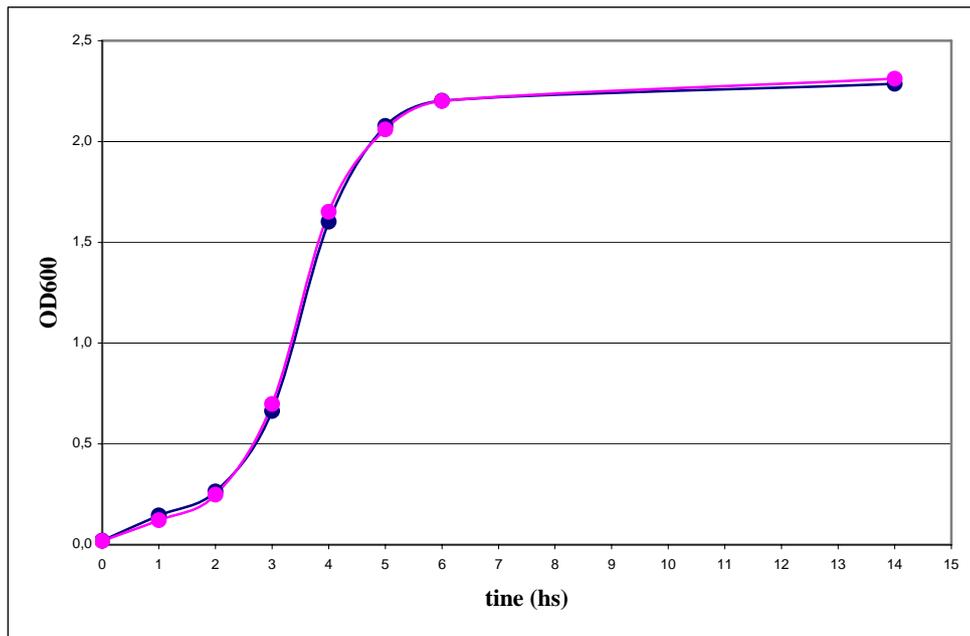
## 4.10. Development of expression and purification protocols of Glycodelin

Preliminary tests were done on the first construct, the first available, in order to decide the conditions for expression of consistent quantities of GdA:

- monitoring cell growth
- expression in the time course
- solubility test

### 4.10.1 Cell growth

The growth of the cells was monitored measuring the OD<sub>600</sub> as a function of time to establish the logarithmic and the stationary phase. A crucial point in protein expression is that induction must occur into the first half of the logarithmic phase (Figure 4.48).

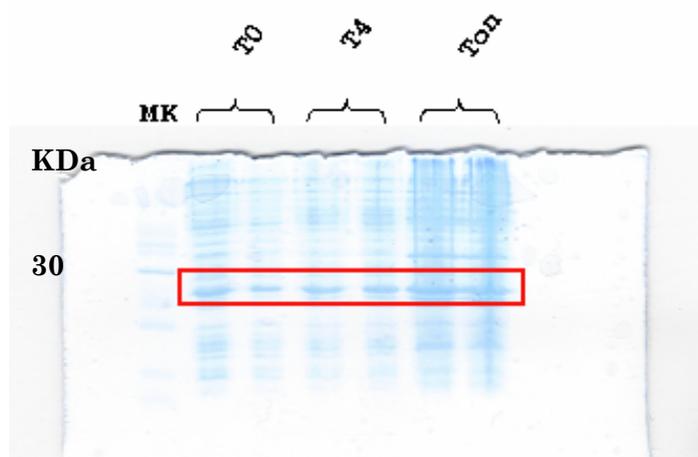


**Figure 4.48.** Growth of *E. coli* cells at 37°C. LB/kan medium was inoculated with a single bacterial colony from a fresh LB/Kan agar plate and incubated with shaking (180 rpm). In the figure is shown OD<sub>600</sub> as a function of time. Measurements on a 200 ml culture sample in a 1L flask are plotted in pink, measurements on a 700 ml in a 3 L flask culture sample are plotted in blue.

From the observation of the graph in figure 4.48 the identified OD<sub>600</sub> range for a correct induction was between 0.5 and 0.8.

#### 4.10.2 Expression Test –Time Course

Cell growth was monitored until the OD<sub>600</sub> of 0.7 was reached. At that time ( $t = 0$ ) the culture was induced with 1 mM IPTG (end concentration). The culture was grown overnight. 1 mL samples for electrophoresis were taken before induction ( $t = 0$ ) and after 4 hours and overnight induction ( $t_4$ ,  $t_{0n}$ ). Cells were centrifuged and the pellet was prepared for SDS PAGE (Figure 4.49).



**Figure 4.49.** Expression test for GdA by coomassie stained SDS-PAGE. Protein samples were taken after 0hs 4hs, and on induction. Lane 1: molecular weight marker; lane2 and 3: T0; lane 4 and 5: T4; lane 5 and 6: T0n

It was observed that protein expression increases in the time course till overnight induction. The overexpression rate, in comparison with that of cL-BABP (see Figure 4.4), is lower.

### 4.10.3 Solubility test

Expression and solubility tests were carried out as described in section 4.2.3. GdA was not detected in the soluble fraction, but was expressed as insoluble inclusion bodies (Figure 4.50). So it was decided to follow overnight induction at 37 °C with induction at OD<sub>600</sub> = 0.5-0.6 adding 1 mM IPTG.

Inclusion bodies are insoluble aggregates of mis-folded protein lacking biological activity, so, the refolding necessary to re-fold the protein from inclusion bodies for the full recovery of biological activity.

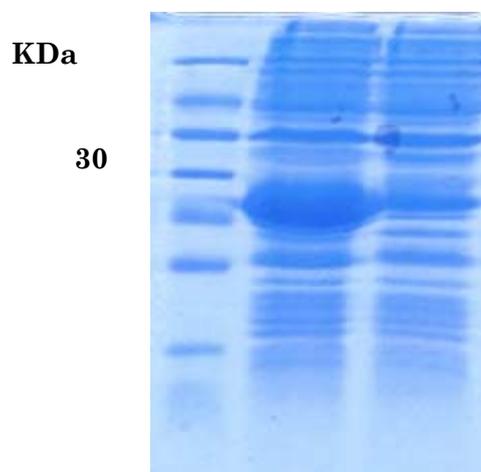


Figure 4.50. Lane 1: molecular weight marker; lane2: total fraction; lane 3: soluble fraction

The work proceeded with a large scale GdA over-expression in 1 litre of LB/Amp using the determined parameters. After overnight induction at 37°C cells were harvested by centrifugation, the supernatant was discarded and the pellet was re-suspended in lysis buffer for cell disruption. The solution was then centrifuged at 8000 g, the supernatant was discarded and the insoluble fraction was dissolved in 8 M Urea solution. Insoluble particles were precipitated by centrifugation.

### 4.10.4 Purification Procedure

The dissolved inclusion bodies were loaded on IMAC chromatography and purification was carried out as described in section 3.4.4. A bound fraction was eluted after linear imidazole gradient; GdA was detected by SDS PAGE analysis (Figure 4.51).



Figure 4.51. SDS PAGE showing the fractions (lane eluted after of imidazole gradient (20-400 mM))

The bound fraction contained a pool of proteins (Figure 4.51, lanes 4-6), so further and different purification step were necessary.

The inclusion bodies from another 1 litre LB preparation were dissolved in 8 M Urea and subjected to some steps of centrifugation at low speed (1000 g) before elution on gel filtration chromatography (Figure 4.52).

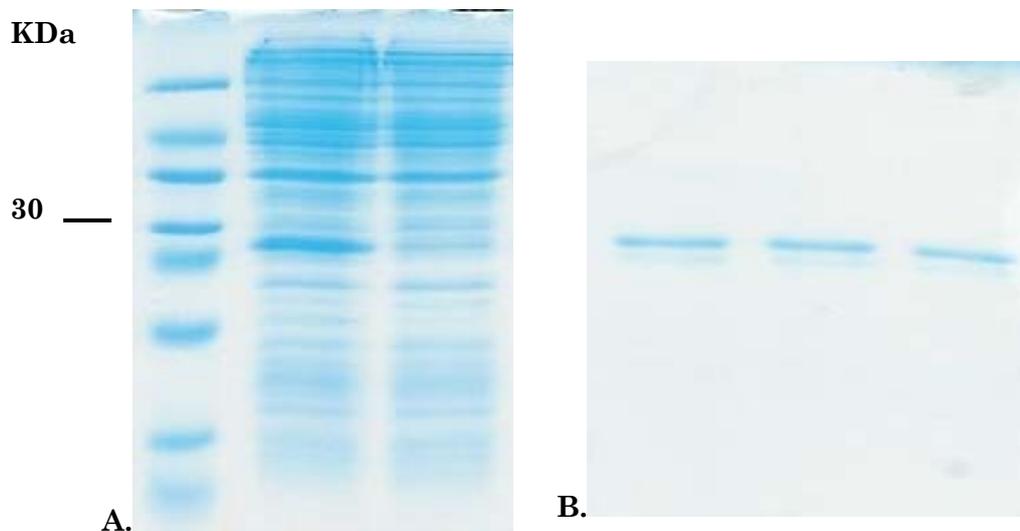


Figure 4.52. SDS PAGE A: lane 1: molecular weight marker, lane 2: 8 M Urea dissolved fraction, lane 3: supernatant after centrifugation at low speed. SDS PAGE B: lanes 1-3: fraction eluted from gel filtration chromatography.

From the analysis of SDS PAGE, a great part of *E. coli* endogenous proteins are removed after centrifugation at low speed. The sample was applied to gel filtration chromatography and GdA was detected in the fractions with elution volume corresponding to a 60 KDa molecule. When Glycodelin is studied by gel filtration is reported (Seppälä *et al.*, 2002) to behave as a homodimeric complex with a molecular mass of 50-60 KDa. The eluted fraction contained only two detectable proteins that migrated in SDS PAGE as molecules with similar size (about 25 000Da) (Figure 4.52 B). The purified fraction was so analysed by bi-

dimensional SDS PAGE-electrophoresis (Figure 4.53); the two bands migrated as molecules with similar isoelectric point (about 5.3).

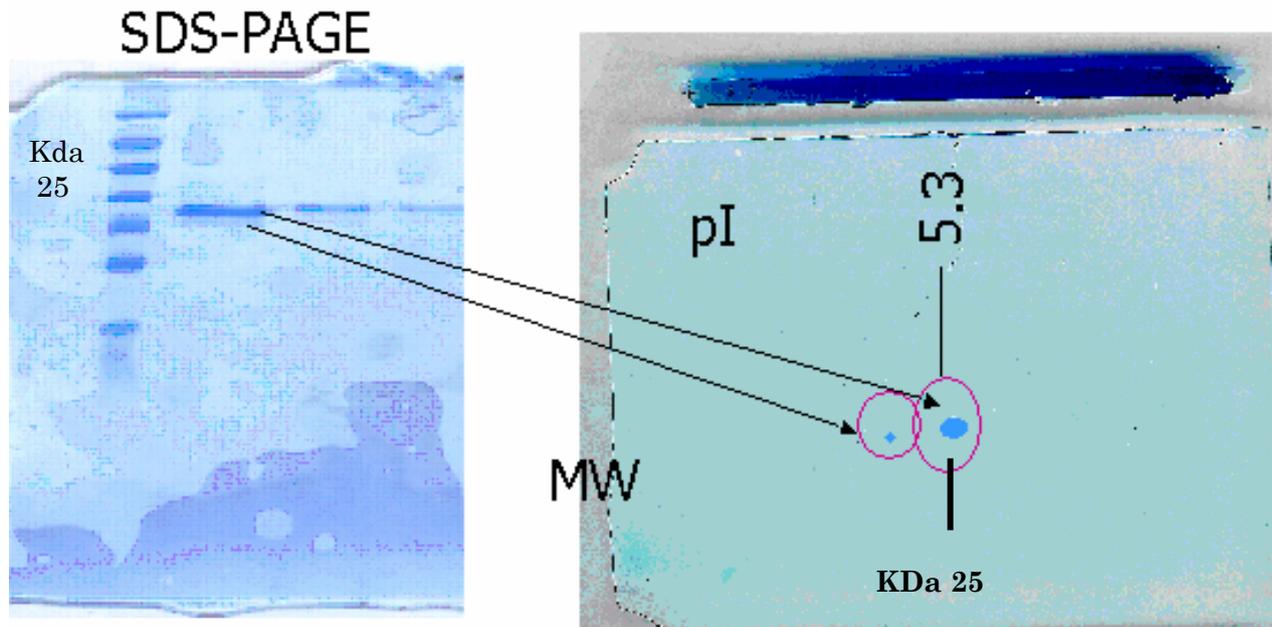


Figure 4.53. SDS PAGE and 2D SDS PAGE of the protein solution eluted from gel filtration chromatography .

Given the similar biochemical features of the two proteins, a further purification of GdA was not achieved.

The obtained sample was subjected to some preliminary characterizations.

#### *Molecular weight mass analysis*

The molecular mass of the protein sample was checked by mass spectrometry.

The analysis of the spectrum indicated the presence of two peaks corresponding to 24504,1 Da, resulted correct, and to 22902,1. The MALDI analysis of the proteins after tryptic digestion revealed that the specie with the lowest molecular mass corresponded to a truncated form of GdA. The truncation was detected at the residue 195. Our hypothesis is that the truncation was caused by the presence of an arginine codified by a rare codon for *E. coli*.

Further analysis and expression tests on the other expression vector will be done in order to obtain pure protein and to determine the state (native or non-native) of Glycodelin.

## 5. Conclusions

The work presented in this thesis concerns the set up of overexpression protocols for the production of pure recombinant cL-BABP wt and several mutants (T91C, H98Y, L21R, R120Q) both unlabelled and labelled with NMR active isotopes carbon-13 and/or nitrogen-15.  $^{13}\text{C}$  and  $^{15}\text{N}$  doubly enriched cL-BABP was employed for the three dimensional structure determination, deposited in the Protein Data B with accession code 1ZRY. Structural studies were then extended to the complex of cL-BABP with chenodeoxycholic acid and a detailed comparison of the dynamic properties of apo and holo proteins led to the identification of an allosteric mechanism, governing bile acid binding. For apo cL-BABP the dynamics analysis revealed the presence of a substantial conformational flexibility, mainly localized in the C-terminal face of the  $\beta$ -barrel, and primarily caused by the protonation/deprotonation equilibrium of the buried H98 located in this flexible C-terminal face. A key role was proposed for H98 in the modulation of the opening/closure of a loop at the protein open end. It is worth mentioning that H98 is highly conserved in the Ileal Lipid Binding Protein family (ILBP), and the comparative structural characterization of other proteins belonging to the ILBP family is in progress, in order to study the details of the conformational switch mechanism controlling the protein activity and function.

This work led to the publication in Journal of Biological Chemistry (2006) of the paper “NMR dynamic studies suggest that allosteric activation regulates ligand binding in chicken liver Bile Acid Binding Protein”. This paper is included in the present thesis.

The mutant H98Y cL-BABP was produced in order to elucidate the role of this histidine on protein conformational flexibility. This mutant was obtained as inclusion bodies and protocols for the correct refolding were optimized. The production of  $^{15}\text{N}$  and/or partially labeled  $^{15}\text{N}$ -His H98Y cL-BABP will be useful to perform further dynamics and interaction studies.

The following site-directed mutants: L21R, R120Q, and T91C were also designed, on the basis of previous NMR characterisation data, and their expression successfully achieved. Preliminary data suggested the presence of substantial conformational changes induced by the L21R mutation and therefore the production of  $^{15}\text{N}$  labeled protein is required, in order to perform *de novo* assignment of the spectra.

The comparative characterization and binding studies on all the produced mutants is in progress in order to evaluate the role of specific residues on binding.

As far as the production of Glycodelin is concerned, different expression protocols were tried, starting from three constructs, differing for the position and length of the His-tagged fusion protein. Preliminary data indicated that only the C-terminal His-tagged construct

lead to the achievement of a protein pure enough for NMR studies.  $^1\text{H}$  NMR experiments are in progress to find out whether the obtained protein is correctly folded.

## 6. References

- Åkerstrom, B., Flower, R. D. & Salier, J. (2000). Lipocalins: unity in diversity. *Biochim. Biophys. Acta.* **1482**, 1-8.
- Bakowies, D. & van Gunsteren, W.F. (2002). Simulations of Apo and Holo-Fatty Acid Binding Protein: Structure and Dynamics of Protein, Ligand and Internal Water, *J. Mol. Biol.* **315**, 713-736.
- Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli*. *Curr. Opin. Biotechnol.* **10**, 411-441.
- Bartels, C., Xia, T., Billeter, M., Güntert, P. & Wüthrich, K. (1995). The program XEASY for computer-supported NMR spectra analysis of biological macromolecules. *J. Biomol. NMR.* **5**, 1-10.
- Beringhelli, T., Goldoni, L., Capaldi, S., Bossi, A., Perduca, M. & Monaco, H. L. (2001). Interaction of chicken liver basic fatty acid binding protein with fatty acids: a  $^{13}\text{C}$  NMR and fluorescence study. *Biochemistry.* **40**, 12604-12611.
- Bianchet, M. A., Bains, G., Pelosi, P., Pevsner, J., Snyder, S. H., Monaco, H. L. & Amzel, L. M. (1996). The three-dimensional structure of bovine odorant binding protein and its mechanism of odor recognition. *Nature Struct. Biol.* **3**, 934-939.
- Boyd, J., Hommel, U. & Campbell, I. D. (1990). Influence of cross-correlation between dipolar and anisotropic chemical shift relaxation mechanism upon the longitudinal relaxation rates of  $^{15}\text{N}$  macromolecules. *Chem. Phys. Lett.* **175**, 477-482.
- Bremer, H. & Dennis, P. P. (1987). In *Escherichia coli and Salmonella typhimurium Cellular and molecular biology*. American Society of Microbiology. Neidhart, F. C., Ed., Washington DC.
- Brownlow, S., Morais Cabral, J. H., Cooper, R., Flower, D. R., Yewdall, S. J., Polikarpov, I., North, A. C. & Sawyer, L. (1997). Bovine beta-lactoglobulin at 1.8 Å resolution-still an enigmatic lipocalin. *Structure.* **5**, 481-95.
- Cai, M., Ying, H., Sakaguchi, K., Clore, G. M., Gronenborn, A. M. & Craigie R. (1998). An efficient and cost-effective isotope labeling protocol for proteins expressed in *Escherichia coli*. *J. Biomol. NMR.*, **11**, 97-102.
- Capaldi, A. P. & Radford, S. E. (1998). Kinetic studies of beta-sheet protein folding. *Curr. Opin. Struct. Biol.* **8**, 86-92.
- Carrel, R. W. & Lomas, D. A. (1995). Conformational disease. *Lancet.* **350**, 134-138.
- Ceciliani, F., Monaco, H. L., Ronche, S., Faotto, L. & Spadon, P. (1994). The primary structure of a basic (pI 9.0) fatty acid binding protein from liver of *Gallus domesticus*. *Trends Comparative Biochem Physiol.* **109B**, 261-271.
- Clore, G.M. & Gronenborn, A.M. (1994). Multidimensional heteronuclear magnetic resonance of proteins. *Methods Enzymol.* **239**, 349-363.

- Cordoba, O. L., Sanchez, E. I. & Santome, J. A. (1999). The main fatty acid-binding protein in the liver of the shark (*Halaetunus bivius*) belongs to the liver basic type. Isolation, amino acid sequence determination and characterization. *Eur. J. Biochem.* **265**, 832-8
- Dallessio, P. M. & Ropson, I. J. (2000).  $\beta$ -Sheet Proteins with nearly identical structures have different folding intermediates. *Biochemistry.* **39**, 860-871.
- Denovan-Wright, E. M., Pierce, M., Sharma, M. K. & Wright, J. M. (2000). cDNA sequence and tissue-specific expression of a basic liver-type fatty acid binding protein in adult zebrafish (*Danio rerio*). *Biochim. Biophys. Acta.* **1492**, 227-232.
- Derome, A.E. (1987). *Modern NMR techniques for chemistry research*. Pergamon, Oxford.
- Di Pietro, S. M., Dell'Angelica, E. C., Veerkamp, J. H., Sterin-Speziale, N. & Santome, J. A. (1997). Amino acid sequence, binding properties and evolutionary relationships of the basic liver fatty-acid-binding protein from the catfish *Rhamdia sapo*. *Eur. J. Biochem.* **249**, 510-517.
- Di Pietro, S. M., Veerkamp, J. H. & Santome, J. A. (1999). Isolation, amino acid sequence determination and binding properties of fatty-acid binding proteins from axolotl (*Ambistoma mexicanum*) liver. Evolutionary relationship. *Eur. J. Biochem.* **259**, 127-134.
- Di Pietro, S. M. & Santome, J. A. (2001). Structural and biochemical characterization of the lungfish (*Lepidosiren paradoxa*) liver basic fatty-acid binding protein. *Arch. Biochem. Biophys.* **388**, 81-90.
- Dobson, C. M. & Fersht, A. R. (1995). Protein folding. *Phil. Trans. R. Soc. Lond. B.* **348**, 1-119.
- Dobson, C. M. (1994). Solid evidence for molten globules. *Curr. Biol.* **4**, 936-940.
- Dobson, C. M. (2001). The structural basis of protein folding and its link with human disease. *Phil. Trans. R. Soc. Lond. B.* **356**, 133-145.
- Dyson, H.J. & Wright, P.E. (2001). Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol.* **339**, 258-270.
- Evans, J.N.S. (1995). *Biomolecular NMR spectroscopy*. Oxford University Press, Oxford.
- Farrow, N. A., Zhang, O., Szabo, A., Torchia, D. A. & Kay, L. E. (1995). Spectral density function mapping using  $^{15}\text{N}$  relaxation data exclusively. *J. Biomol. NMR.* **6**, 153-162.
- Fersht, A. R. (1997). Nucleation mechanism in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

- Flower, R. D. (1996). The lipocalin protein family: structure and function. *Biochem. J.* **318**, 1-14.
- Flower, R. D., North, A. C. T. & Sansom, C. E. (2000). The lipocalin protein family: structural and sequence overview. *Biochim. Biophys. Acta.* **1482**, 9-24.
- Fogolari, F., Ragona, L., Zetta, L., Romagnoli, S., De Kruif, K. G. & Molinari H. (1998). Monomeric bovine  $\beta$ -lactoglobulin adopts a  $\beta$ -barrel fold at pH 2. *FEBS Letters.* **436**, 149-154.
- Fogolari, F., Licciardi, S., Romagnoli, S., Ragona, L., Michelutti, R., Ugolini, R. & Molinari, H. (2000). Electrostatic properties of bovine beta-lactoglobulin. *Proteins Struct. Funct. Genet.* **39**, 317-330.
- Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C.A. & Goto, Y. (2000). Is folding of beta-lactoglobulin non-hierarchical? Intermediate with native-like beta-sheet and non-native alpha-helix, *J. Mol. Biol.* **296**, 1039-1051.
- Forman-Kay, J.D. (1999). The 'dynamics' in the thermodynamics of binding. *Nature Struc. Biol.* **6**, 1086-1087.
- Fujiwara, K., Arai, M., Shimizu, A., Ikeguchi, M., Kuwajima, K. & Sugai, S. (1999). Folding-unfolding equilibrium and kinetics of equine  $\beta$ -lactoglobulin: equivalence between the equilibrium molten globule state and a burst-phase folding intermediate. *Biochemistry.* **38**, 4455-4463.
- Glatz, J. F. C. & Veerkamp, J. H. (1983). A radiochemical procedure for the assay of fatty acid binding by proteins. *Anal. Biochem.* **132**, 89-95.
- Greene, L. H., Chrysina, E. D., Irons, L. I., Papageorgiou, A. C., Acharya, K. V. & Brew, K. (2001). Role of conserved residues in structure and stability: tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily. *Prot. Science.* **10**, 2301-2316.
- Greene, L. H., Grobler, J. A., Malinovskii, V. A., Tian, J., Ravi Acharya, K. & Brew, K. (1999). Stability, activity and flexibility in  $\beta$ -lactalbumin. *Prot. Eng.* **12**, 581-587.
- Harris, R.K. (1983). *Nuclear Magnetic Resonance Spectroscopy*. Pitman, London.
- Higgins, D., Thompson, J., Gibson, T., Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
- Hodson, M. E. & Cistola, D. P. (1997). Ligand binding alters the backbone mobility of intestinal fatty acid binding protein as monitored by  $^{15}\text{N}$  NMR relaxation and  $^1\text{H}$  exchange. *Biochemistry.* **36**, 2278-2290.
- Hodson, M. E. & Frieden, C. (2001). Intestinal fatty acid binding protein: the folding mechanism as determined by NMR studies. *Biochemistry.* **40**, 732-742.

- Honig, B. (1999). Protein folding: from the Levinthal paradox to structure prediction. *J. Mol. Biol.* **293**, 283-293.
- Hore, P.J. (1995). *Nuclear Magnetic Resonance*. Oxford University Press, Oxford.
- Hore, P.J., Jones, J.A. & Wimperis, S. (2000). *NMR: The Toolkit*. Oxford University Press, Oxford.
- Hwang, T. L. & Shaka, A. J. (1995). Water suppression that works. Excitation sculpting using arbitrary waveforms, and pulsed field gradients. *J. Magn. Reson. A* **112**, 275-279.
- Hwang, T. L. & Shaka, A. J. (1998). Multiple-pulse mixing sequences that selectively enhance chemical exchange or cross-relaxation peaks in high-resolution NMR spectra. *J. Magn. Reson.* **135**, 280-287.
- Ikeguchi, M., Kato, I., Shimizu, A. & Sugai, S. (1997). Molten globule state of  $\beta$ -lactoglobulin. *Proteins Struct. Funct. Genet.* **27**, 567-575.
- Kay, L. E., Torchia, D. A. & Bax, A. (1989). Backbone dynamics of proteins as studied by  $^{15}\text{N}$  inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry.* **28**, 8972-8979.
- Kay, L. E., Nicholson, L. K., Delaglio, F., Bax, A. & Torchia, D. A. (1992). Pulse sequence for removal of the effects of cross correlation between dipolar and chemical shift anisotropy relaxation mechanism on the measurements of heteronuclear  $T_1$  and  $T_2$  values in proteins. *J. Magn. Reson.* **97**, 359-375.
- Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M. & Schwalbe, H. (2002). Long-range interactions within a non-native protein. *Science.* **295**, 1719-1922.
- Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins Struct. Funct. Genet.* **6**, 87-103.
- Kuwajima, K. (1996). The molten globule state of  $\beta$ -lactalbumin. *FASEB.* **10**, 102-109.
- Kuwata, K., Shastry, R., Cheng, H., Hoshino, M., Batt, C. A., Goto, Y. & Roder, H. (2001). Structural and kinetic characterization of early folding events in beta-lactoglobulin. *Nature Struct. Biol.* **8**, 151-155.
- Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44-45.
- Lipari, G. & Szabo, A. (1982a). A model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. Theory and range of validity. *J. Am. Chem. Soc.* **104**, 4545-4549.
- Lipari, G. & Szabo, A. (1982b). A model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J. Am. Chem. Soc.* **104**, 4549-4570.

- Lücke, C., Zhang, F., Rüterjans, H., Hamilton, J. A. & Sacchettini, J. C. (1996). Flexibility is a likely determinant of binding specificity in the case of ileal lipid binding protein. *Structure*. **4**, 785-800.
- Marley, J., Lu, M. & Bracken, C. (2001). A method for isotope labeling of recombinant proteins. *J. Biomol. NMR*. **20**, 71-75.
- Messerle, B. A., Wider, G., Otting, G., Weber, C. & Wüthrich, K. (1989). Solvent suppression using a spin-lock in 2D and 3D NMR spectroscopy with H<sub>2</sub>O solutions *J. Magn. Reson.* **85**, 608-613.
- Mishkin, S., Stein, L., Gatmaitan, Z. & Arias, I. M. (1972). The binding of fatty acids to cytoplasmatic proteins: binding to Z-protein in liver and other tissues of the rat. *Biochem. Biophys. Res. Commun.* **47**, 997-1003.
- Molinari, H., Ragona, L., Varani, L., Consonni, R., Zetta, L. & Monaco, H. L. (1996). Partially folded structure of monomeric bovine  $\beta$ -Lactoglobulin. *FEBS Letters*. **381**, 237-243.
- Nikiforovich, G. V. & Frieden, C. (2002). The search for local native-like nucleation centers in the unfolded states of  $\beta$ -sheet proteins. *Proc. Natl. Acad. USA*. **99**, 10388-10393.
- Ockner, R. K., Manning, J. A., Poppenhausen, R. B. & Ho, W. K. L. (1972). A binding protein for fatty acids in cytosol of intestinal mucosa, liver, myocardium, and other tissues. *Science*. **177**, 56-58.
- Pace, C. N. & Sholtz, J. M. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131**, 266-280.
- Peng, J.W. & Wagner, G. (1994). Investigation of protein motions via relaxation measurements. *Methods Enzymol.* **239**, 563-596.
- Perez, M. D., Diaz de Villegas, C., Sanchez, L., Aranda, P., Ena, J. M. & Calvo M. (1989). Interaction of fatty acids with beta-lactoglobulin and albumin from ruminant milk. *J Biochem.***106**, 1094-1107.
- Perez, M. D., Puyol, P., Ena, J. M. & Calvo, M. (1993). Comparison of the ability to bind lipids of beta-lactoglobulin and serum albumin of milk from ruminant and non-ruminant species. *J. Dairy Res.* **60**, 55-63.
- Pervaiz, S. & Brew, K. (1985). Homology of beta-lactoglobulin, serum retinol-binding protein, and protein HC. *Science*.**228**, 335-337.
- Ptitsyn, O.B. (1973). Stage mechanism of the self-organization of protein molecules. *Dokl. Acad. Nauk.* **210**, 1213-1215.
- Qin, B.Y., Creamer, L. K., Baker, E. N. & Jameson, G. B. (1998). 12-bromododecanoid acid binds inside the calyx of bovine  $\beta$ -lactoglobulin. *FEBS Letters*. **438**, 272-278.
- Ragona, L., Pusterla, F., Zetta, L., Monaco, H. L. & Molinari, H. (1997). Identification of a conserved hydrophobic cluster in partially folded bovine  $\beta$ -Lactoglobulin at pH 2. *Folding and Design*. **2**, 281-290.

- Ragona, L., Confalonieri, L., Zetta, L., De Kruif, K. G., Mammi, S., Peggion, E., Longhi R. & Molinari, H. (1999a). Equilibrium unfolding CD studies of bovine  $\alpha$ -lactoglobulin and its 14-52 fragment at acidic pH. *Biopolymers*. **49**, 441-450.
- Ragona, L., Fogolari, F., Romagnoli, S., Zetta, L., Maubois, J. L. & Molinari, H. (1999b). Unfolding and refolding of bovine beta-lactoglobulin monitored by hydrogen exchange measurements. *J. Mol. Biol.* **293**, 953-969.
- Ragona, L., Zetta, L., Fogolari, F., Perez, D. M., Pujol, P., De Kruif, K., Lohr, F., Ruterjans, H. & Molinari, H. (2000). Bovine  $\alpha$ -lactoglobulin: interaction studies with palmitic acid. *Protein Sci.* **9**, 1347-1356.
- Ragona, L., Catalano, M., Zetta, L., Longhi, R., Fogolari, F. & Molinari, H. (2002). Peptide models of folding initiation sites of bovine  $\alpha$ -lactoglobulin: identification of native-like hydrophobic interactions involving G and H strands. *Biochemistry*. **41**, 2786-2796.
- Redfield, C., Shulman, B. A., Milhollen, M. A., Kim, P.S. & Dobson, C. M. (1999).  $\alpha$ -lactalbumin forms a compact molten globule in the absence of disulfide bonds. *Nature Struct. Biol.* **6**, 948-958.
- Roberts, G.C.K. (1993). *NMR of Macromolecules. A Practical Approach*. Oxford University Press, Oxford.
- Rolf, B., Oudenampsen-Kruger, E., Borchers, T., Faegeman, N. J., Knudsen J. & Spencer, F. (1995). Analysis of the ligand binding properties of recombinant bovine liver-type fatty acid binding protein. *Biochim. Biophys. Acta.* **1259**, 245-253.
- Romagnoli, S., Ugolini, R., Fogolari, F., Schaller, G., Urech, K., Giannattasio, M., Ragona, L. & Molinari H. (2000). NMR structural determination of viscotoxin A3 from *Viscum album* L. *Biochem. J.* **350**, 569-577.
- Ropson, I. J. & Frieden, C. (1992). Dynamic NMR spectral analysis and protein folding: identification of a highly populated folding intermediate of rat intestinal fatty acid binding protein by 19F NMR. *Proc. Natl. Acad. USA.* **89**, 7222-7226.
- Sacchettini, J. C., Meininger, T. A., Lowe, J. B., Gordon, J. I. & Banaszak, L. J. (1987). Crystallization of rat intestinal fatty acid binding protein. *J. Biol. Chem.* **262**, 5428-5430
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Sanger, F. & Tuppy, H. (1951a). The amino acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.* **49**, 463-480.
- Sanger, F. & Tuppy, H. (1951b). The amino acid sequence in the phenylalanyl chain of insulin. 1. The investigation of peptides from enzymic hydrolysates. *Biochem. J.* **49**, 481-490.

- Santomè, J. A., Di Pietro, S. M., Cavagnari, B. M., Cordoba, O. L. & Dell'Angelica, E. C. (1998). Fatty acid-binding proteins. Chronological description and discussion of hypotheses involving their molecular evolution. *Trends Comparative Biochem Physiol.* **4**, 23-38.
- Saunders, M., Wishnia, A. & Kirkwood, J.G. (1957). The nuclear magnetic resonance spectrum of ribonuclease. *J. Am. Chem. Soc.* **79**, 3289-3290.
- Scapin, G., Spadon, P., Pengo, L., Mammi, M., Zanotti, G. & Monaco, H. L. (1988). Chicken liver basic fatty-acid binding protein (pI=90). Purification, crystallization and preliminary X-ray data. *FEBS Letters.* **240**, 196-200.
- Schievano, E., Quarzago, D., Spadon, P., Monaco, H. L., Zanotti, G. & Peggion, E. (1994). Conformational and binding properties of chicken liver basic fatty acid binding protein in solution. *Biopolymers.* **34**, 879-887.
- Schievano, E., Mammi, S. & Peggion, E. (1999). Determination of the secondary structural elements of chicken liver fatty acid binding protein by two-dimensional homonuclear NMR. *Biopolymers.* **50**, 1-11
- Shulman, B. A., Kim, P., Dobson, C. M. and Redfield, C. (1997) A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nature Struct. Biol.* **4**, 630-634.
- Shulman, B. A., Redfield, C., Peng, Z.-y., Dobson, C. M. & Kim, P.S. (1995). Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human alpha-lactalbumin. *J. Mol. Biol.* **253**, 651-657.
- Stock, A. (1999). Relating dynamics to function. *Nature* **400**, 221-222.
- Storch, J. & Thumser, A. E. A. (2000). The fatty acid transport function of fatty acid-binding proteins. *Biochim. Biophys. Acta.* **1482**, 28-44.
- Tanford, C., Bunville, L. G. & Nozaki, Y. (1959). The reversible transformation of  $\beta$ -lactoglobulin at pH 7.5. *J. Am. Chem. Soc.* **81**, 4032-4035.
- Thompson, J., Winter, N., Terwey, D., Bratt, J. & Banaszak, L. (1997). The crystal structure of the liver fatty acid binding protein. A complex with two bound oleates. *J. Biol. Chem.* **272**, 7140-7150.
- Thompson, J., Reese-Wagoner, A. & Banaszak, L. (1999). Liver fatty acid binding protein: species variation and the accommodation of different ligands. *Biochim. Biophys. Acta.* **1441**, 117-130.
- Ugolini, R., Ragona, L., Silletti, E., Fogolari, F., Visshers, R. W., Alting, A. C. & Molinari, H. (2001). Dimerization, stability and electrostatic properties of porcine  $\beta$ -lactoglobulin. *Eur. J. Biochem.* **268**, 4477-4488.
- Vasile, F., Ragona, L., Catalano, M., Zetta, L., Perduca, M., Monaco, H. L. & Molinari, H. (2003). Solution structure of chicken liver basic type fatty acid binding protein. *J. Biomol. NMR*, in press.

- Wishart, D. S., Sykes, B. D. & Richards, F. M. (1991). Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **222**, 311-333.
- Wu, S.Y., Pérez, M. D., Puyol, P. & Sawyer, L. (1999).  $\alpha$ -lactoglobulin binds palmitate within its central cavity. *J. Biol. Chem.* **274**, 170-174.
- Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York.
- Yang, D. & Kay, L. E. (1996). Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: application to protein folding. *J. Mol. Biol.* **263**, 369-382.
- Yeh, S., Ropson, I. J. & Rousseau, D. L. (2001). Hierarchical folding of intestinal fatty acid binding protein. *Biochemistry*. **40**, 4205-4210.
- Zidek, L., Novotny, M. V. & Stone, M. J. (1999). Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nature Struct. Biol.* **6**, 1118-1121.

# APPENDIX A 1

## Acronyms, Abbreviations, and alternative names of Lipocalins

<i>Name</i>	<i>Common abbreviation</i>	<i>Alternative names and acronyms</i>
$\alpha_1$ -acid glycoprotein	AGP	orosomucoid (ORM), seromucoid $\alpha_1$ fraction, $\alpha_1$ -S
$\alpha_1$ -microglobulin	A1M	$\alpha_1$ -m, $\alpha_1$ m, protein HC, $\alpha_1$ -microglycoprotein
$\alpha_{2u}$ -globulin	$\alpha_{2u}$	rat $\alpha_2$ -urinary globulin, mouse major urinary protein (MUP)
Apolipoprotein D	apoD	gross cystic disease fluid protein (GCDFP-24), Apocrine secretion odor-binding protein (ASOB-2)
Bacterial lipocalin	Blc	
$\beta$ -lactoglobulin	$\beta$ lg	$\beta$ Lac, BLG
Complement factor 8 $\gamma$ chain	C8 $\gamma$	C8G
choroid plexus lipocalin	cp11	
Extracellular fatty acid binding protein	Ex-FABP	chondrocyte 21 protein (Ch21), P20K, quiescence specific protein (QSP)
Epididymal retinoic acid-binding protein	E-RABP	B/C protein; Epididymal binding protein 1 & 2 (EBP1/EBP2); epididymal secretory protein (ESP1); ERABP; mouse epididymal protein 10 (MEP 10)
Glycodelin	Glc	pregnancy protein 14 (PP14), human pregnancy-associated endometrial protein, $\alpha_2$ globulin ( $\alpha_2$ PEG), chorionic $\alpha_2$ -microglobulin, progesterone-associated endometrial protein (PAEP), $\alpha$ -uterine protein
Neutrophil gelatinase-associated lipocalin	NGAL	human neutrophil lipocalin (HNL), 24p3, SIP24, uterocalin, $\alpha_2$ -microglobulin-related protein, Neu-related lipocalin (NRL)
Odorant-binding protein	OBP	frog Bowman's gland protein
Probasin	PB	pM-40
Prostaglandin D synthase	PGDS	d-trace
Retinol-binding protein	RBP	plasma retinol binding protein, serum retinol binding protein (sRBP)

<b>Tear prealbumin</b>	<b>TP</b>	<b>protein migrating faster than albumin (PMFA); specific tear albumin (STP); tear lipocalin (TL); von Ebner's gland protein (VEG or VEGP); LCN1</b>
------------------------	-----------	--

**Allergens:**

<b>Cockroach allergen</b>	<b>Bla g4</b>	
<b>Mouse urine allergen</b>	<b>Mus m 1</b>	<b>MUP (see above)</b>
<b>Rat allergen</b>	<b>Rat n 1</b>	<b><math>\alpha_{2u}</math>-globulin (see above)</b>
<b>Dog dander allergen 1</b>	<b>Can f 1</b>	
<b>Dog dander allergen 2</b>	<b>Can f 2</b>	
<b>Horse allergen 1</b>	<b>Equ c 1</b>	
<b>Horse allergen 2</b>	<b>Equ c 2</b>	
<b>Major cow dander allergen</b>	<b>Bos d 2</b>	<b>bovine dander allergen (BDA20)</b>

## APPENDIX A 2

Symbol	3-letter	Meaning	Codons
A	Ala	Alanine	GCT, GCC, GCA, GCG
B	Asp, Asn	Aspartic, Asparagine	GAT, GAC, AAT, AAC
C	Cys	Cysteine	TGT, TGC
D	Asp	Aspartic	GAT, GAC
E	Glu	Glutamic	GAA, GAG
F	Phe	Phenylalanine	TTT, TTC
G	Gly	Glycine	GGT, GGC, <b>GGA, GGG</b>
H	His	Histidine	CAT, CAC
I	Ile	Isoleucine	ATT, ATC, ATA
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	TTG, TTA, CTT, CTC, <b>CTA</b> , CTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAT, AAC
P	Pro	Proline	CCT, <b>CCC</b> , CCA, CCG
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	CGT, CGC, <b>CGA</b> , CGG, <b>AGA, AGG</b>
S	Ser	Serine	TCT, TCC, TCA, TCG, AGT, AGC
T	Thr	Threonine	ACT, ACC, ACA, ACG
V	Val	Valine	GTT, GTC, GTA, GTG
W	Trp	Tryptophan	TGG
X	Xxx	Unknown	
Y	Tyr	Tyrosine	TAT, TAC
Z	Glu, Gln	Glutamic, Glutamine	GAA, GAG, CAA, CAG
*	End	Terminator	TAA, TAG, TGA