1. Scientific background

1.1.Introduction

Sex steroids play key roles in the regulation of fertility, reproduction and sexual behavior. They exert their actions at the nucleus *via* specific receptor proteins that function as transcriptional regulators (Mangelsdorf *et al*., 1995) and at the cell surface through a variety of poorly defined non-genomic mechanisms (Revelli *et al*., 1998). Sex steroid hormones are bound by several plasma proteins. Based on ligand specificity and structure, plasma steroid-binding proteins have generally been classified in two categories: corticosteroid-binding globulin (CBG), which binds glucocorticoids and progesterone, and sex hormone-binding globulin (SHBG), which binds estradiol, testosterone and dihydrotestosterone. Both types of globulins are present in most mammals that have been examined, although some mammals apparently have only one type of binding globulin (Wingfield *et al*., 1980; Gayrard *et al*., 1996). In birds, SHBGs have yet to be identified in any species, while CBG is present in all species that have been studied. In reptiles, amphibians, and fishes, some species have SHBG and CBG similar to those reported for mammals, others apparently lack one of two types of binding globulins, and some species have binding globulins whose specificity makes them difficult to classify as either SHBG or CBG (Wingfield *et al*., 1980; Callard and Pasmanik, 1987; Orchinik *et al*., 2000). Besides the highly specific binding of sex hormones by SHBG, albumin binds androgens with low affinity.

The exact role that these binding globulins play in regulation of steroid action is not well understood (Nakhla *et al*., 1990; Hammond, 1995). Binding globulins potentially regulate steroid actions at several levels: (1) rates of hormone metabolism, (2) hormone availability, according to the generally assumed free hormone hypothesis (Mendel, 1992) only steroids not bound to binding globulins leave the circulation and enter target tissues or cross the blood-brain barrier, (3) tissue specific binding to binding globulin receptors on the plasma membrane of target cells, and (4) tissue specific delivery of the steroid. These proposed functions are not mutually exclusive and binding globulins may regulate steroid actions through a variety of mechanisms. For example, plasma steroid-binding globulins may regulate steroid actions by altering the level of free steroid available to tissues, while at the same time

interactions among binding globulins and putative receptors may facilitate steroid delivery to specific target tissues (Damassa and Cates, 1995; Cates and Damassa, 1997; Rosner *et al*., 1999).

1.2. Sex hormone-binding globulin

1.2.1. Gene organization

The gene encoding human SHBG is located on the short arm (p12-13) of chromosome 17 (Berube *et al*., 1990) and comprises at least 8 exons that span approximately 3 kilobases (kb) of genomic DNA (Hammond *et al*., 1989). SHBG is produced as a 402-residue long polypeptide, which includes a typical hydrophobic leader sequence of 29 residues. Plasma SHBG is produced primarily in hepatocytes and a differently glycosylated isoform (Cheng *et al*., 1985), androgen binding protein (ABP), is produced in the Sertoli cells of the testis (French and Ritzen, 1973). The transcript encoding SHBG in hepatocytes and ABP in Sertoli cells is identical. The synthesis of SHBG is stimulated by thyroid hormone (Rosner *et al*., 1984) and estrogens and decreased by androgens (Anderson, 1974). The production of ABP by Sertoli cells is induced by testosterone and follicule-stimulating hormone (Means *et al*., 1980). Several other transcripts result from different exon utilization in sex-steroid-responsive tissues, but the nature of their protein products in these tissues remains to be clarified (Joseph, 1994; Hammond and Bocchinfuso, 1995).

1.2.2. Primary structure and characteristics of SHBG

Human SHBG (alternative name: sex steroid-binding protein, SBP) is a homodimeric plasma glycoprotein that binds both androgens and estradiol with nanomolar affinities (Siiteri *et al*., 1982; Westphal, 1986). The primary structure of human SHBG was first resolved by direct sequence analysis of the purified protein (Walsh *et al*., 1986) and confirmed by the isolation and characterization of a complementary DNA (cDNA) encoding the mature polypeptide (Hammond *et al*., 1987). Human SHBG is a homodimer, and each monomer comprises 373 amino acid residues with a molecular weight of 42 kDa (Walsh *et al*., 1986) and consists of a

tandem repeat of laminin G-like domains (LG-domains) (**Figures 1.1, 1.2**). Early studies on the purified proteins indicated that rat ABP and human SHBG were glycosylated (Danzo *et al*., 1991). It was demonstrated that purified human SHBG contained two biantennary N-linked oligosaccharide chains (residues 351 and 367) and one O-linked oligosaccharide moiety (residue 7) (Avvakumov *et al*., 1983; Danzo *et al*., 1989; Danzo and Black, 1990).
All ABP/SHBG sequences that have been characterized contain four conserved Cys residues at positions 164, 188, 333, and 361. These Cys residues have been shown to form disulfide bridges in the mature protein. The location of two disulfide bridges was established by peptide analysis of unreduced SHBG (Walsh *et al*., 1986). Cys164 and Cys188 form one bridge and Cys333 and Cys361 form another one.
Truncation experiments have shown that the steroid-binding site and main dimerization domain of human SHBG are localized within residues 1 to 205, which comprise the amino-terminal LG domain (Hildebrand *et al*., 1995).

1.2.3. Steroid binding

SHBG binds both types of sex steroid hormones androgens and estrogens (Siiteri *et al*., 1982; Westphal, 1986). Although there is considerable variation in steroid specificity between species, most bind DHT, T and E_2 with high affinity (DHT>T> E_2). Although the binding affinities vary between species, the K_d for DHT is generally between the 1-10 nM. Testosterone and estradiol generally bind with slightly lower affinities, but within one order of magnitude of the DHT affinity (Joseph, 1994). Prior to the crystallographic investigations reported here, several experiments have been done to identify the steroid binding region of SHBG. Various affinity labeling experiments and studies of recombinant SHBG mutants have indicated that a region encompassing and including Met139 in SHBG represents an important component of its steroid-binding site (Grenot *et al*., 1992; Hammond and Bocchinfuso, 1995). Residues Lys134 and Arg135 have been photo-affinity labeled with steroid ligands (Namkung *et al*., 1990) and have been shown to influence the relative binding affinities of different sex steroids across species (Bocchinfuso and Hammond, 1994; Danzo and Joseph, 1994).

```
  1    lrpvlptqsa hdppavhlsn gpgqepiavm tfdltkitkt
 41    sssfevrtwd pegvifygdt npkddwfmlg lrdgrpeiql
 81    hnhwaqltvg agprlddgrw hqvevkmegd svllevdgee
121    vlrlrqvsgp ltskrhpimr ialggllfpa snlrlplvpa
161    ldgclrrdsw ldkqaeisas aptslrscdv esnpgiflpp
201    gtqaefnlrd ipqphaepwa fsldlglkqa agsghllalg
241    tpenpswlsl hlqdqkvvls sgsgpgldlp lvlglplqlk
281    lsmsrvvlsq gskmkalalp plglapllnl wakpqgrlfl
321    galpgedsst sfclnglwaq gqrldvdqal nrsheiwths
361    cpqspgngtd ash
```

Figure 1.1 Primary structure of human SHBG. Each monomer consists of 373 amino acid residues. The steroid-binding site and main dimerization domain of human SHBG are localized within residues 1 to 205, which comprise the amino-terminal LG domain (green).

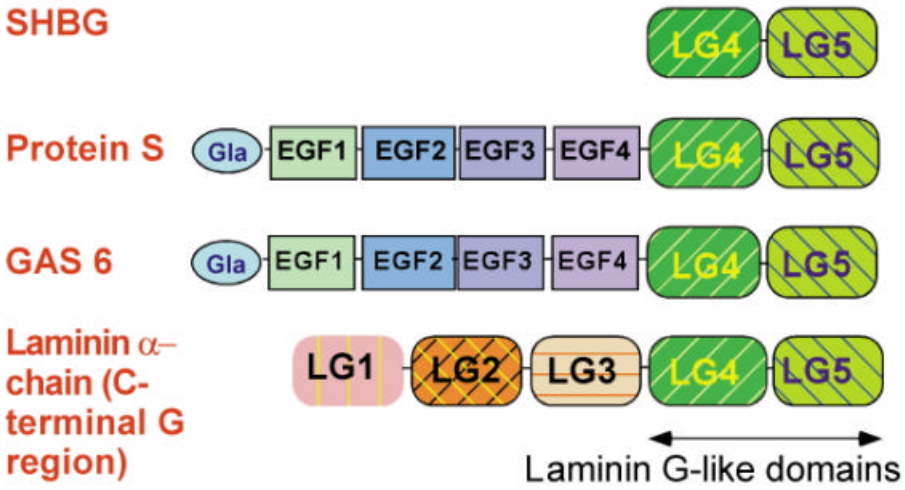


Figure 1.2 Modular architecture of SHBG and its homologues GAS6, protein S and laminin. SHBG consists of a tandem repeat of laminin G-like domains (LG-domains). GAS6 and protein S both contain an amino-terminal GLA-domain and four EGF-like domains. The carboxy-terminal region of laminin α-chain contains five LG-domains. Within each protein, the LG domains share little sequence identity, and the amino-terminal LG domain of SHBG is homologous to the amino-terminal LG domain of GAS6, protein S and the fourth LG domain of laminin α-chain.

The substitution of Ser42 with leucine causes a complete loss of steroid binding of intact human SHBG expressed in CHO cells (Hammond and Bocchinfuso, 1996).
Furthermore, based on the indication that only one steroid molecule is bound per dimer, a model has been proposed in which the steroid is sandwiched between the monomers at the dimerization interface (Petra, 1991).

1.2.4. Functions of ABP and SHBG

SHBG circulates in plasma as a major sex steroid carrier protein. It transports sex steroids in the blood and regulates their access to target cells (Siiteri *et al*., 1982; Hammond, 1993). ABP is secreted by Sertoli cells into the seminiferous tubule lumen and is transported to the epididymis with maturing sperm (French and Ritzen, 1973). ABP is thought to aid in the transport of testosterone to the epididymis and maintain a high androgen level required for epididymal function.
In several target tissues, such as uterine endometrium, MSF-7 breast cancer cells, prostate, male germ cells, testis and epididymis, receptor-binding activity of ABP/SHBG was described (Strel'chyonok *et al*., 1984; Hryb *et al*., 1989; Felden *et al*., 1992; Fortunati *et al*., 1993). Many binding experiments revealed high affinity and low affinity sites for SHBG. Other than their binding properties little is known about the receptor characteristics. Furthermore, the possible physiological role of SHBG has been extended by observations that it triggers cyclic AMP-dependent signaling through binding to specific cell surface receptors in prostate (Nakhla *et al*., 1994) and MCF-7 breast cancer cells (Fortunati *et al*., 1996). The activity of the SHBG receptor appears to be regulated by occupancy of the SHBG steroid-binding site (Rosner, 1990).
Occupation of the steroid binding site of SHBG by a steroid inhibits the ability of SHBG to bind to the membrane receptor. However, if unliganded SHBG is allowed to bind to its receptors on intact cells and the appropriate steroid hormone then is introduced, adenylate cyclase is activated and cAMP increases. This function is specific for steroids with biological activity, such as estradiol, DHT, testosterone. Methoxyestradiol is an endogenous metabolite of estradiol, which is more potent than estradiol in inhibiting receptor binding of SHBG, does

not cause an increase in intracellular AMP (Nakhla *et al*., 1990; Rosner *et al*., 1991). However, the identity of SHBG receptors at the molecular level remains obscure.

1.2.5. Laminin G-like domains

SHBG consists of two laminin G-like domains and, in addition to SHBG, these domains can be found in numerous extracellular proteins (**Figure 1.2**). In these proteins, this domain type binds a diverse range of ligands and is involved in a host of biological functions. LG domains were first described as an independent domain of about 200 residues, repeated five times in the large C-terminal region of laminin α chains (Beck *et al*., 1990). The LG domains of laminin bind to the dystroglycan receptor on muscle cells and anchor the actin cytoskeleton to the extracellular matrix (Hemler, 1999). They also mediate the attachment of *Mycobacterium leprae* to Schwann cells and thus they are implicated in the neural tropism of leprosy (Rambukkana *et al*., 1997). LG domains were later shown to share ~20-25% sequence similarity with domains found in neurexins and SHBG. Neurexins are expressed in hundreds of isoforms on the neuronal cell surface, where they may function as cell recognition molecules (Ushkaryov *et al*., 1992). α-Dystroglycan binding to LG domains was also observed in two other extracellular matrix proteins: the proteoglycans perlecan (Talts *et al*., 1999) and agrin (Bowe *et al*., 1994; Gee *et al*., 1994; Campanelli *et al*., 1996). Agrin plays a fundamental role in the formation of the neuromuscular junction by effecting the clustering of acetocholine receptors (Patthy and Nikolics, 1993). Heparan sulfate proteoglycans (perlecan) are components of all basement membranes. One of their functions is cell attachment, through interacting with integrin and nonintegrin receptors (Noonan *et al*., 1991). LG domains are found in the Drosophila developmental proteins, Slit and Crumb and the Drosophila tumor suppressor Fat (Mahoney *et al*., 1991; Joseph *et al*., 1997). This domain is present in the ligands of the Tyro-3 receptor protein-tyrosine kinase family, i.e. the growth arrest specific protein 6 (GAS6) and Protein S. Recently, GAS6 and Protein S have been implicated in sperm maturation because null mutations in the Tyro-3 family receptors result in defective spermatogenesis (Lu *et al*., 1999). Protein S acts in addition as an anticoagulant cofactor of Protein C and binds tightly to complement protein C4b in plasma (Villoutreix *et al*., 1997). In both GAS6 and Protein S, the LG domains mediate their biological function (He *et al*., 1997).

Thus, the functions associated with LG domains are diverse and include cell signaling, adhesion, migration and differentiation. But, insights into the functional significance of LG domains have been hampered by a lack of structural data.

1.3. The aim of my thesis

The steroid-binding properties of SHBG were extensively studied by numerous biochemical methods. However, to understand these processes in details, i.e. to understand how steroid transport occurs, the knowledge of the three-dimensional structure is essential. The aim of my thesis was to express and purify the protein, produce well diffracting crystals and to solve the crystal structure. For this purpose the protein was overproduced in *E. coli*, purified and successfully crystallized. Then heavy-atom derivatives were obtained and the crystal structure of the N-terminal LG domain of SHBG in complex with the natural ligand 5α-dihydrotestosterone (DHT) was been solved by the multiple isomorphous replacement method.

Subsequently, the crystal structures of SHBG in complex with different natural and synthetic steroids were solved and the steroid-binding pocket of SHBG was studied in more details.

This work forms a basis for further research of SHBG and other LG-domains containing proteins.

In my thesis work I set out to answer the following questions:

- How does the LG-domain fold look like?
- How does steroid transport by SHBG occur?
- How does dimerization of SHBG occur?
- What are the determinants of the steroid-binding specificity of SHBG?
- Zinc appears to modulate steroid-binding affinity. Does the structure help to explain this effect?
- Will the steroid-binding mode in SHBG be similar to binding modes in other steroid-binding proteins?

1.4. Short theory of the isomorphous replacement and molecular replacement methods

Among other crystallographic methods used in this thesis the method of isomorphous replacement and the method of molecular replacement are more specialized and, therefore, they are selected to be shortly introduced here. This short overview is based on the description of these methods given in the textbooks written by Drenth (1999) and Blundell & Johnson (1976).

1.4.1. Phase problem

In case when the structure factors are known, the electron density (ρ(x, y, z)) at the position (x, y, z) in the unit cell can be calculated using the following equation:

$$r(x,y,z) = 1/V \sum_h \sum_k \sum_l F_{hkl} e^{-2pi(hx+ky+lz)}$$

Structure factors are described by their amplitude and phase.

$$F_{hkl} = \left|F_{hkl}\right| e^{ia_{hkl}}$$

F_{hkl}: Structure factor

$\left|F_{hkl}\right|$: Structure factor amplitude of reflection (*h k l*)

a_{hkl} : phase angle

In this equation, the amplitudes can be measured from the diffraction pattern. The phases α, however, cannot be obtained directly from the diffraction pattern and have to be deduced indirectly. This observation is called the phase problem in protein crystallography. Four major methods exist to solve the phase problem in protein X-ray crystallography.

Direct methods in which mathematical calculations based on probability theory can be used to provide phase information. Although these methods work perfectly well for small molecule crystals, it is not easy to extend them successfully to protein crystals.

The molecular replacement method is used when a very similar crystal structure is available. Initial phasing is based on phase approximation taken from the data of the comparable structure.

The isomorphous replacement method in which heavy atoms are introduced into a light atom structure and are used as marker atoms to provide phase information.

The multiple wavelengths anomalous dispersion method in which phase information can be obtained from the information contained in the scattering by an atom whose natural absorption frequency is close to the wavelength of the incident radiation.

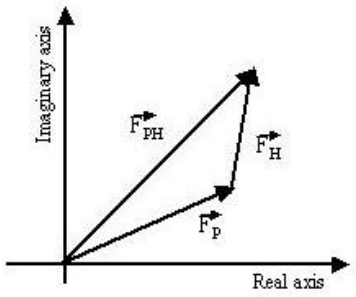
1.4.2. Principles of the multiple isomorphous replacement method

The method of isomorphous replacement is widely used for the X-ray analysis of protein crystals. The method involves five stages.

1) The preparation of heavy atom derivatives. The most widely used procedure includes the soaking of protein crystals in a solution containing the selected heavy-atom compound.

2) The measurement of the intensities of the X-ray diffraction patterns of the native and derivative crystals in order to find the heavy atom contribution to the structure factor of the derivative crystal. The structure factors, $\vec{F}_P$, $\vec{F}_{PH}$ and $\vec{F}_H$ of the protein, derivative and heavy atoms, respectively, are related by $\vec{F}_{PH} = \vec{F}_P + \vec{F}_H$

The relationship is illustrated in the following vector diagram:



However, the phases of $\vec{F}_P$, $\vec{F}_{PH}$, $\vec{F}_H$ can not be measured directly.

3) The successful determination of the heavy atom positions allows for the calculation of the structure factors of the heavy atoms.

$$F_H = \sum_{i=1}^{n} f_i e^{2pi(hx_i+ky_i+lz_i)}$$

where the scattering factor f_i of atom i includes the effect of thermal motion, and n is the number of heavy atoms in the unit cell.

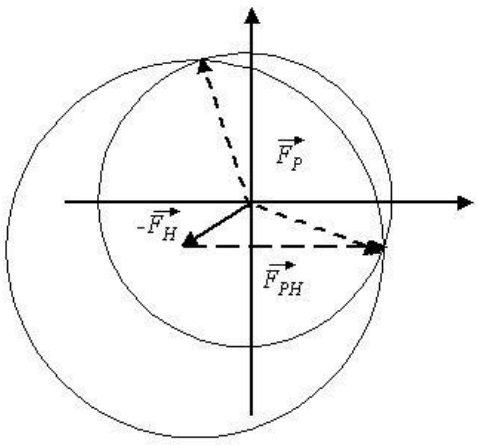
In order to find the coordinates of the heavy atoms, the Patterson function $P(u,v,w)$ should be calculated to give a Patterson map of the heavy atom arrangement in the unit cell. The peak height of a peak of any atom in this map is proportional to the product of the atomic numbers of the atoms responsible for the peak

$$P(u,v,w) = \frac{1}{V}(\Delta|F|_{iso})^2 \cos[2p(hu+kv+lw)]$$

where u, v, w are the relative coordinates in the unit cell, V is the volume of the unit cell. The coefficient $(\Delta|F|_{iso})^2$ is defined by:

$$\Delta|F|_{iso} = |F_{PH}| - |F_P|$$

4) The protein phases are determined using the Harker construction.



The circles intersect at two different points yielding two possibilities for the phases of $\vec{F}_P$. To distinguish between these two possible solutions, a second heavy atom derivative must be generated, which then also gives two possible phase angles. Only one of those will have the same value as one of the two previous phase angles; it therefore represents the correct solution. In practice, more than two different heavy-metal complexes are often needed to give a reasonably good phase determination for all reflections.

1.4.3. The molecular replacement method

The method assumes that a model structure very similar to the unknown structure is available. The technique has been extensively described in the literature (Rossmann, 1962; Crowther, 1967; Rossmann, 1972). Three main steps can be distinguished: finding the orientation of the search model, locating its position; assessing the quality of the solution. The first and second steps are carried out by Patterson search techniques: the position of the model is varied until its calculated Patterson function matches the observed one, corresponding to the unknown structure. The orientation is usually obtained with the rotation function and, given the orientation the displacement is determined using the translation function (Drenth, 1999).

1.4.3.1. The rotation function

The basic principle of the molecular replacement method can be understood by regarding the Patterson function of a protein crystal structure. The Patterson map is a vector map: vectors between atoms in the real structure show up as vectors from origin to maximum in the Patterson map. There are two groups of vectors intermolecular or self- Patterson vectors and intermolecular vectors or cross-Patterson vectors. The self- Patterson peaks all lie in a volume around the origin with a radius equal to the dimension of the molecule. For homologous molecules the self- Patterson vector distribution is very similar. Therefore, if their Patterson functions are superimposed on a correctly rotated version, maximum overlap between the two Patterson maps will occur.

Applying the rotation to the Patterson function P(u) gives the rotated Patterson function $P_r(u_r)$. Rotation function R is an overlap function of P(u) with the rotated version, $P_r(u_r)$ and is defined as

$$R(\alpha,\beta,\gamma) = \int_U P(u) \times P_r(u_r) du$$

α, β, γ are rotation angles,

u are coordinates of the peak in a Patterson map,

U is the volume in which the self-rotation peaks are located.

In the molecular replacement method the two different Patterson maps of the diffraction data and the known structure, which is used as the search model, must be superimposed by the rotation of the model map. To find the right rotation solution means to find overlap between the Patterson function of the diffraction data and the rotated Patterson function of the search model. The product rotation function R depends on the rotation angles and will have a maximum value for correct overlap. Finding right rotation solution depends on a number of parameters like resolution range, shape and size of the region U.

1.4.3.2. The translation function

For the final solution of the molecular replacement method the translation required to overlap one molecule onto the other in real space must be determined, after it has been oriented in the correct way with the rotation function.

The most promising solution from the rotation search are picked up and used in the translation search. The translation search deals with the cross-Patterson vectors as well as space group symmetry. Cross- Patterson vectors between the two molecules can be calculated as

$$P_{1,2}(u)0 = \int_V \rho_1(x) \times \rho_2(x+u) dx$$

The translation function is calculated comparing the cross-Patterson function of the model structure with the observed Patterson function *P*(u) of the diffraction data.

$$T(t) = \int_V P_{1,2}(u,t) \times P(u) du$$

t is the intermolecular vector between local origins of two molecules.
The translation function *T*(t) will have a maximum when the Patterson function of the model structure fits correctly on the observed Patterson function.