2 METHODS

2.1 Introduction into Molecular Modeling

Based on the idea of evolution, all organisms stem from a common ancestry. This further implies that the structures of biological molecules in current organisms also share a common origin. Indeed, it was demonstrated in several cases that proteins with similar sequence share a common fold and quite often also similar, if not the same, functions. Using this information is the basis of homology and comparative modeling of proteins (Srinivasan *et al.*, 1996). Homology modeling is based on homology: a similarity between proteins that have a common ancestor. In the area of homology modeling, proteins that share at least 30-35% identity were considered as homologous and are suitable targets for this procedure. Proteins with less sequence identity were not considered as homologous and are, because of this, targets for comparative modeling approaches.

First studies of homology modeling for model generation (Jones and Thirup, 1986) used the backbone of the protein and exchanged the side chains of the template structure by those of the target sequence. In cases of highly homologous proteins this procedure is still used today, because of its high performance and quite good accuracy in explanation of experimental results (Srinivasan *et al.*, 1996).

2.2 Sequence Alignments

The first step in homology modeling is the search for a homologous protein structure as a template for the model. Sequence alignment investigations are probably the most common tool. On the one hand they allow the search for new protein structures based on sequence similarity to the target sequence, as for example the local alignment search algorithms of BLAST (Altschul *et al.*, 1990) and FastA (Henikoff and Henikoff, 1992; Pearson and Lipman, 1988) do. On the other hand, comparisons of several homologous protein sequences allow conclusions about highly conserved, homologous regions and areas of rather insignificant functional residues (even if no structural data is available).

In search for homologous proteins with known structural data, usually pairwise sequence alignments are used. The local alignment search-algorithms of BLAST and FastA are the most common used. Local alignment searches are trained to find the best fitting regions between different sequences of nucleic acids or proteins. BLAST, for example, offers comparisons between proteins and protein databases (*e.g.* the Protein Data Bank, PDB) using BLASTP, between nucleic acids and nucleic acid databases using BLASTN and furthermore also alignments between nucleic acids and protein databases. Therefore all 6 reading frames of a nucleic acid (3 for each single strand of DNA) are translated into protein sequences and thereafter compared against the protein database. FastA is usually regarded to be more suitable for comparisons between nucleic acids than BLAST and in comparisons of protein families. But in general the algorithm and reliability of both methods are comparable.

The quality of pairwise sequence comparisons is represented by an alignment score S_{align} – the sum of all n scores S_{local} of local alignments between the sequences.

$$S_{align} = \Sigma \left(S^1_{local} + S^2_{local} + \ldots + S^n_{local} \right) \qquad \text{(Equation 1)}$$

Besides pairs of identical residues, mismatches also occur, which are valued by different scoring matrices (*e.g.* PAM, BLOSUM, PSSM). The blocks substitution matrix BLOSUM62, which is probably the most common scoring matrix, was generated by a set of training sequences of proteins that are not more than 62% identical to each other. As a result, to every substitution event (residue exchange) recognized in this training set a number was assigned, which represents the substitution frequency of this event. Hence, BLOSUM62 is a knowledge-based scoring matrix. In general, the scores are highest for identical residues, lower for mismatches of similar properties and lowest for mismatches with far related properties. Nevertheless, some substitution events were never recognized and for this reason

cannot be scored. In such cases a gap is introduced into the alignment. The existence of gaps in an alignment is worse than residues that only partially match, and so the alignment algorithm penalizes gaps. Because gap penalties are not defined as, for example, substitution events are, the size of gap penalties allows refinement of sequence alignments.

Taken together, the calculation of local alignment scores S_{local} is summarized in equation 2.

$$S_{local} = \Sigma (identities + mismatches) - \Sigma (gap\ penalties) \qquad \text{(Equation 2)}$$

Using local alignment search tools in regard to the Protein Data Bank (Berman *et al.*, 2000), finding protein structures as suitable templates with adequate sequence similarity to the target is applied. But very often lots of mismatches or gaps make it difficult to select the best template for modeling, if a template is generally available.

However, if no structural information is found, there are at least two different methods, which may result in a suitable template structure. One method is the use of Profile Hidden Markov Models (Eddy, 1998), which turns a multiple sequence alignment into a position-specific scoring system suitable for searching databases for remotely homologous sequences. Another method is combinatorial approaches of secondary structure prediction and sequence alignments as, for example, the available 3D-PSSM (Kelley *et al.*, 2000). Thereby the secondary structure prediction of the target protein is compared regarding the true and predicted secondary structure of all proteins with a known fold (classified in the SCOP database (Murzin *et al.*, 1995)). These comparisons of secondary structure predictions combined with data from sequence alignments result in far-related protein templates that can be used.

To estimate the information content of a template molecule it is necessary to again compare its sequence against the target. Furthermore, sequences of related proteins have to be added, as for example sequences of the target protein from different species or proteins with similar functions and at least partially a common fold. All of these sequence information used for a multiple sequence alignment, *e.g.* by ClustalW (Thompson *et al.*, 1994a; Thompson *et al.*, 1994b), results in insights about highly and less conserved areas of the target protein. The highly conserved regions are usually related in functionally or structurally important parts of the target protein and therefore have to be found into the template structure. Regions of less conservation are usually not connected to functional or structural importance. Hence, they can be more easily modeled without a template structure. Furthermore these multiple sequence alignments allow conclusions about the evolutionary background of protein families by the use of phylogenetic trees (Feng and Doolittle, 1987).

2.3 Model Generation and Optimization

As already mentioned above, the earliest model generations were made using side chain-substitutions to construct the target's properties on the backbone of template structures. Recently, this strategy was only used for homologous proteins sharing high sequence similarity. In case of GPCRs, a high similarity is given in 7 areas of about 20 to 30 residues that are mainly hydrophobic – the 7 transmembrane helices. Consequently these regions were assembled this way.

The loop structures connecting the transmembrane regions vary in many GPCRs, and in especially compared to the template structure of bovine rhodopsin. Nevertheless, the structural information of loop formation is partly given in the rhodopsin template. This information can be used for loop regions of at least similar length. Loops of different length or where no useful structure information of rhodopsin is available (*e.g.* the fragmentary structure of intracellular loop ICL3) had to be generated another way. Most modeling software packages offer tools for the design of loop structures. In Sybyl6.8 the LOOP SEARCH tool uses a fragment database, which is screened for homologous sequences (of proteins that are deposited in the PDB). In addition to the homology of the target sequence and the fragments, the end-to-end distances between the predecessor and successor residues in the model will be calculated, as well as chances for clashes with other regions of the model structure. This results in an ensemble of loop structures, which have to be analyzed for best fit and interactions with the rest of the model structure. This way, also the N- and C-terminus can be added. In case of larger regions, additional searches against other protein templates are more suitable approaches.

The resulting primary model has to be inspected for failures. Besides the unusual trans-peptide bonds that are frequently the result of automatic loop generations, amino acids with R-configuration or side chain-side chain collisions may have occurred. Additionally, the substitutions of residues lead to a changed packing of aromatic rings and new possibilities for electrostatic contacts. In many cases corrections of these side chain-conformations have to be done by hand. R-configuration and clashes of residues are easily to correct by inversion of Cα and rotations around side chain-torsions, respectively. For repair of unusual backbone torsions (*e.g.* trans-peptide bonds) manual inspection is insufficient. The identification of trans-peptide bonds, clashes and R-configured amino acids as well as recognition of unusual torsion angles can be done by utilities such as the commonly used PROCHECK (Laskowski *et al.*, 1996). It lists several different properties of the protein: the backbone torsion angles φ and ϕ, the side chain-torsions χ_1 and χ_2, planarity of aromatic rings, etc. For comparison of backbone torsion

angles a Ramachandran-plot is used. This plot, which includes every residue without the N-terminal, is generated by φ and ϕ each ranging from –180° to +180°. Based on further studies (Morris *et al.*, 1992) the conformational angles show preferences for values that are expected based on simple energy considerations. The lowest energy is given in the three core regions of the plot, followed by the allowed regions. The disallowed regions are of high energy and therefore unlikely. Most pairs of φ and ϕ will be in the core and favored regions of the plot and only a few will be in the "disallowed" region. Because deviations from these angles may be indicators for potential errors in crystallographic or modeling projects, modeled structures should have at least 90% of their residues in the core region.

The repair of the primary model is followed by certain steps of energy minimization and/or simulated annealing protocols.

2.4 Molecular Dynamics

Molecular dynamics (MD) is generally not used in the prediction of structure or modeling the protein-folding pathway. The most common use of this technique is the simulation of dynamical changes in known structures (*e.g.* in interaction with other molecular structures). Thereby, Newtonian mechanics is applied on molecular systems. In MD simulations force fields, or quantum chemical models, or a mixture of the two is used to simulate the interactions of atomic particles (independently if they are single, in small molecules or larger structures as *e.g.* proteins). In general the following principles are applied:

- Atomic particles are spherical with fixed radii and an assigned net charge
- Interactions are based on springs and classical potentials
- Interactions have to be pre-assigned to specific sets of atoms (bond definition, etc.)
- Interactions determine the spatial distribution of atomic particles and their energies

The potential function computes molecular potential energy as a sum of energy terms describing the deviations of bond length, bond angles and torsions angles in comparison to the equilibrium values (see Equation 3). Additionally, the terms for non-bonded pairs of atoms are applied (see Equation 4).

$$E = E_{bonds} + E_{angle} + E_{dihedral} + E_{non\text{-}bonded} \qquad \text{(Equation 3)}$$

$$E_{non\text{-}bonded} = E_{electrostatic} + E_{van\ der\ Waals} \qquad \text{(Equation 4)}$$

All of these parameters taken together are known as the force field. The backgrounds of such force fields are usually quantum chemical calculations and/or experimental investigations. Minimization of the potential function is known as the Energy Minimization technique. As already mentioned above this technique is used as an optimization of the protein structure to find the local minimum starting from an initial conformation. Energy minimizations result in an optimized arrangement of electrostatic interactions, hydrogen bonds and van der Waals-contacts (based on the initial structure). Although its result is only the finding of a local minimum, this procedure is useful because the main structural features – based on the homology modeling before – are still intact. Physically, this optimization corresponds to an instantaneous freezing of the system.

The potential function as a function of time result is used in Molecular Dynamics simulations by solving Newton's equation of motion (see Equation 5). The integration of Newton's laws of motion with different integration algorithms leads to atomic trajectories, describing the

movement of atoms (and molecules) in space and time.

$$F = ma \qquad \text{(Equation 5)}$$

This technique is also useful for the purpose of generating the global minimum of a protein structure – called Simulated Annealing. The general protocol uses several cycles of heating (molecular dynamics) and cooling (energy minimization) of the protein. Because temperature in MD simulations (such as simulated annealing) is a direct measure of the movement of atoms, the range of temperature allows direct influence on effects of the simulation, *e.g.* prohibition of cis/trans isomerization by use of lower temperatures. After several steps at high temperature, the cool down – the energy minimization – results in reduced atomic movements and in finding new electrostatic and hydrophobic contacts. Every annealing generates a local minimum of the protein, whose structures may vary remarkably from the initial state. By comparison of all minimized structures the global minimum (or at least a local minimum close to it) of the model structure can be found. Very often certain protein fold subfamilies based on similar energies are generated that are very helpful in understanding of mechanisms of the protein.

Furthermore, the introduction of additional non-bonded interactions – called restraints – makes it possible to move atoms rather directly in the simulation. Independent of the used strategy is a pure molecular dynamics procedure or a simulated annealing simulation that can be applied for docking of ligands into proteins or for generation of protein-protein interactions.

2.5 Modeling and Studies of G Protein-Coupled Receptors

In case of this study only one suitable template – the X-ray crystallographic structure of bovine rhodopsin – was available. But because of the sequence studies of all GPCRs, which have been done before (Attwood and Findlay, 1994; Fredriksson *et al.*, 2003), a common building plan as the base of the studied receptors (ETA, ETB, GPR109A, GPR109B) and bovine rhodopsin was given despite the rather small overall sequence identity between ET receptors and rhodopsin (about 22%) or GPR109 receptors and rhodopsin (about 25%).

2.5.1 Endothelin Receptors

Sequence alignments of endothelin receptors of different species sorted by their subtypes ETA and ETB and in alignment with sequence of bovine rhodopsin were made using SeqLab [Wisconsin Package Version 10.2, Accelrys (GCG). San Diego, CA].

Structural models of human ETA and human ETB are based on the X-ray crystallographic structure of bovine rhodopsin using entry 1HZX (Teller *et al.*, 2001) from the PDB (Berman *et al.*, 2000). Models were generated by side chain substitutions in homologous transmembrane regions using SYBYL6.8 [Tripos Inc., St. Louis, MO]. Short loops were added by best fit and homology to fragments of other proteins using LOOP SEARCH tool implemented in SYBYL6.8. The orientation of ECL2 was kept in a sheet-like fold as in the rhodopsin template.

The following remarks should be given about the models:

- The N-terminus was truncated to Cys69 (ETA) and Cys90 (ETB). The C-terminus was truncated to Cys386 (ETA) and Cys403 (ETB).
- Disulfide bridges between the N-terminal tail and ECL3 (Cys69 – Cys341 in ETA / Cys90 – Cys358 in ETB) as well as the conserved one between TMH3 and ECL2 (Cys158 – Cys239 in ETA / Cys174 – Cys255 in ETB).
- Helices TMH5, TMH6 and TMH7 were N-terminally prolonged and helices TMH5 and TMH6 were C-terminally prolonged, too, due to sequence specific differences in the rhodopsin template.
- A proline kink in TMH2 based on the more homologous structure fragment of transmembrane helix TMH6 of sensory rhodopsin II from PDB-entry 1JGJ (Luecke *et al.*, 2001) lead to altered orientation of about 20 degrees outwards compared with the structure of bovine rhodopsin.

After model generation the structures were minimized using the Amber4.1 force field

(Cornell *et al.*, 1995) and Amber95_Protein_ALL charges in SYBYL6.8.

Optimal ligand docking was achieved by a constrained MD simulation procedure, initially performed *in vacuo* with AMBER7 (Case *et al.*, 2002). Experimentally known functionally sensitive side chains were used as anchor points for the few ligand-receptor interaction restraints (see Tab. 2-1). Starting with the empty receptor, the ligand was forced into the binding pocket. Binding procedures were 200fs with harmonic potentials on Cα atoms of receptor residues in transmembrane helices and restrained torsion angles of ligand residues in the helical part. Thereafter the stability of the formed complexes was checked by 2ns of MD in water-vacuum-water box (ter Laak and Kuhne, 1999) without any restraints.

Structural data of the ligands were taken from PDB entries 1EDN (Janes *et al.*, 1994), 1EDP (Andersen *et al.*, 1992), 1V6R (Takashima *et al.*, 2004) and 1SRB (Atkins *et al.*, 1995). Sequence alignments of all peptide ligands were made using SeqLab [Wisconsin Package Version 10.2, Accelrys (GCG), San Diego, CA].

Table 2-1: Data from site-directed mutagenesis experiments on ET receptors taken from literature, and used for MD supported ligand docking as constraints. The effects on binding of these mutations are listed as increased (↑), reduced (↓) or unchanged (→)

Ballesteros' Numbering	Mutations Sites		Effects on Ligand Binding			References	Restraint to ligand property
	hETA	HETB					
1.49	Gly97		ET-1 ↓			(Breu *et al.*, 1995)	
2.53	Tyr129		ET-1 →	ET-2 →	ET-3 ↑	(Lee *et al.*, 1994b; Webb *et al.*, 1996)	
		His150	ET-1 →	ET-2 →	ET-3 →	(Lee *et al.*, 1994b)	
2.64	Lys140		ET-1 ↓			(Adachi *et al.*, 1994a; Adachi *et al.*, 1994b; Breu *et al.*, 1995)	Asp18
		Lys161			ET-3 ↓	(Adachi *et al.*, 1994b)	Asp18
3.26	Lys159		ET-1 ↓			(Breu *et al.*, 1995)	
3.32	Gln165		ET-1 ↓			(Breu *et al.*, 1995)	
3.33		Lys182	ET-1 →	ET-2 ↓	ET-3 ↓	(Lee *et al.*, 1994a)	C-term
5.40	Trp257		ET-1 →			(Imamura *et al.*, 2000)	
		Trp275	ET-1 →			(Imamura *et al.*, 2000)	
5.41	Trp258		ET-1 →			(Imamura *et al.*, 2000)	
		Trp276	ET-1 →			(Imamura *et al.*, 2000)	
5.46	Tyr263		ET-1 →			(Breu *et al.*, 1995)	
6.31		Arg319	ET-1 →			(Abe *et al.*, 2000; Fuchs *et al.*, 2001)	
6.44	Phe315		ET-1 ↓			(Breu *et al.*, 1995)	Trp21
6.55	Arg326		ET-1 →			(Breu *et al.*, 1995)	
7.35	Asp351		ET-1 →			(Breu *et al.*, 1995; Vichi *et al.*, 1999)	

2.5.2 Nicotinic Acid Receptors

To generate a structural model for GPR109A/B, we adopted the X-ray structure of rhodopsin (Teller *et al.*, 2001) from entry 1HZX of the Protein Data Bank PDB (Berman *et al.*, 2000) as a template. Several receptor-specific modifications of the template structure were made based on sequence alignment investigations using SeqLab [Wisconsin Package Version 10.2,

Accelrys (GCG). San Diego, CA]. At the N-terminal tail, the two consecutive cysteines Cys18 and Cys19 form two additional disulfide bridges towards extracellular loop (ECL) 2 (Cys18-ECL2:Cys183) and ECL3 (Cys19-ECL3:Cys266), respectively. For transmembrane helix TMH2 a structural "bulge" of Rhodopsin caused by side chain/backbone interactions of three consecutive threonine residues in the rhodopsin structure would localize Asp85 and Asn86 at the membrane-oriented phase of the helix. New construction of the junction of TMH2 and ECL1 avoids the bulge structure of rhodopsin and considers a proline kink from other TMH structures (Sansom and Weinstein, 2000). In this case a part of TMH6 of sensory rhodopsin II from *Natronobacterium pharaonis* (Luecke *et al.*, 2001) was used for structural refinement. The resulting new conformation is similar to that of chemokine receptor models: the proline of TMH2 is located on the same sequence position as in GPR109A and GPR109B. In addition, a minor change in orientation (10 to 15 degree-twist) of the N-terminal half of TMH5 was generated before the proline kink at Pro200 as a result of different residues compared with the rhodopsin template. The length of ECL3 was extended by an additional helix turn at TMH6 because of more residues in GPR109A/B than in the template structure. Gaps of missing residues in the intracellular loops of the rhodopsin structure were closed using the LOOP SEARCH tool implemented in SYBYL6.8 [Tripos Inc., St. Louis, MO] using GPR109A/B sequence. Concerning ECL2, we started with two different models: one with the original rhodopsin fold in counter-clockwise order of residues around Cys177 and a second model with reversed, clockwise order of ECL2 residues around this cysteine residue by preserving the β-strand motifs. This results in a suitable geometry for pairing additional disulfide bridges. After model generation, the structures were minimized using the Amber4.1 force field and Amber95_Protein_All charges (Cornell *et al.*, 1995). In a first step, the ligands were manually docked according to the potential interaction points suggested by the results obtained with receptor mutants. In a second step, molecular dynamic simulations using Amber7 (Case *et al.*, 2002) in a water-vacuum-water box system (ter Laak and Kuhne, 1999) without any restrains, 1ns duration, periodic boundary conditions, and charges neutralized by adding chlorine ions studied the stability of the ligand in the binding site. The quality of the models and stability was validated by checking the geometry by PROCHECK (Laskowski *et al.*, 1996) and the stability during the molecular dynamics run (overall backbone root mean square deviation, 1.7 Å).

2.5.3 G Proteins

Structural information about considered molecules were taken from 3D databases and modeling. Structures of Gα-subunits Gαi and Gαs are based on entries 1GIA (Coleman *et al.*, 1994), 1GG2 (Wall *et al.*, 1995), 1GP2 (Wall *et al.*, 1995) and 1AZT (Sunahara *et al.*, 1997), 1AZS (Tesmer *et al.*, 1997), 1CJK (Tesmer *et al.*, 1999), respectively, of the Brookhaven Protein Data Bank PDB (Berman *et al.*, 2000). By assembling the NMR-structure of the rhodopsin-bound C-terminal peptide of bovine transducin Gα-subunit from PDB entry 1AQG (Kisselev *et al.*, 1998) these models were additionally refined.

Homology models for Gα-subunits Gαo and Gαq were generated by side chain replacement based on structural information received from the PDB (see above). Necessary multiple sequence alignments for model generation were made using SeqLab [Wisconsin Package Version 10.2, Accelrys (GCG). San Diego, CA]. Energy minimizations for these models were made using AMBER95_PROTEIN_ALL charges in the Amber4.1 force field (Cornell *et al.*, 1995).

Structural information of Gαi-interacting mastoparan-X (MPX) is based on PDB entry 1A13 (Kusunoki *et al.*, 1998). Information about conformations of Gαs-bound mastoparan-S (MPS) was taken from Sukumar and co-workers (Sukumar *et al.*, 1997).

Bradykinin structural information was taken from Mierke and collaborators (Pellegrini *et al.*, 1997; Pellegrini and Mierke, 1997).

Taking advantage of different distances of positive charges found at secretagogues, manual docking studies of the receptor mimetic peptides (RMPs) MPX, Bradykinin and Substance P were performed for compound models at Gαi and for MPS at Gαs. Optimizations of the RMP G-protein complexes were made by energy minimizations of potentially involved side chains using Amber95 force field with AMBER95_ALL charges (Cornell *et al.*, 1995).

Additional smaller peptidomimetics: alkyl-substituted amino acid-derivatives – termed lipoamines – were used from a compound library of Schunack and co-workers (Breitweg-Lehmann *et al.*, 2002; Leschke *et al.*, 1997). All structural models of these compounds were generated within SYBYL6.8 [Tripos Inc., St. Louis, MO] and minimized using Tripos force field (Clark *et al.*, 1989) with charges by Gasteiger and Marsili (Gasteiger and Marsili, 1978).

Scanning the conformational space of lipoamines was done by different search routines implemented in SYBYL (random search, systematic search) depending on the number of rotatable bonds in the head group (amino acid derivative). In most cases the systematic search could not be used because of the expected high output (simplifying search steps of 30 degrees were used; leading to 12^n possible conformations for a number of n bonds). Energy

calculations of the structures of these searches were made using a minimization setting with Tripos force field (Clark *et al.*, 1989), charges by Gasteiger and Marsili (Gasteiger and Marsili, 1978) and a dielectric constant of 80.0 AsV-1m-1 which is similar to those of water. The hydrophobic tail (fatty acid derivative), expected to interact with the membrane, was defined as an aggregate in these minimization set-ups. Conformations where the charges are directed towards a putative membrane were removed from considerations. Next the search results were ordered by their minimum of potential energy and investigated by their distances of positive charges and compared with existing data.

The collected distances of positive charges sorted by the G protein-selectivity of the according compounds were used in finding charge patterns at the comparative models of ETA and ETB receptors.